Article:
Qiu, Z., Xu, Y., Chen, C. et al. (2024) Enhanced disease detection for apple leaves with rotating feature extraction. Agronomy, 14 (11). 2602. ISSN: 2073-4395

https://doi.org/10.3390/agronomy14112602

agronomy

MDPI

Article

Enhanced Disease Detection for Apple Leaves with Rotating Feature Extraction

Zhihui Qiu [1], Yihan Xu [1,*], Chen Chen [1], Wen Zhou [2] and Gang Yu [3]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; qiuzhihui@njfu.edu.cn (Z.Q.); chenchen@njfu.edu.cn (C.C.)
[2] School of Low-Altitude Equipment and Intelligent Control, Guangzhou Maritime University, Guangzhou 510725, China; wenzhou@ustc.edu
[3] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S10 2TN, UK; gyu2@sheffield.ac.uk
* Correspondence: xuyihan@njfu.edu.cn

Abstract: Leaf diseases such as Mosaic disease and Black Rot are among the most common diseases affecting apple leaves, significantly reducing apple yield and quality. Detecting leaf diseases is crucial for the prevention and control of these conditions. In this paper, we propose incorporating rotated bounding boxes into deep learning-based detection, introducing the ProbIoU loss function to better quantify the difference between model predictions and real results in practice. Specifically, we integrated the Plant Village dataset with an on-site dataset of apple leaves from an orchard in Weifang City, Shandong Province, China. Additionally, data augmentation techniques were employed to expand the dataset and address the class imbalance issue. We utilized the EfficientNetV2 architecture with inverted residual structures (FusedMBConv and S-MBConv modules) in the backbone network to build sparse features using a top–down approach, minimizing information loss. The inclusion of the SimAM attention mechanism effectively captures both channel and spatial attention, expanding the receptive field and enhancing feature extraction. Furthermore, we introduced depth-wise separable convolution and the CAFM in the neck network to improve feature fusion capabilities. Finally, experimental results demonstrate that our model outperforms other detection models, achieving 93.3% mAP@0.5, 88.7% Precision, and 89.6% Recall. This approach provides a highly effective solution for the early detection of apple leaf diseases, with the potential to significantly improve disease management in apple orchards.

Keywords: apple leaf disease; object detection; rotated bounding box; attention mechanism; depth-separable convolution

Citation: Qiu, Z.; Xu, Y.; Chen, C.; Zhou, W.; Yu, G. Enhanced Disease Detection for Apple Leaves with Rotating Feature Extraction. *Agronomy* **2024**, *14*, 2602. https://doi.org/10.3390/agronomy14112602

Academic Editor: Marco Scortichini

Received: 4 October 2024
Revised: 1 November 2024
Accepted: 1 November 2024
Published: 4 November 2024

1. Introduction

The apple industry is the largest fruit industry in China, with its production volume and cultivation area ranking first in the country's fruit production. The apple industry plays a vital role in China's economic development and holds an irreplaceable, leading position in the economic growth of certain rural regions [1]. As the apple cultivation area continues to expand, sudden and large-scale disease outbreaks are becoming increasingly frequent [2]. The delayed diagnosis and treatment of diseases can result in substantial economic losses. The effective early management of pests and diseases requires collaboration at multiple levels, with farmers playing a key role in regularly monitoring and diagnosing apple diseases. By identifying diseases early and accurately, and implementing timely control measures, economic losses for fruit farmers can be minimized [3].

With the ongoing advancements in computer vision technology, research into the identification and detection of apple leaf diseases continues to progress. Techniques from image processing and machine learning have been widely employed in the detection of apple leaf diseases [4–6]. These methods primarily involve lesion segmentation and feature

extraction for classification and have achieved significant progress. Mohammed et al. [7] optimized neural networks using genetic algorithms and employed support vector machines for crop disease detection. The detection accuracy for tomato leaf disease was 98.1%. This approach successfully eliminated unnecessary features, significantly improving the detection rates and effectiveness for various diseases. Radha et al. [8] employed support vector machines (SVMs) as classifiers and applied preprocessing techniques, including image loading and contrast enhancement, to detect various plant diseases, achieving an accuracy of 90.6%. Kaur [9,10] proposed a recognition method for grape leaf diseases based on fractional-order Zernike moments (FZMs) combined with an SVM. The performance of this classifier outperformed the others, achieving a classification result of 97.34% at an order of 30. Zou et al. [11] presented a method for recognizing tea diseases relying on spectral reflectance, which included a decision tree-based feature selector and random forest-based tea disease recognizer. The experimental results demonstrated improvements in the recognition accuracy and recall rate. The adulteration of Laochuan tea tree manna was quantitatively analyzed, and the accuracy of the test set was 0.968 [12]. Zhang et al. [13] successfully extracted 38 different color, texture, and shape features by integrating the HSI, YUV, and grayscale models and utilized support vector machine technology to identify three apple diseases. The experimental results demonstrated that this approach achieved an accuracy rate exceeding 90%.

In conventional machine learning approaches, the accuracy of handcrafted feature extraction greatly affects the accuracy of identifying plant leaf diseases and detection [14]. The feature extraction process is both resource-intensive and prone to subjectivity. As artificial intelligence continues to advance, deep learning has demonstrated its power to replace human intelligence [15]. Deep learning excels at learning features that surpass human capabilities, extracting features with richer semantic information than manually designed features [16]. Recently, deep learning has gained widespread use in plant disease detection, surpassing the performance of traditional handcrafted feature techniques. Turkoglu et al. employed three networks, namely AlexNet, GoogleNet, and DenseNet201, to pretrain a multimodal LSTM model. The deep features obtained were fed into the LSTM layers, and the results from the three LSTM layers were classified using a majority voting classifier [17]. Testing on two groups of apple disease images demonstrated that the proposed algorithm achieved good recognition performance. Yu et al. [18] developed a novel two-layer structured model for detecting and identifying diseases in apple leaves. This model organically integrates a classification subnet and a feature extraction subnet, allowing the more accurate extraction of leaf features through training, which achieves 99.94% accuracy on Plant Village. Gao et al. [19] suggested a method utilizing an optimized lightweight YOLOv4 neural network for counting maize seedlings, with an accuracy rate of 96.25%. The method first employed GhostNet for feature extraction and introduced attention mechanisms and the K-means clustering algorithm to enhance the detection accuracy of maize seedling quantity. Then, depth-wise separable convolutions were employed in place of standard convolutions to reduce the network's complexity and make it more lightweight [20]. In the final step, an optimized multi-scale feature fusion network structure was employed to further minimize the total number of model parameters. The experimental results demonstrated the accurate identification of maize seedling quantity using this framework [21].

However, apple leaf diseases exhibit different morphologies and susceptibility to infection, and disease targets often have irregular shapes. There is also similarity between different diseases in local regions.

If the disease positions are not accurately localized during detection, it can easily lead to false positives or false negatives [22]. Accurately identifying the position of the disease is critical for improving the confidence of detection results and enhancing the robustness of the model [23]. To address this, we ensured that high-quality images and precise disease position annotations were used throughout the model training phase. However, it is designed to be robust, meaning it can still effectively detect diseases even in cases where the exact positions are not perfectly marked. This robustness is achieved by enabling

the model to learn broader disease patterns and features beyond just specific localized positions. By doing so, the model can minimize false positives and false negatives while maintaining high accuracy even under less-than-perfect conditions.

This study introduces a detection method for apple leaf diseases utilizing rotating bounding boxes. Then, a Probabilistic IoU (ProbIoU) method for calculating the target similarity is proposed, taking into account the characteristics of the 2D Gaussian distribution [24]. It meets all measurement standards and can represent the real distance between different distributions, effectively improving the accuracy of the oriented bounding box detection. Second, an effective feature extraction backbone combined with SimAM is introduced to capture long-range dependencies and correlations between neighboring regions, enhancing the model's feature extraction capability. Simultaneously, we optimize the feature fusion mode by using global context information and local features to boost detection precision in the neck network.

2. Materials and Methods

2.1. Dataset Setting

2.1.1. Data Acquisition and Annotation

The study area is located within Weifang City (36°.37′ N, 119°, 74′ E), Shandong Province, China. We utilized images from an apple orchard in Weifang City, where more than 40% of the vegetation is made up of apple trees, as shown in Figure 1. The image data used in this study were captured using an aerial drone equipped with a high-definition camera (DJI MINI4). Since the camera's perspective can cause variations in the number and size of targets, the images were collected from multiple angles and distances. The collected images were then filtered to include those representing different stages of disease progression for use in this study. This area is an important apple-growing region and is representative, offering valuable insights for other apple-growing areas.

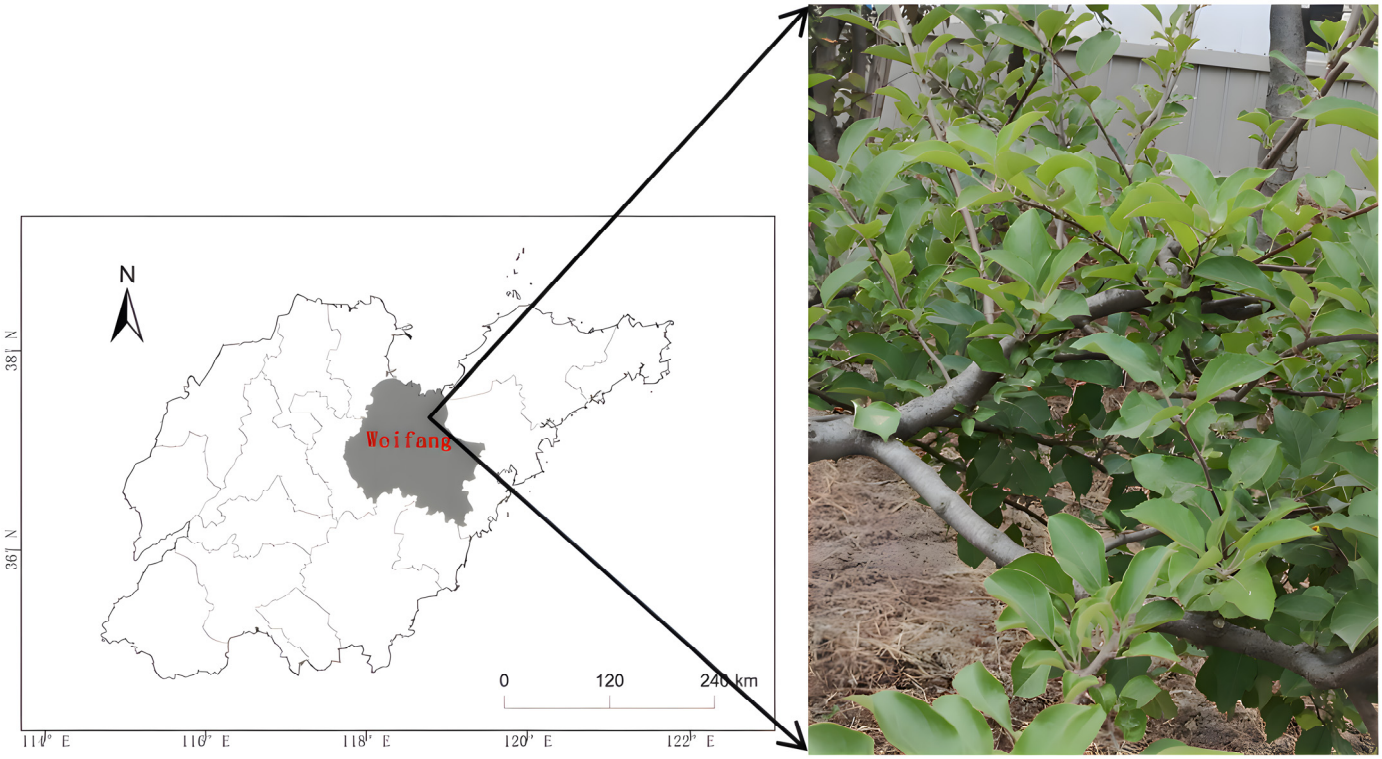


Figure 1. Study area location and corresponding natural color image.

The effectiveness of using Convolutional Neural Networks (CNNs) for apple leaf disease detection is strongly dependent on the dataset's quality and size [25]. With a high-quality dataset, deep learning models can learn more features and exhibit excellent

generalization capabilities. This study selected healthy leaf samples and seven common apple leaf diseases (Alternaria Blotch, Black Rot, Brown Spot, Gray Spot, Mosaic, Rust, Scab) as research subjects [26]. These diseases are widely distributed in apple-growing regions, have high incidence rates, and can survive and spread through multiple channels under different climatic conditions, affecting nearly all major apple-producing areas. Additionally, the characteristics of these diseases are relatively distinct and can be preliminarily identified through changes in the leaves, which facilitates early detection and control. Some of the data came from Plant Village, while others were from the apple orchard we photographed [27]. During the dataset integration process, we applied certain screening criteria based on image clarity, disease typicality, and distribution uniformity. The collected disease images served as experimental data to validate the practicality of the detection method. Images with blur from the initial dataset were removed to prevent interference. When conducting data sample screening, we ensured that the dataset includes images captured under different climatic conditions and growing environments, striving for sample diversity. Such a diverse dataset helps improve the robustness of the model, making it more adaptable in practical applications. After screening, a final set of 3000 images was selected, covering various angles and environmental backgrounds to ensure the model's generalization ability. Figure 2 presents some representative samples.

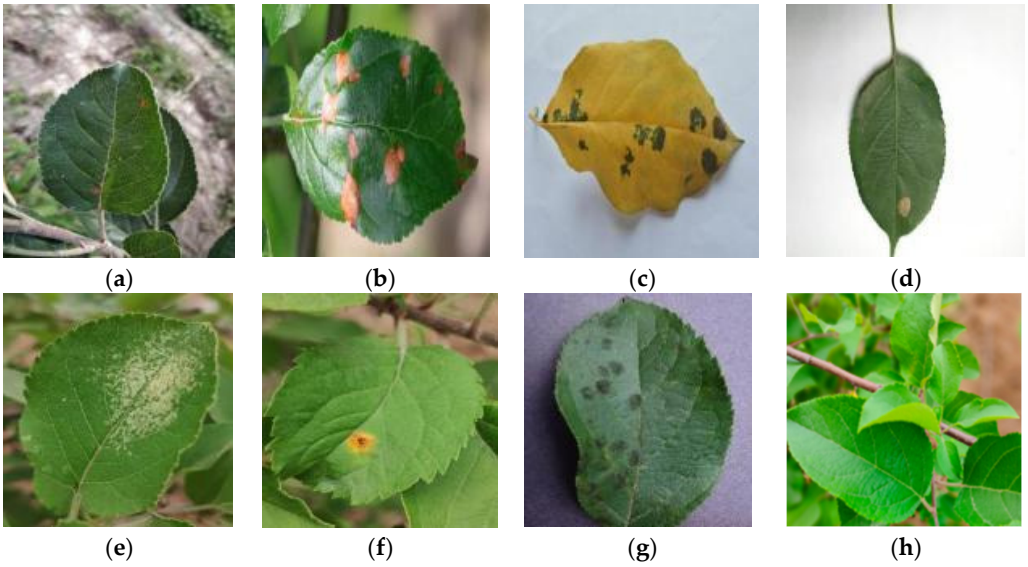


Figure 2. Sample images from the dataset: (**a**) Alternaria Blotch, (**b**) Black Rot, (**c**) Brown Spot, (**d**) Gray Spot, (**e**) Mosaic, (**f**) Rust, (**g**) Scab, and (**h**) healthy apple leaves.

2.1.2. Data Augmentation

Alternaria Blotch showed a gradual transformation from the initial brown round spots to reddish-brown spots, as shown in Figure 2a. During the initial phases of Black Rot, the lesions appear round, measuring 2–3 mm, and are purple. As the disease progresses, they turn dark brown, with a depressed center and raised edges. The middle of the lesion becomes dark gray, densely populated with small black dots, as shown in Figure 2b. The lesions of Brown Spot can be divided into the concentric wheel pattern, needle-hair type, and mixed type. The plaques of the concentric wheel pattern are dark brown round spots with small black spots arranged in a wheel pattern. Needle-awning type lesions are needle-awning outward spread. Mixed-type lesions are widespread, dark brown, and round or irregular in shape, as shown in Figure 2c. The spots of Gray Spot first appear as yellowish-brown round spots with clear edges and then turn grayish-white, as shown in Figure 2d. Mosaic disease spots are mainly divided into Mosaic type, along-the-vein discoloration type, stripe type, and ring type. The Mosaic pattern is present with yellow spots. The vein discoloration pattern is netted and yellowed along the vein, often accompanied by yellow spots. The streaky lesions yellow along the vein of the leaf and extend to the nearby

mesophyll tissue, and the discolored area is wider. The ring pattern has a bright yellow ring or similar ring pattern, but the inner ring is still green, as shown in Figure 2e. Rust disease manifests on leaves as small spots that are yellow or orange. These spots are round or semi-circular and form raised lesions on the leaf surface, as shown in Figure 2f. Scab disease presents as small, round dark brown spots. The color of the lesions gradually changes to black or dark brown, with edges that are typically sharp and ring-like or irregular in shape, as shown in Figure 2g.

After integrating as many apple leaf disease datasets as possible, the dataset's sample count is still limited. Such small samples are more prone to overfitting during subsequent training. To tackle this problem, we employ data augmentation to increase the sample size by generating new samples that follow the same distribution as the original samples. This is a straightforward and effective approach, particularly in preventing overfitting. In this study, before dividing the dataset, we randomly adjust the hue, saturation, and brightness of the sample batches to introduce variability. Additionally, we further increase the sample count by applying geometric transformations such as scaling, flipping, and certain filtering operations.

Each original image is augmented to four times its initial size, resulting in a total of 13,280 images. The dataset was split into training, test, and validation sets in a 7:2:1 ratio, finally forming a dataset that can be used for training and testing. We took particular care to ensure that the augmentation process maintained a balanced representation of each leaf disease, as shown in Table 1. The distribution of augmented images reflects the original dataset's class proportions, thereby reducing the risk of any class imbalance. Furthermore, to further mitigate overfitting, we also considered using additional techniques such as weight regularization and early stopping during the model training phase.

Table 1. Overview of the apple leaf detection dataset.

Type	Original Quantity	Enhanced Quantity
Alternaria Blotch	267	1068
Black rot	225	900
Brown Spot	293	1172
Gray Spot	258	1032
Mosaic	410	1640
Rust	403	1612
Scab	429	1716
Healthy	1035	4140
Total	3320	13,280

2.2. Gaussian Bounding Box and Rotation of Object Detection Box

2.2.1. Gaussian Bounding Box and Probabilistic IoU Method

Object detection represents a core challenge in the field of computer vision, and remarkable advancements have been achieved in recent years owing to deep learning methods [28]. Nevertheless, the majority of object detection approaches utilize horizontal bounding boxes to encode the shape and position of objects, which is inadequate when objects exhibit non-aligned rectangular shapes. This limitation is particularly apparent in the case of apple leaf diseases, where the irregular shapes of diseases and the similarity in textures between different diseases pose significant challenges. Failure to accurately locate the target can easily result in the misdiagnosis of the disease.

In order to determine a two-dimensional Gaussian distribution in a two-dimensional region, its mean μ and covariance $\sum$ need to be calculated. Among them, $\mu = (x_0, y_0)^T$. In the object detection task, you can directly set the parameters of the regression task in the object detection as (x_0, y_0, a, b, c), or you can express the parameters of the regression task

as $(x_0, y_0, a', b', \theta)$, and the latter is more commonly used in object detection. $\sum$ can be expressed as Equation (1):

$$\sum = \begin{bmatrix} a & c \\ c & b \end{bmatrix} = r_{m\theta} \begin{bmatrix} a' & 0 \\ 0 & b' \end{bmatrix} {r_{m\theta}}^T = \begin{bmatrix} a'cos^2\theta + b'sin^2\theta & \frac{1}{2}(a' - b')\sin 2\theta \\ \frac{1}{2}(a' - b')\sin 2\theta & a'\sin^2\theta + b'\cos^2\theta \end{bmatrix}, \quad (1)$$

where $r_{m\theta}$ is a two-dimensional rotation matrix. We begin by assuming that the object region is represented as a continuous two-dimensional binary region Θ, which provides a comprehensive description of its shape. It is assumed that it follows a uniform probability density function, along with the distribution's mean, and the covariance matrix can be calculated as Equation (2):

$$\mu = \frac{1}{N} \int\limits_{x \in \Theta} x, \ \sum = \frac{1}{N} \int\limits_{x \in \Theta} (x - \mu)(x - \mu)^T, \quad (2)$$

where N represents the area of Θ. For horizontal bounding boxes, Θ reduces to a rectangular region whose center is (x_0, y_0). The width of the region is W, and the height is H. In this situation, $\mu = (x_0, y_0)^T$ is simply the center of the rectangular region. $\sum$ can be expressed as Equation (3):

$$\sum = \frac{1}{WH} \int\limits_{-H/2}^{H/2} \int\limits_{-W/2}^{W/2} \begin{bmatrix} x^2 & xy \\ xy & y^2 \end{bmatrix} dxdy = \frac{1}{12} \begin{bmatrix} W^2 & 0 \\ 0 & H^2 \end{bmatrix} \quad (3)$$

For oriented bounding boxes, we can define uncorrelated variances a' and b' based on the sides of the corresponding axis-aligned horizontal bounding boxes and also work out the orientation angle θ. In this case, the covariance matrix $\sum$ can be calculated by Equation (1).

At the beginning, we use the Bhattacharyya Distance to calculate the similarity between different Gaussian bounding boxes. The Bhattacharyya Coefficient B between two probability density functions $\ell(x)$ and $\beta(x)$ is defined as Equation (4):

$$B = \int\limits_{\Re^2} \sqrt{\ell(x)\beta(x)}dx, \quad (4)$$

where $B \in [0, 1]$. If the two distributions are the same, $B = 1$. Based on the above, the Bhattacharyya Distance Bh_D between $\ell(x)$ and $\beta(x)$ can be expressed as Equation (5):

$$Bh_D(\ell, \beta) = -\ln B(\ell, \beta) \quad (5)$$

However, the Bhattacharyya Distance is not considered a true distance metric because it does not adhere to the triangle inequality property. To measure the true distance, the Hellinger Distance H_D is utilized, given by Equation (6):

$$H_D(\ell, \beta) = \sqrt{1 - B(\ell, \beta)}, \quad (6)$$

where $H_D(\ell, \beta) \in [0, 1]$. If the two distributions are the same, $H_D = 0$. In this paper, a method called PIoU is proposed to calculate the similarity between Gaussian distributions. The specific formula for calculating PIoU is Equation (7):

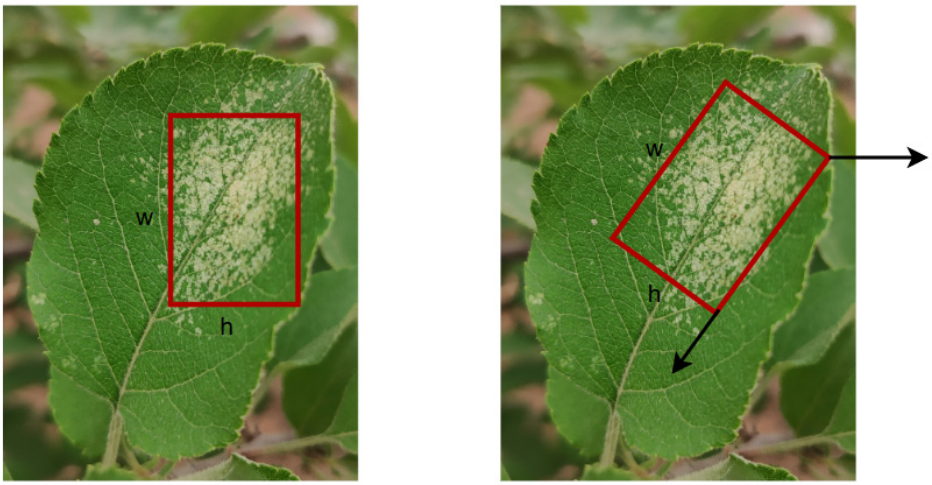
$$PIoU = 1 - H_D(\ell, \beta) \quad (7)$$

2.2.2. Rotation of Object Detection Box

For the diagnosis of diseases in apple leaves, where the targets can have arbitrary orientations, rotated bounding boxes offer a more effective solution than horizontal bounding

boxes [29]. To achieve a more accurate detection of apple leaf diseases, an adaptive thresholding strategy was applied during the training process to select positive and negative samples. This strategy also involves the regression, classification, and prediction of the rotation angle for actual targets.

The conventional horizontal bounding boxes represent the targets mainly with four parameters: the center coordinates (x, y) of the bounding box and its width (w) and height (h). However, for apple leaf disease targets with arbitrary orientations, this representation tends to include a significant amount of background and results in a high level of overlap between closely located bounding boxes. Hence, for apple leaf disease targets, we employ the long-side definition method in the rotated bounding box representation to select the targets. This method primarily consists of five parameters (x, y, w, h, θ). The angle between the longest side w of the rotating frame and the X-axis is θ, $\theta \in [\frac{-\pi}{4}, \frac{3\pi}{4})$. When the long side is above the X-axis, the angle is negative, and when it is below the X-axis, the angle is positive. This representation method can represent the target of an apple leaf disease image with any direction well. The horizontal bounding box and the rotated bounding box are shown in Figure 3a,b, respectively.



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

Figure 3. Methods for representing object detection bounding boxes: (**a**) horizontal box representation; (**b**) oriented box.

For each ground truth bounding box (GT Box) of the apple leaf disease, we incorporate the k-nearest anchor boxes from each level of the feature pyramid, based on their distance to the center of the GT Box, into the candidate sample set. The intersection over union (IoU) was computed between all candidate samples and the GT Box. The mean (m) and variance (n) of the IoU values in this set were calculated, and the adaptive threshold (t) was defined as the sum of the mean and variance: t = m + n. Next, the candidate samples were filtered based on the following criteria: for the preset anchor boxes with an IoU greater than t, if their center points fell within the GT Box, they were assigned as positive samples; otherwise, they were assigned as negative samples. Through this assignment strategy, we ensured that the true apple leaf disease targets had a sufficient number of positive samples for training, enabling the model to better detect the disease targets. As shown in Figure 4, the use of rotating bounding boxes in target detection provides higher confidence, effectively improving the accuracy of apple leaf disease identification and enhancing the generalization ability of the model.

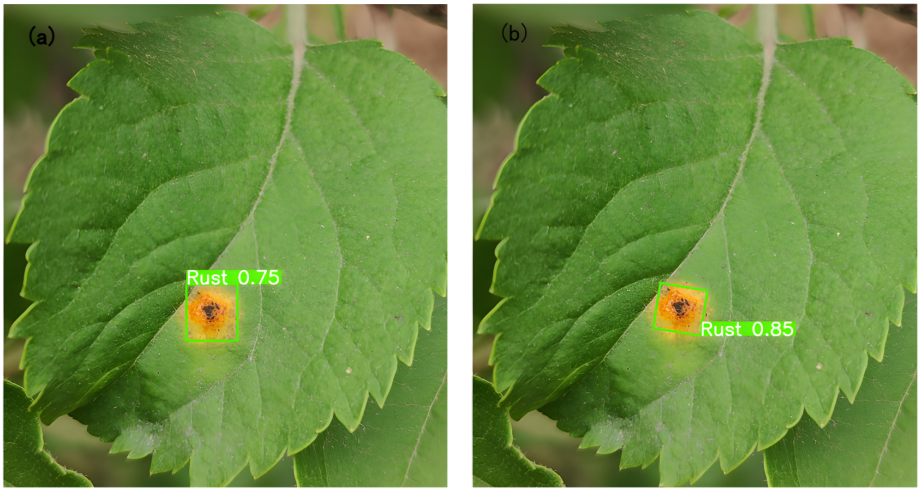


Figure 4. Comparison of detection results using different bounding box methods: (**a**) detection with horizontal bounding box; (**b**) detection with rotating bounding box.

2.3. Architecture of Proposed Network

Given the excellent learning capabilities demonstrated by CNNs, deep learning has been extensively utilized in various object detection tasks [30]. Commonly used object detection networks can be broadly divided into two categories: single- and two-stage networks. Unlike two-stage networks, single-stage networks bypass the generation of region proposals and directly use the backbone network to predict both the category and location of objects. This approach has several advantages, including a faster computation speed and lower computational cost, making it more suitable for real-world scenarios.

In the present research, inspired by the design principles of the YOLO network, we introduce a novel and efficient single-stage convolutional network framework customized for apple leaf disease detection. Our goal was to leverage the advantages of single-stage networks, such as speed and efficiency, to design an effective solution for apple leaf disease detection. The overall network model is shown in Figure 5.

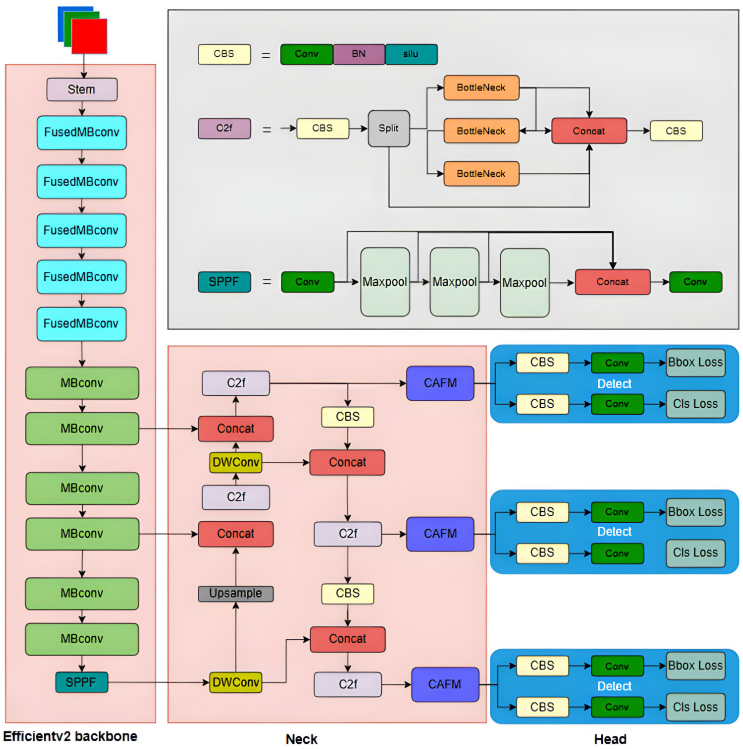


Figure 5. Overall network structure.

2.3.1. Feature Extraction Backbone

In image classification tasks, a convolutional module is typically used to extract image features [31]. The ResNet classification network adopts a residual structure convolutional feature extraction network. However, residual structure convolution uses a dimensionality reduction-then-increase approach to extract network features, which can result in a reduction in feature information during execution.

This research introduced an enhanced backbone network model built upon EfficientNetv2. Unlike the traditional ResNet backbone network, the inverted residual structure FusedMBConv and S-MBConv modules form a sparse feature by first increasing and then decreasing the dimensions, which reduces the information loss. FusedMBConv and S-MBConv are shown in Figure 6. Specifically, the S-MBConv module first expands the input features through a regular convolution containing batchnorm (BN) and SiLU activation functions and then uses depth-wise convolution to minimize the computational load and the number of parameters. Subsequently, it allocates channel weights through an attention mechanism and then reduces the dimensions by inputting a regular convolution containing the BN layer, finally outputting the features after the dropout layer. The S-MBConv module significantly reduced the network parameters and computational expense.

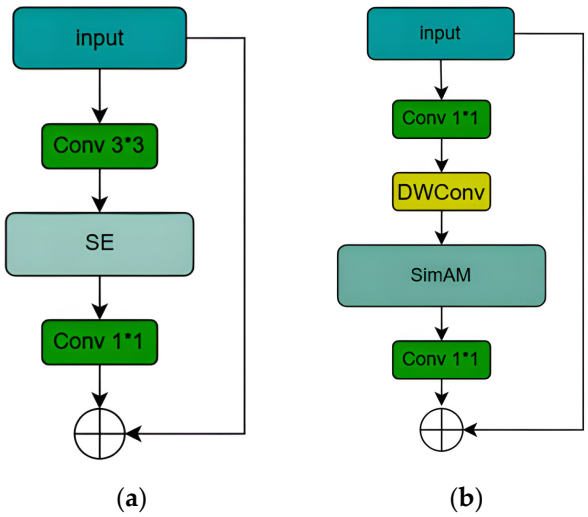


Figure 6. The structure of convolution blocks: (**a**) FusedMBConv; (**b**) S-MBConv.

The selection of the attention mechanism in convolution can significantly impact the model's capacity to extract features. The unimproved MBConv module uses the Squeeze and Excitation (SE) attention mechanism to enhance the extraction of key information. However, the SE attention mechanism only considers modeling the channel relationship to re-evaluate the significance of each channel and is unable to capture attention in the spatial dimension. Consequently, it is more suitable for scenarios with an increased number of channels. This study integrated the SimAM attention mechanism to boost the model's feature extraction capabilities. The SimAM attention mechanism can effectively capture both channel-wise and spatial attention and is more suitable for feature extraction tasks than the SE attention mechanism.

The SimAM attention mechanism is a three-dimensional attention mechanism that combines spatial information and channel information. It evaluates the t-th neuron by defining an energy function e_t^* to determine the significance of each individual neuron. The energy function is defined in Equation (8):

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda}, \tag{8}$$

where λ is the regular norm. t is the t-th neuron on a single channel of the input feature map. $\hat{u}$ is the average of all neurons on a single channel. $\hat{\sigma}^2$ is the variance of all neurons on a single channel. $\hat{\sigma}^2$ and $\hat{u}$ are defined as Equations (9) and (10):

$$\hat{u} = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i, \tag{9}$$

$$\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1} (x_i - u_t)^2 \tag{10}$$

As the value of e_t^* becomes smaller, it indicates that the difference between the neuron and its surrounding neurons is greater, the neuron has higher linear separability, and it is more important. Therefore, by using $1/e_t^*$ as a weight coefficient and applying a sigmoid activation function to limit the value range, a new output feature map can be obtained. And the input features are enhanced by the following Equation (11):

$$\hat{X} = sigmoid(\frac{1}{e_t^*}) \odot X, \tag{11}$$

By integrating the SimAM attention mechanism, the network was empowered to selectively concentrate on and enhance the most meaningful features drawn from both the deep and shallow layers of the backbone network. This focused attention allows the model to more effectively learn and represent the varied characteristics of different target objects, ultimately leading to an improved overall performance.

2.3.2. Feature Mixing Neck Network

In apple leaf disease detection, a major challenge lies in the high similarity of features among certain diseases, which makes it difficult to effectively fuse contextual information and improve the accuracy and efficiency of the detection model. To address this problem, we propose improvements to the neck network. Specifically, we introduce depth-wise separable convolution in PANet, which not only transforms the feature channel numbers of the backbone network but also preserves the spatial properties of the features. Furthermore, the addition of the Convolution and Attention Fusion Module (CAFM) aims to capture long-range dependencies and correlations between domain-specific features, thereby enhancing the fusion of global and local features. By incorporating these innovations, our approach can better leverage the contextual information and improve the overall performance of the apple leaf disease detection model. The enhanced fusion of multi-scale and multi-level features leads to the more accurate and efficient identification of target diseases, even in cases where the visual characteristics are highly similar.

The core of traditional convolution is the ability to simultaneously learn representations along both the channel and spatial aspects of the input data. The convolutional kernels operate across the channels, allowing the extraction of inter-channel correlations while also scanning the spatial dimensions to uncover local patterns and relationships. Compared with traditional convolution, depth-wise convolution is made up of both depth-wise and point-wise convolutions, as illustrated in Figure 7. The convolution kernels of depth-wise convolution operate in a single-channel mode, where each input channel needs to be convolved individually. This leads to an output feature map with a channel count equal to that of the input feature map. Point-wise convolution essentially serves as a channel-wise feature transformation, using a 1×1 convolution kernel to increase the dimensionality of the feature maps. DWConv decomposes the feature extraction into two processes. Compared with the standard volume, the DWConv of Figure 8 significantly reduces the amount of computation and parameters and is more suitable for lightweight detection models.

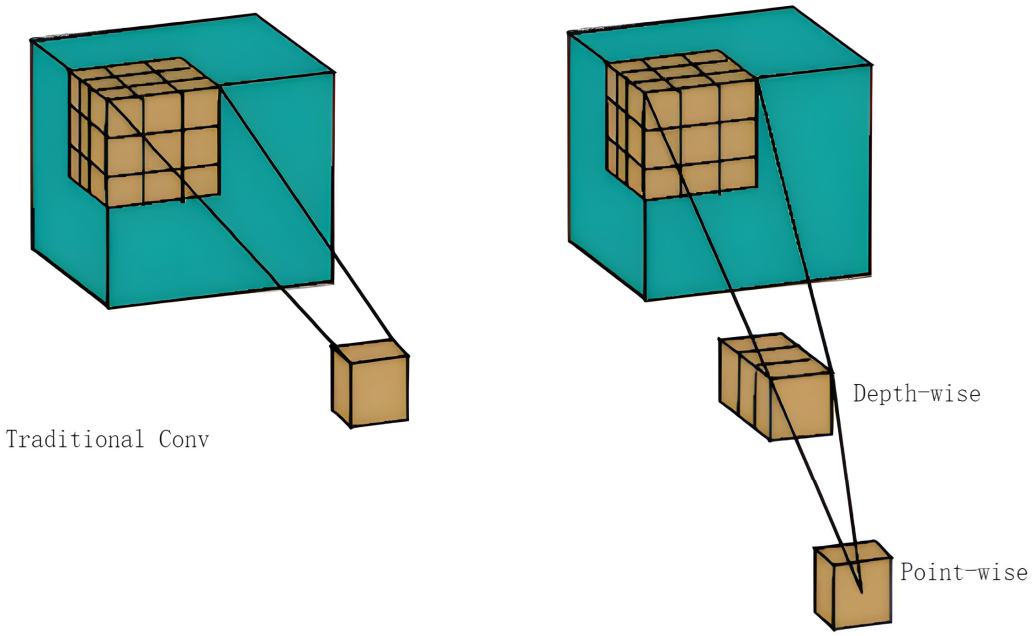


Figure 7. Traditional Conv and depth-wise separable convolution.

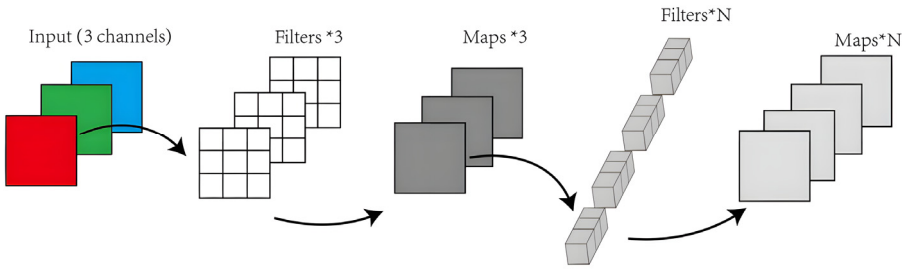


Figure 8. The structure of DWConv.

To strengthen the model's ability to extract features, most deep learning-based methods usually adopt the U-Net architecture to capture both global and local features. The CAFM comprises both a global network and a local network. The structure of CAFM is shown in Figure 9. In the local network, we first use a 1×1 convolution to adjust the channel dimensions, which improves cross-channel interaction and encourages information integration. Following this channel-wise feature transformation, we perform a channel shuffling operation to further mix and blend the channel information. The input tensor is partitioned into groups along the channel dimension by the channel shuffle and then applies depth-wise separable convolution within each group. This induces a shuffling of the channel order, effectively integrating the channel-wise features. The output tensors from each group are merged along the channel dimension to produce a new output tensor, rich in cross-channel interactions and fused information. Finally, we leverage a $3 \times 3 \times 3$ convolution to extract spatially aware features from this enhanced feature representation. The 3D convolution kernel allows us to capture the intricate spatial relationships and patterns within the data, complementing the preceding channel-wise operations. The local network can be expressed as Equation (12):

$$F_{Local} = W_{3\times3\times3}(shuffle(W_{1\times1}(F_{Input}))), \tag{12}$$

where F_{Local} is the output network, $W_{3\times3\times3}$ represents a $3 \times 3 \times 3$ convolution, $W_{1\times1}$ represents a 1×1 convolution, $shuffle$ represents the operation of channel shuffle, and F_{Input} denotes the input feature from the last module.

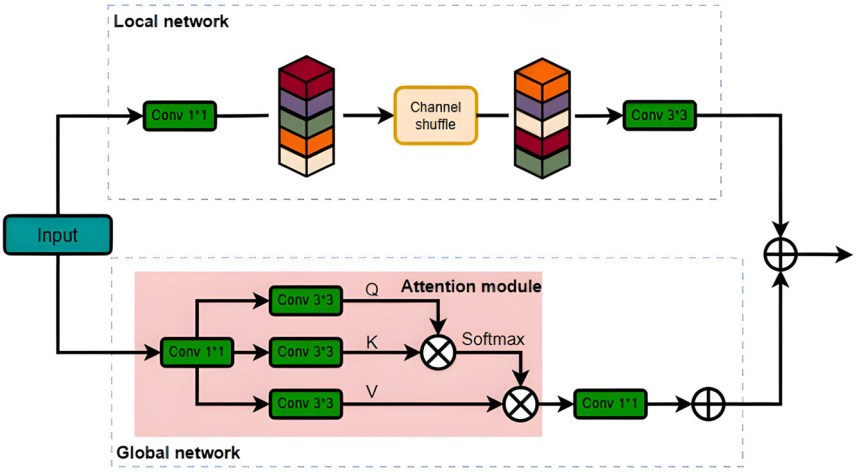


Figure 9. The structure of CAFM.

In the global network, the CAFM first generates a query, key, and value via a 1×1 convolution and 3×3 DWConv operation. These three tensors all have the same shape of $H \times W \times C$. The global network can be described by the following equations:

$$F_{Global} = W_{1\times1} Attention(q, k, v) + F_{Input}, \tag{13}$$

$$Attention(q, k, v) = soft\max(qk^T/\sqrt{d})v, \tag{14}$$

where F_{Global} is the output feature via the global network, and $\sqrt{d}$ represents the square root of the q. To summarize, the output feature of the CAFM can be expressed as Equation (15):

$$F = F_{Global} + F_{Local} \tag{15}$$

2.4. Parameter Settings and Evaluation Metrics

To assess the accuracy and effectiveness of the improved model, we utilized the PyTorch deep learning framework for model design and training. The trained model's weights were then employed to predict apple leaf images. The experiments were conducted using a GeForce RTX 4060 graphics card on a Windows 11 operating system. The manufacturer of the equipment is NVIDIA, and the supplier's address is Taiwan, China. The details of our experiments are presented in Table 2.

Table 2. Experimental environments.

Experimental Environments	Details
Program language	Python3.8
Deep learning frameworks	PyTorch1.12.1
Operating system	Windows 11
CPU	I7-13700KF
GPU	NVIDIA GeForce RTX 4060
CUDA	Toolkit V11.7
Cudnn	8.4.1

The experiments were trained on our enhanced apple leaf dataset. Considering the GPU memory size and the cost of time, we set the batch size of the model to 4 and the initial learning rate to 0.01. According to the SGD optimizer, we adjusted the learning rate to set the weight decay to 0.0005. The training parameters of our model were designed based on YOLOv8, as illustrated in Table 3.

Table 3. Training parameters.

Parameter Name	Parameter Value
Batch size	4
Momentum	0.9
Learning rate	0.01
Weight decay	0.0005
Epoch	300
Optimizer	SGD

To demonstrate the apple leaf disease detection performance of our improved model, we employed the widely used Microsoft COCO evaluation metrics. These metrics offer an in-depth evaluation of the model's effectiveness in object detection tasks. Our model was trained using the training set and tested on the validation set. The main metrics used for evaluation are mAP@0.5, mAP@0.5 : 0.95, Precision (P), Recall (R), and $F1$ score. Accuracy was used to measure the correctness of the predicted number in the detection, and Root Mean Square Error ($RMSE$) was used to evaluate the reliability and stability of the detection. The equations are presented below:

$$P = \frac{TP}{TP + FP}, \tag{16}$$

$$R = \frac{TP}{TP + FN}, \tag{17}$$

$$mAP = \frac{\sum_{n}^{i=1} A_i P}{n}, \tag{18}$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \times 100\% \tag{19}$$

$$Acc = (1 - \frac{1}{n}\sum_{1}^{n} \frac{|t_i - c_i|}{t_i}) \times 100\%, \tag{20}$$

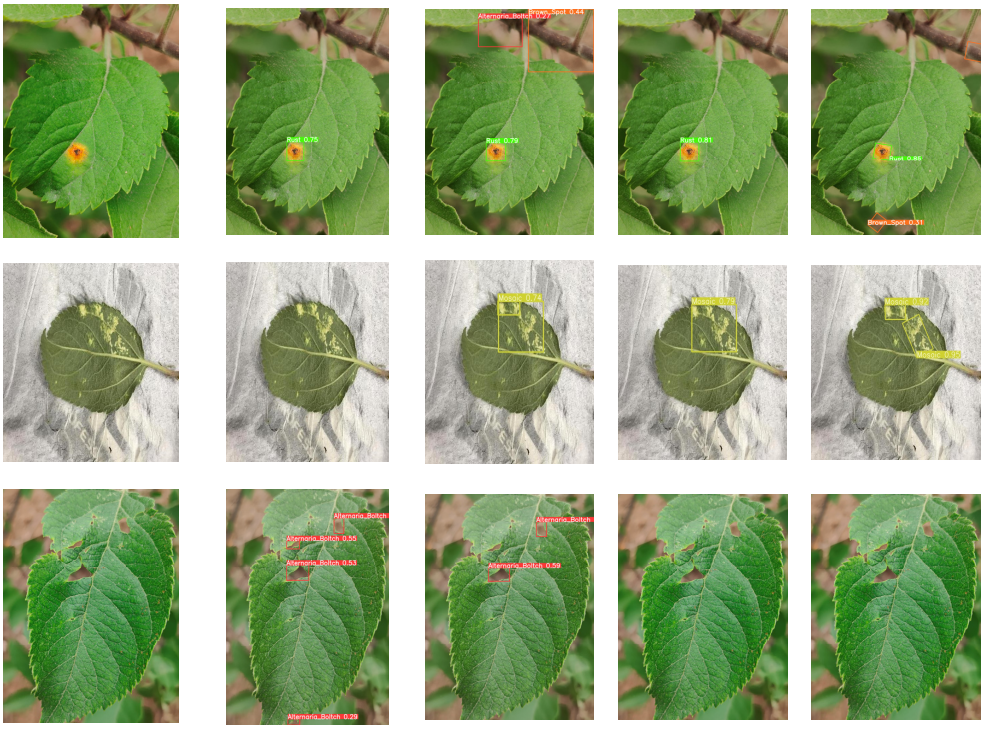
$$RMSE = \sqrt{\frac{\sum_{1}^{n} (t_i - c_i)^2}{\mathrm{n}}} \tag{21}$$

Precision represents the likelihood that the predicted target is accurate. TP represents that the apple leaf disease is detected correctly. FP means that the detection result is wrong. Recall represents the probability of a diseased apple leaf being truthfully predicted. FN denotes that the diseased leaf is incorrectly predicted to be healthy. mAP denotes the Precision accuracy of the detection process. mAP@0.5 represents the average Precision at an intersection over the IoU criterion of 0.5. mAP@0.5 : 0.95 refers to the average mAP value with IoU from 0.5 to 0.95 with a step size of 0.05.

3. Results

3.1. Detection Results from Alternative Methods

Figure 10 shows three images containing diseased apple leaves, a healthy apple leaf, and disease annotation in the test dataset. For apple leaf disease detection, the complexity of the environment and the irregularity of disease are the main problems that affect the detection results. In an environment where the light is insufficient and the field of view is easily blocked, the characteristics of the disease are often easily blocked, which results in information loss, resulting in the false detection and missing detection of the disease. Moreover, if the shape of the disease is irregular, it is difficult for the traditional horizontal box to accurately fit the location of the disease, which leads to a reduction in confidence.



(**a**) Testing images (**b**) Faster-RCNN (**c**) YOLOv5 (**d**) YOLO8 (**e**) Ours

Figure 10. Comparison of testing results using different models: (**a**) original testing images with annotations; (**b**) detection result of Faster-RCNN; (**c**) detection result of YOLOv5; (**d**) detection result of YOLOv8; (**e**) detection results of our proposed model.

To tackle these issues, we utilized our newly designed model for detection, with the results displayed in Figure 10. The quantitative detection result of our model is summarized in Table 4. We found that our model exhibits exceptional performance, particularly in terms of the *mAP*@0.5 (93.3%) and Precision (88.7%). This demonstrates its high performance in locating and predicting apple leaf diseases. The model we proposed also presents a notable Recall of 89.6%, which represents its effectiveness in identifying the most diseased apple leaves. The F1 score, as the harmonic mean of Precision and Recall, achieves a relatively high value of 89.1%. This further validates our model's strength in balancing accuracy and coverage.

Table 4. Detection results obtained from different deep learning models.

Detection Models	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Precision	Recall	F1	Params/M
Faster-RCNN	76.7%	39.5%	67.7%	76.6%	71.9%	-
Efficient Det-D1	79.6%	41.8%	73.8%	68.7%	71.1%	6.6
YOLOv5	85.1%	48.6%	75.3%	82.9%	78.9%	11.7
DETR	82.6%	46.6%	76.4%	73.9%	75.1%	42
YOLOv8	85.6%	50.2%	77.1%	81.3%	79.2%	11.2
Ours	93.3%	53.1%	88.7%	89.6%	89.1%	9.8

To further evaluate the detection performance of the model, we randomly selected 3000 images as a dataset containing 2659 apple leaf disease targets. In this sample, we measured the number of targets detected in the test sample, including potential missed and error detection. The RMSE metric measures the difference between the detection count and the actual label count for each image. After the experiment, we obtained the results shown in Figure 11. The proposed model has the closest count to the original label, and the RMSE value is the smallest (0.368). It has good robustness and generalization ability.

Figure 11. Comparison of RMSE counting results for different models.

3.2. Ablation Experiment

Based on YOLOv8, we substituted the original backbone network with the enhanced EfficientNetV2 network. The reverse residual structure FusedMBConv and S-MBConv modules formed a sparse feature by first increasing and then decreasing, reducing information loss. The addition of the SimAM attention mechanism can capture the channel attention and spatial attention effectively, expand the receptive field effectively, and improve the feature extraction capability of the backbone network. At the same time, the depth-separable product is introduced into PANet, which not only transforms the number of feature channels in the backbone network but also preserves the spatial information of the feature. In addition, the addition of the CAFM is designed to capture remote dependencies and correlations between domain-specific features, thereby enhancing the fusion of global and local features. The improvement from the horizontal frame to the rotating frame provides a new idea for the detection of diseases and pests. Compared with the horizontal frame, the rotating frame can better fit the complex shape of apple leaf diseases, greatly improve the confidence of detection results, and further enhance the robustness of the model. We conducted ablation experiments to explore the reasonableness and effect of the proposed improvement, and the results are shown in Table 5.

Table 5. Detection results obtained from quantitative ablation experiments with YOLOv8 serving as the base network.

Efficientnetv2	SimAM	DWConv	CAFM	Rotating Frame	$AP_{0.5}$	$mAP_{0.5:0.95}$
-	-	-	-	-	85.6%	50.2%
√	-	-	-	-	87.4%	51.6%
-	√	-	-	-	85.9%	50.6%
-	-	√	-	-	86.1%	50.7%
-	-	-	√	-	86.6%	51.2%
-	-	-	-	√	88.2%	51.9%
√	√	-	-	-	86.7%	51.8%
√	-	√	-	-	86.5%	51.7%
√	-	-	√	-	89.6%	52.4%
√	-	-	-	√	90.7%	52.2%
√	√	√	√	√	90.3%	53.1%

4. Discussion

In apple cultivation scenarios, effective progress has been made in pest and disease identification using deep learning methods. However, most experiments have primarily focused on single or dominant leaves, employing horizontal detection boxes. This has resulted in low detection accuracy for early-stage pests and fine, irregularly shaped lesions, as well as limited generalization ability for specific crops. To address issues such as false

positives for early-stage diseases with varying shapes, this study proposes the application of rotated detection boxes, effectively enhancing disease recognition accuracy. To tackle the high similarity of disease features between leaves, we have improved the neck network to strengthen the feature fusion capabilities. Compared to traditional models, the deep detection model proposed in this study demonstrates significant improvements in feature extraction and fusion capabilities, better coping with background interference and the similarities between different diseases.

This paper combines the Plant Village dataset and the on-site photographed dataset and uses data augmentation methods such as Mosaic to augment the dataset and improve the imbalance problem. In the backbone network construction, the EfficientNetV2 structure is adopted. The inverted residual structure FusedMBConv and S-MBConv modules form a sparse feature through an up–down approach, reducing information loss. The addition of the SimAM attention mechanism can effectively capture channel attention and spatial attention, effectively expand the receptive field, and enhance the backbone network's feature extraction ability. In the neck network, depth-separable convolution is introduced, which not only transforms the feature channels of the backbone network but also retains the spatial information of the features. CAFM improves the feature fusion capability of the neck network by combining the global network and the local network.

The experimental results show that the F1 score of this model is 89.1%, which is 17.2%, 18%, 10.2%, 14%, and 9.9% higher than Faster-RCNN, Efficient Det-D1, YOLOv5, DETR, and YOLOv8, respectively, indicating the outstanding performance of our detection model. The improved model also performs well in real orchard disease scenarios. While effective, further improvements in generalization to complex disease features and diverse environments are needed. Future research may explore more diverse datasets and advanced learning techniques like ensemble learning to enhance accuracy and robustness.

5. Conclusions

This research tackles the challenges of small feature sizes, low recognition accuracy, and irregular shapes of apple leaf diseases, as well as the texture similarity between different diseases. We propose a detection method for apple leaf diseases based on rotated bounding boxes. The introduction of rotated boxes effectively helps the model capture irregular pathological features and accurately locate disease positions. Additionally, we introduce a method for calculating target similarity called Probabilistic IoU (ProbIoU), which better computes the intersection over union (IoU) of the rotated bounding boxes, ensuring detection accuracy. We then use the YOLO model as the framework, incorporating the EfficientNetv2 backbone network. The inverted residual structures, the FusedMBConv and S-MBConv modules, create sparse features through a gradual increase and decrease method, reducing information loss. The addition of the SimAM attention mechanism effectively captures channel attention and spatial attention, maintaining the feature extraction capability of the backbone network while keeping the model lightweight. Finally, at the neck network stage, introducing DWConv optimizes the model's parameter count, ensuring a lightweight structure and laying the groundwork for future deployment on mobile devices. The CAFM improves the model's capacity to capture long-term dependencies and correlations between domain-specific features, thereby improving the fusion of global and local features. This enables the model to more precisely and efficiently identify target diseases, even in cases where visual features are highly similar.

An accuracy of 88.7% was achieved by the model, along with a corresponding value of 93.3%, with a model size of 9.8 MB. In upcoming research, we plan to further enhance the model's lightweight design to facilitate easy deployment on edge devices. Additionally, we will enrich the dataset by incorporating more data samples collected under natural conditions, further optimizing the model to reduce interference from complex backgrounds.

Author Contributions: Conceptualization, Z.Q. and Y.X.; formal analysis, investigation, methodology, visualization, and writing—original draft preparation, Z.Q.; writing—review and editing, Y.X., C.C., W.Z. and G.Y.; funding acquisition, Y.X. and C.C.; resources, W.Z. and G.Y.; supervision, Y.X., C.C. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Natural Science Foundation of China, grant number 61601275, and the Introduction of high-level talents and overseas returnee's scientific fund in Nanjing Forestry University, grant number GXL015.

Data Availability Statement: The original contributions presented in this study are included in this article.

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

References

1. Ke, B.; Zhang, Q. Assessing the Efficiency and Potential of China's Apple Exports. *Complexity* **2024**, *2024*, 7816792. [CrossRef]
2. Čirjak, D.; Miklečić, I.; Lemić, D.; Kos, T.; Pajač Živković, I. Automatic Pest Monitoring Systems in Apple Production under Changing Climatic Conditions. *Horticulturae* **2022**, *8*, 520. [CrossRef]
3. Nabi, F.; Jamwal, S.; Padmanbh, K. Wireless Sensor Network in Precision Farming for Forecasting and Monitoring of Apple Disease: A Survey. *Int. J. Inf. Technol.* **2022**, *14*, 769–780. [CrossRef]
4. Khan, A.I.; Quadri, S.M.K.; Banday, S.; Latief Shah, J. Deep Diagnosis: A Real-Time Apple Leaf Disease Detection System Based on Deep Learning. *Comput. Electron. Agric.* **2022**, *198*, 107093. [CrossRef]
5. Singh, S.; Gupta, S.; Tanta, A.; Gupta, R. Extraction of Multiple Diseases in Apple Leaf Using Machine Learning. *Int. J. Image Graph.* **2022**, *22*, 2140009. [CrossRef]
6. Alqethami, S.; Almtanni, B.; Alzhrani, W.; Alghamdi, M. Disease Detection in Apple Leaves Using Image Processing Techniques. *Eng. Technol. Appl. Sci. Res.* **2022**, *12*, 8335–8341. [CrossRef]
7. Moussafir, M.; Chaibi, H.; Saadane, R.; Chehri, A.; Rharras, A.E.; Jeon, G. Design of Efficient Techniques for Tomato Leaf Disease Detection Using Genetic Algorithm-Based and Deep Neural Networks. *Plant Soil* **2022**, *479*, 251–266. [CrossRef]
8. Das, D.; Singh, M.; Mohanty, S.S.; Chakravarty, S. Leaf Disease Detection Using Support Vector Machine. In Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 28–30 July 2020; pp. 1036–1040.
9. Wang, Z.; Cui, J.; Zhu, Y. Review of Plant Leaf Recognition. *Artif. Intell. Rev.* **2023**, *56*, 4217–4253. [CrossRef]
10. Malhi, A.; Apopei, V.; Madhikermi, M.; Mandeep; Främling, K. Smartphone Based Grape Leaf Disease Diagnosis and Remedial System Assisted with Explanations. In *Explainable and Transparent AI and Multi-Agent Systems*; Calvaresi, D., Najjar, A., Winikoff, M., Främling, K., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2022; Volume 13283, pp. 57–71. ISBN 978-3-031-15564-2.
11. Zou, Z.; Wu, Q.; Long, T.; Zou, B.; Zhou, M.; Wang, Y.; Liu, B.; Luo, J.; Yin, S.; Zhao, Y.; et al. Classification and Adulteration of Mengding Mountain Green Tea Varieties Based on Fluorescence Hyperspectral Image Method. *J. Food Compos. Anal.* **2023**, *117*, 105141. [CrossRef]
12. Shukla, A.; Kalnoor, G.; Kumar, A.; Yuvaraj, N.; Manikandan, R.; Ramkumar, M. Improved Recognition Rate of Different Material Category Using Convolutional Neural Networks. *Mater. Today Proc.* **2023**, *81*, 947–950. [CrossRef]
13. Hua, S.; Xu, M.; Xu, Z.; Ye, H.; Zhou, C. Multi-Feature Decision Fusion Algorithm for Disease Detection on Crop Surface Based on Machine Vision. *Neural Comput. Appl.* **2022**, *34*, 9471–9484. [CrossRef]
14. Wani, J.A.; Sharma, S.; Muzamil, M.; Ahmed, S.; Sharma, S.; Singh, S. Machine Learning and Deep Learning Based Computational Techniques in Automatic Agricultural Diseases Detection: Methodologies, Applications, and Challenges. *Arch. Comput. Methods Eng.* **2022**, *29*, 641–677. [CrossRef]
15. Li, X.; Li, C.; Rahaman, M.M.; Sun, H.; Li, X.; Wu, J.; Yao, Y.; Grzegorzek, M. A Comprehensive Review of Computer-Aided Whole-Slide Image Analysis: From Datasets to Feature Extraction, Segmentation, Classification and Detection Approaches. *Artif. Intell. Rev.* **2022**, *55*, 4809–4878. [CrossRef]
16. Yang, Y.; Wu, Z.; Yang, Y.; Lian, S.; Guo, F.; Wang, Z. A Survey of Information Extraction Based on Deep Learning. *Appl. Sci.* **2022**, *12*, 9691. [CrossRef]
17. Turkoglu, M.; Hanbay, D.; Sengur, A. Multi-Model LSTM-Based Convolutional Neural Networks for Detection of Apple Diseases and Pests. *J. Ambient Intell. Humaniz. Comput.* **2022**, *13*, 3335–3345. [CrossRef]
18. Thyagaraj, R.; Satheesha, T.Y.; Bhairannawar, S. Plant Leaf Disease Classification Using Modified SVM with Post Processing Techniques. In Proceedings of the 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, 16–17 June 2023; pp. 1–4.
19. Gao, J.; Tan, F.; Cui, J.; Ma, B. A Method for Obtaining the Number of Maize Seedlings Based on the Improved YOLOv4 Lightweight Neural Network. *Agriculture* **2022**, *12*, 1679. [CrossRef]
20. Ling, L.; Wu, Q.; Huang, K.; Wang, Y.; Wang, C. A Lightweight Bearing Fault Diagnosis Method Based on Multi-Channel Depthwise Separable Convolutional Neural Network. *Electronics* **2022**, *11*, 4110. [CrossRef]

21. Liu, S.; Yin, D.; Feng, H.; Li, Z.; Xu, X.; Shi, L.; Jin, X. Estimating Maize Seedling Number with UAV RGB Images and Advanced Image Processing Methods. *Precis. Agric.* **2022**, *23*, 1604–1632. [CrossRef]
22. Zhao, Y.; Zeng, K.; Zhao, Y.; Bhatia, P.; Ranganath, M.; Kozhikkavil, M.L.; Li, C.; Hermosillo, G. Deep Learning Solution for Medical Image Localization and Orientation Detection. *Med. Image Anal.* **2022**, *81*, 102529. [CrossRef]
23. Harakannanavar, S.S.; Rudagi, J.M.; Puranikmath, V.I.; Siddiqua, A.; Pramodhini, R. Plant Leaf Disease Detection Using Computer Vision and Machine Learning Algorithms. *Glob. Transit. Proc.* **2022**, *3*, 305–310. [CrossRef]
24. Li, Z.; Hou, B.; Wu, Z.; Guo, Z.; Ren, B.; Guo, X.; Jiao, L. Complete Rotated Localization Loss Based on Super-Gaussian Distribution for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5618614. [CrossRef]
25. Vishnoi, V.K.; Kumar, K.; Kumar, B.; Mohan, S.; Khan, A.A. Detection of Apple Plant Diseases Using Leaf Images Through Convolutional Neural Network. *IEEE Access* **2023**, *11*, 6594–6609. [CrossRef]
26. Kejriwal, S.; Patadia, D.; Sawant, V. Apple Leaves Diseases Detection Using Deep Convolutional Neural Networks and Transfer Learning. In *Computer Vision and Machine Learning in Agriculture*; Uddin, M.S., Bansal, J.C., Eds.; Algorithms for Intelligent Systems; Springer: Singapore, 2022; Volume 2, pp. 207–227. ISBN 9789811699900.
27. Hughes, D.P.; Salathe, M. An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics. *arXiv* **2015**, arXiv:1511.08060.
28. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
29. Roy, A.M.; Bose, R.; Bhaduri, J. A Fast Accurate Fine-Grain Object Detection Model Based on YOLOv4 Deep Neural Network. *Neural. Comput. Appl.* **2022**, *34*, 3895–3921. [CrossRef]
30. Turay, T.; Vladimirova, T. Toward Performing Image Classification and Object Detection with Convolutional Neural Networks in Autonomous Driving Systems: A Survey. *IEEE Access* **2022**, *10*, 14076–14119. [CrossRef]
31. Li, G.; Fang, Q.; Zha, L.; Gao, X.; Zheng, N. HAM: Hybrid Attention Module in Deep Convolutional Neural Networks for Image Classification. *Pattern Recognit.* **2022**, *129*, 108785. [CrossRef]