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

SEBF-YOLO: An Improved YOLOv8s for Small Insect Detection

Lai Jiang[1], Rui Xiong[1], and Zhiwu Liao[1*]

[1] School of Computer Science, Sichuan Normal University, Chengdu 610068, Sichuan Province, China
*Corresponding author: 20060097@sicnu.edu.cn

Abstract. To address the low detection accuracy of small insect with blurred features and complex backgrounds in agricultural scenarios, we propose an improved YOLOv8 (You Only Look Once version 8) model, SEBF-YOLO, to tackle the shortcomings of insufficient feature extraction and fusion in the original YOLOv8 for small insect detection. Given that small insects with low pixel occupancy in spatial domains often suffer from feature loss during extraction, a Space-to-Depth(SPD) module is introduced after each convolutional layer in the backbone network to enhance the extraction of fine-grained features for small targets. For the challenges of complex backgrounds and feature blurriness—rooted in the model's inability to distinguish backgrounds and lack of effective multi-scale feature fusion—the C2f_EMA module is added after concatenation layers in the neck network, establishing bidirectional cross-scale connections and adopting a weighted fusion strategy to strengthen critical features of blurred targets by integrating multi-level features. Subsequently, the BiFormer module is introduced after C2f_EMA to leverage dynamic attention mechanisms for weighted focusing on fused feature maps, integrating local details and global contextual information to suppress background interference and enhance target discrimination in complex scenes. Experimental results on a self-built dataset demonstrate that SEBF-YOLO achieves a mean Average Precision (mAP) of 77.3% at an Intersection over Union (IoU) of 0.5, a 4.1% improvement over the original model, providing an effective solution for detecting small insect targets in agricultural environments.

Keywords： YOLOv8, small target detection, attention mechanism[1]

1 Introduction

With the rapid growth of the global population and the intensification of climate change, agricultural production is facing unprecedented challenges, and pest control has become a key issue in ensuring food security and improving agricultural production efficiency. In the complex ecological environment of the fields, it is particularly important to accurately identify and promptly control pests in farmland. Moreover, pest control strategies have been continuously evolving with the development of society and the progress of science and technology [1]. In the past, traditional manual visual inspection methods relied on the visual inspection of experts [2], which had the problems

of low efficiency and high cost, and it was difficult to meet the needs of image blur changes and small target detection in the complex field environment. In recent years, deep learning has developed rapidly [3], and an increasing number of scientific and technical practitioners have begun to pay attention to the application of computers in the agricultural field [4], in order to significantly improve the accuracy and efficiency of pest detection, and further achieve early warning and precise control.

However, field insect detection still faces numerous challenges. Firstly, insect targets are small in size and occupy a low proportion of pixels. Traditional convolutional downsampling is prone to causing feature loss, resulting in an increased rate of missed detections. Secondly, under natural field conditions, pest images encounter challenges such as similarity among species, different scales, pose variations, lighting effects, and occlusion [5]. These factors will further reduce the accuracy of insect detection.

In order to solve the above problems, a target detection model for field insects (SEBF-YOLO) based on the YOLOv8s model is proposed. By introducing the SPD module to reconstruct the convolutional layer, the detection accuracy for small targets is improved; the BiFormer module is introduced to improve the network structure, enhancing the image detection ability in complex environments; the C2f layer is modified to C2f_EMA, which not only improves the efficiency of the model but also further enhances the detection ability for small targets.

The main contributions of this study are as follows:

Proposed Detection Model. A target detection model (SEBF - YOLO) for field insects is proposed, which is based on the YOLOv8s model. To address the problems that the original model struggles to solve, improvement plans are designed from two aspects: feature extraction and feature fusion.

Feature extraction optimization. The SPD module is introduced after each convolutional layer to solve the problem of small - target feature loss caused by traditional convolutional downsampling, thereby retaining more detailed features of insects.

Feature fusion enhancement. The C2f layer after the feature concatenation layer in the head network is replaced with the C2f_EMA layer to solve the lack of a multi - scale fusion mechanism in the original module. The representational ability of blurred insect features is enhanced through a cross - space fusion mechanism. The BiFormer module is introduced after the EMA module, which dynamically focuses on the insect target area. Through local - global feature interaction, it solves the problem that the original model fails to effectively suppress the interference of complex background noise.

Dataset construction. A labeled dataset containing five common types of farmland insects (ants, aphids, corn borers, ladybugs, and spiders) is constructed, covering complex backgrounds such as lighting variations and occlusion.

Section arrangement. Section 2 elaborates on the research progress of target detection models, agricultural insect detection methods, and attention mechanisms. Section 3 expounds on the original model and the improvement methods for it. Section 4 introduces the situation of the self-built dataset and the evaluation indicators. Section 5 presents the experimental results and analysis, including the comparison with the original model, the comparison with the latest different target detection models, ablation experiments, and visualization results. Section 6 summarizes the effectiveness of the model and proposes future research directions.

2 Related Work

2.1 Target Detection Models

In the field of target detection, the mainstream detection models are mainly divided into two categories: two-stage models and one-stage models. The R-CNN series [6] belongs to the classic two-stage algorithms. It generates candidate boxes through a region proposal network. Although it has high detection accuracy, its computational complexity limits its real-time performance. On the other hand, algorithms such as SSD [7] and the YOLO series [8] represent the development of one-stage algorithms. They directly predict the target positions through end-to-end regression, achieving a balance between speed and accuracy. The YOLO series has been continuously iterated and optimized, but in complex field scenarios, it still faces the problems of missed detections of small targets and false detections caused by blurred images.

2.2 Agricultural Insect Detection

With the development of intelligent information technology and deep learning, the research on the recognition of crop insects using Convolutional Neural Network (CNN) and YOLO algorithm has been continuously deepened [9]. For example, Zheng Guo et al. [10] proposed introducing convolutional block attention and feature pyramid into the YOLOv7 algorithm to solve the problem that it is difficult to recognize small-sized harmful rice insects. Gao Jiajun et al. [11] proposed an image segmentation method integrating Swin Transformer to address the problem of difficult recognition and segmentation of multi-larval individual images in complex scenarios. Aiming at the problems of low detection efficiency and poor reliability of cotton diseases and pests, Zhang et al. [12] proposed an algorithm based on the YOLOX network model, which improves the detection accuracy of cotton diseases and pests by introducing an efficient attention mechanism (ECA) and the hard-Swish activation function. Jiao et al. [13] used a deformable residual network to extract the features of pests and adopted a global context-aware module to obtain the regions of interest of pests, achieving good results. However, due to certain difficulties in data collection, the limited types of insects studied, and factors such as blurred images and low pixel values of small targets when both large and small targets are present, for some insects, there are certain challenges in both classification and detection [14].

2.3 Attention Mechanism

The attention mechanism enhances the robustness of the model by focusing on key feature regions. Traditional methods such as SE (Squeeze-and-Excitation) [15] and ECA (Efficient Channel Attention) [16] optimize the feature response through channel weighting, but they lack dynamic adjustment in the spatial dimension. BiFormer proposes a dynamic sparse attention mechanism, which captures long-range dependencies through local window calculation. While reducing the computational load, it improves the target detection performance in complex backgrounds. In this study, by combining the EMA and BiFormer modules, the collaborative optimization of multi-scale and sparse attention is achieved for the first time in insect detection.

3 The YOLOv8 Algorithm and Its Improvements

3.1 Network Model

YOLOv8 is a one-stage object detection model. Based on the depth and width of the network, it offers five models of different sizes: n, s, m, l, and x. The model size increases sequentially, and so does the detection accuracy [17]. Among these five models, YOLOv8s is well-known for its simplicity, high accuracy, and low resource consumption, making it highly suitable as a base model for insect detection tasks. The main structure of YOLOv8s consists of an input end (Input), a backbone network (Backbone), a neck network (Neck), and a head network (Head) [18]. Compared with previous versions of YOLO, YOLOv8 introduces a new C2f module, a new loss function, and improves the feature fusion architecture to enhance object detection accuracy while achieving further lightweight design.

However, there will still be problems such as insufficient feature extraction ability for small-target insects and inefficient multi-scale fusion caused by blurred features. Aiming at the above problems, an insect detection algorithm named SEBF-YOLO based on YOLOv8s is proposed. The improvement is carried out from two aspects: optimization of feature extraction and enhancement of feature fusion. The SEBF-YOLO network model is shown in Figure 1.

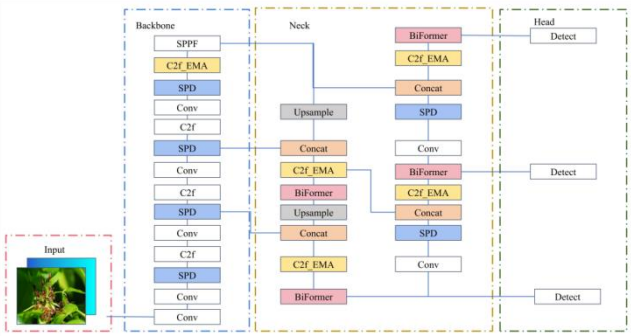

Fig. 1. Structural Diagram of the SEBF-YOLO Model

3.2 Feature extraction optimization

SPD module. In the field of image detection, the YOLO network itself often has low detection accuracy when dealing with low - resolution images or small targets. This is because during the sampling process, it is easy to lose edge features and reduce the texture information of small - target insects. For example, after four downsampling operations in the backbone network of YOLOv8, the resolution of the input image decreases from 640×640 to 20×20, and the pixel information of small targets decays significantly. We considers using the SPD convolution module to optimize the feature extraction ability. The SPD module can retain all information when downsampling the feature maps, thus avoiding the problem of fine - grained information loss caused by traditional convolution and pooling operations.

In the improvement of the YOLOv8 model, the SPD module is chosen to be inserted after the Conv convolutional layer in sequence. The purpose of this is to use the output feature map of the previous convolutional layer as the input of the SPD layer. After transformation, it is then convolved through the subsequent Conv layer, and the number of channels in the output feature map remains unchanged. This insertion method can reduce the spatial dimension without losing information. Compared with traditional convolution operations, it performs better in retaining information within the channels. Therefore, it can significantly improve the model's ability to extract features of small targets.

The SPD module is composed of a space-to-depth layer and a non-strided convolutional layer [19]. During the downsampling process of the feature map, the SPD layer rescales the original image and retains all the information in the channel dimension, thereby reducing the loss of detailed information and enhancing the ability to learn less prominent features.

Specifically, for a given feature map X, downsampling is carried out according to the scaling factor. For example, when the scaling factor is 2, 4 sub-feature maps will be generated, and the shape of each sub-map is (S/2,S/2,C), and the size of X is reduced by a factor of two. Then, these 4 sub-feature maps are concatenated along the channel dimension to form a new intermediate feature map (S/2,S/2,4C). After the SPD layer completes the feature transformation, through a non-strided convolutional layer with D filters (where D is less than 4C), the intermediate feature map is transformed into (S/2,S/2,D). Among them, for the input feature map $X \in \square^{S\times S\times C}$, when the scaling factor is 2, the space-to-depth transformation formula is as shown in Formula 1:

$$\begin{aligned} f_{i,j} &= X[i::scale, j::scale] \\ X' &= concat(f_{0,0}, f_{1,0}, f_{0,1}, f_{1,1}) \in R^{S/2\times S/2\times 4C} \end{aligned} \tag{1}$$

As shown in Formula 1, the original feature map is sampled with a stride of S (scale) to generate four sub-feature maps. Then, all spatial information is retained through dimensional concatenation, avoiding the loss of information during downsampling, as shown in part (a) of Figure 2.

And the formula of the subsequent non-strided convolutional layer is as shown in Formula 2:

$$Y = W * X' + b$$
$$Y \in R^{S/2 \times S/2 \times D}, D < 4C \tag{2}$$

As shown in Formula 2, through a common convolution with a stride of 1, the number of channels is compressed from 4C to D, reducing redundant features while retaining key information. The weight matrix W dynamically adjusts the channel weights during the training process, suppressing noise and enhancing features related to the target. This operation achieves efficient feature recombination and dimensionality reduction while maintaining the spatial resolution.

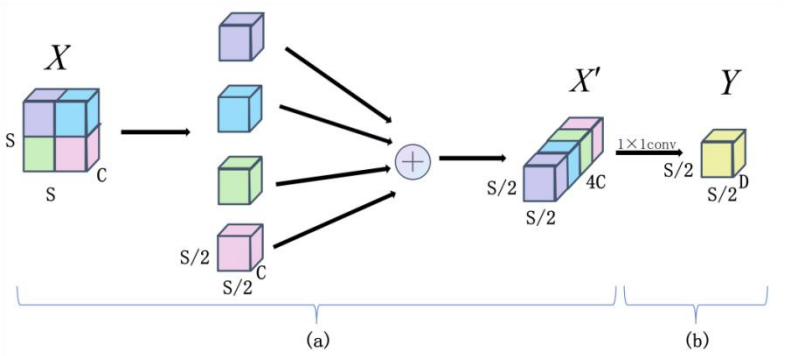

Fig. 2. The Downsampling Process of the SPD Module

3.3 Feature fusion enhancement

C2f_EMA module. In the original C2f layer, the fixed weight allocation struggles to adapt to the multi - scale characteristics of insects. Meanwhile, it is difficult to solve the problem of low detection accuracy caused by feature blurring. To address the above issues, C2f_EMA combined with efficient multi - scale attention is introduced after the splicing layer in the head network. By embedding the efficient multi - scale attention into the cross - stage connection, dynamic weight allocation and cross - spatial feature interaction are achieved. The structural diagram of the C2f_EMA module is shown in Figure 3.

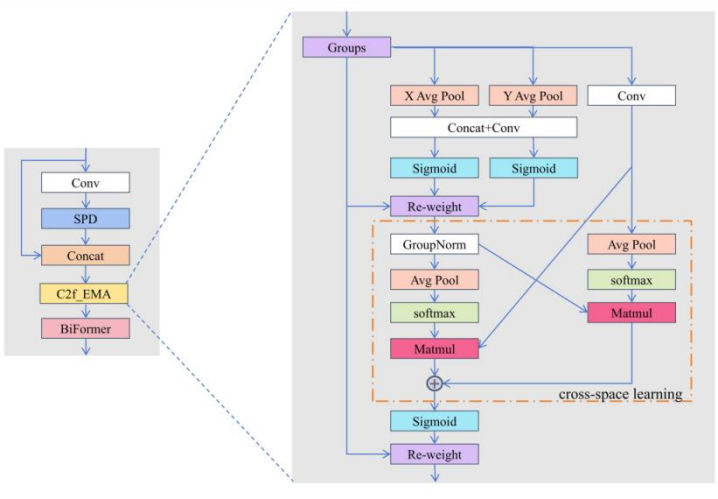

Fig. 3. The Structural Diagram of EMA

The EMA [20] module avoids more sequential processing through parallel sub - structures, reducing the network depth while retaining information in each channel and lowering the computational cost.

In addition, the module also employs global average pooling to encode the global information output from each branch and performs a linear transformation using the Softmax function. Then, it uses matrix dot - product operations to multiply the results of parallel processing, generating a spatial attention map. Finally, the two sets of output features are mapped and added together, and the output with the same size as the input is obtained through the Sigmoid function and multiplication operations.

During the operation process of EMA:

$$z_c^h(h) = \frac{1}{W}\sum_{i=1}^{W} X_c(h, j)$$
$$z_c^w(w) = \frac{1}{H}\sum_{j=1}^{H} X_c(j, w) \tag{3}$$

Formula 3 represents the features after pooling in the horizontal and vertical directions, with dimensions of [B, C, H] and [B, C, W] respectively. Position - sensitive channel feature vectors are generated through one - dimensional global average pooling along the width (W) and height (H) directions. Here, B represents the batch size, and C represents the number of channels in the feature map.

$$A_{channel} = \sigma(Conv_{1\times1}(Concat(z^h, z^w))) \tag{4}$$

Formula 4 is for the generation of channel attention, which is achieved through convolution and aggregation after horizontal and vertical pooling. Among them, $A_{channel}$ is the channel attention weight, with the dimension of [B, C, H, W]. σ represents the

Sigmoid activation function, which normalizes the weight to the range of [0, 1]. $A_{channel}$ generates the channel attention weight through 1×1 convolution and the Sigmoid activation, which is used to dynamically enhance the key channels.

$$F_{3\times3} = Conv_{3\times3}(X_i) \tag{5}$$

Formula 5 is the Conv convolution of the right branch in Figure 3 (the 3×3 branch). It extracts local multi-scale features through 3×3 depthwise separable convolution, expanding the receptive field to capture the details of small targets.

$$\begin{aligned} g_{1\times1} &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} A_{channel}(h,w) \\ g_{3\times3} &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{3\times3}(h,w) \end{aligned} \tag{6}$$

Formula 6 performs global average pooling on the channel attention and multi-scale features to generate a global context vector.

$$\begin{aligned} M_1 &= \mathrm{Softmax}(g_{1\times1}^{\bullet} \cdot F_{3\times3}) \\ M_2 &= \mathrm{Softmax}(F_{1\times1}^{\bullet} \cdot g_{3\times3}) \\ Y &= X \otimes \sigma(M_1 + M_2) \end{aligned} \tag{7}$$

Formula 7 generates a cross - spatial attention map through bidirectional matrix multiplication and Softmax normalization to dynamically fuse channel and spatial information. $\bullet$ represents transpose, and $\otimes$ represents matrix multiplication.

The EMA module achieves information aggregation across multiple spatial dimensions through a cross-space learning mechanism. At the same time, it encodes cross-channel correlations and preserves the details of the spatial structure. This design dynamically assigns feature weights based on the input content. By incurring lightweight computational costs, it enhances the model's ability to focus on key regions, thereby significantly improving the positioning accuracy and generalization performance for small-sized insect targets. It demonstrates unique advantages in fine-grained detection tasks under complex backgrounds.

BiFormer module. In reality, photographing insects is often restricted by equipment and weather conditions. The rapid movement of insects and lighting issues can lead to the loss of texture details. At the same time, noise such as complex backgrounds may be misdetected as small targets. To address the above problems, the BiFormer module is introduced after the EMA module. After fusing multi-scale features, it suppresses background noise through dynamically focusing on regions.

The Bi-Level Routing Attention (BRA) [21] of the BiFormer module can identify the contextual features in images through dynamic sparse computation and region-level semantic filtering, thus improving the detection performance of small targets. The structural diagram of the BiFormer is shown in Figure 4(b), which is composed of an

overlapping patch module and N connected BiFormer Block modules. The structural diagram of the BiFormer Block module is shown in Figure 4(a), which is composed of depthwise separable convolution, layer normalization, the bi-level routing attention mechanism, and the multi-layer perception mechanism through residual operations.

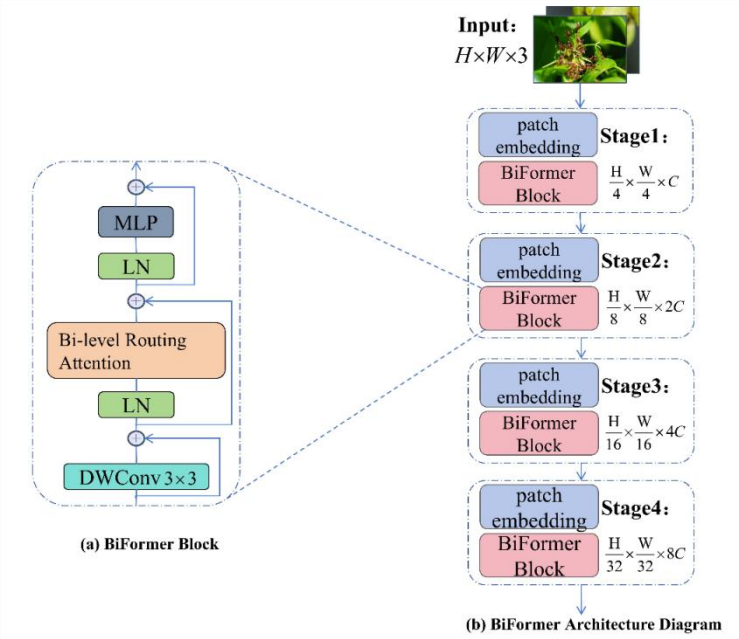

Fig. 4. Structural Diagram of BiFormer Block and BiFormer

The BiFormer Block module can more effectively enhance the perception ability in complex backgrounds and blurred situations of insects through the BRA bi-level routing attention mechanism. The principle is shown in Figure 5. The experimental process mainly includes the following three steps:

Region Partitioning and Projection. Given the input feature map $X \in \mathrm{R}^{H\times W\times C}$, the image is first divided into S×S regions. For example, the yellow shaded image blocks in Figure 5 are the divided region blocks, and each region occupies $\frac{HW}{S^2}$ spatial positions (Tokens).

$$X^{\cdot} = \text{Patchify}(X) \in \mathrm{R}^{S^2\times\frac{HW}{S^2}\times C} \tag{8}$$

The query vector Q, key K, and value tensor V are obtained through linear projection.

$$Q = X^{\text{晦}}W^{q}, K = X\ \ W^{k}, V = X^{?}W^{v} \tag{9}$$

Among them, W^q, W^k, W^v represents the projection weights of the query, key, and value respectively.

Regional-level Routing. By constructing a directed graph, we can find out the regions that each given region should participate in. Calculate the regional relevance matrix and perform Top-k screening.

$$\begin{aligned} A^{\cdot} &= Q^{\text{睫}}(K\ )^{?} \\ I^{\text{睫}} &= TopkInex(A\ ) \end{aligned} \tag{10}$$

Among them, $Q^{\text{睫}}, K$ is the average value of the regional query and key vectors. $A^{\cdot}$ is the adjacency matrix of the correlations between regions obtained from the average value. $I^{\cdot} \in \mathrm{N}^{S^2 \times k}$ represents the index matrix generated by the top k relevant regions that each region focuses on.

Token-level Attention.

$$K^g = gather(K, I^{\text{睫}}), V^g = gather(V, I\ ) \tag{11}$$

$$Attention(Q, K^g, V^g) = soft\max(\frac{Q(K^g)^{\cdot}}{\sqrt{C}})V^g \tag{12}$$

$$BRA(Q, K, V) = Attention(Q, K^g, V^g) + DWConv(V) \tag{13}$$

The Gather operation in Formula 11 is to aggregate the key and value vectors of the selected k regional blocks. Formula 12 calculates the attention weights between spatial positions in the selected regions to focus on details (such as insect limbs). In Formula 13, DWconv represents depthwise separable convolution. Depthwise separable convolution compensates for the local feature information that might be lost in sparse attention. Finally, the attention mechanism and depthwise separable convolution are added together to obtain the final bi - level routing attention.

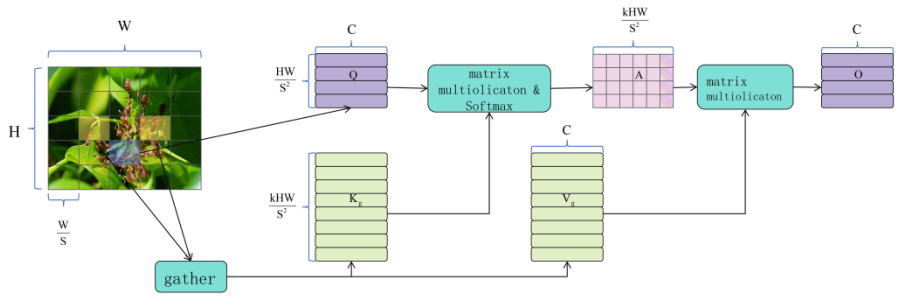

Fig. 5. Structural Diagram of the BRA Module

4 Experimental Data and Settings

4.1 Experimental Environment and Data Sources

This study is based on the Windows 10 system, which is equipped with an NVIDIA GeForce RTX 4060 Ti GPU and an Intel i5 - 12490F CPU. It is implemented using the Python 3.9 programming language on the CUDA 11.3 and PyTorch 1.10.0 deep learning frameworks. The size of the input images is 640×640. The number of training iterations is 100, the initial learning rate is 0.01, the SGD algorithm is used as the optimizer, the weight decay is set to 0.0005, the momentum factor is 0.937, and the Batch Size is 16.

In the experiment, an insect dataset was used, which combined the data collected through on - site shooting at the base of the Sichuan Academy of Agricultural Sciences and publicly available datasets on the Internet. Image data was collected at different times and in different environments. After data augmentation, there were approximately 13,200 images, including those of ants, ladybugs, corn borers, aphids, and spiders. The dataset was randomly divided into a training set, a validation set, and a test set at a ratio of 8:1:1.

Fig. 6. Dataset pictures. Among them, a is an ant, b is a ladybug, c is a corn borer, d is an aphid, and e is a spider. Here, relatively clear pictures are selected to make the dataset pictures visible.

4.2 Evaluation Indicators

In this study, the following indicators are used to evaluate the performance of the model: Precision, Recall, Average Precision (AP), and Mean Average Precision (mAP). The calculation methods of these indicators are as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\, dR \tag{16}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{C}} \sum_{i=1}^{C} \mathrm{AP}_i \tag{17}$$

In Formula 14, TP represents the number of positive samples that are correctly predicted, and FP represents the number of negative samples that are incorrectly predicted as positive. In Formula 15, FN represents the number of positive samples that are incorrectly predicted as negative. Formula 16 represents the average precision at different recall levels. In Formula 17, C represents the total number of categories, and mAP is the average of the average precisions of all categories.

When calculating mAP, different IoU thresholds are taken into account, including $mAP_{0.5}$ and $mAP_{0.5:0.95}$. $mAP_{0.5}$ is the mAP calculated at an IoU threshold of 0.5, while $mAP_{0.5:0.95}$ is the average of the mAP values calculated at IoU thresholds ranging from 0.5 to 0.95. These metrics offer an evaluation of the model's performance at different IoU thresholds.

5 Experimental Results and Analysis

5.1 Comparison between the model before and after improvement

The improved model was compared with the original YOLOv8s model in terms of precision, recall, $mAP_{0.5}$, and $mAP_{0.5:0.95}$. According to the data in Table 1, the SEBF - YOLO model outperforms YOLOv8s in multiple aspects: the precision is increased by 7%, the recall is improved by 2.4%, and $mAP_{0.5}$ is significantly increased by 4.1%. There is also a slight improvement in $mAP_{0.5:0.95}$ compared with the original model. In conclusion, the SEBF - YOLO model has better performance.

Table 1. Comparison of Performance Before and After Improvement

Models	P/%	R/%	$mAP_{0.5}$ /%	$mAP_{0.5:0.95}$/%
YOLOv8s	79.2	68	73.2	44.6
SEBF-YOLO	**86.2**	**70.4**	**77.3**	**45.4**

At the same time, the detections of various types of insects by the models before and after the improvement were also compared, as shown in Figure 7.

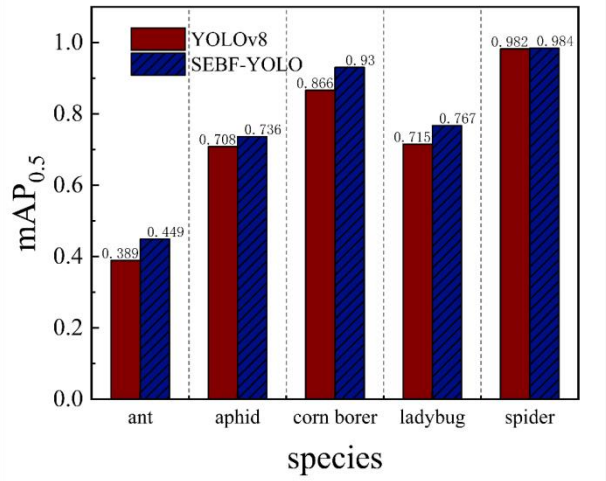

Fig. 7. Bar chart of the comparison of various categories before and after the improvement

As can be seen from the table and the comparison chart, compared with the original model, the SEBF-YOLO model has improved in all aspects. In the detection of various types of insects, both ants and aphids are densely populated and small targets. Meanwhile, SPDconv combines the dynamic attention of BiFormer while avoiding the loss of details (such as the details of antennae) in traditional downsampling, suppressing the complex background noise, and directly improving the performance of object detection. For the corn borer, ladybug, and spider, they are not overly dense in space and often exist as single targets in the image. However, there may be problems of image blurriness. The EMA fuses multi-scale features to enhance the robustness. Combined with the retention of details and the suppression of complex backgrounds, the positioning accuracy of the targets has been improved, which verifies the effectiveness of the model improvement.

5.2 Comparison Experiment of Different Models

To further verify the performance of the SEBF-YOLO model, it was compared with mainstream object detection models such as Faster-Rcnn, YOLOv3, YOLOv5, YOLOv8, YOLOv11, LSKnet, and Hyper-YOLO models. The detection results of each model are shown in Table 2. The precision and recall rate of the SEBF-YOLO model are significantly higher than those of other models. Compared with the Faster-Rcnn, YOLOv3, YOLOv5, YOLOv8, YOLOv11, LSKnet, and Hyper-YOLO models, the $mAP_{0.5}$ of the SEBF-YOLO model has increased by 12.1, 5.1, 2.8, 4.1, 1.7, 2.8, and 4.1 percentage points respectively. And the $mAP_{0.5:0.95}$ also performs better than other models. Overall, it performs significantly better than other models.

Table 2. Comparison Results of Different Models

Model	P/%	R/%	$mAP_{0.5}$ /%	$mAP_{0.5:0.95}$/%
Faster-Rcnn	56.8	69.4	65.2	/
YOLOv3	73.8	69.5	72.2	42.3
YOLOv5	81.3	69.4	74.5	44.1
YOLOv8	79.2	68	73.2	44.6
YOLOv11	84.9	67.9	75.6	44.8
LSKnet	83.8	67.8	74.5	43.4
Hyper-YOLO	77.8	69.5	73.2	42.6
SEBF-YOLO	**86.2**	**70.4**	**77.3**	**45.4**

Through research and comparison with other models, it can be found that as a two-stage model, Faster-Rcnn filters out high-frequency details through its region proposal mechanism, resulting in weak feature capture ability for small targets. It is unable to effectively capture targets such as aphids and ants. For YOLOv3 and YOLOv5, the traditional downsampling method leads to the loss of small target features, and the neck network only supports the feature fusion of adjacent layers, which is unable to compensate for the blurred and broken features, so the results are not ideal. YOLOv8 will not be discussed for now, as it will be analyzed in detail in the subsequent ablation experiments. Regarding YOLOv11, although it is optimized through spatial-channel decoupling, it lacks a dynamic attention mechanism, resulting in a high false positive rate in complex background situations. As for the LSKnet network designed for small targets, it relies more on large kernel convolutions to enhance dependencies and uses a large receptive field to associate surrounding information for detecting small targets. However, it is more suitable for general small target detection, and the similarity between insects and the environment leads to an increased false positive rate. The Hyper-YOLO model introduces hypergraph computing to model high-order feature relationships, but it has poor adaptability to blurred targets. In contrast, the SEBF-YOLO model takes into account the problems of small targets and blurred targets in complex backgrounds and achieves better results.

5.3 Ablation Experiment

Based on the YOLOv8s as the baseline model, the network structure has been improved. To verify the effectiveness of each improved module, an ablation experiment was conducted and compared with the original YOLOv8s model. The experimental results are shown in Table 3.

Table 3. Results of the Ablation Experiment

Models	P/%	R/%	$mAP_{0.5}$ /%	$mAP_{0.5:0.95}$/%
YOLOv8	79.2	68	73.2	44.6
+SPDconv	85.9	65.4	73	44.5
+BiFormer	83.5	65	73.6	44.9
+EMA	78.7	70.3	73.7	43.7
+SPDconv+EMA	83.9	69.1	74.1	42.9
+SPDconv+EMA +BiFormer	**86.2**	**70.4**	**77.3**	**45.4**

According to Table 3, different modules have distinct impacts on the model's performance. Incorporating the SPDconv module into the backbone network increases the model's precision by 6.7 percentage points, while decreasing the recall rate by 2.6 percentage points, $mAP_{0.5}$ by 0.2 percentage points, and $mAP_{0.5:0.95}$ by 0.1 percentage points. SPDconv retains more details through spatial-to-depth transformation, reducing small object misdetections and enhancing precision. However, the added background noise makes the model more conservative in positioning, raising the missed detection rate and lowering recall and average precision. When using BiFormer alone, the model's precision rises by 4.3 percentage points, the recall rate decreases by 3 percentage points, $mAP_{0.5}$ increases by 0.4 percentage points, and $mAP_{0.5:0.95}$ increases by 0.3 percentage points. By dynamically selecting the focused region in complex backgrounds, BiFormer suppresses interference and improves precision. But its limited target screening in blurry images causes missed detections and a lower recall rate. Still, the attention mechanism enhances positioning, thus improving $mAP_{0.5}$. When introducing EMA alone, precision decreases by 0.5 percentage points, the recall rate increases by 2.3 percentage points, $mAP_{0.5}$ increases by 0.5 percentage points, and $mAP_{0.5:0.95}$ decreases by 0.9 percentage points. Cross-scale feature fusion by EMA addresses insufficient feature extraction in blurry images, enhancing small target detection, reducing missed detections, and increasing recall and $mAP_{0.5}$. Yet, its weak ability to suppress false detections in complex backgrounds leads to a slight precision drop. When both SPDconv and EMA are introduced simultaneously, precision increases by 4.7 percentage points, the recall rate increases by 1.1 percentage points, $mAP_{0.5}$ increases by 0.9 percentage points, and $mAP_{0.5:0.95}$ decreases by 1.7 percentage points. They complement each other in detail retention and multi-scale feature fusion. However, the retained details with background noise result in insufficient robustness in complex situations, reducing positioning ability at high thresholds and decreasing $mAP_{0.5:0.95}$. When BiFormer is added on the basis of using both SPDconv and EMA, compared to the original model, precision increases by 7 percentage points, the recall rate increases by 2.4 percentage points, $mAP_{0.5}$ increases by 4.1 percentage points, and $mAP_{0.5:0.95}$ increases by 0.8 percentage points. BiFormer suppresses the background noise from SPDconv, and EMA compensates for blurry targets BiFormer might miss. The three modules complement each other well, outperforming the original model in all indicators.

5.4 Visualization Results and Analysis

In order to verify the performance of the improved YOLOv8s model, both the original model and the improved model were tested on the test dataset. Some of the detection results are shown in Figure 7. The experiments indicate that when the YOLOv8s model identifies insect images in complex backgrounds, there are cases of false detections. There are also issues of missed detections when recognizing some insects. This is because the backbone network of YOLOv8s restricts its global context awareness ability, and the downsampling process of traditional strided convolutions leads to the loss of fine-grained features. In situations where the images are blurry, the proportion of target pixels is low, and the background is complex, missed detections and false detections often occur.

The improved YOLOv8 model has introduced the SPDconv convolution and the BiFormer bidirectional Transformer module. While addressing the issue of detail loss during convolutional downsampling, it suppresses the interference from complex backgrounds. Additionally, the EMA module is embedded, which can fuse feature information across multiple scales and solve the problem of blurry images. The detection results have verified the effectiveness of the SEBF-YOLO model.

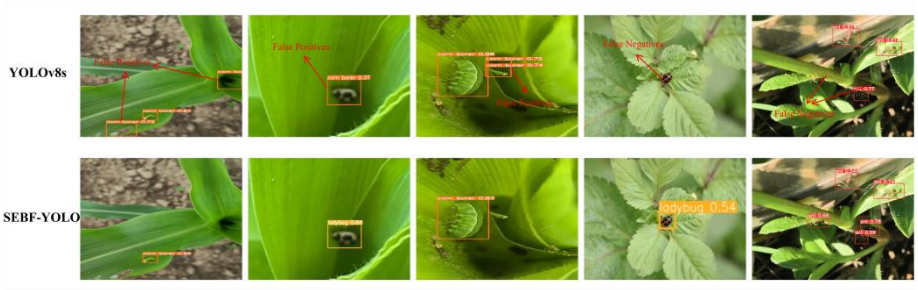

Fig. 8. Detection result graph (After comparing with the original model, it is found that the false positives and false negatives caused by the small target problem and blurring problem in complex backgrounds, which were difficult to solve in the original model, can be better detected in the existing model.)

6 Conclusion

Aiming at the problems of low detection accuracy, false detection, and missed detection caused by blurred insect images and low pixel proportion in complex backgrounds, an improved insect detection method based on YOLOv8s is proposed. Through research and experiments on field insect images, the following conclusions are drawn: By introducing the SPD, BiFormer and EMA during the feature fusion process after feature extraction improvement in the YOLOv8 model, the model's ability to focus on and extract important features is enhanced. While solving the problems, the detection accuracy of the model is improved.

To verify the advantages of the improved SEB - YOLO model, comparative experiments are conducted between the improved model, the original YOLOv8s model, and recent object detection models such as YOLOv5, YOLOv11, LSKnet, and Hyper-YOLO. The results show that the improved YOLOv8 model performs excellently in terms of detection accuracy, recall rate, and average precision at different IoU thresholds.

Although the improved model has made progress in the detection of small objects, there are still issues of missed detections for extremely dense insect images. Future work will focus on optimizing multi-object detection for dense insect images to meet the practical needs of insect detection.

Acknowledgments. This research was funded by Sichuan Science and Technology Program (grant numbers: 24GJHZ0388), the Chengdu Research Base of Giant Panda Breeding (grant numbers: 2020CPB-C09, 2024CPB-B08).

References

1. Hao, D.J., Wang, Y., Dai, H.G., et al. Strategy for Ecological Management of Pests in Plan tations and Prospect for Control Techniques. Journal of Northeast Forestry University 200 4(6), 84–86 (2004).
2. Martineau, C., Conte, D., Raveaux, R., et al. A survey on image-based insect classification. Pattern Recognition 65, 273–284 (2017).
3. Hou, M.L., Lian, F.Y., Qin, Y., et al. Predicting model of stored-grain temperature based o n an improved feature enhanced broad learning. Journal of Henan University of Technolog y (Natural Science Edition). https://doi.org/10.16433/j.1673-2383.202407150001 (last acce ssed: 2025-03-21).
4. Wang, Q.H., Liu, Y.K., Zheng, Q., et al. SMC-YOLO: A High-Precision Maize Insect Pest -Detection Method. Agronomy-Basel 2025, 15(1), 195. https://doi.org/10.3390/agronomy1 5010195.
5. Deng, L., Wang, Y., Han, Z., Yu, R. Research on insect pest image detection and recogniti on based on bio-inspired methods. Biosystems Engineering 2018, 169, 139–148. https://do i.org/[Google Scholar DOI].
6. Sun, X., Wu, P., Hoi, S.C.H. Face detection using deep learning: An improved faster RCN N approach. Neurocomputing 299(19), 42–50 (2018).
7. Liu, W., Anguelov, D., Erhan, D., et al. SSD: Single Shot Multibox Detector. In: Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, LNCS, vol. 9999, pp. 21–3 7. Springer, Heidelberg (2016).
8. Redmon, J., Divvala, S., Girshick, R., et al. You only look once: unified, real-time object d etection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. IEEE, New York (2016).
9. Tan, X.P., Gao, Z.H., Han, H.D., et al. Intelligent Detection of Cells in Fluorescence Image s Based on Improved YOLOv5. Semiconductor Optoelectronics 44(5), 709–716 (2023).
10. Zheng, G., Jiang, Y.S., Shen, Y.L., et al. Recognition of Rice Pests Based on Improved YO LOv7. Journal of Huazhong Agricultural University 42(3), 143–151 (2023). https://doi.org/ 10.13300/j.cnki.hnlkxb.2023.03.017.
11. Gao, J.J., Zhang, X., Guo, Y., et al. Research on the optimized pest image instance segmen tation method based on the Swin Transformer model. Journal of Nanjing Forestry Universi

ty (Natural Sciences Edition) 47(3), 1–10 (2023). https://doi.org/10.12302/j.issn.1000-2006.202206048
12. Zhang, Y.J., Ma, B.X., Hu, Y.T., et al. Accurate cotton diseases and pests detection in complex background based on an improved YOLOX model. Computers and Electronics in Agriculture 2022, 203, 107484. https://doi.org/10.1016/j.compag.2022.107484.
13. Jiao, L., Li, G.Q., Chen, P., et al. Global context-aware-based deformable residual network module for precise pest recognition and detection. Frontiers in Plant Science 2022, 13, 895944. https://doi.org/10.3389/fpls.2022.895944.
14. Yuan, Z.M., Yuan, H.J., Yan, Y.X., et al. Automatic recognition and classification of field insects based on lightweight deep learning model. Journal of Jilin University (Engineering and Technology Edition) 51(3), 1131–1139 (2021). https://doi.org/10.13229/j.cnki.jdxbgxb20200116.
15. Hu, J., Shen, L., Sun, G. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE, Salt Lake City (2018). https://doi.org/10.1109/CVPR.2018.00745.
16. Wang, Q., Wu, B., Zhu, P., et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539. IEEE, Seattle (2020). https://doi.org/10.1109/CVPR42600.2020.01155.
17. Wu, T.H. and Deng, B.G. A Small Target Detection Algorithm for Aerial Images Based on Improved YOLOv8n: LS-YOLO. Telecommunication Engineering. https://doi.org/10.20079/j.issn.1001-893x.241024003 (last accessed: 2025-03-21).
18. Chu, X., Li, X., Luo, B., Wang, X.D., & Huang, S. (2023). Identification method of tomato leaf diseases based on improved YOLOv4 algorithm. Jiangsu Journal of Agricultural Sciences, 39(5), 1199-1208. https://doi.org/10.12345/j.issn.1000-2040.2023.05.001
19. Sunkara, R., Luo, T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 443–459. Springer, Cham (2022).
20. Ouyang, D.L., He, S., Zhang, G.Z., et al. Efficient multiscale attention module with cross-spatial learning. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, New York (2023).
21. Zhu, L., Wang, X.J., Ke, Z.H., et al. BiFormer: vision transformer with Bi-level routing attention. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10323–10333. IEEE, New York (2023).