electronics

MDPI

Article

A Multi-Scale Feature Network for Defect Detection in Substation Equipment

Changzheng Deng [1,*], Jingting Huang [1], Weidong Gao [1] and Bo Chen [2]

[1] College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China; 202408580121827@ctgu.edu.cn (J.H.); 202408080021621@ctgu.edu.cn (W.G.)
[2] Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
* Correspondence: dcz_1980@ctgu.edu.cn

Abstract

To address the challenges in defect detection of substation equipment, including large variations in target scale, complex background interference, and insufficient recovery of edge details, this paper proposes a defect detection method for substation equipment based on multi-scale feature perception. First, an MSEEM module is designed. It integrates multi-scale depthwise convolution, edge prior enhancement, and channel-adaptive calibration to improve the model's ability to represent defects at different scales. Second, a C2SCSA module is constructed to select effective features from both spatial and channel dimensions. This enhances the representation of defect regions under complex backgrounds. Finally, the DySample dynamic upsampling module is introduced during the feature fusion stage to adaptively restore fine-grained information through content-aware sampling, thereby improving bounding-box localization performance. Experiments on the substation equipment defect dataset constructed in this study show that the improved model can accurately identify different types of equipment defects in complex scenarios. Relative to YOLOv11n, MSDFNet increased precision, recall, mAP@0.5, and mAP@0.5:0.95 by 3.3, 7.0, 4.9, and 7.3 percentage points, respectively. These results indicate that the method improves the detection accuracy and localization ability of defects in substation equipment while maintaining low computational overhead.

Keywords: defect detection; substation equipment; multi-scale feature perception; YOLO; deep learning; intelligent inspection

1. Introduction

As a critical hub of the power system, the stable operation of substations is a fundamental prerequisite for ensuring the safe and reliable supply of electricity [1,2]. Many substations are located in remote mountainous areas or exposed outdoor environments. As a result, their equipment is subjected to harsh conditions such as prolonged sunlight, humidity, and weathering. These conditions can easily lead to defects, including corrosion, cracks, and foreign object adhesion [3]. If such defects are not detected and handled in time, they may cause equipment failures or even large-scale power accidents. Therefore, the rapid and accurate detection of defects in substation equipment is essential for reducing fault risks and maintaining the safe and stable operation of the power system [4].

Previous studies on defect detection in substation equipment can generally be divided into two categories: traditional inspection methods and computer vision-based methods.

Check for updates

Academic Editor: Xian-Bo Wang

Received: 23 July 2026
Revised: 8 October 2026
Accepted: 9 October 2026
Published: 21 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Traditional defect detection in substations mainly relies on regular manual inspection. Operation and maintenance personnel conduct on-site inspections according to established schedules. They also use basic portable instruments, such as infrared thermometers and megohmmeters, to collect information on equipment operating conditions [5]. During inspection, visible defects are mainly identified through visual observation. These defects include surface corrosion, insulator flashover, and abnormal oil levels. Inspectors also rely on hearing to determine whether unusual sounds occur during equipment operation. In addition, abnormal phenomena such as a burnt odor may be perceived through smell. Electrical parameters and oil quality are measured using basic instruments. The results are then manually compared with relevant technical standards and operating regulations [6]. This inspection mode has several limitations. It is easily affected by subjective judgment, which may lead to missed detections or false detections. Moreover, with the expansion of power grids and the increasing complexity of substation equipment, traditional manual inspection methods have become insufficient for the operational requirements of modern power systems. Against this background, computer vision technologies with automated and real-time detection capabilities have gradually been widely applied.

Deep learning-based object detection methods are mainly divided into two-stage and one-stage detectors. Two-stage detectors first generate candidate regions that may contain objects. These regions are then classified, and their locations are further refined. Representative two-stage detectors include R-CNN [7], Fast R-CNN [8], Faster R-CNN [9], and Mask R-CNN [10]. Chen et al. [11] proposed a lightweight Faster R-CNN-tiny model. The model uses depthwise separable convolution and transfer learning to optimize multi-scale feature extraction. It achieves high-accuracy and real-time detection of insulator defects. Xin et al. [12] improved the Mask R-CNN algorithm by introducing a global feature pyramid network (GFPN) and the convolutional block attention module (CBAM). This design strengthens feature fusion and object recognition, thereby improving the accuracy and robustness of defect detection. Liu et al. [13] incorporated multi-scale semantic information into an improved Faster R-CNN model. By combining a feature pyramid with deformable convolution, their method was applied to defect detection in substation equipment and showed better adaptability to complex environments. Although two-stage detectors usually have advantages in detection accuracy, they also require higher computational cost, more hardware resources, and stronger data support. These limitations make them less suitable for the real-time requirements of substation equipment defect detection. In contrast, one-stage detectors are more appropriate for scenarios where detection accuracy and inference speed need to be balanced.

Representative one-stage object detectors include the YOLO (You Only Look Once) series [14], SSD (Single Shot MultiBox Detector) [15], and RetinaNet [16]. Zheng et al. [17] proposed an improved fused single shot multibox detector (FSSD) model. By introducing a shallow feature enhancement module and using a clustering algorithm to adaptively optimize the aspect ratios of default boxes, their method improved the detection accuracy of insulators in complex substation scenes. However, the SSD framework itself suffers from computational redundancy and limited generalization ability. In contrast, the YOLO series has more evident advantages in detection speed, end-to-end inference, and engineering deployment. Therefore, it has been widely used in defect detection tasks for power equipment. Dong et al. [18] proposed PHAM-YOLO, which introduces a parallel hybrid attention mechanism into YOLOv5. This design strengthens the model's attention to defective regions on meters. Zhang et al. [19] proposed YOLO-SD. Their method uses ADD-GAN to expand defect image data and combines an attention mechanism with an NWD-CIoU fusion loss function. It achieves accurate detection of multi-scale defect targets in substations. However, this method depends heavily on the quality of generated samples,

and the model still contains a relatively large number of parameters. Wang et al. [20] proposed EAL-YOLO. They designed an ASF2-Neck and a lightweight shared-convolution detection head to improve small-defect detection. However, their method gives limited consideration to the adaptive recovery of defect edge details. Meng et al. [21] used the C2f-OD module to enhance the model's feature representation of irregular defect targets. This improved the detection accuracy of equipment defects to some extent. Nevertheless, the method mainly focuses on enlarging the receptive field and strengthening high-level semantic features. It pays insufficient attention to the recovery of fine-grained textures. Wang et al. [22] proposed a substation defect detection model with enhanced feature extraction and fusion. The model improves small-object detection through adaptive multi-scale feature fusion and a multi-branch detection head. However, the multi-branch detection structure may still introduce additional structural complexity in practical deployment.

Although the methods discussed above have made progress in defect detection for substation equipment, accurate defect recognition remains challenging. This is mainly due to the complex equipment structures and background information in substation scenes. As a result, missed detections and false detections may still occur in practical applications. To address the limitations of current defect detection algorithms, including weak feature extraction under complex backgrounds, limited detection accuracy, and insufficient recovery of boundary details, this paper proposes a defect detection method for substation equipment based on multi-scale feature perception, named the Multi-Scale Discriminative Feature Network (MSDFNet). The main contributions of this study are as follows:

(1) An MSEEM module, namely the Multi-Scale Edge-Aware Enhancement Module, is designed. It strengthens the network's ability to represent objects with scale variations by using branch-based multi-scale convolution and edge-response modeling.
(2) A C2SCSA module, namely C2-based Spatial and Channel Synergistic Attention, is constructed. This module improves the model's ability to automatically highlight defect-related features in complex backgrounds. It also suppresses background noise effectively.
(3) The DySample module is introduced to dynamically adjust sampling positions according to feature content. This further improves the detection performance for defects in substation equipment.

2. Network Structure of YOLOv11n

YOLOv11 [23] is an object detection model released by the Ultralytics team. It adopts a modular network design, which allows flexible configuration of models at different scales. According to differences in network depth and width, YOLOv11 is commonly divided into several versions, including YOLOv11n, YOLOv11s, YOLOv11m, YOLOv11l, and YOLOv11x. Among them, YOLOv11n has fewer parameters and lower computational complexity, making it more suitable for practical deployment. Considering the high requirements of defect detection in substation equipment for real-time performance and deployment convenience, YOLOv11n is selected as the baseline model for improvement in this study. The YOLOv11n model consists of four main parts: the input layer, the backbone, the neck, and the head. The input layer performs preprocessing and data augmentation. It is responsible for data normalization and enhancement before feature extraction. The backbone extracts multi-scale semantic features from the input image. It is jointly composed of the C3k2, C2PSA (Cross-Stage Partial with Spatial Attention), and SPPF (Spatial Pyramid Pooling-Fast) modules. Among these modules, C3k2 serves as the basic feature extraction unit. It adopts an optimized cascaded 3 × 3 convolution strategy, which improves the representation of edge details and local textures. The SPPF module aggregates contextual features under different receptive fields through parallel max-pooling operations. This

provides more consistent global information for subsequent feature fusion and detection. The C2PSA module, located at the end of the network, combines a multi-head attention mechanism with a feed-forward neural network. It further improves the model's ability to extract and represent features. The neck connects the backbone and the output layer. It strengthens the complementarity between deep semantic features and shallow detailed features through an efficient feature fusion mechanism. This structure contains repeated upsampling and concatenation operations, as well as C3k2 modules for further feature refinement. It is therefore suitable for detecting objects at multiple scales. The detection head receives the features output by the neck and performs object classification and bounding box regression. It is mainly composed of detection heads at multiple scales, which correspond to feature maps with different resolutions. This design enables the model to cover objects of various sizes and achieve object detection with both high accuracy and strong real-time performance.

3. Defect Detection Algorithm for Substation Equipment

To address the challenges in defect detection for substation equipment, YOLOv11n is selected as the baseline model in this study. Three improvements are made to enhance its detection performance. First, an MSEEM module is designed in the backbone. By using multi-scale depthwise convolution and an edge-prior enhancement mechanism, this module improves the model's ability to extract features from defects at different scales. It enables the network to capture defect information on power equipment more accurately. Second, to reduce the interference of complex backgrounds on effective defect features, a C2SCSA module is constructed. This module uses a spatial and channel collaborative attention mechanism to further improve detection accuracy and stability. Finally, the DySample module is introduced in the upsampling stage. It replaces conventional fixed interpolation with content-adaptive sampling, which improves the recovery of detailed features. This enhances the localization accuracy of defect boundaries. The overall structure of the improved model is shown in Figure 1.

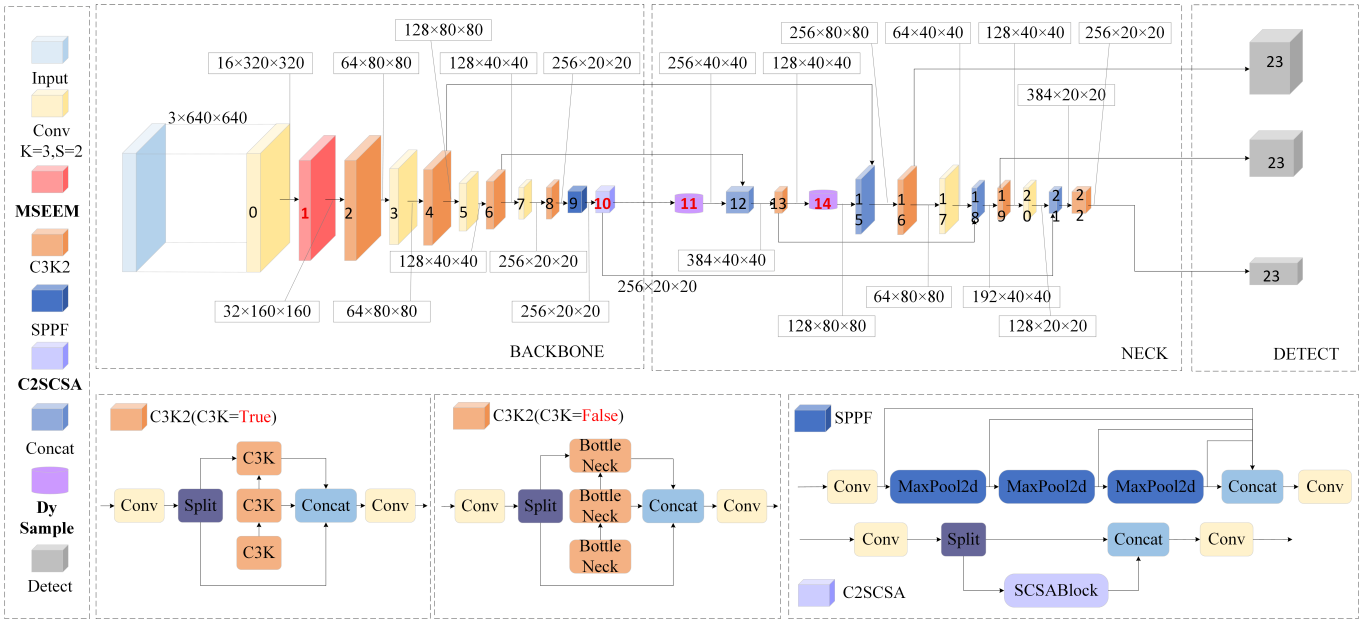


Figure 1. Overall architecture of the MSDFNet.

3.1. MSEEM Module

Due to the diverse types of defects in substation equipment, defect scales often vary in practical detection. This variation is also affected by the distance between the camera and

the equipment, as well as by different shooting angles. To address the large scale differences among defects and the missed detection of small defects, this paper designs a Multi-Scale Edge-Aware Enhancement Module (MSEEM). The proposed module uses global pooling operations at different scales together with scale-specific feature convolutions. In this way, it can capture both local details and contextual information related to defects in substation equipment. Meanwhile, an edge-prior enhancement mechanism is developed, and a lightweight channel-adaptive attention mechanism is introduced. These designs further strengthen the model's ability to extract defect features of different sizes while keeping the computational cost under control. The structure of MSEEM is shown in Figure 2.

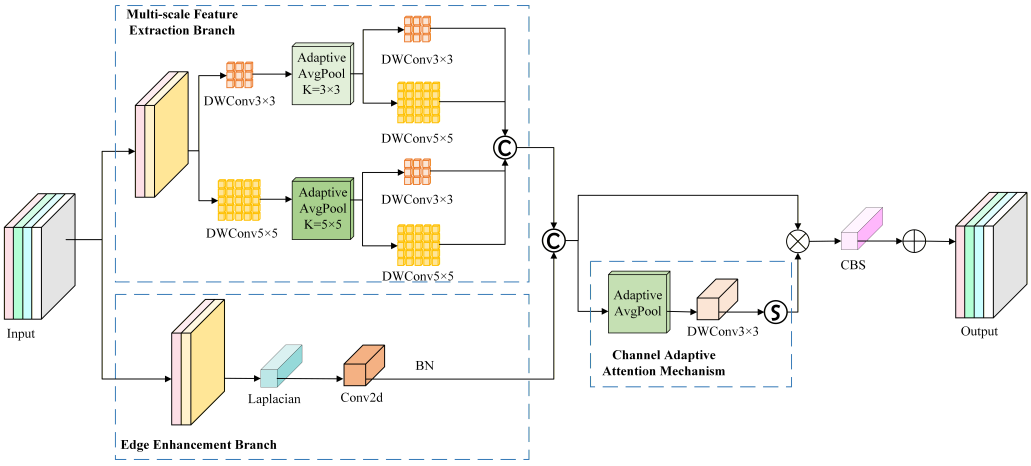


Figure 2. Architecture of the MSEEM module.

Let the input feature be $X \in R^{B\times C_{\text{in}}\times H\times W}$, where B, C_{in}, H, and W denote the batch size, the number of input channels, and the height and width of the feature map, respectively. MSEEM first divides the input feature into upper and lower branches along the channel dimension:

$$X_u = X[:, : \lfloor C_{\text{in}}/2 \rfloor, :, :] \tag{1}$$

$$X_l = X[:, \lfloor C_{\text{in}}/2 \rfloor : , :, :] \tag{2}$$

where X_u is used for multi-scale feature extraction, whereas X_l is used for high-frequency information enhancement. The upper branch X_u is further divided into two sub-features along the channel dimension to construct two scale-specific paths with 3×3 and 5×5 kernels, respectively.

The 3×3 path focuses on capturing local textures and subtle structural variations, such as cracks and corrosion, whereas the 5×5 path provides a larger effective receptive field and can aggregate contextual information over a wider spatial range, making it more suitable for representing relatively large defect regions, such as damage, contamination, and foreign-object coverage. By combining convolutions at different scales, fine-grained local information within smaller receptive fields and contextual information within larger receptive fields can be utilized simultaneously, thereby improving the adaptability of the network to defects with substantial scale variations.

Both paths first employ depthwise convolution for feature extraction, followed by average pooling with kernel sizes of 3 and 5, respectively, to aggregate local information. The stride of both pooling operations is set to 1, with padding values of 1 and 2, respectively; therefore, the spatial dimensions of the feature maps remain unchanged. The pooled

features are subsequently refined using parallel 3×3 and 5×5 depthwise convolutions. The calculation process is expressed as follows:

$$A = \mathrm{AvgPool}_{3,1,1}\left(\mathrm{DWConv}_{3\times3}\left(X_u^{(1)}\right)\right) \tag{3}$$

$$M = \mathrm{AvgPool}_{5,1,2}\left(\mathrm{DWConv}_{5\times5}\left(X_u^{(2)}\right)\right) \tag{4}$$

$$F_m = \mathrm{Concat}_c\left[\mathrm{DWConv}_{3\times3}(A), \mathrm{DWConv}_{5\times5}(A), \mathrm{DWConv}_{3\times3}(M), \mathrm{DWConv}_{5\times5}(M)\right] \tag{5}$$

where $\mathrm{DWConv}(\cdot)$ denotes depthwise convolution and $\mathrm{Concat}_c(\cdot)$ denotes feature concatenation along the channel dimension. Since the total number of channels of the two sub-features obtained by further dividing the upper branch is $C_{\mathrm{in}}/2$, and each sub-feature generates two feature paths, the concatenated multi-scale feature F_m contains C_{in} channels.

In parallel with the multi-scale branch, the lower branch introduces a Laplacian operator to supplement the local high-frequency information in the shallow features. For a multi-channel input, the first three channels are used to construct a single-channel projection feature X_g:

$$X_g = 0.299X_{l,1} + 0.587X_{l,2} + 0.114X_{l,3} \tag{6}$$

A fixed eight-neighborhood discrete Laplacian kernel is then employed to extract high-frequency responses:

$$K_L = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{7}$$

The extracted high-frequency responses are subsequently mapped to C_{out} channels through a 1×1 convolution and Batch Normalization:

$$F_l = \mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left(K_L * X_g\right)\right) \tag{8}$$

where $*$ denotes two-dimensional convolution. The Laplacian convolution uses a stride of 1 and a padding of 1. Its convolution kernel is stored in the network as a fixed parameter and is not updated during training. This branch highlights variations in local responses and supplements high-frequency features for weak-texture and low-contrast defects, such as blurred meter dials, discoloration of silica gel, and ground oil stains. It can also provide detailed information for defects with pronounced structural variations, such as damaged components and foreign objects.

The multi-scale feature F_m and the high-frequency feature F_l are subsequently fused along the channel dimension:

$$F_c = \mathrm{Concat}_c(F_m, F_l) \tag{9}$$

The dimensions of the fused feature are expressed as

$$F_c \in R^{B\times C_f\times H\times W} \tag{10}$$

where the number of channels of the fused feature is

$$C_f = C_{\mathrm{in}} + C_{\mathrm{out}} \tag{11}$$

To further model the response relationships among different types of features, a channel recalibration mechanism is applied after feature fusion. First, global average pooling is used to obtain a global descriptor for each channel:

$$z = \mathrm{GAP}(F_c) \tag{12}$$

The global channel descriptor is then broadcast to the spatial dimensions of the original feature map, followed by a 3×3 depthwise convolution for lightweight feature transformation:

$$q = \mathrm{DWConv}_{3\times3}\Big(\mathrm{Expand}_{H,W}(z)\Big) \tag{13}$$

Softmax normalization is subsequently performed along the channel dimension. The channel weight at each spatial position is defined as

$$\alpha_{b,c,h,w} = \frac{\exp(q_{b,c,h,w})}{\sum_{j=1}^{C_f} \exp\Big(q_{b,j,h,w}\Big)} \tag{14}$$

Since Softmax normalizes the channel weights at each spatial location such that their sum equals 1, the channel weights obtained by Softmax are further multiplied by the number of channels C_f of the fused feature, thereby maintaining the average weight at approximately 1. The resulting weights are then multiplied element-wise with the original fused feature:

$$F_a = (C_f \alpha) \odot F_c \tag{15}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $\mathrm{Expand}(\cdot)$ denotes broadcasting along the spatial dimensions, and $\odot$ denotes element-wise multiplication. Softmax is applied along the channel dimension. Therefore, this process performs feature recalibration according to the relative responses among different channels, enabling the network to adaptively adjust the contributions of the multi-scale and high-frequency features according to the current input.

Finally, the recalibrated feature is processed by a CBS module to perform channel fusion and nonlinear transformation. The CBS module consists of a convolution, Batch Normalization, and a SiLU activation function in sequence, and maps the fused feature to C_{out} channels. Since MSEEM is used to replace a convolutional layer responsible for channel transformation in the backbone network, the input and output channel numbers satisfy $C_{\mathrm{in}} \neq C_{\mathrm{out}}$. Therefore, a 1×1 convolution is employed in the residual branch for channel mapping, and the same stride as that of the main branch is used to align the spatial dimensions. The final output is expressed as

$$Y = \mathrm{SiLU}\Big[\mathrm{BN}\Big(\mathrm{Conv}_{k\times k,s,p}^{C_f \to C_{\mathrm{out}}}(F_a)\Big)\Big] + \mathrm{Conv}_{1\times1,s}^{C_{\mathrm{in}} \to C_{\mathrm{out}}}(X) \tag{16}$$

where k, s, and p denote the kernel size, stride, and padding of the convolution operation in the CBS module, respectively. Through the residual connection, the original input feature participates in subsequent feature extraction together with the enhanced multi-scale feature, allowing effective shallow information to be retained while channel transformation is performed.

In summary, MSEEM enhances local details and contextual information using depthwise convolutions and average pooling at different scales. The Laplacian branch supplements high-frequency features associated with local structural variations, while the channel recalibration mechanism adaptively adjusts the fused features. On the basis of preserving the original feature propagation pathway of the network, MSEEM strengthens the representation capability of the shallow network for defects with substantial scale variations, small objects, and weak-texture characteristics, thereby providing more informative features for subsequent deep semantic extraction and multi-scale feature fusion.

Therefore, embedding the MSEEM module into the shallow feature extraction stage of the YOLOv11 backbone can strengthen the network's perception of defect features in power equipment. It can also effectively reduce missed detections of defects with scale variations in practical detection tasks.

3.2. C2SCSA Module

In defect detection for substation equipment, the complex equipment structures and various background interferences make it difficult for the algorithm to distinguish effective features from interfering ones. The C2PSA module mainly enhances the representation of spatial position features through PSA attention. It can improve the model's focus on key regions to some extent. However, this module mainly emphasizes spatial positional relationships. It makes limited use of the collaboration between spatial information and channel semantics. As a result, defect features and background interference may still be difficult to separate under cluttered conditions, such as uneven illumination or branch occlusion. To address this issue, this study optimizes the C2PSA module by replacing the original Attention operation in the PSABlock with the Spatial and Channel Synergistic Attention (SCSA) mechanism [24]. The resulting unit is denoted as SCSABlock, which is subsequently used to construct the composite C2SCSA module, as illustrated in Figure 3.

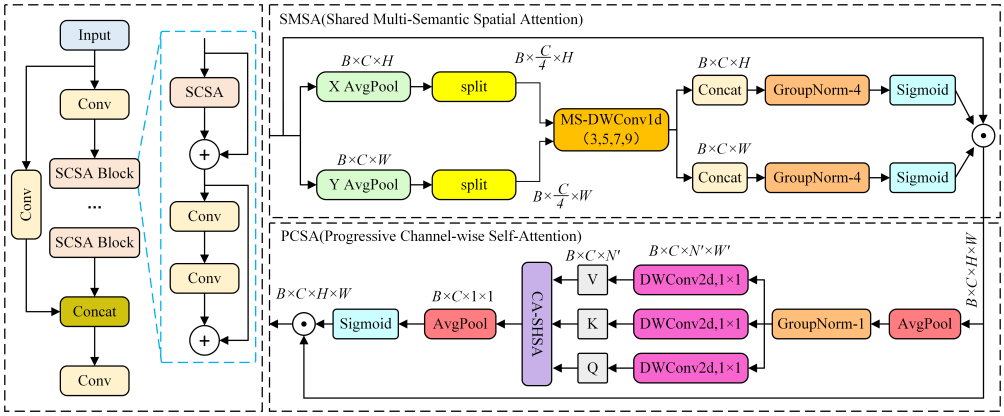


Figure 3. Architecture of the C2SCSA module.

SCSA jointly optimizes the spatial and channel dimensions, enabling the network to simultaneously focus on the spatial location characteristics of defect regions and the semantic responses across different channels. The module consists of Shareable Multi-Semantic Spatial Attention (SMSA) and Progressive Channel-wise Self-Attention (PCSA) connected in series. Spatial attention is first employed to enhance the responses of target regions, after which channel attention further adjusts the importance of different semantic features.

Let the input feature map be defined as

$$X \in R^{B \times C \times H \times W} \tag{17}$$

where B, C, H, and W denote the batch size, number of channels, height, and width of the input feature, respectively.

SMSA first performs average pooling on the input feature along the horizontal and vertical directions to obtain two one-dimensional feature descriptors containing spatial positional information. The channel dimension is then divided into K groups, where $K = 4$. Each group is processed using a one-dimensional depthwise convolution with a different kernel size to capture spatial contextual information over different ranges. The kernel sizes

are set to 3, 5, 7, and 9, respectively. The spatial attention weights along the horizontal and vertical directions are calculated as

$$\mathrm{Attn}_H = \sigma\Big[\,\mathrm{GN}_K\Big(\Phi\big(\,\mathrm{DWConv}_3(X_H^1), \mathrm{DWConv}_5(X_H^2), \mathrm{DWConv}_7(X_H^3), \mathrm{DWConv}_9(X_H^4)\big)\Big)\Big] \tag{18}$$

$$\mathrm{Attn}_W = \sigma\Big[\,\mathrm{GN}_K\Big(\Phi\big(\,\mathrm{DWConv}_3(X_W^1), \mathrm{DWConv}_5(X_W^2), \mathrm{DWConv}_7(X_W^3), \mathrm{DWConv}_9(X_W^4)\big)\Big)\Big] \tag{19}$$

Here, X_H^i and X_W^i denote the horizontal and vertical sub-features of the i-th group, respectively. $\mathrm{DWConv}(\cdot)$ denotes depthwise convolution, $\mathrm{GN}(\cdot)$ denotes group normalization, and $\sigma(\cdot)$ denotes the Sigmoid activation function. By combining convolution kernels of different scales, SMSA can capture both local detailed information and contextual information over a larger spatial range.

After the spatial attention weights are obtained, the horizontal and vertical attention maps are expanded to the corresponding two-dimensional spatial dimensions and applied to the input feature through broadcasting:

$$X_s = X \times \mathrm{Attn}_H \times \mathrm{Attn}_W \tag{20}$$

where $\mathrm{Attn}_H \in R^{B\times C\times H\times 1}$, $\mathrm{Attn}_W \in R^{B\times C\times 1\times W}$, and X_s denotes the spatially enhanced feature.

To further enhance semantic correlations among different channels, PCSA performs adaptive channel recalibration on the spatially enhanced feature. Average pooling is first applied to reduce the spatial resolution and thereby decrease the computational cost of self-attention. The pooled feature is expressed as

$$X_p = \mathrm{Pool}(X_s) \tag{21}$$

where the spatial resolution is reduced from $H \times W$ to $H' \times W'$. Query, key, and value features are subsequently generated from the pooled feature using depthwise convolutions and rearranged into a multi-head representation:

$$Q, K, V \in R^{B\times N_h\times d\times (H'W')} \tag{22}$$

where N_h denotes the number of attention heads and d denotes the channel dimension associated with each attention head.

Unlike spatial attention, PCSA computes correlations along the channel dimension. The dependencies among channels are obtained through matrix multiplication, and the channel-wise self-attention is calculated as

$$X_{\mathrm{attn}} = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{23}$$

where $\sqrt{C}$ is the scaling factor used to prevent excessively large dot-product values from causing unstable training.

Finally, global average pooling is applied to the channel-attention output, followed by the Sigmoid function to generate channel weights. These weights are then applied to the spatially enhanced feature:

$$X_c = X_s \times \sigma(\mathrm{Pool}(X_{\mathrm{attn}})) \tag{24}$$

where X_c denotes the feature after channel recalibration.

Therefore, the overall SCSA operation can be expressed as

$$\mathrm{SCSA}(X) = \mathrm{PCSA}(\mathrm{SMSA}(X)) \tag{25}$$

Through the collaborative optimization of spatial localization and channel semantics, C2SCSA effectively preserves defect-related spatial regions while accurately capturing the corresponding semantic channels. When applied to the end of the backbone network, this module can alleviate the weakening of target features caused by complex background interference in substation scenes. It further improves the detection stability of the network.

3.3. DySample Module

In practical substation equipment inspection, defect targets are often characterized by small scales, irregular boundaries, and rich texture details. Conventional upsampling methods generally rely on fixed interpolation schemes, in which the sampling positions cannot be dynamically adjusted according to the input content. Consequently, problems such as blurred defect boundaries, loss of texture information, and localization deviations may occur.

To improve the recovery of fine-grained defect features, the dynamic sampling upsampling module DySample [25] is introduced in this study. DySample redesigns the upsampling process from the perspective of point sampling. By learning feature-dependent sampling offsets, the sampling positions can be adaptively adjusted according to the input features. The sampling process is illustrated in Figures 4 and 5.

In this study, the LP (low-to-high) mode of DySample is adopted. The upsampling factor is set to $s = 2$, and the number of sampling groups is set to $G = 4$. For an input low-resolution feature,

$$X \in R^{B \times C \times H \times W} \tag{26}$$

where B, C, H, and W denote the batch size, number of channels, height, and width of the input feature map, respectively.

A 1×1 convolution is first used to predict the sampling offset O. According to the implementation, the number of output channels of the offset branch is $2Gs^2$, where G denotes the number of sampling groups and s denotes the upsampling factor. Meanwhile, a dynamic range factor is introduced to constrain the offset range. The final offset is calculated as

$$O' = O \times \sigma(\mathrm{Scope}(X)) + P_0 \tag{27}$$

where $\mathrm{Scope}(X)$ denotes the output of the dynamic range factor branch, $\sigma(\cdot)$ denotes the Sigmoid activation function, and P_0 denotes the initialized sampling positions, which are used to keep the sampling points close to a regular grid distribution during the early stage of training.

After the offset is generated, it is combined with the original regular sampling grid to obtain the dynamic sampling positions:

$$S = P_0 + O' \tag{28}$$

where S denotes the final dynamic sampling grid. To satisfy the input requirements of the grid_sample function, the sampling coordinates are first normalized to the range of $[-1, 1]$. Pixel Shuffle is then used to rearrange the spatial dimensions, mapping the low-resolution offsets to high-resolution sampling positions.

Finally, bilinear resampling is performed according to the dynamic sampling grid using the grid_sample function:

$$X' = \text{GridSample}(X, S) \tag{29}$$

where grid_sample uses bilinear interpolation, with align_corners = False and the boundary padding mode set to border.

Compared with fixed interpolation methods, DySample can dynamically adjust the sampling positions according to the characteristics of target regions. For defect regions with clear boundaries and pronounced structural characteristics in substation equipment, the model can exploit richer contextual information to recover target details. In regions with complex backgrounds, the dynamic sampling mechanism can reduce the introduction of irrelevant information. In addition, the grouped sampling strategy allows different feature groups to learn different sampling patterns, thereby improving the capability of multi-scale feature fusion.

Therefore, DySample can enhance high-resolution feature recovery without significantly increasing the number of model parameters. This contributes to more accurate boundary localization and improved detection of fine-grained texture information for equipment defects in substation scenarios.

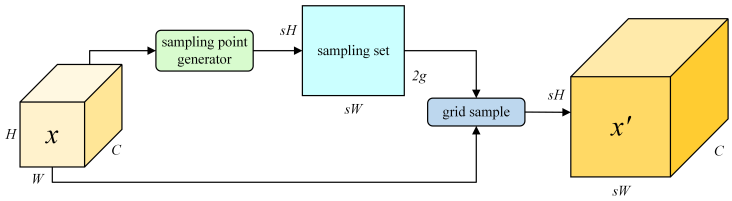


Figure 4. Upsampling process of DySample.

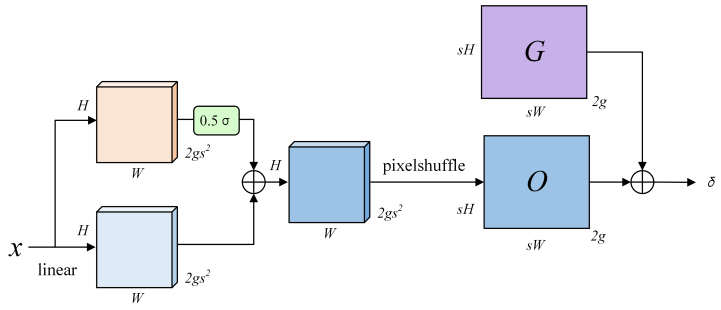


Figure 5. Structure of the dynamic range factor generation method.

4. Experiments

4.1. Experimental Dataset

The dataset used in this study was derived from substation equipment inspection images provided by a national power grid company. The data cover substation scenarios with different voltage levels and equipment layout configurations, including 220 kV and 110 kV outdoor substations. The images were primarily collected by trained operation and maintenance personnel during routine inspection tasks using handheld industrial cameras and unmanned aerial vehicle platforms. Images were captured from different viewing angles and distances, resulting in variations in equipment scale, background structure, and observation perspective. To preserve the engineering authenticity of the dataset, all anomalous samples used in this study were obtained from actual equipment abnormalities or environmental hazards identified during on-site inspections, rather than from artificially simulated defects. The dataset contains a total of 5070 images covering common defects

associated with 10 categories of substation equipment. Each original image contains only one manually verified target instance; therefore, both the number of original images and the number of manually annotated instances are 5070. The number and proportion of annotated instances for each category are presented in Table 1.

Table 1. Number of Annotated Defect Targets by Substation Category.

Defect Category	Number of Defects	Proportion (%)
Blurred meter dial (bj_bpmh)	708	13.96
Damaged meter dial (bj_bpps)	648	12.78
Damaged casing (bj_wkps)	322	6.35
Silicone rubber discoloration (hxq_gjbs)	720	14.20
Abnormal oil level status (ywzt_yfyc)	466	9.19
Ground oil stain (sly_dmyw)	302	5.96
Abnormal cabinet door closure (xmbhyc)	422	8.32
Bird nest (yw_nc)	624	12.31
Suspended foreign object (yw_gkxfw)	548	10.81
Insulator fracture (jyz_pl)	310	6.11

Data annotation was performed using LabelImg (version 1.8.6). To reduce the influence of subjective judgment during manual annotation on label quality, two annotators with relevant domain experience independently annotated all samples according to unified category definitions and annotation rules. For samples on which the two annotators reached agreement, the corresponding annotations were directly adopted. For samples with inconsistent annotations, a third annotator with relevant domain expertise reviewed the results and determined the final labels. For targets with clearly defined boundaries, the bounding boxes were drawn to fit the visible outer contours of the anomalous regions as closely as possible. For irregularly shaped targets, such as oil stains and foreign objects, the minimum enclosing rectangle covering the main visible anomalous region was used. For partially occluded targets whose categories could still be reliably identified, annotation was performed based on the visible region. Samples with severe occlusion, uncertain categories, or insufficient image quality for reliable manual confirmation were excluded from the valid dataset.

To prevent data leakage caused by data augmentation, the dataset was first partitioned at the original-image level before any augmentation was applied. The 5070 original images were divided into training, validation, and test sets at a ratio of 7:1:2, containing 3549, 507, and 1014 images, respectively. During dataset partitioning, images originating from the same inspection sequence or depicting highly correlated acquisition scenes were kept within the same subset whenever possible, thereby reducing the likelihood that the model could exploit fixed backgrounds, equipment appearance, or acquisition viewpoints for memorization. Subject to this constraint, the class composition was also maintained as consistently as possible across the three subsets. After the split was completed, data augmentation, including random-angle rotation, mild noise injection, translation, horizontal flipping, brightness adjustment, and cropping, was applied only to the training set, whereas the validation and test sets were kept unchanged. All augmented samples derived from the training images remained exclusively within the training set, and no augmented derivative was allowed to enter the validation or test sets. This procedure was adopted to avoid evaluation contamination caused by near-duplicate images generated through operations such as rotation, flipping, or cropping. After augmentation, the total number of images actually used in this study was 10,059.

In addition, to further characterize the distribution of targets at different scales across the three data subsets, object instances were categorized as small, medium, or large accord-

ing to their bounding-box areas. Small targets were defined as instances with an area smaller than 32^2 pixels^2, medium targets as instances with an area satisfying $32^2 \leq$ area $< 96^2$ pixels^2, and large targets as instances with an area of at least 96^2 pixels^2. The training set contained 68 small, 814 medium, and 2667 large targets; the validation set contained 8, 127, and 372 targets in the corresponding scale categories; and the test set contained 11, 230, and 773 targets, respectively. Overall, the dataset included 87 small, 1171 medium, and 3812 large target instances.

4.2. Experimental Configuration

All experiments were conducted under the same primary hardware environment. The GPU was an NVIDIA GeForce RTX 3090 with 24 GB of memory. The operating system was Ubuntu 20.04, with Python 3.8, PyTorch 2.0.0, and CUDA 11.8. The training framework was Ultralytics 8.3.0. During training, no pretrained weights were loaded. The model was trained for 200 epochs with a batch size of 32 and an input resolution of 640×640. YOLOv11n and MSDFNet were optimized using SGD. The initial learning rate lr0 was set to 0.01, the final learning rate factor lrf to 0.01, the momentum to 0.937, and the weight decay to 0.0005. A warm-up period of 3 epochs was adopted, with the warm-up momentum set to 0.8 and the warm-up bias learning rate set to 0.1. The random seed was fixed at 0, deterministic training was enabled, automatic mixed precision was enabled, and the early-stopping patience was set to 100 epochs.

The loss-function weights for box regression, classification, and distribution focal loss were set to 7.5, 0.5, and 1.5, respectively. The training-time data augmentation settings included HSV perturbation with hue, saturation, and value factors of 0.015, 0.7, and 0.4, respectively; translation of 0.1; scaling of 0.5; horizontal flipping with a probability of 0.5; and Mosaic augmentation with a probability of 1.0. Mosaic augmentation was disabled during the final 10 epochs, while MixUp was disabled. The nominal batch size was set to 64. With an actual batch size of 32, this corresponded to an effective gradient accumulation factor of approximately 2. Validation was performed throughout training, and the checkpoint with the best validation performance was retained.

4.3. Evaluation Metrics

To comprehensively evaluate the detection performance of the improved algorithm proposed in this paper, Precision (P), Recall (R), and mean Average Precision (mAP) are selected as the main evaluation metrics. These metrics are calculated based on the confusion matrix using statistical terms such as True Positive (TP), False Positive (FP), and False Negative (FN). In addition, the number of parameters (Params) and floating-point operations (FLOPs) are used to evaluate the computational complexity of the model. The corresponding calculation formulas are given as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{30}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{31}$$

$$\mathrm{AP} = \int_0^1 P(R)\,\mathrm{d}R \tag{32}$$

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i, \tag{33}$$

where Precision (P) is used to measure the reliability of the detection results, and Recall (R) is used to evaluate the model's ability to detect real targets. TP denotes the number

of positive samples correctly classified as positive, and a prediction that is successfully matched one-to-one with a ground-truth box is also counted as a true positive (TP). FP denotes the number of negative samples incorrectly classified as positive, and a prediction that fails to match any ground-truth box is also counted as a false positive (FP). FN denotes the number of positive samples incorrectly classified as negative, and a ground-truth box that is not successfully matched by any prediction is also counted as a false negative (FN). N denotes the number of categories. Average Precision (AP) is obtained by integrating the Precision–Recall curve and is used to describe the detection performance for a single category. On this basis, the mean Average Precision (mAP) is calculated by averaging the AP values over all categories. In this study, evaluation is performed at 10 IoU thresholds from 0.50 to 0.95 with an interval of 0.05; i.e., IoU = 0.50:0.05:0.95. mAP@0.5 represents the mean Average Precision over all categories when the IoU threshold is set to 0.5. mAP@0.5:0.95 represents the mean Average Precision over all categories when the IoU threshold ranges from 0.50 to 0.95.

4.4. Ablation Experiments

To verify the effectiveness of the proposed model, ablation experiments were conducted using the same dataset split, training settings, and evaluation protocol. In addition, all FPS measurements were performed on an NVIDIA GeForce RTX 3090 GPU under consistent inference settings to evaluate both detection accuracy and inference efficiency. The experimental results are presented in Table 2.

Table 2. Results of the ablation experiments.

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params/10^6	GFLOPs	FPS
YOLOv11n	83.1	77.2	82.4	53.4	2.58	6.3	150.1
+MSEEM	81.3	82.7	84.9	53.7	2.59	7.0	135.6
+C2SCSA	86.5	79.3	82.9	56.8	2.54	6.3	144.4
+DySample	86.3	79.8	84.9	57.8	2.60	6.3	130.5
+MSEEM+C2SCSA	84.8	80.5	85.6	57.3	2.55	7.0	129.9
+MSEEM+DySample	85.8	82.0	86.6	60.0	2.61	7.1	121.6
+C2SCSA+DySample	86.1	81.4	85.8	58.9	2.55	6.3	125.5
MSDFNet	86.4	84.2	87.3	60.7	2.56	7.0	118.6

The experimental results show that different modules have distinct effects on the evaluation metrics. After introducing the MSEEM module into the shallow feature extraction stage of the backbone, Recall increased from 77.2% to 82.7%, while mAP@0.5 increased from 82.4% to 84.9%, corresponding to improvements of 5.5 and 2.5 percentage points, respectively. This improvement can be attributed to the ability of MSEEM to capture contextual information through multi-scale depthwise convolution and to extract edge and texture characteristics through prior enhancement and adaptive calibration mechanisms, thereby strengthening the network's feature extraction capability for defect targets at different scales. When only C2SCSA was introduced, Precision, Recall, and mAP@0.5:0.95 reached 86.5%, 79.3%, and 56.8%, respectively, with mAP@0.5:0.95 improving by 3.4 percentage points over the baseline. This indicates that the C2SCSA module can effectively mine deep discriminative features through joint calibration in the spatial and channel dimensions. By suppressing complex background interference while enhancing defect-related feature responses, it contributes to a stable improvement in detection accuracy. When only DySample was used to replace the original upsampling method, mAP@0.5 and mAP@0.5:0.95 reached 84.9% and 57.8%, respectively, increasing by 2.5 and 4.4 percentage points over the baseline. This suggests that DySample provides an effective basis for fine-grained feature

recovery in subsequent multi-module combinations by adaptively learning the offsets of sampling points.

The pairwise combination experiments further demonstrate the complementarity among the modules. The mAP@0.5 and mAP@0.5:0.95 values of MSEEM+C2SCSA were 85.6% and 57.3%, respectively; those of MSEEM+DySample further increased to 86.6% and 60.0%; and C2SCSA+DySample achieved 85.8% and 58.9%. Among the two-module combinations, MSEEM+DySample achieved the highest mAP@0.5 and mAP@0.5:0.95, indicating that multi-scale feature enhancement and dynamic sampling exhibit good complementarity for the current task.

When the three modules were used together, the final MSDFNet achieved a Precision of 86.4%, a Recall of 84.2%, an mAP@0.5 of 87.3%, and an mAP@0.5:0.95 of 60.7%. Compared with the YOLOv11n baseline, these metrics were improved by 3.3, 7.0, 4.9, and 7.3 percentage points, respectively. The model contained 2.56×10^6 parameters, while the computational cost increased from 6.3 GFLOPs to 7.0 GFLOPs. These results indicate that the joint use of the three modules improves the overall detection performance without significantly increasing the parameter scale.

In terms of inference efficiency, MSDFNet achieved an inference speed of 118.6 FPS, compared with 150.1 FPS for the YOLOv11n baseline. This shows that the improved model incurs a certain reduction in inference speed but still maintains a relatively high frame rate. The improvement in detection accuracy does not lead to an excessive decrease in FPS, resulting in a favorable balance between detection accuracy and inference speed.

To further verify the rationality of selecting the three modules, replacement experiments were conducted for multi-scale feature enhancement modules, attention modules, and upsampling modules under the same baseline and training conditions. The results are shown in Tables 3–5.

Table 3. Comparison of MSEEM with other feature enhancement modules.

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv11n	83.1	77.2	82.4	53.4
+RFCAConv	84.6	74.6	82.4	54.5
+RFAConv	85.3	74.2	82.9	53.7
+MSEEM	81.3	82.7	84.9	53.7

In the comparison of multi-scale feature enhancement modules, MSEEM achieved a Recall of 82.7% and an mAP@0.5 of 84.9%, both of which were higher than those of RFCAConv and RFAConv [26]. However, its Precision was 81.3%, lower than the 84.6% of RFCAConv and the 85.3% of RFAConv, and its mAP@0.5:0.95 was also not higher than that of RFCAConv. These results indicate that the main advantage of MSEEM lies in improving the detection capability for defect targets rather than simply pursuing higher Precision. In substation equipment defect detection scenarios, this performance characteristic is more consistent with the design objective of reducing missed detections in this study. Therefore, MSEEM was ultimately selected as the multi-scale feature enhancement module.

Table 4. Comparison of C2SCSA with other attention modules.

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv11n	83.1	77.2	82.4	53.4
+SE	87.3	75.1	82.5	55.3
+CA	83.1	77.3	82.0	54.6
+EMA	83.9	78.2	82.3	54.6
+C2SCSA	86.5	79.3	82.9	56.8

In the comparison of attention modules, C2SCSA achieved a Recall of 79.3%, an mAP@0.5 of 82.9%, and an mAP@0.5:0.95 of 56.8%, all of which were the highest among the compared attention modules. The Precision of SE [27] was 87.3%, slightly higher than the 86.5% achieved by C2SCSA. Therefore, the main advantages of C2SCSA lie in its higher recall and overall localization accuracy, which are better aligned with the performance requirements of the current task.

Table 5. Comparison of DySample with other upsampling modules.

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv11n	83.1	77.2	82.4	53.4
+CARAFE	89.7	77.1	83.9	57.8
+EUCB	84.8	76.2	83.0	56.4
+DySample	86.3	79.8	84.9	57.8

In the comparison of upsampling modules, CARAFE [28] achieved the highest Precision of 89.7%, whereas DySample achieved a Recall of 79.8% and an mAP@0.5 of 84.9%, both of which were the highest among the compared methods. DySample and CARAFE achieved the same mAP@0.5:0.95 of 57.8%. These results indicate that, under the current network architecture and dataset conditions, DySample provides a more balanced performance between detection capability and localization accuracy. Therefore, it was adopted in the feature upsampling process of MSDFNet.

To further analyze the model's adaptability to defect targets at different scales, the test set was grouped according to target area, and APsmall, APmedium, and APlarge were calculated separately. Specifically, small targets correspond to an area smaller than 32^2, medium targets correspond to $32^2 \leq$ area $< 96^2$, and large targets correspond to an area greater than or equal to 96^2. The experimental results are presented in Table 6.

Table 6. AP results at different object scales.

Model	AP_S (%)	AP_M (%)	AP_L (%)
YOLOv11n	48.34	45.37	53.79
MSDFNet	68.33	51.21	61.58

Compared with YOLOv11n, the AP_S of MSDFNet increased from 48.34% to 68.33%, corresponding to an improvement of 19.99 percentage points; AP_M increased from 45.37% to 51.21%, an improvement of 5.84 percentage points; and AP_L increased from 53.79% to 61.58%, an improvement of 7.79 percentage points. The improvement was most pronounced for small objects, indicating that the proposed multi-scale feature modeling and feature fusion strategies are particularly beneficial for detecting small-scale defects, while also providing positive gains for medium- and large-scale objects.

To evaluate the influence of training randomness on the experimental results, YOLOv11n, +MSEEM, +MSEEM+C2SCSA, and the final MSDFNet were each trained three times independently under the same experimental settings. The results are reported as the mean $\pm$ standard deviation, as shown in Table 7.

In the repeated experiments, MSDFNet achieved an mAP@0.5 of $87.47 \pm 0.15\%$ and an mAP@0.5:0.95 of $60.87 \pm 0.15\%$, with relatively small overall variation. The corresponding results for YOLOv11n were $82.63 \pm 0.21\%$ and $53.60 \pm 0.20\%$, respectively. These results indicate that the performance fluctuations of the key models were relatively limited under the current repeated-run conditions, demonstrating the stability of the proposed improved model.

Table 7. Results of three repeated training runs for selected key models (mean ± standard deviation).

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv11n	83.37 ± 0.23	77.40 ± 0.20	82.63 ± 0.21	53.60 ± 0.20
+MSEEM	81.57 ± 0.25	83.10 ± 0.36	85.07 ± 0.15	53.87 ± 0.15
+MSEEM+C2SCSA	85.50 ± 0.82	80.00 ± 1.04	85.33 ± 0.25	57.77 ± 0.57
MSDFNet	86.57 ± 0.15	84.33 ± 0.15	87.47 ± 0.15	60.87 ± 0.15

As shown by the class-wise results in Table 8, MSDFNet achieves varying degrees of improvement for most defect categories. In particular, the AP@0.5 values for suspended foreign objects, abnormal cabinet door closure, bird nests, and insulator fracture increase by 12.0, 10.1, 8.9, and 8.8 percentage points, respectively. The AP@0.5 for blurred meter dials increases by 5.2 percentage points, while that for silicone rubber discoloration increases by 3.0 percentage points. Meanwhile, slight decreases are observed for damaged meter dials and abnormal oil level status, by 0.5 and 0.4 percentage points, respectively, whereas the result for damaged casing remains unchanged at 99.5%. The overall advantage of the proposed method mainly arises from improvements across most defect categories, particularly several categories for which the baseline model exhibits relatively lower performance. Together with the increase in overall mAP@0.5 from 82.4% to 87.3%, these results indicate that MSDFNet achieves better overall detection performance for substation defect detection, although performance differences still remain among individual defect categories.

Table 8. Comparison of AP@0.5 for different defect categories.

Defect Category	YOLOv11n (%)	MSDFNet (%)	Change (p.p.)
Blurred meter dial (bj_bpmh)	91.9	97.1	+5.2
Damaged meter dial (bj_bpps)	92.7	92.2	−0.5
Damaged casing (bj_wkps)	99.5	99.5	0.0
Silicone rubber discoloration (hxq_gjbs)	96.2	99.2	+3.0
Insulator fracture (jyz_pl)	56.2	65.0	+8.8
Ground oil stain (sly_dmyw)	87.2	88.7	+1.5
Abnormal cabinet door closure (xmbhyc)	76.5	86.6	+10.1
Suspended foreign object (yw_gkxfw)	57.9	69.9	+12.0
Bird nest (yw_nc)	77.8	86.7	+8.9
Abnormal oil level status (ywzt_yfyc)	87.8	87.4	−0.4

4.5. Comparative Experiments

To verify the effectiveness of the proposed method, comparative experiments were conducted with several mainstream object detection models under the same dataset split, input resolution (640 × 640), test-set evaluation procedure, and primary hardware environment. The compared models included YOLOv5n, YOLOv6n [29], YOLOv8n, YOLOv9t [30], YOLOv10n [31], YOLOv11n, YOLOv12n [32], YOLOv13n [33], and YOLO26n [34]. In addition, the Transformer-based detector RT-DETR-L [35] was included for comparison. For RT-DETR-L, AdamW was used as the optimizer with an initial learning rate of 0.0001, whereas the YOLO-series models adopted the same main training settings as YOLOv11n. The detection results of the different models are presented in Table 9.

Table 9. Comparison of experimental results.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params/10^6	GFLOPs	FPS
YOLOv5n	75.1	70.4	74.3	45.2	2.18	5.8	153.1
YOLOv6n	66.0	60.0	63.2	37.0	4.16	11.5	228.6
YOLOv8n	83.9	73.6	79.4	50.1	2.69	6.8	190.3
YOLOv9t	83.9	74.1	80.9	52.1	1.73	6.4	75.3
YOLOv10n	79.9	67.3	75.5	49.2	2.70	8.2	123.1
YOLOv11n	83.1	77.2	82.4	53.4	2.58	6.3	150.1
YOLOv12n	73.1	73.4	76.6	47.8	2.56	6.3	93.2
YOLOv13n	83.0	74.0	79.6	51.1	2.45	6.2	57.2
YOLO26n	87.8	72.2	82.0	52.4	2.38	5.2	130.3
RT-DETR-L	79.4	69.9	72.9	47.4	32.00	103.5	14.8
MSDFNet	86.4	84.2	87.3	60.7	2.56	7.0	118.6

As shown in Table 9, clear differences can be observed among the detectors in terms of detection accuracy and model complexity. YOLOv5n and YOLOv6n achieve mAP@0.5 values of 74.3% and 63.2%, respectively. YOLOv8n, YOLOv9t, and YOLOv10n show improved overall performance; however, their mAP@0.5:0.95 values are 50.1%, 52.1%, and 49.2%, respectively, which are still lower than the 53.4% achieved by YOLOv11n. With 2.58×10^6 parameters and 6.3 GFLOPs, YOLOv11n achieves a Precision of 83.1%, a Recall of 77.2%, an mAP@0.5 of 82.4%, and an mAP@0.5:0.95 of 53.4%, demonstrating a favorable balance between detection performance and model size.

For YOLOv12n, the Precision, Recall, mAP@0.5, and mAP@0.5:0.95 are 73.1%, 73.4%, 76.6%, and 47.8%, respectively. The corresponding values for YOLOv13n are 83.0%, 74.0%, 79.6%, and 51.1%, while YOLO26n achieves 87.8%, 72.2%, 82.0%, and 52.4%, respectively. Among these models, YOLO26n attains a Precision of 87.8% with 2.38×10^6 parameters and 5.2 GFLOPs, indicating an advantage in terms of precision and model complexity. In comparison, MSDFNet achieves a Precision of 86.4%, which is 1.4 percentage points lower than that of YOLO26n. However, its Recall, mAP@0.5, and mAP@0.5:0.95 reach 84.2%, 87.3%, and 60.7%, respectively, exceeding those of YOLO26n by 12.0, 5.3, and 8.3 percentage points. Compared with YOLOv12n and YOLOv13n, MSDFNet also achieves higher Recall and higher values for both mAP metrics.

For RT-DETR-L, the Precision, Recall, mAP@0.5, and mAP@0.5:0.95 are 79.4%, 69.9%, 72.9%, and 47.4%, respectively, with 32.00×10^6 parameters and a computational cost of 103.5 GFLOPs. In comparison, MSDFNet contains 2.56×10^6 parameters and requires 7.0 GFLOPs, while achieving higher Precision, Recall, and mAP values on the dataset used in this study.

Overall, MSDFNet achieves a Precision of 86.4%, a Recall of 84.2%, an mAP@0.5 of 87.3%, and an mAP@0.5:0.95 of 60.7%. Compared with the YOLOv11n baseline, these four metrics are improved by 3.3, 7.0, 4.9, and 7.3 percentage points, respectively. Meanwhile, the number of parameters is slightly reduced from 2.58×10^6 to 2.56×10^6, while the computational cost increases only from 6.3 to 7.0 GFLOPs. In terms of inference efficiency, the FPS decreases from 150.1 for YOLOv11n to 118.6 for MSDFNet. Although the proposed improved model incurs a certain reduction in inference speed, MSDFNet still maintains reasonable computational complexity and inference efficiency while effectively improving detection accuracy, demonstrating a more balanced performance for substation defect detection.

4.6. Visualization Results and Analysis

Recent interpretable manufacturing frameworks have combined prediction, feature attribution, and independent domain-based corroboration to distinguish meaningful learned

relationships from black-box correlations [36]. To further illustrate the improvement in detection performance achieved by the proposed method, Gradient-weighted Class Activation Mapping (Grad-CAM) [37] is employed to compare the differences in feature-attention regions before and after the model improvement. The heatmaps generated by Grad-CAM visualize the regions on which the model focuses, while the intensity and spatial distribution of the activations reflect the model's response to target-related features, as shown in Figure 6. The visualization results show that the activation regions of the original model are relatively dispersed, with noticeable responses appearing on surrounding equipment structures and background areas, indicating that the model does not always focus accurately on the defect regions. In contrast, the improved model exhibits a more concentrated response, with stronger activations located primarily around the defect targets, enabling more accurate identification of the target regions. These results indicate that MSDFNet has a stronger ability to extract and integrate defect-related features while reducing interference from irrelevant background information. This comparison further supports the advantage of the improved model in enhancing target localization accuracy.

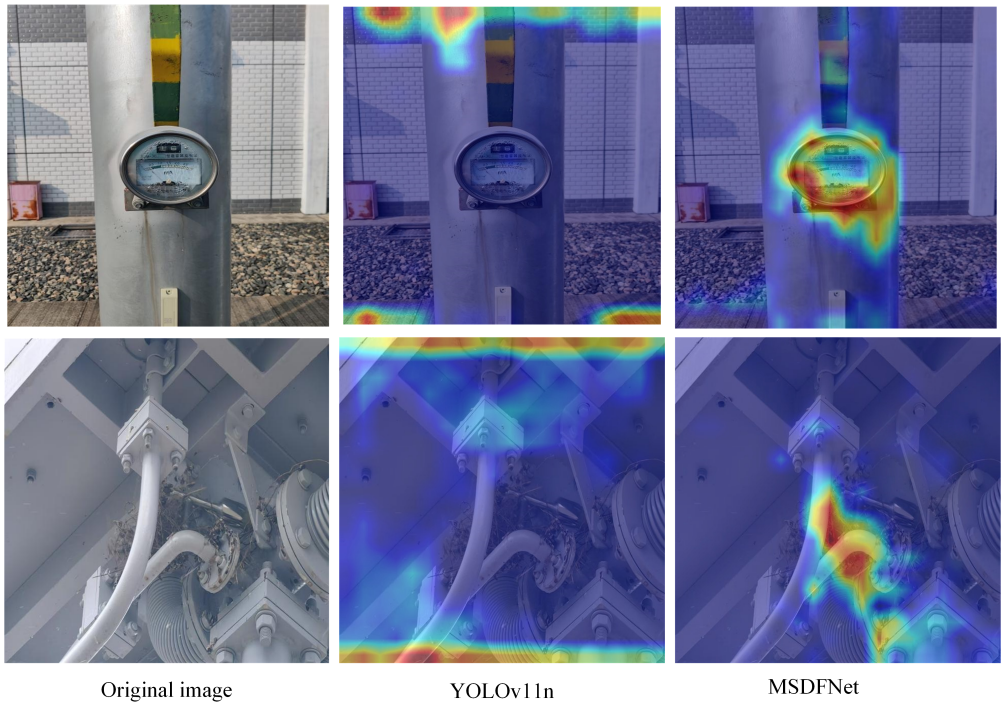


Figure 6. Grad-CAM heat map visualization. (Warmer colors indicate higher activation intensity, whereas cooler colors indicate lower activation intensity.)

To provide an intuitive comparison of the detection performance of YOLOv11n and the proposed MSDFNet under different scenarios, the selected samples cover representative cases involving small-scale defects, complex equipment backgrounds, and low-contrast targets. Several failure cases are also retained to avoid evaluating the models solely on favorable samples. As shown in Figure 7, although YOLOv11n is able to detect the main targets, variations in object scale and interference from complex backgrounds still lead to false detections, missed detections, and inaccurate localization in some scenes. In comparison, MSDFNet can localize defect regions more accurately in practical scenarios and suppress background interference more effectively. Its predicted bounding boxes show a closer correspondence with the target contours, while the confidence scores are generally more stable.

However, the improvements introduced in MSDFNet do not imply that it can achieve higher detection confidence or more accurate localization for all samples. In some scenarios

where defect characteristics are highly similar to the background, missed detections and relatively low confidence scores may still occur.



Figure 7. Visualization results of different detection methods.

For the failure cases shown in Figure 8, although MSDFNet still exhibits the above detection errors, its responses to effective defect features and its localization of target regions are generally improved compared with YOLOv11n, and some false detections and missed detections caused by background interference are alleviated. These results indicate that the proposed improvements can enhance the model's representation capability for defect features in complex scenarios to a certain extent. Therefore, further improving the extraction of weak defect features under extremely complex conditions and reducing the influence of background similarity on detection results remain important issues for future research.

4.7. Comparative Experiments on Generalization Performance

To further evaluate the applicability of the proposed method across different data sources and to reduce the impact of restricted access to the private dataset on experimental verifiability, additional validation was conducted on a publicly available substation equipment defect dataset [38]. The dataset contains 17 categories, with 17,492 images for training, 738 images for validation, and 728 images for testing, resulting in a total of 18,958 images. All compared models were trained and evaluated using the same input size, training strategy, and evaluation metrics as those adopted in the main experiments to ensure a fair comparison. The experimental results are presented in Table 10.

The results show that YOLOv11n achieved a Precision of 75.0%, Recall of 60.1%, mAP@0.5 of 66.3%, and mAP@0.5:0.95 of 41.7% on the public dataset, whereas MSDFNet obtained corresponding values of 74.7%, 64.9%, 69.4%, and 44.1%, respectively. Compared with the baseline model, MSDFNet improved Recall, mAP@0.5, and mAP@0.5:0.95 by 4.8, 3.1, and 2.4 percentage points, respectively, while Precision decreased slightly by 0.3 percentage points. These results indicate that the proposed architecture retains its effectiveness on substation equipment data collected from a different source, particularly in terms of improved target detection capability and localization accuracy, as reflected by the gains in Recall and mAP under stricter IoU thresholds.

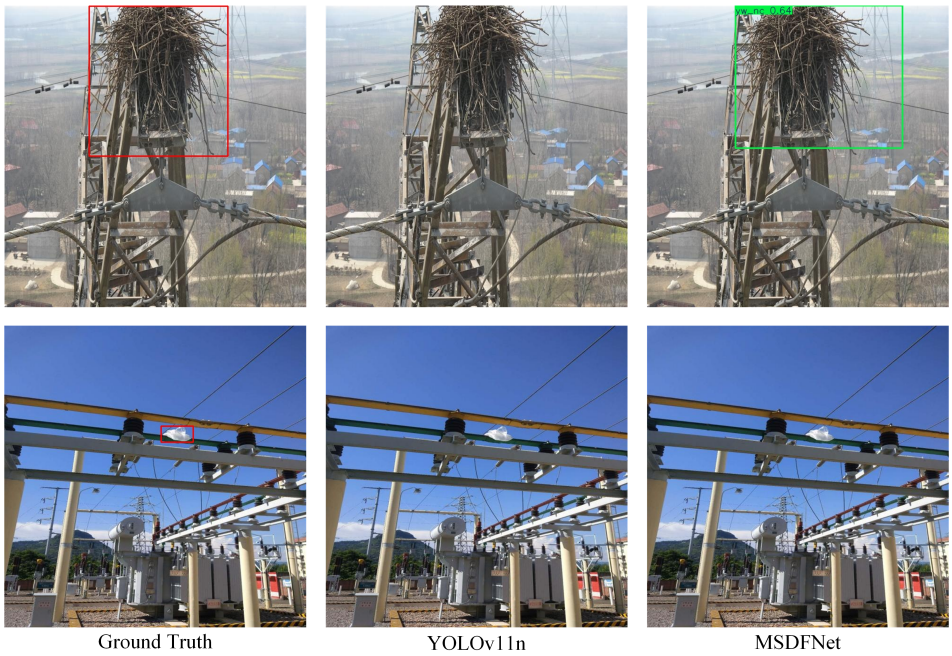


Figure 8. Visualization of failure cases.

Table 10. Comparison of YOLOv11n and MSDFNet on the public dataset.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params/10^6	GFLOPs
YOLOv11n	75.0	60.1	66.3	41.7	2.58	6.3
MSDFNet	74.7	64.9	69.4	44.1	2.56	7.0

5. Discussion and Conclusions

This paper proposes a defect detection method for substation equipment based on multi-scale feature perception. The method is designed to address several challenges in this task, including scale variations among defect targets, interference from complex backgrounds, and insufficient recovery of fine-grained details. First, the backbone network is optimized by designing the MSEEM module. This module significantly enhances the model's perception of defects at different scales, thereby improving detection accuracy. Second, the network structure is further improved by introducing the C2SCSA module. This module strengthens the feature extraction ability of the model under complex background conditions. Finally, the DySample module is incorporated to recover detailed information through adaptive sampling. This improves the model's ability to detect fine-textured defect features. The experimental results show that, compared with the baseline model and mainstream defect detection methods, the improved YOLOv11n network achieves a more balanced performance improvement. It can better meet the practical requirements of defect detection for substation equipment.

Nevertheless, this study still has certain limitations. The private dataset mainly contains pre-screened abnormal samples, while normal equipment images and hard-negative cases are relatively limited. Therefore, the reported results mainly reflect performance on defect-oriented candidate images rather than long-term false-alarm behavior in field inspection. Although additional experiments on a public dataset reduce dependence on a single data source, further validation is still needed across different regions, seasons, equipment types, imaging systems, and challenging conditions such as poor illumination, occlusion, and adverse weather.

Future work will expand the diversity of test data, include more normal and hard-negative samples, and conduct cross-domain and robustness evaluations under broader operating conditions. Repeated multi-seed experiments and statistical analysis will also be introduced. In addition, deployment tests on UAVs, inspection robots, and edge devices will be conducted to evaluate detection accuracy, false-alarm rate, inference latency, and resource consumption in practical inspection scenarios.

Author Contributions: Conceptualization, C.D. and J.H.; methodology, C.D.; software, J.H.; validation, C.D., J.H. and W.G.; formal analysis, C.D.; investigation, J.H. and W.G.; resources, C.D.; data curation, C.D.; writing—original draft preparation, C.D. and J.H.; writing—review and editing, C.D., J.H., W.G. and B.C.; visualization, B.C.; supervision, W.G. and B.C.; project administration, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset presented in this study is not publicly available due to project confidentiality requirements and related restrictions. Access to the dataset may be considered upon request. Researchers who wish to apply for access should contact the corresponding author.

Acknowledgments: The authors would like to thank Weiyi Li for her assistance with language polishing and the preparation of figures and visualizations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Butt, O.M.; Zulqarnain, M.; Butt, T.M. Recent Advancement in Smart Grid Technology: Future Prospects in the Electrical Power Network. *Ain Shams Eng. J.* **2021**, *12*, 687–695. [CrossRef]
2. Luo, L.; Ma, R.; Li, Y.; Yang, F.; Qiu, Z. Image Recognition Technology with Its Application in Defect Detection and Diagnosis Analysis of Substation Equipment. *Sci. Program.* **2021**, *2021*, 2021344. [CrossRef]
3. Cui, Y.; Huang, X.; Zhang, X.; Ye, J.; Zhong, L. A Defects Detection System for Substation Based on YOLOX. In Proceedings of the 2022 IEEE 5th International Electrical and Energy Conference (CIEEC), Nanjing, China, 27–29 May 2022; pp. 4703–4707.
4. Lei, T.; Lv, F.; Liu, J.; Zhang, L.; Zhou, T.; Sun, G. Research on Fault Detection Algorithm of Electrical Equipment Based on Neural Network. *Math. Probl. Eng.* **2022**, *2022*, 9015796. [CrossRef]
5. Lu, S.; Zhang, Y.; Su, J. Mobile Robot for Power Substation Inspection: A Survey. *IEEE/CAA J. Autom. Sin.* **2017**, *4*, 830–847. [CrossRef]
6. Huang, J.; Wang, J.; Tan, Y.; Wu, D.; Cao, Y. An Automatic Analog Instrument Reading System Using Computer Vision and Inspection Robot. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 6322–6335. [CrossRef]
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
8. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.

11. Chen, Y.; Deng, C.; Sun, Q.; Wu, Z.; Zou, L.; Zhang, G.; Li, W. Lightweight Detection Methods for Insulator Self-Explosion Defects. *Sensors* **2024**, *24*, 290. [CrossRef] [PubMed]
12. Xin, M.; Xu, C.; Gao, J.; Wang, Y.; Wang, B. High-Precision Recognition Algorithm for Equipment Defects Based on Mask R-CNN Algorithm Framework in Power System. *Processes* **2024**, *12*, 2940. [CrossRef]
13. Liu, L.; Han, R.; Han, Y.; Qi, D.; Yan, Y. Application of an Improved Faster R-CNN Object Detection Method in the Detection of Suspended Foreign Matters in Substation. *Electr. Meas. Instrum.* **2021**, *58*, 142–146. [CrossRef]
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision–ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
16. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
17. Zheng, H.; Sun, Y.; Liu, X.; Djike, C.L.T.; Li, J.; Liu, Y.; Ma, J.; Xu, K.; Zhang, C. Infrared Image Detection of Substation Insulators Using an Improved Fusion Single Shot Multibox Detector. *IEEE Trans. Power Deliv.* **2021**, *36*, 3351–3359. [CrossRef]
18. Dong, H.; Yuan, M.; Wang, S.; Zhang, L.; Bao, W.; Liu, Y.; Hu, Q. PHAM-YOLO: A Parallel Hybrid Attention Mechanism Network for Defect Detection of Meter in Substation. *Sensors* **2023**, *23*, 6052. [CrossRef] [PubMed]
19. Zhang, N.; Yang, G.; Wang, D.; Hu, F.; Yu, H.; Fan, J. A Defect Detection Method for Substation Equipment Based on Image Data Generation and Deep Learning. *IEEE Access* **2024**, *12*, 105042–105054. [CrossRef]
20. Wang, J.; Sun, Y.; Lin, Y.; Zhang, K. Lightweight Substation Equipment Defect Detection Algorithm for Small Targets. *Sensors* **2024**, *24*, 5914. [CrossRef] [PubMed]
21. Meng, Q.; Fu, T.; Li, K.; Huang, L.; Chen, S. Defect Detection Algorithm for Electrical Substation Equipment Based on Improved YOLOv10n. *IEEE Access* **2025**, *13*, 91409–91422. [CrossRef]
22. Wang, C.; Liu, Z.; Du, J.; Yang, W.; Wang, T. A YOLO-Based Substation Defect Detection Model with Enhanced Feature Extraction and Fusion. *Electr. Power Syst. Res.* **2026**, *252*, 112354. [CrossRef]
23. Jocher, G.; Qiu, J. Ultralytics YOLO11. Ultralytics. 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 30 May 2026).
24. Si, Y.; Xu, H.; Zhu, X.; Zhang, W.; Dong, Y.; Chen, Y.; Li, H. SCSA: Exploring the Synergistic Effects Between Spatial and Channel Attention. *Neurocomputing* **2025**, *634*, 129866. [CrossRef]
25. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to Upsample by Learning to Sample. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 6027–6037.
26. Zhang, X.; Liu, C.; Song, T.; Yang, D.; Ye, Y.; Li, K.; Song, Y. RFAConv: Receptive-Field Attention Convolution for Improving Convolutional Neural Networks. *Pattern Recognit.* **2026**, *176*, 113208. [CrossRef]
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
28. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 29 October–2 November 2019; pp. 3007–3016.
29. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
30. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Proceedings of the Computer Vision–ECCV 2024*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2025; Volume 15089, pp. 1–21.
31. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 107984–108011.
32. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. In Proceedings of the Advances in Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 78433–78457.
33. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
34. Jocher, G.; Qiu, J.; Liu, M.; Lyu, S.; Akyon, F.C.; Kalfaoglu, M.E. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. *arXiv* **2026**, arXiv:2606.03748.
35. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
36. Wu, M.; Yao, Z.; Ye, L.; Ge, J.; Zha, J.; Zhu, Y. A Synergistic Framework Combining Explainable AI and Multiphysical Simulation for Uncovering Process–Structure Relationships in Mask Electrolyte Jet Machining. *Adv. Eng. Inform.* **2026**, *76*, 105089. [CrossRef]

37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
38. substationyolov5. yolov5 Dataset. Roboflow Universe. 2024. Available online: https://universe.roboflow.com/substationyolov5/yolov5-azhje (accessed on 26 August 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.