2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

An Enhanced Multi-Scale Feature Perception and Improved Feature Extraction-Based Algorithm for Cotton Pest and Disease Detection

Zhisheng Wang[1], Lizhen He[1], Jiayu Peng[1], Yiwei Duan[1], Qi Liu[2],
Zhaohui Wang[2], Xin Yang[1], and Jinhai Sa[1(✉)]

[1] School of Softball, Xinjiang University, Urumqi 830091, PR China
[2] School of Business, Henan University of Science and Technology, Luoyang 471023, PR China

Abstract. Aiming at the significant variability of cotton leaf pests and diseases in terms of shape, size, and location distribution in the natural environment, as well as the shortcomings of the existing detection models in terms of parameter optimization and detection efficiency, this paper proposes an algorithm for detecting cotton pests and diseases based on enhanced multi-scale feature sensing and improved feature extraction, named SDRM-YOLO. First, to improve the model's ability to perceive pest and disease feature information and spatial localization accuracy, we propose a Multi-Scale Cross-Space Perception Attention (MCPA), which is a mechanism that effectively enhances the model's ability to focus on key target areas by fusing spatial information at multiple scales. Second, to improve the feature extraction quality of the model, we design the C2f-DCN-RCSOSA (C2f-DR) module, which enables the model to capture the foreground features of the target flexibly and, at the same time, strengthens the focus on the key regions to enhance the feature expression capability. Finally, to reduce the computational complexity of the model and improve the detection speed, we introduce a lightweight network, ShuffleNetv2-RC, into the backbone network to optimize the computational efficiency and maintain a high detection accuracy. The experimental results show that SDRM-YOLO outperforms other state-of-the-art target detection algorithms on both the Cotton Disease Dataset and Cotton Pest Detect Dataset datasets. Compared with the benchmark model YOLOv8n, the mAP50 metrics were improved by 7.3% and 2.5%, significantly enhancing cotton pest detection's accuracy and robustness.

Keywords: Pests and diseases, target detection, feature extraction, multi-scale feature sensing

1 Introduction

As global agriculture has expanded, crop pest and disease monitoring and management have become critical to agricultural output. A lucrative crop essential to the financial systems of many countries worldwide is cotton [1]. Yet, it is regularly subjected to various pests and illnesses during its growth [2], notably impairing cotton quality and output [3]. Traditional pest and disease detection methods primarily depend on manual visual inspection, which is often time-consuming, laborious, and difficult to achieve in large-scale real-time monitoring. In addition, traditional recognition means having difficulty performing manual identification and extraction of features, which increases labor costs and the computational process's complexity. Therefore, intelligent, rapid, and accurate detection of cotton pests and diseases is crucial for disease prevention and pesticide management strategies.

With the rapid progress of deep learning technology, target detection techniques achieve automatic extraction of key features from images and effectively solve the classification and localization challenges of multiple targets in a single image [4]. Target detection methodologies can primarily be categorized into two distinct types: two-stage detection algorithms and single-stage detection algorithms. Two-stage target detection algorithms are represented by R-CNN [5], Fast R-CNN [6], and Faster R-CNN [7], which first generate candidate regions in an image and then subsequently classify and bounding-box regression on these regions [8]. Although such methods excel in detection accuracy, they increase computational complexity and significantly affect the inference speed of the model. In contrast, one-stage target detection algorithms, such as SSD [9], RetinaNet [10], and YOLO [11-14], perform the task of localizing and classifying targets directly on the image. Although these algorithms may be slightly inferior to the two-stage algorithms in detection accuracy, they significantly improve model inference speed. They are more suitable for applications on resource-constrained devices.

In recent years, deep learning technology has made remarkable achievements in image recognition, especially the application of convolutional neural networks, which provide a new solution for automatically detecting crop pests and diseases. As a prominent figure in the realm of object detection, the YOLO series of algorithms has a fast detection speed and reasonable accuracy. YOLOv8n [15] represents one of the latest iterations within its algorithmic series, building upon and refining the strengths of its prior versions. It achieves superior precision in detection and quicker computational performance. However, based on current research findings, particular notable challenges persist when applying YOLOv8n to detect cotton pests and diseases:

(1) In cotton pest and disease detection scenarios, targets appear at different scales, ranging from tiny localized lesions to larger-scale plant damage. The traditional single-scale convolutional kernel is limited to a fixed receptive field, making it difficult to adapt to targets at different scales. It may weaken the accuracy and robustness of detection, especially when the target distribution is complex, or the size varies greatly [16].

(2) In the field of cotton pest and disease detection, it is difficult for the model to accurately capture spots or pests with different shapes and variable locations. The traditional feature extraction paradigm is prone to introducing irrelevant contextual information when extracting features, failing to focus on key areas, thus affecting the detection accuracy and model performance.

(3) Despite YOLOv8n's effective trade-off between detection speed and accuracy, its considerable parameter count and substantial computational demands present difficulties for devices with limited resources.

To address these challenges, this study proposes an algorithm for cotton pest and disease detection based on enhanced multiscale feature sensing and improved feature extraction, and the contributions of this paper are as follows:

(1) To solve the problems of insufficient robustness of model spatial localization perception and the inability of single-scale convolution to adapt to multi-scale targets, we propose the MCPA module. This module significantly improves the model's detection accuracy of pest and disease targets in farmland by introducing convolution kernels of different sizes to extract multi-scale features.

(2) To solve the problems of traditional convolutional sensory field fixation and poor quality of feature extraction, this paper proposes DCNv2 deformable convolution to improve the C2f module and combines it with the RCS-OSA module to enhance the critical region focus. This method improves the match between features and target shape [17].

(3) In order to develop a lightweight backbone for feature extraction, we utilize the ShuffleNetv2-RC architecture. This particular network enhances the precision of pest detection and substantially decreases the quantity of model parameters [18].

2 Related work

YOLO algorithm directly predicts object classification at each position in the feature map, which has obvious advantages in time efficiency and recognition rate and is more suitable for real-time object detection of crop pests and diseases. Hu et al. [19] proposed InterImage as the core operator with deformation convolution. Unlike the traditional CNN scheme, deformation convolution has an effective sensory field and can input and task adaptive spatial domain aggregation. Huang et al. [20] proposed a lightweight Van-YOLOv8 model using a lightweight MobilenetV2 [21] and convolutional block attention modules. Although this method significantly trims down the model's parameter count and lowers its computational demands, it might simultaneously cause a decline in detection accuracy. Chen et al. [22] introduced the Asymptotic Feature Pyramid Network AFPN [23] (AFPN) into YOLOv8 to solve the feature loss problem in multi-scale fusion to address the issue of significant differences in the scales of the detection targets, etc. In addition, the network's feature extraction capability is further improved by replacing the base module in the AFPN with an efficient aggregation network module. However, the improved network is computationally intensive and unsuitable for real-time detection. Vasanthi et al. [24] replaced some of the convolutional layers of

YOLOv8 with the phantom convolution Ghostconv [25]. They introduced a global attention module to effectively capture the global context information of the input feature maps. This enhancement significantly increased detection accuracy; however, the improved model still contains a large number of parameters. Qi et al. [26] improved YOLOv5 by introducing a visual attention model, enhancing feature extraction, and modifying the loss function to account for the overlapping inclusion of the predicted and actual frames. This makes the network converge faster, but the model detection accuracy is not high. Li et al. [27] proposed an enhanced C2f module by integrating the concepts of DenseBlock [28] and the DCF module, thereby improving the extraction of low-level features within the YOLOv8 model. They finally replaced the CBS activation function with the Mish activation function, which effectively solved the problem of gradient disappearance during the training process.

In summary, although the YOLO algorithm and its improved versions have demonstrated significant advantages in crop pest and disease detection, there are still challenges in its application. The models cannot capture features at different scales in the spatial dimension and perform poorly in identifying critical regions. The feature extraction quality of existing models is not high when facing complex disease features. This affects the detection performance to some extent as the complexity of the model increases.

3 Method

3.1 SDRM-YOLO network architecture

The network framework proposed in this study is built on the foundation of YOLOv8n, as shown in Fig.1. Firstly, we integrate a multi-scale cross-spatial perceptual attention mechanism (MCPA) between the neck network and the prediction head, which utilizes convolutional kernels of different sizes and combines spatial information to enhance the channel attention to improve the model's perceptual effect and spatial localization of multi-scale features. Second, aiming to enhance the precision and effectiveness of feature extraction, we reengineered the original C2f module and incorporated the C2f-DCN-RCSOSA module as a substitute for the original module in the neck network. Second, to elevate the quality and accuracy of feature extraction, we redesigned the original C2f module and introduced the C2f-DCN-RCSOSA module to replace the original C2f module in the neck network. This enhancement notably boosts the model's ability to extract features from target objects that exhibit significant variations in shape and position, while directing greater attention toward critical information-rich areas. Ultimately, we utilize the upgraded ShuffleNetV2-RC network as the core framework. This choice helps alleviate the issue of gradient vanishing while maintaining the model's lightweight structure.

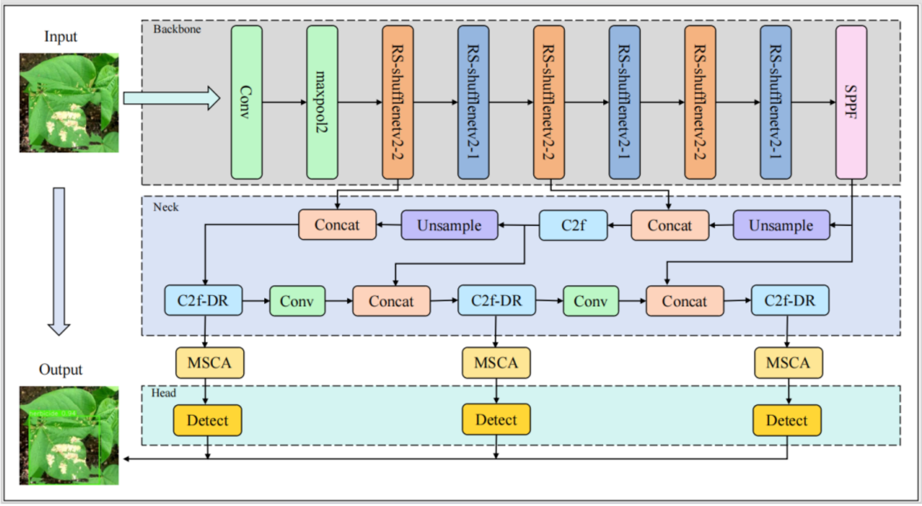

Fig. 1. Overall Structure of SDRM-YOLO Model

3.2 Multi-Scale Cross-Space Perception Attention (MCPA)

In the complex environment of actual farmland cotton leaves, disease areas usually vary significantly in size and often shade each other. In addition, the traditional single-scale convolutional kernel is too complex to adequately capture the features of targets at different scales due to the fixity of its receptive field, especially when the target distribution is complex or the size changes are variable, which may affect the accuracy and robustness of the detection. To address this challenge, we propose the MCPA module and apply it between the neck network and the prediction head to enhance the model's ability to capture the spatial localization perception of features with long-range dependencies.

MCPA is shown in Fig.2. Specifically, the mechanism performs a global averaging pooling operation in both directions on the input feature graph X (size c×h×w). First, the features of each channel are encoded along the horizontal and vertical directions by pooling x_c; two feature mappings, Z^h and Z^w, for the c channel at height h and width w, are obtained, respectively, which enables the model to get the feature information of the target more accurately. Where i and j denote the horizontal and vertical directions, respectively. The formula is:

$$Z_c^h(h) = \frac{1}{w}\sum_{0\le i<w}^{w} x_c\,(h,i) \tag{1}$$

$$\mathrm{Z}_c^{\mathrm{w}}(\mathrm{w}) = \frac{1}{\mathrm{H}}\sum_{0\le \mathrm{j}<\mathrm{H}}^{\mathrm{H}} \mathrm{x}_c(\mathrm{j},\mathrm{w}) \tag{2}$$

Next, these two feature mappings are merged to obtain a representative feature representation. This tensor is then subjected to a 1 × 1 convolutional layer $F_1$for feature mapping nonlinear learning and processed by a nonlinear activation function σ, which ultimately generates an intermediate feature mapping $f_{1\times1}$. The formula is：

$$f_{1\times1} = \sigma(F_{1\times1}([Z^h, Z^w]) \quad (3)$$

Subsequently, the feature mappings after 3x3,5x5 and 7x7 convolution operations are concatenated and summed to generate the final intermediate feature mappings f. Since convolutions of different sizes are able to capture feature information at various scales. Among them, the 3×3 convolution can capture local detailed features, while the 5×5 and 7×7 convolutions can enhance the capture of contextual information and extract a broader range of contextual information. This multi-scale feature extraction can better adapt to diverse pest targets and enhance the model's ability to adapt to complex scenes. The formula is.

$$f = \sigma(F_{3\times3}([Z^h, Z^w]) + F_{5\times5}([Z^h, Z^w]) + F_{7\times7}([Z^h, Z^w]) \quad (4)$$

Then, the intermediate feature map f is divided into two separate feature tensors, f^h and f^w, based on spatial dimensions. These tensors are then passed through 1×1 convolutional layers to match the channel count of the original input X, resulting in the feature tensors $F_h(f^h)$ and $F_w(f^w)$. Ultimately, these tensors undergo a Sigmoid activation function to produce the attention weights gh and gw for the height and width dimensions, respectively. The formula is as follows：

$$g^h = \sigma(F_h(f^h)) \quad (5)$$

$$g^w = \delta(F_w(f^w)) \quad (6)$$

Finally, the original input feature map X is weighted by element-by-element multiplication with the channel attention weights to obtain the weighted feature map output after processing by the MCPA module, Eq:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) x + y = z \quad (7)$$

The MCPA attention structure is shown in Fig. 2.

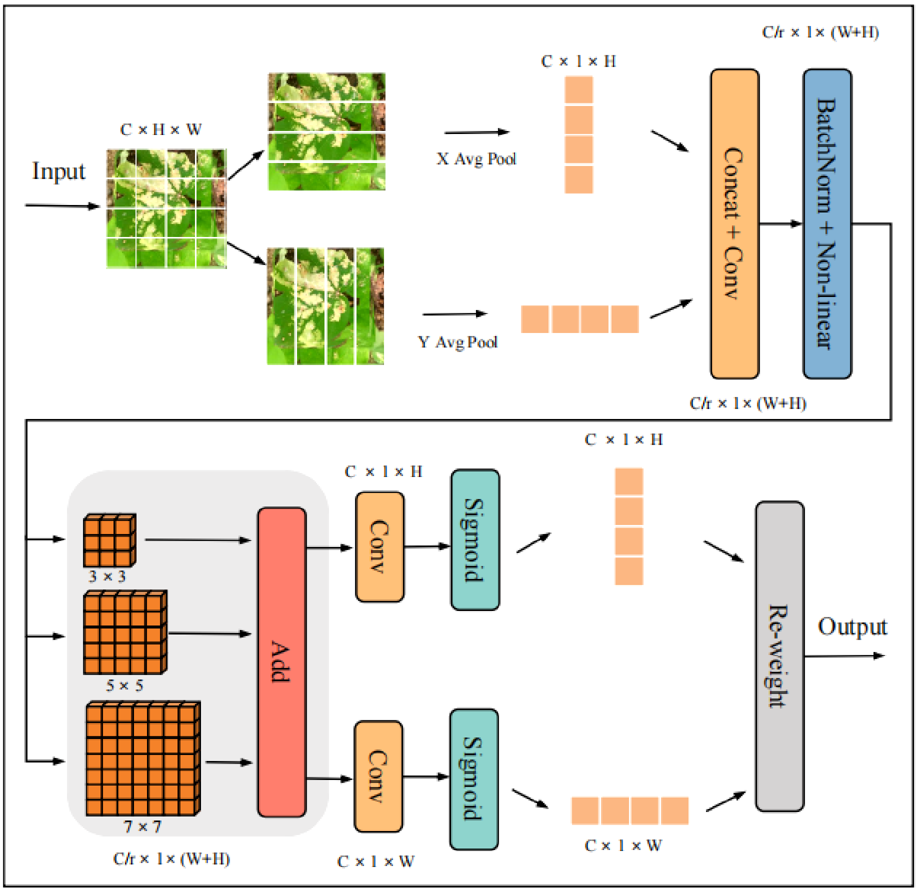

Fig. 2. MCPA module structure diagram

Through this process, the multi-scale features output from the feature extraction module can be adaptively enhanced and fused to generate highly discriminative feature representations. These representations provide richer semantic information and spatial details for the prediction header, ultimately realizing more accurate pest and disease detection in complex farmland environments.

3.3 C2f-DCN-RCSOSA module (C2f-DR)

Since the receptive field of the traditional convolutional kernel is fixed, it is difficult for the model to flexibly adjust the receptive field when dealing with irregular shapes and geometrical variations of objects. In addition, for targets to be detected with different sizes and shapes of cotton leaf diseases, the target features extracted by the static convolution kernel often contain a large amount of irrelevant contextual information, affecting the detection accuracy. Conversely, deformable convolution enhances traditional convolution by incorporating a learnable offset mechanism. This enables the convolution kernel to dynamically adjust its sampling positions based on the target's unique geometry and spatial configuration. Consequently, the convolution operation becomes more attuned to the object's outline, resulting in more accurate and robust feature representation.

To address the limitations of standard convolution operations, we incorporated the DCNv2 layer [30] into the C2f module to boost its capabilities. Specifically, we created the C2f-DCNv2 module by replacing a key convolutional layer in the Bottleneck with the DCNv2 layer.Given that deformable convolution offers substantial flexibility in sampling positions, we believe it is particularly well-suited for application in the neck of the network. By integrating deep and shallow features, the neck network ensures that the output feature maps possess detailed information and abundant semantic content while eliminating unnecessary noise. This process facilitates learning more rational deformation paths by the deformable convolution.

Traditional convolution involves a two-stage process: initially, a convolution kernel K with a defined size is applied to sample the input feature map x; second, a weighted summation of the sampled values and the convolution kernel parameter w is performed. Through this process, the value of the convolution result at any position P_0 on the resultant feature map y is:

$$y(p_0) = \sum_{P_n \in R} w(P_n).\, x(P_0 + P_n) \tag{8}$$

Among them, P_n is an enumeration of positions in K.

Conversely, the deformable convolution introduces a positional offset $\Delta \mathrm{P_n}$to Eq. (8), and for each position in K, an offset P_0 is learned.

$$y(p_0) = \sum_{P_n \in R} w(P_n).\, x(P_0 + P_n) \tag{9}$$

The deformable convolution is sampled at position $\mathrm{P} = \mathrm{P_0} + \mathrm{P_n} + \Delta\mathrm{P_n}$, where the offset pn is obtained by learning and is usually fractional, so x(p) is obtained by bilinear interpolation:

$$x(P) = \sum_q G(q, p).\, x(q) \tag{20}$$

Where q denotes an integer position in the feature map and G denotes a bilinear interpolation operation.

The C2f-DCNv2 module can more precisely capture features that correspond to the object's shape by dynamically adjusting its sampling strategy. This adjustment allows the module to better align with the object's geometry, enhancing the network's overall robustness to object deformation. In addition, to further enhance the model's focus on key areas of cotton leaf diseases, we use the high-quality features extracted by DCNv2 as inputs to the Reduced Channel Spatial Object Attention (RCSOSA) module [31]. The RCSOSA module leverages a spatial attention mechanism to compute the significance weights for each location, allowing the model to concentrate on areas that are most likely to hold crucial information and disregard irrelevant background details. In addition, RCSOSA aggregates spatial information from multiple directions through the spatial attention mechanism, ensuring that the model comprehensively understands the target and its surroundings. This mechanism further enhances the model's attention to critical areas. In both mechanisms, DCNv2's offset learning adapts the receptive field to the target contour, while RCSOSA further strengthens the semantic key points through orientation-sensitive attention.The structure of C2f DR module is shown in the following fig. 3:

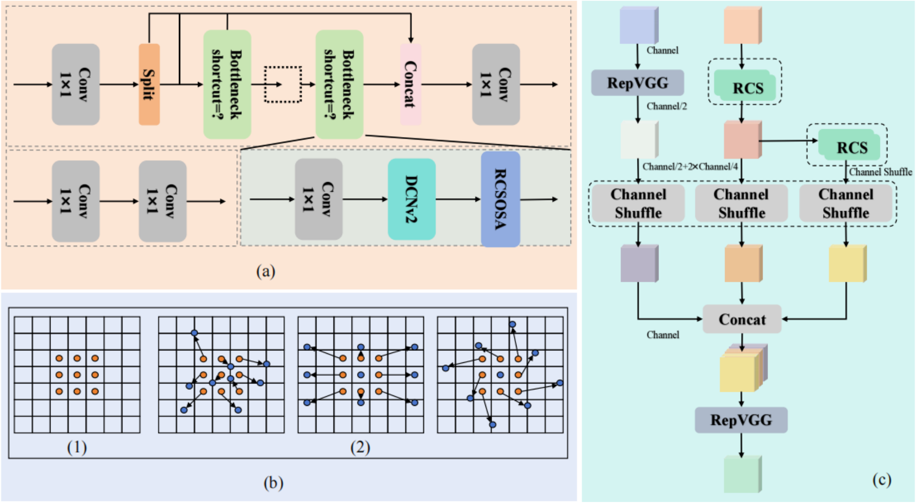

Fig. 3. (a) shows the C2f-DR model improvement location map, (b) shows the sampling location comparison between traditional convolution (1) and deformable convolution (2), and (c) shows the structure of the RCSOSA model

3.4 Residual-Mixed scrubbing cooperative network（Shufflenetv2-RC）

In the context of crop cultivation, continuous surveillance of cotton diseases is essential for both field and laboratory settings. These scenarios usually face a major challenge: mobile devices have limited computational resources to support the efficient operation of complex models. Although the traditional ShuffleNetV2 is lightweight through channel segmentation and depth-separable convolution, its linear stacking structure has a fundamental flaw in building deep networks. When the network depth exceeds 50 layers, the gradient decays exponentially in backpropagation, resulting in gradient degradation rather than mere disappearance.

This study proposes the ShuffleNetv2-RC module to address this bottleneck, as shown in Fig. 4 below. ShuffleNetv2-RC further proposes a dual-path feature interaction architecture, which innovatively combines the advantages of channel attention and spatial convolution and solves the gradient vanishing and exploding problems in deep network training by introducing residual connectivity. Specifically, the primary path adopts a “compression-excitation” optimization structure. Firstly, the channel is compressed by 1×1 convolution. After swish nonlinear activation, 3×3 depth-separable convolution is used for efficient spatial feature extraction, and finally, the dimensional reconstruction is achieved by 1×1 convolution. A batch normalization layer is introduced to optimize the gradient flow. The auxiliary path, on the other hand, is designed with an adaptive gating mechanism, and when the dimensionality of the input and output channels do not match, grouped convolution (groups=4) is used to realize efficient feature projection, which effectively reduces the computational overhead compared with the standard 1×1 convolution.

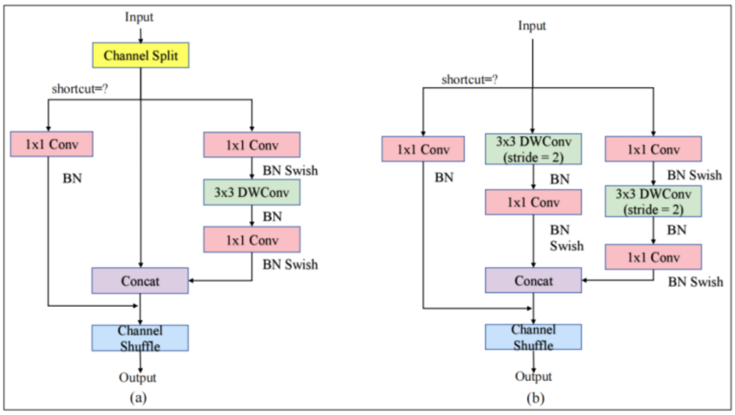

Fig. 4. (a) is the basic unit, (b) is the downsampling unit

4 Experiment

4.1 Data sets and evaluation indicators

This paper uses two datasets for experiments: the Cotton Disease Dataset and the Cotton Pest Detect Dataset.

Cotton Disease Dataset: The photos of cotton leaf pests and diseases are from the China Agricultural Pests and Diseases Research Image Library. The format is.jpg, and after re-annotation, the length and width of the photos are 320 pixels. The disease samples in the data set include Alternaria Leaf Spot, Curl Leaves, Red Spot, bacterial blight, foliar disease, and herbicide, which are common in Xinjiang, China. The dataset was split using a script into 1094 training, 143 validation, and 285 testing photos.

Cotton Pest Detect Dataset: This dataset contains 2319 photographs in 6 categories: blight, curl, grey mildew, healthy, leaf spot, and wilt. The images are divided according to 7:1:2, where the training set is 1585, the validation set is 235, and the test set is 499 sheets.

Precision indicates the proportion of detected targets that are proper targets. Recall measures the proportion of all actual targets that have been correctly detected. Parametric quantity denotes the total count of all trainable parameters within the model. GFLOPs serve as a metric for assessing the model's computational complexity. mAP is another key evaluation metric, representing the mean of average precision (AP). For each category, the AP value can be calculated as the area under the precision-recall curve for all samples in that category across different thresholds. The mapped value is then obtained by averaging the AP values of all categories, and it is computed as follows:

$$P = \frac{TP}{TP+FP} \tag{31}$$

$$R = \frac{TP}{TP+FN} \tag{42}$$

$$AP = \int_0^1 P(R)\, dR \tag{53}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{1} AP_i \tag{64}$$

$$params = O(\sum_{i=1}^{n} M_i^2 * K_i^2 * C_{i-1} * C_i) \tag{75}$$

$$GFLOPs = O(\sum_{i=1}^{n} K_i^2 * C_{i-1}^2 * C_i + \sum_{i=1}^{n} m^2 * C_i) \tag{86}$$

False positives (FP) represent the number of samples where annotation frames were generated but were either in the wrong position or had incorrect category labels. True positives (TP) denote the number of lesion regions that were accurately detected. False negatives (FN) refer to the number of samples without annotation frames generated within the lesion regions. Additionally, K indicates the size of the convolution kernel, C represents the number of channels, M signifies the size of the input image, O stands for the number of constant orders, N represents the number of defects, and I denotes the number of iterations. AP is the area under the precision-recall curve, while mAP is the mean of AP values across different categories.

4.2 Experimental environment

The server operating system used for the experiments is Ubuntu 20.04.5 LTS. The computing resources are a single NVIDIA GeForce RTX 4090 (24G) graphics card. The compiler platform is PyCharm, with Python 3.8.10, PyTorch 2.0.0, and CUDA 11.8 as the deep learning framework.

4.3 Ablation experiment

An ablation study was performed on the Cotton Disease Dataset to assess the effectiveness of the SDRM-YOLO model introduced in this paper for detecting cotton pests and diseases.

First, to evaluate the effectiveness of the proposed MCPA, it was benchmarked against SE, CBAM, and CA attention mechanisms. The results of these experiments are presented in Table 1. The MCPA module demonstrated a notably higher mAP50 index compared to the other attention modules.

Table 1. Ablation experiments of MCPA module

	P	R	mAP50	GFLOPs	Params
SE	76.1	73.7	79.1	7.9	3.01
CBAM	82.3	72.5	79.8	8.2	3.02
CA	83.2	70.9	79.5	7.9	3.02
MCPA	71.0	75.2	80.1	8.1	3.16

Then, Initially, the DCNv2 and the C2f module in the neck network were integrated, resulting in a 0.5% increase in the model's mAP50 index. Subsequently, incorporating the RCSOSA attention mechanism further enhanced the module's performance. Ultimately, compared to the original model, the mAP50 index was improved by 1.1%. The detailed experimental results are presented in Table 2.

Table 2. Ablation experiment of C2f-DR module

Neck	P	R	mAP50	GFLOPs	Params
YOLOv8n	77.5	73.7	79.1	8.1	3.00
C2f-DCNv2	76.6	74.1	79.6	8.2	3.06
C2f-DCN-RCSOSA	74.0	75.5	80.2	8.5	3.49

Finally, the backbone network part is improved by introducing the ShuffleNetv2 module, which utilizes the channel rearrangement mechanism and the depth-separable convolution technique to enhance the model detection accuracy, resulting in a 1.1% increase in its mAP50 metric. Later, residual connection and new activation function Swish is introduced to optimize the design, and it is verified that the improved ShuffleNetv2-RC module improves the mAP50 by 0.5% compared with the original ShuffleNetv2 module. The experimental results are shown in Table 3 below.

Table 3. Ablation experiment of ShuffleNetv2-RC module

Backbone	P	R	mAP50	GFLOPs	Params
YOLOv8n	77.5	73.7	79.1	8.1	3.00
Shufflenetv2	82.8	69.3	80.2	8.1	2.79
ShuffleNetv2-RC	80.8	74.5	80.7	7.8	2.79

After independently performing ablation tests on each design module individually, we plan to incorporate these newly designed modules into the model framework to assess their contribution to the overall model performance. Using YOLOv8n as the baseline model, we verified the effect of the modules on the model performance by adding the MCPA, C2f-DR, and ShuffleNetv2-RC modules, respectively, and kept all hyperparameters consistent in the experiments, the results of which are shown in Table 4 below. First, the innovative MCPA module was introduced between the neck and the prediction head, demonstrating excellent efficiency in multi-scale disease feature fusion and achieving a 5% improvement in mAP50 performance. Then, we incorporated the C2f-DR module in the network's neck (Neck) region. This improved the model to more accurately recognize the contours and boundaries of complex targets, resulting in a 1.1% improvement in mAP50 performance. Finally, we enabled the mAP50 to achieve the highest accuracy of 86.4% by integrating the optimized ShuffleNetv2-RC module into the backbone network. Compared to the baseline model, SDRM-YOLO achieves a 7.3% performance improvement in the key metric of mAP50 and minimizes the quantity of model parameters and computational intensity.

Table 4. Ablation experiments of SDRM-YOLO model

ShuffleNetv2-RC	C2f-DR	MCPA	P	R	mAP50	GFLOPs	Params
			77.5	73.7	79.1	8.1	3.00
√			80.8	74.5	80.7	7.8	2.79
	√		74.0	75.5	80.2	8.5	3.49
		√	71.0	75.2	80.1	8.1	3.16
√		√	82.5	76.1	83.8	7.9	2.98
√	√		81.2	76.5	82.3	8.1	3.28
	√	√	82.6	76.8	84.1	8.4	3.62
√	√	√	85.2	78.7	86.4	7.8	2.92

4.4 Comparative experiment

To fully evaluate the performance of the SDRM-YOLO model proposed in this study, we conducted extensive comparative experiments on two different datasets. In these experiments, we selected a variety of state-of-the-art including SSD, Faster R-CNN, YOLOv5n, YOLOv6n, YOLOv7-tiny, Efficient [32], YOLOv8n (as a benchmark model), DETR [33], YOLOv10n [34], and YOLOv11n [35] target detection models are analyzed competitively. The experimental outcomes presented in Table 5 demonstrate that SDRM-YOLO exhibits superior performance on the Cotton Disease Dataset compared to other models included in the comparison across several critical evaluation metrics. Particularly in the three core assessment indicators of precision (P), recall (R), and mAP50, the SDRM-YOLO model's performance outstrips the latest state-of-the-art YOLOv11 series models. This finding fully validates the outstanding capabilities of SDRM-YOLO in object detection. Fig. 5 illustrates the detection results of the YOLOv8n and SDRM-YOLO models when applied to six different pests within the Cotton Disease Dataset. As demonstrated by the experimental data, the improved model possesses an excellent ability to recognize partially occluded targets, thus significantly reducing the number of undetected category samples.

Table 5. Comparative experiments on Cotton Disease Dataset

Model	P	R	mAP50	GFLOPs	Params
SSD	80.1	69.2	70.5	62.8	27.12
Faster R-CNN	81.8	71.4	70.9	136.2	40.35
YOLOv5n	82.4	76.8	72.2	4.5	2.02
YOLOv6n	81.9	73.6	73.2	9.1	4.24
YOLOv7-tiny	78.0	67.3	72.7	3.2	6.07
EfficientDet	80.4	76.6	83.8	5.1	11.10
YOLOv8n	77.5	73.7	79.1	8.1	3.00
DETR	81.1	76.8	82.1	98.6	8.15
YOLOv10n	82.3	77.0	82.8	6.8	2.73
YOLOv11n	83.7	77.3	84.9	7.2	9.42
SDRM-YOLO	85.2	78.7	86.4	7.8	2.92

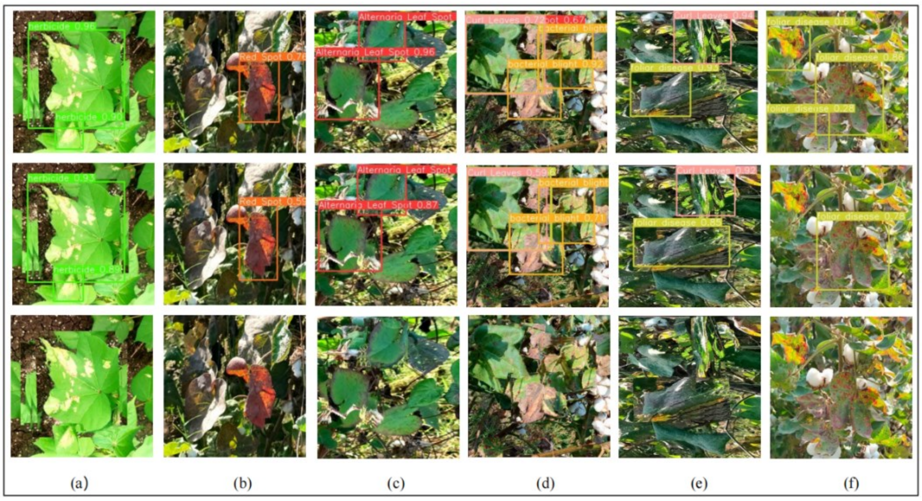

Fig. 5. Visual test results on the Cotton Disease Dataset， the rows from bottom to top represent the SDRM-YOLO detection result graph, YOLOv8n detection result graph, and original graph, respectively. Results of disease detection for each category: （a)：herbicide (b):Red Spot (c): Alternaria Leaf Spot (d):bacterial blight (e): Curl Leaves (f):foliar disease

Table 6. Comparative experiments on Cotton Pest Detect Dataset

Model	P	R	mAP50	GFLOPs	Params
SSD	80.2	65.3	71.5	62.8	27.12
Faster R-CNN	82.6	65.9	73.0	136.2	40.35
YOLOv5n	84.5	66.3	74.1	4.5	2.02
YOLOv6n	76.8	66.4	75.2	9.1	4.24
YOLOv7-tiny	78.0	67.3	74.7	3.2	6.07
EfficientDet	82.1	68.7	76.5	5.1	11.10
YOLOv8n	83.6	69.3	75.5	8.1	3.00
DETR	81.6	68.6	75.9	98.6	8.15
YOLOv10n	83.4	69.6	76.1	6.8	2.73
YOLOv11n	84.4	70.0	76.6	7.2	9.42
SDRM-YOLO	86.7	71.7	78.0	7.8	2.92

To further examine the generalization performance of the SDRM-YOLO model, we implemented additional comparison experiments on the Cotton Pest Detect Dataset. The results of the experiments are detailed in Table 6 below.6 In the comparison of SDRM-YOLO with the models SSD, Faster R-CNN, YOLOv5n, YOLOv6n, YOLOv7-tiny, Efficient, YOLOv8n (as the benchmark model), DETR, YOLOv10n, and YOLOv11n, the key metric of mAP50 metric achieves significant improvements of 6.5%, 5%, 3.9%, 2.8%, 3.3%, 1.5%, 2.5%, 2.1%, 1.9% and 1.4%, respectively. In addition, SDRM-YOLO demonstrated significant enhancement in three other evaluation metrics. These results further confirm the ability of SDRM-YOLO to generalize on different datasets. Fig. 6 presents the results of the visualization and analysis of YOLOv8n with the enhanced SDRM-YOLO model on the Cotton Pest Detect Dataset for six categories of pests and diseases. As can be seen from the figure, the improved model shows significant recognition accuracy advantages, surpassing the original model's performance.

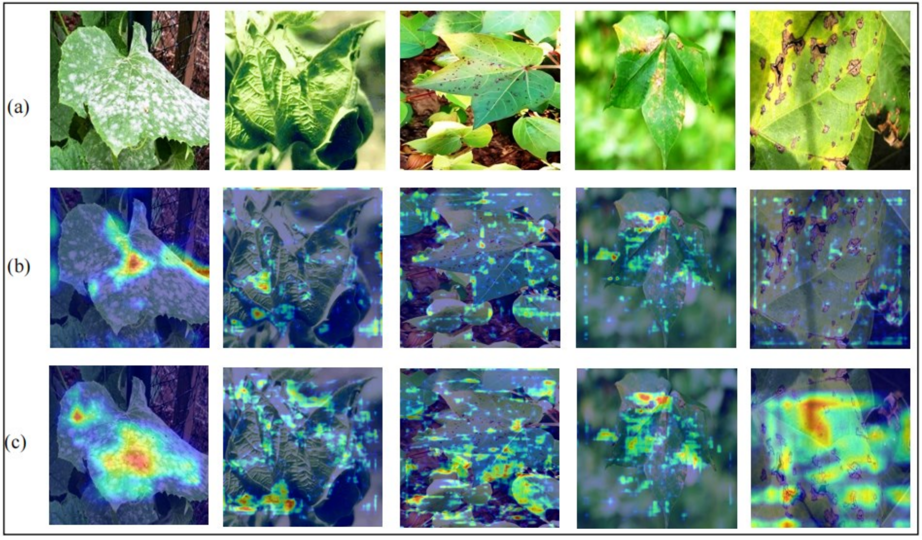

Fig. 6. Comparison chart of thermal effects on Cotton Pest Detect Dataset, where (a) is the original image, (b) is the YOLOv8n detection thermal map, and (c) is the SDRM-YOLO detection thermal map

5 Conclusions

In this study, we propose an improved SDRM-YOLO model to increase cotton pest and disease detection accuracy in complex contexts. We achieve significant performance improvements by optimizing the YOLOv8n model in multiple ways. Specifically, first, we integrated an innovative multi-scale cross-spatial perceptual attention mechanism (MCPA) between the neck network and the prediction head of the model, which utilizes three different-sized convolutions to strengthen the channel attention, thus improving the model's ability to capture multi-scale features and the accuracy of spatial localization. Next, we introduce the C2f-DR module to replace the original C2f layer, and this optimization improves the feature extraction efficiency of the model when dealing with targets with significant variations in shape and location and enhances attention to critical regions. Finally, we adopt the optimized ShuffleNetv2-RC as the backbone network, which not only reduces the model parameters but also effectively mitigates the phenomenon of gradient vanishing and improves the training efficiency and performance of the model. With the above improvements, the new model has more obvious performance improvement on both the Cotton Disease Dataset and Cotton Pest Detect Dataset, and the model complexity is reduced compared to the benchmark model. However, the C2f-DR module in this study is characterized by a high parameter count. Future research endeavors will concentrate on developing target detection algorithms that aim to reduce the module's parameter quantity while preserving the model's detection accuracy.

Acknowledgments. This work was funded by National Natural Science Foundation Project (Grant number: 62266043), High resolu tion Earth observation system of the State Administration of Science, Technology and Industry for National Defense (Grant numbers: 95-Y50G37-9001-22/23), and Tianshan Innovation Team (Grant numbers: 2023D14012).

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Manavalan R. Towards an intelligent approaches for cotton diseases detection: A review[J]. Computers and Electronics in Agriculture, 2022, 200: 107255.
[2] John M A, Bankole I, Ajayi-Moses O, et al. Relevance of advanced plant disease detection techniques in disease and Pest Management for Ensuring Food Security and Their Implication: A review[J]. American Journal of Plant Sciences, 2023, 14(11): 1260-1295.
[3] Li R, He Y, Li Y, et al. Identification of cotton pest and disease based on CFNet-VoV-GCSP-LSKNet-YOLOv8s: a new era of precision agriculture[J]. Frontiers in Plant Science, 2024, 15: 1348402.
[4] Khan S, Yairi T. A review on the application of deep learning in system health management[J]. Mechanical systems and signal processing, 2018, 107: 241-265.
[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[6] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[7] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
[8] He Y, Zhu C, Wang J, et al. Bounding box regression with uncertainty for accurate object detection[C]//Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2019: 2888-2897.
[9] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
[10] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[12] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[13] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arxiv preprint arxiv:1804.02767, 2018.
[14] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arxiv preprint arxiv:2004.10934, 2020.
[15] Wang Z, Hua Z, Wen Y, et al. E-YOLO: Recognition of estrus cow based on improved YOLOv8n model[J]. Expert Systems with Applications, 2024, 238: 122212.
[16] Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11963-11975.
[17] Peng J, Lv K, Wang G, et al. MLSA-YOLO: a multi-level feature fusion and scale-adaptive framework for small object detection[J]. The Journal of Supercomputing, 2025, 81(4): 528.
[18] Zhou H, Chen J, Niu X, et al. Identification of leaf diseases in field crops based on improved ShuffleNetV2[J]. Frontiers in Plant Science, 2024, 15: 1342123.
[19] Hu X, Chen J, Chen Y. RegMamba: An Improved Mamba for Medical Image Registration[J]. Electronics (2079-9292), 2024, 13(16).
[20] Huang M, Wei H, Zhai X. A Study on Enhancing Chip Detection Efficiency Using the Lightweight Van-YOLOv8 Network[J]. Computers, Materials & Continua, 2024, 79(1).
[21] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.
[22] Chen H, Tao J. Utilizing improved YOLOv8 based on SPD-BRSA-AFPN for ultrasonic phased array non-destructive testing[J]. Ultrasonics, 2024, 142: 107382.
[23] Yang G, Lei J, Tian H, et al. Asymptotic feature pyramid network for labeling pixels and regions[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024.
[24] Vasanthi P, Mohan L. Ensemble of ghost convolution block with nested transformer encoder for dense object recognition[J]. Biomedical Signal Processing and Control, 2024, 88: 105645.
[25] Li L, Wang Z, Zhang T. Gbh-yolov5: Ghost convolution with bottleneckcsp and tiny target prediction head incorporating yolov5 for pv panel defect detection[J]. Electronics, 2023, 12(3): 561.
[26] Qi J, Liu X, Liu K, et al. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease[J]. Computers and electronics in agriculture, 2022, 194: 106780.
[27] Li P, Zhou J, Sun H, et al. RDRM-YOLO: A High-Accuracy and Lightweight Rice Disease Detection Model for Complex Field Environments Based on Improved YOLOv5[J]. Agriculture, 2025, 15(5): 479.
[28] Bao L, Yang Z, Wang S, et al. Real image denoising based on multi-scale residual dense block and cascaded U-Net with block-connection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 448-449.
[29] Zhai X, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre-training[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 11975-11986.
[30] Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9308-9316.
[31] Zhang X, Shen B, Li J, et al. Lightweight Algorithm for Rail Fastener Status Detection Based on YOLOv8n[J]. Electronics, 2024, 13(17): 3399.

[32] Sze V, Chen Y H, Yang T J, et al. Efficient processing of deep neural networks: A tutorial and survey[J]. Proceedings of the IEEE, 2017, 105(12): 2295-2329.

[33] Li F, Zhang H, Liu S, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13619-13627.

[34] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[35] Jegham N, Koh C Y, Abdelatti M, et al. Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors[J]. arxiv preprint arxiv:2411.00201, 2024.