Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **11**, Number **1**, February **2026**

Substation Equipment Defect Detection Based on Attention Mechanism and Multi-Scale Feature Fusion

Jin-Ling Yang[1,*], Lei-Lei Chen[1], An-Hong Wang[1], Lin Li[1]

[1]School of Electronic Information Engineering,
Taiyuan University of Science and Technology, ShanXi 030024, China
yangjl@tyust.edu.cn, s202215210605@stu.tyust.edu.cn, ahwang@tyust.edu.cn, 2023077@tyust.edu.cn

Lin-Jian Chen[2]

[2]Huanjiang Laboratory, ZheJiang 311800, China
1326891880@qq.com

Jun-Hui Hou[3]

[3]Department of Computer Science,
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
1746328732@qq.com

[*]Corresponding author: Jin-Ling Yang

Received October 14, 2024, revised March 27, 2025, accepted May 19, 2025.

Abstract. *In the complex inspection environment of substations, the detection targets exhibit multi-scale characteristics and some defect targets are highly similar to the background, which leads to false detections and missed detections during inspections. This paper proposes substation equipment defect detection algorithm by YOLOv8, which incorporates attention mechanisms and multi-scale feature fusion to reduce the false detection rate and missed detection rate during inspections. Firstly, to suppress complex background interference, the Swin Transformer attention mechanism module is introduced into the traditional YOLOv8 network to enhance the prominence of defect targets. Secondly, to address the issue of insufficient feature representation leading to missed and false detections, a multi-scale adaptive weighted fusion module is designed to enhance the network's ability to obtain more comprehensive and accurate information from the feature maps. Thirdly, an FS-IoU is proposed to replace the CIoU in the original model, combining the advantages of Focal-IoU and Shape-IoU to reduce the occurrence of overlapping boxes and further improve the model's detection accuracy. The experimental results show that the average accuracy of substation equipment is improved by 3.8 % through the improved algorithm, and then the effectiveness of the algorithm is proposed.*
Keywords: Substation equipment defect detection, Attention mechanism, Multi-scale feature fusion, YOLOv8, Loss function

1. **Introduction.** The safe operation of various equipment in substations is crucial for the overall safety and productivity of the power system. Equipment or components within substations may exhibit defects such as damaged respirators, discolored silicone, blurred or cracked dials, or the presence of foreign objects and nests, which could potentially lead to power outages or accidents [1]. Traditionally, defect detection in substations relies on manual inspections, where problems are recorded and reported manually. This method is time-consuming, And it is easy to miss and misdetect. With the rapid advancement of artificial intelligence technologies, intelligent inspection systems for substations has

become increasingly important [2]. Leveraging deep learning techniques to enhance the performance of defect detection in substation equipment has emerged as one of the key research areas in the field of power vision [3].

Deep learning-based object detection algorithms can be categorized into two types: two-stage detection algorithms and single-stage detection algorithms. The two-stage detection algorithms first extract candidate regions from an image and then perform feature extraction and detection on these regions. Representative algorithms include Region-based Convolutional Neural Network (R-CNN) [4], Fast R-CNN [5], and Faster R-CNN [6]. Li et al. [7] used Faster R-CNN for defect detection in infrared images of substation equipment. Yin et al. [8] achieved detection and identification of four common defects in substations using an improved Faster R-CNN. However, two-stage object detection algorithms are often characterized by larger model sizes and poorer real-time performance. On the other hand, single-stage detection algorithms do not require candidate region generation; instead, they directly predict the classes and locations of objects. Common algorithms include SSD [9] and the YOLO series [10, 11, 12]. These algorithms offer both high detection accuracy and good real-time performance. Currently, many researchers have applied single-stage detection algorithms to the field of power inspection, achieving favorable detection results. For example, Li et al. [13] proposed a YOLO-AFB multi-target detection algorithm for substations, which is based on attention mechanisms and feature balancing to address the difficulties of multi-target recognition in complex substation environments. Zhao et al. [14] introduced an improved YOLOX-based algorithm for detecting appearance defects in substation instruments, addressing the issues of diverse defect characteristics and difficulty in feature extraction. Yan et al. [15] proposed a lightweight defect recognition method based on an improved YOLOv5-LITE for rapid localization and identification of distribution component defects. Pei et al. [16] proposed an improved YOLOv8 algorithm to address the problem of reduced detection performance in different environments. This was achieved by introducing a Transformer attention mechanism into the backbone network and applying a multi-attention mechanism detection head network, which enhanced detection accuracy for transmission line defects across various environments. Overall, for object detection in complex backgrounds, there is still room for improvement in feature extraction capabilities.

Contribution. To address this, the paper proposes a YOLOv8-based method for detecting defects in substation equipment and components, incorporating attention mechanisms and multi-scale feature fusion. The main contributions are:

(1) For low significance of defect target in complex background, this paper introduces Swin Transformer attention mechanism in YOLOv8 backbone network to enhance the prominence of defect target.

(2) In this paper, the feature fusion structure is changed to BiFPN structure, and a multi-scale feature adaptive weighted fusion module is designed on the basis of BiFPN structure, so as to efficiently complete the feature fusion between different scales.

(3) The proposed FS-IoU loss function is introduced to enhance the aggregation of low-quality anchor frames by focusing on the shape and size of the bounding frame. These improvements collectively enhance the model's performance in defect detection for substation equipment.

2. **Problem description.**

In the analysis of defect detection results for equipment components at a substation using inspection robots, issues with the algorithm model, including missed and false detections, have been identified. This is particularly concerning for critical equipment components where missed detections could pose significant safety risks. For

example, missed detection of oil leaks in a 35kV/0.4kV transformer can result in insufficient internal oil, causing overheating under normal load, which affects the transformer's stable operation and lifespan. Additionally, oil spills on the ground create a fire hazard. Similarly, missed detection of discoloration or damage to the silicone of a breather can impair the transformer's oil insulation, leading to internal risks. These defects could result in transformer accidents or economic losses. Therefore, this paper aims to improve the YOLOv8 algorithm model to enhance the detection accuracy for transformer oil leaks, silicone discoloration or damage, and other defects such as instrument damage, blurriness, and bird nests, thus better achieving unmanned intelligent substation inspection.

3. **Algorithm Improvement.** The algorithm framework for equipment defect detection at substations proposed in this paper is shown in Figure 1.

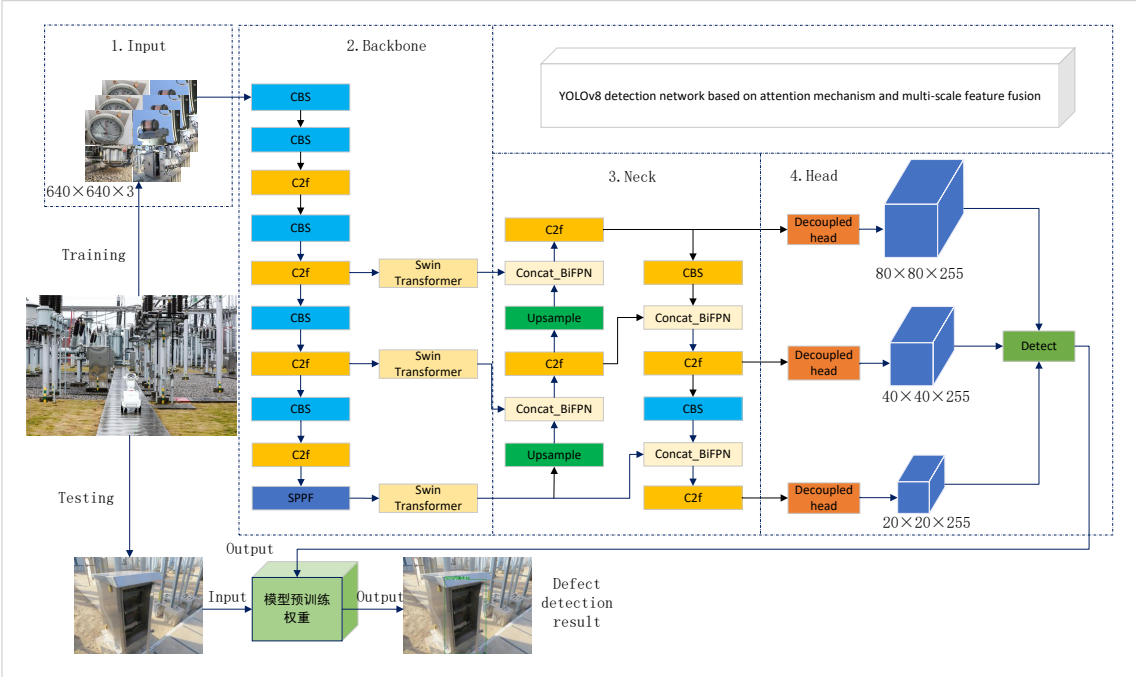

Figure 1. Detection framework of the proposed algorithm

The process for intelligent defect detection in substations is as follows: 1) Use unmanned vehicles to capture field defect images, then divide the images into training and testing sets; 2) Annotate defects in the training images; 3) Train the improved algorithm program on a GPU unit for a certain period to obtain the best model pre-trained weights; 4) Load the model pre-trained weights onto the onboard card of the unmanned vehicle to implement unmanned defect detection at the substation.

3.1. **Swin Transformer.** Due to the complex background of the substation, the defect targets have low prominence. By integrating the Swin Transformer [17] into the feature extraction backbone network, this module can efficiently capture both global and local feature information, enhancing the ability to extract low-prominence features. The Swin Transformer has two types of attention modules: the Window Multi-Head Self-Attention Module (W-MAS) and the Shifted Window Multi-Head Self-Attention Module (SW-MSA). The main components of the Swin Transformer structure include W-MAS, SW-MSA, LN, and MLP, as shown in the figure below.

The overall process is as follows: After feature input, normalization is performed first, followed by W-MSA operation. The output is then added to the initial input to obtain new

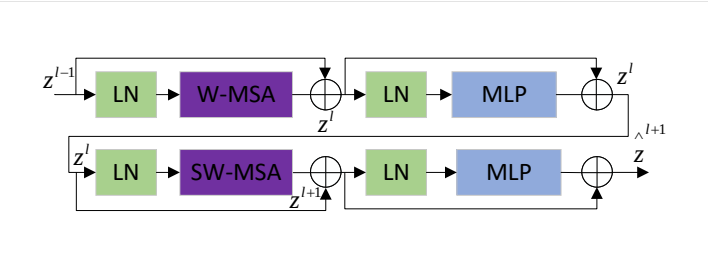

Figure 2. Diagram of Swin Transformer structure

features, which are normalized again before being processed by the MLP layer. Finally, the MLP layer is added to the result from the previous normalization to obtain new features for input into the next module. In the next module, the W-MSA from the previous steps is replaced with SW-MSA, and the process is repeated to achieve the final result.

The core of the Swin Transformer is W-MAS and SW-MAS, with the shifted window mechanism illustrated in Figure 3. MSA grows linearly with the size of the input image,

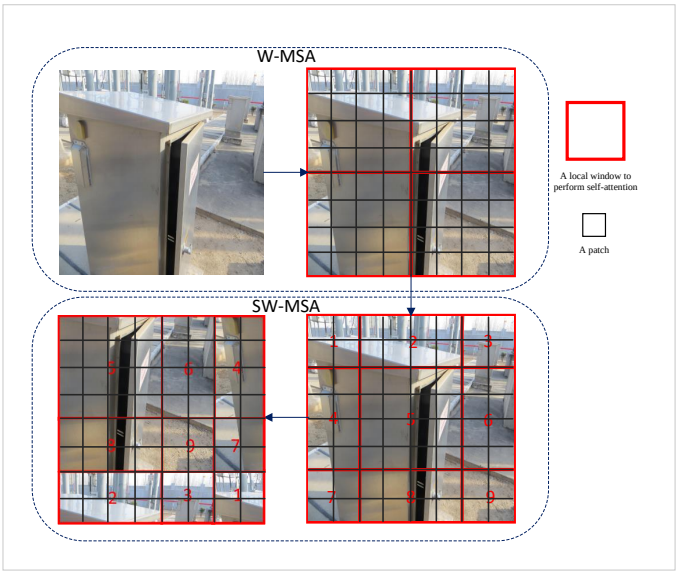

Figure 3. Diagram of relocating shifted windows

while global self-attention results in quadratic complexity. W-MSA reduces the sequence length by partitioning the image into 4 windows, thereby effectively decreasing the computational complexity. Assuming the image size is and each window contains patches, the computational complexity of the global MSA module and W-MSA is given by Formula (1):

$$\begin{aligned} \Omega(MSA) &= 4hwC^2 + 2(hw)^2C \\ \Omega(W-MSA) &= 4hwC^2 + 2M^2hwC \end{aligned} \tag{1}$$

W-MSA effectively addresses memory and computational issues, but it lacks communication between windows; each self-attention operation is confined to small windows and cannot capture information from other windows, limiting global understanding. To address this, a shifted window partitioning method (SW-MSA) is introduced, dividing the image into 9 windows. Each window includes at least two of the windows defined by W-MSA, enhancing connections between independent windows. This method reduces information loss and facilitates information exchange and global modeling, thus ensuring model accuracy.

3.2. **Multi-Scale Adaptive Feature Fusion Module.** Effective feature fusion can enhance object detection accuracy. Feature pyramids facilitate feature fusion, with different feature pyramid structures illustrated in Figure 4. The Feature Pyramid Network (FPN [18]) introduces top-down channel paths to fuse multi-scale features from level 3 to level 7, improving the semantic representation of multi-scale features, as shown in Figure 4(a). This method is limited by the unidirectional information flow of the top-down path. To address this limitation, FPN+PAN [19] is used, which adds an additional bottom-up aggregation process. This process combines shallow feature maps (low resolution but weak semantic information) with deep feature maps (high resolution but rich semantic information) and transmits features along specific paths. FPN transfers semantic information from high to low dimensions, while PAN transfers semantic information from low to high dimensions. This operation further enhances multi-scale feature representation and improves detection accuracy. YOLOv8 uses the FPN+PAN structure in the Neck layer, featuring both top-down and bottom-up characteristics, as shown in Figure 4(b). This structure is complex and computationally intensive, and it does not directly fuse features between different levels, limiting feature fusion capability. To address these issues, a more efficient BiFPN [20] feature fusion structure is adopted, as shown in Figure 4(c). This structure removes nodes with only one feature input, reducing computational load with minimal impact on feature fusion, and connects input nodes with lower-level nodes to achieve feature fusion across different levels, thereby enhancing feature representation capability.

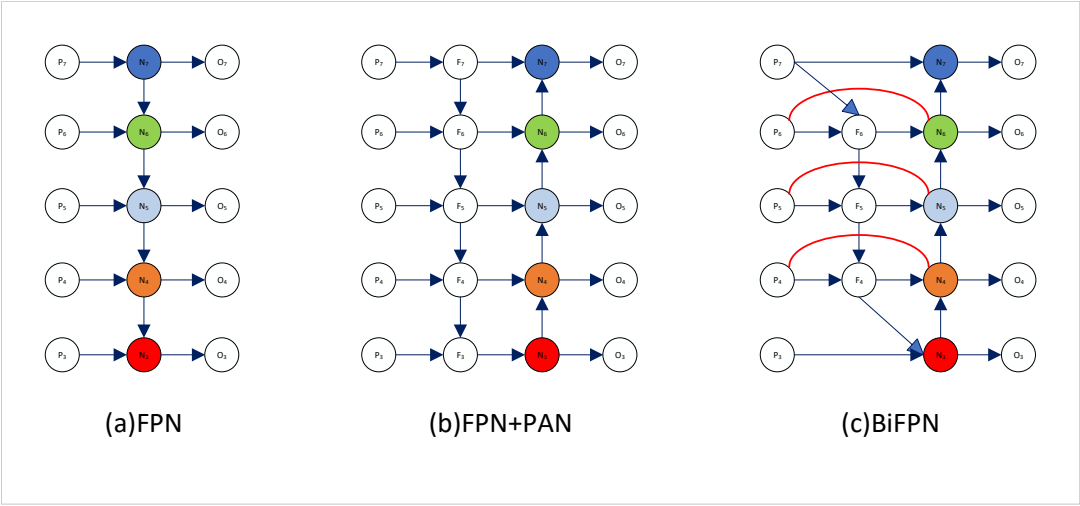

Figure 4. Structure of different characteristic pyramids

After introducing the BiFPN feature fusion structure in the Neck layer, features of the same scale can be efficiently fused, but features of different scales cannot be effectively fused. Therefore, based on the idea of Adaptive Spatial Feature Fusion (ASFF [21]), this paper designs a Multi-Scale Adaptive Weighted Feature Fusion Module (MAWFF), as shown in Figure 5.

The main idea of this module is divided into two steps: First, identity scaling between different scales. Initially, set the features of input layer Level $l \in \{1, 2, 3\}$ to x_l . YOLOv8 has feature maps at three different input levels with varying scales and channel numbers. Therefore, it is necessary to adjust the upsampling and downsampling for each scale to align the features of other level $n(n \neq l)$ with the scale of Level x_l .

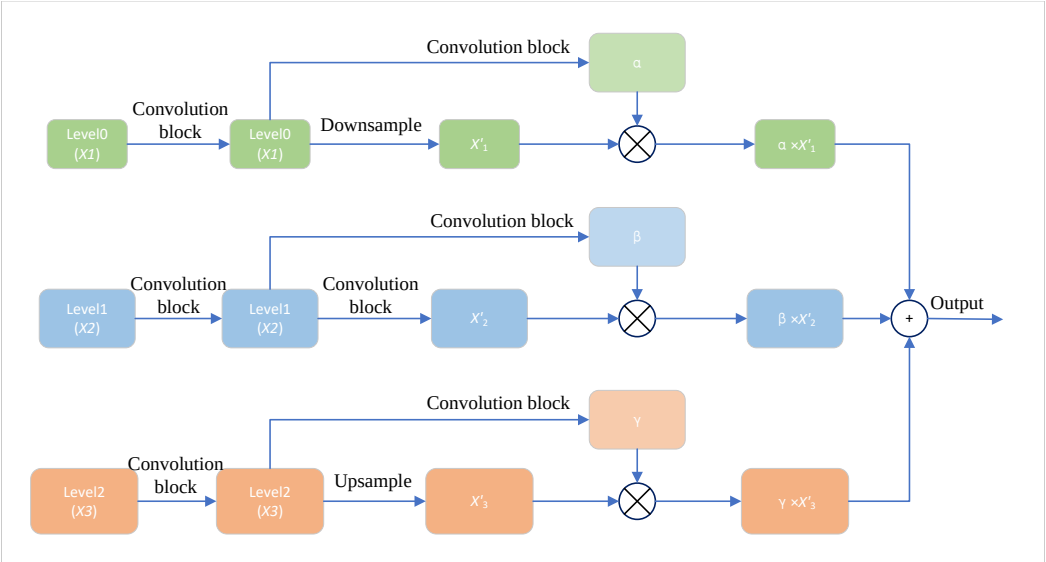

Figure 5. MAWFF module structure

(1) In MAWFF-1, the feature map of Level 2 undergoes a 1×1 convolution operation and is upscaled by 2 times to get $x^{2\to1}$, while the feature map of Level 3 undergoes a 1×1 convolution operation and is upscaled by 4 times to get $x^{3\to1}$.
(2) In MAWFF-2, the feature map of Level 1 undergoes a 1×1 convolution operation with a stride of 2 to get $x^{1\to2}$, and the feature map of Level 3 undergoes a 1×1 convolution operation and is upscaled by 2 times to get $x^{3\to2}$.
(3) In MAWFF-3, the feature map of Level 1 undergoes both max pooling and a 3×3 convolution operation with a stride of 2 to get $x^{1\to3}$, while the feature map of Level 2 undergoes a 3×3 convolution operation with a stride of 2 to get $x^{2\to3}$.

In the second step, adaptive weighted fusion is performed. After completing the first step, three new features x'_1, x'_2, x'_3 are obtained. Simultaneously, weight coefficients α, β, γ, and α, β, γ are learned from each input layer. These weight coefficients, obtained through convolution operations on the feature maps of Level 1 to Level 3, satisfy the following relationship:

$$\alpha_{ij}^{l} + \beta_{ij}^{l} + \gamma_{ij}^{l} = 1, \quad \alpha_{ij}^{l}, \beta_{ij}^{l}, \gamma_{ij}^{l} \in [0, 1] \tag{2}$$

Finally, the new features are multiplied by the weight coefficients and summed to obtain the new fused feature for level l, as shown in the following formula:

$$\text{Feature}_o = \alpha_{ij}^{l} \otimes x_{ij}^{1\to l} + \beta_{ij}^{l} \otimes x_{ij}^{2\to l} + \gamma_{ij}^{l} \otimes x_{ij}^{3\to l} \tag{3}$$

3.3. **Loss function.** In the YOLO series, the loss function is used to measure the difference between the predicted bounding boxes and the ground truth boxes, and to update the model parameters based on these differences. The bounding box loss function plays a crucial role in object detection, and improving it can enhance the precision of model training. In the original YOLOv8 object detection framework use the CIoU loss function. This function considers only the geometric relationship between the predicted boxes and the ground truth boxes, calculating the loss based on the aspect ratio, overlapping area, and distance between the centers of the boxes. However, this approach overlooks the impact of the intrinsic properties of the bounding boxes, such as their shape and size, on bounding box regression.

To address this limitation, this paper proposes an improved loss function-FS-IoU, which combines the advantages of Focaler-IoU [22] and Shape-IoU [23] loss functions to improve

model performance. The Focaler-IoU method reconstructs the IoU loss using a linear interval mapping approach to adaptively focus on different regression samples in various detection tasks. Shape-IoU considers the intrinsic shape and size of the bounding boxes. By integrating Focaler-IoU with Shape-IoU, the FS-IoU loss function is developed. FS-IoU can adaptively adjust the loss function based on the difficulty of the samples and the shape and size of the bounding boxes, achieving more precise optimization for bounding box regression. The calculation structure is illustrated in Figure 6.

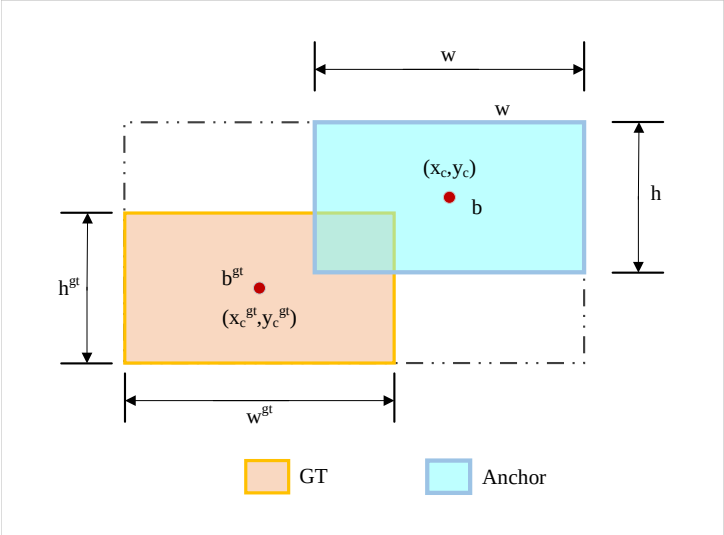

Figure 6. Diagram of FS-IoU calculation structure

The calculation formula is as follows:

$$IoU^{focaler} = \begin{cases} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \le IoU \le u \\ 1, & IoU > u \end{cases} \tag{4}$$

The weight coefficients for the horizontal and vertical directions are as follows:

$$\begin{aligned} ww &= \frac{2 \times (w^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \\ hh &= \frac{2 \times (h^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \end{aligned} \tag{5}$$

Here, scale is the scale factor related to the size of the detection targets in the dataset, typically ranging from 0 to 1.5. As the target size decreases, the absolute shape has a greater impact on the IoU value, and the scale value increases accordingly.
The formula for the distance loss function is as follows:

$$\text{distance}^{shape} = hh \times \frac{(x_c - x_c^{gt})^2}{c^2} + ww \times \frac{(y_c - y_c^{gt})^2}{c^2} \tag{6}$$

The formula for the shape loss function is as follows:

$$\Omega^{shape} = \sum_{t=w,h} \left((1 - e^{-\omega_t})^{\theta} \right) = \left(1 - e^{-\omega_w}\right)^{\theta} + \left(1 - e^{-\omega_h}\right)^{\theta} \tag{7}$$

$$\begin{cases} w_w = hh \times \frac{|w - w^{gt}|}{\max(w, w^{gt})} \\ w_h = ww \times \frac{|h - h^{gt}|}{\max(h, h^{gt})} \end{cases} \tag{8}$$

Therefore, the corresponding bounding box regression loss is as follows:

$$\text{Loss}_{FS-IoU} = 1 - IoU^{focaler} + distance^{shape} + 0.5 \times \Omega^{shape} \tag{9}$$

4. **Experiments and results analysis.** Model of this experiment are based on the PyTorch 1.12.0 deep learning framework, with PyCharm as the development tool. Training parameters include a batch size and number of workers both set to 8, with a total of 200 epochs for training.

Table 1. Experimental platform configuration parameters

Configuration Name	Version Parameters
GPU	NVIDIA GeForce GTX 1650 (4GB)
CPU	Intel Core i7-12700F @2.10 GHz
Operating System	Windows 10
CUDA	11.3
OpenCV	3.4.6
Python	3.9.0

4.1. **Experiment Data.** The dataset used in the experiment was collected from a substation site in recent years. Some images of fault targets are shown in Figure 7.

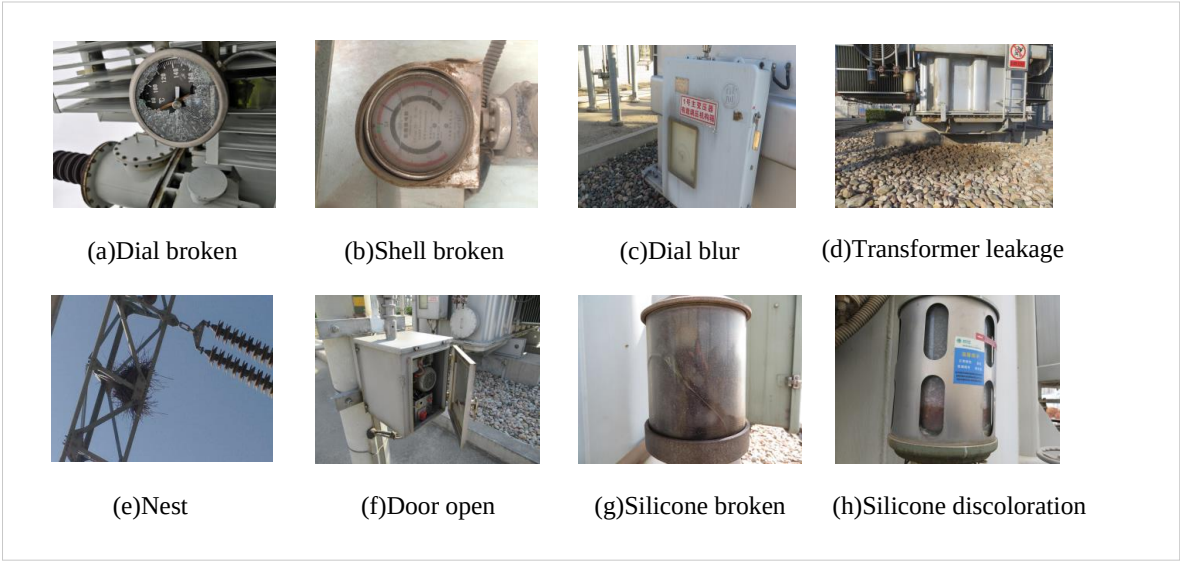

Figure 7. Substation equipment fault targets

The dataset used in this study consists of 7,000 images of substation equipment defects. All images were annotated using the Make Sense AI annotation tool, with a total of 10 defect categories. The definitions of these categories are shown in Table 2. The samples for each defect category were divided into training, validation, and test sets in an 8:1:1 ratio, with 5,600 images allocated for training, 700 images for validation, and 700 images for testing.

4.2. **Evaluating indicator.** To objectively evaluate the performance of the YOLOv8 network with attention mechanisms and multi-scale feature fusion in substation equipment defect detection, this study selects precision, recall, and mean average precision (mAP) as evaluation metrics.

Table 2. Faults types and quantity

Fault Type	Label Name	Number of Failures
Error list number	bj-dsyc	789
Shell broken	bj-wkps	523
Dial blur	bj-bpmh	869
Dial broken	bj-bpps	723
Silicone discoloration	hxq-gjbs	1174
Silicone broken	hxq-gjtps	106
Nest	yw-nc	883
Suspended solids	yw-gkxfw	729
Door open	xmbhyc	383
Transformer leakage	sly-dmyw	833

Precision is defined as the ratio of correctly detected objects to all detected objects, as shown in Formula (10). N_{TP} represents the number of correctly identified samples, while N_{FP} represents the number of false positives and missed detections.

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

Recall is defined as the ratio of correctly detected targets to all actual targets, as shown in Formula (11). N_{FN} represents the number of false negatives (missed detections) in the samples.

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

Mean Average Precision (mAP@0.5) is defined as the mean of the average precision values across all categories. Here, N represents the number of sample labels. The formula for mAP is shown in Equation (12).

$$m_{AP}@0.5 = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{12}$$

4.3. **Improved method effect comparison.**

4.3.1. **Comparison Experiment on Attention Mechanisms.** To evaluate the performance of the Swin Transformer attention mechanism in substation equipment fault detection, comparative experiments were conducted under identical network positions and experimental conditions. The attention mechanisms compared include SE [24], ECA [25], CBAM [26], and Swin Transformer. The experimental results are presented in Table 3.

Table 3. Comparison with different attention mechanisms

Attention Mechanisms	P/%	R/%	mAP@0.5/%	GFLOs/G	Params/M
SE	66.1	62.6	64.8	8.3	3.01
ECA	66.8	62.3	65.2	8.9	3.06
CBAM	67.3	63.9	65.6	8.5	3.02
Swin Transformer	69.7	64.2	67.5	21.9	3.45

From Table 3, it can be observed that the Swin Transformer attention mechanism module increases both computational complexity and the number of parameters, but it

significantly improves the model's detection performance. Specifically, when incorporating the Swin Transformer attention mechanism during training, the model's accuracy increased by 3.6 %, 2.9 %, and 2.4 % compared to using SE, ECA, and CBAM attention mechanisms, respectively.The SE attention mechanism generates channel-wise weights by performing global average pooling on the input image features and passing them through two fully connected layers. The ECA mechanism is an improvement over SE, removing the fully connected layers and replacing them with a 1×1 convolutional kernel, thus reducing the number of parameters and making the module more lightweight. The CBAM mechanism first applies channel attention and then spatial attention to the input image features, producing attention weights. In contrast, the Swin Transformer attention mechanism enhances the model's performance by combining W-MSA and SW-MSA self-attention models. This approach improves information exchange between neighboring windows and allows the model to better learn similar information in the surrounding neighborhood, thereby boosting performance in multi-scale feature detection tasks and increasing the model's detection accuracy.

4.3.2. **Comparison of Loss Functions.** To verify the effectiveness of the FS-IoU loss function, this study compares the Shape-IoU loss function with DIoU [27], GIoU [28], SIoU [29], Shape-IoU, and Focaler-EIoU, based on a model incorporating the first two improved modules. The comparison results, as shown in Table 4, indicate that the model using the FS-IoU loss function detects object locations more accurately and quickly, showing superior performance in defect detection.

Table 4. Comparison with different loss functions

IoU Loss	**P/%**	**R/%**	**mAP@0.5/%**
DIoU	67.9	63.1	67.6
GIoU	67.5	62.9	66.8
SIoU	68.2	63.5	67.3
Shape-IoU	68.6	63.5	67.5
Focaler-IoU	68.8	63.8	67.3
FS-IoU	70.1	64.6	68.6

4.3.3. **Ablation Study.** To validate the effectiveness of each module in the improved algorithm and demonstrate the efficacy of the proposed methods, an ablation study was conducted. Under identical experimental conditions, the Swin Transformer attention mechanism module (S-T), multi-scale feature fusion module (MAWFF), and FS-IoU loss function were added sequentially. The experimental results are shown in Table 5.

Table 5. Ablation Experiment

S-T	**MAWFF**	**FS-IoU**	**P/%**	**R/%**	**mAP@0.5/%**
—	—	—	67.6	62.2	64.8
✓	—	—	69.7	64.2	67.5
—	✓	—	68.6	63.9	66.3
—	—	✓	68	61.9	65.4
✓	✓	—	69.6	64.1	68.1
✓	✓	✓	70.1	64.6	68.6

As shown in Table 5, incorporating each of the three improved algorithms individually into the network framework enhances detection performance. Additionally, combining

these improvements within the detection network further optimizes the results. The final model achieved a 2.5 % increase in precision and a 2.4 % increase in recall, with the average precision mAP@0.5 improving by 3.8 %.

4.3.4. **Model Recognition Performance.** The original YOLOv8 algorithm and the improved algorithm were used for defect detection of substation equipment or components. The images show the detection results. In Figure 8, the first and second rows of panels (a) (c) display the detection results of defect targets in substations using the original YOLOv8 algorithm and the improved algorithm, respectively.

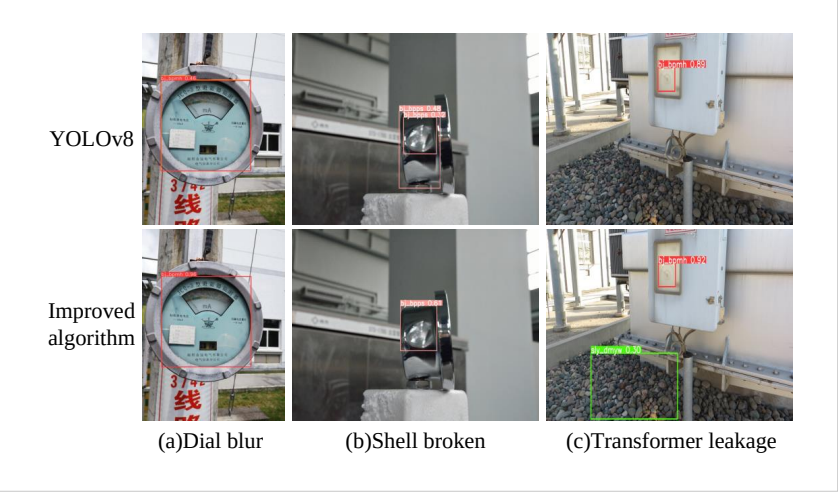

Figure 8. Diagram of recognition effect

From Figure 8(a), it can be seen that due to the unclear features and range of the defect target on the dial, and the lack of deep feature extraction, the original YOLOv8 algorithm shows lower confidence in defect detection. The improved algorithm enhances feature extraction capabilities by delving into deeper information, significantly increasing detection confidence for such defects. Figure 8(b) demonstrates that the original YOLOv8 algorithm struggles with inaccurate bounding box predictions when facing significant changes in target scale. The improved algorithm addresses this issue by extracting and merging multi-scale features and refining the loss function, which enhances the accuracy of bounding box predictions. Figure 8(c) shows that, when dealing with oil leakage faults, the original YOLOv8 algorithm fails to detect these faults due to the similarity between the oil leakage and the ground background. The improved algorithm optimizes feature transfer, enhances defect feature extraction, and reduces missed detections, successfully identifying oil leakage faults and validating the effectiveness of the proposed algorithm improvements.

5. **Conclusion.** To reduce false positives and missed detections in substation equipment defect detection and achieve unmanned intelligent substation inspections for a substation, this paper improves the original YOLOv8 algorithm. First, the Swin Transformer attention mechanism module is introduced to enhance feature extraction capabilities in complex backgrounds. Second, the original model's feature fusion module is replaced with BiFPN, the Neck layer add a multi-scale adaptive weighted feature fusion module, efficiently completing feature fusion across different scales. Finally, FS-IoU is proposed as the bounding box loss function, focusing on the shape and size of the bounding box to compute loss, making bounding box regression more accurate and further improving model precision in defect detection. The proposed algorithm is validated using collected

substation equipment defect data. Experimental results show that the improved algorithm significantly increases average precision when dealing with defects that are similar to their surroundings, exhibit large scale variations, or are obscured, effectively identifying substation equipment defects and proving its effectiveness in practical detection.Detecting defect targets that are extremely similar to their background, such as silicone discoloration or equipment oil leakage, the algorithm has improved detection performance to some extent. However, false positives or missed detections still occur, which remains a topic for further research.

REFERENCES

[1] H. Liu, X. Han, and Y. Mao, "Substation defect detection based on self-supervised learning," *Computer Systems Applications*, vol. 32, no. 05, pp. 112–122, 2023.

[2] J. He, G. Luo, M. Cheng, Y. Liu, Y. Tan, and M. Li, "A research review on application of artificial intelligence in power system fault analysis and location," *Proceedings of the CSEE*, vol. 40, no. 17, pp. 5506–5516, 2020.

[3] Z. Zhao, S. Feng, Y. Xi, J. Zhang, Y. Zhai, and W. Zhao, "The era of large models:a new starting point for electric power vision technology," *High Voltage Engineering*, vol. 50, no. 05, pp. 1813–1825, 2024.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[5] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[6] S. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[7] W. Li, K. Xie, X. Liao, X. Li, and H. Wang, "Intelligent diagnosis method of infrared image for transformer equipment based on improved faster rcnn," *Southern Power System Technology*, vol. 13, no. 12, pp. 79–84, 2019.

[8] Z. Yin, R. Meng, X. Fan, B. Li, and Z. Zhao, "Typical visual defect detection system of substation equipment based on edge computing and imporoved faster r-cnn," *China Sciencepaper*, vol. 16, no. 03, pp. 343–348, 2021.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[10] X. Zhang, H. Wang, D. Zhou, J. Li, and H. Liu, "Abnormal detection of substation environment based on improved yolov3," in *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol. 1. IEEE, 2019, pp. 1138–1142.

[11] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 213–10 224.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[13] B. Li, Y. Li, X. Zhu, S. Wang, L. Qu, J. Zeng, H. Liu, and Y. Tian, "Multi-target detection in substation scence based on attention mechanism and feature balance," *Power System Technology*, vol. 46, no. 06, pp. 2122–2132, 2022.

[14] Z. Zhao, D. Ma, S. Ying, and G. Li, "Appearance defect detection algorithm of substation instrument based on improved yolox," *Journal of Graphics*, vol. 44, no. 05, pp. 937–946, 2023.

[15] H. Yan, J. Wan, Z. Pan, J. Zhang, and R. Ma, "Defect identification of distribution components based on improved yolov5-lite lightweight," *High Voltage Engineering*, vol. 50, no. 05, pp. 1855–1864, 2024.

[16] S. Pei, H. Zhang, C. Hu, W. Yang, and Y. Liu, "The defect detection method for cross-environment power transmission line based on er-yolo algorithm," *Transactions of China Electrotechnical Society*, vol. 39, no. 09, pp. 2825–2840, 2024.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[19] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[20] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 781–10 790.

[21] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection. arxiv 2019," *arXiv preprint arXiv:1911.09516*, 1911.

[22] H. Zhang and S. Zhang, "Focaler-iou: More focused intersection over union loss," *arXiv preprint arXiv:2401.10525*, 2024.

[23] ——, "Shape-iou: More accurate metric considering bounding box shape and scale," *arXiv preprint arXiv:2312.17663*, 2023.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 534–11 542.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[28] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.

[29] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022.