International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 3452-3463

https://doi.org/10.70517/ijhsa463291

Optimization and Accuracy Enhancement of Target Detection Algorithm Based on Improved Convolutional Neural Network Structure

Tao Wang[1], Yuming Xue[1], Luoxin Wang[1,*], Tianen Li[2] and Hongli Dai[1]
[1] Institute of New Energy Intelligence Equipment, Tianjin Key Laboratory of Film Electronic & Communication Devices, School of Integrated Circuit Science and Engineering, Tianjin University of Technology, Tianjin, 300384, China
[2] Institute of Mechanical Engineering, Baoji University of Arts & Science, Baoji, Shaanxi, 721013, China

Corresponding authors: (e-mail: orwellx@tjut.edu.cn).

Abstract The study takes the target detection algorithm based on convolutional neural network (YOLOv5) as the optimization object, for the problem of limited sensory field existing in the standard convolution, a mask module conforming to the characteristics of the distribution of the effective sensory field is designed to adjust the convolutional kernel weights, and a kind of improved deformable convolution (MDC) is proposed, and the MDCYOLO detection model is constructed. Small and large target detection experiments are performed on the insulator dataset and Vis Drone2019-DET dataset, respectively. The experimental results show that the detection accuracy of MDCYOLO is greatly improved compared to Y0L0v5 using standard convolution, which also reduces the computation of the detection model and improves the detection speed. The detection accuracy of the MDCYOLO model outperforms that of other mainstream models, regardless of whether it performs small target detection or large target detection. The target detection optimization method based on improved convolutional neural network structure designed in this paper has obvious advantages in detection accuracy and speed.

Index Terms Convolutional neural network, YOLOv5, MDC, Target detection

I. Introduction

Along with innovations in information technology and the spread of the Internet, the world is rapidly moving into a more intelligent era. This transformation is accompanied by the generation of a large amount of information and data, which puts higher demands on the ability to process and analyze these data. In this context, the application of Convolutional Neural Networks (CNNs) is particularly important. CNNs, in deep learning techniques, specialize in processing image data and have become an effective tool for tasks such as target detection [1]. The target detection technology can provide support for automatic driving, security monitoring, face detection and other fields, especially face detection technology, which is particularly critical due to its wide application in key fields such as security, identity verification and healthcare [2]-[4]. The core of target detection technology lies in identifying and localizing specific objects in an image, which not only needs to classify the types of objects present in the image, but also accurately mark the location of each object. By using anchor frames to identify the specific location of an object, it is possible to understand the surrounding environment more accurately and react accordingly [5]-[8]. Applying CNN structures to the task of target detection demonstrates powerful applications by being able to learn advanced feature representations from the data to achieve accurate recognition and localization of targets [9], [10].

However, despite the remarkable achievements of CNN target detection algorithms, they face many challenges. First, complex scene environments and target deformations make the task of target detection exceptionally difficult, e.g., the target may have multiple poses, occlusions, or lighting variations, all of which affect the accuracy and robustness of the detection algorithms [11]-[13]. Secondly, traditional CNN structures often suffer from insufficient local sensing ability and inadequate utilization of contextual information, resulting in limited detection performance in complex scenes [14]-[16]. In view of this, it is of great significance to explore the problem of optimizing the structure of convolutional neural networks applied to the task of target detection, focusing on their innovations and breakthroughs in improving accuracy, robustness and efficiency.

To address the aforementioned challenges, researchers have continuously proposed various innovative solutions. Literature [17] shows that the complex interaction between human body and objects is a major point of difficulty faced by target detection algorithms, for this reason, a deep convolutional neural network (DCNN) learning model based on Black Widow Optimization (BWO) is proposed for target detection and tracking of video images, which exhibits high detection performance. Literature [18] addresses the problem of inaccurate localization of target

detection systems supported by traditional high-capacity neural networks and proposes to improve CNNs using Bayesian optimization and structural prediction methods in order to improve the localization accuracy of target detection models. Literature [19] designed a global convolutional neural network structure combined with optimized nonlinear activation function, which can fully grasp the local and global information of the feature maps of different layers in the network, and applied it to the task of target recognition in the field of warehousing and logistics, which significantly improves the detection accuracy of the model. Literature [20] examined the application of convolutional neural network based on YOLOv2 network structure in the task of target detection, and the training and testing found that it has a significant detection effect for small targets in complex scenes. Literature [21] proposes a multi-scale deformable convolutional target detection network to meet the challenge of target detection for dense and randomly transformed objects, and shows that the proposed deep convolutional network can effectively extract the multi-scale features of the detected objects under geometrical deformation, and makes an outstanding contribution to the improvement of recognition accuracy. Literature [22] generates CNN network architectures capable of providing accurate and fast feedback on spam-detected targets by adapting multiple types of single-subject detectors (SSDs) and region suggestion networks (RPNs) and combining different loss optimization methods and other techniques. Although the above methods improve the efficiency and reliability of deep learning models in the field of target detection, the enhancement of the network's ability to extract target features and enhance its generalization ability and robustness in complex scenarios is weak, and further research is urgently needed.

Convolutional neural network based on standard convolution has the problem of limited sensory field, this paper proposes an improved deformable convolution (MDC) based on channel attention, and uses MDC to replace the standard convolution on the basis of YOLOv5, and proposes an improved MDCYOLO model. Starting from the distribution characteristics of the effective sensory field, two masks for adjusting the weight distribution of deformable convolution are designed, and then SE channel attention is introduced to make the model learn different mask feature distributions in different channels of the feature map. Finally this paper compares the detection accuracy and speed of the MDCYOLO model and the existing target detection model on the insulator dataset and the Vis Drone2019-DET dataset to verify the feasibility of the improved method.

II. Method

YOLOv5 is a state-of-the-art target detection algorithm based on convolutional neural networks, which utilizes the feature extraction capability of convolutional neural networks to achieve target localization and classification quickly and accurately. While the standard convolution suffers from the problem of limited sensory field, accordingly, an improved deformable convolution (MDC) is proposed in this chapter, which is further validated on the YOLOv5 model and the improved model is called MDCYOLO.

II. A. Convolutional Neural Networks

Deep learning convolutional neural network (CNN) is a neural network model in computer vision. It is widely used in tasks such as image recognition, target detection, and image segmentation [23]. The structure of convolutional neural network is shown in Figure 1. The main components of convolutional neural network include convolutional layer, pooling layer and fully connected layer. The convolutional layer is the core of the convolutional neural network, which is mainly responsible for the convolution operation on the input data to extract the features of the input data. The pooling layer is mainly responsible for downscaling the output of the convolutional layer to reduce the parameters and computation of the model and prevent overfitting. The fully connected layer is mainly responsible for transforming the output of the pooling layer into the final classification result.

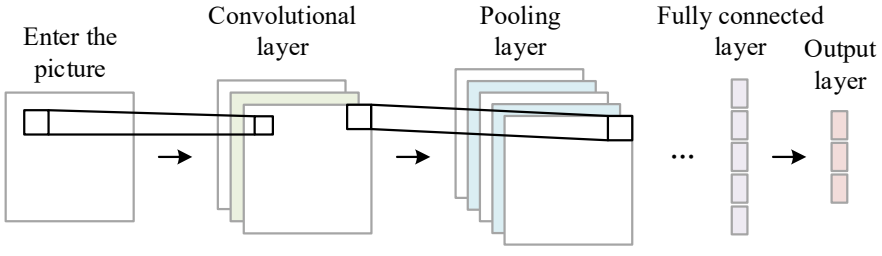


Figure 1: Convolutional neural network

The training process of convolutional neural network is as follows:

(1) Forward propagation

In the forward propagation stage, the input data X is passed through the layers of the network, and each layer l performs specific transformations $f^{(l)}$ on the data, which include convolution operations, activation functions, pooling operations, and so on. Eventually, the network outputs the predicted value $\hat{Y}$, which is transformed into a probability distribution in the classification task by means of the softmax function. The mathematical representation of forward propagation is:

$$\hat{Y} = f^{(L)}(f^{(L-1)}(\cdots f^{(L)}(X))) \tag{1}$$

where L denotes the number of network layers.

(2) Loss function calculation

The predicted output $\hat{Y}$ of the model is obtained and the loss function $L(\hat{Y},\hat{Y})$ is used to calculate the difference between the model output and the true label Y. The choice of loss function depends on the specific task, e.g., cross-entropy loss is used for classification tasks. The loss function provides an optimization objective for model training:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i(Y_i,\hat{Y}_i) \tag{2}$$

where N is the number of samples.

(3) Backpropagation

Backpropagation is the process of calculating the gradients $\nabla_w L$ and $\nabla_b L$ of the loss function with respect to the parameters of the network (weights W and bias b) using the chain rule. These gradients provide information on how to adjust the parameters to minimize losses. For each layer L, the computed gradients indicate how the loss function varies with small changes in the parameters of that layer.

(4) Parameter Updates

Finally, using an optimization algorithm, the network parameters are updated based on the computed gradients. The general form of the update formula is:

$$W^{(l)} = W^{(l)} - \eta\nabla_w L^{(l)} \tag{3}$$

$$b^{(l)} = b^{(l)} - \eta\nabla_b L^{(l)} \tag{4}$$

where η is the learning rate, which can determine the step size of the parameter update.

These four stages constitute an iterative cycle of CNN training. This cycle continues throughout the training process until the model's performance on the validation set is no longer significantly improved or a predetermined number of iterations is reached.

II. B.YOLOv5 algorithm

YOLO is a real-time target detection model based on deep learning. The core idea of YOLO algorithm is to utilize the whole graph as the input to the network and regress the bounding box and category probabilities in the output layer. The main feature of YOLO algorithm is to solve the target detection problem as a regression problem, predicting the category and location of the target directly in the forward propagation, which effectively improves the speed of target detection [24].

YOLOv5, as a popular detection model of YOLO series, the overall structure includes several parts such as feature extraction network, feature fusion module, prediction header, etc., and its model structure is shown in Fig. 2. The Input structure is mainly responsible for preprocessing the original input image so that the image can be processed correctly by the model. Backbone is responsible for converting the input image into a high-level abstract feature representation [25].

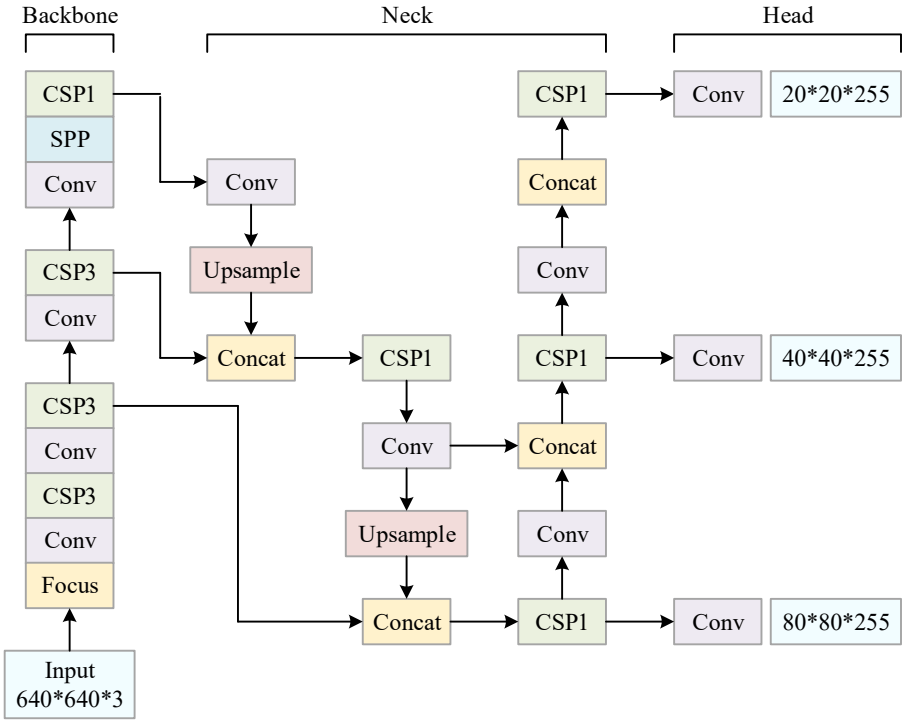


Figure 2: YOLOv5 structural diagram

The loss function of YOLOv5 consists of three components, including the bounding box regression loss, the target confidence loss, and the category prediction box loss.

(1) Bounding box regression loss

This loss is used to measure the difference between the bounding box predicted by the model and the true bounding box. In YOLOv5, the bounding box is represented by the center point coordinates (x, y) and width and height (w, h). The loss function usually uses mean square error (MSE) or smoothed L1 loss to calculate the difference between the predicted value and the true value.YOLOv5 uses CIoU as a measure of the box loss, which is an extension of IoU that takes into account the distance from the centroid of the box, the aspect ratio, and the overlap area.The CIoU loss can be expressed as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{5}$$

where IoU denotes the intersection and concurrency ratio of the predicted frame to the real frame, $\rho^2(b, b^{gt})$ is the Euclidean distance between the center point of the predicted frame b and that of the real frame b^{gt}, c is the length of the diagonal line containing the smallest region of closure of the two frames, v takes into account the concurrency of the width-to-height ratios, and α is the weighting parameter .

(2) Target confidence loss

The target confidence loss measures the difference between the predicted probability distribution and the true probability distribution. The target confidence represents the probability that a target exists within a certain bounding box. In YOLOv5, for each bounding box, a target confidence is predicted. For a bounding box that contains a target, the true target confidence is 1. For a bounding box that does not contain a target, the true target confidence is 0. The loss function usually uses a binary cross-entropy loss to calculate the difference between the predicted value and the true value. The target confidence loss can be expressed as:

$$L_{obj} = -(y_0 \log(\hat{y}_0) + (1 - y_0) \log(1 - \hat{y}_0)) \tag{6}$$

where y_0 is the true label indicating whether the object exists or not, and $\hat{y}_0$ is the confidence level of the model's prediction that the object exists.

(3) Category prediction loss

The category prediction loss is used to measure the difference between the predicted category distribution and the true category labels. The cross-entropy loss is usually used to calculate the category loss, and the category prediction loss function can be expressed as:

$$L_{class} = -\sum_{c=1}^{C} y_{o,c} \log(\hat{y}_{o,c}) \tag{7}$$

The total loss function of YOLOv5 is a weighted sum of the above component losses and can be expressed as:

$$L_{total} = \lambda_{box} L_{CIoU} + \lambda_{class} L_{class} + \lambda_{obj} L_{class} \tag{8}$$

Among them, $\lambda_{box}, \lambda_{class}$ and λ_{obj} are the weighting coefficients used to balance the contributions of different loss components.

The loss function of YOLOv5 is able to integrate the accuracy of the detection frame, the accuracy of the category prediction, and the existence of the object, which can improve the overall performance and accuracy of the model while maintaining high speed detection.

II. C.Improving deformable convolutional networks

II. C. 1) Improving deformable convolution

Existing research work on the effective receptive field of convolutional neural networks shows that an important factor affecting the performance of computer vision tasks such as classification and detection is the effective receptive field of the model. In this paper, based on DCNv2, we design a mask that conforms to the distribution law of the effective perceptual field to adjust the weight distribution of convolution kernel, the mask presents the characteristics of gradual decay from the center region to the edge of the feature map, so as to realize the simultaneous adjustment of convolution kernel shape and weight, and this improved deformable convolution is named as MDC [26]. The MDC first inputs the feature map and adjusts the weights of the corresponding convolution kernel by mask, then learns 2N offsets in the x and y directions at the grid points of the feature map participating in the convolution operation through the deformable convolution v2, and finally restricts the degree of offset of the deformable convolution by the modulation parameter. Compared with standard convolution, MDC can adapt to object pose changes and can extract important feature information in images more effectively. In this paper, we provide two ideas to design the mask module, one is to let the mask decay as a polynomial function from the center region to the edges, and the other is to let the mask decay as a two-dimensional Gaussian distribution function from the center region to the edges, according to the two ideas of the above design of the mask module are called as (PM) and (GM), respectively.

PM assigns corresponding weights based on the distance between the corresponding pixel point and the center point on the feature map. The closer the pixel point is to the center point, the higher its weight value is. To calculate the weight of point (i, j) on the input feature map x, we multiply the difference between the coordinates of i, j and the coordinates of the center point of the feature map, and divide by the area of the corresponding feature map area. In this way, the closer the pixel point on the feature map is to the center point, the larger the value is taken, in order to comply with the law of decay from the center to the edge, we invert the calculation result and add a constant 1 in front of it, assuming that the size of the input feature map is $W \times H$, and the pixel point k is located at (i, j), and the corresponding weight on the mask can be calculated by Equation (9):

$$\lambda_k = I - \left(\frac{(i - \frac{W}{2}) \cdot (j - \frac{H}{2})}{\frac{W}{2} \cdot \frac{H}{2}} \right)^n \tag{9}$$

As mentioned earlier, the effective receptive field exhibits a Gaussian-like distribution in the field of view domain, while the standard two-dimensional normal distribution is mathematically expressed as shown in Equation (10):

$$f(x, y) = (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1} \tag{10}$$

$$\exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right] \tag{11}$$

GM calculates the weights of the corresponding pixel points on the feature map according to the two-dimensional Gaussian normal distribution, in order to simplify the calculation, this paper assumes that the directions of the feature map x and y are independent of each other, i.e., the correlation coefficient $\rho = 0$, and the coordinates of the center pixel point on the feature map are (0, 0). In addition the constant coefficient in front of exp is abbreviated as A, and the corresponding mask weight calculation formula is shown in Equation (12):

$$\lambda_k = A \exp\left[-\frac{1}{2}\left(\frac{(x-W/2)^2}{\sigma_l^2}+\frac{(y-H/2)^2}{\sigma_2^2}\right)\right] \tag{12}$$

In this paper, we use mask to adjust the weight of convolution kernel on the basis of deformable convolution v2 in DCNv2, i.e., the input feature map x first uses mask to adjust the distribution of feature information, enhance the useful information of the objects within the effective receptive field, and weaken the useless information of the edge region, so as to exclude the interference of the background and the impurity to improve the feature extraction ability of the network, and then after deformable convolution v2 to enhance the network's The deformable convolution v2 enhances the network's ability to adapt to the object's posture and shape, and finally makes the network's sensory field fall near the target's region. For the input feature map x, the convolution kernel w, the output feature map y on the point p can be expressed through the formula (13):

$$y_p = \sum_{k-j}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \lambda_k \tag{13}$$

where k denotes the points of the grid on the convolution kernel, Δp_k denotes the offset of the grid points learned by the deformable convolution, Δm_k denotes the modulation parameter of the deformable convolution, and λ_k denotes the weight value of the corresponding point on the mask.

II. C. 2) Channel Attention Module

The current PM generates corresponding weights for all channels of the feature map in the same way, while different channel dimensions of the feature map contain feature information of different importance, especially in target detection, where the model localizes and detects according to the important feature information embedded in the channels, while some channels contain useless information that can interfere with the results. To address this problem, this paper introduces the SE channel attention mechanism to learn an importance level characterization of PM on each channel, so as to improve MDC's ability to filter and process the information.

Learning the importance characterization of each channel of mask using SE is shown in Fig. 3.SE channel attention can be divided into two Tai parts: compression and excitation. Among them, the compression part needs to compress the global spatial information in the spatial dimension first, then learn the importance of different channels in the channel dimension, and finally assign different weights to each channel through the incentive part. For the input feature map x with dimension $H \times W \times C$, the compressed part will first use global average pooling to compress the feature map to the $1 \times 1 \times C$ dimension, and the compressed feature map will be sent to the fully connected layer to obtain the importance weight matrix S of each channel, and then use S to assign the weight of the double x.

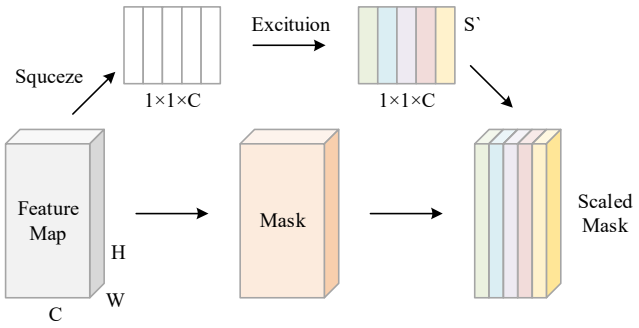


Figure 3: The use of SE representation of the learning of mask channels

For MDC, in this paper, we first use global average pooling to compress the PM of $H \times W \times C$ dimension into $1 \times 1 \times C$ dimension, and then go through two fully connected layers to get the importance matrix S* of different channels of the PM, and then multiply S* with the mask module and then adjust the weights of the different channels,

and the MDC formula based on the attention of the channel is as follows The MDC calculation formula based on channel attention is shown in Equation (14):

$$y_p = \sum_{c-l}^{C} \sum_{k-l}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \lambda_k S_C \tag{14}$$

where $_c$ denotes each channel and S_c denotes the importance weight value of each channel.

II. C. 3) Improved deformable convolutional network based on YOLOv5

Following the idea of verifying the effectiveness of deformable convolution in the DCN family of models, this paper uses MDC to replace the standard convolution to verify the effectiveness of the improved deformable convolution in the Y0L0v5 model network [27].The backbone network in YOLOv5 doesn't use convolution alone, but puts the convolution conv and the batch-regularized BN and SiLU activation functions together, called the Conv module, and in this paper, we replace the conv operation in the Conv module with MDC and call the improved Conv module $Conv_{MDC}$. In addition, the C3 module, as an important feature extraction unit in Y0L0v5, mainly adopts the CSP structure, and the module is mainly divided into two branches, and the outputs of the Conv and BottleNeck modules cascaded on the main branch and the outputs of the Conv on the sub-branch are connected through Concat, and then the outputs are obtained through a Conv. Further, this paper replaces the Conv module in the C3 structure with $Conv_{MDC}$, and refers to the improved C3 module as $C3_{MDC}$. The YOLOv5 model can be divided into n,s,m,l,x structures according to the needs of different usage scenarios, and this paper subsequently conducts experiments on the YOLOv5n network. Experiments are conducted and the YOLOv5 model using MDC is called MDCYOLO, and the structure of MDCYOLO network is shown in Fig. 4.

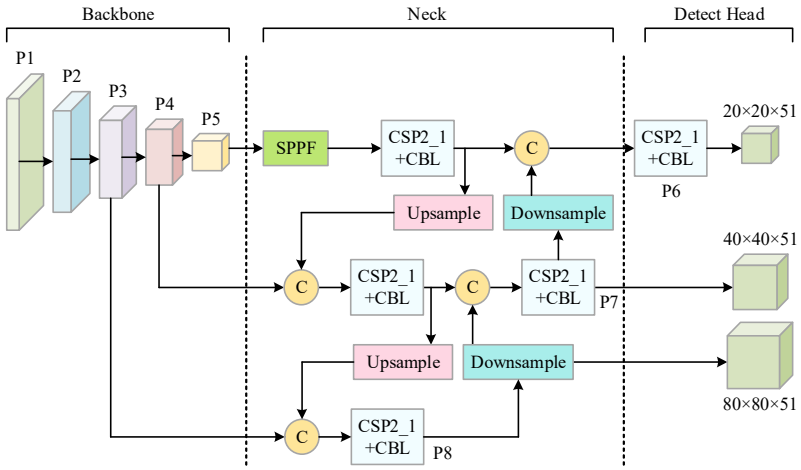


Figure 4: MDCYOLO Network structure

III. Results and Discussion

In this chapter, two experiments, small target detection and large target detection, are designed respectively to verify the effect of MDCYOLO model optimization and accuracy improvement.

III. A. Assessment of indicators

There are many evaluation indexes for the detection accuracy of the object detection model, and the experimental results in this paper mainly evaluate the model from the precision, recall and mAP@0.5.

Precision is mainly used to evaluate whether the prediction target is accurate, and recall is mainly used to evaluate whether all the detection targets are found. The formulas for calculating precision and recall are shown in Eq. (15) and Eq. (16):

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{15}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{16}$$

where TP refers to the number of correct detections (true positives), FP refers to the number of false detections (false positives), and FN refers to the number of missed detections (false negatives).

The AP value is the area under the P-R curve consisting of recall in the horizontal coordinate and accuracy in the vertical coordinate, and the formula is shown in Equation (17):

$$AP = \int_0^1 PdR \tag{17}$$

The formula for the mAP value is shown in equation (18):

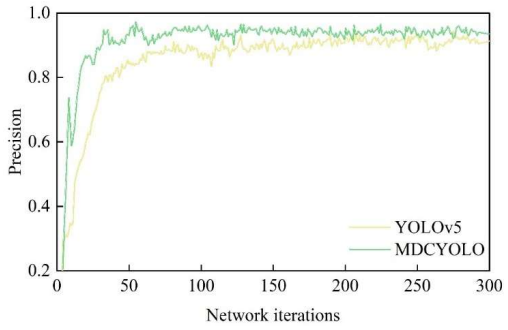
$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{18}$$

where N refers to the total number of detection categories, which is 2 in this article. mAP@0.5 is the average value of an AP when IoU=0.5.

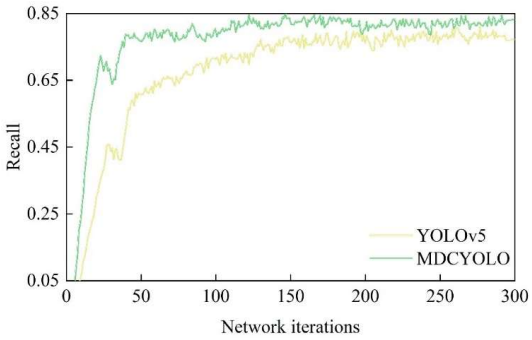
In target detection, detection speed is also one of the evaluation indicators. Commonly used speed metrics are frames per second (FPS) and floating-point arithmetic (GFLOPs).

III. B. Network Model Training

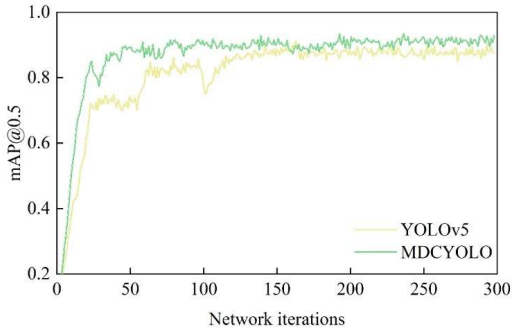
The experiments use Stochastic Gradient Descent (SGD) optimizer [28] to train the MDCYOLO model. At the beginning of the training, in order to prevent the instability of the target detection model due to the large selection of the learning rate, Warmup is used to warm up the learning rate to maintain the stability of the detection model in the deeper layers. The initial learning rate during the experimental training is 0.01, the final learning rate is 0.1, the momentum is 0.937, and the batch size is 32. To prevent model overfitting, the optimizer weight decay is set to 0.0005, and the optimal model weights are obtained after 300 iterations of training. The learning rates of the normalization, weight and bias layers in the experiment are lr0, lr1 and lr2, respectively.



(a) Precision



(b) Recall



(c) mAP@0.5

Figure 5: Experimental results of each model in the insulator data set

III. C. Small Target Detection Experiment

III. C. 1) Data sets

There are a wide variety of datasets related to small target defect detection, this experiment collects and organizes the existing insulator images in the electric power inspection, and finally summarizes a total of 1,366 insulator images, including 820 images of normal insulators and 546 images of defective insulators. The number of existing samples is small, the target information is not obvious, it is difficult to complete the task of target detection, so the existing dataset for image segmentation, image enhancement, normalized size and other data preprocessing. And with the help of dataset annotation tool LabelImg, the collected and organized insulator dataset is labeled with labels, and the labeled dataset is organized into YOLO format, which is randomly divided into training set, validation set, and test set in the ratio of 7:2:1.

III. C. 2) Ablation experiments

In order to better analyze the influence of the improved method on the defect detection of small targets, the detection models of YOLOv5 and MDCYOLO were trained under the same hyperparameters. The experimental training results of YOLOv5 and MDCYOLO detection models on the insulator dataset are shown in Figure 5. (a)~(c) represent the accuracy, recall and mAP@0.5 of the two detection models on the insulator dataset, respectively. The MDCYOLO detection model has a significant improvement trend in the three evaluation indicators of accuracy, recall and mAP@0.5. It shows that the MDCYOLO model has greatly improved the detection accuracy of insulator defects.

The main purpose of the ablation experiment was to verify the effectiveness of the improved object detection model. The size of the insulator images entered by the network is 640×640, and the test results are shown in Table 1 and Figure 6. It can be seen that the mAP@0.5 and precision of the MDCYOLO model proposed in this paper are 90.2% and 93.6%, respectively, the FPS of the model is 20.6, and the GFLOPs are reduced to 50.9% of the traditional YOLOv5 model. It can be seen that after replacing the standard convolution with MDC in the Y0L0v5 model network, the detection speed of the model is greatly improved without affecting the detection accuracy, and the complexity of the detection model is also reduced.

Table 1: Ablation experiment results

Network model	Precision (%)			FPS	GFLOPs
	Whole	Insulator	Insulator defect		
YOLOv5	86.9	89.6	88.9	83.6	16.9
MDCYOLO	93.3	92.8	93.6	20.6	8.6

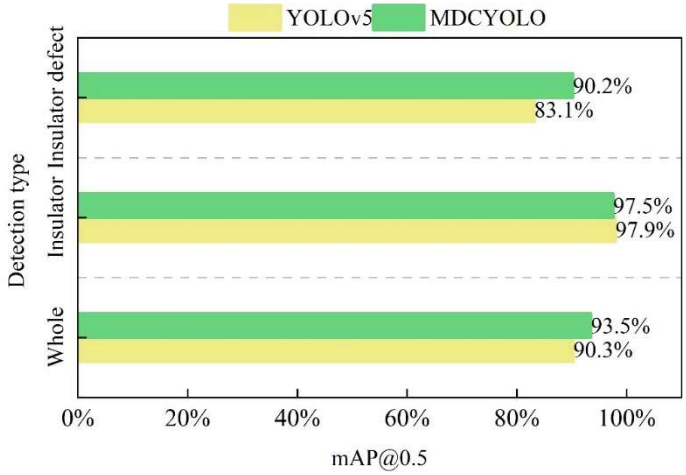


Figure 6: mAP@0.5 of different models

III. C. 3) Comparison experiments

In order to further verify the performance advantages of the MDCYOLO model in the detection of small target defects, different detection models are used for testing, and the improved MDCYOLO model based on YOLOv5 is compared with other mainstream models Faster-RCNN, YOLOv3 and YOLOv5 detection models, and the experimental

comparison results are shown in Table 2. From the experimental results, it can be seen that the Faster-RCNN detection model has the highest recall, but the detection accuracy and mAP@0.5 are very low, which is very large, indicating that the detection speed is the slowest. The accuracy of the MDCYOLO detection model is higher than that of the YOLOv5 detection model, slightly lower than that of the YOLOv3 detection model, and the recall rate is basically close to that of the YOLOv3 detection model, but its floating-point operation is about 94.4% lower than that of the YOLOv3 detection model and about 49.1% lower than that of the YOLOv5 detection model, indicating that the MDCYOLO detection model is significantly better than the YOLOv3 detection model and the YOLOv5 detection model in terms of detection speed. In terms of FPS, the MDCYOLO detection model is smaller than other networks. The experimental results show that the MDCYOLO detection model has a great improvement in the detection speed while ensuring the detection accuracy.

Table 2: Performance comparison results of various models

Network model	Precision (%)	Recall (%)	mAP@0.5 (%)	FPS	GFLOPs
Faster-RCNN	70.2	97.1	77.9	1080.4	1231.8
YOLOv3	94.6	84.9	83.1	133.9	154.6
YOLOv5	88.9	78.3	83.1	83.6	16.9
MDCYOLO	93.6	84.3	90.2	20.6	8.6

III. D. Large Target Detection Experiment

III. D. 1) Data sets

In this thesis, the Vis Drone2019-DET dataset is used as the large target detection dataset, which covers a wide range of street scenes in nearly 15 cities in China, with the farthest distance spanning thousands of kilometers, including urban, suburban, highway, and rural environments, and also includes scenes of pedestrians, cars, bicycles, and other transportation vehicles in different brightness and darkness under the weather conditions of peak pedestrian commuting and weekend pedestrian scenes. It also includes scenes of pedestrians in cars, bicycles, and other means of transportation under different weather conditions, during peak commuting hours, and on weekends when pedestrians are scarce. The collected images were manually annotated with about 600,000 target frames and centroids to ensure the rigor and completeness of the dataset.

The VisDrone2019-DET dataset has a total of 10 categories, including pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and cars. The distribution of the number of tags corresponding to the different categories is shown in Figure 7. Pedestrians (32.5%) and cars (33.3%) make up the vast majority of the dataset.

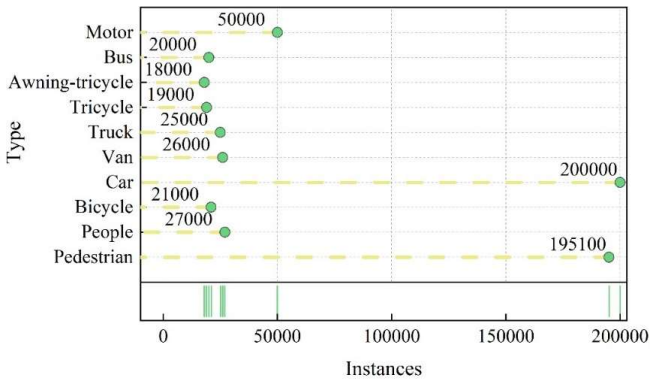


Figure 7: The number of tags corresponding to different categories

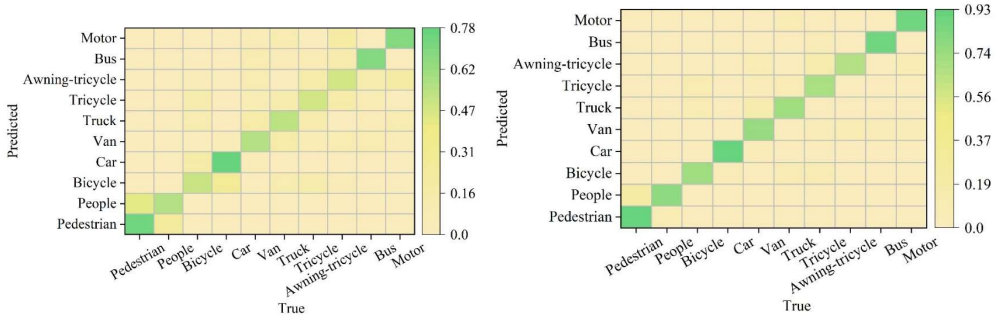
III. D. 2) Ablation experiments

The results of the evaluation metrics before and after the improved deformable convolutional incorporation into the YOLOv5 model in the Vis Drone2019-DET test set are shown in Table 3.The MDCYOLO model compared to the original YOLOv5 model improved the accuracy rate by 15.9% in the test set, and the floating-point computation decreased to 12.6.The MDCYOLO model outperforms both the speed of detection and the accuracy of the large target detection in the YOLOv5 model.

Table 3: Ablation experiment results

Network model	Precision (%)	Recall (%)	mAP@0.5 (%)	FPS	GFLOPs
YOLOv5	83.2	81.6	82.6	87.6	19.9
MDCYOLO	96.4	93.2	90.6	26.9	12.6

Confusion matrix plots are generally used in target detection to measure the magnitude of image classification accuracy. In the confusion matrix, the color blocks of different colors indicate the experimental results corresponding to different categories, and the depth of the color blocks corresponding to the values on the diagonal line is related to the accuracy of the corresponding category. The confusion matrices derived from the different models in the ablation experiments are shown in Fig. 8, and (a) and (b) denote the confusion matrices derived from the YOLOv5 and MDCYOLO models, respectively.Car has the highest detection accuracy in each model due to the fact that there are the largest number of Car categories in the dataset, and the largest number of cars captured in the aerial photographs. The detection accuracy of other categories like Pedestrian and Tricycle in the dataset is also proportional to the percentage of that category in the dataset, except for Bus, which has a higher value of detection accuracy but does not account for a large percentage of the dataset, guessing the reason may be that the bus itself is long and has a larger target, which makes it easier to detect compared to other targets. Comparing the depths of the corresponding color blocks on the diagonal of (a) and (b), it is obvious that the color of figure (b) is darker, indicating that the detection accuracy of the MDCYOLO model is higher.



(a) YOLOv5 (b) MDCYOLO

Figure 8: The confusion matrix diagram of different models in the experiment

III. D. 3) Comparison experiments

The experimental results of different models in the Vis Drone2019-DET validation set are shown in Table 4. In object detection, the accuracy, recall, mAP@0.5 and detection speed of the MDCYOLO model are excellent. The image quality contained in the validation set is better than that of the test set, with fewer occlusion targets, resulting in better detection.

Table 4: Performance comparison results of various models

Network model	Precision (%)	Recall (%)	mAP@0.5 (%)	FPS	GFLOPs
Faster-RCNN	71.2	76.1	71.9	1063.4	1563.8
YOLOv3	84.6	84.6	86.1	124.9	163.1
YOLOv5	84.9	83.6	83.4	82.3	19.6
MDCYOLO	97.6	94.6	91.4	25.8	11.3

IV. Conclusion

In this paper, MDC is used to replace the standard convolution in the YOLOv5 algorithm to construct an improved MDCYOLO model. Two experiments, small target detection and large target detection, were designed to verify the optimization effect of the YOLOv5 algorithm based on the improved convolutional neural network structure. In the small target detection experiment, the mAP@0.5 and accuracy reached 90.2% and 93.6%, respectively, and the number of frames per second (FPS) and floating-point operations (GFLOPs) of the model decreased to 20.6 and 8.6. Compared with the Faster-RCNN, YOLOv3 and YOLOv5 models, the MDCYOLO model has obvious advantages in the detection accuracy and speed of small target datasets. In the large target detection experiment, the accuracy of the MDCYOLO model in the Vis Drone2019-DET test set, and the YOLOv5 model improved by

15.9%, which is also excellent in all indicators compared with the mainstream models. The proposed method is effectively verified on two datasets.

References

[1] Cong, S., & Zhou, Y. (2023). A review of convolutional neural network architectures and their optimizations. Artificial Intelligence Review, 56(3), 1905-1969.

[2] Feng, D., Harakeh, A., Waslander, S. L., & Dietmayer, K. (2021). A review and comparative study on probabilistic object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(8), 9961-9980.

[3] Alamri, F. (2025). Comprehensive study on object detection for security and surveillance: A concise review. Multimedia Tools and Applications, 1-32.

[4] Chen, W., Huang, H., Peng, S., Zhou, C., & Zhang, C. (2021). YOLO-face: a real-time face detector. The Visual Computer, 37, 805-813.

[5] Leonard, J. K. (2019). Image classification and object detection algorithm based on convolutional neural network. Science Insights, 31(1), 85-100.

[6] Yao, H., Fan, Y., Liu, Y., Cao, D., Chen, N., Luo, T., ... & You, Z. (2024). Development and optimization of object detection technology in civil engineering: A literature review. Journal of Road Engineering.

[7] Thakare, P., & V, R. S. (2024). Advanced pest detection strategy using hybrid optimization tuned deep convolutional neural network. Journal of Engineering, Design and Technology, 22(3), 645-678.

[8] Becherer, N., Pecarina, J., Nykl, S., & Hopkinson, K. (2019). Improving optimization of convolutional neural networks through parameter fine-tuning. Neural Computing and Applications, 31, 3469-3479.

[9] Rodriguez-Conde, I., Campos, C., & Fdez-Riverola, F. (2022). Optimized convolutional neural network architectures for efficient on-device vision-based object detection. Neural Computing and Applications, 34(13), 10469-10501.

[10] Dhillon, A., & Verma, G. K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 9(2), 85-112.

[11] Loussaief, S., & Abdelkrim, A. (2018). Convolutional neural network hyper-parameters optimization based on genetic algorithms. Int. J. Adv. Comput. Sci. Appl, 9(10), 252-266.

[12] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2020). Improved inception-residual convolutional neural network for object recognition. Neural Computing and Applications, 32(1), 279-293.

[13] Zhang, S., Shao, Z., Huang, X., Bai, L., & Wang, J. (2021). An internal-external optimized convolutional neural network for arbitrary orientated object detection from optical remote sensing images. Geo-Spatial Information Science, 24(4), 654-665.

[14] Jaganathan, T., Panneerselvam, A., & Kumaraswamy, S. K. (2022). Object detection and multi-object tracking based on optimized deep convolutional neural network and unscented Kalman filtering. Concurrency and Computation: Practice and Experience, 34(25), e7245.

[15] Wang, K., & Liu, M. (2021). A feature-optimized Faster regional convolutional neural network for complex background objects detection. IET Image Processing, 15(2), 378-392.

[16] Cheng, G., Han, J., Zhou, P., & Xu, D. (2018). Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Transactions on Image Processing, 28(1), 265-278.

[17] Mukilan, P., & Semunigus, W. (2021). Human object detection: an enhanced black widow optimization algorithm with deep convolution neural network. Neural Computing and Applications, 33(22), 15831-15842.

[18] Zhang, Y., Sohn, K., Villegas, R., Pan, G., & Lee, H. (2015). Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 249-258).

[19] An, F. P., Liu, J. E., & Bai, L. (2022). Object recognition algorithm based on optimized nonlinear activation function-global convolutional neural network. The visual computer, 38(2), 541-553.

[20] Dong, E., Zhu, Y., Ji, Y., & Du, S. (2018, August). An improved convolution neural network for object detection using YOLOv2. In 2018 IEEE international conference on mechatronics and automation (ICMA) (pp. 1184-1188). IEEE.

[21] Cao, D., Chen, Z., & Gao, L. (2020). An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks. Human-centric Computing and Information Sciences, 10(1), 14.

[22] Melinte, D. O., Travediu, A. M., & Dumitriu, D. N. (2020). Deep convolutional neural networks object detector for real-time waste identification. Applied Sciences, 10(20), 7301.

[23] Mukilan, P. & Semunigus, Wogderess. (2022). Human and object detection using Hybrid Deep Convolutional Neural Network. Signal, Image and Video Processing,16(7),1-11.

[24] Hao Meng & Jieqing Tan. (2024). An Improved Algorithm for Small Target Detection based on YOLO. Frontiers in Computing and Intelligent Systems,10(3),18-22.

[25] Zukai Sun,Ruzhi Xu,Xiangwei Zheng,Lifeng Zhang & Yuang Zhang. (2024). A forest fire detection method based on improved YOLOv5. Signal, Image and Video Processing,19(1),136-136.

[26] Zhang Shihao,Li Yang,Song Mingzhao & Ma Xiao. (2023). Concrete surface crack detection method based on improved deformable convolution with YOLOv5.(eds.) Qilu University of Technology,Shandong Academy of Sciences (China)

[27] Liu Kang,Lv Zhongliang,Xia Kewen,Zhou Chuande,Lu Zhenyu,Zuo Hailun... & Chen Xuanlin. (2023). Improved YOLOv5 based on deformable convolution and efficient decoupled head for pill surface defect detection.(eds.) Chongqing University of Science and Technology (China)

[28] Benjamin Gess,Sebastian Kassing & Nimit Rana. (2024). Stochastic Modified Flows for Riemannian Stochastic Gradient Descent. SIAM Journal on Control and Optimization,62(6),3288-3314.