Electrical Power and Energy Systems 172 (2025) 111088

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Optimizing power system edge computing with a high-performance and light-weight YOLO-based substation equipment defect detection network

Qian Wang [a], Rui Liu [a], Sichen Qin [a,*], Jiawei Pu [b], Rong Shi [c], Yulu Wang [c]

[a] Xi'an University of Technology, Xi'an 710048, China
[b] Ultra High Voltage Company of State Grid Shaanxi Electric Power Co., Ltd., Xi'an 710026, China
[c] State Grid Shaanxi Electric Power Company Economic Research Institute, Xi'an 710065, China

ARTICLE INFO

Keywords:
C3-DCNv2
CAM
Defect detection
Substation equipment
Sigma-CIoU

ABSTRACT

In the current fast-paced development of the power industry, traditional inspection is insufficient to meet the requirements of substation operation and maintenance and safety control. The unattended automatic operation and management mode can achieve remote monitoring functions, but a large amount of equipment data obtained during the inspection process still needs to be transmitted back to the monitoring main station for further analysis and judgment by the operation and inspection personnel. YOLOv5 is a universal object detection model, but there is still great room for improvement in multi-object detection problems such as unbalanced sample quality, mutual object occlusion, and poor distinguishability between the object and background. Therefore, this paper has made targeted improvements to the YOLOv5n loss function, backbone network and neck network, designed a new loss function Sigma-Complete Intersection over Union (CIoU) and C3-DCNv2 modules and integrated the Context Augmentation Module (CAM) into the model neck network. YOLO-Substation-tiny (YOLO-SS-tiny) high-performance lightweight model is proposed in this paper and constructs a defect detection dataset for substation equipment for multiple comparative experiments. The results show that the YOLO-SS-tiny reduces the total loss by 24.34 %, achieves mean average precision (mAP) of 69.1 %, which is 6.5 % higher than that of the original model, and reaches a detection speed of 274.2 FPS, based on increasing the parameter number by one-third of the original model. Compared with multiple advanced detection models, YOLO-SS-tiny has the best comprehensive performance and can better meet the requirements of the intelligent inspection of the distribution network in the new stage.

1. Introduction

Against the backdrop of rapid development in today's power system, active distribution networks are gradually becoming a key force in the transformation of the power industry due to their high reliability, optimized power quality, and efficient utilization of controllable resources. Especially with the large-scale integration of renewable energy into the grid, mainly in the form of distributed power generation, it not only promotes the flexibility and diversity of energy structure but also significantly reduces costs. However, this trend has also brought unprecedented challenges to the observability and controllability of the power system and has put forward more stringent requirements for the monitoring accuracy and real-time performance of power equipment [1,2]. Substations, as the core nodes connecting the backbone power grid and the distribution network, are expanding in scale and experiencing a significant increase in the number of equipment [3–5]. This further highlights the limitations of traditional manual inspections, such as high safety risks, and rising detection error and omission rates, which make it difficult to meet the needs of efficient operation and safety control in substations [6–8]. Therefore, the intelligent transformation of the power industry is imperative, and there is an urgent need for an efficient, accurate, and deep intelligent inspection solution that can penetrate the edge of the power system. The intelligent patrol technology for power system edge computing can reduce data transmission delay and improve system response speed through data processing and analysis at the edge, which is invaluable for ensuring the stable operation of the power grid and improving operation and maintenance efficiency.

* Corresponding author.
E-mail addresses: qianqian82@126.com (Q. Wang), ruiliu991112@163.com (R. Liu), qinsc31@126.com (S. Qin), pujiawei.1997@qq.com (J. Pu), shirong013@163.com (R. Shi), 1425142810@qq.com (Y. Wang).

https://doi.org/10.1016/j.ijepes.2025.111088
Received 16 April 2025; Received in revised form 12 August 2025; Accepted 4 September 2025
Available online 30 September 2025

In the field of object detection algorithms, most research has primarily been applied to datasets with a large number of high-quality training samples, such as the COCO dataset and the VOC dataset. By training on these datasets, it is relatively easy to obtain neural network models with high detection performance [9–11].The improvement ideas for the YOLO series models mainly include enhancing the loss function, incorporating various network modules, refining or replacing the backbone network, refining or replacing the neck network, refining the YOLOHead detection head, and utilizing k-means++ to recluster the prior boxes. Wang et al. proposed an improved vehicle visual object detection model, VV-YOLO, based on YOLOv4 [12]. The author used an improved K-means++algorithm to re-cluster the prior boxes and added a coordinate attention mechanism, CA-PAN network to the neck network, achieving multidimensional modeling of image feature channel relationships and improving the extraction efficiency of complex image features. Wu et al. proposed an efficient and lightweight road damage detection algorithm, YOLO-LWnet, for mobile terminal devices [13]. The author designed a new lightweight module, LWC, and used it as the basic building unit to replace the backbone and feature fusion network in YOLOv5. Its model outperforms state-of-the-art real-time detectors in balancing detection accuracy, model size, and computational complexity. Song et al. proposed an improved tea disease detection model, TSBA-YOLO [14]. The author added the Transformer self-attention mechanism enhanced model to the backbone network of YOLOv5, replaced the neck network with a BiFPN feature fusion network, added a random attention SA attention mechanism, and optimized the loss function with SIoU. The proposed model has a higher detection accuracy than the mainstream object detection model and achieves real-time detection speed. Zhora Gevorgyan et al. proposed an improved SIoU loss function based on the IoU loss function, optimizing the training and inference processes of object detection algorithms [15]. PP-YOLO, an improved YOLOv3 detection model proposed by Long et al., achieves a good balance between effectiveness and efficiency [16]. In studies adding attention blocks to the YOLO series of networks, various improved modules have emerged successively, such as Efficient Channel Attention (ECA), Efficient Squeeze-and-Excitation block (ESE), dual-attention networks based on Squeeze-and-Excitation (SE) attention blocks (A2Attention), Global Attention Mechanism (GAM), and Polarized Self-Attention (PSA) based on Convolutional Block Attention Module (CBAM) attention blocks (CBAM, attention block) [17–23]. However, existing research has mostly focused on improving detection accuracy and efficiency, neglecting the issue of the imbalanced contribution of training samples to the loss, which results in prediction bounding boxes being unable to effectively perform regression tasks, affecting the convergence speed of the loss function. Research on object detection technologies in the power industry has achieved certain results in detecting transmission lines, insulators, and other equipment [24–26]. However, it primarily focuses on detecting a single type of defect and lacks a public dataset containing multiple identification types, limiting the conduct of comparative studies. Additionally, issues such as poor distinguishability between equipment and backgrounds in actual substations, as well as mutual occlusion between equipment, have not been fully considered. Overall, existing achievements have optimized the training and inference process of object detection algorithms, proposed various improvement schemes, and achieved fruitful results. However, there are still some common problems in the current research on the application of object detection technology in the field of power: firstly, each improved model has a single type of defect detection target for substation equipment, and lacks a publicly available dataset that includes a large number of recognition types, making it difficult to conduct comparative studies. Secondly, existing research has not fully considered the impact of practical problems such as poor distinguishability between equipment and background in actual substations, and mutual occlusion between equipment. The detection performance of the model is severely affected by the complex environment of substations. Thirdly, most of the existing research only focuses on improving detection speed or accuracy, without considering the balance between speed and accuracy, which makes it difficult to meet the actual inspection needs of substations.

To address the aforementioned issues, algorithm design must balance two core elements: performance and parameter quantity. On the one hand, the algorithm is required to possess excellent detection capabilities to ensure efficient and accurate monitoring in complex and varied power system environments. On the other hand, given the limited resources of edge devices, the parameter quantity of the algorithm model must be strictly controlled to reduce the computational burden, improve deployment efficiency, and ensure seamless integration and stable operation on the edge side. In the context of substation equipment defect detection, the complex environment of the equipment, poor distinguishability between detection objects and backgrounds, severe occlusion between objects, and the lack of computers with powerful computing capabilities in most substations make model training difficult and limit detection performance. Therefore, optimizing the model network and maximizing its performance while ensuring the conversion cost of substation computing power becomes an urgent problem to solve. Furthermore, data indicates that the parameter quantities of YOLOv5n and YOLOv8n are 1.7 and 3.1, respectively, and their computational demands are 4.3 and 8.2, respectively. YOLOv5n has lower computational demands. Thus, although YOLOv8 offers superior overall performance, YOLOv5′s performance is sufficient for some application scenarios. In resource-limited substation environments, YOLOv5′s lightweight model remains highly competitive.

Therefore, this paper proposes a new loss function, Sigma-CIoU, which aims to suppress the contribution of low-quality samples to the loss and enhance the contribution of high-quality samples. Meanwhile, DCNv2 is used to replace the standard convolution in the original model backbone C3 module, resulting in the C3-DCNv2 module, and the CAM module is integrated into the neck network of the original YOLOv5n model. By collecting and constructing a substation equipment defect dataset, ablation experiments, real-time performance tests, Gram-CAM heatmap visualization comparison experiments, and horizontal comparison experiments are conducted to verify the impact of the aforementioned designs on the performance of the original YOLOv5n model.

2. Materials and methods

This paper proposes the YOLO-SS-tiny model with improvements to the model loss function, backbone network and neck network based on YOLOv5n.

2.1. Basic model

YOLOv5-7.0 version is the latest version of YOLOv5 released by the Ultralytics team in 2022, which proposed a new basic model YOLOv5n. Table 1 is the performance comparison of different YOLOv5 models in each version, where the number of parameters (Para.) of YOLOv5n (1.78 M) is about one-fourth of the lightest model YOLOv5s (7.08 M) in the previous version. The testing results on the COCO dataset show that the YOLOv5n model mAP (45.7 %) is 11.1 % lower than that of YOLOv5s (56.8 %).

Table 1
Performance comparison of different YOLOv5 models.

Model	mAP/%	Inference speed/s	Para./M
n	45.7	45	1.9
s	56.8	98	7.2
m	64.1	224	21.2
l	67.3	430	46.5
x	68.9	766	86.7

2.2. Improved loss function

To address the problem of unbalanced loss contribution in the substation equipment defect dataset, this paper proposes a new loss function Sigma-CIoU to replace the complete cross-ratio loss function used in the YOLOv5n model location loss. The total loss function of the basic model YOLOv5n consists of the location loss function (L_{box}), the objectness loss function (L_{obj}), and the classes loss function (L_{cls}). Eq. (1) is the location loss function. Eq. (2) 0is the default computation equation of the CIoU loss function used in YOLOv5n, and the model also comes with Generalized IoU (GIoU) and Distance-IoU (DIoU) losses for switching during training. The calculation equation for the aspect ratio of the detection box scale is shown in Eq. (3).

$$L_{box} = \lambda_{IoU} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} L_{CIoU} \tag{1}$$

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

$$V = \frac{4}{\pi^2} \cdot \left(arctan \frac{w_{gt}}{h_{gt}} - arctan \frac{w}{h} \right)^2 \tag{3}$$

where B is the predicted box; b_{gt} is the ground truth bounding box; S is the parameter for the number of grid divisions in the feature map; L_{DIoU} is Distance-IoU loss function; ρ denotes the Euclidean distance between the center point of the predicted box and the center point, indicating the diagonal distance that can simultaneously minimize the closed region C; α is the weight coefficient; V is the aspect ratio penalty term; w_{gt} is the width of the ground truth bounding box; h_{gt} is the height of the ground truth bounding box, w is the width of the predicted box, and h is the height of the predicted box.

In the original YOLOv5 model, L_{obj} and L_{cls} use the binary cross entropy for calculation by default. There are three implementations for calculating L_{box}: GIoU, DIoU, and CIoU. CIoU is used by default during the model training. In the predicted box regression calculation, low-quality samples often account for a large part of the loss function. The contribution of low-quality samples to the loss function is much larger than that of high-quality samples, which is very detrimental to the model training. However, whether it is the CIoU, GIoU and DIoU that come with the basic network, or the latest SIoU and EIoU, they are based on IoU, optimize the loss function of the predicted box and the ground truth bounding box to degenerate back to IoU in some specific location relationship by adding penalty terms related to the location and size information of the prediction box and the ground truth bounding box, while the problem of the unbalanced contribution of training samples to loss is ignored, which can't effectively describe the goal of predicted box regression, and seriously affects the convergence rate of the loss function and the accuracy of regression results.

In order to reduce the adverse effects caused by the above problems, the CIoU loss function is improved in this paper by using IoU to participate in calculating the loss, and the CIoU loss function is nonlinearly weighted to obtain the Focal-CIoU loss function, which aims to suppress the contribution of low-quality samples to the loss, and its expression is shown in Eq. (4).

$$L_{Focal-CIoU} = IoU^{\gamma} L_{CIoU} \tag{4}$$

On the basis of suppressing the contribution of low-quality samples to the loss, this paper further optimizes the Focal CIoU loss function to strengthen the contribution of high-quality samples to the loss. The Sigmoid function is used to treat the intersection and union ratio nonlinearly instead of the square of IoU. The value range of the complete intersection and union ratio is [0,1], and the complete intersection and union ratio is limited to the range of the independent variable of the Sigmoid function by α and β, that is [-10,10]. Then the loss weights are optimized to design the Sigam-CIoU loss function, whose expression is shown in Eq. (5).

$$L_{Sigma-CIoU} = \sigma(\alpha IoU - \beta) L_{CIoU} \tag{5}$$

where σ is the Sigmoid function, whose expression is shown in Eq. (6).

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

Fig. 1 is the comparison of the curves of the complete intersection and union ratio, the Focal-CIoU, and the Sigam-CIoU loss function. It can be intuitively observed that both the Focal-CIoU and the Sigam-CIoU loss function outperform the complete intersection and union ratio adopted by the basic network in terms of balancing the contribution of low-quality and high-quality samples to the loss. The difference is that in the range [0,0.2] corresponding to the low-quality samples, the loss value of the Sigma-CIoU loss function is close to 0, which suppresses the low-quality samples more significantly than Focal-CIoU; in the range [0.6,1] corresponding to the high-quality samples, the loss value of the Sigam-CIoU loss function is higher than that of the Focal-CIoU loss function, which strengthens the contribution of high-quality samples to the loss.

2.3. Improved backbone network

In order to reduce the useless contextual information in the backbone network feature extraction due to the poor distinguishability between the substation detection object and the background, and to extract the shallow features of the images more effectively, this paper proposes an improved scheme: the C3-DCN module is designed by replacing the traditional convolution of the C3 module with the deformable convolution (DCNv2) and replacing the C3 module in the original model backbone network (layers 6 and 8) in the original model backbone.

The YOLOv5 backbone uses a standard convolution module consisting of three parts: traditional convolution, normalized layer, and Silu activation function. Eq. (7) is the calculation equation for traditional convolution. The convolution layer is used to extract feature information in the images. The traditional convolution usually uses a rectangular convolution kernel with fixed size and weight to perform feature learning and down-sampling operations on the feature map. It is found that different locations in the same feature layer correspond to objects of different scales and shapes, and the traditional convolution in YOLOv5n uses regular grid point sampling, which leads to the difficulty of adapting the model to geometric deformation, thus having certain limitations in object detection.

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{7}$$

where x is the feature map; w is the convolution kernel parameter; p_0 is the coordinate of the convolution kernel center in respect to the upper left corner of the input feature map; and p_n is the offset of the R element in respect to the convolution kernel center; y is the input result of the

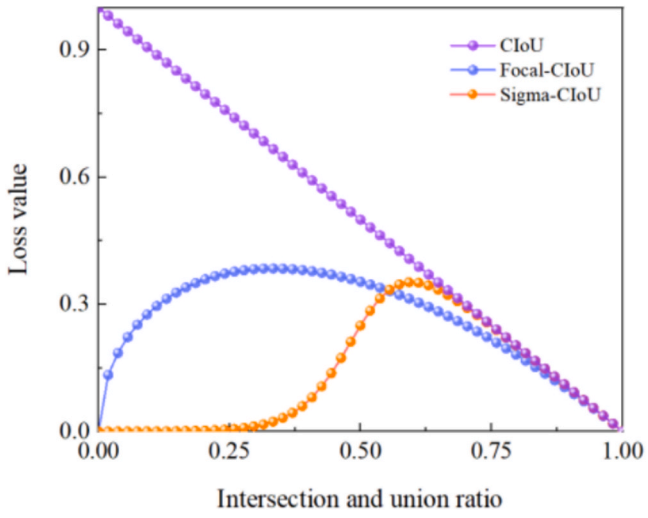


Fig. 1. Comparison of loss functions.

convolution operation.

To weaken this limitation, the deformable convolution v1 extends the convolution kernel by introducing offset variables, so that the convolution kernel can be sampled randomly near the current location and the model feature extraction process can be more focused on the effective information region, effectively overcoming the shortcomings of insufficient sampling of the fixed rectangular structure and improving the network ability to simulate object deformation. Eq. (8) is the calculation equation of the deformable convolution DCNv1.

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{8}$$

where Δp_n is the location offset vector.

However, DCNv1 may introduce useless contexts (regions) to interfere with the feature extraction of the detection model, which will obviously degrade the detection performance of the model. In substation environment where the distinguishability between the detection object and the background is poor, the defects of DCNv1 will be amplified, resulting in severe degradation of the model detection performance. To further enhance the network ability to adapt to deformable objects, DCNv2 introduces a modulation mechanism based on DCNv1, which, in addition to allowing the model to learn the offset of sampling points, also adds a new weight for learning each sampling point, aiming to mitigate the interference of irrelevant factors. Eq. (9) is the calculation equation of the deformable convolution DCNv2. DCNv2 adds weight coefficients to DCNv1, assigning different weights to the offsets of sampling points in the input feature map obtained from the convolution calculation, removing irrelevant contextual information and achieving more accurate feature extraction. As a result, the deformable convolution network DCNv2 can not only adjust the learned offsets but also optimize the input features, meanwhile, making certain regions in the images not affect the model output.

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_k \tag{9}$$

where Δm_k is the offset weight and takes the value range of [0,1].

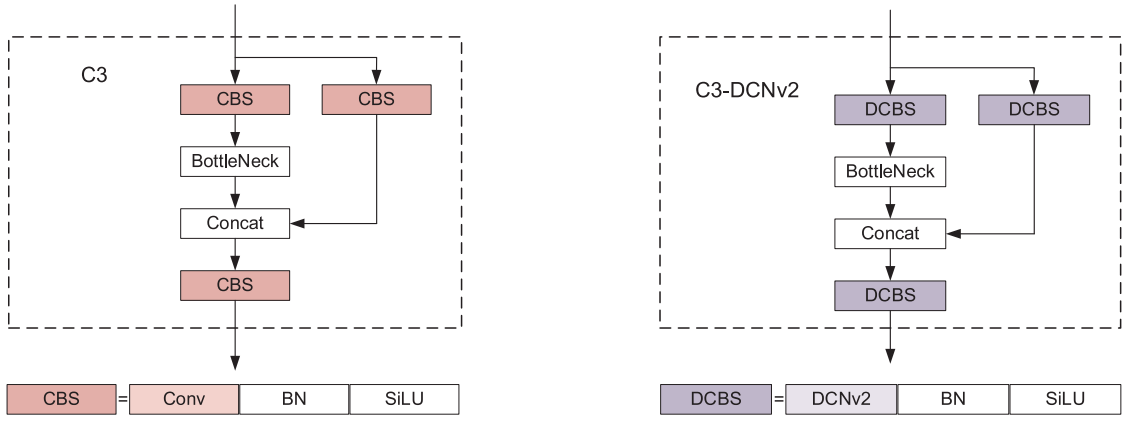
As shown in Fig. 2, this paper introduces the deformable convolution DCNv2 into the backbone network of YOLOv5n, using DCNv2 to replace the regular convolution of the standard convolution CBS in the C3 module to form the C3-DCNv2 module. Fig. 3 shows the modified best-performing backbone network structure.

2.4. Improved neck network

The occlusion problem in object detection refers to the object being occluded by other objects or obstructions, resulting in the model being unable to detect and identify the object correctly, which is an important challenge in practical applications, especially in substations with complex environment. In the process of substation equipment inspection, the phenomenon of mutual occlusion of detection objects is common due to the complex environment, the equipments often being occluded by each other, and the limitations of objective factors such as the shooting angle and the location of the detection objects. For the problem that the detection objects are considered as background and cannot be detected due to the occlusion phenomenon, this paper proposes an improved solution: incorporating CAM into the neck network.

As shown in Fig. 4(a), the SSP module first halves the input channel through a standard convolution module, and then performs max pooling with convolution kernel sizes of 5, 9, and 13, respectively, to adaptively fill different convolution kernel sizes. Finally, the results of the three max pooling operations are concatenated with the data that has not been pooled, resulting in a feature map with twice the number of channels after merging. SPPF has improved SPP to enhance the efficiency of the model. This module uses three cascaded max pooling layers to extract deeper feature maps, and then fuses the feature maps through concatenation operations. As shown in Fig. 4(b), the Spatial Pyramid Pooling-Fast (SPPF) first concatenates the feature maps that have undergone a standard convolution module, one-time pooling, two pooling, and three pooling, and then uses the standard convolution module to extract features. In the fast pyramid pooling module, although the feature maps have been pooled multiple times, their size and number of channels have not changed, so the pooled output feature maps can be fused in the channel dimension. The SPPF module in the YOLOv5 backbone network serializes the three maximum pooling operations with a convolution kernel size of five, which speeds up the operation by reducing the repetitive operations. However, the equipment environment in the substations is quite complex, and problems such as mutual occlusion between detection objects are serious. In this environment, the model should detect all the devices in the station as much as possible, and the accuracy and recall metrics are very important. Each pooling operation in the SPPF module is performed on the basis of the previous operation, and the repeated use of the pooling layer loses more object detail information, resulting in some objects being considered as the background and unable to be detected, which is the main reason why YOLOv5n has difficulty in meeting the accuracy requirements in this environment.

Compared with the SPPF module, each convolution output in CAM contains a larger range of information that can increase the perceptual field of the YOLOv5n network. In the initial improvement strategy, this paper uses CAM to replace the fast pyramid pooling module in the original YOLOv5n backbone network, but the difference in the number of model parameters before and after the replacement is too large to make this improvement scheme practical.



(a) C3 (b) C3-DCNv2

Fig. 2. Structures of C3 and C3-DCNv2 module.

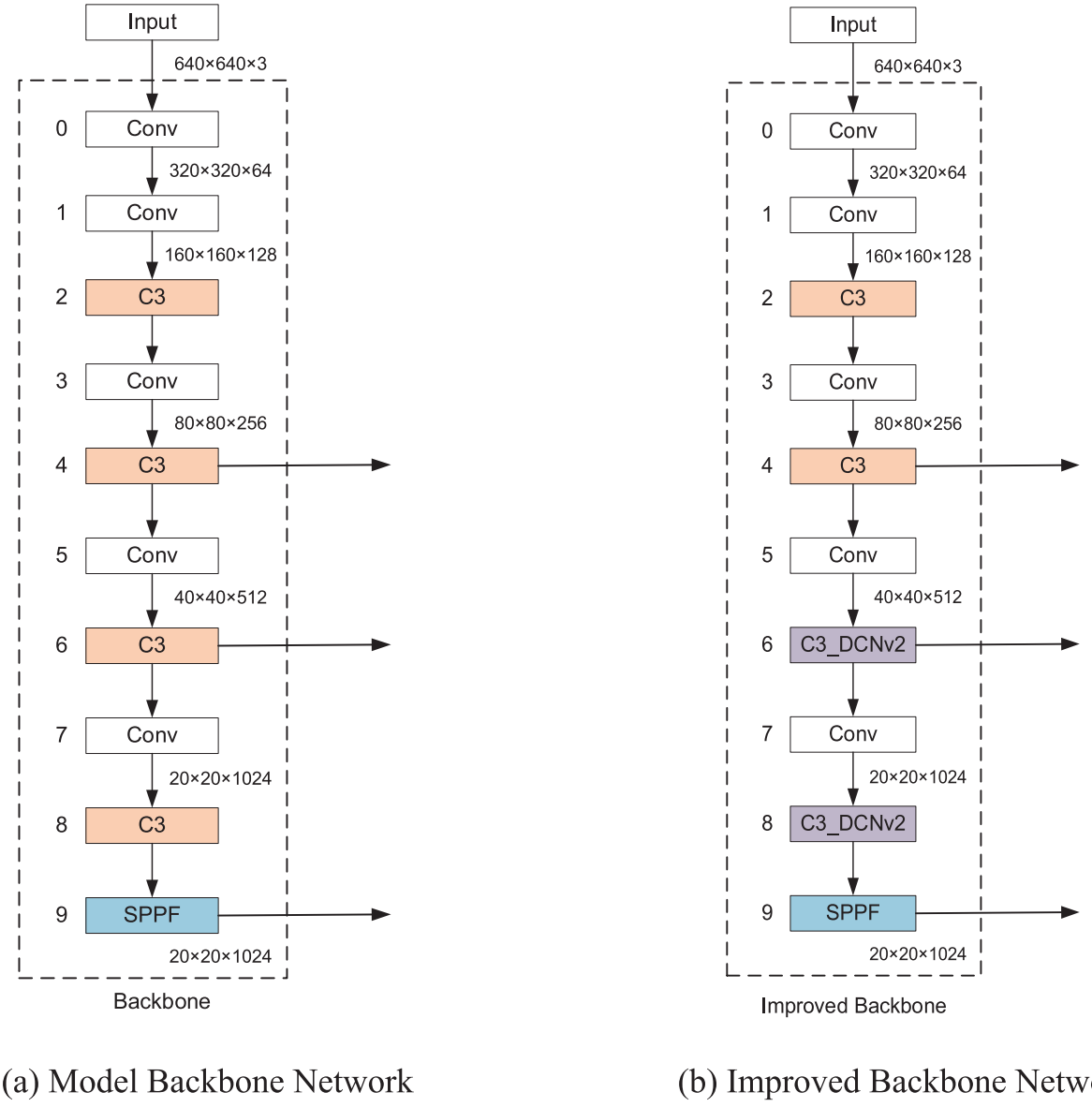


Fig. 3. Comparison diagram of the model structure before and after improvement.

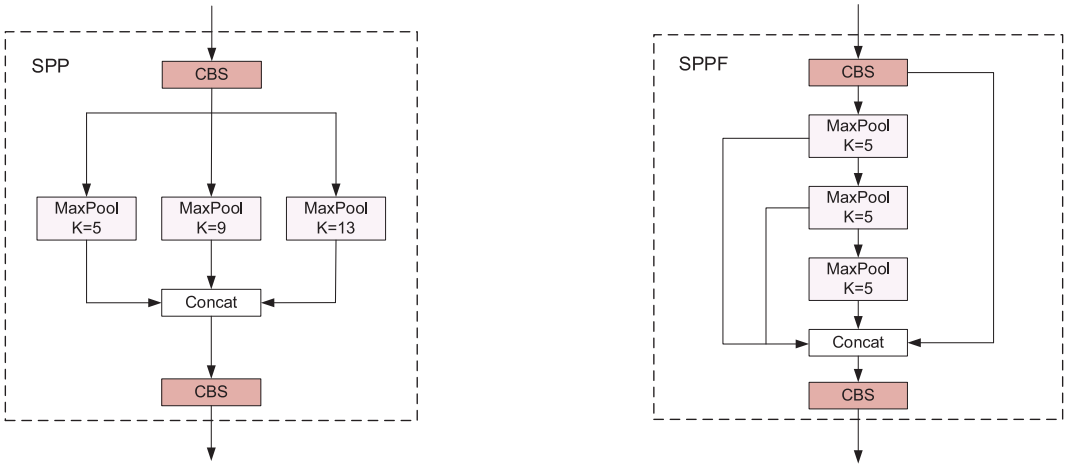
In the study of the neck network, this paper finds that the C3 module (layer 23), which is in the last layer of the neck network, contains more location information but lacks semantic information. This is due to the fact that the neck network has a tandem information transmission structure, and the number of channels is reduced by down sampling before feature fusion, which makes the C3 module unable to obtain sufficient semantic information. As shown in Fig. 5, CAM is incorporated between layers 10 and 23 of the neck network, and the features in the fast pyramidal pooling module are extracted by using the null convolution with different null convolution rates and injected into the C3 module (layer 23) of the neck network to supplement the context information, so as to optimize the YOLOv5n neck network. This paper conducts comparative experiments on CAM with three different fusion methods to further study their effects on the model performance.

Three structures of CAM are shown in Fig. 6. All three structures of CAM perform a null convolution with a null convolution rate of 1, 3, and 5 for the input feature layer, respectively. The number of convolution kernels is set to one-fourth of the number of channels in the feature map in order to avoid excessively large parameters during training. The number of channels is expanded back to the initial number by (1 × 1) convolution to obtain three feature layers with the same size but different perceptual fields. Finally, the feature extraction is completed by different fusion methods. The difference lies in the way of feature fusion. The CAM-W module using weighted feature fusion is shown in Fig. 6(a), the CAM-C module using cascade operation fusion is shown in Fig. 6(b), and the CAM-A module using adaptive feature fusion is shown in Fig. 6(c), which is similar to the implementation principle of the SE attention mechanism. Specifically, the CAM-A module obtains the adaptive weights of the input channels by convolution operations and Softmax activation function, and fuses the contextual information extracted by the three convolution operations into the output feature map. The structure diagram of the improved model YOLO-SS-tiny network proposed in this paper is shown in Fig. 7.

3. Results and Discussion

3.1. Dataset

The identification of substation equipment defects requires the collection of relevant images, and this paper takes a 220 kV substation equipment as the research object, through the collection of field inspection data and historical data, and the collected 6301 substation



(a) SPP (b) SPPF

Fig. 4. Structures of SPP and SPPF module.

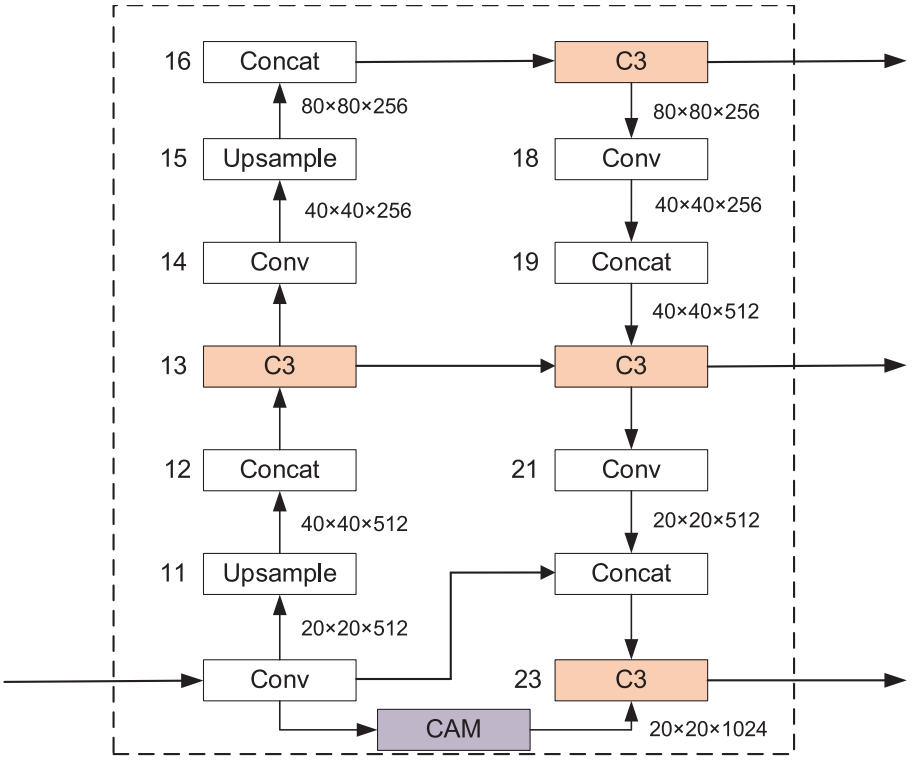


Fig. 5. Structure of the improved model.

images mostly contain a single defective object, a small number of them have multiple detection objects in the same environment. Defect detection categories are divided into 11 types, as shown in Table 2 indicating defects of each category corresponding to the sticky note ID and their number.

This paper uses the LabelImg tool to make labels in XML format and converts the label files in XML format into TXT format so that it can be recognized by the YOLO algorithm, finally divides the data set into a training set, test set, and validation set in the ratio of 7:2:1. Fig. 8 demonstrates part of the substation image data.

The category distribution of the training set used in this validation experiment is shown in Fig. 9(a), and the detection categories are more evenly distributed. Fig. 9(b) shows adaptive anchor boxes jointly generated by 11 defect detection categories. Most of the anchor boxes are of medium scale, and there are more small-size anchor boxes than large-size anchor boxes, and the difference in aspect ratio is not significant. Fig. 9(c) and (d) show the location offsets of the predicted boxes output by the YOLOv5 model object detection head, where x and y are the coordinates of the center point of the predicted boxes. Most of the predicted boxes are distributed at the center of the images, and the detection objects are mostly of medium and small sizes.

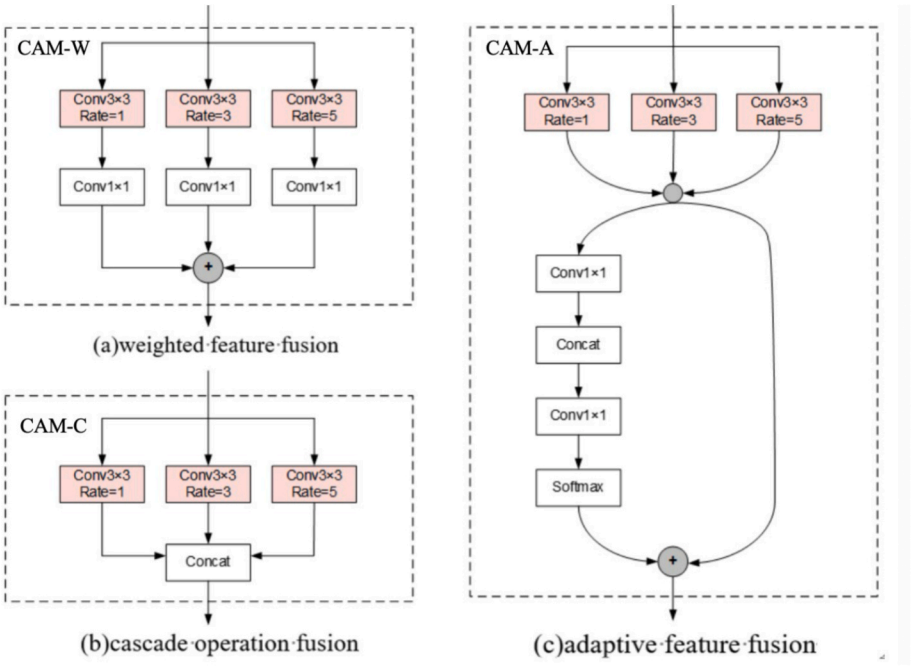


Fig. 6. Three structures of CAM module.

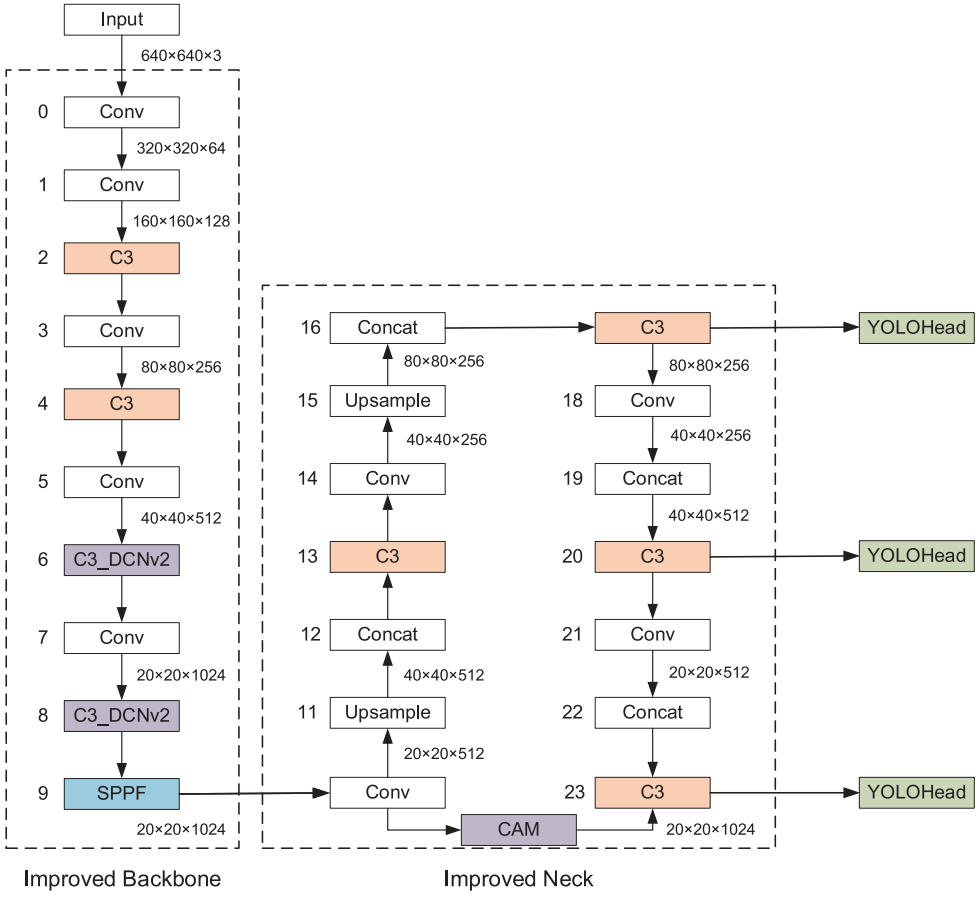


Fig. 7. The structure of YOLO-SS-tiny.

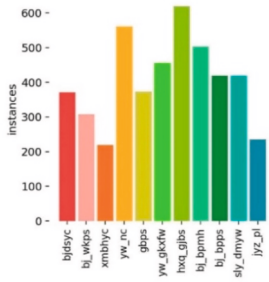
Table 2
Substation image dataset.

Ref	Type	Label	Number
0	Abnormal meter reading	bjdsyc	789
1	Shell damage	bj_wkps	523
2	Abnormal closure of the box door	xmbhyc	383
3	Nest	yw_nc	883
4	Damaged cover plate	gbps	654
5	Hanging suspended solids	yw_gkxfw	729
6	Respirator silicone discoloration	hxq_gjbs	1174
7	Blurred dial	bj_bpmh	869
8	Damaged dial	bj_bpps	723
9	Oil stains on the ground	sly_dmyw	833
10	Insulator rupture	jyz_pl	410

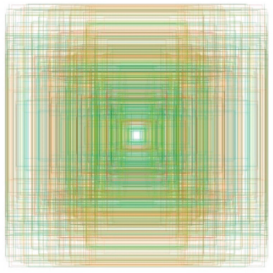
Table 3
Hyperparameter settings.

Hyperparameter	Parameter settings
Image size	640 × 640
Epochs	200
Batch size	16
Optimizer	SGD
Initial Learning rate (lr0)	0.01
Cycle Learning rate (lrf)	0.01
Momentum	0.937
Weight decay	0.0005

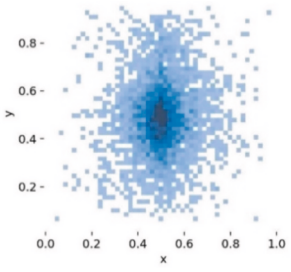
Fig. 8. Part of the substation image data.



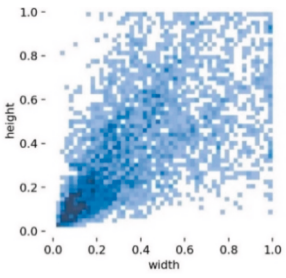
(**a**) Category distribution of training set



(**b**) Anchor boxes scale



(**c**) Center point distribution of ground truth bounding boxes



(**d**) Length and width dimensions of ground truth bounding boxes

Fig. 9. Image dataset of the substation.

Table 4
Experimental environment.

Environmental configuration	Name	Information
Hardware configuration	CPU	Intel(R) Core(TM) i9-12900H
	GPU	NVDIA GeForce RTX 3060
	Memory	16.0 GB
Software environment	Operating system	Windows 11
	Development environment	PyCharm
	Python	3.8.13
	Pytorch	13.1
	CUDA	11.3
	cuDNN	8.3.2

3.2. Experimental environment

The hyperparameter settings and the experimental environment of this paper are shown in Table 3 and Table 4.

3.3. Metrics of evaluation

Precision (P), Recall (R), and Mean Average Precision (mAP) are used as evaluation metrics for the substation equipment defect detection model. The equations for calculating P, R, and mAP are shown in Eqs. (10)–(12). The above evaluation metrics are closely related to the intersection and union ratio detection threshold, which is 0.5 in this experiment, i.e., the prediction is correct when the overlap area between the detection box and the ground truth bounding box exceeds 50 % of the total area of the two boxes, and it is wrong when it is lower than 50 %.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$mAP = \frac{\sum_{i=1}^{N} \int_0^1 PRdR}{n} \tag{12}$$

where *TP* represents the positive class as a positive class; *FP* represents the negative class as a positive class; *FN* represents the positive class as a negative class; *n* is the category number.

In addition, this paper chooses the number of parameters to represent the required arithmetic power of the model, and Frame Per Second (FPS) to verify the real-time performance of the modified model.

3.4. Results and analysis

3.4.1. Experiment of improved loss function

In order to verify the regression advantages of the improved Sigma-CIoU function on the model in this paper, the complete intersection and union ratio, Focal-CIoU, and the improved Sigma-CIoU function are chosen to calculate the L_{box} in YOLOv5 for comparison experiments, and the results are summarized as shown in Tables 5 and 6 and Fig. 10.

Table 5 shows the model performance metrics obtained in the last round of training results in 200 rounds. Combined with Fig. 10(a), it can be concluded that, compared with the default complete intersection and union ratio loss function used in the original YOLOv5 training, the L_{box} of Focal-CIoU is reduced by 22.65 %, and the total loss is reduced by only 13.86 %. The average accuracy of the model is reduced by 1.4 %. The actual training results of the improved Sigma-CIoU function in this paper outperform the above two loss functions. Compared with the default CIoU loss function used in the original YOLOv5, the L_{box} is reduced by 35.37 % and the total loss is reduced by 19.74 %. The significant advantage is achieved in the convergence speed.

Table 5
Comparison of experimental results 1 of loss function.

Loss function	Loss Boxes	Total
CIoU	0.038303	0.064629
Focal-CIoU	0.029626	0.055672
Sigma-CIoU	0.024755	0.051871

Table 6
Comparison of experimental results 2 of loss function.

Loss function	P/%	R/%	mAP/%	Para./M
CIoU	0.701	0.6	62.6	1.78
Focal-CIoU	0.662	0.609	61.2	1.78
Sigma-CIoU	0.634	0.642	62.8	1.86

Table 6 shows the model performance metrics with the best performance in the 200-round training results. Combined with Fig. 10(b), it can be concluded that the original YOLOv5, which uses the complete intersection and union ratio loss function to calculate the L_{box}, has an average accuracy of 62.6 %. After using Focal-CIoU to replace the complete intersection and union ratio, the average accuracy decreases; after using the improved loss function proposed in this paper, Sigma-CIoU, the average accuracy is improved by 0.2 % compared with the original YOLOv5. Combined with the total loss value curve of the model shown in Fig. 10(b), it can be intuitively seen that the loss value of Sigma-CIoU decreases fastest among the three loss functions and the final convergence value is lower. In summary, the improved Sigma-CIoU loss function in this paper makes the model converge significantly faster while ensuring the accuracy of the original algorithm.

3.4.2. Experiments of the improved backbone network

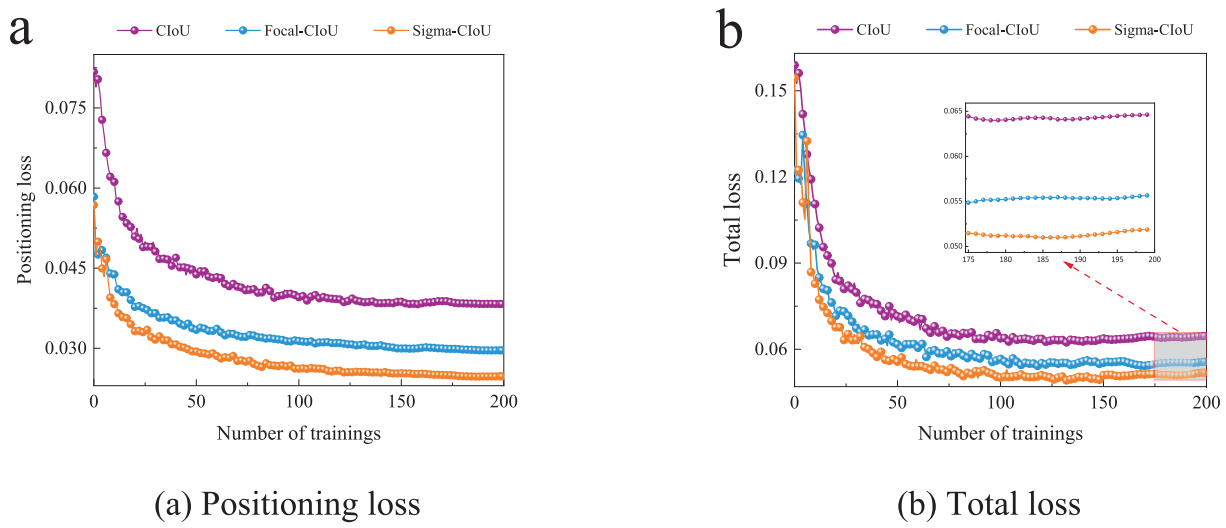
In the backbone improvement experiments, this paper uses C3-DCN modules to replace C3 modules in the original YOLOv5 algorithm backbone network, and conducts comparative experiments on replacing the C3 module in the last layer (layer 8), the last two layers (layers 6 and 8), and the last three layers (layer 4, 6 and 8), as shown in Table 7. The experimental dataset still uses real images from a certain substation as mentioned earlier.

Combined with Fig. 11, it can be concluded that, compared to YOLOv5n, the performance gains are the greatest when the C3-DCN module is used to replace the last two layers (layers 6 and 8) of the backbone network, with an average accuracy improvement of 3.4 % and 13.8 % increase in the number of model parameters, which is within the acceptable range.

3.4.3. Experiment of improved neck network

Initially, this paper uses CAM-W, CAM-A and CAM-C to replace SPPF in the original YOLOv5n backbone network. Table 8 shows the parameter test of the model before and after the replacement. It can be concluded that the arithmetic volume of the replacement model has increased by more than twice that of the original model, which indicates that such an improvement scheme has no practical value. Therefore, in the neck network improvement experiments, this paper adds CAM between layer 10 and layer 23 of the YOLOv5n network and conducts comparison experiments on the algorithms after adding three CAMs with different structures, and the experimental results are shown in Table 9.

Combined with Fig. 12, it can be concluded that CAM-A with the addition of the adaptive feature fusion structure has the largest performance improvement on the original network model, with an average accuracy increase of 5.3 % and the model parameter number increase of 27.8 % compared to the original YOLOv5n, which is within the acceptable range.



(a) Positioning loss (b) Total loss

Fig. 10. Results of loss function improvement experiments.

Table 7
Backbone improvement experimental results.

Replace the location layer of C3 module	P/%	R/%	mAP/%	Para./M
no	0.701	0.6	62.6	1.78
8	0.688	0.622	63.4	1.81
6、8	0.712	0.635	66.0	1.86
4、6、8	0.671	0.632	64.9	1.87

Table 9
Neck improvement experimental results.

Added module	P/%	R/%	mAP/%	Para./M
No	0.701	0.6	62.6	1.78
CAM-W	0.734	0.626	65.6	2.34
CAM-C	0.734	0.626	65.6	2.34
CAM-A	0.717	0.65	67.9	2.27

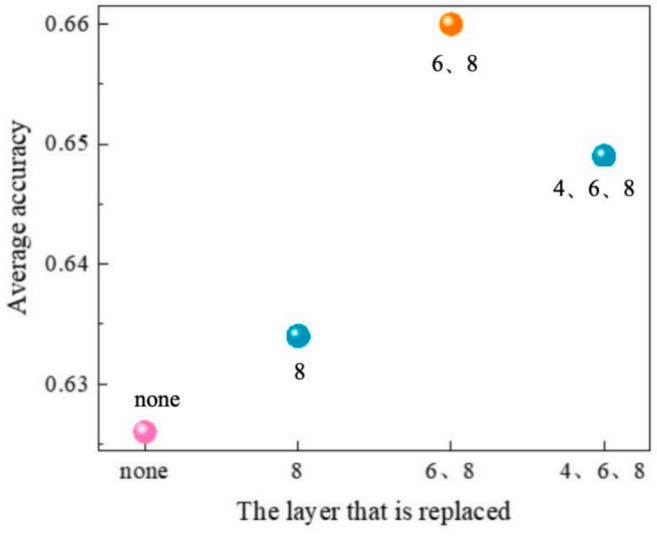


Fig. 11. The experiment results of C3 substitution.

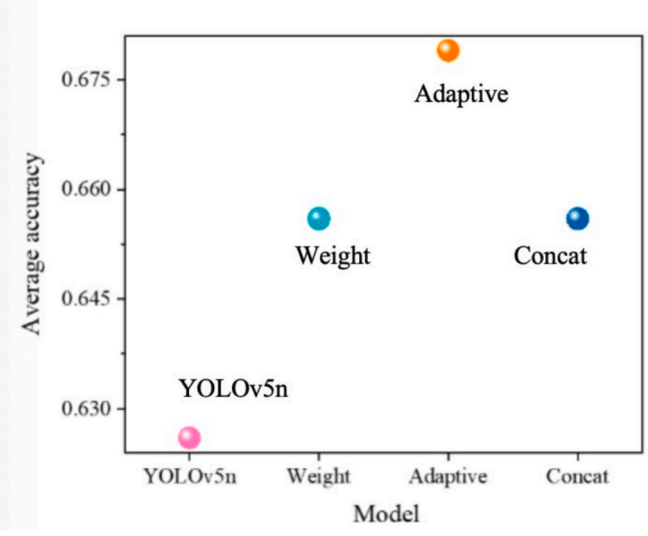


Fig. 12. Improvement experiments results of Neck network.

Table 8
Module parameter test.

Module	Para./M	Arithmetic number/G
SPPF	1.78	4.3
CAM-W	3.58	5.7
CAM-A	3.65	5.7
CAM-C	3.59	5.7

3.4.4. Ablation experiments

To verify the effectiveness of each improved module in the YOLO-SS-tiny model, multiple rounds of ablation experiments were conducted in

Table 10
The ablation experimental results.

Model	Total loss	P/%	Re/%	mAP/%	Para./M
YOLOv5n	0.06463	70.1	60.0	62.6	1.78
+Sigma-CIoU	0.051871	63.4	64.2	62.8	1.78
+Sigma-CIoU + C3-DCN	0.047483	71.4	65.7	67.4	1.86
+Sigma-CIoU + C3-DCN + CAM (YOLO-SS-tiny)	0.0489	77.2	66.5	69.1	2.35
YOLOv5s	0.0605	73.6	64.5	67.4	7.05

this paper, and the results are shown in Table 10. The experimental results show that after using the Sigma-CIoU loss function designed in this paper, the L_{box} of the model is reduced by 35.37 % and the total loss is reduced by 19.74 %, and a significant advantage in the convergence speed is achieved. This is because Sigma-CIoU redefines the loss of the predicted box to better enhance the contribution of high-quality samples to the loss while suppressing the contribution of low-quality samples to the loss, reflecting the performance of cutting irrelevant data and making the convergence performance of the model effectively improved. On this basis, the improved module C3_DCNv2 of this paper is used to replace layers 6 and 8 of the YOLOv5n model backbone network, and the improved model achieves an average accuracy of 67.4 % on the basis of increasing the number of parameters of the original model by 13.8 %, which is 3.4 % better than the basic model because C3_DCNv2, compared to the C3 module of the YOLOv5n model backbone network, can better adapt to the detection object in the complex environment of the substation and achieve a more accurate and effective extraction of the shallow features of the image. Finally, based on the above improvement strategies, the CAM module is incorporated between the 10th and 23rd layers of the basic model of the neck network. Finally, the proposed model in this paper achieves an average accuracy of 69.1 %, which is 5.3 % better than that of the basic model, on the basis of increasing the number of parameters of the original model by 27.8 %. The results of the ablation experiments prove the effectiveness of each improved module in the YOLO-SS-tiny model.

Fig. 13 shows the comparison of the detection performance of the three models YOLOv5n, YOLOv5s, and the proposed YOLO-SS-tiny. In terms of the model loss value, the total loss (0.0489) of the improved model YOLO-SS-tiny proposed in this paper is 24.34 % lower than that of the original YOLOv5n (0.06463) model, and 19.17 % lower than that of the YOLOv5s model (0.0605). In terms of model accuracy, the average accuracy of the improved model YOLO-SS-tiny proposed in this paper (69.1 %) is 6.5 % higher than that of the original YOLOv5n (62.6 %) model and 1.6 % higher than that of the YOLOv5s model (67.4 %). In terms of model size, the improved model YOLO-SS-tiny has 32 % more parameters (2.35 M) than the original YOLOv5n (1.78 M) model, which is one-third of the YOLOv5s model (7.05 M). The comprehensive performance of the improved model YOLO-SS-tiny is better than that of YOLOv5n and YOLOv5s.

3.4.5. Thermodynamic diagram visualization

To show the detection effect more intuitively, the detection performance of three models, YOLOv5n, YOLOv5s, and YOLO-SS-tiny proposed in this paper, is tested using Grad-CAM, and the visualized thermodynamic diagrams of the three models in several environment are shown in Fig. 14. The red areas of all thermodynamic diagrams represent the areas where the model attention is focused, and the darker the red, the higher the level of attention.

The original dataset shown in Fig. 14(a) displays the actual scene images of the substation. (b) The heatmap of the YOLOv5n model shown shows its attention distribution towards the target. In object detection with complex backgrounds, attention is relatively scattered. (c) Compared to YOLOv5n, the results of the YOLOv5s model show that it can concentrate on the target in certain scenarios, but still has certain limitations when dealing with complex backgrounds and small targets. The results of the YOLO-SS tiny model shown in Figure (d) indicate that all the red regions in the YOLO-SS tiny feature maps are concentrated on the defect target of the substation, indicating that the extraction effect of this model is better than that of the YOLOv5n and YOLOv5s models, providing more attention to the detection target and better detecting objects. Specifically, in the first and second row scenarios, YOLO-SS tiny has higher confidence in detecting targets with complex backgrounds and outperforms the other two models in terms of detection performance; In the third scenario, the YOLO-SS tiny model performs better in detecting small targets; In the fourth scene, for bird's nest targets with overlapping and occluded areas, the red areas of the YOLO-SS tiny model feature map are concentrated on the bird's nest targets, demonstrating better feature extraction capabilities.

3.4.6. Cross-sectional comparison experiments

To further validate the detection performance of the YOLO-SS-tiny model, we compare the proposed YOLO-SS-tiny model with other advanced object detection methods, including FasterR-CNN in the two-stage object detection model RCNN, SSD-MobileNetv2 in the single-stage object detection model SSD, and YOLO series in YOLOv7-tiny and YOLOv8n, the difference in the number of parameters of the above models is not significant.

The experimental results are shown in Table 11. It can be concluded that the proposed lightweight model YOLO-SS-tiny achieves the highest detection accuracy with the least parameter number, and the detection speed is 274.2 FPS, which is second only to YOLOv8n and 5.9 % faster than that of YOLOv7-tiny, and the comprehensive detection performance of the proposed method is better than that of other similar methods.

4. Conclusion

With the rapid promotion of the "double carbon" target and the upgrading of active distribution networks, the scale of substations is expanding and the number of substation equipment is increasing,

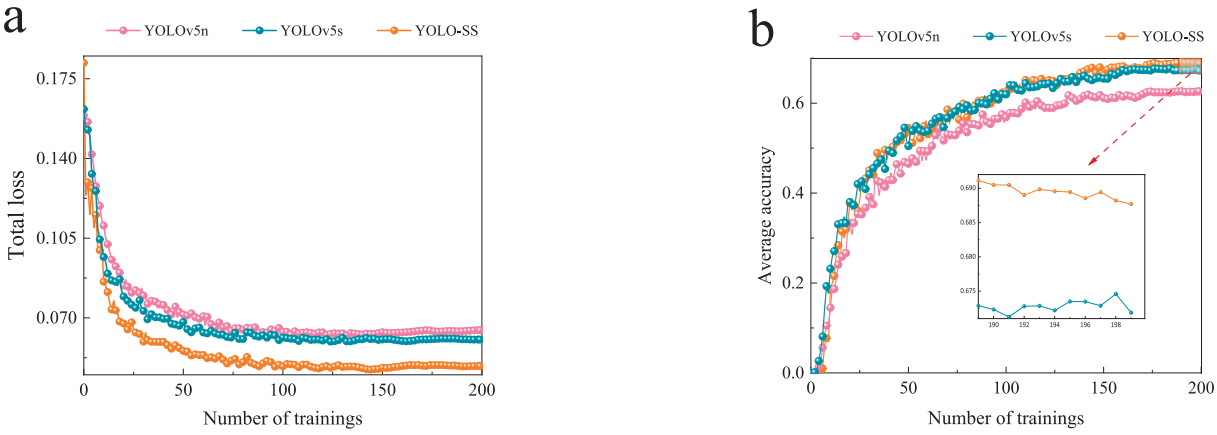


Fig. 13. Performance comparison.

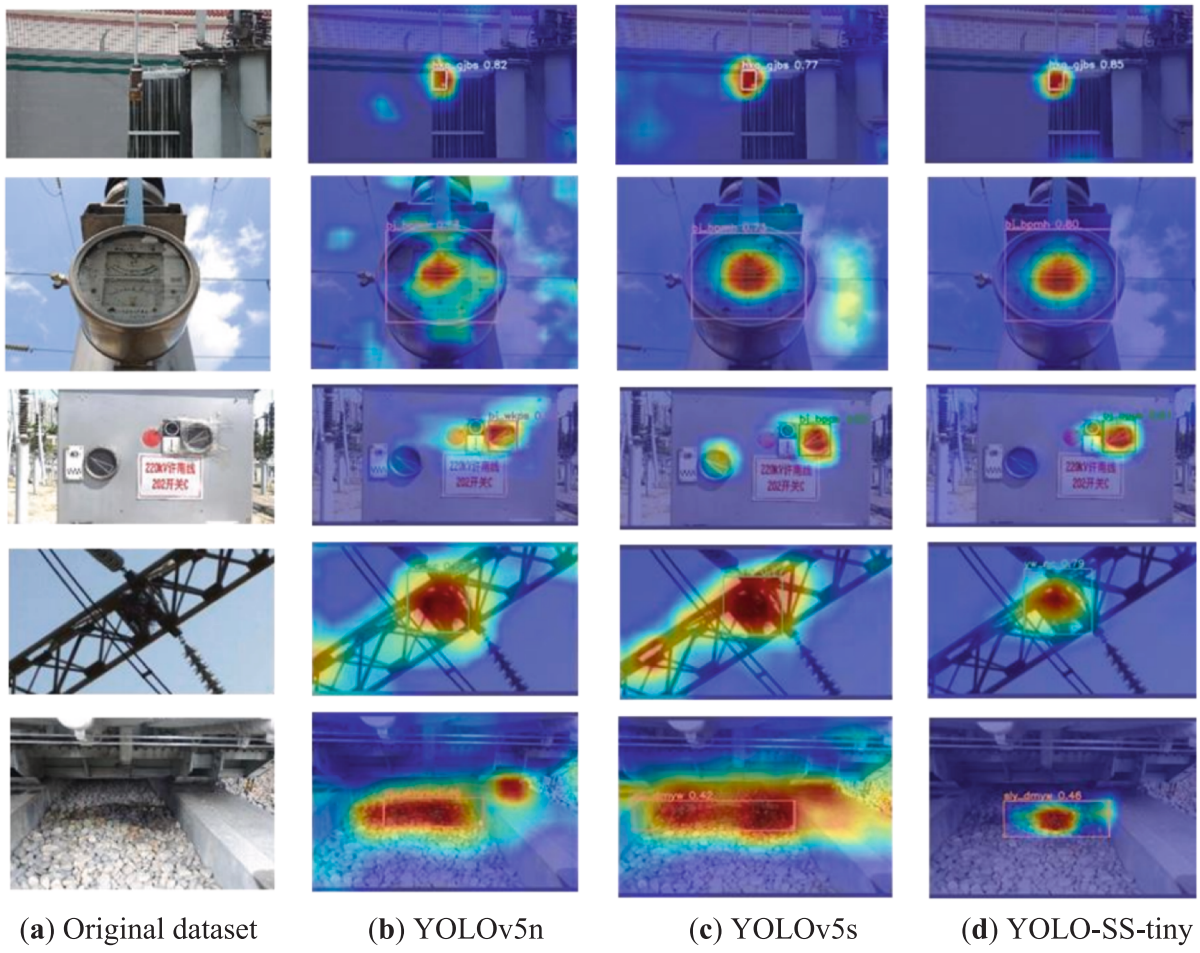


Fig. 14. Comparison of Grad-CAM thermodynamic Maps of Different Models in Actual Substation Scenarios.

Table 11
Comparison with other advanced object detection methods.

Model	mAP/%	Para./M	FPS
Faster-RCNN	40.3	138.36	—
SSD-MobileNetv2	44.2	8.8 M	84.7
YOLOv7-tiny	61.1	6.04 M	259.0
YOLOv8n	68.3	3.16 M	333.3
YOLO-SS-tiny	69.1	2.35 M	274.2

making it difficult for the traditional manual on-site inspection to meet the operation and maintenance of substations and safety control. In addition, the complex equipment environment and limited arithmetic power resources in the substations have seriously hindered the intelligent development of substations and power systems.

This paper presents an improved model and experimental analysis based on YOLOv5n and proposes a lightweight model, YOLO-SS-tiny, for substation equipment defect detection. Compared with the basic model, YOLO-SS-tiny can effectively balance the contribution of low-quality and high-quality samples to losses and has better comprehensive detection performance in substation environment with poor distinguishability between detection objects and backgrounds and serious mutual object occlusion, which provides a new implementation idea for intelligent inspection in the new stage of active distribution network development and has good engineering practice significance. In future research, we will use more advanced methods to improve and optimize the model for target features, background environments, and other different scenarios. At the same time, it is compared and validated with many other excellent object detection models to continuously improve the model's detection performance.

CRediT authorship contribution statement

Qian Wang: Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Rui Liu:** Writing – original draft, Resources, Investigation, Data curation. **Sichen Qin:** Writing – review & editing, Software, Funding acquisition, Data curation. **Jiawei Pu:** Writing – original draft, Visualization, Software, Investigation. **Rong Shi:** Validation, Supervision, Project administration. **Yulu Wang:** Visualization, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Projects Nos. 52577033, 52107029, the China Postdoctoral Science Foundation (No. 2021M692599), the Shaanxi Provincial Key R&D Plan Fund (No. 2023-YBGY-132), and the Shaanxi Provincial Department of Education Youth Innovation Team Research Plan Project (No. 23JP102).

Data availability

Data will be made available on request.

References

[1] Liu W, Gu W, Li P, Cao G, Shi W, Liu W. Non-iterative semi-implicit integration method for active distribution networks with a high penetration of distributed generations. IEEE Trans Power Syst 2021;36(1):438–50. https://doi.org/10.1109/TPWRS.2020.3003367.

[2] Xiao H, Pei W, Deng W, Ma T, Zhang S, Kong L. Enhancing risk control ability of distribution network for improved renewable energy integration through flexible DC interconnection. Appl Energy 2021;284(116387). https://doi.org/10.1016/j.apenergy.2020.116387.

[3] Pan JS, Wang HJ, Nguyen TT, Zou FM, Chu SC. Dynamic reconfiguration of distribution network based on dynamic optimal period division and multi-group flight slime mould algorithm. Electr Pow Syst Res 2022;208:107925. https://doi.org/10.1016/j.epsr.2022.107925.

[4] Zhao Z, Cai W, Wang Y, Xiong J, Liu W. Day-ahead electricity pricing mechanism considering the conflict between distribution network congestion and renewable produce. Int Trans Electr Energy Syst 2021;31(12). https://doi.org/10.1002/2050-7038.13218.

[5] Liu W, Wu Q, Wen F, Østergaard J. Day-ahead congestion management in distribution systems through household demand response and distribution congestion prices. IEEE Trans Smart Grid 2014;5(6):2739–47. https://doi.org/10.1109/TSG.2014.2336093.

[6] Qinzhang S, Bin L and Zhuangli H U: Discussion on Intelligent Operation and Maintenance Method of High Voltage Cable Line Based on Integrated Monitoring. In: Electric Engineering. (2019).

[7] Wang, H., Han, J., Zhang, K., Wang, F. and Fan, S: Study on Operation and Maintenance Management Mode of High-Voltage Transmission Lines. In: Application of Intelligent Systems in Multi-modal Information Analytics. 929, pp 1353-1357 (2019). 10.1007/978-3-030-15740-1_168.

[8] Cui, Y., Huang, X., Zhang, X., Ye, J. and Zhong, L: A Defects Detection System for Substation Based on YOLOX. In: Proceedings of the 2022 IEEE 5th International Electrical and Energy Conference (CIEEC); May. pp 4703–4707 (2022). 10.1109/CIEEC54735.2022.9846606.

[9] Minar, M.R. and Naher, J: Recent Advances in Deep Learning: An Overview (2018). 10.13140/RG.2.2.24831.10403.

[10] Sun, R: Optimization for Deep Learning: Theory and Algorithms (2019). 10.48550/arXiv.1912.08957.

[11] Zou, Z., Chen, K., Shi, Z., Guo, Y. and Ye, J: Object Detection in 20 Years: A Survey. In: Proceedings of the IEEE. 111(3), pp 257-276 (2023). 10.48550/arXiv.1905.05055.

[12] Wang Y, Guan Y, Liu H, Jin L, Li X, Guo B, et al. VV-YOLO: a vehicle view object detection model based on improved YOLOv4. Sensors 2023;23:3385. https://doi.org/10.3390/s23073385.

[13] Wu C, Ye M, Zhang J, Ma Y. YOLO-LWNet: a lightweight road damage object detection network for mobile terminal devices. Sensors 2023;23:3268. https://doi.org/10.3390/s23063268.

[14] Song W, Suandi SA. TSR-YOLO: a Chinese traffic sign recognition algorithm for intelligent vehicles in complex scenes. Sensors 2023;23:749. https://doi.org/10.3390/s23020749.

[15] Gevorgyan Zhora: SIoU Loss: More Powerful Learning for Bounding Box Regression. (2022). arXiv abs/2205.12740.

[16] Long X, Deng K, Wang G, Zhang Y and Wen S: PP-YOLO: An Effective and Efficient Implementation of Object Detector (2020). arXiv.2007.12099.

[17] J. Hu, L. Shen and G. Sun: Squeeze-and-Excitation Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp 7132-7141 (2018). 10.1109/CVPR.2018.00745.

[18] Wang QL, Wu BG, Zhu PF, Li PH, Hu QH. ECA-Net: Efficient Channel attention for Deep Convolutional Neural Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020. https://doi.org/10.1109/CVPR42600.2020.01155.

[19] Y. Lee and J. Park: CenterMask: Real-Time Anchor-Free Instance Segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp 13903-13912 (2020). arXiv:1911.06667.

[20] Chen, Y.P., Kalantidis, Y., Li, J.S.; Yan, S.; Feng, J.S: Aˆ2-Nets: Double Attention Networks. (2018). arXiv:1810.11579.

[21] Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S: CBAM: Convolutional Block Attention Module. In: Computer Vision – ECCV 2018. Lecture Notes in Computer Science. 11211 (2018).

[22] Liu, Y.C., Shao, Z. and Hoffmann, N: Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. (2021). arXiv:2112.05561.

[23] Liu, H.J., Liu, F.Q., Fan, X.Y. and Huang, D: Polarized Self-Attention: Towards High-Quality Pixel-Wise Regression. (2021). arXiv:2107.00782.

[24] Sadykova D, Pernebayeva D, Bagheri M, et al. IN-YOLO: Real-time detection of outdoor high voltage insulators using UAV imaging. IEEE Trans Power Delivery 2019;35(3):1599–601. https://doi.org/10.1109/TPWRD.2019.2944741.

[25] Hou Y.M. and Di J.M: Application of Improved Scale Invariant Feature Transform Accurate Image Matching in Target Positioning of Electric Power Equipment. In: Proceedings of the CSEE. 32(19) 2012 134-139.

[26] Shao JX, Yan YF, Qi DL. Substation switch detection and state recognition based on hough forests. Auto Electric Power Sys 2016;40(11):115–20. https://doi.org/10.7500/AEPS20150524001.