Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 3, 2024

Detection of Small Object based on Improved-YOLOv8

Yingying Tan *, Jinpeng Song, Chen Chu

Department of Computer, North China Electric Power University, Baoding, Hebei, China
*** Corresponding author:** Yingying Tan (Email: 15073395727@163.com)

Abstract: An improved YOLOv8 model is proposed to address the issue of poor recognition performance caused by their low resolution and weak feature representation in small object detection task. Firstly, to extract a richer set of low-level features from images, a WT_Conv module is designed to fuse the feature components extracted by WT (Wavelet Transform) with those extracted by convolutional layer. Secondly, based on the idea that shallow and deep feature maps contain information at different scales, a MS (Multiscale Separation) module is designed to preserve the features of small objects separated from shallow layer and transfer the salient features of large objects to the deeper layers, effectively solving the problem of inconsistent feature expression caused by the direct fusion of shallow and deep feature maps. Finally, we introduce the DE (Detail Enhancement) module capable of fusing adjacent feature maps to process the small-objects features separated by the MS module, enhancing feature representation for small objects. Experiment results on UAVOD-10 and Small Object datasets show that our model achieves a mAP improvement of 9.5% and 2.3% respectively over the baseline, and it also shows a significant advantage over other comparative models, affirming the effectiveness of the proposed model for small object detection tasks.

Keywords: Small Object Detection; YOLOv8; Wavelet Transform; Multiscale Separation; Feature Fusion.

1. Introduction

Object detection has made remarkable achievements in many fields, but small object detection is still a very difficult task. In the image, small objects generally have small coverage areas and low resolution, making them easy to be affected by environmental factors like occlusion and illumination changes. This results in their detection performance being far worse compared to that of large and medium objects. However, there is a great demand for small target detection in modern society. For example, in the industrial field, small object detection algorithms are used to accurately locate tiny defects on material surfaces; In agriculture, they are used to detect small crops and pests in the field for effective field management; In the field of aerial remote sensing, it is required to be able to accurately identify the object with the size of only dozens of pixels from satellite images. Therefore, small object detection has important practical significance.

In recent years, many new algorithms for small object recognition have been proposed. For example, in terms of data preprocessing technology, data enhancement strategies such as Mosaic [1] and Copy-Pasting [2] are applied to dataset to improve the proportion of small objects in the image, so as to enhance the learning ability of the algorithm model for small object features. In the optimization technology of neural network structure, attention mechanism [3-4] is introduced into the network to enhance the model's attention to local information. In terms of positive and negative sample matching technology, the positive sample region is redefined by rotating the box [5] to improve the detection accuracy of dense small objects. In terms of feature fusion, there are NET mechanism [6] for reconfiguring features to eliminate scale confusion, and RiLFE module [7] for fusing multiple feature maps to enrich the feature information of small objects. Various feature pyramid methods have also been applied, such as PAFPN [8], BiFPN [9], NAS-FPN [10], etc. The common purpose of these feature pyramid methods is to fully integrate the feature maps of different scales, so as to generate more expressive feature maps.

Object detection algorithms can be classified into Two-Stage algorithms and One-Stage algorithms. YOLO [11] is one of the classical single-stage algorithms. Among them, YOLOv8 model not only has the advantage of fast detection speed, but also has better detection performance for large and medium-sized objects than mainstream two-stage object detection algorithms. The PAFPN structure in YOLOv8 directly connects the shallow feature map of the feature extraction network and the deep feature map of the fusion network horizontally to enrich the details of the features of the detection layer. However, there is a significant difference in information between the feature maps with large scale variations. In shallow layers, feature maps contain various scale target information, while feature maps loses a lot of small object feature in deep layers. If the two layers are directly sampled to the same scale and then fused, the differences will be ignored, resulting in inconsistent representations, and information of large objects will weakening the small-object features thus interfere the detection of small objects.

As a conventional algorithm, Wavelet Transform transforms the image from the spatial domain to the wavelet domain. There are many researches that combine wavelet transform and convolutional neural network in image denoising, image reconstruction, object detection and other tasks. For example, in literature [12], wavelet transform is used to obtain wavelet residual images to guide network training, which significantly improves the de-noising effect. In literature [13], the neural network reconstructs high-resolution images by learning the Wavelet Transform coefficients of low-resolution face images, effectively enhancing the clarity of face images. [14] enhances object detection performance by introducing a Haar wavelet-based lossless feature encoding block into the downsampling operation of the network. [15] extract the defect images component by Wavelet Transform to achieve better segmentation effect. Small object image usually has the problem of blurred edge and low distinction from foreground.

A richer and more comprehensive feature representation can be constructed by combining the edge and texture extracted by wavelet transform with the features extracted by convolution.

Based on the above analysis, this paper designs an improved algorithm for small target detection. The main contributions are as follows:

(1) Wavelet-transform convolution (WT-Conv) module is designed, and wavelet transform technology is introduced into feature extraction network to extract high-frequency and low-frequency components in images and enrich low-level features.

(2) Multiscale Separation (MS) module is designed to achieve information stripping of objects of different scales.

(3) A Detail Enhancement (DE) module is designed to fuse small object features from adjacent feature maps, thereby increasing the semantic information of channels and enhancing the communication between channels to enhance the feature expression of small objects.

2. Methodology

The YOLOv8 network model consists of three components: the Backbone for feature extraction, the Neck for feature fusion, and the Detect head for object detection. The network structure of the improved YOLOv8 model proposed in this paper is illustrated in Fig. 1. Firstly, modifications are made to the Backbone structure of the meta-model by replacing the C_1 layer with a WT_Conv module, which fuses wavelet-transformed feature components with original image features. Secondly, MS modules and DE modules are introduced between the original Backbone and Neck. Two MS modules divide features from shallow layer in Backbone to retain details of small objects while transferring large-objects feature information to deeper layers in the network. The two DE modules enhance small object features and fuse them with bottom-up information flow in the Neck section to provide more direct small object features for detection at D_1 layer. Additionally, a detect layer at D_3 is employed specifically for detecting large objects.

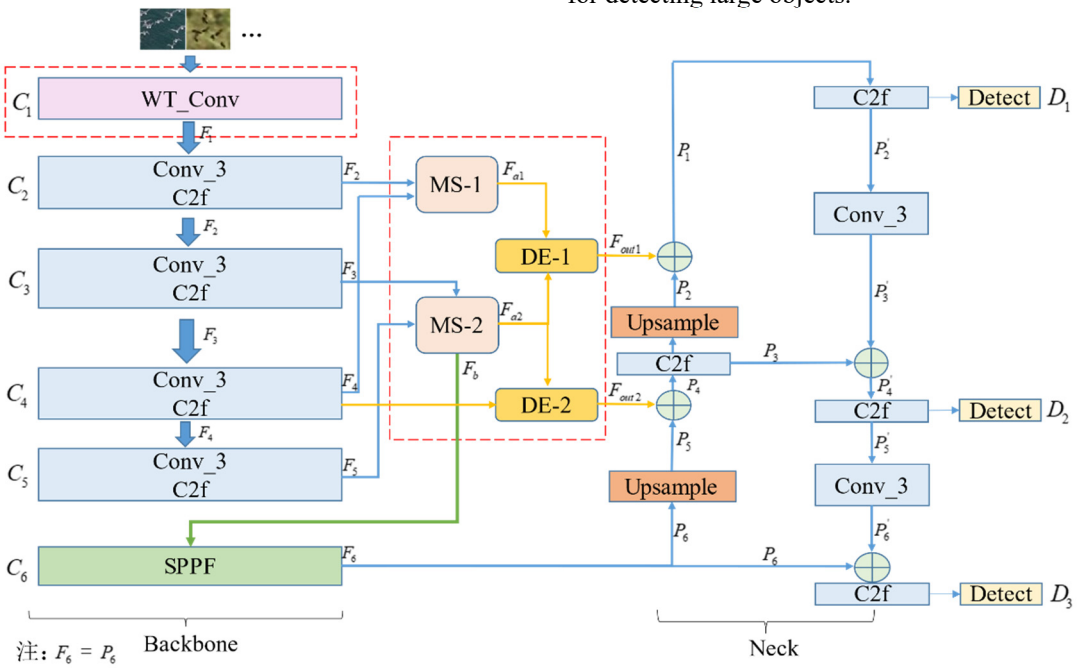


Fig 1. The network structure of improved-YOLOv8 model

2.1. WT_Conv Module

Wavelet Transform (WT) is a classical image processing algorithm, which can transform an image from the spatial domain to the wavelet domain to extract the edge, texture and other features of the image. Small objects usually have blurred edges and low differentiation from the foreground, so extracting these fine features is particularly important for small object detection. By incorporating wavelet transform into the convolutional neural network, the feature richness can be improved without increasing the number of parameters. The WT_Conv module structure is shown in Fig. 2. It is necessary to convert the RGB image to the gray image as the input of the wavelet transform, and finally obtain the approximation coefficients (cA), horizontal detail coefficients (cH), vertical detail coefficients (cV) and diagonal detail coefficients (cD). The process is expressed by the formula:

$$cA, cH, cV, cD = WT(gray\_img) \quad (1)$$

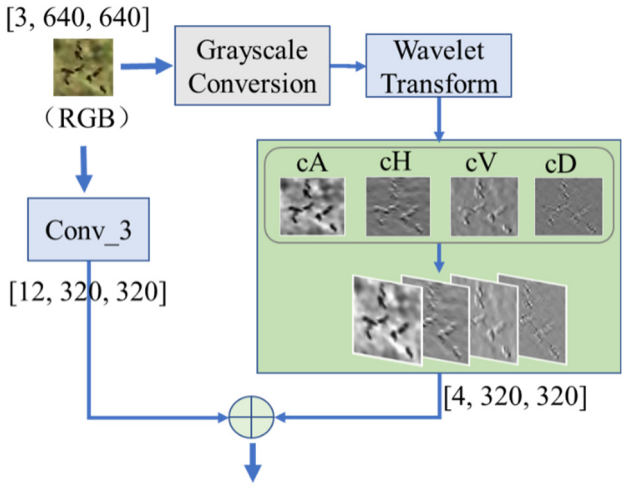


Fig 2. WT_Conv module

The four coefficients are of equal magnitude, with the width and height dimensions being half that of the original

image. They are combined in the channel dimension to yield an output with size [4, 320, 320]. Simultaneously, a 3×3 convolution is employed to transform the input vector from [3, 640, 640] to [12, 320, 320]. Finally, the outputs after both the convolution operation and wavelet transform processing are concatenated to obtain a vector of size [16, 320, 320], which serves as input for the subsequent layer.

2.2. MS Module

The YOLO model incorporates multiple convolution and downsampling operations, resulting in deep feature maps that encompass rich semantic information but lacks fine-grained details of small objects. Conversely, the shallow feature maps contain information about objects of varying scales. To address the deficiency in detail within the deep feature map, YOLOv8 employs the Neck structure depicted in Fig 3. However, this approach introduces scale-confusion due to directly fusing the feature maps F_3 and F_4 which come from shallow layers C_3 and C_4 with P_2 and P_5, respectively. And large-objects features in shallow feature maps will affect the detect of small objects.

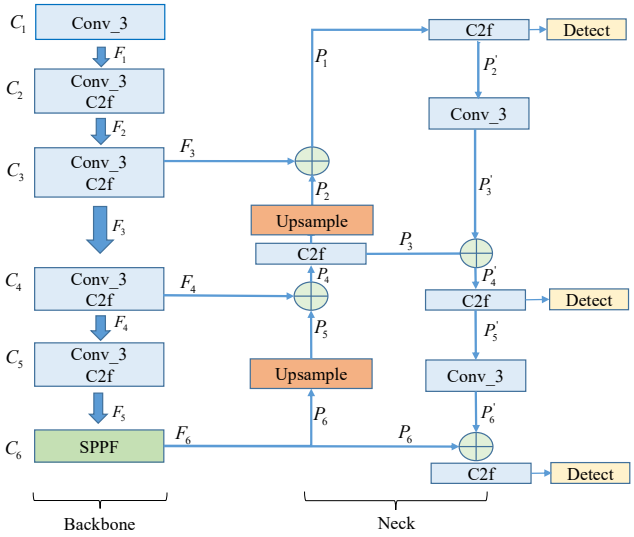


Fig 3. Structure of YOLOv8 model

Inspired by the idea of separating large object information from shallow layers mentioned in literature [6] to reduce scale confusion, the MS module is designed, as shown in Fig. 4. Adjacent feature maps contain complementary feature information of the same object. In order to avoid the feature information loss of scale overlapping parts caused by subtraction operation, the input of MS module is BackBone's hierarchical features Fi and Fi+2. Fi represents feature map from shallow layer, Fi+2 represents feature map from deep layer. The process of multi-scale separation is divided into four steps, of which (1) and (2) are shown in the yellow box in the figure:

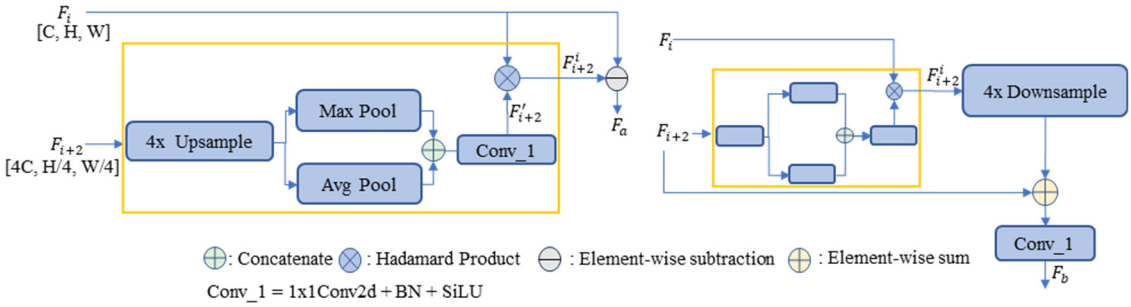


Fig 4. MS module

(1) First, F_{i+2} is upsampled to the same size as F_i. The two feature maps are spliced after maximum pooling and average pooling respectively, and then 1×1 convolution is performed to generate the gate F'_{i+2} of each channel. F'_{i+2} can adjust feature activation thresholds at different scales in the channel dimension.

(2) Perform a Hadamard product between the variable F'_{i+2} and F_i to obtain the feature map F^i_{i+2}. In this process, based on the gating signal F'_{i+2} generated by the deep feature information, the response intensity in the shallow feature map can be adjusted. Therefore, it is considered that F^i_{i+2} mainly includes large object features from F_i that need to be removed.

(3) By subtracting F^i_{i+2} from F_i, it means that the big target feature in F_i is removed, and the small target feature map F_a can be obtained.

(4) Downsample F^i_{i+2} to a size consistent with F_{i+2}. The two vectors are directly added for fusion, and then 1×1 convolution is used to promote cross-channel information exchange, which transfers the shallow large object feature to the deep neural network, and finally outputs the large object feature F_b.

Backbone of YOLOv8 contains four layers of the same structure, so this paper sets up two MS modules in the improved model, named MS-1 and MS-2. The separated shallow small object features F_{a1} and F_{a2} are the inputs of the two DE modules respectively. And the large object feature information output by MS-2 is transmitted to the Backbone deep layer as the input of C_6 layer.

2.3. DE Module

Compared with large and medium objects, it is difficult to detect small objects in pictures with low resolution and fuzzy details, which is more difficult to be represented and learned by neural networks. Therefore, a module DE designed to

enhance the features of small objects can fuse adjacent shallow features and indirectly alleviate the problem of unbalanced feature representation.

As shown in Fig.5, F_j and F_{j+1} come from two adjacent shallow layers, respectively, which contain much richer information about small objects than the deep feature maps. The whole process of detail enhancement is as follows:

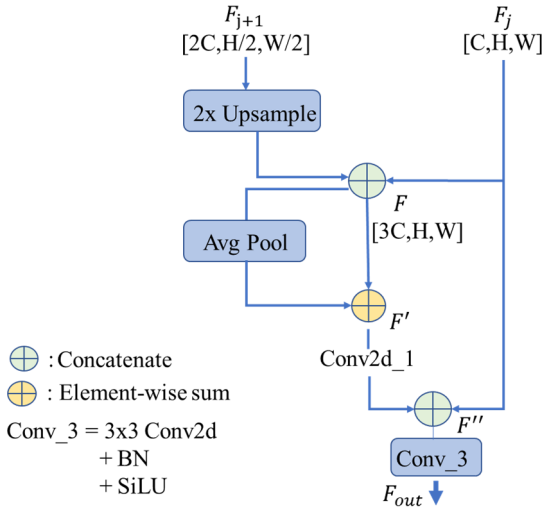


Fig 5. DE module

(1) F_{j+1} is up-sampled and then concatenated along the channel dimension to obtain feature F. F is augmented by adding itself after global average pooling, resulting in F', thus enhances the semantic information of each channel.

(2) The channel information of feature graph F' is fully fused using a 1×1 convolution, and then combined with F_j along the channel dimension to obtain F''. Considering that feature maps fusion may cause inconsistent expression of features [7], a 3×3 convolution kernel is employed to eliminate noise from the fused feature vector F'', yielding the final output as F_{out} of the DE module.

The improved model also incorporates two DE modules, named DE-1 and DE-2, which enhance the small object features from the two MS modules and the C_4 layer output. The DE module receives and enhances the feature information of small objects from C_2, C_3 and C_4 layers. The output of DE module can provide rich semantic and detailed information of small objects to the detection layer, thus improving the small object detection performance of the model.

3. Experiment and Result

3.1. Dataset Introduction and Processing

The experimental data are obtained from the open-source remote sensing dataset called UAVOD-10 [16] and the small object dataset [17] , both are divided into training, validation, and test sets according to 7:2:1. The UAVOD-10 dataset consists of a total of 844 images labeling 10 types of targets such as buildings, boats, vehicles, and wells, which are sufficiently varied in size, shape, and texture, and are disturbed by imaging conditions and complex environments.

The small target dataset Small Object has three categories of targets such as bees, flying insects, and ornamental fish, and the targets account for fewer pixels, and some of them have occlusion problems and inconsistent densities. Due to the limited number of images in the original dataset, three enhancement methods called Flip, Hue and Brightness are used to expand the dataset to 900 images, and the enhancement effect is shown in Fig. 6.

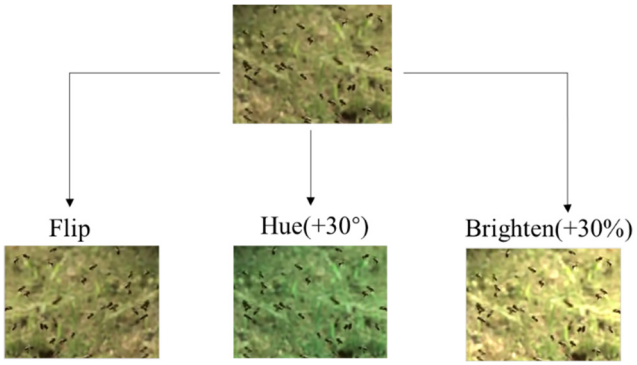


Fig 6. Examples of image enhancement

3.2. Experimental Parameter Settings

This paper builds a neural network model based on the PyTorch framework, using PyCharm 2021 Professional as the development tool, with an NVIDIA GeForce RTX 2060 GPU and 12GB of VRAM. The Python version is 3.8.18, and the PyTorch version is 1.12.1.

The experimental setup specifies a Batch Size of 16, with an initial learning rate of 0.01 and a final learning rate of 0.0001. Observing the loss curve in Fig. 7, it's noted that the validation loss begins to stabilize within the iteration range of 150 to 200, suggesting that the network training is sufficiently thorough. Consequently, the number of epochs is set to 200.

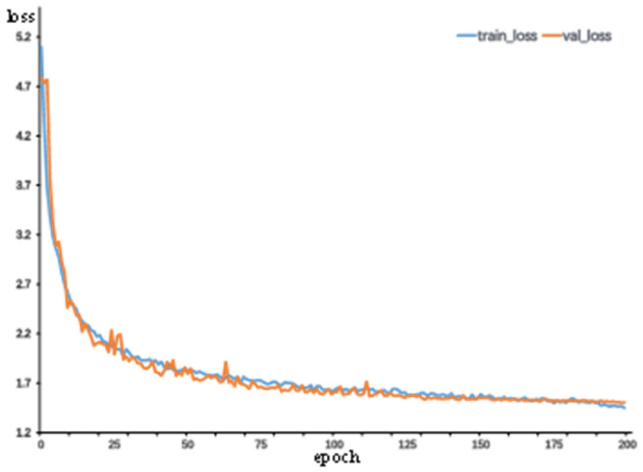


Fig 7. Training loss curve and validation loss curve

3.3. Analysis of Experimental Results

In this study, we've selected the YOLOv8n model as our benchmark and compared it with seven other mainstream algorithms, including the two-stage classic model Faster-RCNN, the single-stage models from the YOLOv5 to YOLOv8 series, and the small dense object detection models TPH-YOLOv5 [18], Drone-YOLO [19], and YOLOv8-BiFPN. Table 1 presents a comparison of the total detection accuracies for these models, as well as the detection accuracies for each category, based on the UAVOD-10 dataset.

Table 1. Comparison of the accuracy of each model on the UAVOD-10 dataset

Model	mAP/%	building	ship	vehicle	prefabricated house	well	Cable tower	pool	cultivation mesh cage	quarry
Faster-RCNN	24.3	32.4	6.4	0	13.4	8.2	25.9	38.9	42.7	50.9
YOLOv5n	27.4	61.4	5.4	0	70.4	4.4	5.8	16.6	41.8	40.7
YOLOv5s	33.6	67.9	6.3	0.1	76.9	5.1	18.5	24.9	55.5	46.8
YOLOv6n	26.7	59.4	9.2	0	74	22.3	2	1.4	30.8	32.4
YOLOv7-tiny	37.1	63.6	7.1	0	73.5	13.2	25.4	24.8	54.9	71.7
YOLOv8n	37.8	68.1	18.5	0.19	79.4	38.7	13.7	25.0	59.8	37.0
TPH-YOLOv5n	22.9	65.6	2.91	0	76	3.2	18.9	0	23.7	15.3
TPH-YOLOv5s	33.6	72	12	0.1	77.4	24.3	33.6	0	38.7	44.1
Drone-YOLO	38	69.3	35.1	0.1	80.9	42.1	28.3	19.2	41.8	25
YOLOv8-BiFPN	44.1	71.8	22.9	0.24	75.1	32.5	34.9	50.1	66.7	62.2
Ours	47.3	71.9	31.6	0.21	78.6	39.3	29	66.6	66.7	52.7

The improved model achieved a 9.5% increase in mAP compared to the benchmark model YOLOv8 on the UAVOD-10 test set. Different categories showed varying degrees of improvement in detection accuracy. Among them, the categories "pool" and "quary" have the largest pixel area and are easily have its features extracted by the WT_Conv module at different scales and directions, so their detection accuracies see the most improvement, with increases of 41.6% and 15.7%.

Following are categories of cable-tower and ship, with an average edge length proportion of only 0.02 to 0.03, whose detection accuracy increases of 15.3% and 13.1%. This is attributed to the MS module which prevent small object from being interfered with by other scales object and DE module which enhances small targets expression. The category with the smallest pixel area is vehicle, with an average side length ratio of 0.01, which is extremely difficult to observe even with the human eye and our model do not perform well on this category. our model achieves the highest mAP, which is 3.2%, 9.3%, 24.4% and 13.7% higher than other small object detection models like YOLOv8-BiFPN, Drone-YOLO, TPH-YOLOv5n and TPH-YOLOv5s, respectively and achieves optimal or suboptimal detection results in the categories of building, ship, well, pool, and cultivation-mesh-cage.

In order to prove the effectiveness of the improved model for Small Object detection tasks, experiments were conducted on a small object dataset named Small Object, and the results were shown in Table 2.

Table 2. Comparison of the accuracy of each model on the Small Object dataset

Model	mAP/%	fish	fly	honeybee
Faster-RCNN	69.6	77	59	72.8
YOLOv5n	88	96.6	71.6	95.7
YOLOv5s	91.3	97	80.1	96.9
YOLOv6n	88.5	94.9	77.7	92.9
YOLOv7-tiny	92.6	97.4	84	96.4
YOLOv8n	92.8	97.5	85.3	95.6
TPH-YOLOv5n	93.3	96.3	88.9	94.8
Drone-YOLO	94.7	97.6	91.5	95.1
YOLOv8-BiFPN	94	96	90	96.2
ours	95.1	97.9	91.5	95.8

Due to the simplicity of the Small Object dataset, the detection accuracy of the benchmark model has reached a higher level of 92.8%, so the improvement is smaller than that of the UAVOD-10 dataset, and the overall mAP improvement is 2.3%. Among them, the detection accuracy of fish, fly and honeybee three categories increased by 2.3%, 6.2% and 0.2% respectively. Among all the models involved in the comparison, the overall mAP value of the model proposed in this paper is the highest, which is 95.1%, and it achieves the best detection performance in the detection tasks of fish and fly, and also achieves good performance in the detection of honeybee. Based on the above analysis, the improved model presented in this paper shows excellent ability of small target detection.

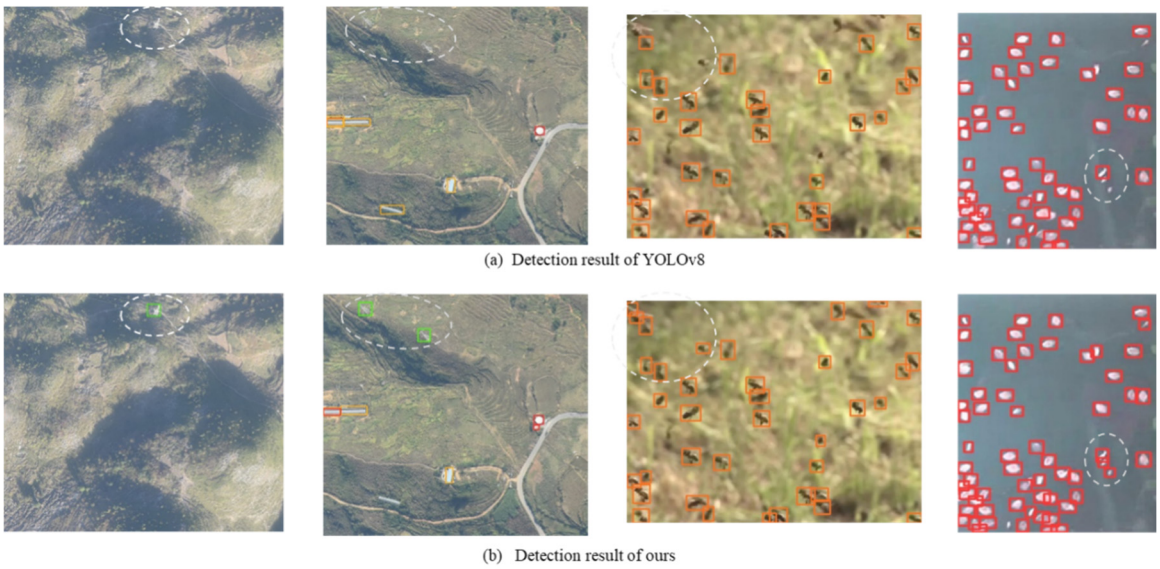
As shown in Fig. 8, the results of small object detection in part of the dataset were visually compared between the proposed improved model and the baseline model. To avoid the obstruction caused by the label text, only the detection box is drawn in the figure, and the difference between the two is marked with a white oval box. As you can see, the improved model successfully identified several small objects that the benchmark model had previously missed.

4. Discussion

4.1. Ablation Experiment

In order to verify the effectiveness of the improved algorithm model in this paper, ablation experiments were designed on the UAVOD-10 dataset for the proposed WT_Conv module, MS module and DE module, and the

results are shown in Table 3.



(a) Detection result of YOLOv8

(b) Detection result of ours

Fig 8. Comparison of detection results between improved model and baseline

On the basis of the baseline model, three models a, b and c were designed to constitute the ablation experiment. Model a adds WT_Conv module to the C_1 layer of the baseline model, which can extract more low-level features from different scales, enrich the representation of features, and its mAP is increased by 1.7%. Model b adds MS-1 and MS-2 modules based on a, and fuses P_2 and P_5 by downsampling the separated small target feature graph twice, and transfers the separated large object features to the deep layer of the network to reduce the interference to the small object feature information, and its mAP increases by 4.1%. The final model c is based on b, which adds DE-1 and DE-2 modules to strengthen the feature representation of small objects, and the mAP of the model is improved by 3.7%. As can be seen from Table 3, the detection accuracy of the proposed improved method has achieved the highest degree of improvement, and its effectiveness has been proved.

Table 3. The results of the ablation experiment

Model	WT_Conv	MS	DE	mAP/%
Baseline				37.8
a	√			39.5
b	√	√		43.6
C(Ours)	√	√	√	47.3

4.2. Algorithm Complexity Analysis

The complexity of the algorithm is also an important factor to consider when evaluating the performance of the model, which represents the operational efficiency and resource consumption of the algorithm. In order to evaluate the complexity of the improved model more comprehensively, the number of model parameters and the amount of computation are taken as evaluation indicators, and the improved model is compared with a series of models mentioned in the comparison experiment above. The results are shown in Table 4.

The results show that the parameters of the improved model are 4.4M and the calculation consumption reaches 19.3G, which is higher than that of the baseline model YOLOv8n, but the mAP is significantly improved. Compared with other models, the parameters of the improved model are at a moderate level, but the calculation cost is a little larger, but the accuracy of the improved model is obviously higher than other models. In summary, the improved model can greatly improve the detection accuracy of small objects with acceptable increase of parameters and calculation cost.

Table 4. Comparison of Model Complexity

Model	Magnitude of Parameters/M	FLOPs/G	mAP/%
Faster-RCNN	137	370	24.3
YOLOv5n	1.78	4.2	27.4
YOLOv5s	7.05	15.8	33.6
YOLOv6n	4.92	11.4	26.7
YOLOv7-tiny	6.03	13.1	37.1
YOLOv8n	3.16	8.9	37.8
YOLOv8-BiFPN	3.17	8.9	44.1
Drone-YOLO	3.4	17.6	38
TPH-YOLOv5n	2.32	6.2	22.9
TPH-YOLOv5s	9.2	23.3	33.6
Ours	4.4	19.3	47.3

5. Conclusion

In order to solve the problem of poor recognition of small objects caused by scale confusion, an improved YOLOv8 model is proposed in this paper. Firstly, the WT_Conv module is designed in the model, which uses wavelet transform to extract low-level features and increase the feature information. Secondly, the MS module is designed to separate the features of objects at different scales, so as to effectively deal with the problem of inconsistent feature expression caused by the direct fusion of deep and shallow feature maps. At the same time, a DE module is designed to fuse the separated shallow feature maps to enhance the expression of small object features. Experimental results on UAVOD-10 and Small Object datasets show that the detection performance of this model is significantly improved compared with the baseline model, and it also has greater advantages compared with other comparative network models.

However, the MS module proposed in this paper simply

classifies the detected objects according to their sizes, while the scale range of various target objects in the real scene is varied. In the future, we can try to design lighter model structures in the case of finer scale division.

Acknowledgments

I am deeply grateful for the support of my friends and colleagues during the writing of this thesis. Their encouragement and feedback were crucial. I also extend my thanks to the providers of the Small Object and UAVOD datasets for their invaluable contributions to my research. Additionally, I appreciate the ideas and insights from all authors whose works I have cited, whose scholarship has significantly influenced my own.

References

[1] BOCHKOVSKIY A, WANG C Y, LIAO H Y M.YOLOv4: Optimal Speed and Accuracy of Object Detection [EB/OL]. 2020. https://arxiv.org/ abs/2004.10934.

[2] KISANTAL M, WOJNA Z, MURAWSKI J, et al. Augmentation for Small Object Detection [EB/OL]. 2019. https: // arxiv.org/abs/1902.07296.

[3] LI Li-xia, WANG Xin, WANG Jun, et al. Small object detection algorithm in UAV image based on feature fusion and attention mechanism[J]. Journal of Graphics, 2023,44(04):658-666.

[4] LI Qingyuan, DENG Zhaohong, LUO Xiaoqing, et al. SSD Object Detection Algorithm with Attention and Cross-Scale Fusion[J]. Journal of Frontiers of Computer Science and Technology, 2022,16(11): 2575-2586.

[5] MA J, SHAO W, YE H, et al. Arbitrary-Oriented Scene Text Detection via Rotation Proposals[J]. IEEE Transactions on Multimedia, 2017,99: 1-1.

[6] LI Y, PANG Y, SHEN J, et al. NETNet: Neighbor Erasing and Transferring Network for Better Single Shot Object Detection [J]. IEEE, 2020.

[7] HUANG S, LIU Q. Addressing Scale Imbalance for Small Object Detection with Dense Detector[J]. Neurocomputing, 2022, 473:68-78.

[8] LIU S, QI L, QIN H, et al. Path Aggregation Network for Instance Segmentation[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768.

[9] TAN M, PANG R, LE Q V. Efficientdet: Scalable and Efficient Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10781-10790.

[10] GHIASI G, LIN T Y, LE Q V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7036-7045.

[11] REDMON J, FARHADI A, YOLO9000: Better, Faster, Stronger [C] //IEEE Conference on Computer Vision & Pattern Recognition, Honolulu: IEEE Press, 2017: 6517-6525.

[12] B W, Y J, C Y J. Beyond Deep Residual Learning for Image Restoration: Persistent Homology-guided Manifold Simplification [C]// Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 145-153.

[13] HUANG H, HE R, SUN Z, et al. Wavelet-SRNet: A Wavelet-Based CNN for Multi-scale Face Super Resolution[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[14] LI Lin, JIN Zhixin, YU Xiaolei, et al. Road Vehicle and Pedestrian Detection Based on YOLOv9 for Haar Wavelet Downsampling [J/OL]. Computer Engineering and Applications, 2024:1-9.

[15] HE Guohuan, ZHU Jiangping. WT-U-Net++: Surface Defect Detection Network based on Wavelet Transform[J]. Journal of Computer Applications, 2023, 43(10): 3260-3266.

[16] HAN W, LI J, WANG S, et al. A Context-scale-aware Detector and a New Benchmark for Remote Sensing Small Weak Object Detection in Unmanned Aerial Vehicle Images[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 112:102966.

[17] ZHENG M, LEI Y, CHAN A B. Small Instance Detection by Integer Programming on Object Density Maps[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3689-3697.

[18] ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778-2788.

[19] ZHANG Z. Drone-YOLO: an Efficient Neural Network Method for Target Detection in Drone Images[J]. Drones, 2023, 7(8): 526.