



OPEN ACCESS

EDITED BY

José Bruno Malaquias,
Federal University of Paraíba, Brazil

REVIEWED BY

Robson Nascimento,
Federal University of Paraíba, Brazil
Allef Souza Silva,
Federal University of Paraíba, Brazil

*CORRESPONDENCE

Hainie Zha

✉ zhahn@aqnu.edu.cn

Xueyong Chen

✉ xueyongchen@fafu.edu.cn

RECEIVED 17 February 2025

ACCEPTED 18 April 2025

PUBLISHED 20 May 2025

CITATION

Wang X, Hu C, Wang X, Zha H, Chen X,
Yuan S, Zhang J, Liao J and Ye Z (2025)
Research on multi class pests identification
and detection based on fusion attention
mechanism with Mask-RCNN-CBAM.
Front. Agron. 7:1578412.
doi: 10.3389/fagro.2025.1578412

COPYRIGHT

© 2025 Wang, Hu, Wang, Zha, Chen, Yuan,
Zhang, Liao and Ye. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Research on multi class pests identification and detection based on fusion attention mechanism with Mask-RCNN-CBAM

Xingwang Wang^{1,2,3}, Can Hu³, Xufeng Wang³, Hainie Zha^{1*},
Xueyong Chen^{2*}, Shanshan Yuan⁴, Jing Zhang⁵, Jianfeng Liao⁵
and Zhangying Ye⁶

¹Anhui Province Key Laboratory of Smart Monitoring of Cultivated Land Quality and Soil Fertility Improvement, Anqing Normal University, Anqing, China, ²College of Mechanical and Electrical Engineering, Fujian Agriculture and Forestry University, Fuzhou, China, ³School of Mechanical and Electrical Engineering, Tarim University, Alar, China, ⁴Bayin'guoleng Mongol Autonomous Prefecture Qimo County Agricultural and Rural Development Service Center, Quality and Safety Inspection and Testing Center for Agricultural Products, Korla, China, ⁵Anhui Yi Gang Information Technology Co., Anhui Eagle Information Technology Co., Ltd, Anqing, China, ⁶Institute of Agricultural Bio-Environmental Engineering, College of Biosystems Engineering and Food Science, Zhejiang University, Anqing, China

This study addresses challenges in agricultural pest detection, such as false positives and missed detections in complex environments, by proposing an enhanced Mask-RCNN model integrated with a Convolutional Block Attention Module (CBAM). The framework combines three innovations: (1) a CBAM attention mechanism to amplify pest features while suppressing background noise; (2) a feature-enhanced pyramid network (FPN) for multi-scale feature fusion, enhancing small pest recognition; and (3) a dual-channel downsampling module to minimize detail loss during feature propagation. Evaluated on a dataset of 14,270 pest images from diverse Chinese agricultural regions (augmented to 7,000 samples and split into 6:1:3 training/validation/test sets), the model achieved precision, recall, and F1 scores of 95.91%, 95.21%, and 95.49%, respectively, outperforming ResNet, Faster-RCNN, and Mask-RCNN by up to 2.67% in key metrics. Ablation studies confirmed the CBAM module improved F1 by 5.5%, the FPN increased small-target recall by 6%, and the dual-channel downsampling boosted AP@50 by 3.1%. Despite its compact parameter size (63.87 MB, 1.39 MB lighter than Mask-RCNN), limitations include reduced accuracy in low-contrast scenarios (e.g., foggy fields) and GPU dependency. Future work will focus on lightweight deployment for edge devices and domain adaptation, offering a robust solution for intelligent pest monitoring systems that balance accuracy with computational efficiency.

KEYWORDS

Mask-RCNN-CBAM, attention mechanism, feature enhanced pyramid network, dual channel downsampling, pest extraction, deep learning

1 Introduction

Globally, pests represent one of the most significant challenges to agricultural production, with crop yield losses attributed to pests estimated between 20% and 40% annually. These pests not only threaten food security but also impose substantial economic losses, exacerbate hunger, and pose potential ecological risks (Suganya Kanna et al., 2023). In recent years, China has experienced steady growth in the production of major crops, such as rice, wheat, and maize, as well as an increase in the yield of economic crops like cotton, soybeans, peanuts, fruits, and tea. However, the impact of climate change has amplified the pest and disease pressures, negatively affecting crop yields in certain regions. The frequency of pest and disease outbreaks in China has quadrupled since the 1970s, with rising temperatures further accelerating pest growth and reproduction, particularly during the night (Mendoza et al., 2022).

During the cultivation of crops, different crops suffer from different pest problems. The most common pests include the second generation rice borer, fall armyworm, rice leaf roller, corn borer, thrips, cotton bollworm, locusts, and cutworms, which not only reduce crop yields and cause economic damage but, if not controlled in time, can lead to complete crop failure (Jiang et al., 2019). Therefore, rapid identification and efficient control of major pests are critical for reducing yield losses and ensuring food security. To safeguard food security, China established the National Agricultural Technology Extension Service Center in 1995, which includes pest and disease forecasting and control agencies responsible for pest and disease management. As of 2023, there were 24,800 national agricultural technology extension agencies in China. In the early years, agronomists primarily relied on manual pest counting to predict pest outbreaks. This method not only required significant human labor but was also time consuming and inefficient. With technological advancements, the introduction of intelligent pest monitoring lamps has solved the problem of manual counting by using trap lamps that attract specific pests with light of a particular wavelength. These lamps use industrial cameras to capture images of pests at predetermined intervals, remotely upload these images to the corresponding network ports for processing, and employ object detection algorithms to classify and count the pests in the images. The key difficulty in intelligent pest monitoring systems lies in the pest image detection and recognition algorithms, which commonly use object detection algorithms in computer vision for pest detection (Wei et al., 2022).

Currently, research on pest detection primarily focuses on improving the efficiency of recognition and classification algorithms, especially with deep learning models that have achieved outstanding performance in agricultural pest detection. For example, YOLO series algorithms, SSD (Single Shot MultiBox Detector), and Faster-RCNN (Faster Region based Convolutional Neural Networks) offer a good balance between speed and accuracy in pest detection (Zongwang et al., 2021; Yunong et al., 2023; Li et al., 2024; Zhu et al., 2024). Additionally, the introduction of attention mechanisms, multimodal fusion, and super resolution techniques has made significant progress in improving pest detection performance (Wang et al., 2020; Hu et al.,

2023; Khalid et al., 2023). However, challenges remain in detecting small target pests (minute or densely distributed pests), mainly due to difficulties in extracting small target pests, the impact of background interference, and poor performance in complex environments (Wang et al., 2020; Suganya Kanna et al., 2023). To address these issues, research needs to focus on specific technical improvements for small target pest detection and explore efficient, lightweight models suitable for agricultural production scenarios.

This study focuses on the classification capabilities of object detection algorithms for small target pest categories. Compared with traditional machine learning methods, deep learning based image recognition models can achieve higher accuracy. Among various deep learning network models, Mask-RCNN (mask region based convolutional neural network) (He et al., 2017) has an advantage in pest detection and recognition accuracy (Li et al., 2022). Mask-RCNN is an instance segmentation model based on Faster-RCNN, adding a mask branch for pixel segmentation. It uses anchor boxes for classification and regression while also incorporating pixel level segmentation and classification, resulting in more accurate classifications. In recent years, many researchers have focused on insect detection and recognition with most choosing Mask-RCNN as the base recognition algorithm. For instance, research by Deepika et al. (Deepika and Arthi, 2022; Rong et al., 2022; Kasinathan and Uyyala, 2023) achieved an accuracy rate of over 92% for single-target or single-class detection in each image, while Mendoza and Denver (2022) achieved high accuracy for multi-class detection in one image. In Liu et al.'s (2023) study, detection of four sparsely distributed classes achieved up to 92% accuracy, but non target similar pests and high density cases were not considered, leaving room for improvement. Thus, Mask-RCNN is effective for multi target classification in a single image, but further improvements are needed for multi class small target detection.

This paper addresses the challenges of high intra class similarity, varying target scales, and complex backgrounds in small target pest images captured by pest monitoring lamps. These challenges result in low detection precision, false positives, and missed detections. Inspired by the attention mechanism, multi scale feature transmission, and Mask-RCNN algorithms, this study proposes a fusion attention based Mask-RCNN-CBAM (Convolutional Block Attention Module) object recognition network. This network tackles the issues of complex backgrounds and insufficient multi-scale feature extraction, combining attention mechanisms and multi-level semantic features to achieve precise extraction of small target pests. The proposed method offers high efficiency and accuracy, providing an optimized and innovative solution for pest identification in pest monitoring lamps.

2 Materials and methods

2.1 Image data acquisition

This study uses a self-developed intelligent pest monitoring lamp (YG-L1200) for pest sampling. This lamp uses a light source

with a wavelength range of 320–680 nm (Nanometer) during night operation. Due to the weak penetration of ultraviolet light (320–400 nm) and its inability to cover large agricultural fields, visible light with stronger penetration is used to attract pests over greater distances, drawing them towards the detection lamp. The pests attracted to the light will collide with the glass screen in the device and fall into an infrared treatment chamber. After entering the chamber, pests are effectively killed by infrared heat treatment without damaging their bodies. The pests are killed within 3–5 minutes of falling, and regularly collected pests are transferred to a drying chamber for further heat treatment before being stored in a collection box. To improve pest-killing efficiency, two infrared heating boxes are used, and two movable doors are alternately opened to ensure that each pest undergoes at least one infrared heating cycle. The processed pests will fall onto a shooting platform at preset time intervals. The platform is equipped with a vibration device and a small conveyor belt. Once the pests fall onto the platform, they are evenly distributed on the conveyor belt and moved to the shooting area. The system camera captures images of the pests, and these photos are transmitted to the image acquisition system. The captured pest carcasses are then moved to the collection device.

To ensure the diversity of experimental samples, pest trapping and sampling points were set up in multiple regions across China, as shown in Figure 1. Each pest monitoring light deployed at the sampling point can cover an area of 2 hectares. Since different regions have varying crops, the types of pest infestations also differ. To facilitate comprehensive pest monitoring, the collected pest data from various crops are categorized and recorded, as shown in Table 1.

2.2 Image data classification and processing

The data set used in this experiment is collected by the pest monitoring lights. The pest monitoring lights are equipped with a

20 megapixel high definition industrial camera, which is mounted above the workbench in a docking mode to capture high resolution images of pests attracted by the light source. The images captured by the camera are then promptly uploaded to local storage. A sample image captured by the camera is shown in Figure 2.

Due to variations in the image quality captured by the pest monitoring lights, and the fact that certain pest species may appear in large numbers during the same season, this could significantly affect the quality of the data set. To address this issue, the training data set is first filtered to remove low quality image samples. In deep learning algorithms, high resolution images are required for better recognition accuracy. Therefore, images with resolutions lower than 4096×2160 are excluded from the data set. Similarly, images containing incomplete leaves or pests are also discarded. Next, images from different seasons are extracted in specific proportions to ensure an adequate number of images for each pest species, thus meeting the deep learning training requirements. Some sample data information is shown in Table 2.

2.3 Experimental data set construction

Due to the dense collection of images of pests, this paper constructs a multi-type pest data set. Before the experiment, all images are uniformly processed by cropping each image into 9 equal parts, with each cropped image set to a resolution of 1824 pixels × 1216 pixels. To ensure diversity in the experimental samples, data augmentation techniques are employed to increase the data set size (Yuqi et al., 2023), including horizontal flipping, vertical flipping, 90° rotation, 180° rotation, 270° rotation, and noise addition. Through these operations, the data set size is increased seven fold, resulting in a total of 7000 images. Labeling is done using the LabelMe tool, and each image's label is saved as a .json file in the directory where the image is stored. For data set training using Mask-RCNN, the data set format follows the COCO format, which includes image information, annotation details, and category definitions, as shown in Figure 3. Finally, the data set is split into

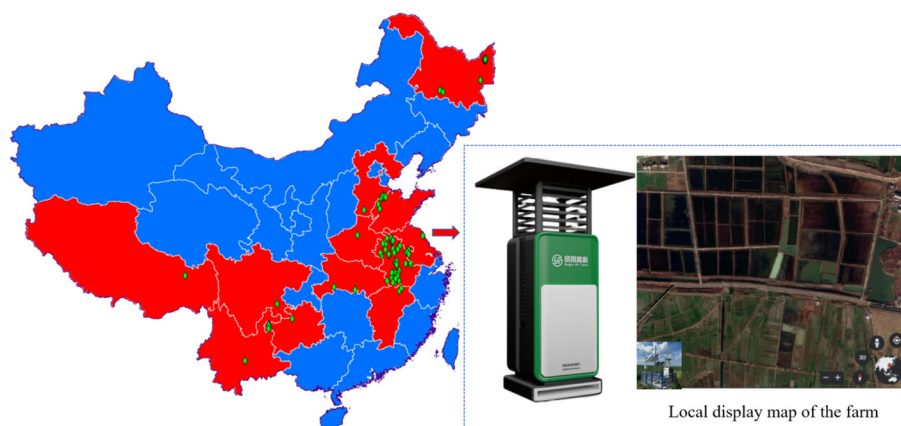


FIGURE 1
Pest monitoring light deployment map.

TABLE 1 Deployment and distribution of pest monitoring lights and collected pest data.

Region	Quantity (units)	Longitude and latitude range	Main crops	Major pests (insect scientific name)
Heilongjiang Province	14	1、 133°06'E,47°20'N 2、 132°56'E,47°22'N 3、 133°02'E,47°20'N 4、 133°00'E,47°17'N 5、 133°12'E,47°23'N 6、 133°13'E,47°23'N 7、 133°05'E,47°14'N 8、 133°06'E,47°18'N 9、 133°07'E,47°16'N 10、 133°07'E,47°20'N 11、 131°34'E, 45°44'N 12、 131°35'E, 45°44'N 13、 126°32'E, 45°45'N 14、 126°13'E, 45°58'N	Cold-resistant rice, spring wheat	Nilaparvata lugens, Sogatella furcifera, Empoasca vitis, Cnaphalocrocis medinalis, Scirpophaga excerptalis, Scirpophaga excerptalis, Chilo suppressalis, Spodoptera litura, Hylastinus obscurus, Melolontha melolontha, etc.
Shandong Province	11	1、 116°44'E, 37°38'N 2、 116°51'E, 37°48'N 3、 116°36'E, 37°36'N 4、 116°51'E, 37°39'N 5、 116°48'E, 37°44'N 6、 116°12'E, 37°05'N 7、 117°00'E, 37°41'N 8、 117°00'E, 37°42'N 9、 117°08'E, 37°37'N 10、 116°18'E, 36°40'N 11、 116°17'E, 36°40'N	Winter wheat, corn, grapes, apples	Spodoptera litura, Hylastinus obscurus, Melolontha melolontha, Gryllotalpa orientalis, Ostrinia nubilalis, Caelifera, Bemisia tabaci, Parthenolecanium corni, Lobesia botrana, Thrips tabaci, Cydia pomonella, Grapholita molesta, Sternorrhynchus spp, etc.
Henan Province	1	113°31'E, 34°36'N	Winter wheat, rice, corn	Nilaparvata lugens, Sogatella furcifera, Empoasca vitis, Cnaphalocrocis medinalis, Scirpophaga excerptalis, Scirpophaga excerptalis, Chilo suppressalis, Spodoptera litura, Hylastinus obscurus, Melolontha melolontha, Thrips tabaci, Agrotis ipsilon, Ostrinia nubilalis, Caelifera, Bemisia tabaci, etc.
Anhui Province	66	1、 117°05'E, 32°56'N 2、 117°00'E, 32°56'N 3、 117°04'E, 32°59'N 4、 116°47'E, 30°08'N 5、 116°29'E, 30°34'N 6、 116°47'E, 30°08'N 7、 117°01'E, 30°39'N 8、 117°07'E, 33°53'N 9、 116°53'E, 30°56'N 10、 117°11'E, 30°12'N 11、 116°05'E, 33°34'N 12、 116°22'E, 33°29'N 13、 116°06'E, 33°34'N 14、 115°46'E, 32°33'N 15、 116°43'E, 30°19'N 16、 117°46'E, 33°12'N 17、 117°01'E, 33°36'N 18、 117°02'E, 33°39'N 19、 118°35'E, 31°21'N 20、 115°59'E, 30°18'N 21、 116°34'E, 30°57'N 22、 116°32'E, 30°05'N 23、 116°34'E, 30°57'N 24、 118°37'E, 31°31'N 25、 117°03'E, 33°56'N 26、 117°03'E, 30°31'N 27、 115°59'E, 30°18'N 28、 116°15'E, 33°01'N 29、 116°52'E, 30°28'N 30、 117°07'E, 33°42'N 31、 117°33'E, 32°53'N 32、 116°58'E, 30°50'N	Rice, winter wheat, corn, tea	Nilaparvata lugens, Sogatella furcifera, Empoasca vitis, Cnaphalocrocis medinalis, Scirpophaga excerptalis, Scirpophaga excerptalis, Chilo suppressalis, Spodoptera litura, Hylastinus obscurus, Melolontha melolontha, Thrips tabaci, Agrotis ipsilon, Ostrinia nubilalis, Caelifera, Bemisia tabaci, Euproctis chrysorrhoea, Ectropis obliqua, Setora nitens, Adoxophyes honmai, Cydia pomonella, etc.

(Continued)

TABLE 1 Continued

Region	Quantity (units)	Longitude and latitude range	Main crops	Major pests (insect scientific name)
		33、118°19'E, 32°29'N 34、114°34'E, 36°45'N 35、117°48'E, 33°10'N 36、117°07'E, 32°38'N 37、116°37'E, 34°00'N 38、118°39'E, 32°42'N 39、118°36'E, 32°39'N 40、117°49'E, 33°09'N 41、118°38'E, 32°40'N 42、117°05'E, 30°33'N 43、117°07'E, 33°42'N 44、117°06'E, 30°33'N 45、118°28'E, 32°29'N 46、117°02'E, 31°46'N 47、117°21'E, 31°03'N 48、117°06'E, 31°42'N 49、117°07'E, 31°43'N 50、117°06'E, 31°40'N 51、116°15'E, 33°01'N 52、118°36'E, 32°18'N 53、117°11'E, 33°44'N 54、117°12'E, 33°43'N 55、117°33'E, 32°53'N 56、115°42'E, 32°49'N 57、117°15'E, 31°52'N 58、117°10'E, 31°34'N 59、117°10'E, 31°35'N 60、117°10'E, 31°34'N 61、117°12'E, 30°48'N 62、117°06'E, 30°33'N 63、116°52'E, 30°23'N 64、118°07'E, 31°33'N 65、116°33'E, 32°29'N 66、116°32'E, 32°35'N		
Yunnan Province	1	101°59'E, 24°03'N	Rice, oranges, tangerines, tea	<i>Nilaparvata lugens</i> , <i>Sogatella furcifera</i> , <i>Laodelphax striatellus</i> , <i>Cnaphalocrocis medinalis</i> , <i>Scirpophaga excerptalis</i> , <i>Chilo partellus</i> , <i>Sitophilus oryzae</i> , <i>Spodoptera litura</i> , <i>Aphis citricola</i> , <i>Liriomyza</i> spp., <i>Spodoptera exigua</i> , <i>Helicoverpa armigera</i> , <i>Thrips tabaci</i> , <i>Ceratitis capitata</i> , Cutworm, <i>Tephritis</i> spp., <i>Plutella xylostella</i> , <i>Clostera anachoreta</i> , etc.
Tibet Autonomous Region	1	95°34'E, 30°55'N	Barley, rapeseed	<i>Locusta migratoria tibetensis</i> , <i>Mamestra brassicae</i> , <i>Pieris rapae</i> , <i>Plutella xylostella</i> , <i>Agrotis ipsilon</i> , <i>Toxocera</i> spp., <i>Raphanus sativus</i> , etc.
Hebei Province	1	114°34'E, 36°45'N	Winter wheat, corn, peaches	<i>Spodoptera litura</i> , <i>Hyles euphorbiae</i> , <i>Melanotus</i> spp., <i>Gryllotalpa orientalis</i> , <i>Ostrinia nubilalis</i> , <i>Locusta migratoria</i> , <i>Laodelphax striatellus</i> , <i>Planococcus ficus</i> , <i>Erythroneura vitis</i> , <i>Thrips tabaci</i> , <i>Cydia pomonella</i> , <i>Carposina sasakii</i> , <i>Aphis pomi</i> , etc.
Sichuan Province	1	104°55'E, 28°58'N	Rice, tangerines, tea	<i>Nilaparvata lugens</i> , <i>Sogatella furcifera</i> , <i>Laodelphax striatellus</i> , <i>Cnaphalocrocis medinalis</i> , <i>Scirpophaga excerptalis</i> , <i>Chilo partellus</i> , <i>Sitophilus oryzae</i> , <i>Spodoptera litura</i> , <i>Aphis citricola</i> , <i>Liriomyza</i> spp., <i>Spodoptera exigua</i> , <i>Helicoverpa armigera</i> , <i>Thrips tabaci</i> , <i>Ceratitis capitata</i> , Cutworm, <i>Tephritis</i> spp., <i>Plutella xylostella</i> , <i>Clostera anachoreta</i> , etc.
Guizhou Province	7	1、106°22'E, 27°40'N 2、104°06'E, 27°12'N 3、104°08'E, 26°54'N 4、104°14'E, 26°54'N 5、104°06'E, 27°08'N	Rice, rape, tea	<i>Sogatella furcifera</i> , <i>Laodelphax striatellus</i> , <i>Cnaphalocrocis medinalis</i> , <i>Scirpophaga excerptalis</i> , <i>Chilo partellus</i> , <i>Sitophilus oryzae</i> , <i>Spodoptera litura</i> , <i>Aphis citricola</i> , <i>Liriomyza</i> spp., <i>Spodoptera exigua</i> , <i>Helicoverpa armigera</i> , <i>Thrips tabaci</i> , <i>Ceratitis capitata</i> , etc.

(Continued)

TABLE 1 Continued

Region	Quantity (units)	Longitude and latitude range	Main crops	Major pests (insect scientific name)
		6、 104°06'E, 26°50'N 7、 104°06'E, 27°35'N		
Hubei Province	2	1、 110°38'E, 30°16'N 2、 112°42'E, 29°52'N	Rice, tea, oranges, tobacco	Nilaparvata lugens, Sogatella furcifera, Laodelphax striatellus, Cnaphalocrocis medinalis, Scirpophaga excerptalis, Chilo partellus, Sitophilus oryzae, Spodoptera litura, Aphis citricola, Liriomyza spp., Spodoptera exigua, Helicoverpa armigera, Thrips tabaci, Ceratitis capitata, Cutworm, Tephritis spp., Plutella xylostella, Clostera anachoreta, etc.
Jiangxi Province	2	1、 117°03'E, 29°18'N 2、 117°18'E, 29°30'N	Rice, tea, oranges, tobacco	Nilaparvata lugens, Sogatella furcifera, Laodelphax striatellus, Cnaphalocrocis medinalis, Scirpophaga excerptalis, Chilo partellus, Sitophilus oryzae, Spodoptera litura, Aphis citricola, Liriomyza spp., Spodoptera exigua, Helicoverpa armigera, Thrips tabaci, Ceratitis capitata, Cutworm, Tephritis spp., Plutella xylostella, Clostera anachoreta, etc.
Jiangsu Province	2	1、 120°24'E, 33°42'N 2、 120°23'E, 33°40'N	Rice, wheat, corn, rape	Nilaparvata lugens, Sogatella furcifera, Laodelphax striatellus, Cnaphalocrocis medinalis, Scirpophaga excerptalis, Chilo partellus, Sitophilus oryzae, Spodoptera litura, Aphis citricola, Liriomyza spp., Spodoptera exigua, Helicoverpa armigera, Thrips tabaci, Ceratitis capitata, Cutworm, Tephritis spp., Plutella xylostella, Clostera anachoreta, etc.

training, validation, and test sets in a 6:1:3 ratio. Since this study only labels pests that cause significant crop damage, pests causing minor damage, beneficial insects, and non-crop pests are not labeled. The labeling style is shown in Figure 4. Figure 4a shows a sample of the original image, while Figure 4b shows the corresponding labeled image.

2.4 Mask-RCNN network

The network structure used in this paper was based on the classic Mask-RCNN architecture. Mask-RCNN has a simple structure and can be used for various tasks, such as object detection, semantic segmentation, instance segmentation, and

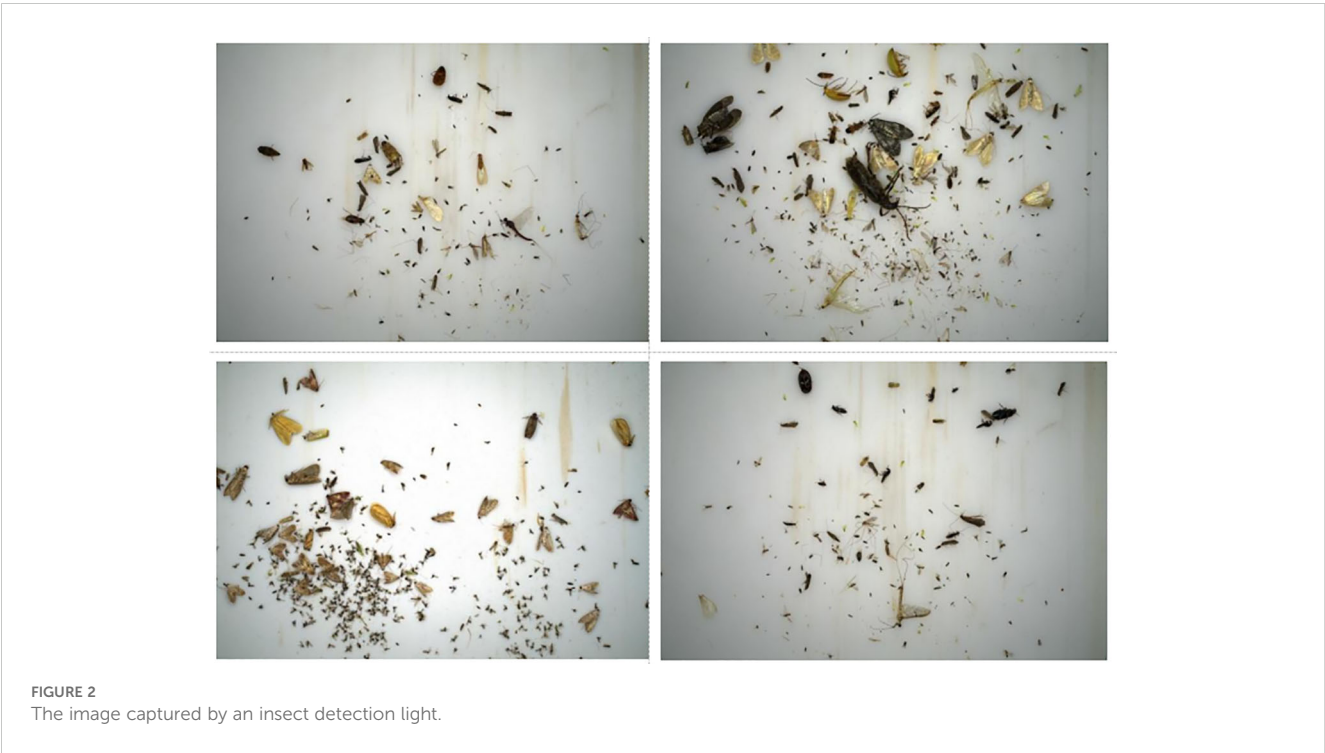





















TABLE 2 The sample size and image information for various pest species.

Number	Pest	Acronym	Image	Number of labeled samples (bounding box)	Percentage of bounding box samples
1	<i>Cicadella viridis</i> (Linnaeus)	CV		6171	4.33%
2	<i>Thaia rubiginosa</i> (Kuoh)	TR		9008	6.31%
3	<i>Nephotettix cincticeps</i>	NC		8982	6.29%
4	<i>Nilaparvata lugens</i>	NL		8765	6.14%
5	<i>Sogatella furcifera</i>	SF		6548	4.59%
6	<i>Laodelphax striatellus</i>	LS		7254	5.08%
7	<i>Echinocnemus squamous</i> (Billberg)	ES		6875	4.82%
8	<i>Verania discolor</i> (Fabricius)	VD		4521	3.17%
9	Thysanoptera	TP		9468	6.63%
10	<i>Helicoverpa armigera</i>	HA		7251	5.08%
11	<i>Naranga aenescens</i> (Moore) (♂)	NA_M		6790	4.75%
12	<i>Naranga aenescens</i> (Moore) (♀)	NA_F		7741	5.43%
13	<i>Inazuma dorsalis</i> (Motschulsky)	ID		8894	6.23%
14	<i>Nezara viridula</i> (Linnaeus)	NV		9274	6.50%

(Continued)

TABLE 2 Continued

Number	Pest	Acronym	Image	Number of labeled samples (bounding box)	Percentage of bounding box samples
15	<i>Anomala corpulenta</i> (Motsehulsiy)	AC		6744	4.72%
16	<i>Cnaphalocrocis medinalis</i> (Guenee) (♀)	CM_F		7133	5.00%
17	<i>Cnaphalocrocis medinalis</i> (Guenee) (♂)	CM_M		6674	4.68%
18	<i>Chilo suppressalis</i>	CS		7213	5.05%
19	<i>Sesamia inferens</i>	SI		7396	5.18%
	Total			142,702	100%

♂ represents male individuals, ♀ represents female individuals.

human pose recognition. Its network structure was shown in Figure 5.

The Mask-RCNN model extends the Faster-RCNN framework by adding a fully connected segmentation network for semantic segmentation, offering both detection and extraction functions. Unlike Faster-RCNN, which uses VGG as the backbone feature extraction network, Mask-RCNN uses ResNet50 and ResNet101 as the backbone feature extraction networks. It also incorporates a feature pyramid network (FPN) into the backbone structure, where different backbone combinations lead to feature layers of different sizes. The Region Proposal Network (RPN) then generates anchor boxes at each point in the effective feature layer to perform rough screening, producing local feature layers. Afterward, Region of Interest (ROI) Align is applied to these local feature layers, which are passed into classification and regression models as well as the Mask model for classification and mask generation, ultimately outputting classification and segmentation results.

2.5 Mask-RCNN network improvements

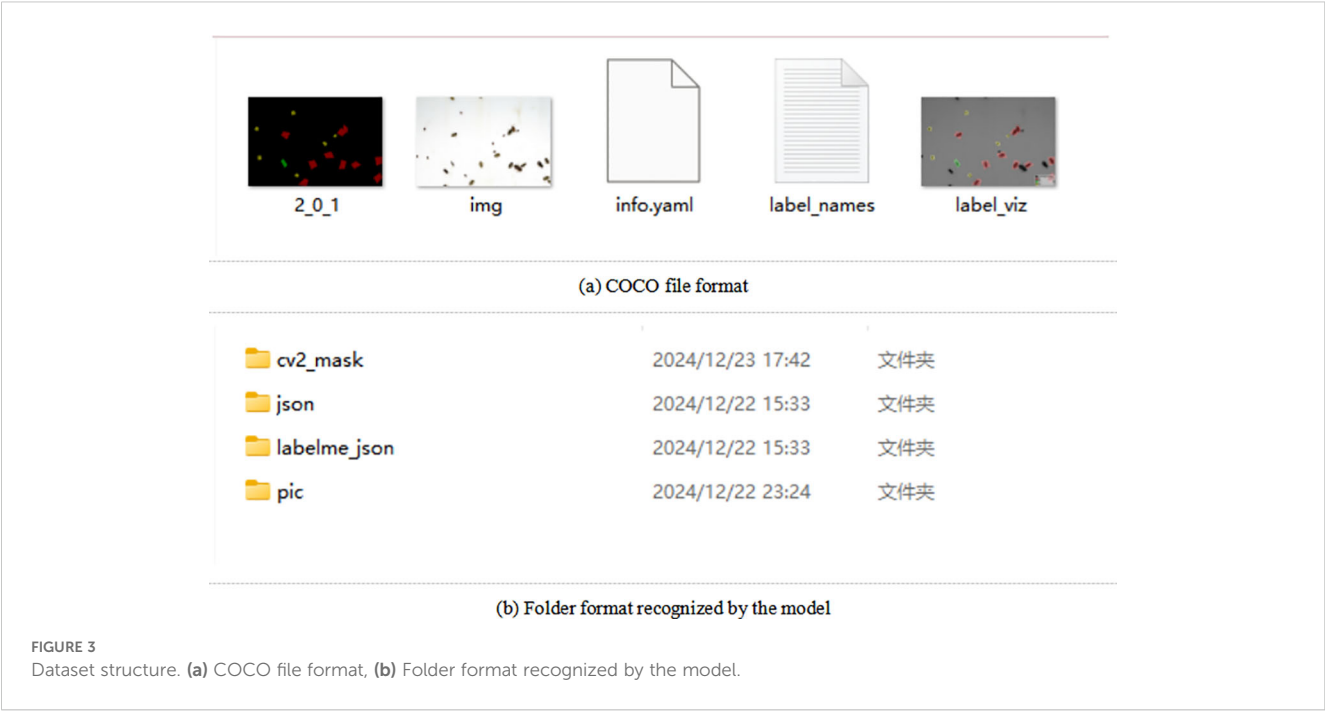
The proposed Mask-RCNN-CBAM deep learning network is based on the original Mask-RCNN structure and introduces the CBAM (Convolutional Block Attention Module) attention mechanism at the feature processing stage. The CBAM module enhances channel and spatial attention to weigh the feature information of different regions in the image, allowing the model to learn and amplify the target features. Compared to the original network, the enhanced feature pyramid module adds bottom up

feature transfer branches to transmit shallow features to deeper layers and fuse them. The features passed from the upper layers are fused with the shallow spatial information and deep semantic information. The feature transmission process utilizes dual channel down sampling convolution operations to reduce detail feature loss during down sampling, retaining image detail features and improving the network's feature extraction and detail optimization capabilities.

As shown in Figure 6, in the input stage, the image passes through the ResNet101 backbone feature extraction layers, and the resulting features are sent to the CBAM attention mechanism module. The attention mechanism helps the model focus on processing important regions. The channel attention module and spatial attention module are connected in sequence, which improves the network's ability to learn complex object features and avoids false positives and false negatives in complex backgrounds. The network also introduces a multi scale feature fusion pyramid module that performs feature transmission at different scales, mitigating the impact of multi scale feature deficiencies on target detection (Yu et al., 2022). To further address the information loss issue caused by traditional down sampling methods, a dual channel down sampling module is introduced. This module uses two channels to retain information and extract features, thereby improving the model accuracy.

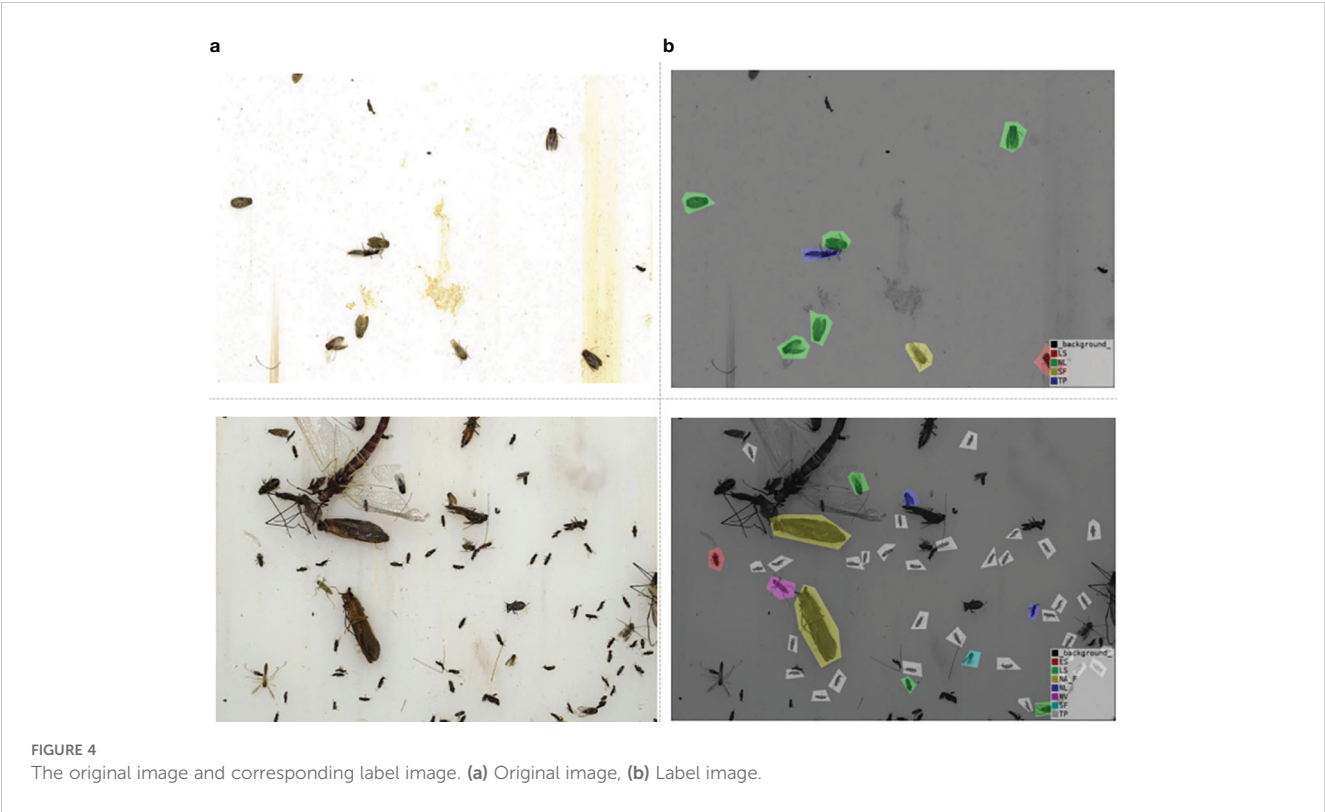
2.6 CBAM attention mechanism module

CBAM (evolutionary block attention module) (Woo et al., 2018; Wang et al., 2021) is an optimization algorithm, which combines



the channel attention module and the spatial attention module. The image is extracted from the backbone feature by backbone and sent to the CBAM module. The feature will first aggregate the features generated by the channel attention with the input features through the channel attention module to generate the final channel attention feature. The generated features are input into the spatial feature map, and the final attention feature is obtained through pooling and

convolution connection. This mechanism is very effective in processing multi-scale feature extraction tasks, which can highlight the effective features of the target and reduce redundant information. In this paper, we use the relationship between channel attention of features to generate a channel attention graph. Each channel of the feature represents a special feature detector, and the attention of different channel feature channels will be given



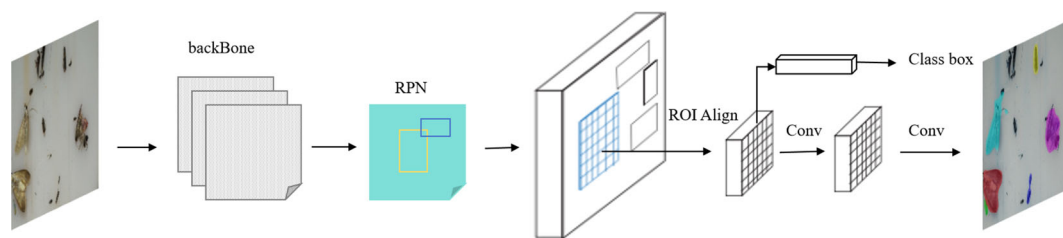


FIGURE 5
Mask-RCNN network framework.

corresponding weight coefficients. The channel attention will focus on meaningful features, compressing the spatial dimension of the input feature mapping to calculate the channel attention more efficiently, which can generate region proposal features for regions with rich features. As shown in Figure 7a, the channel attention module utilizes the maximum pooled output and average pooled output of the shared network. Process: (1) Aggregating the spatial information of the channel feature map through the average pooling and maximum pooling operations to generate two different spatial context descriptors: F_{avg}^c and F_{max}^c , representing the average pooling feature and the maximum pooling feature respectively; (2) Transfer the generated features to a shared network to generate a $1 \times 1 \times C$ channel attention map $Mc(F)$. The shared network consists of multi-layer perceptron MLP and a hidden layer. To reduce the cost of parameters, the activation size of the hidden layer is set to C/r , where C is the number of neurons, r is the attenuation rate, and the activation function was ReLu. After the shared network was applied to each descriptor, the sum of feature elements is combined and the feature vector is output. In short, the calculation of channel attention is shown in Equation 1:

$$MC(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F_{max}^c)) + W_1(W_0(F_{avg}^c))) \quad (1)$$

Where: σ was sigmoid function; W_0 and W_1 where two shared weights for input characteristics.

Different from channel attention, the spatial attention mechanism focuses on the location of features, and mainly uses the spatial relationship of features to generate a spatial attention feature map. Spatial attention and channel attention are complementary. In order to calculate spatial attention, we first use the average pooling and maximum pooling operations of channels, and connect them to generate effective feature descriptors. Then the feature descriptor is used to generate the spatial feature map $Ms(F)$ through a convolution layer. As shown in Figure 7b, the process: (1) By using two pooling layers to aggregate the channel information of the feature map, two two-dimensional features F_{avg}^c and F_{max}^c where generated, representing the average pooling feature and the maximum pooling feature respectively; (2) The features where connected and convolved

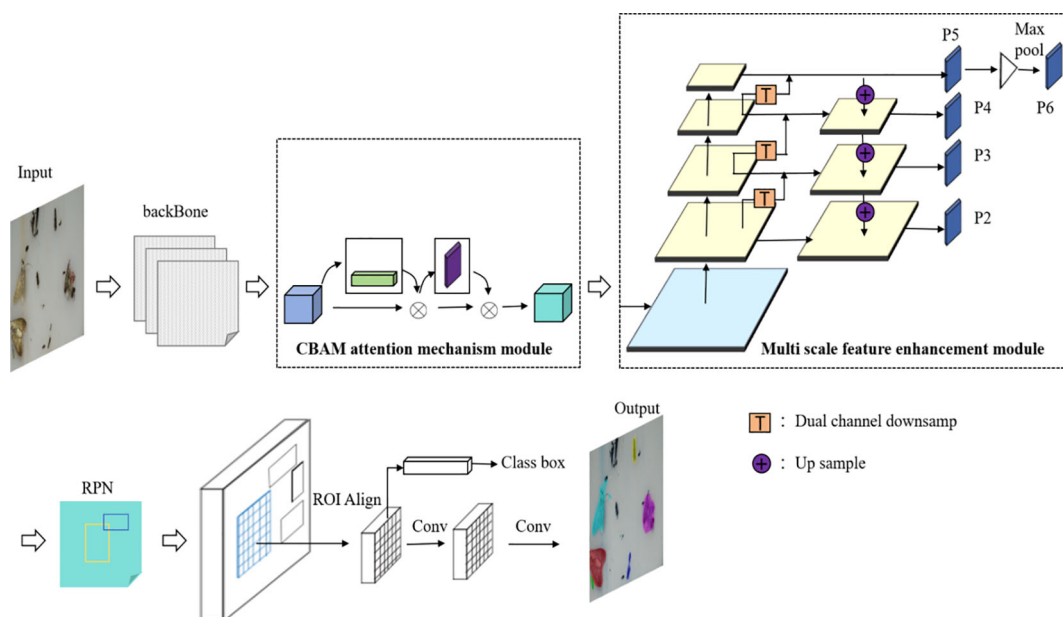
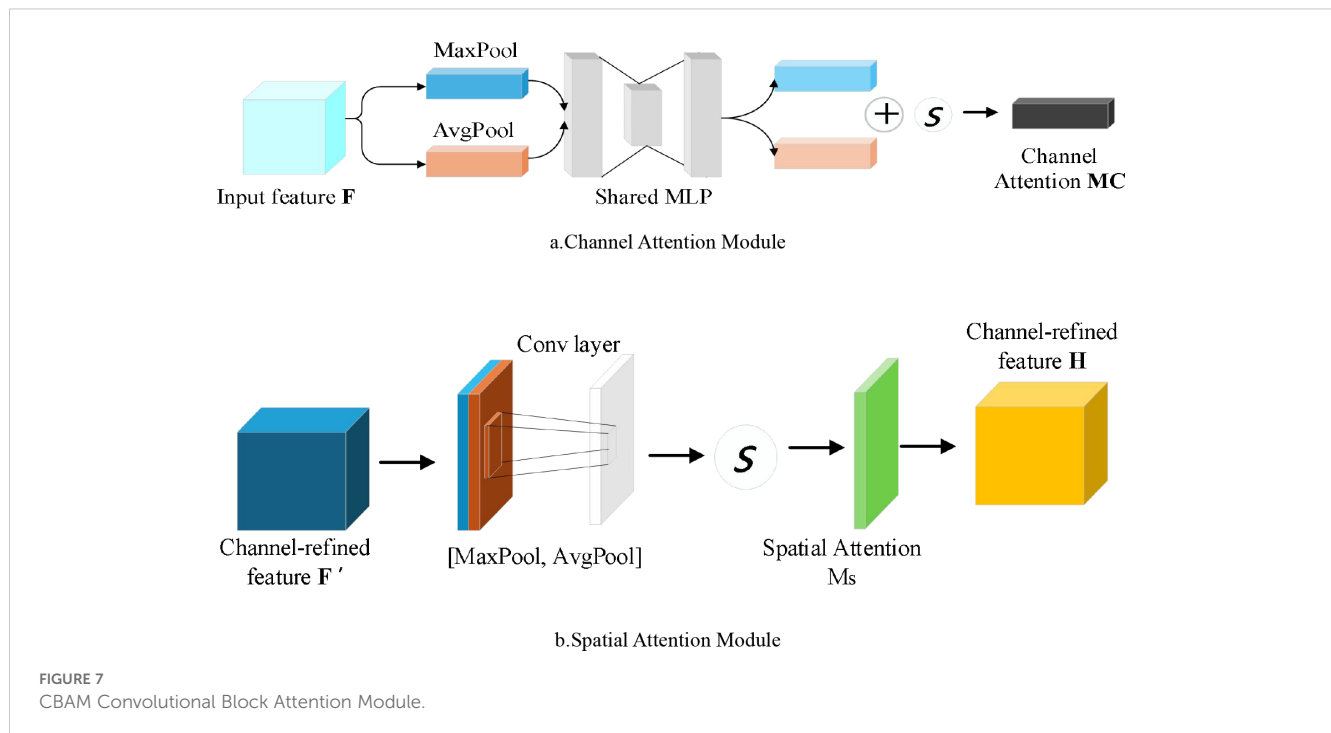


FIGURE 6
Improved Mask-RCNN network framework.



through the standard convolution layer, and input into the sigmoid function to generate the spatial attention feature map. In short, the calculation of spatial attention is shown in Equation 2:

$$\begin{aligned} M_s(F) &= \sigma(f^{7 \times 7}(\text{AvgPool}(F))); \\ (\text{MaxPool}(F)) &= \sigma(f^{7 \times 7}(F_{avg}^c; F_{max}^c)) \end{aligned} \quad (2)$$

Where: σ was sigmoid function; $f^{7 \times 7}$ represents the convolution operation with a filter size of 7×7 .

2.7 Feature enhanced FPN module

Figure 8 illustrates the entire feature transmission process of the FPN network, which consists of two parts: top down down sampling feature extraction and bottom up up sampling feature extraction. Down sampling extracts deeper image features with stronger semantic information, resulting in higher resolution features. Up sampling shallow features often contain rich spatial information about the object, which is crucial for locating the target in the image (Yu et al., 2022; Yuqi et al., 2023).

For example, when the input image feature is $1024 \times 1024 \times 3$, a convolutional layer with a stride of 2 is applied for dimensionality reduction. Since convolutional down sampling can cause some feature loss, an Identity Block module is used during the sampling process to enhance the network, allowing deeper layers to learn more complex features. The features are passed through the block to obtain features C_i , $i \in (\text{Jiang et al., 2019; Mendoza et al., 2022; Wei et al., 2022})$, and the C5 feature is up sampled and fused with

the C4 feature from the previous layer. After up-sampling, a convolutional operation is performed on the fused features to generate new features P4. After the bottom up up sampling process, a pooling operation is performed on the C5 feature to reduce information redundancy and prevent over fitting, producing new feature layers P_j , $j \in (\text{Jiang et al., 2019; Zongwang et al., 2021; Mendoza et al., 2022; Wei et al., 2022})$, which contain more semantic and spatial information.

2.8 Dual-channel down sampling module

The down sampling module is a method used to reduce the resolution of an image or feature map. However, commonly used down sampling methods often lead to the loss of detailed information. To reduce the feature loss during the down sampling process, this paper improves the down sampling module. The improved dual-channel down sampling module combines two transition modules. The left branch applies a 2×2 max-pooling operation followed by a 1×1 convolutional layer, while the right branch applies a 1×1 convolutional layer followed by a 3×3 convolutional layer with a stride of 2×2 . These two branches stack their results and output them together. Compared to traditional methods, this module better captures and processes the input image's features, improving object detection performance, and reduces the feature map's size without changing its depth. This improved down sampling module effectively reduces information loss and helps the network retain important detail information. The structure is shown in Figure 9.

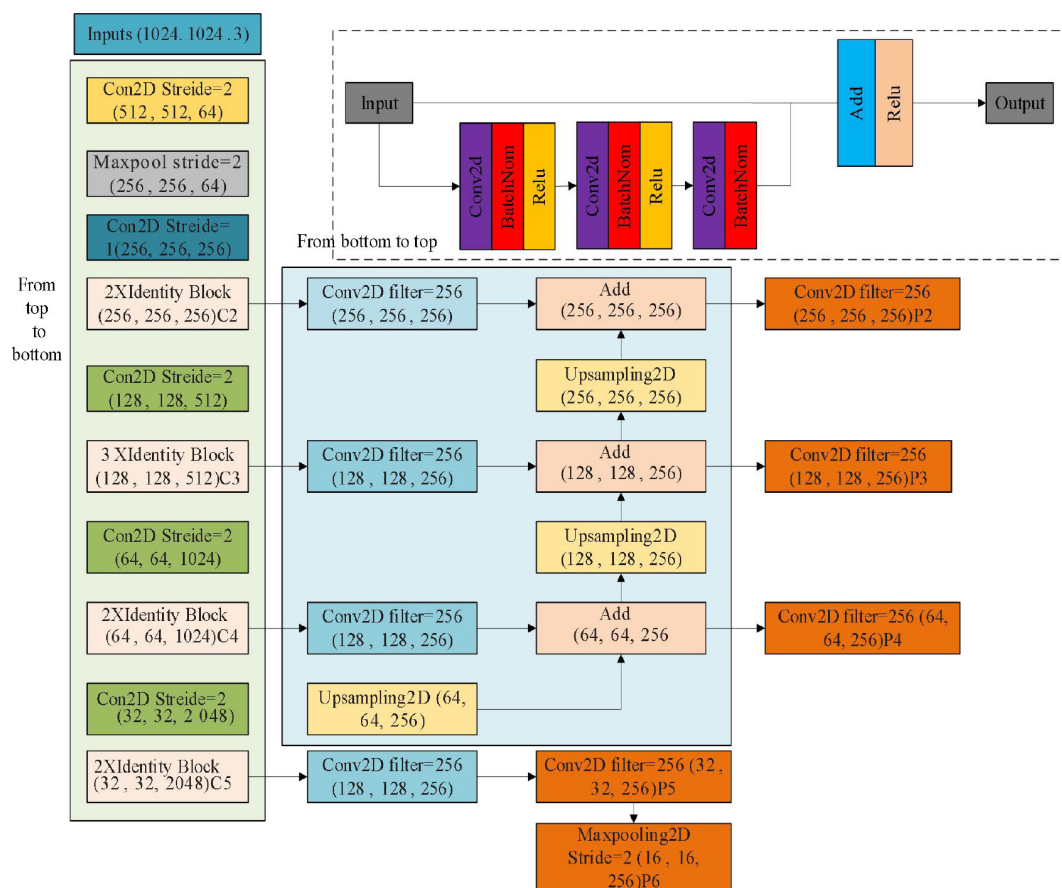


FIGURE 8
FPN characteristic conversion process.

2.9 Loss function

The loss function of Mask-RCNN-CBAM is a multi-task loss function that combines the losses of classification, localization, and segmentation masks, as shown in Equation 3:

$$L = L_{cls} + L_{box} + L_{mask} \quad (3)$$

Where: L_{cls} represents the classification loss, which is used to determine which category each ROI belongs to; L_{box} represents the bounding box offset loss, which is used to regress the boundary box of each ROI; L_{mask} represents the pixel segmentation mask generation loss, which is used to generate a mask for each ROI and each category.

3 Experiment and result analysis

3.1 Experimental environment

The programming language used in the experimental environment is Python 3. 8.10, the deep learning framework uses Tensorflow2.4.0, and the hardware environment is configured with Intel (R) Core (TM) i7-12700K × 20, The operating system is

Windows 10, the graphics card is NVIDIA GeForce RTX3080, and the graphics card driver uses Cuda11.6 and Cudnn9.6. The initial learning rate is set to 0.001, and the learning momentum is set to 0.9. The small batch size is set to 128. The weight falloff is set to 0.0005. The total number of iterations is also set to 300.

3.2 Evaluation index

In this paper, the Intersection over Union (IoU), Precision, Recall, and F1-score are adopted as evaluation metrics to comprehensively assess the performance of the model in pest detection. The IoU quantifies the overlap between the model-predicted region and the ground truth region, directly reflecting the accuracy of target localization (e.g., bounding boxes or pixel areas). For tasks like object detection or image segmentation, precise spatial localization is essential beyond mere classification correctness, and IoU serves as a critical indicator of positional accuracy. Precision measures the proportion of true positive samples among all predicted positives, emphasizing the model's ability to minimize false positives (misdetctions), making it particularly vital in scenarios where erroneous positive predictions must be strictly avoided. Recall, on the other hand, evaluates the ratio of correctly identified pests to the total actual pests, highlighting the model's capacity to reduce false

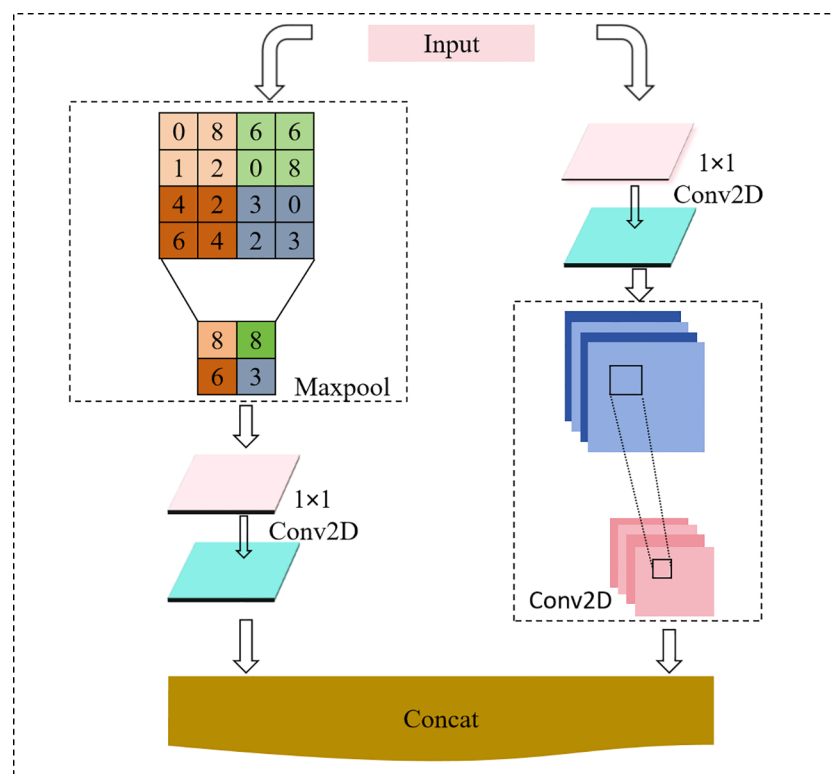


FIGURE 9
Dual-channel down sampled module.

negatives (missed detections). High Recall is prioritized when the cost of overlooking true positives is significant, such as in agricultural monitoring where missing pests could lead to severe consequences. The F1-score, as the harmonic mean of Precision and Recall, balances these two metrics and provides a robust evaluation in imbalanced class scenarios or cases requiring trade-offs between false positives and false negatives. Together, these metrics address localization accuracy (via IoU), classification reliability (via Precision and Recall), and overall robustness (via F1-score), ensuring a holistic assessment of the model's effectiveness, as detailed in Equations 4–7.

$$IoU = \frac{\text{Intersection}}{\text{Union}} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

Where: Intersection refers to the area of overlap between the model's predicted region and the ground truth region; Union refers to the area of the union between the model's predicted region and

the ground truth region; TP represents the pixels correctly classified as pests; FP refers to the positive samples that are incorrectly classified.

3.3 Experimental results analysis

To validate the accuracy and effectiveness of the Mask-RCNN-CBAM network in pest detection, this paper introduces three classic networks for comparative analysis: ResNet (Wang et al., 2025), Faster-RCNN, and Mask-RCNN. ResNet addresses the vanishing gradient problem in deep networks by introducing residual connections, allowing information and gradients to efficiently propagate through deep networks, which enables the training of very deep neural networks and significantly improves model performance. Faster-RCNN introduces a Region Proposal Network (RPN) and shared convolutional features to achieve end to end training, efficiently generate candidate regions, and provide high precision and efficiency in object detection. Mask-RCNN builds upon Faster-RCNN by adding a fully connected segmentation network for semantic segmentation and introducing the ROI Align module to accurately align pixels and handle the semantic segmentation problem. To validate the effectiveness of the proposed Mask-RCNN-CBAM network, it is compared with these classic networks. In order to more intuitively demonstrate the

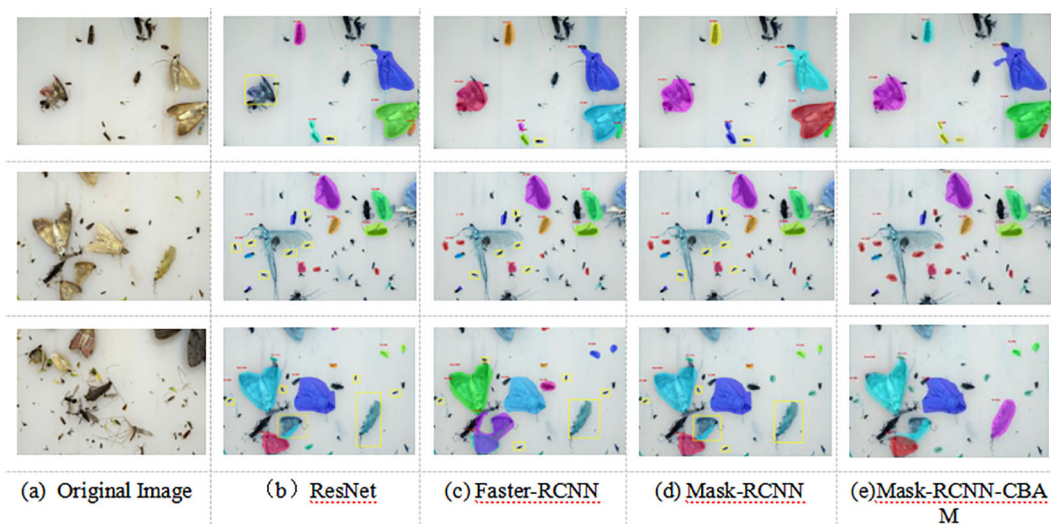


FIGURE 10
Visualization of Comparison Results for Classic Networks Experiment. (a) Original Image, (b) ResNet, (c) Faster-RCNN, (d) Mask-RCNN, (e) Mask-RCNN-CBAM.

performance of the improved model, the test set data is input into all four networks, and the pest detection results for each image are output in an end to end manner. The comparison of the recognition results from each model is shown in Figure 10.

Figure 10 shows the pest recognition results from three sets of images in the pest data set. The methods mentioned in this paper all perform well in pest recognition overall. However, there are differences in the model's ability to recognize pests in cases of complex backgrounds and dense pest populations. From the overall performance, the Mask-RCNN-CBAM model proposed in this paper achieves the best recognition results. The yellow boxes in Figure 10 represent the results of each method, showing false positives and false negatives in comparison to the proposed method.

In the first row, the ResNet network slightly struggles with mask extraction accuracy for individual pests in dense populations, resulting in less smooth masks. The ResNet network's optimization of detailed features for similar pests is weaker, leading to instances of pest segmentation being connected. Both the Faster-RCNN and Mask-RCNN networks show false negatives when detecting dense pests.

In the second and third rows of Figure 10, the ResNet and Faster-RCNN models perform poorly in extracting pests from structures that are complex or densely distributed. All three models ResNet, Faster-RCNN, and Mask-RCNN experience some false negatives and false positives, with the Mask-RCNN network showing the most severe false negatives. In terms of mask smoothness and segmentation quality, ResNet performs better than Faster-RCNN but is weaker than Mask-RCNN and the proposed Mask-RCNN-CBAM network.

The method proposed in this paper performs better at detecting pest dense areas and extracting pests. The Mask-RCNN-CBAM network effectively filters complex backgrounds and is more sensitive to pest features. In various pest detection tasks, the Mask-RCNN-CBAM network consistently shows good recognition

performance, with higher accuracy and smoother masks. This is because the proposed network integrates a dual channel attention mechanism that assigns greater weight to pest features while reducing the impact of background features, effectively distinguishing between pest and background features. Furthermore, the feature enhanced FPN (Feature Pyramid Network) merges shallow and deep features of the image, enriching the detail information and significantly improving the network's detection accuracy and efficiency.

To evaluate the effectiveness of the proposed method, a quantitative analysis of the experimental results was conducted. ResNet, Faster-RCNN, Mask-RCNN, and Mask-RCNN-CBAM models were tested on the pest dataset, and Table 3 presents the results from the test set. From Table 3, it can be observed that the proposed method outperforms the other three methods in terms of Precision, MIoU, and F1 score. Compared with ResNet, Precision, MIoU, and F1 increased by 0.73%, 1.38%, and 0.11%, respectively; compared with Faster-RCNN, the three metrics improved by 2.65%, 5.14%, and 3.35%, respectively; and compared with Mask-RCNN, the metrics improved by 2.44%, 0.6%, and 1.21%, respectively. Regarding the Recall metric, the proposed method showed improvements of 2.67% and 1.64% over Faster-RCNN and Mask-RCNN, respectively, while the improvement over ResNet was not significant. This could be due to the insufficient dataset size, leading to less obvious learning results.

From Table 3, it can be observed that the proposed method outperformed the other three methods in terms of MIoU, Recall, Precision, and F1 Score. MIoU (Mean Intersection over Union) reflects the model's ability to localize pests accurately, while Recall (95.21%) indicates a high coverage of true pest instances. The F1 Score (95.49%) demonstrates a balanced trade-off between precision and recall, crucial for dense pest detection in complex backgrounds.

Additionally, as shown in Table 3, the proposed network has a parameter size of 63.87MB, which is 1.32MB smaller compared to the Mask-RCNN network. This reduction is due to the optimization of

TABLE 3 Comparison of evaluation metrics for three classical network models.

Method	MIoU (%)	Recall (%)	Precision (%)	F1 Score (%)	Parameters (MB)
ResNet	90.07	95.16	95.18	95.38	7.77
Faster-RCNN	86.31	92.45	93.26	92.14	87
Mask-RCNN	90.85	93.57	93.47	94.28	65.19
Mask-RCNN-CBAM	91.45	95.21	95.91	95.49	63.87

TABLE 4 Comparison of indicators before and after integrating the attention mechanism module.

Method	Training Time (s/epoch)	F1(%)	AP(50)/ (%)	AR(small)/ (%)
With module	25	58.8	76.3	35.0
Without module	24	53.3	70.0	32.0

AP@50 (%) is the average accuracy when iou=0.5; AR (small) is the average recall rate of target<32 × 32 pixels.

certain parameters during the convolution process in the improved dual-channel attention module and the feature-enhanced FPN module. Compared to the original network, the proposed method not only uses fewer parameters but also extracts and segments pests more accurately, resulting in better detection performance. This demonstrates that the proposed improvement method achieves a better balance between segmentation accuracy and efficiency.

4 Discussion

4.1 Impact of attention mechanism module ablation on extraction results

To evaluate the impact of the attention mechanism module on extraction results, an ablation experiment was conducted. The base network for the experiment was the Mask-RCNN-CBAM, and the data set used was the constructed pest data set. The results of the ablation experiment quantitatively validating the effectiveness of the attention mechanism module on pest extraction are shown in Table 4.

Table 4 shows that after incorporating the attention mechanism module, the network's F1 score increased by 5.5%, and the accuracy of small target extraction improved by 3%. After the attention mechanism module was added, the network's ability to extract important feature information and allocate weights was optimized, making the network more sensitive to pest features and enhancing the effectiveness of feature extraction. Additionally, there was a slight change in the network's running time before and after the addition of the attention mechanism, with the running time being 1 second slower

before the module was added, although the efficiency of network operation remained similar. The experimental results confirm that the attention mechanism can enhance the accuracy of target detection, improve feature extraction performance, help the model focus on important features, reduce the sensitivity to noise or irrelevant information, minimize over fitting, and speed up convergence.

4.2 Impact of multi-scale feature fusion module ablation on extraction results

The FPN network enhances feature information and makes full use of multi-scale features. In this paper, improvements were made to the FPN network by adding top-down branches to better explore the image's detailed features. To validate the effectiveness of the feature enhanced pyramid module introduced by the proposed method, the quantitative experimental results are shown in Table 5. The table indicates that after adding the module, the network's F1 score increased by 0.4%, the AP value improved by 3.3%, and the recall rate for small target pests such as thrips and leaf hoppers increased by 6%. However, in the experiment, the network's running time was slightly higher after adding the module. This could impact the model's running efficiency when processing large datasets, but the model's accuracy significantly improved. Through this ablation experiment, it can be observed that the multi-scale feature fusion pyramid module enables the Mask-RCNN network to increase the model's receptive field, enrich the information in the feature layers, better understand object instances in the image, and adapt to targets at different scales, thereby improving detection and segmentation accuracy.

TABLE 5 Comparison of indicators before and after integrating the multi-scale feature fusion pyramid module.

Method	Runtime (s)	F1(%)	AP(50)/ (%)	AR(small)/ (%)
With module	50	55.2	74.1	41.3
Without module	46	54.8	70.8	35.0

4.3 Impact of dual channel down sampling module ablation on extraction results

To validate the effectiveness of the proposed dual channel down sampling module, this paper compares the performance of the max pooling down sampling and dual channel down sampling modules in the network. The dual channel down sampling module uses a 2×2 max pooling operation, followed by a 1×1 convolution compression module, and combines it with another 1×1 convolution followed by a 3×3 convolution kernel with a 2×2 stride. This approach reduces the down sampling stride in the feature transmission process and stacks convolution layers to decrease the number of channels. The experimental results are shown in Table 6. The dual-channel down sampling module improves the overall network's F1 score, AP value, and AR(small) by 0.8%, 3.1%, and 0.8%, respectively. The experiment demonstrates that the dual-channel down sampling module effectively reduces information loss, further enhancing the network's ability to retain feature information during feature transmission.

4.4 Model complexity and efficiency analysis

As demonstrated in Section 2.3, the introduction of the attention mechanism module, feature enhanced FPN module, and improved dual channel down sampling module does not increase the model's parameter count and, in fact, slightly reduces the model's complexity compared to the original model. Ablation experiments show that the inclusion of the CBAM attention mechanism module enhances the network's ability to extract contextual information from remote sensing images. The addition of the feature-enhanced pyramid module further improves the fusion of deep and shallow feature information. The design of the dual-channel down sampling module effectively reduces feature loss during the transmission process. These improvement modules optimize the network's ability to extract and analyze features for remote sensing image target detection tasks, positively impacting the efficiency and accuracy of remote sensing image target recognition.

Regarding model runtime efficiency, the introduction of the attention mechanism module and feature enhanced FPN module has a minimal effect on the model's processing time. The implementation demonstrates that the proposed Mask-RCNN-CBAM network performs quite well in terms of runtime efficiency for pest extraction tasks.

TABLE 6 Comparison of indicators before and after integrating the dual-channel down sampling module.

Method	F1(%)	AP(50)/ (%)	AR(small)/(%)
With module	54.5	73.2	35.1
Without module	53.7	70.1	34.3

4.5 Comparison with other studies

Compared to recent studies, our method demonstrates significant advantages. For instance, Li et al. (2024) achieved an F1 Score of 92.14% using Faster R-CNN on similar pest datasets, while our model improved this metric to 95.49%. SSD-based methods (Zongwang et al., 2021) in small pest detection (AR (small)=41.3% vs. 35.7%). Additionally, the parameter size of our model (63.87MB) is notably smaller than Wang et al.'s (2020) (87MB), indicating higher computational efficiency for field deployment. The integration of CBAM and dual-channel downsampling uniquely addresses small-target pest detection in dense backgrounds, a challenge less explored in prior work (Liu et al., 2023). Additionally, the integration of CBAM addresses background interference more effectively than the baseline Mask-RCNN (Liu et al., 2023), as evidenced by the 2.44% improvement in Precision. These advancements highlight the practical relevance of our approach for pest monitoring systems.

4.6 Limitations of the model

Although this study has achieved promising results, the proposed Mask RCNN CBAM model has certain limitations. Firstly, its generalization ability may be limited to pest species and environmental conditions similar to the training dataset. For example, in field experiments in foggy rice fields, the accuracy of the model decreased by 4.2% due to reduced image contrast. Secondly, although the parameter size is reduced by 1.39MB compared to Mask RCNN, the model still requires a GPU with ≥ 8 GB of VRAM for efficient inference, which limits its deployment on edge devices. These limitations highlight the necessity of future work in domain adaptation and model lightweighting. Therefore, our next research goal is to study a more lightweight detection model, so that it can run faster, and can be deployed on small-scale devices such as mobile phones to realize the application of remote monitoring system that can be used in intelligent agriculture.

5 Conclusion

In order to address the issues of false positives and false negatives in pest extraction caused by complex backgrounds and dense pest stacking, this paper proposes a Mask-RCNN-CBAM pest extraction network that integrates the attention mechanism. By incorporating the CBAM attention mechanism into the feature processing process and enhancing multi-scale feature extraction and fusion through a feature pyramid module, the network's ability to extract contextual information is strengthened, and the loss of image detail features is reduced. Experimental results show that the Mask-RCNN-CBAM network performs excellently in extracting targets on the pest data set, achieving higher precision and better performance, particularly under complex backgrounds and dense pest conditions. It achieves the highest performance in Precision, F1

score, and Recall, with lower false positive and false negative rates, demonstrating that the network model is reliable and applicable.

Compared to other pest extraction methods, the Mask-RCNN-CBAM network can better extract pest feature information and optimize detail information. However, during detection, there are still some false negatives due to the similarity between the texture of some pests and the background. Future work will continue to focus on the attention mechanism, enhancing the network's ability to extract and optimize image detail information, further improving the pest extraction capabilities of the proposed method.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

XiW: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft. CH: Conceptualization, Data curation, Methodology, Resources, Supervision, Writing – review & editing. XuW: Formal analysis, Investigation, Writing – review & editing. HZ: Validation, Writing – review & editing. XC: Supervision, Writing – review & editing. SY: Resources, Writing – review & editing. JZ: Validation, Writing – review & editing. JL: Data curation, Investigation, Validation, Writing – review & editing. ZY: Writing – review & editing, Methodology, Visualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work is supported by

the Major Project of Anhui Province University Innovation Team Project, Digital Agriculture Innovation Team (2023AH010039), Anhui Province University Science Research Project (2022AH051029), the Major Science and Technology Project of Anhui Province (202203a06020007), Anqing Normal University High-level Talents Introduction Project, and the guidance of researchers Xie Chengjun and Zhang Jie from the Institute of Intelligent Machinery, Hefei Institute of Materials Science, Chinese Academy of Sciences.

Conflict of interest

Authors JZ and JL were employed by the company Anhui Yi Gang Information Technology Co.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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.

References

- Deepika, P., and Arthi, B. (2022). Prediction of plant pest detection using improved mask FRCNN in cloud environment. *Measure.: Sens.* 24, 100549–100557. doi: 10.1016/j.measen.2022.100549
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask-RCNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy. 2980–2988. doi: 10.1109/ICCV.2017.322
- Hu, Y., Deng, X., Lan, Y., Chen, X., Long, Y., and Liu, C. (2023). Detection of rice pests based on self-attention mechanism and multi-scale feature fusion. *Insects* 14, 280. doi: 10.3390/insects14030280
- Jiang, M., Li, X., Xin, L., and Tan, M. (2019). The impact of paddy rice multiple cropping index changes in southern China on national grain production capacity and its policy implications. *J. Geogr. Sci.* 74 (1), 1773–1787. doi: 10.1007/s11442-019-1689-8
- Kasinathan, T., and Uyyala, S. R. (2023). Detection of fall armyworm (spodoptera frugiperda) in field crops based on Mask-RCNN. *Sig. Image Video Process.* 17, 2689–2695. doi: 10.1007/s11760-023-02485-3
- Khalid, S., Oqaibi, H. M., Aqib, M., and Hafeez, Y. (2023). Small pests detection in field crops using deep learning object detection. *Sustainability* 15, 6815. doi: 10.3390/su15086815
- Li, K.-R., Duan, L.-J., Deng, Y.-J., Liu, J.-L., Long, C.-F., and Zhu, X.-H. (2024). Pest detection based on lightweight locality-aware faster-RCNN. *Agronomy* 14, 2303.
- Li, W., Zhu, T., Li, X., Dong, J., and Liu, J. (2022). Recommending advanced deep learning models for efficient insect pest detection. *Agriculture* 12 (7), 1065–1073. doi: 10.3390/agriculture12071065
- Liu, S., Fu, S., Hu, A., Pan, M., Hu, X., Tian, X., et al. (2023). Research on insect pest identification in rice canopy based on GA-Mask-RCNN. *Agronomy* 13, 2155–2173. doi: 10.3390/agronomy13082155
- Mendoza, C. D. P., and Garcia, D. (2022). Black soldier fly or wasp: An instance segmentation using Mask-RCNN[EB/OL]. doi: 10.13140/RG.2.2.13110.78401
- Mendoza, C. D. P., and Garcia, D. (2022). Sequential variation analysis of forest pest disasters in China from 1998 to 2018. *Sci. Silvae Sinicae* 58, 134–143.
- Rong, M., Wang, Z., Ban, B., and Guo, X. (2022). Pest identification and counting of yellow plate in field based on improved Mask R- CNN. *Discr. Dynam. Nat. Soc.* 28, 1155–1164. doi: 10.1155/2022/1913577
- Suganya Kanna, S., Premalatha, K., Kavitha, K., Vijayakumar, M., Dheebakaran, G., and Bhuvaneswari, K. (2023). Effect of Weather Parameters on of Pests and Diseases in Groundnut and Castor in Salem District of Tamil Nadu, India. *Int. J. Environ. Climate Chang.* 13 (10), 3652–3659. doi: 10.9734/IJECC/2023/V13I103035
- Wang, J., Qiao, X., Liu, C., Wang, X., Liu, Y. Y., Yao, L., et al. (2021). Automated ECG classification using a non-local convolutional block attention module. *ComputerMethods Progr. Biomed.* 203, 106006. doi: 10.1016/j.cmpb.2021.106006

- Wang, F., Wang, R., Xie, C., Yang, P., and Liu, L. (2020). Fusing multi-scale context-aware information representation for automatic in-field pest detection and recognition. *Comput. Electron. Agric.* 169, 105222. doi: 10.1016/j.compag.2020.105222
- Wang, Q., Zhang, S. Y., Dong, S. F., Zhang, G. C., Yang, J., Li, R., et al. (2020). Pest24: A large-scale very small object data set of agricultural pests for multi-target detection. *Comput. Electron. Agric.* 175, 105585. doi: 10.1016/j.compag.2020.105585
- Wang, J., Zhang, B., Yin, D., and Ouyang, J. (2025). Distribution network fault identification method based on multimodal ResNet with recorded waveform-driven feature extraction. *Energy Rep.* 13, 90–104. doi: 10.1016/j.egy.2024.12.012
- Wei, Z., He, H., Youqiang, S., and Wu, X. (2022). AgriPest-YOLO: A rapid light-trap agricultural pest detection method based on deep learning. *Front. Plant Sci.* 13, 2022. doi: 10.3389/fpls.2022.1079384
- Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In: V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (eds) *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science()*, vol 11211. (Cham: Springer International Publishing), 3–19. doi: 10.1007/978-3-030-01234-2_1
- Yu, M., Chen, X., Zhang, W., and Liu, Y. (2022). Building extraction on high-resolution remote sensing images using attention gates and feature pyramids tructure. *J. Geo-Inform. Sci.* 24, 1785–1802. doi: 10.12082/dqxxkx.2022.210571
- Yunong, T., Shihui, W., En, L., Yang, G., Liang, Z., and Tan, M. (2023). MD-YOLO: Multi-scale Dense YOLO forsmall target pest detection. *Comput. Electron. Agric.* 213, 108233. doi: 10.1016/j.compag.2023.108233
- Yuqi, C., Xiangbin, Z., Yonggang, L., Wei, Y., and Ye, L. (2023). Enhanced semantic feature pyramid network for small object detection. *Signal Process.: Image Commun.* 113, 116919. doi: 10.1016/j.image.2023.116919
- Zhu, X., Jia, B., Huang, B., Li, H., Liu, X., and Seah, W. K. G. (2024). “Pest-YOLO: A Lightweight Pest Detection Model Based on Multi-level Feature Fusion,” in *Advanced Intelligent Computing Technology and Applications. ICIC 2024. Lecture Notes in Computer Science*, vol. 14865. Eds. D. S. Huang, C. Zhang and W. Chen (Springer, Singapore).
- Zongwang, L., Huifang, J., Tong, Z., Fuyan, S., and Xu, H. (2021). Small object recognition algorithmof grain pests based on SSD feature fusion. *IEEE Access* 9, 43202–43213. doi: 10.1109/ACCESS.2021.3066510