Proceedings of Machine Learning Research vol 245:1–12, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Detection Method of Forest Pests Based on Attention Mechanism and Lightweight YOLOv5

Kehao Cha 981608731@QQ.COM
School of Software, Dalian Jiaotong University, Dalian, Liaoning, 116028, China

Xudong Song
School of Computer Science, Dalian Jiaotong University, Dalian, Liaoning, 116028, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

In view of the forestry pest identification research is less, manual identification time-consuming and labor-intensive low accuracy. An attention-based and lightweight YOLOv5 forestry pest identification method was proposed. First of all, the traditional backbone network CSPDarknet is modified to the improved ShuffleNetV2, which simplifies the network structure and makes the network more lightweight; Secondly, the hybrid attention mechanism CBAM (Convolutional Block Attention Module) is introduced to increase the perception ability of cyberspace and channel features while keeping the parameters and calculation load basically unchanged. Finally, the loss function is replaced by WIoU to improve model training to speed up model convergence and improve regression accuracy. The average detection accuracy of the improved model is 89.9%, which is 3.6% higher than that of the original YOLOv5s algorithm; The parameters decreased by 3213050 and the calculation amount decreased by 7.8. The improved model improves the detection accuracy and reduces the parameters and calculation amount. Compared with other advanced algorithms, the algorithm in this paper has excellent performance, which can provide reference for forest pest identification and management.

Keywords: forestry pests, YOLOv5, CBAM, ShuffleNetV2, WIoU

1. Introduction

Studies of today have revealed that forest pests are a grave danger to forest health and ecosystem steadiness. It is essential to detect and observe the existence and spread of pests quickly and precisely for taking prompt control steps. However, due to the variety, complex morphology and concealment of forest pests in the natural environment, traditional pest detection methods face many challenges. The advent of deep learning technology in recent times has presented a novel approach to the challenge of pest identification.

Traditional methods rely on artificial feature design and machine learning, which requires domain experts to extract features and use classifiers for pest detection. However, there are great limitations, such as the need for expert knowledge and time-consuming. In contrast, deep learning methods can automatically learn feature represent as Faster-RCNN (Ren et al., 2016), SSD (Liu et al., 2016), YOLO (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020). In addition, the deep learning method has better generalization ability and can adapt to challenges such as pest diversity and background interference. At present, many scholars have done a lot of research on pest detection in agriculture and forestry. Xiao et al. (2021) and other the traditional convolution neural network Alexnet, removed local response normalization and normalized it after

convolution layer, and adopted global average pooling and PReLU activation function optimization model, which provided a new idea for intelligent identification of agricultural pests.

Proposed by Lei et al. (2022) and others, a multi-scale attention residual network model supplanted ordinary convolution with multi-scale convolution, and incorporated an attention mechanism selective kernel unit into the residual structure, thus augmenting the receptive field and suppressing the Sun et al. (2022) and colleagues; enhancement of detection speed and accuracy, as well as the resolution of leakage and misdetection issues through data preprocessing enhancements. lightweight network, attention module and Focal Loss. It meets the accuracy and speed requirements of real-time detection of forestry pests. Sun et al. (2023) and others proposed an improved YOLOv4 algorithm for stored grain pest detection. A YOLOv5-based forest pest image detection algorithm is proposed in this paper, which seeks to solve the issues of small targets overlapping and occlusion, as well as improve detection accuracy and speed. This is accomplished by utilizing K-means clustering and spatial pyramid pooling for multi-scale detection. The main improvement points are as follows:

(1) To address the issue of high parameter count and challenging deployment of the model, this paper replaced the backbone network with the enhanced ShuffleNetV2 (Ma et al., 2018). Specifically, this paper replaced the convolution operation in the Shuffle_Block with grouped convolution using 4 groups to reduce the module's parameter count.

(2) In response to the model's inadequate feature extraction capability, this paper enhanced the C3 neck network by incorporating the CBAM (Woo et al., 2018) mechanism, forming the C3CBAM module. This modification aims to intensify the model's attention towards the target and enhance detection accuracy.

(3) To address issues with inaccurate anchor frame matching and slow convergence speed encountered by the CIoU model, this paper propose modifying the original CIoU loss function to WIoU (Zhang et al., 2022). Additionally, this paper employ a monotonic focusing mechanism to compute anchor frame features, aiming to enhance anchor frame matching accuracy.

The aforementioned enhancements have significantly improved the model's capability to detect forest pests and have yielded promising results on the forest pest dataset from Peking University. Experimental findings have demonstrated its practical efficacy and value.

2. Model Introduction

As shown in Figure 1, in this paper, the backbone network of YOLOv5 is replaced by ShuffleNetV2, which is a lightweight neural network model designed to reduce the computational complexity and number of parameters of the model while maintaining high accuracy. Its core structure is depthwise convolution and channel shuffle operations. The Shuffle_Block unit is the basic building block of ShuffleNetV2. It is based on group convolution, which divides the input feature map into several groups and then applies depth-separable convolution operation to each group. In depth-separable convolution, each input channel is first convolved independently, and then the results are combined between channels. This design can effectively reduce the amount of computation and the number of parameters, and maintain the expressive power of the feature map to some extent. Channel rearrangement is another key operation of ShuffleNetV2 for exchanging information about the feature map within the Shuffle_Block unit. It divides the channels of the input feature map into groups and cross-mixes the channels of different groups. Through the channel rearrangement

operation, different channels in the feature map can communicate and blend with each other to enhance the feature expression capability.

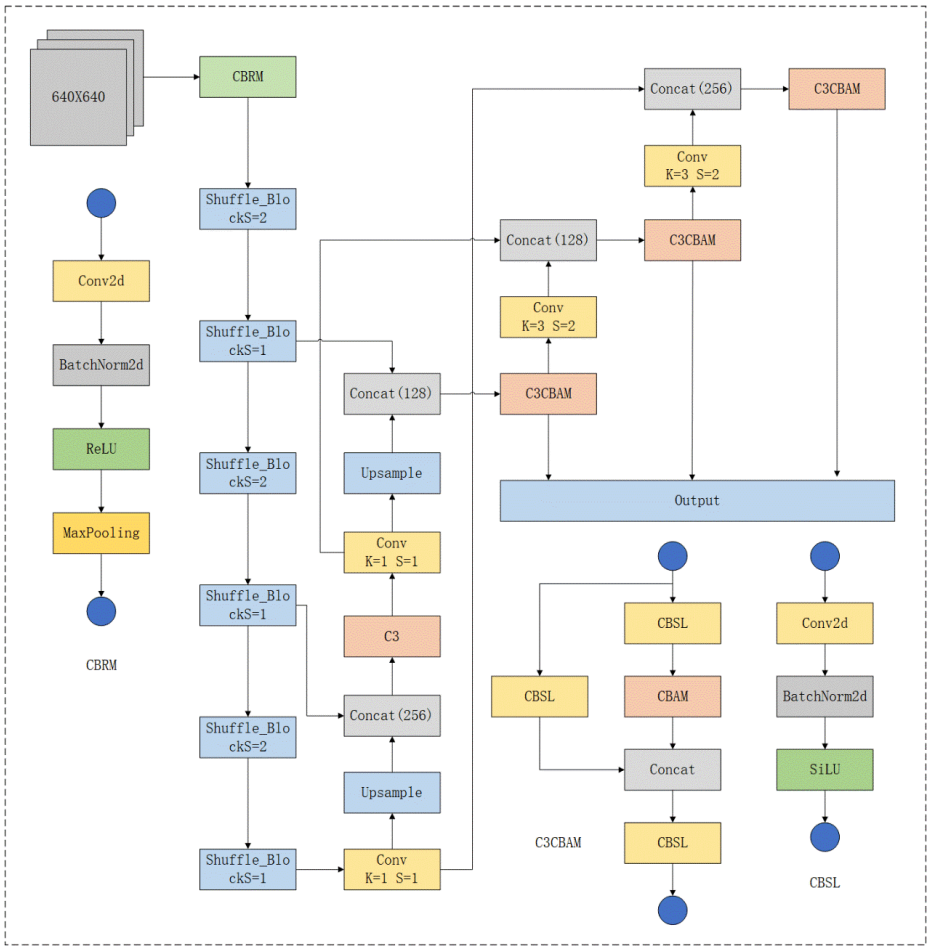

Figure 1: Algorithm structure in this paper.

2.1. Lightweight Feature Extraction

Table 1 shows the various parts of the improved and reconstructed backbone network, in which the CBRM module performs feature extraction and down-sampling operations on input data. Secondly, Shuffle_Block module constructs two branches according to the value of step size, and if the step size is greater than 1, it will take down-sampling process. If the step size is equal to 1, no down sampling is performed. This paper has been significantly enhanced by the implementation of group convolution in Shuffle_Block, replacing the traditional ordinary convolution with group convolution with a group number of 4. The advantage of this is that it can simplify module operation, reduce module parameters.

Table 1: ShuffleBlock.

Layer	Output Size	Stride	Output channels
Input	640×640	–	3
CBRM	160×160	–	32
Shuffle_Block2	80×80	2	128
Shuffle_Block1	80×80	1	128
Shuffle_Block2	40×40	2	256
Shuffle_Block1	40×40	1	256
Shuffle_Block2	20×20	2	512
Shuffle_Block1	20×20	1	512

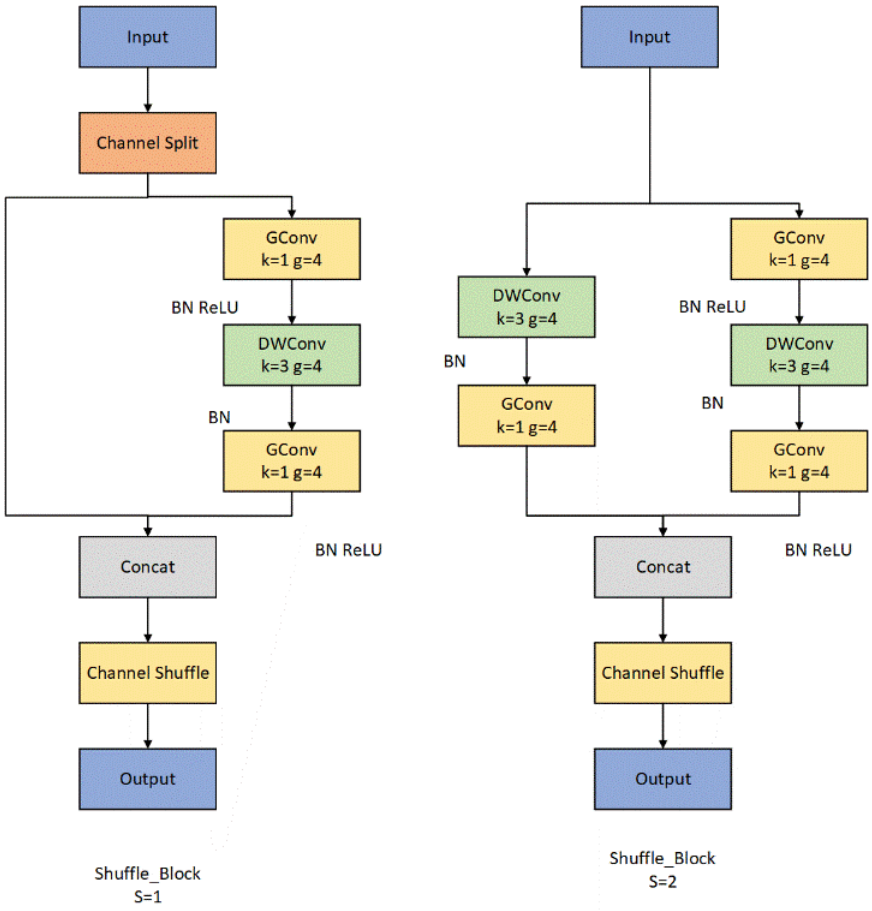

Figure 2: Shuffle_Block module.

Figure 2 is the basic unit of the improved ShuffleNetV2 feature extraction network. Shuffle_Block1 splits the input feature channels into two groups. To eliminate superfluous activities, one group abstains from processing, while the other executes a single deep separable convolution and two 1×1 group convolution and normalization operations to guarantee that the number of channels

in the left and right groups is equal. Subsequently, Concat operation is executed on both groups. Finally, the order is randomly disrupted in the channel dimension, thus strengthening the fusion of channel feature information. Shuffle_Block2 no longer dividing channels, has adopted down-sampling operation in both the left and right branches, thus diminishing the size of the feature map and enlarging its dimension. The two units together constitute the basic structure of ShuffleNetV2, which makes the computation and parameters of ShuffleNetV2 network relatively low, thus realizing a lightweight feature extraction network.

2.2. Adding Attention Mechanism

CBAM, short for Convolutional Block Attention Module, represents an attention-enriching mechanism tailored for convolutional neural networks (CNNs). Its principal function revolves around adaptively discerning the significance of features across various spatial locations within an image, thereby amplifying the efficacy of feature representation. Given the petite scale and inconspicuous attributes characterizing forest pests, the incorporation of the CBAM attention mechanism serves to heighten the expression of pertinent features associated with the target, thereby augmenting detection precision.

The CBAM structure is succinctly illustrated in Figure 3, delineating its architectural configuration. Concretely, the operational sequence of CBAM unfolds through the following steps:

(1) Firstly, the feature map is input into the CBAM module. Subsequently, two distinct operations, namely F maximum pooling and F average pooling, are conducted on the feature map.

(2) The channel attention feature map is determined by the formula F_C, which yields two channel features that are then input into MultiLayer Perceptron (MLP) for a series of convolution operations. Subsequently, the convolved elements are added one by one and inputted into Sigmoid function to activate.

$$F_C = Sigmoid(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{1}$$

(3) Obtain the channel attention characteristic map, multiply F_C the input characteristic map F and the channel attention characteristic F_C map, and obtain the middle section characteristic map by the formula F':

$$F' = F \times F_C \tag{2}$$

(4) After average and maximum pooling operations, F' is spliced into the input middle feature map. Subsequently, 7×7 convolution is reduced to a channel, prompting the Sigmoid function to be activated. Finally, F_S sub-modes are used to obtain the spatial attention feature map.

$$F_S = Sigmoid(C^{7\times7}[AvgPool(F); MaxPool(F)]) \tag{3}$$

(5) Obtain the spatial attention feature mapF_S, multiply the spatial attention feature F_S map with the middle F' feature map, and obtain the CBAM attention mechanism feature map by the following formula F''

$$F'' = F_S \times F' \tag{4}$$

In this paper, CBAM and BottleNeck module of YOLOv5s are replaced to form C3CBAM module, replacing C3 module of PANet part of YOLOv5s, which not only reduces the number of parameters, but also improves the feature extraction ability of the network.

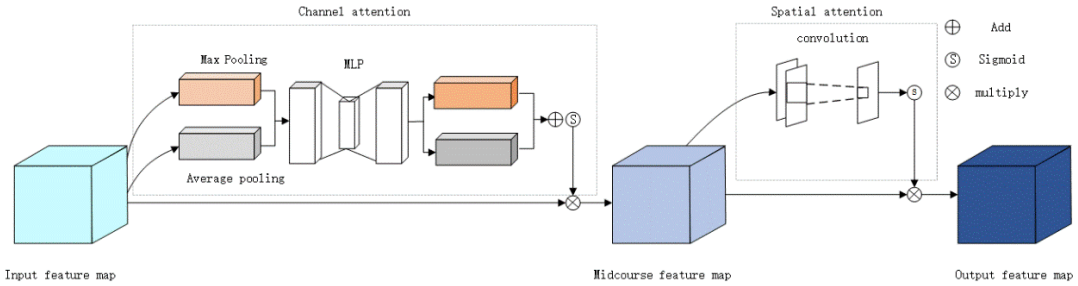

Figure 3: CBAM Module.

2.3. Improvement of Loss Function

The training process of target detection employs IoU (Intersection over Union), a tool created to gauge the correspondence between the prediction box and the real box. The loss function of YOLOv5 is CIoU (Complete IoU), which combines the concepts of IoU and CIoU. IoU only considers the position information of bounding boxes, while ignoring other important factors, such as the size, shape and overlap degree of bounding boxes. An extra adjustment factor, CIoU, is introduced that takes into consideration the magnitude and convergence of boundary boxes. It is described as:

$$CIoU = IoU - \left(\frac{p^2\left(b, b^{gt}\right)}{C^2} + av \right) \tag{5}$$

Here is $p^2(b, b^{gt})$ the distance between the center point of the prediction box and the real box C^2, the diagonal distance between the prediction box and the real box av, and the penalty factor of the length-width ratio between the prediction box and the real box, a where is the weight functiovn, which can measure the consistency of width w_{gt} and height, and a below v is the calculation formula of:

$$a = \frac{v}{(1 - IoU) + v} \tag{6}$$

$$v = \frac{4}{\pi}\left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w}{h} \right)^2 \tag{7}$$

By using the above formula, CIoU can measure the matching degree of bounding boxes. The advantage of CIoU is that the aspect ratio of frames is considered, which solves the problem that the aspect ratio of DIoU is different but the center points overlap, which leads to insufficient discrimination. Reflecting the aspect ratio difference instead of the real difference expressed by width and height and confidence, the penalty function will fail when the prediction box width and height meet certain conditions, thus impacting the optimization similarity of the network - a disadvantage.

Hence, this paper introduces the Wise-IoU (WIoU) loss function. WIoU is formulated based on a dynamic non-monotonic focusing mechanism, which employs ”outlier” instead of IoU to assess the quality of anchor frames. Additionally, it offers a sophisticated gradient gain allocation strategy, aimed at reducing the competitiveness of high-quality anchor frames and mitigating the adverse gradients generated by low-quality anchor frames. This enables WIoU to prioritize normal quality anchor frames and enhance the overall detector performance. Notably, WIoU comprises three versions: WIoU V1, WIoU V2, and WIoU V3. In this study, WIoU V3 is selected, representing an

optimized and enhanced iteration built upon WIoU V1 and WIoU V2. The formula for WIoU V1 is delineated as follows:

$$L_{IoU} = 1 - IoUL_{IoU} \tag{8}$$

$$L_{WIoUV1} = R_{WIoU}L_{IoU} \tag{9}$$

$$R_{WIoU} = exp\left(\frac{(x - x_{gt})^2 - (y - y_{gt})^2}{\left(W_g^2 + W_g^2\right)^*}\right) \tag{10}$$

where the W_g width H_g and height of the minimum bounding box are represented. To prevent R_{WIoU} a gradient that hinders convergence W_g, H_g and to separate it from the calculation diagram (superscript * indicates this operation). Because it effectively eliminates the factors hindering convergence, it does not introduce new measures, such as aspect ratio.

WIoUV2 adds monotone focusing coefficient to WIoUV1 and constructs L_{IoU}^{y*} the following formula:

$$L_{WIoUV2} = L_{IoU}^{\gamma*}L_{WIoUV1}, \gamma > 0 \tag{11}$$

In the training process of the model $L_{IoU}^{\gamma*}$, it decreases L_{IoU} with the decrease, which will lead to slow convergence speed in the later training period. Therefore, the introduced L_{IoU} mean value is used as the normalization factor:

$$L_{WIoUV2} = \left(\frac{L_{IoU}^*}{\overline{L_{IoU}}}\right)^{\gamma} L_{WIoUV1} \tag{12}$$

In the above formula L_{IoU}, the average running value of momentum m is added. Because of the normalization factor, the gradient gain $\left(\frac{L_{IoU}^*}{\overline{L_{IoU}}}\right)^{\gamma}$ is kept at a high level, which ensures the convergence speed in the later training period.

WIoUV3 introduces a dynamic non-monotonic focusing coefficient based on WIoUV2. The composition formula is as follows:

$$\beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \epsilon [0, +\infty] \tag{13}$$

$$L_{WIoUV3} = rL_{WIoUV1}, r = \frac{\beta}{\delta\left(\alpha\right)^{\beta-\delta}} \tag{14}$$

In the above equation β denotes the outlier, where α and δ denote the hyperparameters controlling the gradient gain r. Here α = 1.9, δ = 3. In this paper we use exactly WIoUV3.

3. Experimental Results and Analysis

3.1. Introduction of Forest Pest Data Set

As shown in Table 2, this paper is based on the publicly available dataset of forestry pests in the family of small stupidity from Beijing Forestry University, which contains a total of 2,183 images in six categories, of which 1,693 are training sets, 245 are test sets, and 245 are test sets. The six categories are red fat small and large stupid, pine twelve-toothed small stupid, Huashan pine large and small stupid, spruce eight-toothed small stupid, four-eyed small stupid, and six-toothed small stupid. The images are all on a single white background, with a resolution of 1236×1236 scaled to 640×640, using fill-in light and natural light, with and without alcohol in the collector, with and

without alcohol, with the number of pests divided into crowded and sparse, and with the lens focus divided into precise and fuzzy when taking pictures, so that the dataset can contain as much as possible all kinds of complexity to increase the accuracy of the recognition rate.

Table 2: Number of labels.

Insect name	Sample size
Leconte	2216
Boerner	1595
Armandi	1765
Coleoptera	2091
Acuminatus	953
Linnaeus	1727

3.2. Experimental Environment

Experiments comparing the proposed method to the YOLOv5s model were conducted to evaluate the feasibility and efficacy of these enhancements. The experiments were conducted in a cloud platform environment using an Ubuntu system with 15 vCPUs, an Intel (R) Xeon (R) Platinum 8358P CPU @2.60 GHz processor, 80GB of memory, and an RTX A5000(24GB) graphics card. Using PyTorch Python development framework, experiments of 32 batches ran for 100 epochs.

In evaluating detection accuracy, we employed Average Precision (AP) and mean Average Precision (mAP). Additionally, Precision, Recall, model parameters, and floating-point operations (GFLOPs) were utilized to gauge detection accuracy. Collectively, these metrics offer a thorough comprehension of the detection models being examined.

$$P = \frac{TP}{TP + FP} \cdot 100\% \tag{15}$$

$$R = \frac{TP}{TP + FN} \cdot 100\% \tag{16}$$

$$AP = \int P\left(R\right) \tag{17}$$

$$MAP = \frac{1}{C} \sum_{J}^{C} \left(AP\right)_j \tag{18}$$

3.3. Ablation Test

To assess the efficacy of the alterations suggested in this paper, the YOLOV5s network is enhanced and epoch 100 rounds are chosen to incrementally incorporate various enhanced experimental data for comparison. For the convenience of table making and understanding, abbreviations will be adopted below, such as ShuffleNet abbreviated as SF, C3CBAM abbreviated as CM, and WIoU abbreviated as W. The Table 3 reveals that, following the initial enhancement of the backbone network, the models precision reaches 0.875, the parameter amount drops to 3802211 and the computation

amount also drops to 8.0 - a remarkable decrease. Most of the decreased parameters and computation quantity are attributed to the deep separable operation adopted in ShuffleNet and the grouping convolution improvement adopted in this paper. Then, C3CBAM module is added on this basis. The models precision is augmented further upon the introduction of the CBAM attention mechanism. Map50 reaches 0.895, Map95 reaches 0.641, and the accuracy increases. The recall rate has little change. The WIoU loss function is, at last, incorporated to augment the precision of anchor frame forecasting. The average accuracy (Map50) of the final model reaches 0.899 from 0.863, the Precision (Precision) reaches 0.832 from 0.818, the Recall (Recall) reaches 0.833 from 0.818, and the model parameters decrease by 3213050. The calculation amount decreases by 7.8.Experimental findings demonstrate that the enhanced ShuffleNet network not only reduces its size but also preserves satisfactory detection outcomes. Then CBAM attention mechanism is introduced to deepen the attention of the network to the spatial and channel characteristics. The addition of the WIoU loss function to augment convergence and training effect was finally completed. This ablation experiment verified the efficacy and practicality of this improvement in forest pest image detection.

Table 3: Ablation experiment.

Model	Map50	Map95	Precision	Recall	Parameters	GFLOPs
Baseline	0.863	0.628	0.818	0.818	7029004	15.8
SF	0.861	0.626	0.828	0.798	3802211	8.0
SF + CM	0.895	0.641	0.831	0.823	3815954	8.0
SF + CM + W	0.899	0.651	0.832	0.833	3815954	8.0

3.4. Contrast Test

To assess the detection capability of the network suggested in this paper, this paper has chosen sophisticated algorithms such as Faster R-CNN and SSD to be compared on the training set of the Faster R-CNN data set at Beijing Forestry University. It can be seen from Table 4 that Faster R-CNN has good average accuracy and few parameters, only 477758 parameters, which has great advantages for the deployment of the model in terminals and mobile terminals, but its problem is that the detection speed is slow, only 15 frames per second, which obviously cannot meet the application requirements. The FPS of SSD has only slightly augmented to 23, barely satisfying the applications needs; however, the mean precision has dropped by nearly 7%, and the parameters have been significantly enhanced, which is evidently impossible. The average accuracy of the baseline model in this paper is basically the same as that of the two-stage Faster R-CNN, but the parameters are large, FPS rises sharply to 75 frames per second, and the detection speed is considerable, but at the same time, the parameters reach 7 million, which is not convenient for deployment on mobile devices or terminal platforms. The improved model average accuracy is augmented by 3.6%, its parameter quantity is decreased by 3213050, and its FPS is 64 frames per second - a remarkable performance in comparison to Yolov5s.

Table 4: Comparative Test.

Model	Map	Parameters	FPS
Faster R-CNN	0.866	477758	15
SSD	0.798	3737867	23
Yolov5s	0.863	7029004	75
Text model	0.899	3815954	64

Select three pest detection pictures in the data set, which are the original picture and the original YOLOv5 algorithm and the improved YOLOv5 algorithm for pest detection. As can be seen from Figure 4, the improved YOLOv5 algorithm is more accurate in detection categories, and the original YOLOv5 algorithm often has the problem of category detection errors, while the algorithm in this paper is not only more accurate in classification, but also higher in accuracy rate than the original algorithm. Experiments show that the proposed improvement significantly reduces the probability of false detection and improves the classification accuracy of the model.

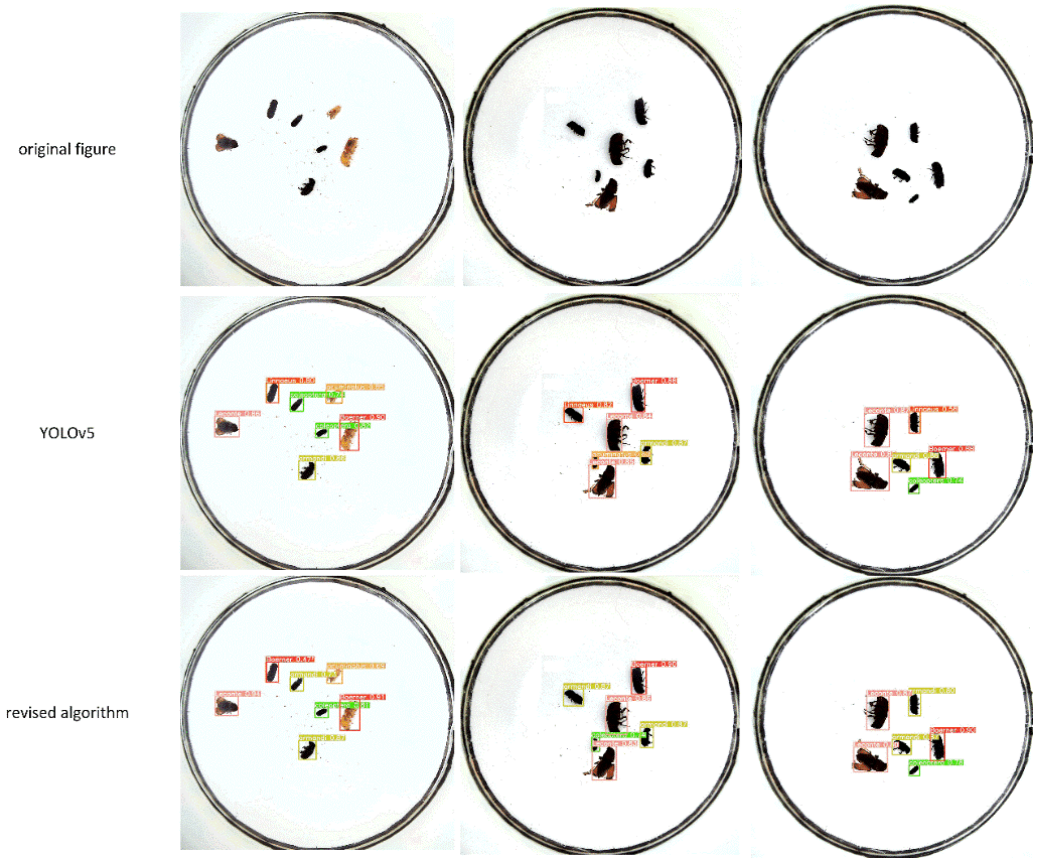

Figure 4: Picture comparison.

4. Conclusion

This paper present an attention-based and lightweight approach to forestry pest identification, utilizing YOLOv5, in response to the difficulty of identifying small and hard-to-find forest pests, a task that is often time-consuming, laborious, and prone to mistakes when done manually. Experimental results demonstrate that the improved model achieves an average accuracy of 89.9% and operates at a detection speed of 64 frames per minute. Verifying the efficacy and effect of our suggested model enhancements, ablation experiments and analyses of the enhanced experiments were conducted. The main network enhancement experiment confirms that ShuffleNetv2 can efficiently process data while maintaining model performance. Furthermore, the attention mechanism experiment verifies that CBAM enhances the networks feature extraction capabilities without increasing computational burden. Meanwhile, the loss function experiment demonstrates that WIoU accelerates convergence speed and boosts model performance without altering model size. Visual results from the experiments reveal superior detection accuracy compared to the original model. Furthermore, performance advantages over some advanced algorithms are verified. Future research will explore the feasibility of deploying the lightweight model in practical applications.

References

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Chengmin Lei, Shaomin Mou, Wenjie Sun, and Enquan Cui. Image recognition of peach pests based on multi-scale attention residual network. *Journal of Shandong Agricultural University (Natural Science Edition)*, (02):53, 2022.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37, 2016.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Haiyan Sun, Yunbo Chen, Dingwei Feng, Tong Wang, and Xingquan Cai. Forest pest detection method based on attention model and lightweight yolov4. *Computer Application*, 42(11):3580–3587, 2022.

Yan Sun, Xuehua Song, Jing Chen, and Siwei Jiang. Detection algorithm of stored grain pests based on improved yolov4. *Computer and Digital Engineering*, 51(6):1217–1222, 2023.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

X Xiao, H Yang, W Yi, Y Wan, Q Huang, and J Luo. Application of improved alexnet in image recognition of rice pests. *Science Technology and Engineering*, 21:9447–9454, 2021.

Yi-Fan Zhang, Weiqiang Ren, Zhang Zhang, Zhen Jia, Liang Wang, and Tieniu Tan. Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing*, 506:146–157, 2022.