Current Science
CS 5 (6), 5439-5453 (2025)
Received 02 Nov 2025 | **Accepted 17 November** 2025 | **Online Available 24 November** 2025
https://doi.org/10.52845/CS/2025-5-6-13

Current Science

Current Science

OPEN ACCESS

ISSN (O) 2795-8639

Original Article

Research on an Improved Small Target Detection Algorithm for YOLOv8

Tinghang Guo[1,2], Xiaohan Li[1]*, Xin Ji [1], Zuanping Qin[12], Guangda Lu[12], Yu Han[1,2], Runze Li[1,2]

[1]School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China

[2]Tianjin Key Laboratory of Information Sensing & Intelligent Control, Tianjin University of Technology and Education, Tianjin 300222, China

***Corresponding Author: Xiaohan Li**

Abstract:
In order to solve the problems of missed detection and low detection accuracy due to the changes in the shape, appearance and position of objects in the imaging process, as well as the complex influence of lighting conditions and occlusion factors, an improved YOLOv8 algorithm based on Swin Transformer was proposed. By introducing modules including Focus, deeply separable convolutional DwConv, c2, etc., the computation and parameters are reduced, the receptive field and feature channels are increased, and the Swin Transformer module is used to extract visual features to capture the context information of small target objects to enhance the feature representation. In addition, the loss function of the original network is replaced by the WIOU loss function to optimize the model and improve the accuracy of small target detection. Comparative experiments based on public datasets show that compared with the YOLOv8 algorithm, the improved model has improved the accuracy of small target detection P, recall R and average accuracy of mAP@0.5, which enhances the ability of intelligent robots to identify small targets in complex environments and provides important support for the technological progress of related industries.

Keywords: small target detection; deep learning; improved YOLOV8 model; Loss;

1. Introduction

With the continuous change of computer vision technology, deep learning-enabled target detection methods for intelligent traffic monitoring, industrial quality inspection, automated driving, mobile payment and aerial photography and mapping have realized large-scale applications, providing strong basic technical support for research in related fields [1-4]. The current mainstream detection architectures are mainly divided into two categories: two-stage and one-stage, in which the two-stage detection algorithms adopt the progressive processing flow of "region generation-feature extraction-categorization decision", and the typical representatives include the RCNN series (RCNN [5], Fast RCNN [6], and Feature Pyramid Network (FPN [7])), and the RCNN series. . This architecture effectively improves the detection accuracy through the candidate region screening mechanism, but it consumes more than 200ms on average when processing 1080P images, which limits the real-time performance. Further, single-stage detection algorithms, using end-to-end detection mode, typical YOLO series [8] and SSD [9] achieve response through multi-scale feature fusion, which has obvious advantages in speed and computational efficiency and is suitable for application scenarios that require real-time

detection, but is limited by the feature resolution attenuation, and the detection accuracy of the small targets is significantly reduced.

Traditional target detection techniques mainly contain three core aspects: region selection, feature analysis and classification decision. However, region sampling-based methods suffer from sample redundancy and insufficient targeting, resulting in low computational efficiency and limited detection performance. Although current AI techniques have significantly improved the overall level of target detection, the recognition of small targets still faces serious challenges. Such targets are usually characterized by two types of features: one is physical size limitation, i.e., the projected area of the target in the image plane is generally lower than the 32*32 pixel threshold; the other is spatial occupancy limitation [10], which refers to the proportion of the target in the original image that does not reach the preset baseline value of the object. No matter what kind of definition criteria, small targets face difficulties such as weak feature characterization ability, significant background noise interference, and stringent spatial localization accuracy requirements, which pose multiple challenges to the robustness of the detection system. It is worth noting that deep learning-based small target detection algorithms have gradually replaced traditional algorithms, showing breakthroughs in industrial inspection, remote sensing and mapping, medical imaging, and other application scenarios Deep learning-based small target detection models have gradually replaced the traditional model, Wang et al. proposed a deep feature extraction network based on residual modes [11], which achieved better detection accuracy and robustness in infrared small target detection.Bai et al. utilized an improved multi-target tracking algorithm [12] to reduce omissions caused by occlusion. Due to the success of the YOLO series, many researchers improved the algorithm based on YOLO.Li and Shen proposed an infrared small target detection method based on super-resolution and deep learning [13], which improved the accuracy of small target detection.Shen et al. introduced a deformable convolutional C2f (DCN_C2f) module [14] based on YOLOv8 for self adaptive network field adjustment.Zhang et al. proposed a YOLOv7-tiny-based algorithm for small target detection in UAV aerial images [15].

With this background, this paper carries out an improvement study based on the YOLOv8 algorithm and proposes the DFSTF-YOLOv8 model.Firstly, the Focus and DWconv modules are added [16], which can capture huge contextual information and understand the structure and characteristics of the background; meanwhile, the C2 module is introduced to help the model perceive and analyze the small-target objects, and to improve the model's ability of multiscale; The feature fusion is increased by introducing Swin transformer [17], which solves the problem of its lack of contextual and semantic information; the WIOU [18] loss function is used to replace the CIOU loss function, and the focus of the localization loss is modified to enhance the convergence and generalization ability of the YOLOv8 model; finally, the effectiveness of the algorithm is verified through experiments.

1 Optimised Design of YOLOv8 for Small Target Detection

The detection head of YOLOv8 adopts the current mainstream decoupled head structure, which separates the prediction branch from the regression branch, and the two branches use different loss functions, using anchor-free instead of anchorbased.YOLOv8 is combined with Swin Transformer, and introduces the Focused Attention, Depth Separable Convolution , and C2 module, the advantages brought by this optimisation are as follows, 1) High localisation accuracy: in complex background detection, small target detection is crucial for security monitoring, unmanned vehicles, precision instrument

detection and other fields. And it helps to improve the detection accuracy and reduce the false alarm rate.2) Real-time performance: the YOLO series of algorithms are known for their high efficiency, which is suitable for scenarios requiring real-time feedback, such as video surveillance and autonomous driving, ensuring a balance between processing speed and detection quality.3) Good robustness: the model still maintains good detection performance under changing environmental conditions, which is especially important for outdoor applications. Fig. 1 shows the network model proposed in this paper.

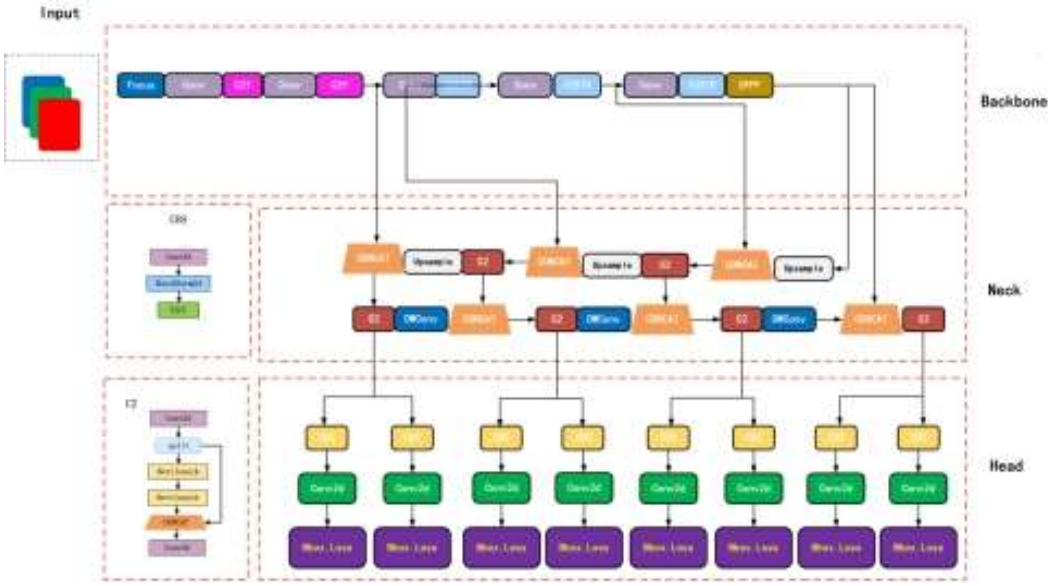


Fig.1:Optimized YOLOv8 architecture

1.1 Dwconv Module

For small target object detection, due to its small size, the convolution operation in traditional convolutional neural networks usually involves a large number of parameters and computations, and its spatial positional features are easily lost after multiple convolution and pooling operations. In this paper, we introduce the Dwconv module, which is used in equipment assembly in industrial contexts, where robots usually need to deal with a large number of small target objects for detection. by decomposing the standard convolution into deep convolution and point-by-point convolution, Dwconv is able to reduce the amount of computation while maintaining sufficient feature expression capability, which helps the model to better capture the positional features of the small targets, allowing the model to be more lightweight while maintaining the accuracy. more lightweight, making it possible to run on devices with limited computational resources and increase processing speed. This approach enables the model to effectively learn the spatial and channel information in the input feature map, which enhances the model's expressive and perceptual capabilities, and also improves the robot's intelligence level and operational efficiency.

In this paper, the original Conv is replaced by DWconv in the Neck part to make the model more lightweight while maintaining sufficient feature capability of spatial position information, thus enabling the model to better understand the structure and features of the background and achieve accurate localisation of small target objects, while the inference speed of the model is improved to better meet the needs of mobile and embedded devices, and the module is also more suitable for the application scenarios that require

lightweight and high-speed reasoning.The DWconv schematic is shown in Figure 2.

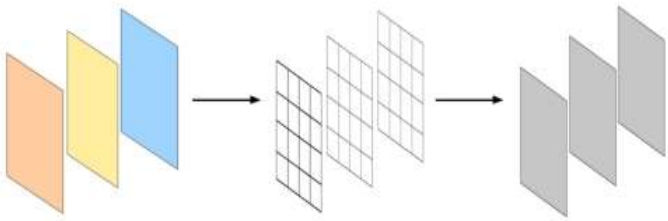


Fig.2:Depth separable convolution schematic diagram

1.2 Focus Module

Small target objects occupy fewer pixels in the image, and direct feature extraction may lead to information loss.Focus module is an attention mechanism module designed for object detection tasks, which increases the number of channels by slicing operations on the image, enabling the network to better capture the detailed information features of small targets.

In this paper, the Focus module is introduced in the YOLOv8 backbone network, which improves the model's attention to important feature regions by applying a lightweight convolution operation to the input feature map, which is subsequently partitioned into multiple smaller sub-regions. These sub-regions are then weighted and fused by learning the generated weights to generate the final attention feature map. Enabling the model to effectively focus on important feature regions. The schematic diagram of the focusing module is shown in Fig. 3.

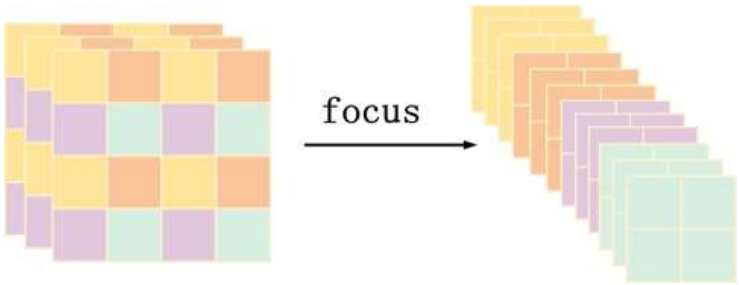


Fig.3:Principle diagram of focusing module

1.3 C2 Module

The C2f module in the original YOLOv8 network performs segmentation and fusion of the feature maps at an early stage of the network, which may cause the feature information of small targets to be lost or insufficiently salient in the pre-processing and may create the problem of gradient vanishing or gradient explosion on small target objects. The C2 module, on the other hand, extracts high-level semantic features and enhances their representativeness to improve the performance of computer vision tasks. It outperforms the C2f module in terms of memory footprint and inference speed.

From Fig. 4 it can be seen that the C2 module splits the input feature map into two parts. These two parts of the features are subjected to separate convolution operations, one part is directly convolved and the other part may go through some additional convolutional layers. The processed two parts of the features are merged by element summing or other fusion methods to generate a new feature map.

In this paper, after replacing the C2f module with

the C2 module in the Neck part, the network processes the features of small targets more directly and efficiently, which reduces the complexity of the feature fusion, and allows the network to focus more on the small but important features and improves the gradient flow, which improves the accuracy and efficiency of the detection of the small target objects, and makes the training of the small target objects more stable.The structure of the C2 module is shown in Fig. 4.

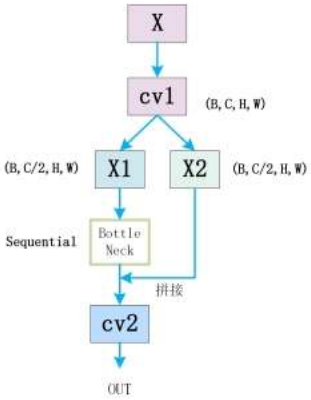


Fig.4 : C2 module structure diagram

1.4 C3STR Module

For the problem that most of the object features that the small target objects in the image are supposed to have are constantly lost in multiple convolution operations, therefore, this paper introduces the C3STR module into the backbone network of the YOLOv8 algorithm, which borrows the idea of Swin Trasformer in the operation of feature fusion. The core idea is to establish the global dependence on different spatial locations of the feature map through the self-attention mechanism, and to enhance the semantic information and feature representation of small target objects with the help of the window self-attention module. It is able to perform adaptive feature interaction for each location in the feature graph to capture the contextual information around the object. The module contains paired Window Multihead Self-Attention Modules, Sliding Window Multihead Self-Attention Modules, and Multi-Layer Perceptron Mechanisms, and each of them is internally connected using residual connections. In industrial scenarios, for small target detection under precision equipment assembly, this module improves the resolution of small target detection through local window processing without sacrificing too much computational resources, better captures the fine features of small targets, and reduces false and missed detections.The structure of C3STR is shown in Fig. 6. The computational process of the multi-head self-attention mechanism is as follows.

$$Attetion(Q,K,V) = SoftMax(QK^T/\sqrt{d} + B)V \quad (1)$$

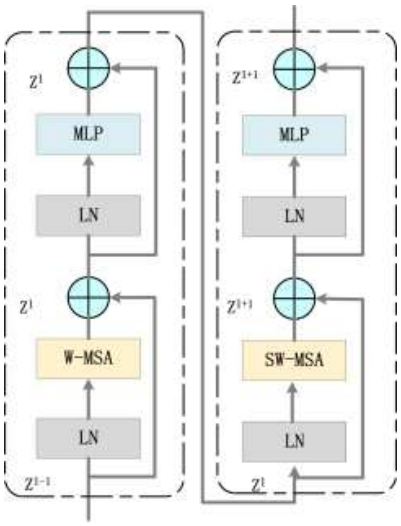


Fig.5: C3STR structure diagram

1.5 WIoU Loss Function:

By analysing the loss functions used in the YOLO series of algorithms, this study found that the localisation loss function is usually a mean square error function. This function is regressed using only the coordinate and aspect information of the predicted and actual frames. Therefore, this approach is considered to be one of the main reasons for the instability observed during model training. Considering the inherent characteristics of target detection, this study needs to identify the predicted frame that has the highest degree of overlap with the actual target frame. This criterion most effectively meets the goal of target detection [23]. Therefore, this paper proposes to shift the localisation loss function from measuring the distance between predicted and real frames to evaluating the intersection and integration set (IOU) between them. This modification aims to better meet the basic requirements of target detection. Given the presence of low-quality examples in the training data, it is important to consider the impact of geometric factors (e.g., distance and aspect ratio) on the penalties associated with these examples. These factors can exacerbate the negative impact of low-quality examples and ultimately lead to a degradation of the model's generalisation performance. Traditional IOU calculation methods only consider the ratio of intersections and crosstalk between detection frames without considering their positional relationship in the image. A well-designed loss function should mitigate the effect of geometric factors when there is an overlap between anchored frames and aspect ratios. In the case where the anchored frame is closely aligned with the target frame, the effect of geometric factors on the penalty should be reduced. In addition, minimising interventions during training will improve the generalisation ability of the model. In order to improve the accuracy of measuring the coverage of the detection frame, some researchers have proposed a method called 'Wise IOU' (WIOU) [19]. Based on this, the construction of distance attention relies on the utilisation of the distance metric. This model replaces the original loss function with the WIOU loss function, which consists of a two-tier attention mechanism with the following formula:

$$L_{WIOUwi} = R_{WIOU} L_{IOU} \quad (2)$$

$$R_{WIOU} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \quad (3)$$

$$L_{IOU} = 1 - IOU = 1 - \frac{W_i H_i}{S_u} \quad (4)$$

Eq. Significantly amplifies the , significantly reduces the Rmov of normal-quality anchor frames, and reduces the emphasis on centre-of-mass distances when anchor frames overlap well with object frames. To mitigate the convergence barrier caused by gradients in Rmov , the computational mapping separates Wg and Hg (denoted by superscript *). In the literature [13] the authors propose a method that effectively solves the obstacles to convergence. The introduced method does not introduce any new metrics. Instead, it utilises intersection and concatenation (IOU) to quantify the level of overlap between predicted and real frames in the object detection task. The overlap region depicted in Fig. 7 is measured by the IOU metric and the area is denoted as:

$$S_u = wh + w_{gt}h_{gt} - W_i H_i \quad (5)$$

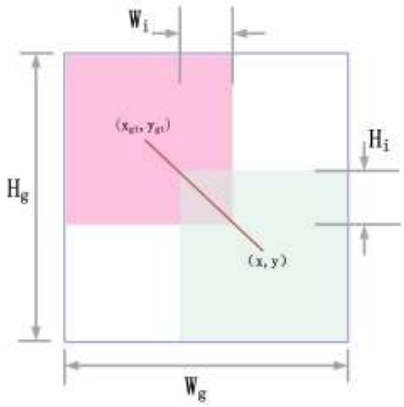


Fig.6:The overlapping area between the predicted and actual boxes

2 Experimental Results and Analysis

In order to evaluate the effectiveness and accuracy of the model in recognising small target objects. In this study, the FloW-Img dataset is selected for experimental validation. FloW-Img [20] is the world's first dataset for surface floating trash detection based on unmanned vessel viewpoints, which contains 2,000 images and 5,271 labelled targets. In this dataset, small-size targets (size less than 32×32 pixels) occupy a large proportion, so it is very suitable as a dataset for small target detection. The images are divided into training and test sets in the ratio of 8:2. In order to expand the applicability of the model, the validity of the model was verified in the customised dataset and VisDrone2019 dataset [21], respectively, where the two commonly used everyday objects in the customised dataset are apples and small scissors, which represent regular and irregular objects, respectively. The custom dataset consists of images of apples and scissors, and the images are enhanced with the data enhancement technique, which takes 35 original custom images and enhances them to a total of 395 images using the data enhancement technique, which is suitable for target detection experiments in a laboratory environment.VisDrone2019 is a high-resolution image dataset of aerial photographs taken by a

drone, with an image resolution of 2000×2000×1000 pixels. with an image resolution of 2000 × 1500.The training portion of this dataset contains 6471 images with a total of 343205 labels, with an average of 53 instance objects per image, and most of the objects are very small in size (less than 32 pixels), which makes it ideal for dataset studies on dense small target detection.

All the labelled information of the dataset is converted to TXT format for subsequent model training. Mosaic data augmentation is used throughout the training process to enrich the training samples and enhance the generalisation performance of the model. The commonly used evaluation indexes in the field of deep learning are used for evaluation as shown in Table 1, and the higher value of each evaluation index represents the higher accuracy of the model.

Tab.1: Symbols and formulas for all evaluation indicators

Symbol	Meaning	Assessment indicator	formula
TP	Numberof samples predicted to be positive by the model that are also actually positive	Precision	$P=\frac{TP}{FP+TP}$ Precision is the proportion of true positive samples out of all samples predicted as positive by the model. High precision is defined as a high proportion of true samples among the positive predictions made by the model. $R=\frac{TP}{FN+TP}$ Recall is the ratio of the number of samples correctly identified as positive categories by the model to the number of all samples identified as positive categories. It measures the ability of the model to identify positive categories. A high recall means that the model does a good job of finding all true positive samples. $F1=\frac{2\,\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$ The F1-score combines the accuracy and recall of the model. It takes values between 0 and 1, and the closer it is to 1, the better the model's performance.
FP	Number of samples predicted as positive by the model but actually negative	Recall	
TN	Number of samples that are predicted to be negative by the model and are actually also negative	F1-score	
FN	Number of samples predicted by the model to be negative and actually positive		
IOU	Calculate the ratio of the intersection area to the concatenation area of two bounding boxes to assess the degree of overlap	Union ratio IOU	$\mathrm{IOU}=\frac{A\cap B}{A\cup B}$ The IOU intersection and join ratio is a quantitative measure used to assess the degree of overlap in the results of an object detection algorithm. The calculation involves comparing the area of the detection result with the intersection and splice of the area of the actual label. The intersection and concatenation ratio can be defined as the quotient of the intersected area and the spliced area. The accuracy of the detection results increases as the value of the

			IOU calculation formula increases.
MAP0.5	Mean accuracy ratios calculated with an IOU threshold of 0.5	Average Precision	$AP = \int_0^1 P(r)dr$ $mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$ Average Precision (AP) is calculated as the average precision of each recall point and represents the area under the precision-recall (P-R) curve. The mean accuracy (mAP) is calculated as the average of the mean accuracy (AP) values for all categories N.
MAP0.5:0.95	Mean precision averages calculated at IOU thresholds from 0.5 to 0.95 (in 0.05 intervals)	Mean Average Precision Mean MAP	

The experimental environment used in this paper is using PyTorch based on PyTorch 1.10.1, Cuda11.3,, with NVIDIA GeForce RTX 3090 for the graphics card, Ubuntu 20.04 for the operating system, and Python 3.9 for the programming language.The number of training iterations used on both the customised dataset and the public dataset is 300 epochs, and in the The training phase is optimised using Stochastic Gradient Descent (SGD) algorithm.The SGD algorithm uses an initial learning rate of 0.01, a momentum factor of 0.937, and a weight decay factor of 0.0005.The input image is normalised to a size of 640×640.

2.1 Analysis of Experimental Results

2.1.1 Ablation Experiment

In this paper, a series of improvements are made on the original YOLOv8 model, and then ablation experiments are conducted to explore the enhancement effect brought by each module, the dataset is adopted in the FloW-Img dataset, the specific module improvements and experimental results are shown below.

1 WIOU Improvement

In order to further verify the advantages of WIOU [22] over other bounding box loss functions, this paper optimises the original loss function calculation method of YOLOv8 to make it applicable to WIoU and conducts comparative experiments. The experiments are all compared on the YOLOv8 model, and the specific experimental results are shown in Table 2. The recall of the model is improved by 4.8 percentage points, the precision is improved by nearly 2.9 percentage points, the mAP at 0.5 IoU is improved by 3.7 percentage points, and the overall evaluation index F1 score is improved by 4 percentage points.

Tab.2 : Experimental results of YOLOv8+WIOU (FloW-Img dataset)

modelling	Precision(%)	Recall rate (%)	mAP@0.5(%)	mAP@0.5:0.95(%)	F1-score
YOLOv8	79. 5	71.5	78.6	37.7	75.2
YOLOv8+WIOU	82.4	76.3	82.3	39.8	79.2

2 Validity of Attention Mechanisms.

In order to demonstrate that the proposed FOCUS is superior to the currently used attention mechanisms, we experimentally validate it in YOLOv8 using different attention mechanisms. The detection results of YOLOv8 when different attention mechanisms are introduced are shown in Table 4. It is obvious from Table 3 that the model works best when FOCUS is introduced as the attention mechanism in YOLOv8. The values of F1-score, mAP0.5%, and mAP0.5:0.95% are nearly 5.7%, 5.5%, and 4.6% higher, respectively, after the introduction of FOCUS than without

FOCUS. with an FPS of 625 (frames per second), the FPS is also an important performance metric, which indicates how fast the model is able to process and output video frames. A high FPS detection system provides more real-time monitoring and analysis capabilities. Moreover, the slicing operation is performed before the image enters the backbone of the network, resulting in efficient feature extraction. This design helps to reduce computation and memory consumption while improving detection performance. In summary, the introduction of FOCUS in YOLOv8 is effective and beneficial.

Tab.3 : Detection results of YOLOv8 when different attention mechanisms are introduced (FloW-Img dataset)

modelling	Precision(%)	Recall rate (%)	mAP@0.5(%)	mAP@0.5:0.95(%)	FPS (frames per second)	F1-score
YOLOv8	83.5	71.5	78.6	37. 7	410.6	75.2
YOLOv8+GE	84.2	77.2	83.4	41.8	625	80.5
YOLOv8+SE	85.8	76.2	83.8	41.5	625	80.7
YOLOv8+FOCUS	**85.9**	**76.4**	**84.1**	**42.3**	625	80.9

3 Effectiveness of Each Module in This Algorithm

In order to evaluate the effectiveness of each module in this algorithm, ablation experiments are conducted based on the custom dataset and the public dataset respectively to explore the enhancement effect brought by each module, and the results are shown in Table 5, in which DW, and STF denote the two new modules proposed in this paper, WIOU is the loss function of the replacement, and √ denotes that the module is being used. For the convenience of description, we name the models adding new modules in Table 4 as S1, S2, S3, and this algorithm in order. The experiments were conducted to evaluate the performance using metrics such as accuracy (%), recall (%), mAP@0.5 (%), mAP@0.5:0.95 (%), and F1 score. The results show that the mAP is increased from 78.6% to 86.2% using STF module and DW module combined.

Tab. 4: Ablation experimental results (FloW-Img dataset)

modelling	FOCUS	DW	WIOU	STF	Precision (%)	Recall rate (%)	F1-score	mAP@0.5(%)	mAP@0.5:0.95(%)
YOLOv8					83.5	71.5	75.2	78.6	37. 7
S1	√				85.9	76.4	80.8	84.1	42.3
S2	√	√			84.4	80.2	82,2	85.1	42.9
S3	√	√	√		85.8	79.2	82.3	85.9	43.3
ours	√	√	√	√	85.2	**81**	**83**	**86.2**	**44.1**

2.1.2 Comparison of the Performance of the Present Algorithm with Other Mainstream Algorithms

In order to further test the performance advantage of the present algorithm in small target detection, this paper compares the improved network model with other mainstream target detection models in

a comparative experiment. As shown in Table 5, the accuracy of the present algorithm is 85.2%, the recall is 81%, the mAP is 86.2% at 0.5 IoU, the mAP is 44.1% at 0.5:0.95 IoU, and the F1-score is nearly 83%. Improved algorithmic model achieves higher mAP@0.5 Although the YOLOv8 model is a little lower than the most primitive YOLOv3 model in all metrics, its network structure is simpler than that of YOLOV3. In addition, YOLO with the addition of the STF [23] module shows significant improvement in all target detection metrics and outperforms the previous YOLO model.

Tab.5 : Comparison of Different Algorithms (FloW-Img Dataset)

modelling	Precision(%)	Recall rate (%)	mAP@0.5(%)	mAP@0.5:0.95(%)	F1-score
YOLOv3	84.8	76.3	83	42.5	80.3
YOLOv5	81.6	71.4	78	36.8	76.1
YOLOv8	83.5	71.5	78.6	37．7	77
YOLOv8_CA_BOT3_ghost	85	75.4	83	40.7	79.9
YOLOv8_DW_CA_BOT3_focus_C2	85.3	75.4	3.1	40.1	80
ours	**85.2**	**81**	**86.2**	**44.1**	**83**

In order to further verify the effectiveness of the improved model, this paper did comparison experiments on other datasets, the results are shown in Table 6, from the experimental results can be seen compared to YOLOv8 algorithm, the present algorithm in the other two datasets have a different degree of improvement On a customised dataset mAP@0.5 increased 1.9%; in the small target is more intensive in the VisDrone2019 dataset mAP@0.5 increased 12.7% and compared to the literature [24] The algorithms in the mAP@0.5 increased 6.8%, these significant results show that the present algorithm has achieved significant improvement in detection accuracy. Therefore, the present algorithm proves to be suitable for accurate detection of small target objects and offers a broad application prospect in this field.

Tab.6 The detection results of this algorithm on other datasets

dataset	modelling	mAP@0.5(%)	mAP@0.5:0.95(%)
Customised data sets	YOLOv8 ours	88.6 90.5	86.6 87.6
VisDrone2019	YOLOv8 Literature Algorithm [25] ours	31.9 37.8 **44．6**	18.4 22.4 **27.2**

In order to get a better feel for the data, this paper visualises the data as a bar chart (Figure 7)

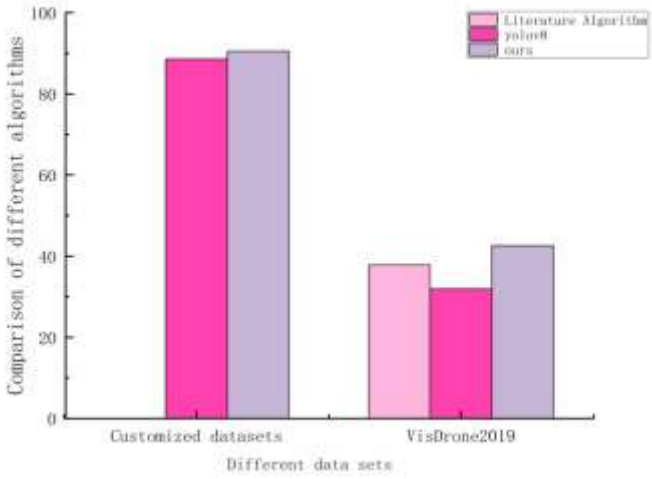
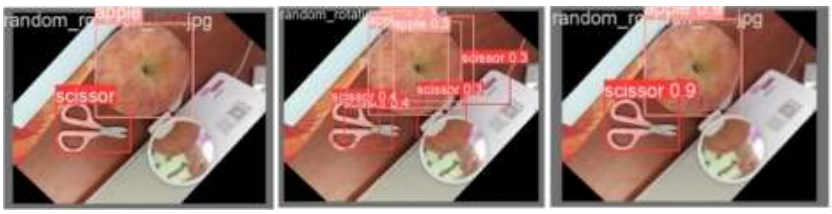


Fig.7: Comparison results of different datasets

In order to be able to intuitively feel the effect of the improved algorithm, the improved algorithm of this paper, and the prediction result images of the original algorithmic framework of YOLOv8 are compared, and three sets of complex scene images are selected from the custom dataset, FloW-Img dataset and VisDrone2019 dataset for testing. The experimental results are shown in Figures 8 to 10. the left image shows the detection results of the original algorithm. The right image shows the detection effect of the improved algorithm. In Figure 8, for the detection of irregular objects in the self-defined dataset, the original algorithm has omission and re-detection for irregular small targets, and the improved algorithm improves it to some extent; in Figure 9, the original algorithm successfully identifies the small targets in the far distance, and there is omission for the small objects in the sparse background in the near distance, while the improved algorithm can successfully detection; Figure 10 for the dense target scene, the improved algorithm can detect more targets than the original algorithm. In summary, the effectiveness of the improved module can be proved.



a) Real bounding box b) YOLOv8 algorithm c) This algorithm

Figure8 : Comparison of detection results of irregular small objects in a custom dataset

b) Real bounding box b) YOLOv8 algorithm c) This algorithm

Figure.9 : Comparison of sparse small target object detection results of FloW-Img public dataset

c) Real bounding box b) YOLOv8 algorithm c) This algorithm

Figure.10 : Comparison of detection results in dense scenarios of VisDrone2019 public datasets

3 Summary and Prospect

Aiming at the problems of misdetection, re-detection and low detection accuracy in the small target detection task, the self-defined dataset and the public dataset are adopted as the small target dataset, which contains richer background information, which can help the model learn more robust background suppression ability, this paper, on the basis of YOLOv8 algorithm, makes a series of improvements, and proposes the DFSTF for small target detection- YOLOv8 model. The specific improvements are as follows:

(1) Replacing the original Conv with DWconv in the Neck part makes the model lighter while maintaining sufficient feature capability, so that the model better understands the structure and features of the background and achieves accurate localisation of small target objects.

(2) Introducing the Focus module into the YOLOv8 trunk section, which improves the model's focus on important feature regions by applying lightweight convolution operations to the input feature map.

(3) Replacing the C2f module with the C2 module reduces the complexity of feature fusion and makes the network more focused on small but important features, which improves the detection accuracy and efficiency of small target objects.

(4) The C3STR module is introduced into the backbone network of the YOLOv8 algorithm to provide state-of-the-art accuracy by effectively combining transformer-based attention to enhance feature representation while using an efficient convolutional design to maintain reasonable detection speeds and introducing WIoU loss.

In order to improve the generalisation of the model to various datasets, the improved algorithmic model is applied to two other datasets with significant improvements, however, there are limitations in terms of model size and complexity. The introduction of other modules resulted in the DFSTF-YOLOv8 model not being lightweight enough for a highly resource-constrained environment. To address the above issues, future work will focus on further compressing the model design on the one hand and, based on this, further improving the accuracy and reducing the inference time of the model and exploring supplementary data sources. On the other hand, for the small target detection task in the physical world, the robot will identify and grab the target objects in the self-defined dataset as a way to improve the effectiveness and accuracy of this algorithm in the physical world.

Tinghang GUO：Responsible for the supervision of thesis

Xiaohan LI（corresponding author）：2295915597@qq.com, wrote the main text of the manuscript and drew Figures 1-10 and all tables.

Xin JI: Responsible for the supervision of thesis

Zuanping QIN: Responsible for the supervision of thesis

Guangda LU: Responsible for the supervision of thesis

Yu HAN: Responsible for the supervision of thesis

Runze LI: Responsible for the supervision of thesis

All authors reviewed the manuscript.

Funding Declaration:

Tianjin Municipal Education Commission Scientific Research Program Project (2021KJ011)

Data Availability: All data are available by contacting the corresponding author

References.

1. Xinyang L ,Xiaowan H, Xingye C, et al. Spatial redundancy transformer for self-supervised fluorescence image denoising[J].Nature Computational Science,2023,3(12):1067-1080.
2. Yudong Cao, Xinlin Liao, Xin Chen, et al. A visual target detection model incorporating deep active learning[J/OL]. Journal of Jilin University(Engineering Edition),2024,12(2):1-8. https://doi.org/10.13229/j.cnki.jdxbgxb.20240223.
3. Li D, Wang R, Wang Y, et al. A highly robust track surface defect detection method based on machine vision[J]. Journal of Railway Engineering,2024,41(05):11-18.
4. Vijayakumar A, Vairavasundaram S. YOLO-based Object Detection Models: A Review and its Applications[J]. Multimedia Tools and Applications, 2024, 83(35): 83535-83574.
5. C.H. Chen, S.F. Wang, SZ. Huang, An improved faster RCNN-based weld ultrasonic atlas defect detection method, Meas[J]. Control,(2023),56(3-4):832–843.
6. Y.F. Fan, S.J. Tian, Q.H. Sheng, et al., A coarse-to-fine vehicle detection in large SAR scenes based on GL-CFAR and PRID R-CNN[J]. Remote Sens, 2023,44 (8):2518–2547.
7. W.S. Sheng, X.F. Yu, J.Y. Lin, et al., Faster R CNN target detection algorithm integrating CBAM and FPN[J], Appl.Sci,22023,13(12): 6913(1-18).
8. Yang L ,Noguchi T ,Hoshino Y .Development of a pumpkin fruits pick-and-place robot using an RGB-D camera and a YOLO based object detection AI model[J].Computers and Electronics in Agriculture,2024,227(P2):109625-109625.
9. Lee C ,Lee S ,Moon G , et al. ReZNS: Energy and Performance-Optimal Mapping Mechanism for ZNS SSD[J].Applied Sciences,2024,14(21):9717-9717.
10. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-sign detection and classification in the wild[J]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2016;21–26 .pp. 2110–2118.
11. L.Z. Wang, Y.M. Zhang, Y.B. Xu, et al., Residual depth feature-extraction network for infrared small-target detection[J], Electronics,2023,12(12):2568(1-20).
12. Z.Y. Zhang, P. Gao, S.X. Ji, et al., Infrared small target detection combining deep spatial-temporal prior with traditional priors[J], IEEE Trans. Geosci. Remote Sens,2023,61:5004718(1-18).
13. R.H. Li, Y. Shen, YOLOSR-IST: a deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO[J], Signal Process. 208 (2023) 108962 (1-12).
14. L. Shen, B. Lang, Z. Song, DS-YOLOv8-Based object detection method for remote sensing images[J], IEEE Access,2023,11:125122–125137, https://doi.org/10.1109/access.2023.3330844.
15. G. Zhang, et al., Small target detection algorithm for UAV aerial images based on improved YOLOv7-tiny[J], Engineering Science and Technology,2023,1–14,https://doi.org/10.15961/j.jsuese.202300593.
16. Han Q, Fan Z, Dai Q,et al. On the connection between local attention and dynamic depth-wise convolution. 2021.
17. Shi M ,Zheng D ,Wu T , et al.Small object detection algorithm incorporating swin transformer for tea buds.[J].PloS one,2024,19(3):e0299902-e0299902.
18. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision tra

nsformer using shifted windows. In: Proceedi ngs of the IEEE/CVF international conference on computer vision; p. 10012–10022.
19. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision tra nsformer using shifted windows. In: Proceedi ngs of the IEEE/CVF international conference on computer vision;.p. 10012–10022.
20. Yi Y. Goat Based on Improved YOLOv3 Res earch on Object Detection of Dairy. 2022;.
21. CHENG Y, ZHU J, JIANG M, et al. Flow: a d ataset and benchmark for floating waste detect ion in inland waters[C]// Proceedings of the I EEE/CVF International Conference on Comp uter Vision, 2021: 10953-10962.
22. DU D, ZHU P, WEN L, et al. Visdrone-det20 19: the vision meets drone object detection in image challenge results[C]// Proceedings of th e IEEE/CVF International Conference on Co mputer Vision Workshops, 2019.
23. Tong Z, Chen Y, Xu Z, et al. Wise-IoU: boun ding box regression loss with dynamic focusin g mechanism[J]. arxiv preprint arxiv:2301.100 51, 2023.
24. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision tra nsformer using shifted windows. In: Proceedi ngs of the IEEE/CVF international conference on computer vision; p. 10012–10022.
25. Wang Guoming,Jia Daiwang. Optimisation of small target detection model based on YOLOv 8[J/OL]. Computer Engineering,1-10[2024-11 -28].https://doi.org/10.19678/j.issn.1000-3428 .0070027.