Enhancing YOLOv8 Object Detection with Shape-IoU Loss and Local Convolution for Small Target Recognition

Qi Zhang[1*], Jingyuan Zhang[2], Shilei Yang[2]
[1]Engineering Training Center, Xi'an University of Science and Technology, Xi'an, 710600, China
[2]School of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an, 710600, China
E-mail: xustzhangqi@163.com

Keywords: YOLO, loss function, local convolution, object detection, distinguish, computer

In light of the prevailing challenges, including the suboptimal recall rate and the substantial computing demands of the You Only Look Once (YOLO) series target detection algorithm, the study proposes an innovative approach to enhance the algorithm's loss function. This enhancement is achieved by introducing the SIOUs loss function, which is derived from the YOLOv8 network architecture. Additionally, the study proposes an adjustment to the calculation method of the weight of the loss function, employing the SlideLoss loss function. Then, this paper introduces SPD-Conv and PConv local convolution structure to improve the Backbone layer and proposes a computer object detection model based on loss function and local convolution. It was found that the prediction accuracy of the study model was above 90% on both the training and validation sets. Compared with the conventional YOLOv8 network, the average accuracy of the study model was increased by 13.19%. In the COCO and ImageNet datasets, the number of parameters, FLOPs, reasoning speed, and other indicators of the proposed model were better than the traditional YOLOv8. This indicated that the proposed algorithm could significantly improve the detection speed under the condition of improving the detection accuracy, realizing the balance between the two, and being more effective and reliable. The example results showed that the average coverage rate and the center position error of the research model were increased by 5.7% and 2.3%, respectively, which could effectively improve the operation accuracy. The research results can provide high detection accuracy for the target identification problems in various fields and have important application value in medical image analysis, industrial quality inspection, and other fields.

Povzetek: Razvita je izboljšana YOLOv8 arhitektura z novo Shape-IoU izgubo in lokalnimi konvolucijami, ki znatno poveča natančnost zaznave majhnih tarč ob zmanjšani računski zahtevnosti.

1 Introduction

The rapid advancement of big data, computing power, and deep learning algorithms has ushered artificial intelligence technology into a new era of big data and deep learning. In this process, artificial intelligence technology has been extensively applied in various industries and commercial sectors, providing substantial convenience to people's daily lives. Additionally, it has played a pivotal role in promoting technological innovation and development across multiple domains [1-2]. As a key task in computer vision, object detection has shown great potential for applications in multiple scenarios such as autonomous driving, intelligent monitoring, and medical image analysis. Currently, Computer Object Detection (COD) technology is moving in diversified directions [3-4]. In recent years, due to the increasing demand for object detection speed, efficient object detection algorithms have emerged. Single-stage detector object detection algorithms have attracted a lot of attention due to their excellent performance. This kind of algorithm transforms object detection tasks into regression problems and achieves real-time detection by simultaneously completing object classification and detection through a single forward propagation. Compared to traditional two-stage detectors, single-stage detectors do not require the generation of candidate regions for reclassification. Instead, they directly perform dense sampling on the feature map and use Convolutional Neural Networks (CNN) for classification and position regression. LiH et al. introduced deterministic aware pseudo labels for object detection and improved the semi-supervised object detection algorithm to address the issue of imbalanced positioning accuracy and amplification categories when using pseudo labels. The results showed that this method improved the performance of the most advanced SSOD on COCO and PASCAL VOC, with an average accuracy increase of 1-2%, and was orthogonal complementary to most existing methods [5]. To reduce the noise in the generation of pseudo labels and minimize the negative impact of noisy pseudo labels on model training, Yang J adopted a comprehensive pseudo label denoising method. In unsupervised 3D object detection adaptive noise training, this method was the most advanced in performance among all evaluation settings, greatly exceeding the corresponding baseline. Even in the KITTI 3D object detection benchmark, its performance has improved by 9.6% to 38.16% compared to fully supervised prediction based on object priors [6]. In view of the problems existing in the detection of small and medium Remote Sensing (RS) targets, Weiya et al. proposed a RS Small Object Detection (SOD)

method based on cross-layer fusion and weighted receptive field. To solve the problem of feature loss in deep layers of this method, a cross-layer attention fusion module was designed. The background noise was effectively filtered by introducing the double-layer routing attention. To enhance the ability of the model to sense multi-scale objects, especially small-scale ones, a weighted spatial pyramid pooling module of multi-receptive field voids was introduced. The experimental results clearly demonstrated the significant advantages of the model in RS SOD, surpassing the performance of the current mainstream model [7]. Faced with inherent challenges such as multi-scale, limited target area, and complex background in RS images, Zhong et al. proposed a RS image object detection method with multi-scale fusion dynamic head. This method has improved accuracy by 6.3% compared to traditional models, clearly demonstrating its significant advantages [8]. To solve the safe operation of the power grid, an improved You Only Look Once (YOLO) algorithm was introduced by Ya et al. Through experimental verification, in the test set, the unimproved YOLO algorithm's recall rate and detection accuracy were 68.3% and 88.8%, respectively, while the recall rate and detection accuracy of the improved model reached 82.7% and 91.4%. The crane recognition accuracy was improved by 2.9%, with a significant improvement [9]. To enhance the effectiveness of SOD in UAV aerial images, Xiang et al. proposed an enhanced YOLO algorithm based on multi-scale spatial context. This algorithm achieved significant results, with an Mean Absolute Precision (mAP) of 3.0% higher than the baseline method, demonstrating satisfactory performance in SOD [10]. YANG et al. constructed a cascaded model grounded on an improved YOLO structure to address difficulty in combining environmental semantic information at various scales in underwater detection algorithms. The proposed algorithm was capable of detecting aquatic targets and the evaluation index of common objects in context has improved by 4.34% [11]. To improve the generalization performance of the model and the application of data augmentation technology in target detection, Abdulghani et al. proposed a multiple data enhancement method. The experimental results showed that each data enhancement technology brought different degrees of improvement, and the method improved the mAP value of all objects by 13%, effectively improving the performance of the target detection algorithm [12]. Ananthakrishnan et al. developed an improved detection framework built on YOLOv5 to address inadequate lighting and poor performance of automated object detection under nighttime lighting conditions. It was used for effective detection under uneven lighting conditions at night. This architecture achieved higher results on mAP while reducing model size and total parameters. In terms of model size, the proposed model was 11.24% lighter and 12.38% lighter in terms of parameters [13]. The above research results are summarized as shown in Table 1.

Table 1: Summary table of literature reviews

Authors	Research Methods	Experimental Results
Li H et al. [5]	We introduce deterministic aware pseudo labels for target detection to improve semi-supervised target detection calculation	Improved state-of-the-art SSOD performance on COCO and PASCAL VOC, with improved average accuracy of 1-2%
Yang J et al. [6]	Unsupervised domain-adaptive denoising self-training in 3D object detection	Increase in performance by 9.6% 38.16%
Weiya et al. [7]	Proposed a YOLO-based method with cross-layer fusion and weighted receptive field.	The model showed significant advantages in small object detection,outperforming mainstream models.
Zhong et al. [8]	Introduced a multi-scale fusion dynamic head method for object detection in RS i	Improved accuracy by 6.3%compared to traditional models.
Ya et al. [9]	Developed an improved YOLO algorithm for power grid safety	Achieved 91.4%detection accuracy,with crane recognition accuracy improved by 2.9%
Xiang et al. [10]	Proposed an enhanced YOLO algorithm with multi-scale spatial context for UAV aerial image	Increased mAP by 3.0%compared to baseline methods.
YANG et al. [11]	Designed a cascaded model based on improved YOLO for underwater object detection	Improved evaluation metrics by 4.34%compared to baseline networks.
Abdulghani et al. [12]	Investigated data augmentation techniques for object detection in limited datasets.	Improved mAP by 13% of all objects.
Ananthakishnan et al. [13]	Proposed an improved YOLOv5 framework for object detection under	Achieved higher mAP, reduced model size by 11.24%, and parameters by 12.38%.

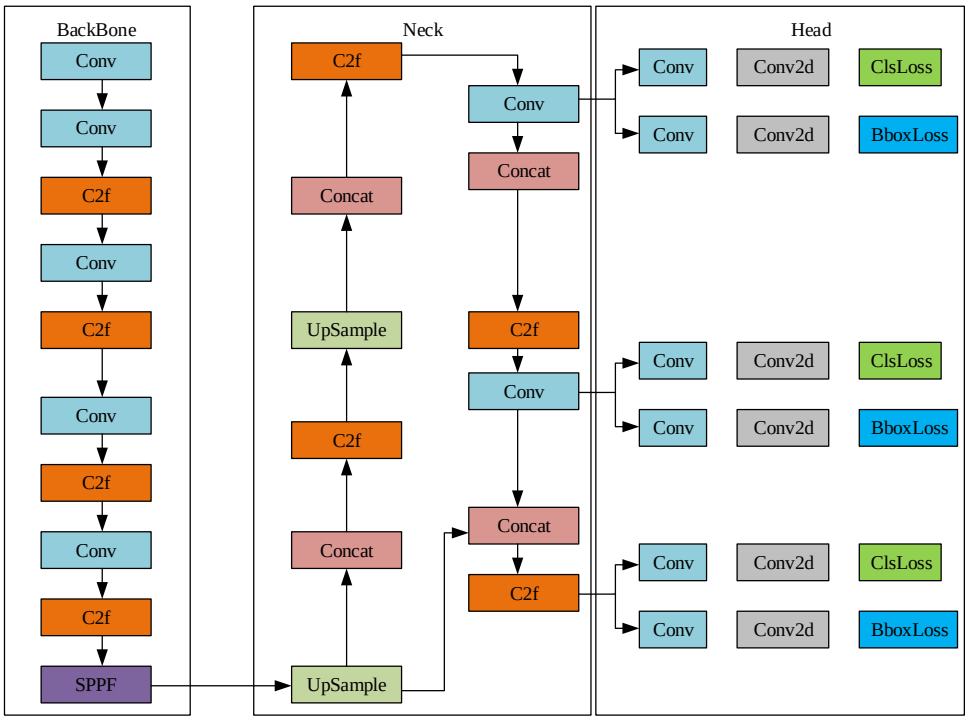
	uneven lighting conditions at night	

In summary, research on YOLO algorithms has achieved certain results. The success of the YOLO algorithm is attributed to its end-to-end training method, fast detection velocity, low background false detection rate, and strong generality and generalization capacity. Meanwhile, the YOLO algorithm also adopts a multi-scale feature fusion strategy, effectively handling object detection problems at different scales. However, the current YOLO series still has shortcomings. On the one hand, compared to other algorithms, YOLO has a lower recall rate, which may result in missing some targets. This is also due to its approach of treating object detection tasks as a regression problem, resulting in relatively conservative predictions of the target. On the other hand, it has a high demand for computing resources. To achieve high performance, the YOLO series models require high computing resources, which limits the application of YOLO in some resource-constrained scenarios. Based on this, the study innovatively proposes a modified YOLOv8 target detection model based on loss function and local convolution. In this model, based on the YOLOv8 network, the SIOU loss function is introduced to improve the algorithm. Meanwhile, the SlideLoss function is used to adjust the weight calculation method of the loss function to solve the problem of low recall rate and accuracy of YOLOv8 target detection. Then, a SPD Convolution (SPD-Conv) and a more efficient local convolution structure (Partial Convolution, PConv) are introduced to improve the Backbone layer to solve the problem that YOLOv8 needs high computing resources. The hypotheses to be verified in this study mainly include four aspects. (1) Whether SIOU loss function can improve the recall rate and precision of YOLOv8 network; (2) Whether the introduction of SPD-Conv can improve the SOD performance of YOLOv8 network; (3) Whether the combination of SPD-Conv and PConv reduces the calculation cost while maintaining accuracy; (4) Whether the combination of SIOU loss function, SPD-Conv, and PConv can balance the detection accuracy and efficiency of YOLOv8 network. Through the above in-depth analysis, the research is expected to improve the detection accuracy of YOLOv8 algorithm, reduce the computing resources, and promote the application value of YOLO series algorithm in object detection and recognition.

2. Methods and materials

2.1 Construction of an improved object detection model based on loss function

At present, COD technology is widely used in various fields. However, current computer detection technology can no longer meet the needs of reality [14-16]. In response to the problems existing in current object detection algorithms, this study improves the YOLOv8 algorithm from the aspects of local convolution structure and objective function, and proposes a COD model suitable for small target scenes. The reason why YOLOv8 is chosen as the sole baseline in this study is that, firstly, the algorithm is the latest version of the YOLO series. Compared to YOLOv5 and YOLOv7, it has significantly improved detection accuracy, speed, and model efficiency, representing the state-of-the-art in target detection. Secondly, YOLOv8 has undergone multiple optimizations in architecture design and training strategies, such as a more efficient backbone network and a more flexible loss function, providing a stronger foundational framework for research. In addition, YOLOv8 has an open source community and abundant tools and resources, making it easy for researchers to conduct experiments and validation. Choosing YOLOv8 as the benchmark can more intuitively demonstrate the effectiveness of the improved method and ensure that the research results are consistent with the current technological frontier, thereby enhancing the practicality and reference value of the research. The traditional YOLOv8 structure is shown in Fig.1.



Conv: Represents a conventional convolutional layer, Concat： Represents the connection layer, UpSample represents the upsampling operation

Figure 1: Traditional YOLOv8 structure

In Fig.1, Conv represents the traditional convolutional layer, Concat represents the connected layer, and the UpSample represents the upsampling operation. The traditional YOLOv8 structure includes Backbone, Neck, and Head layers. Among them, the Backbone layer effectively extracts feature information of the input object through a series of convolution and transpose convolution operations. Meanwhile, to reduce network size and improve overall performance, this layer also incorporates the design of residual connections and bottleneck structures [17-18]. This section uses convolution module, Cross Stage Physical Bottleneck with 2Convolutions (C2f) module, and Spatial Pyramid Pooling Fast (SPPF) module as the basic building blocks. The Neck layer is mainly responsible for fusing feature maps from different stages in the Backbone to enhance feature expression ability. Its internal structure includes components such as upsampling module and Path Aggregation Network (PAN). The Head layer focuses on the localization and category determination of object detection. The Head layer adopts a separated design, which means that classification tasks and detection tasks are handled separately by independent classification heads and detection heads. Both parties bear the optimization of classification loss and localization loss separately. Specifically, the classification loss uses VFL loss, while the localization loss combines DFL loss and CIoU loss. The traditional YOLOv8 algorithm combines localization loss and classification loss when calculating loss, with the total loss representing the loss of the entire algorithm. The total loss function is given by equation (1).

$$CIoU = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{1}$$

In equation (1), b and b^{gt} are the center point of the prediction and real boxes. c is the diagonal distance between the closed areas of two rectangular boxes. ρ represents the calculated Euclidean distance. α is the weight coefficient. v measures the ratio consistency between the width and height of two rectangular frames. Compared with IoU, this loss function measures the positioning results by including the overlap area, center point distance, and Aspect Ratio (AR) as part of the loss, making the model's positioning results more accurate. However, the CIoU loss calculation used in this method is relatively complex because it not only considers the Intersection over Union (IoU) ratio itself but also introduces multiple factors such as center point distance and AR. This complexity may result in high computational costs during the training process, especially for large-scale datasets, where training time may significantly increase. In the formula of the CIoU loss function, there are some terms related to the AR of the predicted and true boxes. When the AR of the predicted box differs greatly from the real box, the values of these terms may become very large, leading to the problem of gradient explosion [19-20]. For this issue, this study improves the YOLOv8 function. Firstly, SIOU not only considers IoU but also combines the angle information of the target to make the

loss function smoother and avoid the gradient explosion problem caused by the excessive difference in length and width ratio. Furthermore, SIOU has proven to be particularly effective in detecting small targets. This efficacy stems from SIOU's ability to capture shape and position characteristics of these targets with a high degree of precision, enhancing the system's overall detection accuracy. Compared to CIoU, SIOU performs better in terms of computational efficiency and stability, making it particularly suitable for handling large-scale datasets and complex models. Meanwhile, SIOU exhibits higher robustness and accuracy in SOD tasks [15]. The angle loss Ω of this function is described in equation (2).

$$\Omega = 1 - 2 \times \sin^2\left(\arcsin\left(\frac{h}{\chi}\right) - \frac{\pi}{4}\right) \tag{2}$$

In equation (2), h is the height difference between the actual and predicted box center points. χ is the distance between two center points. The distance loss ∂ of the SIOU function also takes into account the angle loss, as shown in equation (3).

$$\partial = \sum_{t=x,y} 1 - e^{-\gamma \rho_t} \tag{3}$$

In equation (3), ρ_t is the square value where the difference between the actual box and the predicted box is less than the min bounding rectangle width of both boxes, $\gamma = 2 - \Omega$. Among them, w is the width of the minimum bounding rectangle of the actual box and the predicted box. h_1 is the height of the rectangle. (o_{ax}, o_{ay}) and (o_{px}, o_{py}) are the coordinates of the center points of the actual and predicted boxes. The description of the shape loss ℓ of SIOU is shown in equation (4).

$$\ell = \left(1 - e^{-w_1}\right)^{\theta} + \left(1 - e^{-h_2}\right)^{\theta} \tag{4}$$

In equation (4), w_1 and h_2 are the width and height of the prediction boxes. θ is the correlation of shape loss, within the range of $[2,6]$. This range is determined based on previous experience [19]. Finally, the loss description formula for SIOU is shown in equation (5).

$$L = 1 - IOU + \frac{\partial + \ell}{2} \tag{5}$$

This study can replace the original CIoU loss of YOLOv8 with SIoU loss to improve the convergence speed and accuracy. Then, in response to the problem of detecting more details or smaller target categories in practical applications, this study further utilizes SlideLoss to adjust the weight calculation method of the loss function, making the algorithm network more inclined to learn difficult samples.

The weight allocation scheme of the SlideLoss function is shown in Fig.2.

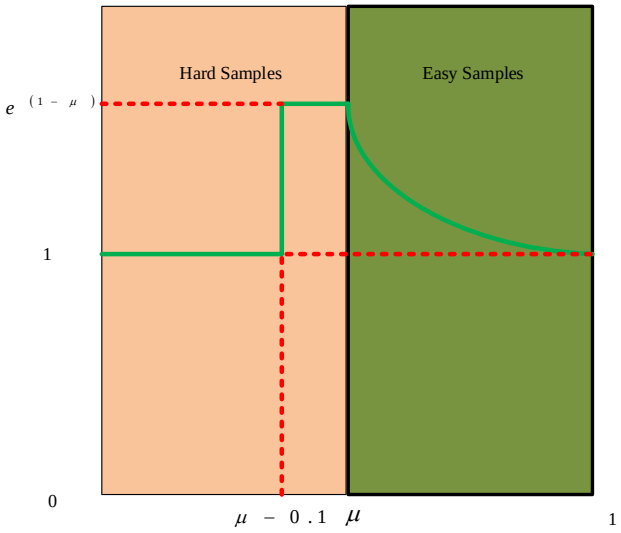


Figure 2: Weight allocation scheme for SlideLoss function

Slide Loss uses the average IoU value of all bounding boxes as the threshold μ. The determination of this threshold depends on the characteristics of the entire dataset and is dynamically adjusted to changes in the data. For samples in object detection tasks, this study classifies them based on their IoU values with real boxes for ease of classification. Specifically, when the IoU value is below the preset threshold, the sample is considered a difficult sample with a higher weight, thus occupying a more important position in loss calculation. On the contrary, samples with IoU values higher than the threshold are considered simple samples, with relatively smaller weights and smaller contributions to the loss. Samples located outside the $[\mu, \mu - 0.1]$ boundary region between different categories are more prone to misclassification. Therefore, assigning the maximum weight to these samples can make the classifier pay more attention to these difficult samples during training, thereby improving overall classification performance. However, samples located near decision boundary $[\mu, \mu - 0.1]$ have the highest weight, indicating that they generate the most significant losses. However, when the IoU value is below a certain level $\mu - 0.1$, these samples are usually considered extremely difficult or false positive abnormal situations, and their weights are not adjusted at this time [20]. The SlideLoss strategy is based on this principle, which prioritizes and invests more resources in learning difficult samples, thereby effectively improving the overall IoU value and enhancing the detection precision.

2.2 Construction of object detection model based on local convolution improvement

After improving the loss function of the YOLOv8 algorithm, it is necessary to further enhance the computational efficiency of YOLOv8, suppress the frequent memory access of neural networks, especially deep convolutions, and reduce redundant calculations and memory access. Therefore,

this study introduces SPD-Conv and PConv structure to improve the Backbone layer. The Backbone network of YOLOv8 utilizes a convolution module with 2 steppes for downsampling, gradually expanding the receptive field and generating multi-scale feature maps. However, this downsampling process inevitably causes the loss of some detailed information in the feature map. Especially for small objectives with low-resolution and limited information, this information loss situation can seriously affect their detection performance and may cause a significant decrease in detection accuracy. Meanwhile, YOLOv8's convolutional layers perform poorly in handling small targets. This is because after multiple convolution operations, the feature information of small targets is easily diluted, leading to a decrease in detection accuracy. In addition, the convolutional structure of YOLOv8 is more sensitive to background noise, especially in complex backgrounds, which can easily lead to false positives or false negatives. The article introduces a novel Convolutional module called SPD-Conv in the initial downsampling stage to replace the original convolution layer with a stride of 2, ensuring that information is not lost during the downsampling process. The network structure is shown in Fig.3.

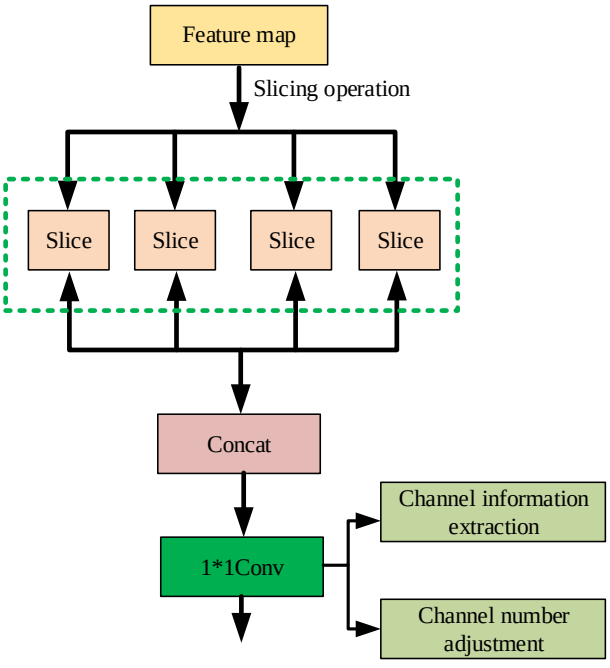


Figure 3: SPD-Conv structure.

In Fig.3, the SPD-Conv structure starts from the input feature map and is segmented into multiple slices by slice operation. These sections are then spliced with Concat together to form a new feature map. Next, the spliced feature maps are processed by 1×1 convolution to adjust the channel number or fused channel information. Meanwhile, the structure also contains the channel information extraction and channel number adjustment module for further processing the channel information. This design enhances the feature extraction capability and flexibility of the model. The workflow of the SPD-Conv module first involves segmenting the feature map into four parts, implementing slicing operations, and then recombining these parts and applying 1×1 convolution kernels. This step is to extract information across channels and adjust the number of channels while maintaining extremely low information loss throughout the entire process. The C2f module is the core component of the Backbone layer, which relies on multiple layers of DarknetBottleneck modules to deeply explore the features of the feature map. In addition, the C2f module integrates the extracted feature information into the output by introducing multi-layer skip connections and concatenation techniques. This method not only effectively extracts features at different levels but also prevents the occurrence of "gradient vanishing/exploding" phenomenon. Considering the redundant computing problem caused by this method, this study introduces PConv to improve the C2f module and calls the improved module C2f-Pairial Convolution (C2f-PC). The structure of C2f-PC is shown in Fig.4.

In the structure of Fig.4, the input first passes through a Split operation to segment the feature graph. Then, the feature graph enters the PConv layer for spatial feature extraction, and the PConv layer can capture the local features in the image more effectively through the local convolution operation. Then, the feature map through multiple PConv layers are spliced (Concat) with the output of conventional convolution (Conv) layers to integrate different levels of feature information. This improved structure helps to improve the performance of YOLOv8 in the target detection task and enhance the feature extraction capability and detection accuracy. This improved structure helps to enhance the performance of YOLOv8 in object detection tasks, strengthen feature extraction capabilities, and improve detection accuracy. Among them, studies have shown that PConv can reduce Floating Point Operations (FLOPs) by about 30%-50% in typical convolutional layers, depending on the size of the input feature map and the number of channels. In addition, PConv can reduce memory access by 20%-40% by reducing redundant calculations and optimizing data access modes [21-22]. PConv effectively reduces the amount of FLOPs required by limiting the number of channels involved in computation, thereby reducing computational complexity. Compared to traditional convolution, PConv can significantly reduce memory access, which is particularly advantageous for devices with limited Input/Output (I/O) resources. Although it only processes a subset of input channels, these selected channels can still play a role in the subsequent pointwise convolutional layers, ensuring that feature information can flow and be utilized between all channels. The final COD model suitable for difficult sample detection is obtained through the above improvements, as shown in Fig.5.

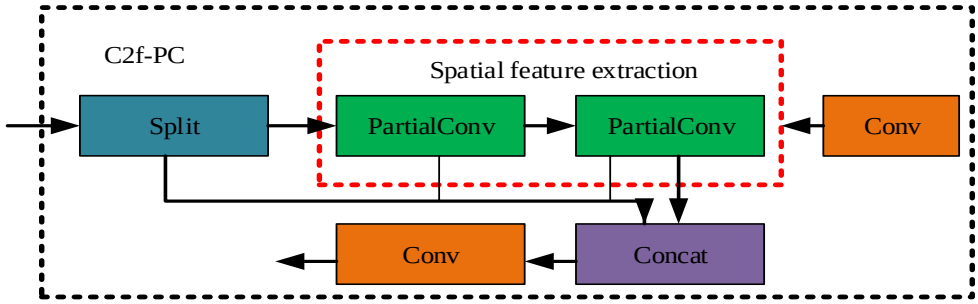


Figure 4: C2f-PC structure

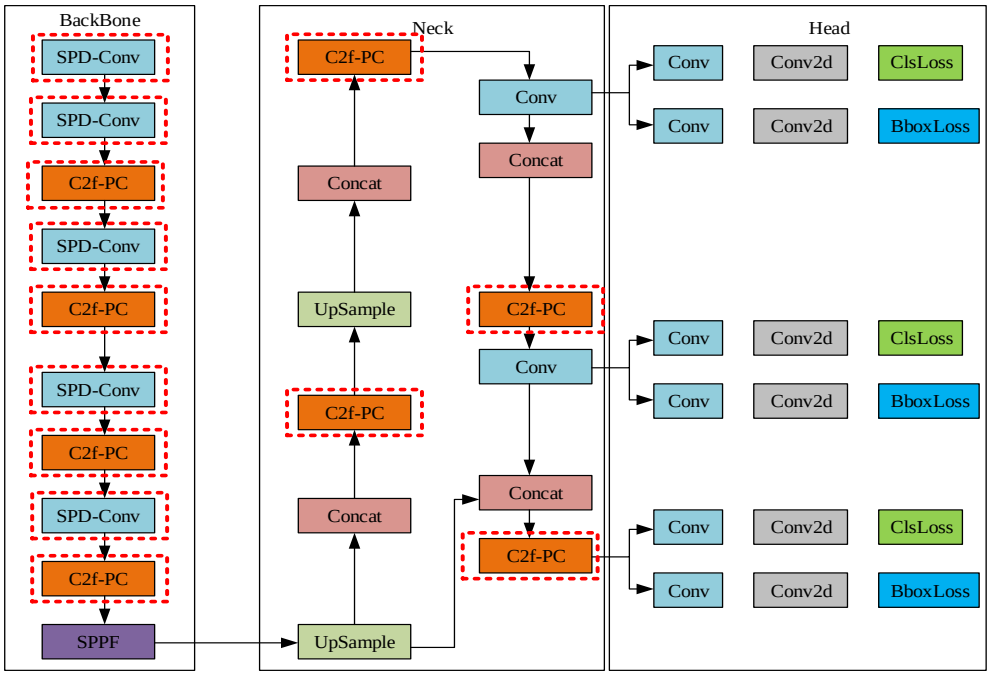


Figure 5: Computer object detection model

In the proposed COD model, SPD-Conv can achieve downsampling without sacrificing information integrity, thereby enhancing the accuracy of SOD. In contrast, PConv utilizes redundant information in feature maps to reduce computational burden and memory access requirements. Specifically, PConv only applies regular convolution to a portion of the input channel to extract spatial features, while keeping the rest channels unchanged. The pseudo-code of the improved COD algorithm based on the loss function and local convolution is shown in Fig.6.

```
// Improved YOLOv8 Algorithm Pseudocode

// Initialize network with improved backbone
function initializeNetwork() {
    backbone = createBackboneWithPConvAndSPDConv()
    head = createDetectionHead()
    return combine(backbone, head)
}

// Create backbone using Partial Conv (PConv) and SPD-Conv
function createBackboneWithPConvAndSPDConv() {
    // Define layers with PConv and SPD-Conv
    layers = []
    for i in range(numLayers):
        if usePConv(i):
            layers.append(createPConvLayer())
        else if useSPDConv(i):
            layers.append(createSPDConvLayer())
        else:
            layers.append(createStandardConvLayer())
    return connectLayers(layers)
}

// Loss function using SIOU and SlideLoss
function computeLoss(predictions, targets) {
    siouLoss = computeSIOU(predictions, targets)
    slideLoss = computeSlideLoss(predictions, targets)
    totalLoss = combineLosses(siouLoss, slideLoss)
    return totalLoss
}

// Training loop
function trainNetwork(data, epochs) {
    network = initializeNetwork()
    for epoch in range(epochs):
        for batch in data:
            predictions = forwardPass(network, batch.input)
            loss = computeLoss(predictions, batch.targets)
            backpropagate(network, loss)
            updateWeights(network)
    return network
}

// Main execution
data = loadData()
epochs = setNumberOfEpochs()
trainedNetwork = trainNetwork(data, epochs)
```

Figure 6: A pseudo-code diagram of an improved COD algorithm based on loss function and local convolution

3 Results

3.1 Model performance testing

To test the effectiveness of the proposed model, this paper focuses on actual monitoring video data of ground service processes at multiple high traffic airports in China, and conducts detailed annotation work based on standard documents issued by the Civil Aviation Administration of China. By writing Python scripts, this study effectively screens and removes small and non-core monitoring targets (This represents a less important goal than core processes such as aircraft take-off and landing) from the data, accurately constructing a dataset suitable for training and testing object detection algorithms. In the processing of airport flight support surveillance video, due to the long video time, this study adopts a fixed time interval strategy (1 frame every 5 seconds) to extract keyframes. For high frame rate videos, the extraction frequency can be appropriately reduced and increased to ensure that important information is not missed.

However, the keyframe extraction operation inevitably introduces a large number of images with highly similar content, leading to the problem of data redundancy. To improve the quality of the data set, the study implements data filtering steps after the data collection. First, the degree of repetition of the image content is quantified by calculating the similarity between the images. Next, a similarity threshold is set. When the similarity between two images exceeds the threshold, they are considered highly similar and only one image is retained. In addition, the study also introduces a time interval strategy to ensure that adjacent frames are not excessively filtered in time, thereby preserving key dynamic information in the video. To further optimize the filtering effect, the K-means clustering algorithm is used to group the images, and only the most representative images are retained in each group. Finally, through this series of data filtering steps, the study successfully removes the redundant images in the data set, significantly improving the diversity and quality of the data. After labeling all the data using the above annotation method, the final COCO dataset consists of 15,118 images, and the dataset is divided in a 4:1 ratio, of which 12,105 are used for training and 3,013 for testing. The ImageNet dataset consists of 24,005 images, divided in a 4:1 ratio, with 19,204 images used for training and 4801 images used for testing. During this algorithm evaluation, mAP is utilized to assess the detection accuracy, computational inference time is taken to evaluate the speed, and FLOPs and parameter count are used to evaluate the computational complexity and complexity of the model. Among them, mAP is calculated by combining Precision and Recall (R). The pixel resolution is 1920*1080. Table 2 shows the configuration of environmental parameters.

In Fig.7 (a), the loss in two datasets gradually decreases with the iteration of learning times. When the last training finished, the training set's loss rate decreases from 0.1800 to 0.1084, and that in the validation set from 0.1362 to 0.0915, showing an improvement in generalization ability. In Fig.7 (b), the prediction accuracy is above 90%, and as the iteration increases, the final accuracy is 96.93% and 98.21%. To test the effectiveness of each part of the algorithm improvement, ablation experiments are performed to evaluate the impact of different improvement strategies on the performance of algorithm under the same conditions. This ablation experiment uses YOLOvs8 as a benchmark and testes it on a self-built dataset. No pre-trained weights are utilized in the tests, as listed in Table 3.

Table 2: Experimental environment and parameter settings

Category	Operating environment
Operating system	Win 7 operating system
Developer Components	JDK 7, Eclipse
CPU	Intel (R) Core (TM) i9-10900X
GPU	Nvidia RTX 3090 GPU
Platform software environment	Python3.8, PyTorch1.10.0
Programming Language	Java
Batch size	8
Batch size	0.001
Batch size	Adam

Firstly, the study model is trained on the training set and the validation set of the COCO data set. The results are shown in Fig.7.

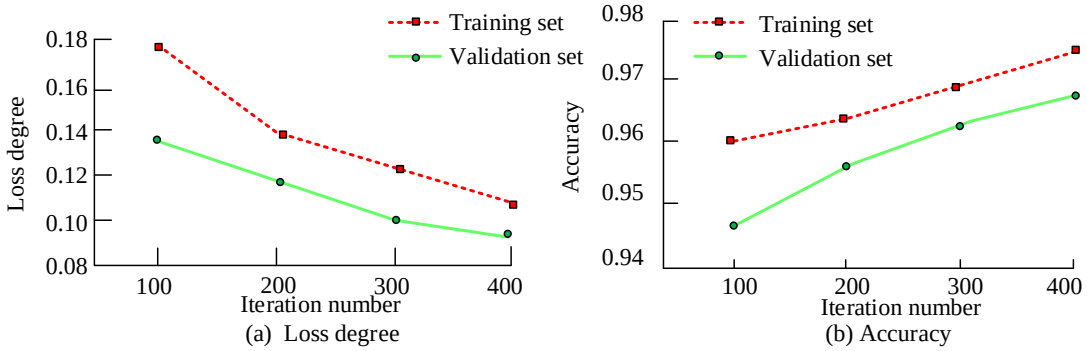


Figure 7: Training results of the research model

Table 3: Results of model ablation experiment

Model		YOLOv8-base	YOLOv8-SPD-Conv	YOLOv8-C2f-PC	YOLOv8-SIoU	Research model
mAP	First	14.3	15.2	14.4	14.5	16.5
	Second	14	15.1	14.7	14.2	16
	Third	14.9	15.6	14.7	15.4	16.4
	Mean	14.4	15.3	14.6	14.7	16.3
	Promotion rate	/	6.25%	1.39%	2.08%	13.19%
AP50	First	27.1	29.1	29.3	27.7	31.2
	Second	27.5	28.9	27.1	28.1	31.8
	Third	27.9	29.6	27.3	28.2	32.7
	Mean	27.5	29.2	27.9	28	31.9
	Promotion rate	/	6.18%	1.45%	1.82%	16.00%
P value		/	$P<0.05$	$P>0.05$	$P>0.05$	$P<0.05$

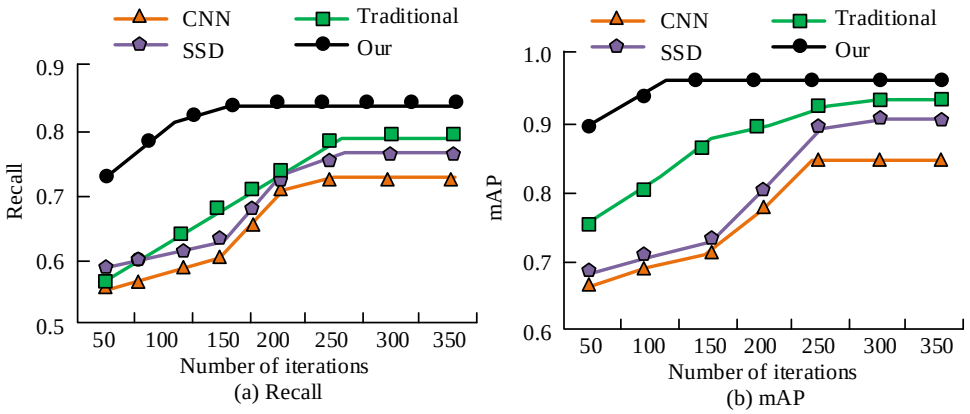


Figure 8: Comparison results of different detection algorithms

In Table 2, after introducing the SPD-Conv structure, mAP and AP50 improve by 6.25% and 6.18% respectively compared to the basic model, effectively solving the problem of reduced detection accuracy caused by information loss in

the benchmark model. After using C2f PC, mAP and AP50 increase by 1.39% and 1.45% compared to the baseline model. This module can enhance the ability of C2f module to learn and express features, and add a SOD layer to strengthen the detection precision. After improving the loss function, the mAP and AP50 increase by 2.08% and 1.82% compared to the basic model, which can accelerate the convergence velocity. The mAP and AP50 of the research model with improved modules increase by 13.19% and 16.00%, with $P<0.05$ indicating statistically significance. The research model can improve the performance of dense object detection and enhance the accuracy. This study selects Single Shot Multi-Box Detector (SSD), CNN, and traditional YOLOv8 as comparative algorithms to test their performance. The corresponding recall rate and mAP results are shown in Fig.8.

In Fig.8 (a), the maximum recall rates for research algorithms, traditional YOLOv8, SSD, and CNN are 0.84, 0.78, 076, and 0.72. Among them, the research algorithm has the best recall performance and converges the fastest, with a 7.7% improvement compared to CNN. In Fig.8 (b), the maximum mAP values of the research algorithm compared to traditional YOLOv8, SSD, and CNN are 0.97, 0.94, 088, and 0.82. Among them, the mAP performance of the research algorithm is the best, with a 3.2% improvement compared to CNN. This indicates that the detection accuracy of the research algorithm is better. To further test the effectiveness, the study tests it on the ImageNet and COCO datasets, as shown in Fig.9.

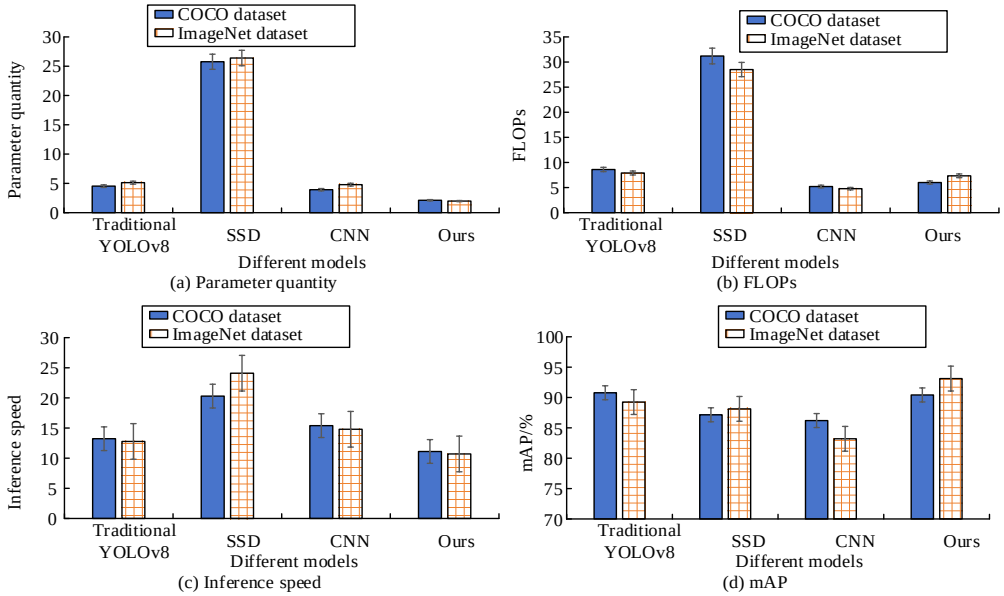


Figure 9: Model test results under different datasets

Figs.9 (a) to (d) show the test values for parameter quantities, FLOPs, inference speed, and mAP of four algorithms. In 9 (a), the proposed algorithm has parameter counts of 2.12 and 1.98 on the COCO and ImageNet datasets, respectively, showing improvements of 53.2% and 61.3% compared to YOLOv8. In 9 (b), the FLOPs of the research algorithm on two datasets are 6.0 and 7.3, which are 30.2% and 7.09% higher than YOLOv8. In 9 (c) and (d), the inference speed and mAP of the designed algorithm on two datasets are 11.1s and 10.7s, 90.7% and 93.1%, which are 16.0% and 16.2%, 14.3% and 13.2% higher than the traditional YOLOv8. The above test results are $P<0.05$, and all the results are statistically significant. This indicates that research algorithms can significantly improve detection speed while enhancing detection accuracy, achieving a balance between the two and making target detection more effective and reliable.

To further verify the robustness of the proposed algorithm in complex environments, the study selects the nuScenes dataset for testing. Among them, the nuScenes dataset is a large-scale autonomous driving dataset. It contains rich sensor data (such as cameras, LiDAR, radar), marked with targets such as vehicles, pedestrians, traffic signs, etc., suitable for tasks such as target detection and tracking. The study processes the dataset in the same way, yielding the nuScenes dataset consisting of 18,425 images, of which 14,740 are used for training and 3,685 for testing. The model has been tested using the same parameters, and the test results are shown in Table 4.

Table 4: Detection results of different models under the nuScenes dataset

Model		Parameter quantity	FLOPs	Inference speed /s	mAP
SSD		4.47	8.9	13.8	78.8
CNN		5.21	7.2	14.9	76.9
Conventional YOLOv8		3.72	8.5	12.3	82.2
Research algorithm		2.33	6.8	9.8	94.2
Compared to the conventional YOLOv8	Promotion rate	37.4%	20.0%	20.3%	14.6%
	P value	$P<0.05$	$P<0.05$	$P<0.05$	$P>0.05$

In Table 4, under the nuScenes dataset, the number of parameters, FLOPs, inference speed, and mAP of the proposed algorithm reach 2.33, 6.8, 9.8s, and 94.2%, respectively, which are all better than the comparison algorithm. Compared with the traditional YOLOv8 algorithm, the various indexes of the proposed algorithm increase by 37.4%, 20.0%, 20.3%, and 14.6%, respectively. Among them, the *P* value of each index except mAP is less than 0.05, which is statistically significant. The results show that the algorithm can balance the detection accuracy and efficiency in complex environment and effectively identify the detection target. The results of the confusion matrix for the different models in the test set of the nuScenes dataset are shown in Table 5.

Table 5: Confusion matrix analysis of the different detection models

Different detection models		The actual case	The actual counterexample	Accuracy rate	Recall	F1 score
SSD	Forecast the case	TP=800	FP=200	0.80	0.84	0.82
	Predicting counterexample	FN=150	TN=1850			
CNN	Forecast the case	TP=750	FP=250	0.75	079	0.77
	Predicting counterexample	FN=200	TN=18D0			
Conventional YOLOv8	Forecast the case	TP=850	FP=150	0.85	0.89	0.87
	Predicting counterexample	FN=100	TN=1900			
Research algorithm	Forecast the case	TP=900	FP=100	0.90	0.95	0.92
	Predicting counterexample	FN=50	TN=1950			

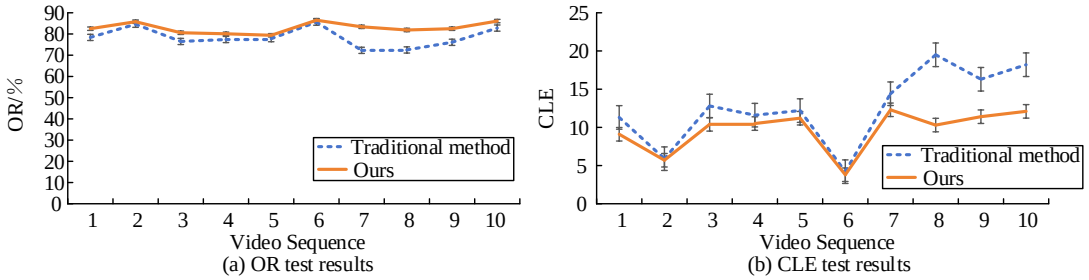
From Table 5, SSD performs moderate and both FP and FN are high, indicating that the model has some problems in false reporting and under-reporting. CNN performs poorly with higher FP and FN, indicating that the model has limited performance in the target detection task. The traditional YOLOv8 performs better with lower FP and FN, indicating that the model has high accuracy in the target detection task. The improved YOLOv8 shows the best performance and the lowest FP and FN, indicating that the improved model has significant advantages in reducing false and under-reporting. In addition, the improved YOLOv8 has obvious advantages, with an accuracy rate of 0.90, which is higher than other models, indicating that the FP is lower. The recall rate is 0.95, which is higher than other models, indicating that the under-reporting rate. The F1 score is 0.92, which has the best comprehensive performance, indicating that it achieves a

better balance between precision rate and recall rate. The improved methods (such as SIoU loss function, SPD-Conv, etc.) can still maintain high detection accuracy and robustness in complex scenarios.

3.2 Example application analysis

To validate the practical performance, this paper applies it to the operation system of a certain fish and shrimp fishing vessel, and compares the detection results of a target crayfish by the research model with the statistical results of traditional methods. Ten video clips are randomly selected from the valid ones that contained fishing crayfish in baskets. Traditional methods are used to count the catch of crayfish in 10 video segments. Ten randomly selected video clips of catching crayfish are inputted into the research model for statistical analysis of the number of crayfish caught. Video clips have a resolution of 1920x1080 pixels with a duration of 10s. This time, Overlap Rate (OR) and Center Location Error (CLE) are chosen as assessment indexes. The detection comparison of traditional and research methods are displayed in Fig.10.

In Fig.10 (a), in terms of OR indicators, the OR of the research models all reaches over 80%, and their average detection results improve by 5.7% compared to traditional methods, $P<0.05$. In Fig.10 (b), in the CLE indicator, the average CLE of the research model is 9.7, which is 2.3% higher than the traditional method. This means that the research method performs greater and can assist in statistical analysis of fishing boat operations, improving operational accuracy. The results of comparing the detection performance of the fish and shrimp fishing vessel operation system using research algorithms with the detection of marine biological targets before use are exhibited in Table 6.



(a) OR test results (b) CLE test results

Figure 10: Traditional and research methods for fish and shrimp fishing vessel detection results

Table 6: Marine biological target detection effectiveness before and after using research algorithms

Category	Pre-use accuracy (%)	Post-use accuracy (%)	*P* value
Close group	97.524	99.322	$P<0.05$
Long distance group	90.396	92.432	$P>0.05$
Dark light group	87.804	89.146	$P>0.05$
Fuzzy group	80.676	84.588	$P<0.05$
Multi-target group	78.408	81.832	$P>0.05$
Twisted group	93.744	97.096	$P<0.05$
Mean	87.912	90.63	$P<0.05$

In Table 6, overall, the detection accuracy of each group is lifted utilizing the research algorithm, with an average increase of 2.7%. Among the tested samples, the closest group has the highest accuracy. Due to the high image clarity and good lighting of the close range group, its accuracy reaches 99.3%, an increase of 2.2% compared to before use. The lowest is the multi-target group, where some images have target occlusion, greatly affecting the accuracy of detection, only 81.8%, but still 3.4% higher than the detection accuracy before use. This result indicates that the research method can effectively identify detection targets and improve the efficiency of fishing operations for fishing vessels.

4 Discussion

The proposed improved YOLOv8 target detection model, which integrates the SIoU loss function and local convolution, demonstrated superior performance compared to traditional YOLOv8 and other YOLO-based models. Experimental results indicated that the improved loss function increased mAP and AP50 by 2.08% and 1.82%, respectively, aligning with findings from Weiya et al. [7], who proposed a RS SOD method based on cross-layer fusion and weighted receptive fields. This improvement was attributed to the introduction of SIOU loss function and SlideLoss function. The SIOU function optimized bounding box regression by combining target shape information, while the SlideLoss function enhanced dense object detection by dynamically adjusting loss weights. Additionally, the integration of SPD-Conv and PConv in the Backbone layer further boosted the model's performance. SPD-Conv increased mAP and AP50 by 6.25% and 6.18%, respectively, while PConv added 1.39% and 1.45%, surpassing the cascade model proposed by YANG et al. [11]. SPD-Conv reduced downsampling loss, improving SOD, while PConv enhanced computational efficiency. These modifications not only elevate detection accuracy but also significantly reduce computational complexity, achieving a balance

between precision and efficiency.

The proposed method not only performed well on the COCO and ImageNet datasets, but also demonstrated its applicability in a broader range of scenarios. For example, in RS images, UAV aerial images, and target detection tasks in complex backgrounds, the models all showed high detection accuracy and robustness. This benefits from the enhancement of SOD by SPD-Conv and the full utilization by the SIOU loss function [16]. In addition, the introduction of PConv enabled the model to operate efficiently in a constrained computational resource environment, further expanding its application scenarios. In target detection tasks, there was often a trade-off between computational complexity and detection accuracy. However, by introducing PConv and SPD-Conv, the computational complexity of the model was significantly reduced while improving the detection accuracy. The experimental results showed that the improved model achieved significant optimization in both FLOPs and the number of parameters, while also improving the inference speed. This trade-off made models more competitive in real-time detection tasks, such as autonomous driving, video surveillance, and drone target detection.

5 Conclusion

At present, traditional object detection techniques can no longer satisfy the demands of reality. In response to the issues of low recall and large computational resources in the current YOLOv8 algorithm, this study proposed a YOLOv8 model based on loss function and local convolution improvement. The model first introduced SIOU to enhance the loss function of the algorithm and used SlideLoss to adjust its weight calculation method. Then, this study introduced SPD-Conv and more efficient PConv to improve the Backbone layer. The results found that the research model had high prediction accuracy on the training set and validation set. After introducing SPD-Conv structure, using C2f-PC, and improving the loss function, the mAP and AP50 indexes of the model were significantly improved compared with the basic model. This indicated that the model could significantly improve the performance and accuracy of intensive target detection. At the same time, on the COCO and ImageNet datasets, compared with the traditional YOLOv8 algorithm, the number of parameters and FLOPs have increased, but the inference speed has been improved, and mAP has also been significantly improved. This achieved a balance between detection accuracy and detection speed, improving the reliability and effectiveness of object detection. In the case test, the average OR and CLE of the research model improved by 5.7% and 2.3%, which can effectively improve the accuracy of the task. However, this study found that the accuracy is greatly affected by the monitoring scenario during the small sample object detection experiment. Future research will attempt to obtain more scene data, have surveillance video data with a wider time span, and use larger computing resources and more sophisticated models to conduct in-depth research on this specific problem. Furthermore, the study only conducted experiments with fixed parameters, and future work will discuss head-to-head training comparisons with different parameter settings.

References

[1] Carmen Gheorghe, Mihai Duguleana, Razvan Gabriel Boboc, Cristian Cezar Postelnicu. Analyzing Real-Time Object Detection with YOLO Algorithm in Automotive Applications: A Review. Computer Modeling in Engineering & Sciences,2024,141(12):1939-1981. DOI: 10.32604/cmes.2024.054735

[2] Mahmoud Atta Mohammed Ali, Tarek Aly, Atef Tayh Raslan, Mervat Gheith, Essam A. Amin. Advancing Crowd Object Detection: A Review of YOLO, CNN and ViTs Hybrid Approach. Journal of Intelligent Learning Systems and Applications,2024,16(3):175-221. DOI: 10.4236/jilsa.2024.163011

[3] Zhao M, Li W, Li L. Single-frame infrared small-target detection: A survey. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 87-119. DOI: 10.1109/MGRS.2022.3145502

[4] Lei F, Tang F, Li S. Underwater target detection algorithm based on improved YOLOv5. Journal of Marine Science and Engineering, 2022, 10(3): 310.

[5] Li H, Wu Z, Shrivastava A. Rethinking pseudo labels for semi-supervised object detection. Proceedings of the AAAI conference on artificial intelligence. 2022, 36(2): 1314-1322. DOI: 10.1609/aaai.v36i2.20019

[6] Yang J, Shi S, Wang Z. St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(5): 6354-6371. DOI: 10.48550/arXiv.2103.05346

[7] Weiya Shi, Shaowen Zhang, Shiqiang Zhang. CAW-YOLO: Cross-Layer Fusion and Weighted Receptive Field-Based YOLO for Small Object Detection in Remote Sensing. Computer Modeling in Engineering & Sciences,2024,139(6):3209-3231.

[8] Zhongyuan Zhang, Wenqiu Zhu. YOLO-MFD: Remote Sensing Image Object Detection with Multi-Scale Fusion Dynamic Head.Computers, Materials & Continua, 2024,79(5):2547-2563.

[9] Yakui Liu, Xing Jiang, Ruikang Xu, Yihao Cui, Chenhui Yu, Jingqi Yang, Jishuai Zhou. A Novel Foreign Object Detection Method in Transmission Lines Based on Improved YOLOv8n.Computers, Materials & Continua,2024,79(4):1263-1279. DOI: 10.32604/cmc.2024.048864

[10]Xiangyan Tang, Chengchun Ruan, Xiulai Li, Binbin Li, Cebin Fu. MSC-YOLO: Improved YOLOv7 Based on Multi-Scale Spatial Context for Small Object Detection in UAV-View.Computers, Materials & Continua,2024,79(4):983-1003. DOI: 10.32604/cmc.2024.047541

[11]YANG Yuyi, CHEN Liang, ZHANG Jian, LONG Lingchun, WANG Zhenfei. UGC-YOLO: Underwater

Environment Object Detection Based on YOLO with a Global Context Block. Journal of Ocean University of China,2023,22(3):665-674. DOI: 10.1007/s11802-023-5296-z

[12]Abdulghani M.Abdulghani, Mokhles M. Abdulghani, Wilbur L. Walters, Khalid H. Abed. Multiple Data Augmentation Strategy for Enhancing the Performance of YOLOv7 Object Detection Algorithm.Journal on Artificial Intelligence,2023,5(1):15-30. DOI: 10.32604/jai.2023.041341

[13]Ananthakrishnan Balasundaram, Anshuman Mohanty, Ayesha Shaik, Krishnadoss Pradeep, Kedalu Poornachary Vijayakumar, Muthu Subash Kavitha. Zero-DCE++Inspired Object Detection in Less Illuminated Environment Using Improved YOLOv5.Computers, Materials & Continua,2023,77(12):2751-2769. DOI: 10.32604/cmc.2023.044374

[14]Li M, Dong H, Zhang F. A method for top view pedestrian flow detection based on small target tracking. Informatica, 2024, 48(11):1813.-1830. DOI: 10.31449/inf.v48i11.6033

[15]Liu K, Sun Q, Sun D. Underwater target detection based on improved YOLOv7. Journal of Marine Science and Engineering, 2023, 11(3): 677. DOI: 10.1109/ICICML57342.2022.10009683

[16]Yang R, Li W, Shang X. KPE-YOLOv5: an improved small target detection algorithm based on YOLOv5. Electronics, 2023, 12(4): 817. DOI: 10.3390/electronics12040817

[17]Xie Z, Wu G. Optimized Method for Basketball Game Judging by Integrating Faster-RCNN with LK Algorithm. Informatica, 2024, 48(23): 17-31: 3367.-3372. DOI: 10.31449/inf.v48i23.6696

[18]Chen G, Hou Y, Cui T. YOLOv8-CML: A lightweight target detection method for Color-changing melon ripening in intelligent agriculture. Scientific Reports, 2024, 14(1): 14400. DOI: 10.1038/s41598-024-65293-w

[19]Hui Y, You S, Hu X. SEB-YOLO: An Improved YOLOv5 Model for Remote Sensing Small Target Detection. Sensors, 2024, 24(7): 2193. DOI: 10.3390/s24072193

[20]He J, Luo J, Fu C. The Occlusive Basketball Player Detection Algorithm Based on Posture Recognition Assisted Feature Alignment. Informatica, 2024, 48(21):114-11. DOI: 10.31449/inf.v48i21.6695

[21]Wang Y, Tian Y, Liu J,. Multi-stage multi-scale local feature fusion for infrared small target detection Remote Sensing, 2023, 15(18): 4506-4520. DOI: 10.3390/rs15184506

[22]Ling S, Chen L, Wu Y. ACANet: Attention-based context-aware network for infrared small target detection. The Journal of Supercomputing, 2024, 80(12): 17068-17096. DOI: 10.1007/s11227-024-06067-z