Citation: Xu E, Chenkao Liu, Jin Zhou, et al. Apple leaf disease identification model based on improved MobileNetV3-Small. Journal of Harbin Institute of Technology(New Series). DOI:10.11916/j.issn.1005-9113.24065

Apple Leaf Disease Identification Model Based on Improved MobileNetV3-Small

*Xu E**, *Chenkao Liu*, *Jin Zhou*, *Wei Song*, *Qi Yan*, *Song Wang*

(*College of Information Science and Technology*, *Bohai University*, *Jinzhou 121013*, *Liaoning*, *China*)

Abstract: To enhance the recognition accuracy of current network models for apple leaf diseases, a lightweight model that leverages an enhanced MobileNetV3-Small architecture is introduced in this study. The improved model utilizes MobileNetV3-Small, a lightweight architecture with fewer parameters, serving as the primary network for feature extraction. It integrates a weighted bi-directional feature pyramid network that fuses multi-scale features, thereby enhancing the model's capacity to detect disease characteristics across various scales. Additionally, an efficient multi-scale attention mechanism is integrated to mitigate the influence of complex background noise in natural environments, further improving disease recognition accuracy. The experiment utilizes the AppleLeaf9 public dataset to classify healthy apple leaves and eight distinct disease types. The results indicate that, when using the augmented dataset, the improved model achieves a recognition accuracy of 95.98%, with only 1.72M parameters, 123.16M FLOPs, and an inference time of just 14.10 ms. Compared with eight other lightweight neural network models, including MobileNetV2, ShuffleNet_v2_1.5×, ResNet50, MobileNetV3-Large, EfficientNet-B0, MobileNetV3-Small, MobileNetV4-Conv-Small, and MobileNetV4-Conv-Medium, the improved model demonstrates superior accuracy. In particular, the proposed model achieves a recognition accuracy improvement of 0.93 percentage points compared with the baseline MobileNetV3-Small model. The optimized model introduced in this study effectively improves the accuracy in identifying diseases in apple leaves, while maintaining a low parameter count and fast inference speed, thus offering a novel approach for deploying disease recognition models on agricultural electronic devices.

Keywords: deep learning; feature fusion; attention mechanism; lightweight; disease identification

CLC number: TU399 **Document code**: A **Article ID**: 1005-9113(2025)00-0000-11

0 Introduction

Apple is one of the important cash crops in our country[1]. Diseases affecting apple leaves have significantly hindered the healthy growth of the apple industry[2]. Conventional approaches for detecting apple leaf diseases primarily depend on farmers' expertise and practical experience, which cannot provide real-time prevention and control of fruit trees. As a result, the disease is often not discovered until it has severely impacted apple quality, and in some cases, it even causes a significant reduction in apple production[3]. In recent years, advances in artificial intelligence have led to the extensive application of deep neural networks in disease recognition, as these networks can automatically extract features directly from original input data[4]. Zhang et al.[5] introduced a hybrid attention network designed for recognizing citrus diseases, which combines a frequency-domain attention mechanism and an SE attention module within the ResNet backbone. Additionally, they employed a large convolution kernel to enlarge the receptive field. Zhang et al.[6] used the maximum inter-class variance method based on super green feature to segment the image, and applied transfer learning and a residual network to construct a millet disease recognition model, achieving 98.2% recognition accuracy for four types of millet diseases. Liang et al.[7] improved the SqueezeNet network by introducing a spatial attention mechanism and a dense connection module, constructing two models, SqueezeNet1 and SqueezeNet2, which achieved recognition accuracies of 89.60% and 94.37%, respectively, on the apple leaf disease dataset. Chen et al.[8] adopted DenseNet as the backbone architecture and substituted traditional convolutional layers with depthwise separable convolutions to decrease model

Received 2024-11-04

Sponsored by the China National Key R&D Program during the 13th Five-year Plan Period (Grant No.2019YFD0901605)

Corresponding author: E Xu, Ph.D, Professor. Email: exu21@163.com.

parameters. They also introduced an enhanced Mobile-DANet framework, integrating both spatial and channel attention mechanisms, achieving a recognition accuracy of 98.5% on the open Plant Village maize disease dataset and 95.86% on the local dataset with complex background conditions. Hu et al.[9] developed a model named LE-EfficientNet aimed at identifying grape leaf diseases. Their approach enhanced the EfficientNet-B0 architecture by integrating a large kernel attention module and an efficient channel attention mechanism, thereby improving feature extraction capabilities related to disease detection. Experimental outcomes demonstrated that this method achieved a recognition accuracy increase of 1.58 percentage points on the Plant Village grape leaf disease dataset. Guo et al.[10] integrated the ET attention module within the MobileNetV3 architecture, optimized the model's fully connected layer and operator, and combined it with a transfer learning strategy to construct an apple leaf disease recognition model, achieving a disease recognition accuracy of 95.62%.

The studies mentioned above demonstrate that deep neural networks can achieve remarkable results in recognizing plant leaf diseases, even under conditions involving complicated backgrounds. Nevertheless, deep neural networks typically contain numerous parameters, resulting in high memory consumption on agricultural electronic devices deployed within farming environments, which limits their application in agricultural production. In addition, there are some problems in apple leaf disease recognition in natural environment, such as disease feature size, different shapes, scattered feature distribution and complex picture background. To address the aforementioned issues, this study presents MobileNet-BiFPN-EMA, a lightweight model designed for apple leaf disease recognition. The proposed model enhances recognition accuracy while keeping the parameter count low, reduces the memory burden on agricultural electronic devices.

1 Experimental Dataset and Preprocessing

The AppleLeaf9 dataset, created by Yang et al.[11], is used for the experiments in this paper. To guarantee precise labeling of images in the AppleLeaf9 dataset, Yang et al. invited agricultural disease experts to screen the data and remove mislabeled images. A total of 14582 images were obtained, 94% of which were taken in natural environments with complex backgrounds. Only 2.5% of the images are from the PVD dataset with static image backgrounds, making the dataset more representative of real-world application scenarios and enhancing the model's capability to generalize effectively in real-world environments.

Fig. 1 illustrates sample images for each disease included in the AppleLeaf9 dataset. The dataset comprises healthy leaf samples and eight distinct disease categories: alternaria leaf spot, brown spot, frogeye leaf spot, gray spot, mosaic, powdery mildew, rust, and scab. Table 1 provides detailed information for each disease type. To assess the performance of the model, the dataset is divided into training, validation, and test sets following a 3∶1∶1 ratio, containing 8754 training images, 2916 validation images, and 2912 test images, respectively. To enhance data diversity, six online data augmentation methods are applied to the training set images, including random cropping, random brightness and contrast adjustment, random rotation, random Gaussian blur, horizontal flipping, and vertical flipping, using the transforms provided by torchvision. The validation and test set images are not augmented to ensure the model's practicality. Fig. 2 illustrates a representative example of a pre-processed image.

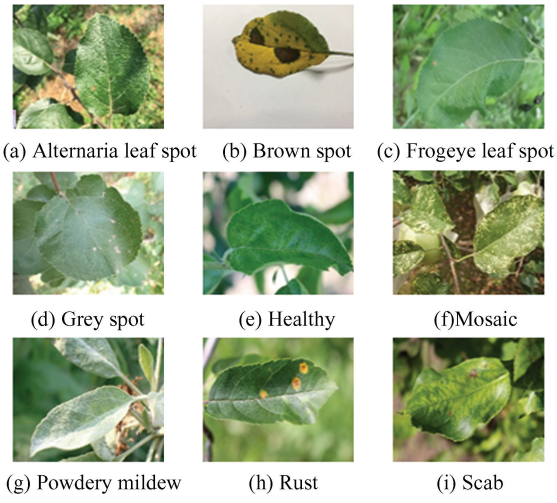


Fig.1 Example images from dataset

2 Construction of Apple Leaf Disease Recognition Model

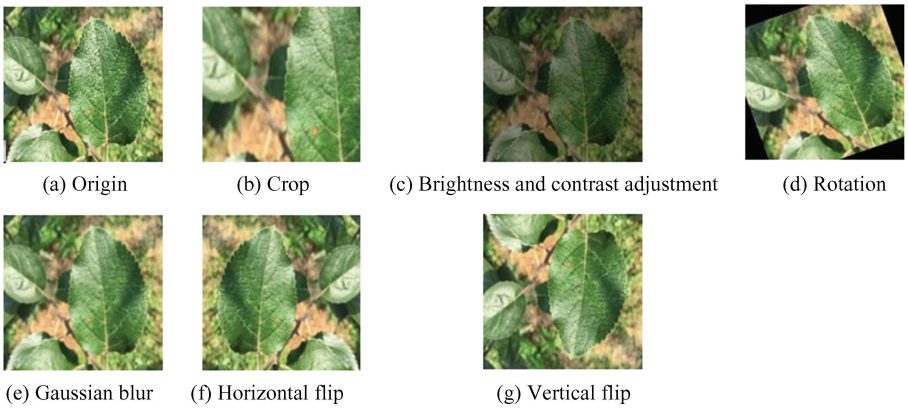
2.1 Principle of MobileNetV3-Small

MobileNetV3-Small[12] is a lightweight neural

network architecture that traces its origins back to 2017, when Google researchers first proposed version V1[13]. Then, based on version V2[14], MobileNetV3-Small was designed using neural network architecture search technology, making it more accurate and efficient than versions V1 and V2. MobileNetV3-Small incorporates Depthwise Separable Convolution (DSConv), Inverted Residual Block, SE attention mechanism[15] (Squeeze-and-Excitation, SE), and uses the h-swish activation function instead of ReLU. Separable convolution effectively decreases the parameter count by dividing the conventional convolution operation into two distinct steps: depthwise convolution (Dwise) and pointwise convolution (Pwise).

Table 1 The distribution of dataset

Categories	Label	Quantity
Alternaria leaf spot	A_1	417
Brown spot	A_2	411
Frogeye leaf spot	A_3	3181
Grey spot	A_4	339
Healthy	A_5	516
Mosaic	A_6	371
Powdery mildew	A_7	1184
Rust	A_8	2753
Scab	A_9	5410
Sum	–	14582



(a) Origin (b) Crop (c) Brightness and contrast adjustment (d) Rotation

(e) Gaussian blur (f) Horizontal flip (g) Vertical flip

Fig.2 Example images of dataset pre-processing

MobileNetV3-Small introduces a redesigned activation function, substituting the original ReLU with Hardswish, an enhanced variant of the swish activation function. The swish function is smoother than ReLU and can better capture the nonlinear features in images, thereby effectively improving the model's accuracy[16-17]. However, the exponential calculation involved in the swish function leads to increased computational complexity in the model. The Hardswish function approximates the swish function while requiring fewer computing resources, which makes it appropriate for scenarios with limited resources.

As shown in Fig. 3, MobileNetV3-Small incorporates the SE attention mechanism into the inverted residual module of version V2, forming a new block that serves as the core module of MobileNetV3-Small. The inverted residual module structure shown on the left side of Fig. 3 utilizes depthwise separable convolution and incorporates residual connections inspired by ResNet[18]. In the residual module of ResNet, the input feature map dimension is initially decreased and subsequently increased after convolution. Conversely, the inverted residual module first expands the input feature map dimension and then compresses it following convolution operations. The SE attention mechanism is a network based on channel attention. By using the SE module, the model can autonomously identify and learn the significance of various channels within feature maps. By assigning different weights to the channels, the model pays less attention to those with lower weights, thus suppressing less informative features and allocating higher attention to channels with greater importance, thereby enhancing the features of channels with more information. The SE attention module is composed of a global average pooling layer, two fully connected layers, and utilizes ReLU and Hsigmoid activation functions, as shown on the right side of Fig. 3.

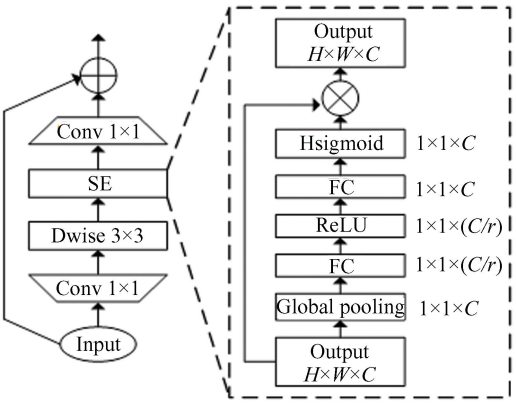


Fig.3 MobileNetV3-Small kernel block

Fig. 4 illustrates the architecture of the MobileNetV3-Small model, where the first Bneck is a depthwise separable module, and the rest of the Bneck blocks are as shown in Fig. 3. When the Bneck-based inverted residual module has the same number of input and output channels, and the stride required by the depthwise convolution operation is 1, the inverted residual modules use residual connections. Table 2 provides the detailed network parameters for the MobileNetV3-Small model. In Table 2, NBN signifies the absence of a convolutional layer following the batch normalization layer, #out indicates the number of output channels for each module, SE specifies whether the SE attention mechanism is utilized, and NL denotes the type of activation function used, #H is the h-swish function, #R is the ReLU function, and s is the stride length required for the depthwise convolution operation, k denotes the number of output channels in the model.

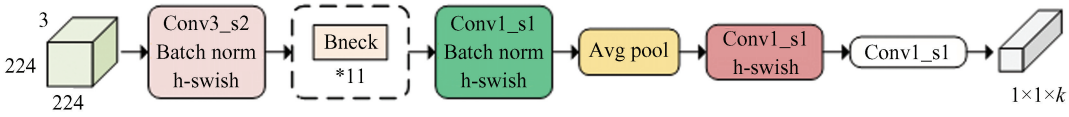


Fig.4 Structure diagram of MobileNetV3-Small

Table 2 Parameters of MobileNetV3-Small

Input	s	NL	SE	#out	Operator
$224^2\times3$	2	#H	–	16	Conv2d, 3×3
$112^2\times16$	2	#R	√	16	Bneck1,3×3
$56^2\times16$	2	#R	–	24	Bneck2,3×3
$28^2\times24$	1	#R	–	24	Bneck3,3×3
$28^2\times24$	2	#H	√	40	Bneck4,5×5
$14^2\times40$	1	#H	√	40	Bneck5,5×5
$14^2\times40$	1	#H	√	40	Bneck6,5×5
$14^2\times40$	1	#H	√	48	Bneck7,5×5
$14^2\times48$	1	#H	√	48	Bneck8,5×5
$14^2\times48$	2	#H	√	96	Bneck9,5×5
$7^2\times96$	1	#H	√	96	Bneck10,5×5
$7^2\times96$	1	#H	√	96	Bneck11,5×5
$7^2\times96$	1	#H	√	576	Conv2d,1×1
$7^2\times576$	1	–	–	–	Pool,7×7
$1^2\times576$	1	#H	–	1024	Conv2d 1×1,NBN
$1^2\times1024$	1	–	–	k	Conv2d 1×1,NBN

2.2 Multi-scale Feature Fusion Network

In convolutional neural networks, shallow features with a large scale have small receptive fields, allowing them to capture rich detailed information and specific location information, which is useful for detecting small-sized targets. On the other hand, deep features with a small scale have large receptive fields, enabling them to capture the overall structure of image features and rich semantic information, which is beneficial for detecting large-sized targets. Due to the loss of some image details during multiple downsampling operations through the network, the feature information of small targets is lost. Nevertheless, the weighted bi-directional feature pyramid network[19] (BiFPN) effectively captures Multi-scale feature information with only a minor increase in parameters, thereby preserving small-target details and boosting their detection performance. Fig. 5 presents the corresponding network architecture. BiFPN uses 1×1 convolution to connect feature maps with different channel numbers horizontally, resulting in feature maps that maintain an identical channel count. Multi-scale features are subsequently integrated via top-down and bottom-up fusion paths. In the top-down pathway, by fusing deep-layer features with shallow-layer features, the positional information from shallow layers is preserved, and the semantic information from deep layers is propagated downward. The bottom-up path effectively transfers the target location information of shallow features upward, enabling the feature maps derived from the network to encompass abundant location and semantic information. Since feature maps at different scales focus on objects of different sizes and contribute differently to feature fusion, BiFPN introduces bidirectional cross-scale connections and fast normalized fusion for weighted feature fusion to learn feature maps. Its calculation formula (1) is as follows:

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \cdot I_i \tag{1}$$

where O is the output feature map, I_i is the i-th input feature map, w_i is the learnable weight of the i-th input feature map, and $\sum_j w_j$ is the weight sum of all input feature maps. The ReLU activation function is applied to $\sum_j w_j$ to ensure that the value is greater than or equal to 0, and ε is set to 0.0001 to avoid gradient vanishing caused by the denominator being 0. After fast normalization, the weight range of the feature map is (0,1). Taking P_3^{out} as an example, the output feature map is derived through Eqs. (2) and (3):

$$P_3^{\text{td}} = \text{DSConv}\left(\frac{w_1 \cdot P_3^{in} + w_2 \cdot \text{Resize}(P_4^{\text{td}})}{w_1 + w_2 + \varepsilon}\right) \quad (2)$$

$$P_3^{\text{out}} = \text{DSConv}\left(\frac{w_1' \cdot P_3^{\text{in}} + w_2' \cdot P_3^{\text{td}} + w_3' \cdot \text{Resize}(P_2^{\text{out}})}{w_1' + w_2' + w_3' + \varepsilon}\right) \quad (3)$$

where P_3^{td} and P_4^{td} are the intermediate feature maps of the 3rd and 4th feature layers in the top-down feature fusion path, P_3^{out} and P_2^{out} are the output feature maps of the 3rd and 2nd feature layers in the bottom-up feature fusion path. Resize represents the up-sampling or down-sampling operation, w is the connection weight between the feature maps, and DSConv is the depthwise separable convolution. The primary goal of employing DSConv is to lower the network's parameter count.

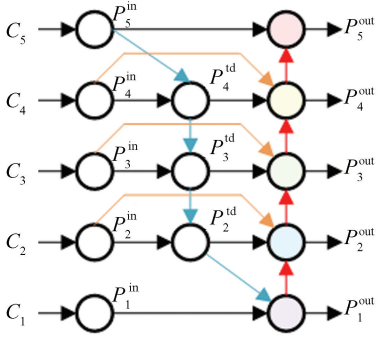


Fig.5 Structure diagram of BiFPN

2.3 Efficient Multi-Scale Attention

Since the apple planting environment is typically outdoors, in most of the apple leaf disease images gathered during the experiment, complex backgrounds obstruct disease recognition and impair the model's capacity to detect smaller disease details. In this work, an Efficient Multi-scale Attention[20] (EMA) module is integrated to strengthen the model's robustness against background interference, allowing it to focus more effectively on disease regions. The network architecture is illustrated in Fig. 6.

EMA reorganizes a portion of the input feature map's channel dimension into the batch dimension by grouping features. The input feature map $\boldsymbol{X} \in \mathbf{R}^{C\times H\times W}$ is evenly divided into G sub-feature maps along the cross-channel dimension, ensuring that the spatial semantic features are evenly distributed among the sub-features, which can be represented as $\boldsymbol{X} = [\boldsymbol{X}_0, \boldsymbol{X}_i, \cdots, \boldsymbol{X}_{G-1}]$, $\boldsymbol{X}_i \in \mathbf{R}^{(C//G)\times H\times W}$, with $G \ll C$ generally being used. The symbol "//" in "$C//G$" indicates rounding down.

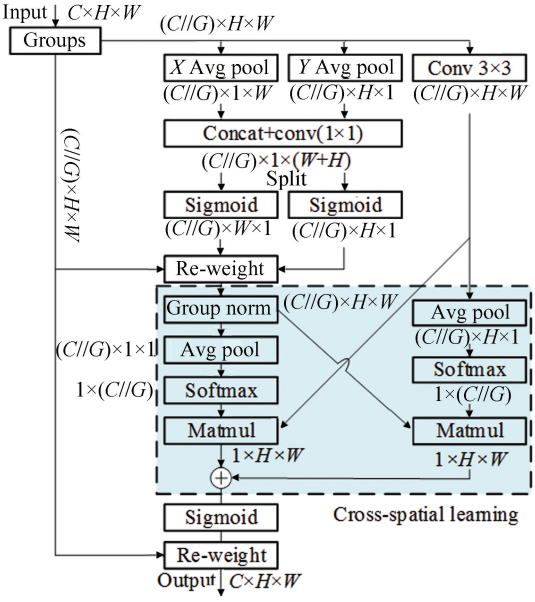


Fig.6 Structure diagram of EMA

The multi-scale parallel network substructure is used to process sub-features, model cross-channel information interaction, and derive attention weight descriptors from the sub-feature maps. The parallel structure labels the subnetwork where the X Avg pool path and Y Avg pool path are located as 1×1 branches, and the subnetwork where the 3×3 convolution operation is located as 3×3 branches. The 1×1 branches encode the channels using one-dimensional global average pooling along the X-axis and Y-axis directions of the sub-feature map. The encoded channel information is concatenated along the Y-axis direction of the sub-feature map using a shared 1×1 convolution. The output from the 1×1 convolution is directly split into two feature tensors, and two channel attention maps are generated using the Sigmoid activation function. The original input sub-feature maps are then aggregated by multiplication to produce the output tensors of the 1×1 branches. In order to facilitate cross-channel interaction between the two pathways in the 1×1 branch, the model

focuses on important channel features, minimizing the loss of significant channel information. The 3 × 3 branch employs a single 3×3 convolution to capture diverse receptive fields, thereby boosting the ability to extract multi-scale features.

EMA uses cross-spatial learning to effectively model short-distance and long-distance dependencies between features. The output tensor of the 1 × 1 branch, after group normalization, is encoded using two-dimensional global average pooling, and the output tensor $\boldsymbol{O}_1 \in \mathbf{R}_1^{1\times(C//G)}$ is activated by the Softmax function. The matrix dot product of $\boldsymbol{O}_1$ and the 3×3 branch output tensor $\boldsymbol{O}_3 \in \mathbf{R}_3^{(C//G)\times HW}$, after the dimensional reshaping process, it produces the initial spatial attention map, capturing spatial information at various scales. The 3 × 3 branch output tensor is activated by the Softmax function after 2-D global average pooling to produce the output tensor $\boldsymbol{T}_3 \in \mathbf{R}_3^{1\times(C//G)}$. The 1 × 1 branch output tensor $\boldsymbol{T}_1 \in \mathbf{R}_1^{(C//G)\times HW}$ is dimensionally reshaped and then multiplied by $\boldsymbol{T}_3$ using the matrix dot product to generate a second spatial attention graph, collecting more precise spatial information. The two generated spatial attention maps are added, then activated using the Sigmoid function, and finally, it is multiplied by the original sub-feature map, yielding the EMA's final output. Cross-space learning integrates global context information and local features, enabling EMA to highlight global context relationships while capturing pixel-level correlations.

2.4 MobileNetV3-Small Model Improvement

To improve MobileNetV3-Small's feature extraction performance, this study uses MobileNetV3-Small as the backbone network and presents an enhanced MobileNet-BiFPN-EMA model that integrates BiFPN with EMA. Its structure is shown in Fig.7.

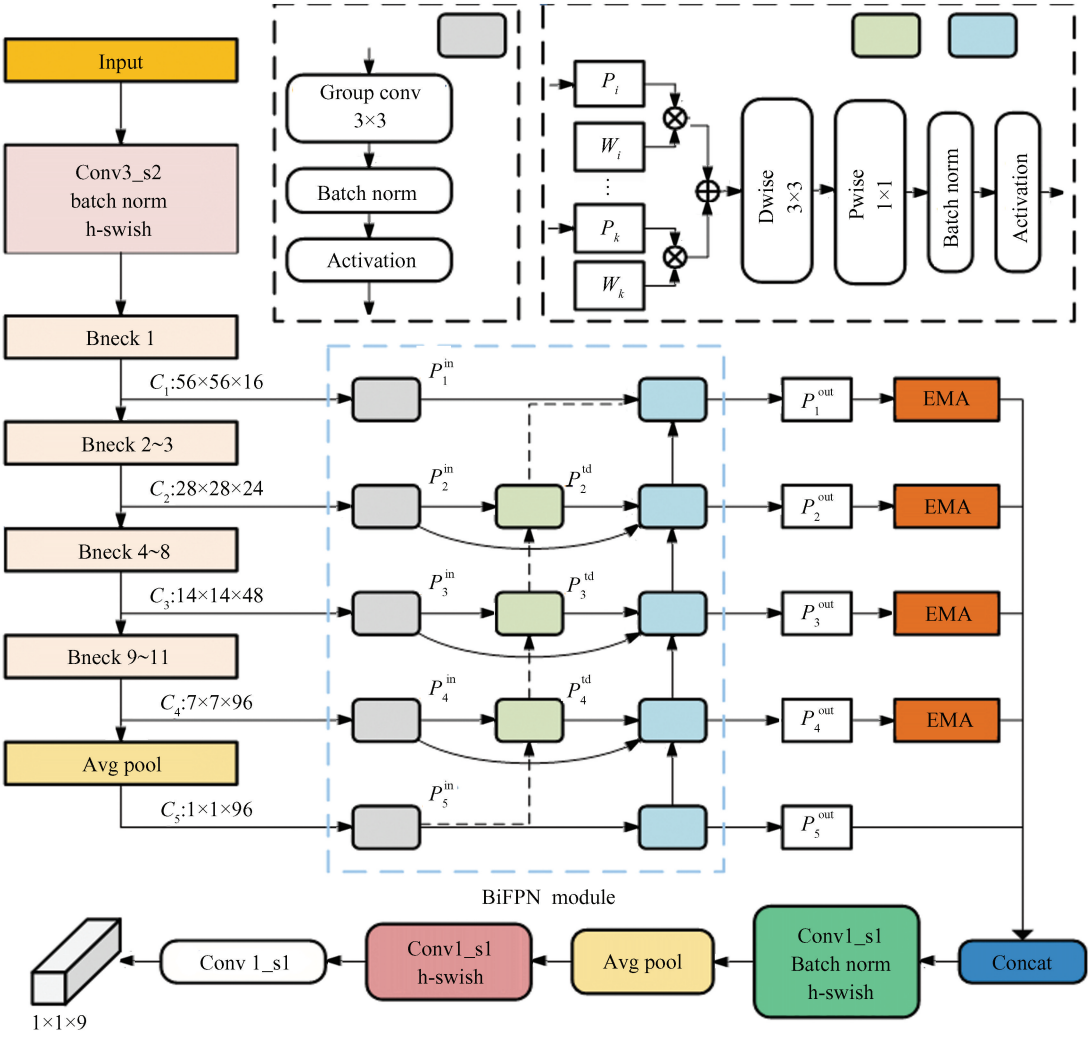


Fig.7 Structure diagram of MobileNet-BiFPN-EMA

Since nine disease categories need to be identified in this paper, the classification output number of the MobileNetV3-Small model is set to 9 in the improved model. The size of the feature map output by the Bneck11 module is reduced to a pixel size using adaptive average pooling to obtain the feature map C_5, which is input to the BiFPN module through lateral connections with C_1 to C_4.

To strengthen the local texture extraction capabilities within the BiFPN module, a 3×3 group convolution with reduced parameters is employed for lateral connections to consistently adjust the channel dimension of the C_1 through C_5 feature maps extracted from the backbone network to 64. Since the EMA module adopts cross-channel information interaction and captures pixel-level fine-grained features in the image through cross-space learning, the model can more precisely pinpoint various feature locations of the target object, reducing the impact of background interference. Therefore, the EMA module is added after the output features of the BiFPN module, from maps $P_1^{\text{out}} \sim P_4^{\text{out}}$, to further augment the model's capability to distinguish disease regions from intricate backgrounds across multiple scales, thereby enhancing overall performance.

3 Experiment and Result Analysis

3.1 Experimental Environment and Parameter Setting

The experiment ran on a 64-bit Windows 10 operating system, utilizing the PyTorch 2.3.1 deep learning framework, the CUDA 11.8 parallel computing architecture, and the cuDNN 8.7.0 library. PyCharm served as the development environment, and Python 3.8.18 was used as the programming language. The CPU model was Intel © Core™ i5-10200H CPU @ 2.40GHz, and the GPU model was GTX 1650.

The initial learning rate was set to 0.001, with the model trained using the Adam optimizer. The batch size was established at 16, the input image size was adjusted to 224×224, and the number of training epochs was specified as 160. The dropout layer's deactivation rate was set to 0.2 to prevent overfitting.

3.2 Experimental Results of the Improved Model

The confusion matrix is used to better present the model's test results, as shown in Fig. 8. The row labels of the confusion matrix represent the true labels of the samples, while the column labels represent the predicted values of the model for the samples. The diagonal values in Fig. 8 indicate the number of correct identifications of each disease category by the model. From the confusion matrix, it can be seen that the improved model has lower recognition accuracy for leaf spot disease and gray spot disease but higher recognition accuracy for other diseases. Specifically, the improved model misidentifies 6 leaf spot disease samples as rust, 4 leaf spot disease samples as gray spot disease, and 5 gray spot disease samples as leaf spot disease. This may be due to the similar contours and sizes of the three diseases, which increase the likelihood of misidentification by the model.

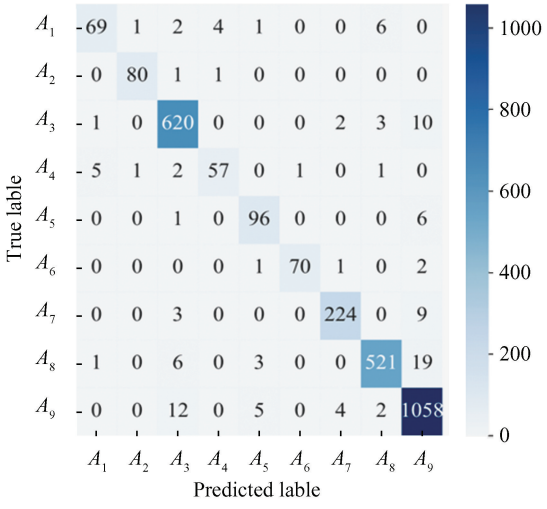


Fig.8 Confusion matrix of the predictions by the improved model

3.3 Ablation Experiment

To assess the effectiveness of the proposed improvements, the BiFPN module is incorporated into the MobileNetV3-Small network, creating the MobileNet-BiFPN model. To further evaluate the effectiveness of the EMA Attention Mechanism proposed in this paper, the SE Attention Mechanism and the Coordinate Attention (CA) Mechanism were substituted for the EMA Attention Mechanism in the MobileNet-BiFPN-EMA model, resulting in the MobileNet-BiFPN-SE model and the MobileNet-BiFPN-CA model. Table 3 displays the test results for each model, where the inference time denotes the duration required for the model to process a single image. After adding the BiFPN module to the benchmark model, the model´s accuracy improves by 0.58 percentage points, with the average recall and average F1-score rising by 2.08 and 0.22 percentage

points, respectively. The model's parameter count increases by 0.19 M and model size increases by 0.82 MB, which is relatively small and can still handle the memory limitations of agricultural electronic equipment. The FLOPs of the model increased by 41.6 M, and the inference time increased by 6.96 ms.

Table 3 Results of the ablation experiments

Model	Accuracy (%)	Params (M)	FLOPs (M)	Average precision (%)	Average recall (%)	Average F1-Score (%)	Inference time (ms)	Model size (MB)
MobileNetV3-Small	95.05	1.53	61.18	94.80	91.44	93.04	5.10	6.23
MobileNet-BiFPN	95.67	1.72	102.78	93.08	93.52	93.26	12.06	7.05
MobileNet-BiFPN-EMA	95.98	1.72	123.16	95.09	93.17	94.09	14.10	7.07
MobileNet-BiFPN-SE	95.60	1.73	102.79	94.24	92.75	93.44	12.31	7.09
MobileNet-BiFPN-CA	95.78	1.73	103.55	94.03	92.44	93.18	13.61	7.09

The MobileNet-BiFPN-EMA model achieves an accuracy of 95.98%, surpassing the MobileNet-BiFPN model by 0.31 percentage points. The average recall and average F1-score increase by 2.01 and 0.83 percentage points, respectively, demonstrating that the EMA attention mechanism enhances model accuracy without adding extra parameters. Compared with the MobileNetV3-Small model, the MobileNet-BiFPN-EMA model experienced an increase of 61.98 M in FLOPs, a 9 ms rise in inference time, and a 0.84 MB increase in model size.

Compared with the improved model using the SE attention mechanism and the CA attention mechanism, the improved model using the EMA attention mechanism has fewer parameters and a smaller model size. Additionally, the model accuracy is increased by 0.38 percentage points compared with the SE attention mechanism and by 0.2 percentage points compared with the CA attention mechanism. This indicates that the EMA attention mechanism is more suitable for the recognition task than the SE and CA attention mechanisms.

The loss curve of the training set for the model before and after improvement is shown in Fig. 9. The training loss value of the improved model decreases the fastest in the early stages and is lower than that of the MobileNetV3-Small and MobileNet-BiFPN models in the later stages, gradually stabilizing at 0.13.

This paper uses Grad-CAM technology to visualize the improvements in the model. Grad-CAM creates a heatmap by merging gradient information with feature maps from convolutional neural networks, enabling a clear identification of the image regions the model focuses on during decision-making. The redder the color of a region in the heatmap, the more attention the model gives to that region. The bluer the region, the less attention the model gives to it[21]. Fig. 10 displays the heatmap both before and after the model improvements. As shown in Fig. 10, when the feature distribution of leaf diseases is more dispersed, the improved model can capture the disease region more completely; when the feature distribution of leaf diseases is more concentrated, the improved model can locate the disease feature region more accurately, mainly because the BiFPN module improves the model's capacity to capture features at multiple scales. The EMA attention mechanism enables the model to precisely focus on key disease features.

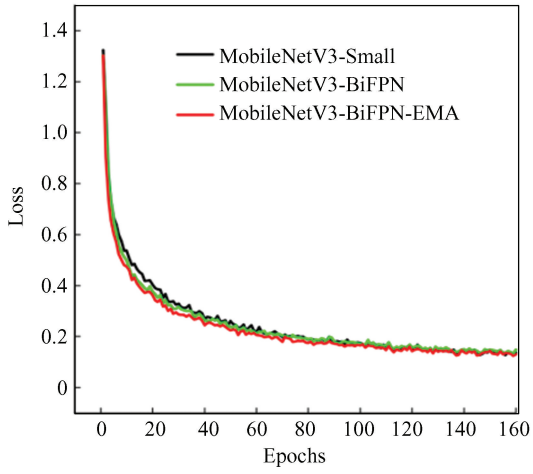


Fig.9 Comparison of the training set loss curves of the model before and after improvement

3.4 Comparative Experiment

To assess the improved model's performance in recognizing apple leaf diseases, eight lightweight neural network models, including ShuffleNet_v2_1.5×, MobileNetV2, ResNet50, MobileNetV3-Large, EfficientNet-B0, MobileNetV3-Small, MobileNetV4-Conv-Small and MobileNetV4-Conv-Medium, were selected for comparison experiments. Each model was trained on the same training dataset for 160 epochs with identical hyperparameter settings and in the same

hardware environment. Table 4 presents the experimental results for the test set. As shown in Table 4, the accuracy of all comparison models exceeded 94%, with EfficientNet-B0 achieving the highest accuracy at 95.90% among the eight models. However, this model exhibited higher FLOPs and model size.The FLOPs of the benchmark model MobileNetV3-Small were 14.87% of those of EfficientNet-B0, and its model size was 38.10% of that of EfficientNet-B0, while achieving an accuracy of 95.05%. These results demonstrate that MobileNetV3-Small delivers high accuracy with lower resource consumption. The improved model achieves the highest accuracy of 95.98%, with lower parameter count, FLOPs, inference time, and model size compared with all other models, except the benchmark model. Furthermore, the improved model surpasses the other models in average accuracy and F1-score. When compared with MobileNetV2, it reduces the parameter count and FLOPs by 23.21% and 62.25%, respectively. When compared with EfficientNet-B0, it achieves reductions of 57.21% and 70.07%, respectively, demonstrating its suitability for memory-constrained device environments. Furthermore, the model's inference time is only 14.10 ms, this allows for the swift and precise identification of apple leaf diseases.

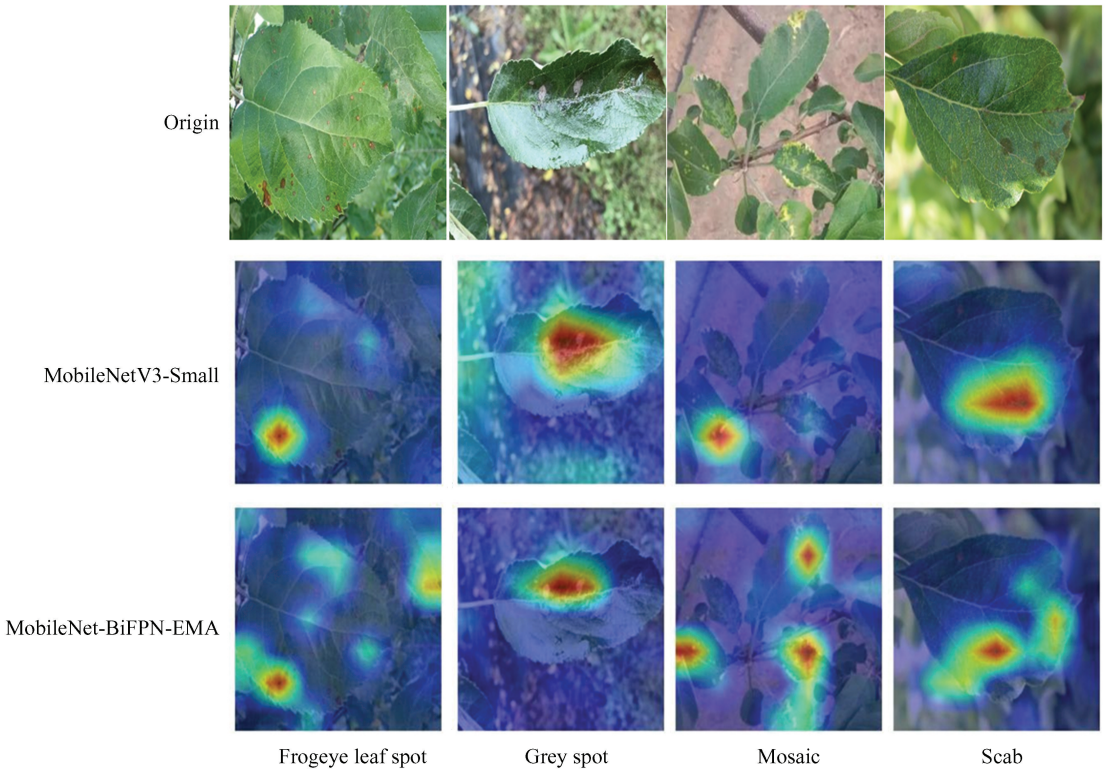


Fig.10 Comparison of the heatmaps before and after model improvement

Table 4 Results of the contrast experiments

Model	Accuracy (%)	Params (M)	FLOPs (M)	Average precision (%)	Average recall (%)	Average F1-Score (%)	Inference time (ms)	Model size (MB)
MobileNet-BiFPN-EMA	95.98	1.72	123.16	95.09	93.17	94.09	14.10	7.07
ShuffleNet_v2_1.5×	95.71	2.49	308.82	94.56	92.48	93.47	7.04	10.14
MobileNetV2	95.78	2.24	326.28	93.00	92.83	92.84	5.87	9.16
ResNet50	95.71	23.53	4132.00	93.89	93.26	93.53	8.15	94.40
MobileNetV3-Large	95.78	4.21	232.97	93.53	93.87	93.62	6.84	17.04
EfficientNet-B0	95.90	4.02	411.56	93.63	93.21	93.39	8.97	16.35
MobileNetV3-Small	95.05	1.53	61.18	94.80	91.44	93.04	5.10	6.23
MobileNetV4-Conv-Small	94.37	3.03	309.00	92.23	90.35	91.14	14.23	12.30
MobileNetV4-Conv-Medium	95.40	8.42	854.63	92.68	92.46	92.52	8.91	34.07

3.5 Data Augmentation Experiment

Plant disease identification often encounters the challenge of limited sample data. This study employs six online data augmentation techniques—random cropping, brightness and contrast adjustment, random rotation, Gaussian noise addition, horizontal flipping, and vertical flipping—to process the original images in multiple ways, simulating different acquisition conditions helps increase the diversity of training samples. These augmentations enable the model to learn more comprehensive features during training, thereby improving its generalization ability.To evaluate the effectiveness of data augmentation, we compared the experimental results of the benchmark model MobileNetV3-Small and the improved model MobileNet-BiFPN-EMA using non-augmented training datasets with those obtained from augmented training datasets. Fig. 11 displays the loss curves of the models during training, while Table 5 presents the results from the test set. As shown in Fig. 11, the loss curves of both the benchmark and improved models in the augmented dataset exhibit faster convergence during the first 40 epochs, smoother convergence in later stages, and lower loss values compared to the unaugmented dataset, which shows slower convergence. As shown in Table 5, the use of augmented data results in a 4.56 percentage point improvement in the accuracy of the benchmark model. Additionally, the average accuracy, recall, and F1-scores improve by 6.56,5.82, and 6.29 percentage points, respectively. The enhanced model shows a 3.16 percentage point increase in accuracy, with average accuracy, recall, and F1-scores improving by 5.16, 4.49, and 4.83 percentage points, respectively. These results demonstrate that data augmentation significantly enhances the model´s disease identification capability.

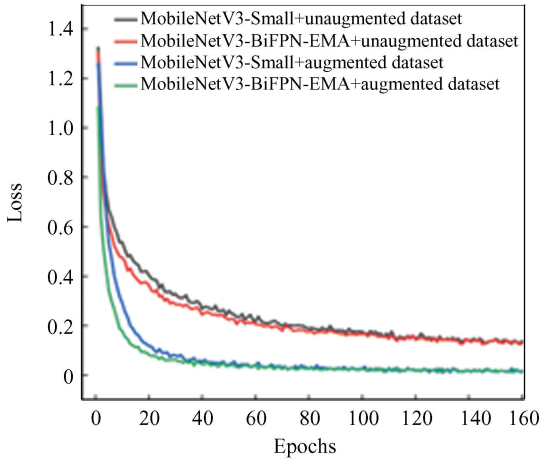


Fig.11 Comparison of the training set loss curves of the model before and after improvement

Table 5 Classification results of data augmentation experiment

Model	Accuracy(%)	Average precision(%)	Average recall(%)	Average F1-Score(%)
MobileNetV3-Small + unaugmented dataset	90.49	88.24	85.62	86.75
MobileNet-BiFPN-EMA+ unaugmented dataset	92.82	89.93	88.68	89.26
MobileNetV3-Small+ augmented dataset	95.05	94.80	91.44	93.04
MobileNet-BiFPN-EMA+ augmented dataset	95.98	95.09	93.17	94.09

4 Conclusions

This paper presents a lightweight, improved model—MobileNet-BiFPN-EMA—based on the MobileNetV3-Small architecture, aimed at enhancing the recognition accuracy of apple diseases. By incorporating the BiFPN module and the EMA attention mechanism, the model effectively leverages both shallow and deep features, thereby improving its feature extraction performance. Comparative experimental results reveal that, when tested on the AppleLeaf9 dataset, the improved model outperforms the original MobileNetV3-Small model, with increases of 0.93, 0.29, 1.73, and 1.05 percentage points in accuracy, average precision, average recall, and average F1-score, respectively. Furthermore, the model's parameter count and size are only 1.72 M and 7.07 MB, respectively, indicating an optimal balance between model size and recognition performance, making it highly suitable for memory-constrained environments. Ablation experiment showed that the improved model better identified scattered disease features after the BiFPN module was integrated, and the identification of disease features was further improved by adding the EMA attention mechanism. Data enhancement experiment showed that recognition accuracy could be effectively improved by using the enhanced training set.

This study has several limitations:

1) First, the model encounters difficulties in accurately identifying three specific diseases: spot leaf drop, gray spot disease, and rust disease. Additionally, the dataset is limited in the diversity of

apple leaf diseases it includes. To address these limitations, we plan to expand the dataset by collecting disease images from both online and offline sources.

2) While the current dataset contains only one type of crop disease, future efforts will focus on acquiring images of various crop diseases from natural environments, which will enhance and validate the model's generalization capability across different agricultural diseases.

3) The size and computational complexity of the improved model can still be further reduced. In subsequent stages, optimization techniques, including model pruning and quantization, will be explored to refine the model. Additionally, the optimized model will be deployed in agricultural electronic equipment to better identify and address challenges encountered in production practices.

References

[1] Huo X X, Liu T J, Liu J D, et al. China apple industry development report (abridged edition). China Fruit & Vegetable, 2022, 42(2): 1-6. DOI: 10.19590/j.cnki.1008-1038.2022.02.001.

[2] Wang S T, Wang Y N, Cao K Q. Occurrence of and research progress in important apple disease in china in recent years. Plant Protection, 2018, 44(5): 13-25, 50. DOI: 10.16688/j.zwbh.2018300.

[3] Kamilarisk A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 2018, 147: 70-90. DOI: 10.1016/j.compag.2018.02.016.

[4] Yu M, Guo Z Y, Wang Y. Review of computer vision-based plant disease identification techniques. Science Technology and Engineering, 2024, 24(23): 4811-4823.

[5] Zhang F K, Jin X B, Lin G, et al. Hybrid attention network for citrus disease identification. Computers and Electronics in Agriculture, 2024, 220: 108907. DOI: 10.1016/j.compag.2024.108907.

[6] Zhang H T, Luo Y M, Tan L. Research on millet disease identification based on transfer learning and residual network. Journal of Henan Agricultural Sciences, 2023, 52(12): 162-171. DOI: 10.15933/j.cnki.1004-3268.2023.12.018.

[7] Liang X M, Gao S P, Liu Z D. Exploration of identifying apple leaf diseases using lightweight convolutional neural network model. China Plant Protection, 2024, 44(4): 41-49.

[8] Chen J D, Wang W H, Zhang D F, et al. Attention embedded lightweight network for maize disease recognition. Plant Pathology, 2021, 70(3): 630-642. DOI: 10.1111/ppa.13322.

[9] Hu S W, Deng J X, Wang H Y, et al. Grape leaf disease identification method based on improved EfficientNetB0 model. Modern Electronics Technique, 2024, 47(15): 73-80. DOI: 10.16652/j.issn.1004-373x.2024.15.012.

[10] Guo H P, Cao Y Z, Wang C S, et al. Recognition and application of apple defoliation disease based on transfer learning. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(3): 184-192.

[11] Yang Q, Duan S K, Wang L D. Efficient identification of apple leaf diseases in the wild using convolutional neural networks. Agronomy, 2022, 12(11): 2784.

[12] Howard A, Sandler M, Chu G, et al. Searching for MobilenetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 1314-1324. DOI: 10.1109/ICCV.2019.00140.

[13] Howard A G, Zhu M L, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861, 2017.

[14] Sandler M, Howard A, Zhu M L, et al. MobilenetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520. DOI: 10.1109/CVPR.2018.00474.

[15] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[16] Ramachandran P, Zoph B, Le Q V. Searching for activation functions. arXiv preprint arXiv: 1710.05941, 2017.

[17] Elfwing S, Uchibe E, Doya K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks, 2018, 107: 3-11. DOI: 10.1016/j.neunet.2017.12.012.

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[19] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10781-10790. DOI: 10.1109/CVPR42600.2020.01079.

[20] OuYang D L, He S, Zhang G Z, et al. Efficient multi-scale attention module with cross-spatial learning. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5. DOI: 10.1109/ICASSP49357.2023.10096516.

[21] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 618-626. DOI: 10.1109/ICCV.2017.74.