pagepress

Fine-grained recognition algorithm of crop pests based on cross-layer bilinear aggregation and multi-task learning

Juquan Ruan, Shuo Liu, Wanjing Mao, Shan Zeng, Zhuoyi Zhang, Guangsun Yin

School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, China

Abstract

Fine-grained recognition of crop pests is a crucial concern in the field of agriculture, as the accuracy of recognition and generalization ability directly affect the yield and quality of crops. Aiming at the characteristics of crop pests with a wide variety of species, small inter-class and large intra-class differences in external morphology, as well as the problems of uneven sample distribution and noisy labels in fine-grained image datasets under complex environments, we propose a fine-grained recognition model of crop pests (MT-MACLBPHSNet) based on cross-layer bilinear aggregation and multi-task learning, which consists of three key modules: the backbone network module, the cross-layer bilinear aggregation module, and the multi-task learning module. A new union loss function is designed in the primary task of the multi-task learning module, which is used to alleviate the two problems existing in the model training fine-grained image datasets. The experimental results show that the model effectively balances the model complexity and recognition accuracy in a comparative analysis with several existing excellent network models on the IP102-CP13 dataset, with the recognition accuracy reaching 75.37%, which is 7.06% higher than the baseline model, and the F1-score reaching 67.06%. Additionally, the generalization of the model is also verified on the IP102-VP16 dataset, and the model outperforms most of the models in terms of recognition accuracy and generalization ability, which can provide an effective reference for fine-grained recognition of crop pests.

Correspondence: Shuo Liu, School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430048, China. E-mail: 874477154@qq.com

Contributions: all authors made a substantive intellectual contribution, read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no competing interests, and all authors confirm accuracy.

Acknowledgements: this work was supported by Wuhan Polytechnic University in the project under Grant 2023Y42, Hubei's Key Project of Research and Development Program under Grant 2023BBB046, excellent young and middle-aged scientific and technological innovation teams in colleges and universities of Hubei Province under Grant T2021009, and NSFC-CAAC under Grant U1833119.

Received: 7 November 2023.
Early view: 15 March 2024.

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

As a largely agricultural country, agriculture is one of the most important basic industries in China. According to the comprehensive analysis of relevant factors such as the source base of pests, cultivation practices, and climate trends, it is predicted that major crop pests and diseases in China will show an obvious recurrence trend in 2023. This will severely threaten the safety of food production and cause significant economic losses to agricultural production. At present, crop pests are still one of the main hazardous factors affecting agricultural production and need to be monitored, given early warning, and controlled on time to protect the yield and quality of crops. However, crop pests are characterized by a wide variety of species, inter-class similarity in external morphology, and intra-class variability, which makes their classification more challenging. Therefore, with the increasing level of agricultural modernization and intelligence, how to accurately, scientifically, and efficiently achieve fine-grained recognition of crop pests has become a key task in the safe production of smart agriculture. Crop pest diagnosis based on traditional machine vision technology mainly relies on manual intervention in feature design, which is both time-consuming and energy-consuming and prone to misjudgment. In recent years, the rapid development of deep learning technology has allowed it to automatically extract crop pest features for recognition in an end-to-end manner, effectively avoiding the subjective factors of manual feature extraction. It has made good progress and been widely used in the research field of conventional classification and fine-grained recognition of crop pests. For example, Li *et al.* (2020) first introduced multiple preprocessing methods to remove the natural background of pests, and then, with the help of a fine-tuned GoogLeNet (Szegedy *et al.*, 2015) network model, performed ten classes of crop pest conventional classification and achieved better recognition results than the ResNet-101 (He *et al.*, 2016) model. Thenmozhi and Reddy (2019) proposed a classification model of crop pests based on a deep convolutional neural network and transfer learning (Li *et al.*, 2022), and the experimental results indicated that it was effective in classifying pests in field crops. Nanni *et al.* (2020) proposed an automatic classifier based on the fusion of the saliency method and a convolutional neural network to classify 10 species of pests and achieve better recognition accuracy. Li *et al.* (2020) proposed a convolutional neural network (CNN) model based on optimized GoogLeNet to identify corn borer pests, which effectively improved the recognition performance. Wang *et al.* (2021) embedded the improved convolutional attention module into AlexNet, MobileNetV2 (Sandler *et al.*, 2018), and other CNNs for fine-grained recognition of crop pests and diseases, which resulted in a certain improvement in accuracy. Wei *et al.* (2022) proposed a multi-scale feature fusion network-based approach to crop pest classification that achieved better classification performance on a dataset of twelve types of pests. Zhang *et al.* (2023) designed a structural optimization transmission network (SOT-Net) for hyperspectral image (HSI) and light detection and ranging (LiDAR) data classification, where the information transmission process consists of spectral structure branches optimized using a cross attention mechanism and geometrical structure (spatial and elevation) branches optimized by developing a symmetrical dual-modes propagation module, and integrates self-aligned regularization into the multi-source collaborative classification task during inference, which enhances the robustness of the feature extraction and classification process. Wang *et al.* (2023) designed a multistage self-guided separation network (MGSNet) for remote-sensing scene classification based on the target-background separation strategy and the introduction of contrastive regularization (CR) and achieved good classification performance on three benchmark test sets. Wang *et al.* (2023) proposed a representation-enhanced status replay network (RSRNet) for multisource remote-sensing image classification, which combines modal augmentation and semantic augmentation to learn the structural features of the embedding space, fuses multisource information with the cross-modal interactive fusion (CMIF) method, and utilizes the status replay strategy to mitigate the bias of the decision boundary of the classifier, making it superior. Zheng *et al.* (2023) designed a deep residual network based on multi-scale feature extraction to recognize rice pests and achieved better classification results on a dataset of twenty-two types of common rice pests. Although recognition algorithms of crop pests based on deep neural networks have made great progress in recent years, most of them are only single tasks to study conventional classification and fine-grained recognition of crop pests, and their recognition accuracy and generalization ability need to be further improved. Moreover, fine-grained recognition of crop pests remains an exceptionally challenging task, and one of the important research hotspots is how to achieve discriminative feature extraction and characterization of crop pests' interclass differentiation and intra-class similarity. Finally, due to the high computational complexity and number of parameters in most of the existing models, they fail to achieve an effective balance between model complexity and recognition accuracy and are difficult to be deployed and applied in practical production.

Aiming at the above challenges, we propose a crop pest fine-grained recognition model based on cross-layer bilinear aggregation and multi-task learning using HOR-Shuffle-CANet (Ruan and Liu, 2023) as the baseline model and use the IP102-CP13 dataset and the IP102-VP16 dataset as research objects to validate, which is intended to provide new ideas to meet the practical needs of smart agriculture production. The contributions of this paper are summarized as follows:

- A lightweight CNN model (MT-MACLBPHSNet) based on cross-layer bilinear aggregation and multi-task learning is proposed for fine-grained recognition of crop pests. In particular, the backbone network module improves the baseline model with activation function optimization, depth-wise convolutional kernel enlargement, and embedded improved pyramid split attention to extract pest fine-grained features at multiple scales; the cross-layer bilinear aggregation module improves the fine-grained feature representation of the model by fusing different levels of features with Hadamard product operations; and the multi-task learning module employs image feature reconstruction as an auxiliary task to collaborate with the primary task of fine-grained classification of pests, complementing the feature information between the two tasks to enhance the recognition effect.
- A new union loss function based on the combination of the softmax equalization loss and the bi-tempered logistic loss learning strategy is designed to optimize the training model, which is used to alleviate the problems of long-tailed distribution and noisy data for the fine-grained image dataset in the field environment, and to make the model have strong generalization performance.
- Comprehensive experiments were conducted with the extremely challenging IIP102-CP13 dataset and IP102-VP16 dataset as experimental materials, and abundant experimental results demonstrate the superiority of the MT-MACLBPHSNet model, which efficiently balances the number of model parameters, floating-point computation, and recognition accuracy, and is characterized by its convenient migration and easy deployment.

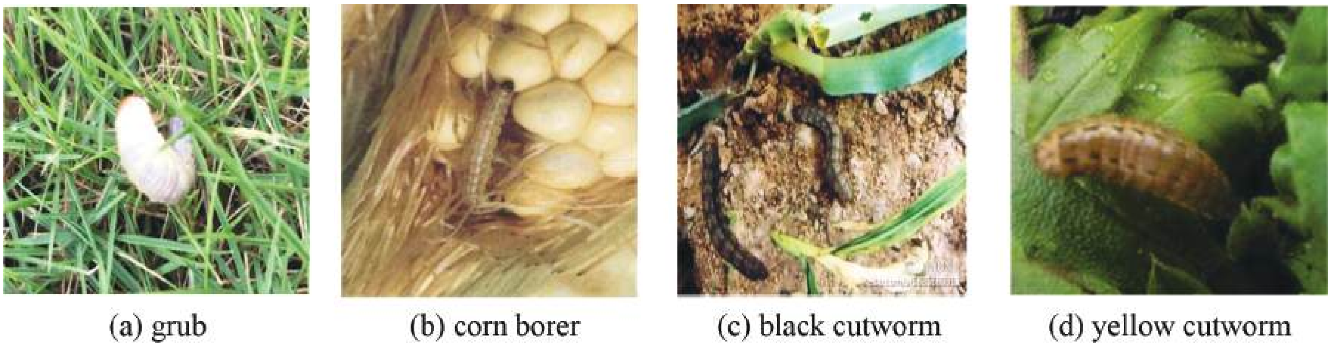

(a) grub
(b) corn borer
(c) black cutworm
(d) yellow cutworm

Figure 1. Sample images from the IP102-CP13 dataset.

Materials and Methods

Dataset

To validate the effectiveness and feasibility of the proposed fine-grained recognition algorithm of crop pests, the IP102-CP13 dataset was constituted with images of 13 classes of corn pests with the highest imbalance ratio (IR) of farmland crops from the largest and most challenging open-source benchmark for pest images in field environments, IP102 (Wu *et al.*, 2019), which was utilized as a research subject for ablation study and comparison study, as shown in Figure 1. Each class of crop pest images in this dataset exhibits inter-class similarity, intra-class diversity, and long-tailed distribution, which can accurately describe the real and complex agricultural practical production environment. Consequently, this dataset holds good practical significance and can better evaluate the performance of crop pest fine-grained recognition algorithms. Among them, detailed information about labels, species names, and numbers of images in the IP102-CP13 dataset is provided in Table 1, which will be used for experiments.

To verify the generalization ability of the proposed fine-grained recognition algorithm of crop pests, the 16 classes of vitis pest images with the highest IR of economic crops in IP102 were used to constitute the IP102-VP16 dataset, which was used as the research material for the generalization study.

As a majority of images in the above datasets have varying sizes, to reduce recognition errors caused by irregular image sizes and to meet the input image size requirements of subsequent network models, the images are uniformly cropped and adjusted to 224×224, and then the online data augmentation of the images is continued on this basis.

Data augmentation is a common method for extending data in deep learning. The use of online data augmentation can avoid the problem of storage load brought about by traditional offline data augmentation methods while increasing the random diversity of samples. In this paper, online data augmentation includes various operations such as random rotation, horizontal flipping, vertical flipping, color perturbation, noise addition, *etc*.

Network architecture

The MT-MACLBPHSNet recognition model is composed of three main components: the backbone network module, the cross-layer bilinear aggregation module, and the multi-task learning module. The overall architecture of this model is illustrated in Figure 2.

When a crop pest image to be recognized is input to the MT-MACLBPHSNet model:

- It first enters the backbone network module for feature extraction. This involves passing through a shared encoder composed of one convolutional layer and three CA-Res modules to extract shallow-level common features. Subsequently, downsampling is performed using the max-pooling layer that retains essential features, followed by multi-scale extraction of the image's more fine-grained deep-level semantic features by the interactive layers of CA-Stage and IPSA modules, and subsequently, a convolutional layer is used to blend these features. Among them, the three CA-Stages are modular layers composed of Shuffle-CA Unit2 and Shuffle-CA Unit1, and the numbers of Unit2 and Unit1 in the modular layers are 1:3, 1:7, and 1:3.
- It then enters the cross-layer bilinear aggregation module for feature fusion. Here, the shallow-level common feature maps output F_S from the shared encoder, the deep-level semantic feature maps output F_D from the CA-Stage2 modular layer, and the global feature maps output F_G from the final Conv2 convolutional layer are sequentially fused across layers using Hadamard product operations, and the feature vectors are obtained by subsequent transformations.
- Finally, it enters the multi-task learning module for fine-grained classification of pests and image feature reconstruction. The primary task is composed of a fully connected layer with 13 neurons, which generates the predicted category labels of the input pest image. While the auxiliary task is divided into two parts: a shared encoder and a decoder, the decoder is composed of one transposed convolutional layer (TConv) and one convolutional layer (Conv), which accomplish the feature reconstruction of the input image. The whole model is trained by jointly optimizing the two tasks with a weight combination of a new

Table 1. IP102-CP13 dataset information.

Labels	Species name	Training	Validation	Testing
0	Grub	516	86	258
1	Mole cricket	989	165	495
2	Wireworm	532	88	267
3	White margined moth	88	14	45
4	Black cutworm	512	85	257
5	Large cutworm	294	49	148
6	Yellow cutworm	287	48	144
7	Red spider	317	53	159
8	Corn borer	1018	170	510
9	Army worm	642	107	322
10	Aphids	2456	409	1229
11	Potosiabre vitarsis	339	56	170
12	Peach borer	414	69	208
Total count	8404	1399	4212	

union loss function and a mean square error loss function as the target loss function of the multi-task learning module.

Backbone network module

We use the MAHSNet model that we designed as the backbone network of the MT-MACLBPHSNet model to efficiently extract features from crop pest images. The MAHSNet model is a simple and lightweight convolutional neural network, with its primary parameters shown in Table 2.

Activation function optimization

The activation function plays a crucial role in deep convolutional neural networks, which are mainly used to introduce non-linearity and enhance the expressive ability of the network. The widely used non-linear activation function ReLU has properties such as accelerating convergence and mitigating gradient vanishing (Li and Yuan, 2017). As can be seen from Figure 3, when the neuron activation value is negative, the ReLU activation function completely truncates the information flow to achieve non-linearity, which results in the gradient not being able to be updated and is prone to permanent neuron necrosis, making the network unable to learn during back propagation, while the Mish activation function (Misra, 2019) allows for the existence of a slight flow of the gradient, retaining more information to flow into the neural network for learning. When the neuron activation value is positive, the Mish activation function is smoother than ReLU, with the gradient gradually converging towards 1. It not only inherits the advantages of the ReLU activation function but is also easier to optimize and helps to improve the generalization performance of the model. Therefore, the use of the Mish activation function in deep neural networks is better than the ReLU activation function in terms of accuracy, and we adopt the Mish activation function instead of ReLU in the Shuffle-CA Unit module, as shown in Figure 4.

Depth-wise convolution kernel enlargement

In the typical network design process, depth-wise convolution (DWConv) usually uses a 3×3 convolution kernel to perform convolution operations in the depth direction for each channel of the input. As illustrated in Figure 4, to capture more detailed features of the image, we substitute the original 3×3 convolution kernels with larger 5×5 convolution kernels. Importantly, this does not significantly increase the number of parameters. The use of a larger kernel allows for the coverage of wider input areas, which in turn increases the network's receptive field (Luo *et al.*, 2016) and introduces more non-linear operations, which helps to improve the expressive ability of the network and learn more complex semantic feature information.

Improved pyramid split attention

To enable the model to capture discriminative pest features in

Table 2. Details of MAHSNet network parameters.

Name	Filter/Stride	Output size
Image	-	224×224×3
Conv1	3×3/2	112×112×24
CA-Res1	-	112×112×24
CA-Res2	-	112×112×24
CA-Res3	-	112×112×24
MaxPool	3×3/2	56×56×24
CA-Stage1	-	28×28×48
IPSA1	-	28×28×48
CA-Stage2	-	14×14×96
IPSA2	-	14×14×96
CA-Stage3	-	7×7×192
Conv2	1×1/1	7×7×1024

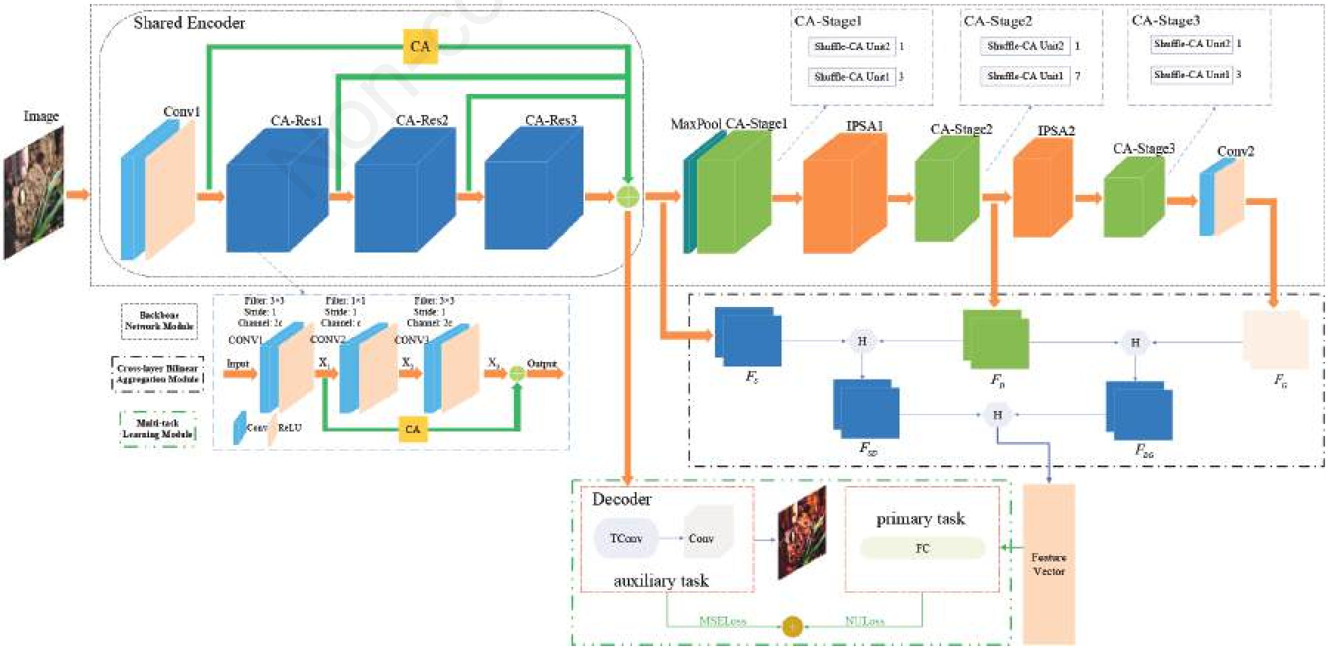


Figure 2. MT-MACLBPHSNet network architecture.

the channel dimension, we introduce a global max-pooling branch based on channel attention (ECANet) (Wang *et al.*, 2020) and design an efficient channel attention module (MPECA) that mixes global max-pooling and global average-pooling strategies.

As shown in Figure 5, the MPECA module consists of three branches. The two branches first go through a global average-pooling operation (retaining global features) and a global max-pooling operation (focusing on locally significant features), respectively, and then adaptively determine the kernel size *k* to perform one-dimensional convolution cross-channel interactions for information fusion, and finally perform the sigmoid function activation to generate weight information Z_1 and Z_2. Then, these weights are summed element-wise to get the new weight aggregation weight information Z_3, which is element-wise multiplied with the input feature map of another branch to generate the ultimate weighted output feature map. This process can suppress ineffective features and highlight effective ones, thereby achieving feature filtering in the channel dimension.

At the same time, to further enhance the network's ability for more fine-grained multi-scale feature extraction, we designed the Improved Pyramid Split Attention (IPSA) module based on hybrid pooling efficient channel attention with the original PSA module (Zhang *et al.*, 2022), as illustrated in Figure 6.

Certainly, the implementation steps of the IPSA module are as follows:

Step 1 - The input feature map is divided into multiple branches using the Improved Split and Concat (ISPC) module. Subsequently, group convolutions with multi-scale convolution kernels are applied to extract multi-scale features focusing on spatial information on each channel feature map. This produces different scale feature maps F_i, which have the same channel dimension $C'=C/S$:

$$\left.\begin{array}{c} F_i = Conv(K_i \times K_i, G_i)(X) \\ K_i = 2\times(i+1)-1, i=0,1,\ldots,S-1 \\ G_i = i+1 \end{array}\right\} \quad (1)$$

Where K_i represents the size of the i-th convolution kernel and G_i represents the size of the i-th group in the group convolution.

Step 2 - The MPECA Weight operation generates channel

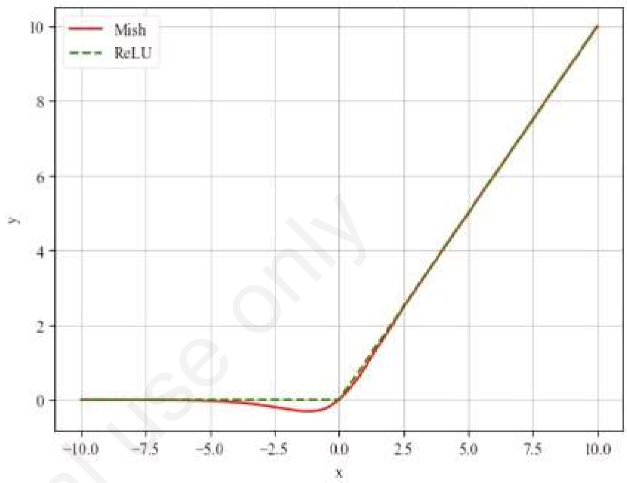


Figure 3. Comparison of Mish and ReLU activation functions.

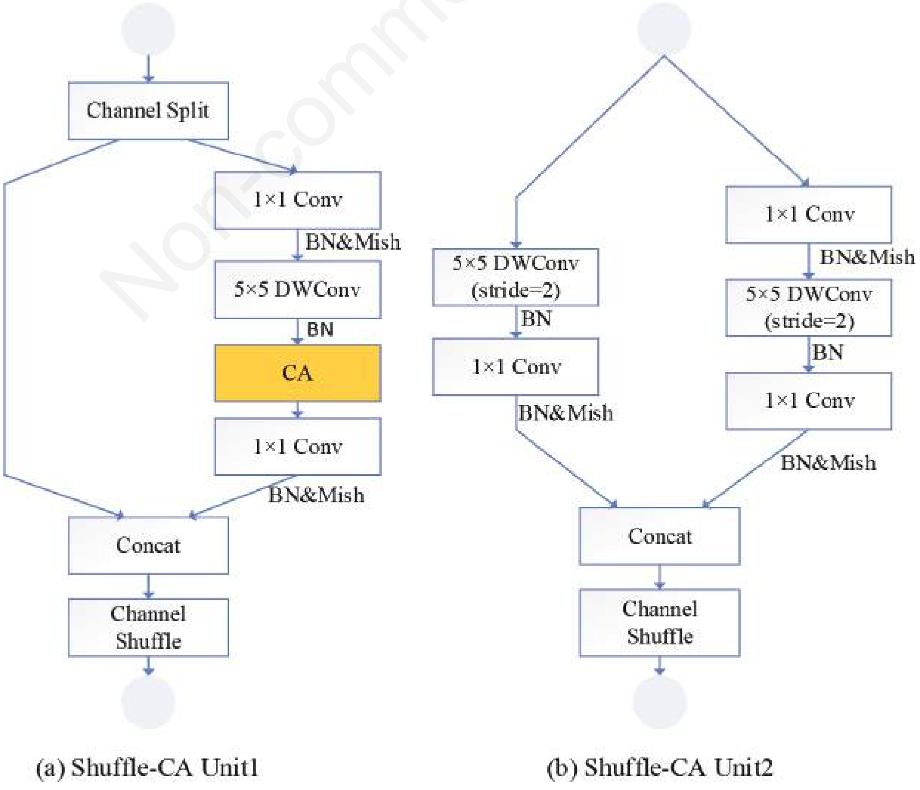


Figure 4. Shuffle-CA unit.

attention weights for different scale feature maps F_i. These weights are then concatenated dimensionally to form the entire multi-scale channel attention weight vector;

Step 3 - The softmax operation is applied to recalibrate the weights of the multi-scale channel attention vector, resulting in recalibrated weights w after multi-scale channel interaction.

Step 4 - Element-wise multiplication is performed on the feature maps of the corresponding scales F_i and recalibrated weights w. This yields the feature maps N_i after being weighted by multi-scale channel attention.

Step 5 - The recalibrated feature maps N_i are concatenated dimensionally and finally yield feature maps N rich in multi-scale information.

Cross-layer bilinear aggregation module

Conventional classification CNN only uses fully connected layers to capture global semantic information from images, which limits the feature representation capability of the network. On the other hand, bilinear CNNs with higher-order feature interactions utilize the outer product operation of the features to acquire second-order information from images, which is more discriminative and robust compared to first-order features, presenting obvious advantages in fine-grained image classification (Lin *et al.*, 2015; Yuan *et al.*, 2022; Kim *et al.*, 2016). However, bilinear CNNs also merely take the features extracted from the last convolutional layer as image representations, which is not enough to describe the various semantic information of the object at a fine-grained level. Furthermore, it ignores inter-layer feature interaction relationships and the interconnectivity of fine-grained feature learning. Additionally, the outer product operation is prone to dimensional explosion. Therefore, we are inspired by the pooling idea of weakly supervised fine-grained classification CNNs (Yu *et al.*, 2018)

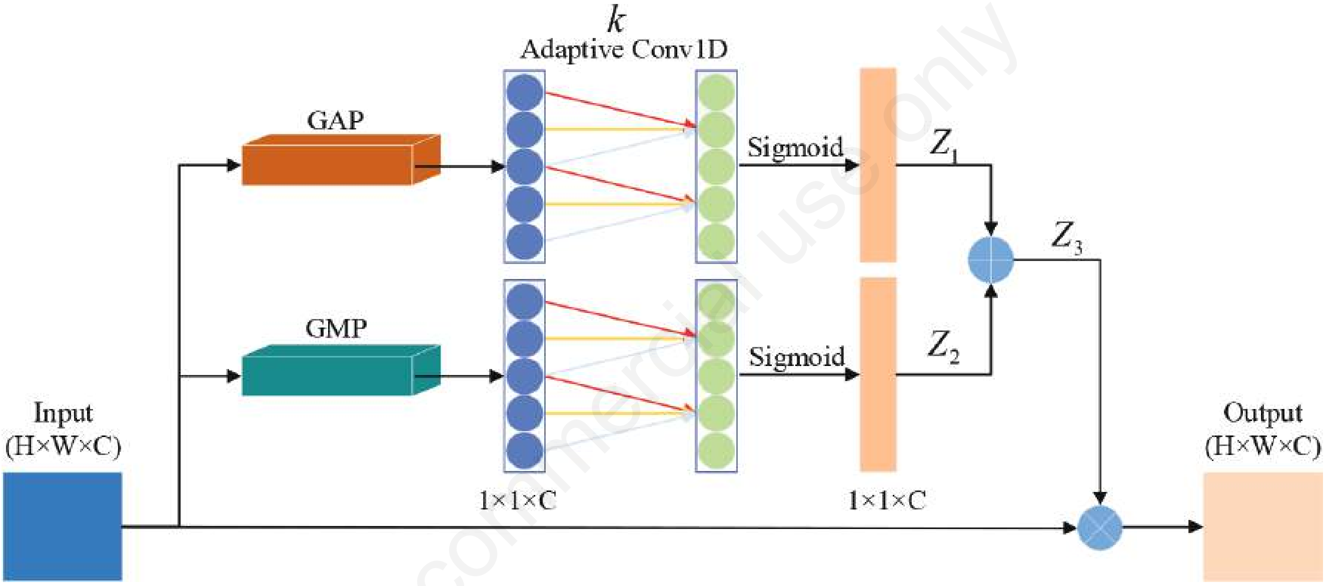


Figure 5. Structure of the MPECA module.

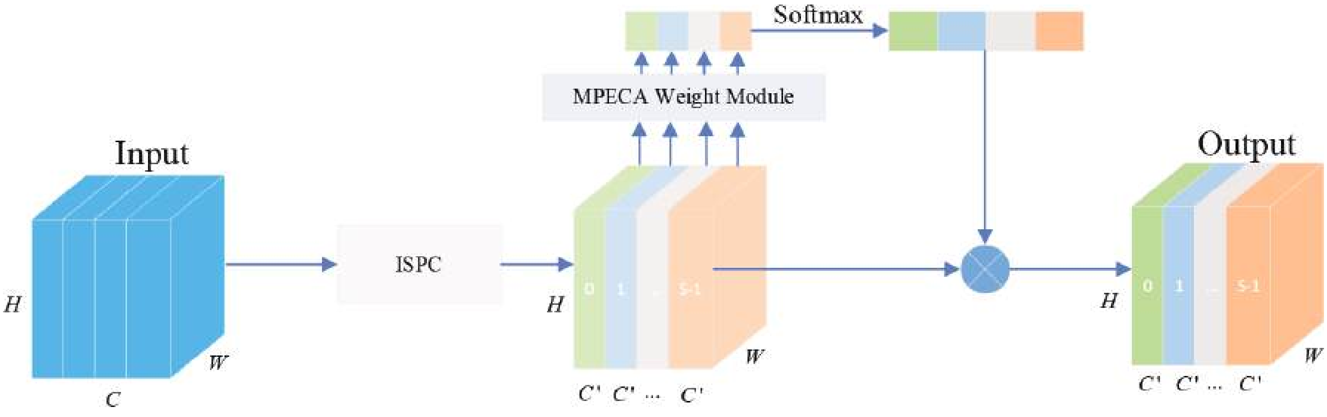


Figure 6 Structure of the IPSA module.

and propose a method of cross-layer bilinear aggregation to effectively fuse shallow-level common features, deep-level semantic features, and global features generated in the backbone network module. Through Hadamard product operations, the features from two different hierarchical structures are element-wise fused in turn, which extracts paired second-order bilinear feature information within the same network across layers. This cross-layer modeling, associated with paired local features, can effectively achieve inter-layer feature interaction and advantageous information complementation. Such mechanisms are of great significance in fine-grained image classification. Importantly, Hadamard product operations entail element-wise multiplication of corresponding channels in two feature maps, significantly reducing computational complexity.

As shown in Figure 7, the steps of the cross-layer bilinear aggregation module are as follows:

Step 1 - Matching shallow-level common feature maps F_S and deep-level semantic feature maps F_D with differing resolutions into the same dimensions. To achieve this, a method involving max-pooling is used to reduce the resolution of feature maps F_S. Subsequently, a 1×1 convolutional layer is employed to expand the features from different hierarchical features into the high-dimensional space. This approach not only retains the crucial information of shallow-level common features but also ensures the spatial information of deep-level semantic features.

$F_S \in R^{H \times W \times C}$ and $F_D \in R^{H \times W \times C}$ represent the matched pairs of hierarchical feature maps are then used Hadamard product operations to effect bilinear fusion, yielding shallow-deep fusion feature maps $F_{SD} \in R^{H \times W \times C}$.

Step 2 - Matching deep-level semantic feature maps F_D and global feature maps F_G with different resolutions into the same dimensions. To achieve this, bilinear interpolation is employed to enhance the resolution of feature maps F_G. Similar to the previous step, a 1×1 convolutional layer is employed to expand features from different hierarchical features into high-dimensional space. $F_D \in R^{H \times W \times C}$ and $F_G \in R^{H \times W \times C}$ represent the matched pairs of hierarchical feature maps are then used Hadamard product operations to achieve bilinear fusion, yielding deep-global fusion feature maps $F_{DG} \in R^{H \times W \times C}$.

Step 3 - Using similar operations as above, shallow-deep fusion feature maps $F_{SD} \in R^{H \times W \times C}$ and deep-global fusion feature maps $F_{DG} \in R^{H \times W \times C}$ are aggregated to generate classification fusion feature maps $F_N \in R^{H \times W \times C}$.

Step 4 - The classification fusion feature maps F_N undergo spatial summation pooling to obtain the integrated information and are finally converted into feature vectors ζ by performing the symbolic square root transformation operation and the L2 normalization operation. This transformed feature vector ζ will be fed into the subsequent multi-task learning module.

The above computational process is represented as Eq. (2).

$$\begin{cases} F_{SD} = F_S \circ F_D \\ F_{DG} = F_D \circ F_G \\ F_N = F_{SD} \circ F_{DG} \\ \xi(I) = \sum_{j=1}^{C} F_N(1:H \times W, j) \\ x = vec(\xi(I)) \\ y = sign(x)\sqrt{|x|} \\ z = y / \|y\|_2 \end{cases} \tag{2}$$

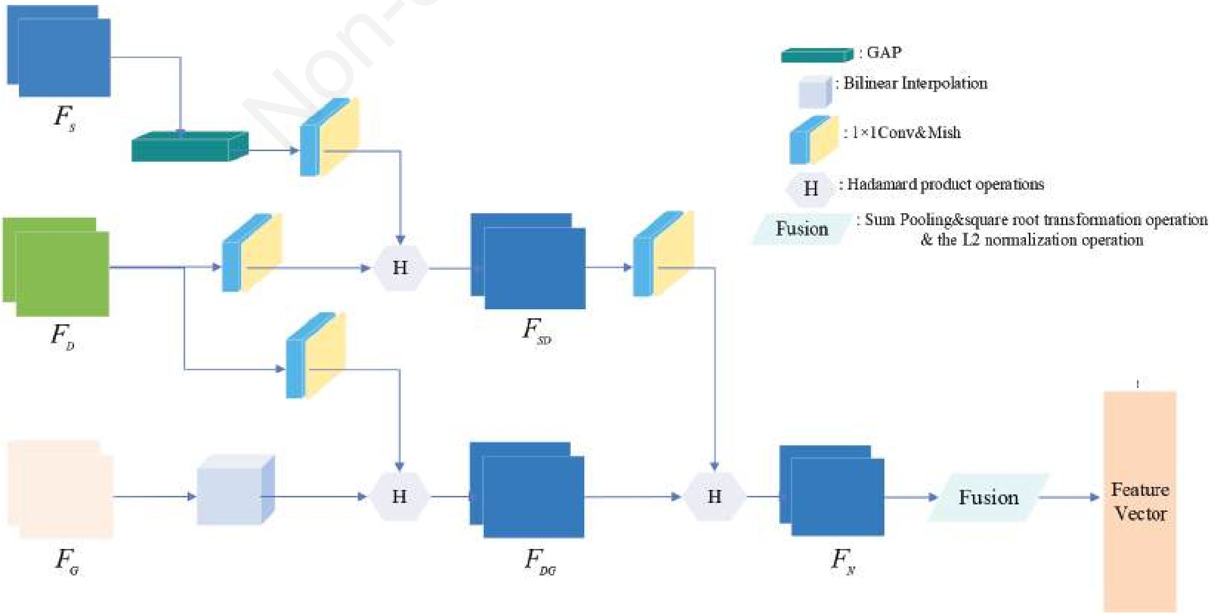

: GAP
: Bilinear Interpolation
: 1×1Conv&Mish
H : Hadamard product operations
Fusion : Sum Pooling&square root transformation operation & the L2 normalization operation
F_S
F_D
F_G
H
F_{SD}
F_{DG}
F_N
Fusion
Feature Vector

Figure 7. Structure of the cross-layer bilinear aggregation module.

Multi-task learning module

Multi-task learning involves simultaneously considering multiple related tasks in one or more models. Its objective is to utilize the inner correlation among tasks to enhance the learning performance of single tasks (Zhang *et al.*, 2020; Zhang and Yang, 2021). As shown in Figure 2, to improve the recognition accuracy using the multi-task learning framework, we adopt a commonly used hard parameter sharing mechanism in multi-task learning. A multi-task learning module is designed with fine-grained classification of pests as the primary task and image feature reconstruction as the auxiliary task. These two different tasks are mutually reinforcing and together provide gradient information to the shared encoder whose features are shared so that it can better enhance the model's generalization ability.

The objective loss function of the multi-task learning module is as follows:

$$L = \tau L_{NU} + \psi L_{MSE} = \tau L_{NU} + \psi \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Where L_{NU} represents the new union loss function for the primary task; L_{MSE} represents the mean squared error loss function for the auxiliary task; n is the total number of samples in the pest image dataset; y_i and $\hat{y}_i$ are the actual and predicted values of the features, respectively; τ and ψ are two hyperparameters that maintain the weights of the losses of the two tasks in the overall loss. In multi-task learning scenarios, the performance of the network structure is affected by the loss of each task. A simple linearly weighted summation of the losses of each task might result in consistent scaling, but the weighting hyperparameters are difficult to determine, and the model's performance is highly sensitive to the choice of weights. Therefore, we employ a strategy based on the uncertainty of variance to automatically learn the relative weights of different tasks, which serves as the basis for weighted loss in multi-task learning (Cipolla *et al.*, 2018; Wang *et al.*, 2022). Table 3 summarizes the training process of the multi-task learning MT-MACLBPHSNet algorithm. During training, the training parameters that need to be set first are the initial learning rate, the iteration index, the total number of iterations, *etc*.

The calculation formula during the gradient backpropagation process is as follows:

$$\frac{\partial L^t}{\partial x_i^t} = \tau \frac{\partial L_{NU}^t}{\partial x_i^t} + \psi \frac{\partial L_{MSE}^t}{\partial x_i^t} \tag{4}$$

The updated formula for the weight W is as follows:

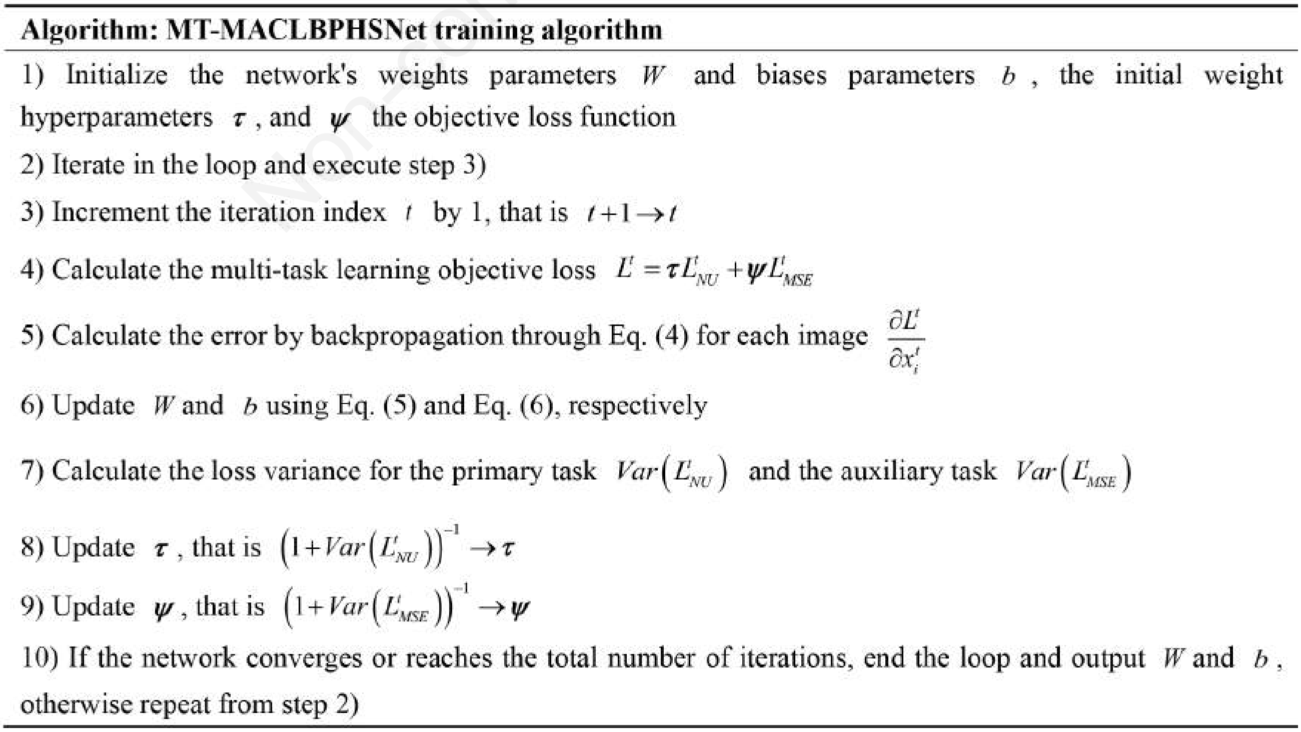
$$W^{t+1} = W^t - \eta^t \frac{\partial L^t}{\partial W^t} \tag{5}$$

The updated formula for the bias b is as follows:

$$b^{t+1} = b^t - \eta^t \frac{\partial L^t}{\partial b^t} \tag{6}$$

Among them, the new union loss function for the primary task is a combined learning strategy based on the softmax equalization loss (Tan *et al.*, 2020) and the bi-tempered logistic loss (Amid *et al.*, 2019). The softmax equalization loss function can effectively mitigate the existence of a long-tailed distribution of the dataset and can focus on fewer classes of data samples to make the network training fairer for each class, which in turn enhances the model's recognition accuracy. While the bi-tempered logistic loss

Table 3. Training process for the MT-MACLBPHSNET algorithm.



```
Algorithm: MT-MACLBPHSNet training algorithm
1) Initialize the network's weights parameters $W$ and biases parameters $b$, the initial weight hyperparameters $\tau$, and $\psi$ the objective loss function
2) Iterate in the loop and execute step 3)
3) Increment the iteration index $t$ by 1, that is $t+1 \rightarrow t$
4) Calculate the multi-task learning objective loss $L^t = \tau L_{NU}^t + \psi L_{MSE}^t$
5) Calculate the error by backpropagation through Eq. (4) for each image $\frac{\partial L^t}{\partial x_i^t}$
6) Update $W$ and $b$ using Eq. (5) and Eq. (6), respectively
7) Calculate the loss variance for the primary task $Var(L_{NU}^t)$ and the auxiliary task $Var(L_{MSE}^t)$
8) Update $\tau$, that is $(1+Var(L_{NU}^t))^{-1} \rightarrow \tau$
9) Update $\psi$, that is $(1+Var(L_{MSE}^t))^{-1} \rightarrow \psi$
10) If the network converges or reaches the total number of iterations, end the loop and output $W$ and $b$, otherwise repeat from step 2)
```

function can address two shortcomings of the standard logistic loss function in training with noisy datasets, such as noisy data and outliers, this enhances the model's adaptability to noise. Thus, the new union loss function combines both these loss functions by configuring the weights, which can take into account their respective characteristics and play the role of combined supervised learning.

The definition of the softmax equalization loss function is given by the Equation: $L_{SEQ} = -\sum_{i=1}^{C} y_i \log(\tilde{p}_i)$

$$\begin{cases} y_k = (1-\boldsymbol{\varepsilon}\times C/(C-1))\times y_k + \boldsymbol{\varepsilon}/(C-1) \\ \tilde{p}_i = \dfrac{e^{z_i}}{\sum_{k=1}^{C} \tilde{w}_k e^{z_k}} \\ \tilde{w}_k = 1-\boldsymbol{\beta} T_\lambda(f_k)(1-y_k) \\ TR(\boldsymbol{\lambda}) = \dfrac{\sum_i^C T_\lambda(f_i) N_i}{\sum_i^C N_i} \end{cases} \tag{7}$$

Where ε is the weight factor of label smoothing regularization (LSR) in the range of 0≤ε≤1, C is the number of classes; β is a random variable used to balance the contribution of positive and negative samples, with a probability parameter γ taking the value of 1 and a probability parameter 1-γ taking the value of 0; f_i is the frequency of the class i in the dataset; $T_\lambda(x)$ is a threshold function that outputs 1 when $x<\lambda$ is met, and 0 otherwise; λ is a parameter used to distinguish between tailed classes and others, and TR is the tail ratio used to set the value of λ.

The definition of the bi-tempered logistic loss function is given by Eq. (8).

$$L_{BT} = \sum_{j=1}^{C}\left(y_j\left(\log_{t_1} y_j - \log_{t_1} \hat{y}_j\right) - \frac{1}{2-t_1}\left(y_j^{2-t_1} - \hat{y}_j^{2-t_1}\right)\right) \tag{8}$$

$$\begin{cases} \log_{t_1}(x) := \dfrac{1}{1-t_1}\left(x^{1-t_1}-1\right) \\ \exp_{t_2}(x) := \left[1+(1-t_2)x\right]_+^{1/(1-t_2)} \\ \hat{y}_j = \exp_{t_2}\left(\hat{a}_j - \boldsymbol{\varphi}_{t_2}(\hat{a})\right) \\ y_j = (1-\boldsymbol{\varepsilon}\times C/(C-1))\times y_j + \boldsymbol{\varepsilon}/(C-1) \\ s.t. \sum_{j=1}^{C} \exp_{t_2}\left(\hat{a}_j - \boldsymbol{\varphi}_{t_2}(\hat{a})\right) = 1 \end{cases} \tag{9}$$

Where t_1 is the temperature parameter, and t_2 is the tail weight parameter. The bi-tempered logistic loss reduces to the standard cross-entropy loss when $t_1=t_2=1$; the bi-tempered logistic loss becomes bounded, preventing large-margin noise samples from pushing the decision boundary too far when $0\leq t_1\leq 1$; the bi-tempered logistic loss exhibits heavy tails when $t_2>1$, helping to keep the decision boundary away from small-margin noise samples; $\hat{a}_j$ is the linear activation of the class j, and $\boldsymbol{\varphi}_{t_2}(\hat{a})$ is the normalized value for each sample.

The definition of the new union loss function is given by Eq. (10).

$$L_{NU} = L_{SEQ} + \boldsymbol{\rho} L_{BT} \tag{10}$$

Where L_{SEQ} represents the softmax equalization loss function with label smoothing regularization (LSR), L_{BT} represents the bi-tempered logistic loss function with LSR, and ρ is a weight hyperparameter that acts as the adjustment coefficient between the two loss functions. After several trials, this study automatically searches for the optimal weight hyperparameter ρ=0.83 through Bayesian optimization.

Results and Analysis

Configuration

All experiments were conducted on a GPU cloud server configured with 80 GB of RAM, an AMD EPYC 7642 48-Core processor, an NVIDIA GeForce RTX 3090 GPU, and 24 GB of video memory. The software environment included Python 3.8 and PyTorch 1.9.1, an open-source deep-learning computing framework.

Considering the performance of the test device, the number of samples per batch was set to 64, and the number of iterations was set to 300. The initial learning rate was set to 0.001, and the learning rate is updated by the warmup and cosine annealing decay strategies. The optimization of the loss function is performed using the AdamW algorithm.

Among them, the two hyperparameters (γ=0.75, λ=0.00043) of the softmax equalization loss function with LSR and the two hyperparameters (t_1=0.8, t_2=1.2) of the bi-tempered logistic loss function with LSR are used for the new union loss function of the primary task.

Ablation analysis of the Backbone MAHSNet model

We conducted ablation experiments to validate the feasibility of the MAHSNet model of the backbone network. Using the HOR-Shuffle-CANet as the baseline model architecture, we introduced the Mish activation function, enlarged the depth-wise convolution kernels, and embedded the IPSA module for training on the IP102-CP13 dataset. With the number of parameters (Params), the amount of floating-point operations (FLOPs), the F1-score, and accuracy as evaluation metrics (Vujović, 2021).

From Table 4 it can be observed that the introduction of the Mish activation function led to the accuracy of the model increas-

Table 4. Comparison of ablation the MAHSNET model.

Model	Params/M	FLOPs/G	F1-score/%	Accuracy/%
Baseline	0.40	0.35	58.74	68.31±0.18
Baseline + Mish	0.40	0.35	61.42	70.63±0.11
Baseline + DW	0.42	0.36	60.81	69.28±0.20
Baseline + Mish + DW	0.42	0.36	61.72	70.97±0.14
MAHSNet	0.49	0.39	62.12	71.44±0.15

ing by 2.32% without increasing the model's parameters or computational complexity. After enlarging the depth-wise convolution kernels to 5×5, the recognition accuracy of the model also obtained an improvement of 0.97%, although the number of model parameters and computation was slightly increased. The inclusion of the IPSA module enhanced the network's feature extraction capability, resulting in an effective accuracy improvement of 71.44%. Ultimately, under the premise of only increasing the number of parameters and computation amount, the MAHSNet model demonstrated favorable recognition performance compared to the baseline, with an F1-score of 62.12 and a recognition accuracy improvement of 3.13%. The above analysis shows that the designed backbone network structure in this study is effective and feasible.

Analysis of different loss functions to optimize the MAHSNet model

To verify the performance of our designed new union loss function for training the MAHSNet model, four different loss functions, including the cross-entropy loss function (CELoss), the softmax equalization loss function (SEQLoss), the bi-tempered logistic loss function (BTLoss), and the new union loss function (NULoss), are selected for comparative analysis, respectively. Figure 8 show the curves of training loss and validation accuracy of the MAHSNet model optimized with different loss functions over iterations. Additionally, the performance comparison of the MAHSNet model is optimally trained using different loss functions on the test dataset, as illustrated in Table 5.

As can be seen from Figure 8a, it's evident that under the same number of iterations, the optimized training model using the NULoss exhibits better advantages than the rest of the loss function. It can better guide the model's training, resulting in relatively low training losses and faster convergence speeds. From Figure 8b, it can be seen that under the same strategy, the model optimized with the NULoss shows relatively higher recognition accuracy on the validation set compared to the rest of the loss functions.

From the comparison of the experimental results in Table 5, it can be observed that the softmax equalization loss function is more advantageous than the cross-entropy loss function, which can alleviate the problem of imbalance in the distribution of the number of each class in the long-tailed dataset. Although the approach using the bi-tempered logistic loss function is less outstanding in terms of the model's recognition accuracy, it focuses more on noisy data, which can facilitate model learning to handle datasets with such characteristics. Ultimately, the NULoss, which is based on a combination of the SEQLoss and BTLoss strategies, optimizes the training of the MAHSNet model with a higher F1-score and recognition accuracy. This loss function incorporates the respective advantages of the above two loss functions, making it effective in mitigating both the long-tailed distribution and adapting to noisy data.

Ablation analysis of the MT-MACLBPHSNet model

To further investigate the effectiveness of the MT-MACLBPHSNet model, the cross-layer bilinear aggregation module and the multi-task learning module are sequentially added for ablation experiments based on the training of the MAHSNet model optimized with the new union loss function.

From Table 6, it can be seen that the MACLBPHSNet model with the added cross-layer bilinear aggregation module integrates features from different layers, which effectively enhances the fine-grained classification accuracy by 4.90% compared to the baseline model. Through the joint learning of the multi-task learning module, it can enable the MT-MACLBPHSNet model to further explore the hidden common feature data between different tasks, thus improving the recognition accuracy of the model more effectively. Eventually, the MT-MACLBPHSNet model outperformed the accuracy by 2.16% over the single-task model, enhancing the recognition accuracy over the baseline model to 75.37% and F1-score up to 67.06% on the IP102-CP13 test dataset. Thus, this demonstrates the feasibility and effectiveness of the MT-MACLBPHSNet model.

Comparison study

To further examine the performance of the MT-MACLBPHSNet model for the fine-grained recognition of crop pests, a comparison with the classical convolutional neural network models such as ResNet-34, GoogLeNet, similar and excellent lightweight CNN models like SqueezeNet, MobileNetV3-Small (Howard *et al.*, 2019), ShuffleNetV2 (Ma *et al.*, 2018), EfficientNet-B0 (Tan and Le, 2019), GhostNet (Han *et al.*, 2020), FasterNet-T1 (Chen *et al.*, 2023) and fine-grained image classification models like BLCNN, BLShuffleNet, CBP (Gao *et al.*, 2016), HBP, MC-Loss (Chang *et al.*, 2020), etc. These models were both comparative tests on the IP102-CP13 dataset in terms of

Table 5. Comparison of performance using different losses.

Model	Loss	F1-score/%	Accuracy/%
MAHSNet	CELoss	62.12	71.44±0.15
	SEQLoss	62.24	72.13±0.09
	BTLoss	62.17	71.68±0.12
	NULoss	62.68	72.18±0.10

Table 6. Comparison of ablation the MT-MACLBPHSNet model.

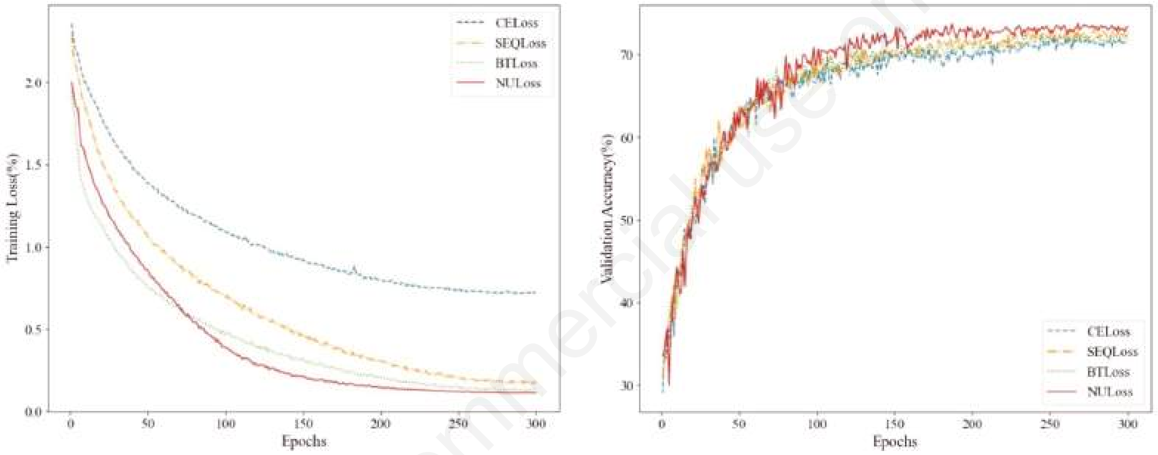
Model	Params/M	FLOPs/G	F1-score (%)	Accuracy (%)
Baseline	0.40	0.35	58.74	68.31±0.18
MAHSNet	0.49	0.39	62.68	72.18±0.10
MACLBPHSNet	0.88	0.46	64.31	73.21±0.13
MT-MACLBPHSNet	0.89	0.59	67.06	75.37±0.11

Params, FLOPs, Weighted File Size (WFS), Precision, Recall, F1-score, and Accuracy. The results are illustrated in Table 7. Meanwhile, Figure 9 was carried out to visualize the curve of validation accuracy of similar and excellent lightweight network models with the number of iterations.

As can be observed from Figure 9, network models with different architectures and the number of layers have different effects on the fine-grained recognition of corn pests. The MT-MACLBPHSNet model that we proposed can extract the fine-grained features of the corn pest images more adequately than the rest of the similar and lightweight network models, and it has a higher validation accuracy. Moreover, as indicated by Table 7, the F1-score and testing accuracy of the MT-MACLBPHSNet model are higher than most of the CNN models mentioned above, and second only to the MC-Loss model. However, the parameter count of the MT-MACLBPHSNet model is 0.89 million and the size of the weight file is 10.56 MB, which is only slightly higher than lightweight networks like SqueezeNet and ShuffleNetV2, and maintains a reasonable level of floating-point operations. In addition, the bilinear fusion of the outer product operation produces a large amount of feature redundancy that affects the model decision, resulting in the poor performance of the customized BLCNN model and the BLShuffleNet model based on the bilinear network structure. Therefore, the MT-MACLBPHSNet model that we proposed effectively balances the number of model parameters, floating-point operations, and test accuracy. It proves to be more suitable for the fine-grained classification task of corn pests and meets the application requirements for deployment on mobile devices.

Visualization analysis

Model visualization can intuitively reflect the areas of concern of the model, and class activation mapping (CAM) can emphasize the regions more important for model inference than other visualization techniques. Therefore, to demonstrate the effectiveness of



(a) Curves of training loss

(b) Curves of validation accuracy

Figure 8. Curves of training loss and validation accuracy.

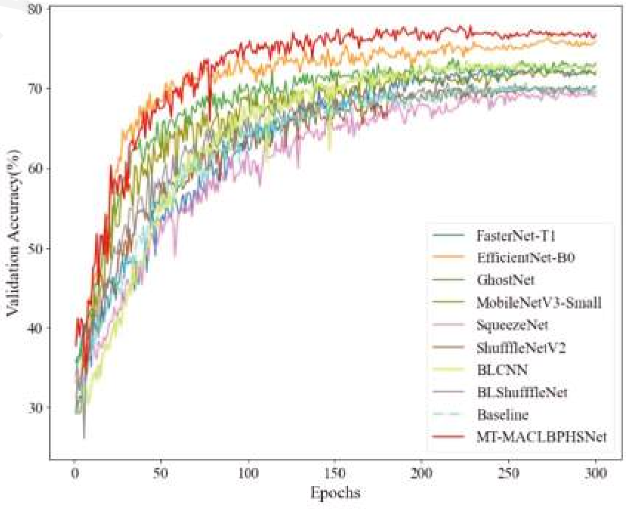
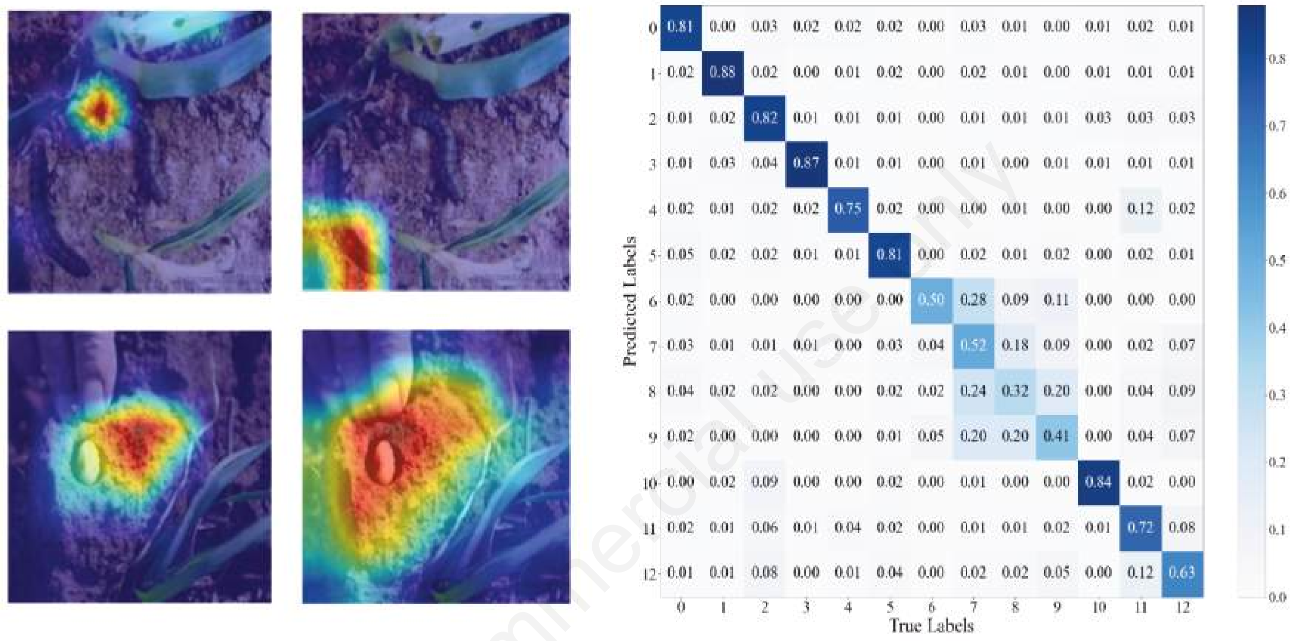


Figure 9. Comparison of variation curves for validation accuracy.

the MT-MACLBPHSNet model, the new Gradient Class Activation Mapping (Grad-CAM) technique is employed to visualize the class activation maps of both the baseline model and the MT-MACLBPHSNet model on the corn pest test dataset, as shown in Figure 10 a,b. The first row of corn pests is grubs, while the second row is small groundhogs. Among them, different colors in the activation features represent varying levels of attention in different regions. In other words, the redder the color, the higher the attention, which is more conducive to the fine-grained recognition of corn pests.

From Figure 10b, it is evident that when using the same original corn pest image, the MT-MACLBPHSNet model, which leverages the cross-level bilinear aggregation module to fuse features from different levels and incorporates multi-task learning with joint image feature reconstruction, tends to focus on more areas of corn pest features in the discrimination than the baseline model. This indicates that there are more bases for the decision-making of the model, resulting in more accurate recognition.

To analyze the effectiveness of the MT-MACLBPHSNet model for fine-grained recognition of corn pests, Figure 10c pres-



	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.81	0.00	0.03	0.02	0.02	0.02	0.00	0.03	0.01	0.00	0.01	0.02	0.01
1	0.02	0.88	0.02	0.00	0.01	0.02	0.00	0.02	0.01	0.00	0.01	0.01	0.01
2	0.01	0.02	0.82	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.03	0.03	0.03
3	0.01	0.03	0.04	0.87	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.01	0.01
4	0.02	0.01	0.02	0.02	0.75	0.02	0.00	0.00	0.01	0.00	0.00	0.12	0.02
5	0.05	0.02	0.02	0.01	0.01	0.81	0.00	0.02	0.01	0.02	0.00	0.02	0.01
6	0.02	0.00	0.00	0.00	0.00	0.00	0.50	0.28	0.09	0.11	0.00	0.00	0.00
7	0.03	0.01	0.01	0.01	0.00	0.03	0.04	0.52	0.18	0.09	0.00	0.02	0.07
8	0.04	0.02	0.02	0.00	0.00	0.02	0.02	0.24	0.32	0.20	0.00	0.04	0.09
9	0.02	0.00	0.00	0.00	0.00	0.01	0.05	0.20	0.20	0.41	0.00	0.04	0.07
10	0.00	0.02	0.09	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.84	0.02	0.00
11	0.02	0.01	0.06	0.01	0.04	0.02	0.00	0.01	0.01	0.02	0.01	0.72	0.08
12	0.01	0.01	0.08	0.00	0.01	0.04	0.00	0.02	0.02	0.05	0.00	0.12	0.63

(a) Baseline (b) MT-MACLBPHSNet (c) Confusion matrix for MT-MACLBPHSNet

Figure 10. Visualization of corn pests.

Table 7. The recognition performance of different models on the IP102-CP13 test dataset.

Model	Backbone	Params/M	FLOPs/G	WFS/MB	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
ResNet-34	-	21.29	3.68	128.00	68.18	65.53	66.69	75.15±0.13
GoogLeNet	-	5.62	1.52	64.50	66.61	63.84	65.10	74.02±0.10
FasterNet-T1	-	6.33	0.86	72.60	63.30	60.03	61.44	70.36±0.15
EfficientNet-B0	-	4.02	0.40	46.40	65.71	64.67	65.08	73.87±0.09
GhostNet	-	3.92	0.15	45.20	62.45	61.42	61.86	70.64±0.13
MobileNetV3-Small	-	1.54	0.06	17.80	63.12	59.82	61.21	69.65±0.12
SqueezeNet	-	0.74	0.73	8.54	61.13	55.48	57.75	67.48±0.20
ShuffleNetV2	-	0.36	0.04	5.29	59.79	57.81	58.56	67.63±0.16
BLCNN	VGG-16	18.12	15.35	207.42	63.65	60.16	61.64	70.42±0.13
BLShuffleNet	ShuffleNetV2	14.32	1.19	160.26	61.03	57.74	59.11	68.18±0.17
CBP	VGG-16	14.82	15.34	169.64	64.01	61.83	62.77	71.94±0.24
HBP	ResNet-18	17.51	2.13	200.56	66.65	64.98	65.76	74.56±0.17
MC-Loss	ResNet-18	14.35	3.62	164.44	68.50	66.02	67.18	75.83±0.18
MT-MACLBPHSNet	MACLBPHSNet	0.89	0.59	10.56	68.52	65.87	67.06	75.37±0.11

ents the confusion matrices of the MT-MACLBPHSNet model for the recognition results on the test dataset in the form of a standardized matrix. Finally, the recognition precision, recall, and F1-score of the 13 classes of corn pests are calculated as the performance evaluation metrics of the model, as illustrated in Table 8.

The experimental results indicate that the MT-MACLBPHSNet model achieves better recognition results than the baseline model on most of the corn pests, but by indexing to view labels 8 (corn borer) and 9 (army worm), there may be a high degree of similarity between the different growth stages of some of the images. The recognition accuracy for these two pest categories is slightly lower compared to other categories, and so there is still room for improvement.

Generalization study

To further evaluate the generalization performance of the MT-MACLBPHSNet model, the IP102-VP16 dataset was used as the experimental material. The model was compared with similar and excellent lightweight CNN models such as SqueezeNet, MobileNetV3-Small, ShuffleNetV2, EfficientNet-B0, GhostNet, and FasterNet-T1, as well as fine-grained image classification models like BLCNN, BLShuffleNet, CBP, HBP, and MC-Loss.

As can be seen from Table 9, the MT-MACLBPHSNet model that we proposed similarly obtains a higher F1-score and accuracy than most of the remaining good network models outside of MC-Loss on the IP102-VP16 dataset, which further verifies that the MT-MACLBPHSNet model has certain generalization ability and scalability.

Conclusions

Based on the practical demands of agricultural production, we have designed a fine-grained recognition model of crop pests (MT-MACLBPHSNet) based on cross-layer bilinear aggregation and multi-task learning, according to the characteristics of crop pest fine-grained image datasets. Additionally, a combined learning strategy based on the softmax equalization loss and the bi-tempered logistic loss has been designed as a new union loss function to optimize the training of the model. After a large number of experimental demonstrations, the MT-MACLBPHSNet model effectively balances the relationship between the number of parameters, the floating-point operations, and the performance of the model under the premise of guaranteeing recognition accuracy and generalization performance. Finally, the model achieves a recognition accuracy of 75.37% and an F1-score of 67.06% on the IP102-CP13 test dataset. Impressively, the model's parameter count is merely 0.89 million, its computational complexity amounts to 0.59 billion floating-point operations, and its weight file size is a mere 10.56MB. These results firmly demonstrate the model's exceptional performance in fine-grained recognition of crop pests. Furthermore, the model's attributes include strong generalization abilities, ease of transferability, and seamless deploy-

Table 8. Recognition precision, recall, and F1-score for the IP102-CP13 test dataset.

Species name	Precision (%)	Recall (%)	F1-score (%)
Grub	81.2	77.8	79.5
Mole cricket	88.9	91.5	90.2
Wireworm	82.5	83.7	83.1
White margined moth	87.2	86.0	86.6
Black cutworm	75.6	76.2	75.9
Large cutworm	80.8	75.0	77.8
Yellow cutworm	50.7	47.9	49.3
Red spider	52.4	52.8	52.6
Corn borer	31.7	28.1	29.8
Army worm	41.0	41.4	41.2
Aphids	83.9	62.4	71.6
Potosiabre vitarsis	72.1	71.9	72.0
Peach borer	62.7	61.7	62.2

Table 9. The recognition performance of different models on the IP102-VP16 test dataset.

Model	Backbone	Params/M	FLOPs/G	WFS/MB	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
FasterNet-T1	-	6.34	0.86	72.65	74.61	63.97	66.92	79.52±0.10
EfficientNet-B0	-	4.02	0.41	46.48	78.24	66.16	70.58	82.88±0.09
GhostNet	-	3.92	0.15	45.20	71.02	62.64	65.87	78.59±0.12
MobileNetV3-Small	-	1.54	0.06	17.80	77.21	60.96	66.79	79.06±0.14
BLCNN	VGG-16	18.15	15.35	209.42	72.63	61.89	65.76	77.91±0.15
CBP	VGG-16	14.83	15.35	169.92	74.65	63.85	67.88	81.04±0.09
HBP	ResNet-18	17.52	2.13	200.66	76.48	66.70	71.21	82.95±0.05
MC-Loss	ResNet-18	14.36	3.62	164.46	79.25	69.70	73.84	84.98±0.10
MT-MACLBPHSNet	MACLBPHSNet	0.89	0.59	10.56	78.96	67.09	71.43	83.87±0.07

ment readiness.

However, the MT-MACLBPHSNet model has the limitation of limited recognition performance when different pest growth stages are highly similar. Although the MT-MACLBPHSNet model employs cross-layer bilinear aggregation and multi-task learning, these methods may not be able to completely solve the problem of the high similarity in different pest growth stages. This is because these methods mainly focus on improving the model's characterization and generalization ability, while there may still be certain deficiencies in dealing with the problem of the high similarity in different pest growth stages.

In future research endeavors, the focus will be on expanding the collection of fine-grained image datasets of crop pests, encompassing a broader range of characteristics such as inter-class similarity and intra-class diversity. Additionally, there will be an exploration and integration of novel network architectures and learning strategies to further optimize lightweight network models. The overarching goal is to further optimize lightweight network models to continue improving the accuracy and robustness of the high similarity problem in different pest growth stages within the field of fine-grained recognition of crop pests, and to promote the development of modern agriculture.

References

Amid, E., Warmuth, M.K.K., Anil, R., Koren, T. 2019. Robust bi-tempered logistic loss based on bregman divergences. Proc. 32nd Adv. Neural Inf. Process. Syst., Vancouver. p. 14987-96.

Chang, D., Ding, Y., Xie, J., Bhunia, A.K., Li, X., Ma, Z., Xu, M., Guo, J., Song, Y.Z. 2020. The devil is in the channels: mutual-channel loss for fine-grained image classification. IEEE Trans. Image Process 29:4683-4695.

Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C., Chan, S. G. 2023. Run, don't walk: chasing higher FLOPS for faster neural networks. Proc. 36th IEEE Conf. Comput. Vis. Pattern Recog., Vancouver. pp. 12021-31.

Cipolla, R., Gal, Y., Kendall, A. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Proc. 31st IEEE Conf. Comput. Vis. Pattern Recog., Salt Lake City. pp. 7482-91.

Gao, Y., Beijbom, O., Zhang, N., Darrell, T. 2016. Compact bilinear pooling. Proc. 29th IEEE Conf. Comput. Vis. Pattern Recog., Las Vegas. pp. 317-26.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C. 2020. GhostNet: More features from cheap operations. Proc. 33rd IEEE Conf. Comput. Vis. Pattern Recog. Seattle. pp. 1577-1586

He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep residual learning for image recognition. Proc. 29th IEEE Conf. Comput. Vis. Pattern Recog, Las Vegas. pp. 770-8.

Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H. 2019. Searching for MobileNetV3. Proc. 17th IEEE Int. Conf. Comput. Vis., Seoul. pp. 1314-24.

Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T. 2016. Hadamard product for low-rank bilinear pooling. arXiv:1610.04325.

Li, C., Zhen, T., Li, Z. 2022. Image classification of pests with residual neural network based on transfer learning. Appl. Sci. 12:4356.

Li, J., Che, G., An, Y. 2020. Image recognition of Pyrausta nubilalis based on optimized convolutional neural network. J. South China Agric. Univ. 41:110-116.

Li, Y., Wang, H., Dang, L. M., Sdeghi-Niaraki, A., Moon, H. 2020. Crop pest recognition in natural scenes using convolutional neural networks. Comput. Electron. Agric. 169:105174.

Li, Y., Yuan, Y. 2017. Convergence analysis of two-layer neural networks with relu activation. Proc. 30th Adv. Neural Inf. Process. Syst., Long Beach. pp. 597-607.

Lin, T.Y., RoyChowdhury, A., Maji, S. 2015. Bilinear CNN models for fine-grained visual recognition. Proc. 15th IEEE Int. Conf. Comput. Vis., Santiago. pp. 1449-57.

Luo, W., Li, Y., Urtasun, R., Zemel, R. 2016. Understanding the effective receptive field in deep convolutional neural networks. Proc. 29th Adv. Neural Inf. Process. Syst., Barcelona. pp. 4898-906.

Ma, N., Zhang, X., Zheng, H.T., Sun, J. 2018. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. Proc. 15th Eur. Conf. Comput. Vis., Munich. pp. 122-38.

Misra, D. 2019. Mish: A self regularized non-monotonic activation function. arXiv:1908.08681.

Nanni, L., Maguolo, G., Pancino, F. 2020. Insect pest image detection and recognition based on bio-inspired methods. Ecol. Inform. 57:4356.

Ruan, J., Liu, S. 2023. Lightweight recognition of crop pests based on high-order residual and attention mechanism. Comp. Syst. Appl. 32:104-115.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proc. 31st IEEE Conf. Comput. Vis. Pattern Recog., Salt Lake City. pp. 4510-20.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. 2015. Going deeper with convolutions. Proc. 28th IEEE Conf. Comput. Vis. Pattern Recog., Boston. pp. 1-9.

Tan, J., Wang, C., Li, B., Li, Q., Quyang, W., Yin, C., Yan J. 2020. Equalization loss for long-tailed object recognition. Proc. 33rd IEEE Conf. Comput. Vis. Pattern Recog., Seattle. pp. 11659-68.

Tan, M., Le, Q. 2019. EfficientNet: rethinking model scaling for convolutional neural networks. arXiv 1905.11946v5.

Thenmozhi, K., Reddy, U.S. 2019. Crop pest classification based on deep convolutional neural network and transfer learning. Comput. Electron. Agric. 164:104906.

Vujović, Ž. 2021. Classification model evaluation metrics. Int. J. Adv. Comput. Sci. Appl. 12:599-606.

Wang, D., Wang, J., Ren, Z., Li, W. 2022. DHBP: A dual-stream hierarchical bilinear pooling model for plant disease multi-task classification. Comput. Electron. Agric. 195:106788.

Wang, J., Li, W., Wang, Y., Tao, R., Du, Q. 2023. Representation-enhanced status replay network for multisource remote-sensing image classification. IEEE Trans. Neural Netw. Learn. Syst. 2023. Online Ahed of Print.

Wang, J., Li, W., Zhang, M., Tao, R., Chanussot, J. 2023. Remote-sensing scene classification via multistage self-guided separation network. IEEE Trans. Geosci. Remote. Sens. 61:1-12.

Wang, M., Wu, Z., Zhou, Z. 2021. [Fine-grained identification research of crop pests and diseases based on improved CBAM via attention].[Article in Chinese]. Trans. Chin. Soc. Agric. Mach. 52:239-247.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proc. 33rd IEEE Conf. Comput. Vis. Pattern Recog., Seattle. pp. 11531-9.

Wei, D., Chen, J., Luo, T., Long, T., Wang, H. 2022. Classification of crop pests based on multi-scale feature fusion. Comput. Electron. Agric. 194:106736.

Wu, X., Zhan, C., Lai, Y.K., Cheng, M.M., Yang, J. 2019. IP102: a large-scale benchmark dataset for insect pest recognition. Proc. 32nd IEEE Conf. Comput. Vis. Pattern Recog., Long Beach. pp. 8779-88.

Yu, C., Zhao, X., Zheng, Q., Zhang, P., You, X. 2018. Hierarchical bilinear pooling for fine-grained visual recognition. Proc. 15th Eur. Conf. Comput. Vis., Munich. pp. 595-610.

Yuan, P., Qian, S., Zhai, Z., Fernán Martínez, J., Xu, H. 2022. Study of chrysanthemum image phenotype on-line classification based on transfer learning and bilinear convolutional neural network. Comput. Electron. Agric. 194:106679.

Zhang, M., Li, W., Zhang, Y., Tao, R., Du, Q. 2023. Hyperspectral and LiDAR Data Classification Based on Structural Optimization Transmission. IEEE Trans. Cybern. 53:3153-3164.

Zheng, X., Zheng P., Wang, W., Cheng, Y., Su, Y. 2023. Rice pest recognition based on multi-scale feature extraction depth residual network. J. South China Agric. Univ. 44:438-446.

Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D. 2022. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. Proc. 16th Asian Conf. Comput. Vis., Macau. pp. 541-57.

Zhang, Y., Liu, J., Zuo, X. 2020. Survey of multi-task learning. Chin. J. Comput. 43:1340-1378.

Zhang, Y., Yang, Q. 2021. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng. 34:5586-5609.