



ARTICLE

High-Performance and Lightweight Detection Network for Substation Equipment Defect Detection

Hongliang Tian, Xiaoke Liu^{*}, Bolin Song and Chenying Pei

Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education (Northeast Electric Power University), Jilin, China

^{*}Corresponding Author: Xiaoke Liu. Email: 2202400391@neepu.edu.cn

Received: 16 March 2026; Accepted: 26 May 2026; Published: 23 July 2026

ABSTRACT: In the intelligent inspection of power systems, the detection of equipment defects is confronted with problems such as low background discrimination, multi-scale morphological differences, and the difficulty in identifying small targets and fine-grained defects, which makes it hard for existing models to balance detection accuracy and computational efficiency. To address this, this study proposes an improved lightweight detection framework, GRID-YOLO. This framework enhances the semantic discrimination ability of the backbone network for complex defects by introducing a cross-stage hierarchical multi-cognitive spatial attention module (C2MSA), designs an enhanced multi-scale bidirectional feature pyramid network (EMFPN) to achieve efficient alignment and preservation of cross-level features, and builds a hybrid attention reparameterization fusion module (HAFusion) to dynamically filter and enhance heterogeneous features. Additionally, a grouped distribution focal loss head (GDFLHead) is adopted to strengthen the collaborative optimization between classification and regression tasks. In this study, a dedicated dataset for substation equipment defect detection was constructed, and extensive comparative experiments were conducted. The results show that GRID-YOLO achieves an mAP₅₀ of 78.3%, which is 6.1 percentage points higher than the baseline model. The mAP₅₀₋₉₅, Recall, and F1-score reach 52.3%, 73.6%, and 77.6%, respectively, and the inference speed reaches 506 FPS. Compared with various advanced detection models, GRID-YOLO demonstrates strong overall performance and can better meet the actual needs of substation equipment defect detection in the intelligent inspection scenarios of power systems.

KEYWORDS: Substation equipment defect detection; hybrid attention; multi-scale feature fusion; GRID-YOLO

1 Introduction

In the intelligent inspection of power systems, the perception of substation equipment status and defect identification are crucial for ensuring the safe operation of equipment and improving maintenance efficiency [1]. With the continuous expansion of substation equipment scale and the increasingly complex operating environment, inspection objects exhibit characteristics such as a wide variety, dense distribution, and significant dynamic changes in operating status, which impose higher requirements on the accuracy, real-time performance, and stability of defect detection [2]. Compared with traditional inspection methods that rely on human experience, intelligent detection methods can more efficiently obtain abnormal information of equipment and provide technical support for status assessment and maintenance decision-making. However, the shortcomings of manual inspection in terms of safety, efficiency, and consistency remain prominent, especially in large-scale equipment scenarios, making it difficult to meet the actual needs of intelligent operation and maintenance in modern substations [3]. Therefore, researching efficient and

accurate equipment defect detection methods is of significant engineering value for promoting the intelligent development of power system inspections.

Although deep learning provides a new technical path for substation equipment defect detection, this task still faces significant challenges. On the one hand, equipment defects are often manifested as local anomalies, subtle texture changes, or weak boundary features, which are easily overlooked during feature extraction. On the other hand, inspection images often have problems such as poor background distinction, equipment occlusion, and significant scale variations, increasing the difficulty of defect detection and localization. Therefore, how to enhance the ability to identify and locate subtle defects under the resource-constrained application conditions of substations and achieve a balance between detection performance and model efficiency remains a key issue in substation equipment defect detection.

To address the above problems, this paper proposes a lightweight detection framework, GRID-YOLO, for substation equipment defect detection. By collaboratively designing key detection components, it aims to improve the ability to identify subtle defects and balance detection performance with model efficiency. The main contributions of this paper are as follows:

- (1) By enhancing the multi-scale semantic modeling capability of the backbone network and introducing the hierarchical multi-cognitive spatial attention module (C2MSA), the fine-grained discrimination effect of defect targets under conditions of poor background discrimination is improved.
- (2) An enhanced multi-scale bidirectional feature pyramid network (EMFPN) is constructed to address the problems of insufficient cross-layer information interaction and easy loss of detail features in multi-scale object detection, thereby enhancing the alignment and transmission capabilities of features at different levels.
- (3) To enhance the ability to retain effective information during the heterogeneous feature fusion process, a hybrid attention reparameterization fusion module (HAFusion) is designed to improve the effectiveness of cross-layer feature integration.
- (4) In the detection output stage, a grouped distribution focal loss head (GDFLHead) is designed. Through probabilistic bounding box representation and a distribution-focused loss mechanism, the consistency of classification and localization is strengthened, thereby improving the localization accuracy of small-scale and blurred defects.

The remaining sections of this paper are organized as follows: [Section 2](#) introduces related work on substation equipment defect detection; [Section 3](#) details the proposed GRID-YOLO method and its core modules; [Section 4](#) presents the experimental setup, comparison experiments, and ablation experiment results, along with analysis and discussion; [Section 5](#) summarizes the entire paper and looks ahead to future research directions.

2 Related Work

In substation inspection research, early methods mainly relied on manual inspection and traditional image processing techniques. Manual inspection can make judgments based on field experience, but in large-scale inspection tasks, it is easily affected by operator expertise, workload, and environmental conditions, resulting in low efficiency, strong subjectivity, and a high risk of missed detections [4]. To improve detection automation, Jabid and Ahsan [5] proposed an insulator detection and defect classification method based on rotation-invariant local directional patterns, using handcrafted texture features and an SVM classifier to identify insulator regions and analyze their defects. Yu et al. [6] introduced texture and shape priors into an active contour model to enhance the stability of insulator segmentation under complex backgrounds. These traditional methods can be effective when target structures are relatively regular and imaging conditions

are stable. However, their performance depends heavily on handcrafted features, threshold rules, or prior assumptions, and their robustness and generalization ability remain limited under background interference, equipment occlusion, illumination variation, and subtle target abnormalities.

In recent years, deep learning-based object detection models have improved target representation capability in complex scenarios through end-to-end feature learning [7]. Related studies have gradually shifted from handcrafted feature design to feature representation enhancement, multi-scale feature fusion, and lightweight structural optimization, aiming to meet the practical requirements of detection accuracy, robustness, and inference efficiency in substation inspection tasks. For example, Wang et al. [8] combined CNN and ViT to enhance the feature extraction ability of small targets, but their design focus was mainly on small target perception, and the collaborative modeling of multiple types of defects and cross-layer features was still insufficient. Dong et al. [9] introduced parallel hybrid attention to enhance defect perception in complex backgrounds, but mainly focused on attention enhancement, and the adaptability to defects with significant scale changes was still limited. Wang et al. [10] improved the inference efficiency by using a lightweight FasterNet backbone, but the replacement of the backbone might weaken the feature extraction ability of subtle defects in complex backgrounds. Qian et al. [11] improved multi-scale defect detection efficiency by introducing a lightweight feature fusion network into LFF-YOLO, but the method still showed limited adaptability to defects with significant shape and scale variations. You and Zhao [12] enhanced the detection ability of small target insulator defects through multi-scale bidirectional feature fusion, but their scene applicability was relatively concentrated, and there were still limitations in lightweight collaborative detection of multiple types of substation equipment defects. Wang et al. [13] effectively improved the inference efficiency by implementing reparameterization training-inference decoupling through the ResNet-RepConv module, but retaining detailed information of subtle defects during cross-layer feature fusion remained a challenge. Zhou et al. [14] introduced a lightweight shared detail enhancement detection head to improve the detail capture ability of transformer oil leakage, but their detection head still mainly focused on enhancing shared convolutional features, and the optimization of localization quality for defects with ambiguous boundaries was still insufficient. Ji et al. [15] enhanced the multi-scale insulator defect feature fusion ability by introducing ASFF in the detection head, but the optimization of cross-layer detail information retention and consistency in localization quality for small-scale and ambiguous defects was still insufficient. Ding et al. [16] introduced an additional small target detection head in GC-YOLO to improve the detection accuracy of small target defects in complex backgrounds, but the additional detection branch might bring certain computational overhead.

Although the aforementioned methods have made certain progress in the detection of defects in substations or power equipment, most of them still focus on optimizing a single aspect, such as backbone network enhancement, attention model building, feature fusion improvement, detection head adjustment, or lightweight structure reconfiguration, and have not yet formed a systematic solution in terms of diversified defect identification and the coordinated balance of detection accuracy and computational efficiency. The advantages and disadvantages of the existing representative methods are shown in Table 1.

Table 1: Comparison of representative methods for substation equipment defect detection.

Method	Advantage	Limitation
Rule-based image processing methods	Relatively simple to implement and effective in specific scenarios	Depend heavily on handcrafted features and show limited robustness to complex backgrounds and noise interference

(Continued)

Table 1 (continued)

Method	Advantage	Limitation
Handcrafted feature and classifier-based methods	Can combine texture, shape, and other handcrafted features for defect recognition and fault discrimination	Limited generalization ability and insufficient adaptability to complex scenes and diverse defect patterns
EAL-YOLO [8]	Combines CNN and ViT to improve feature representation capability	High computational complexity
PHAM-YOLO [9]	Introduces a parallel hybrid attention mechanism to enhance defect feature perception under complex backgrounds	Limited adaptability to multi-scale defects
YOLO-SS-Large [10]	Introduces the lightweight backbone design of FasterNet to improve inference efficiency	Feature extraction capability remains insufficient in complex backgrounds
LFF-YOLO [11]	Employs a lightweight feature fusion network to improve multi-scale defect detection efficiency	Limited adaptability to defects with significant variations in shape and scale
YOLOLS [13]	Improves inference efficiency through a reparameterization-based lightweight design	Limited preservation of subtle defect information during feature fusion
IALF-YOLO [17]	Introduces an improved S-CBAM attention mechanism to strengthen attention to critical defect features	Limited ability to model global contextual attention

In industrial visual inspection tasks, lightweight neural networks have attracted extensive attention due to their small number of parameters, low computational cost, and high inference efficiency [18]. Existing studies usually reduce model complexity through lightweight backbone networks, depthwise separable convolutions, channel shuffle, structural reparameterization, and lightweight detection heads, thereby improving their applicability in resource-constrained scenarios. For substation equipment defect detection, lightweight design helps improve the response efficiency and practical applicability of intelligent inspection systems [19]. However, model lightweighting is often accompanied by a decline in feature representation capability. Especially under conditions of strong background interference, easily confused subtle defects, and significant target scale variations, problems such as insufficient feature extraction, inadequate cross-layer information preservation, and decreased localization accuracy may still occur. Therefore, how to enhance the recognition capability for subtle defects while maintaining the advantages of lightweight models, and achieve coordinated optimization between detection accuracy and computational efficiency, remains a key issue that needs to be further addressed in current research.

3 The Proposed GRID-YOLO Algorithm

To address the issues of strong background interference, dense small targets, and significant scale differences in defects in substation inspection scenarios, this paper improves upon YOLOv11n and proposes the GRID-YOLO algorithm. After the inspection images are input into the network, the backbone network first extracts the initial features, and C2MSA is used to enhance semantic expression in complex backgrounds. Subsequently, EMFPN performs cross-scale interaction and fusion of multi-layer features to improve the consistency of features at different scales. On this basis, HAFusion further strengthens the integration of cross-layer features to enhance the effectiveness of the fusion representation. Finally, to accurately map the optimized features to detection results, GDFLHead adopts probabilistic bounding box representation and distribution-focused loss mechanisms, strengthening the collaborative optimization of classification and localization tasks under a lightweight architecture, significantly improving the localization accuracy of small-scale and blurred defects. Through the above processing, GRID-YOLO forms an overall optimization process from feature extraction, feature interaction to detection output, improving detection performance while considering model complexity and inference efficiency, providing a feasible solution for intelligent substation inspection in resource-constrained scenarios. The overall structure is shown in Fig. 1.

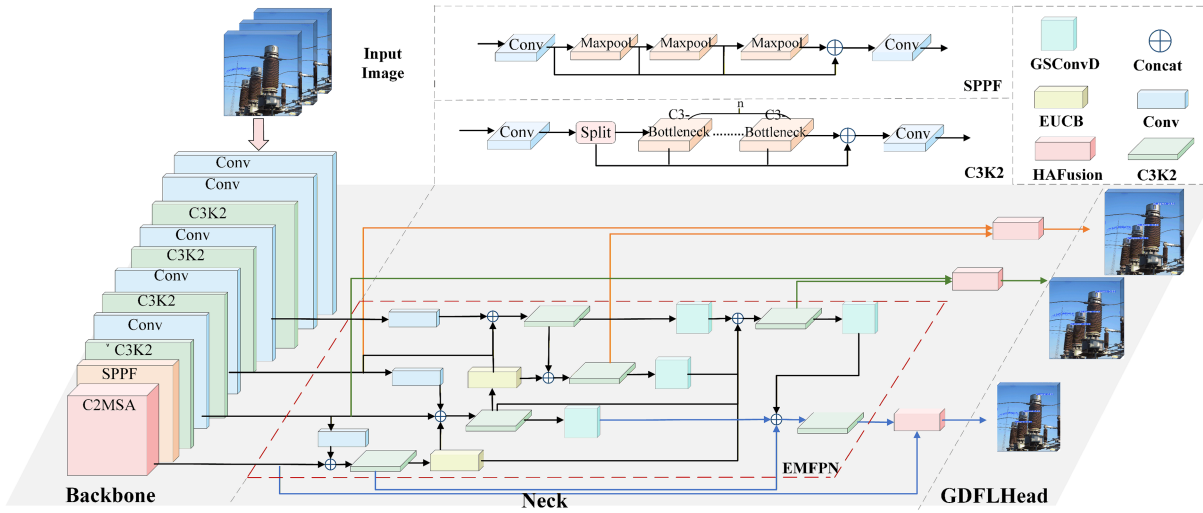


Figure 1: GRID-YOLO algorithm framework.

3.1 C2MSA Module

To address the challenges of semantic confusion between defect targets and complex backgrounds and large scale differences in substation scenarios, existing attention mechanisms (e.g., channel-wise SE [20] and spatial CBAM [21]) often focus on only a single dimension and are difficult to jointly model multi-scale contexts. Therefore, we propose the C2MSA module, whose core is to construct a hierarchical processing flow through a multi-cognitive visual adapter (MCA), successively performing feature recalibration, parallel multi-scale context perception, and spatial-channel collaborative attention screening, thereby achieving adaptive and multi-dimensional enhancement of deep semantic features to improve the model's discrimination ability for complex defects. The overall structure of this module is shown in Fig. 2.

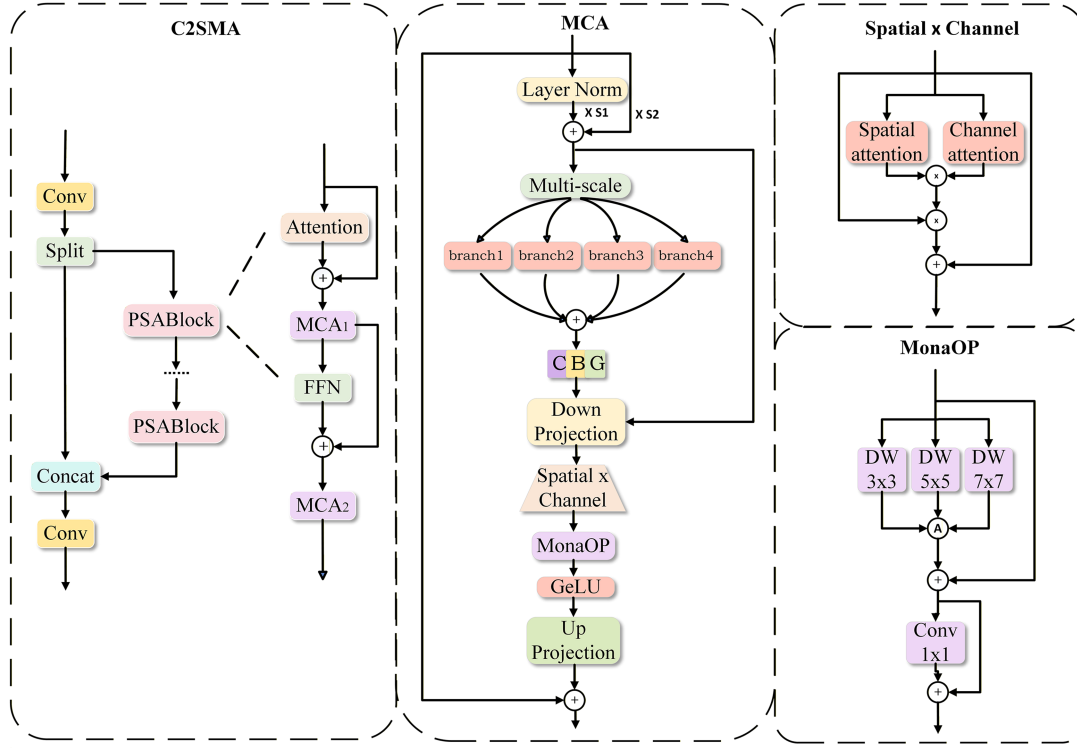


Figure 2: C2MSA module structure diagram.

As the core component of C2MSA, MCA constructs a hierarchical feature enhancement process. Given an input feature $x_0 \in \mathbb{R}^{B \times C \times H \times W}$, where B , C , H , and W denote the batch size, number of channels, height, and width, respectively, the input feature is first recalibrated by layer normalization and learnable channel-wise scaling parameters, as shown in Eq. (1):

$$x_1 = s_1 \otimes \text{LN}(x_0) + s_2 \otimes x_0 \quad (1)$$

where $\text{LN}(\cdot)$ represents the layer normalization operation, s_1 and s_2 are learnable channel scaling parameters. In this paper, $\otimes$ denotes element-wise multiplication, with broadcasting used for dimension alignment when necessary, rather than tensor product. Then, the recalibrated feature x_1 is fed into four parallel branches to capture multi-scale contextual information. Three branches adopt 3×3 convolutions with different dilation rates, while the fourth branch extracts global context through global average pooling followed by 1×1 convolution and upsampling, as shown in Eqs. (2) and (3):

$$B_i = \text{Conv}_{3 \times 3, d=i}(x_1), \quad i \in \{1, 2, 3\} \quad (2)$$

$$B_4 = \text{Up}(\text{Conv}_{1 \times 1}(\text{gap}(x_1))) \quad (3)$$

Here, $\text{Conv}_{3 \times 3, d=i}$ denotes a 3×3 convolution with dilation rate i , $\text{gap}(\cdot)$ denotes global average pooling, and $\text{Up}(\cdot)$ denotes upsampling to the spatial size of x_1 . The outputs of the four branches are concatenated along the channel dimension and fused by a Convolution-Batch Normalization-Gaussian Error Linear Unit (Conv-BN-GeLU) block. A residual connection is introduced to preserve the original feature information, and the fused feature is then compressed by the lower projection layer. The process is shown in Eq. (4):

$$Z = D_l(\text{CBG}(\text{Cat}[B_1, B_2, B_3, B_4]) + x_1) \quad (4)$$

where $\text{Cat}[\cdot]$ denotes channel-wise concatenation, $\text{CBG}(\cdot)$ denotes the Conv-BN-GeLU block, and $D_l(\cdot)$ denotes the lower projection layer. After that, spatial and channel attention maps are generated from Z to collaboratively enhance the feature response. The attention-enhanced feature is defined as shown in Eq. (5):

$$F_{att} = Z \otimes (A_s \otimes A_c) + Z \quad (5)$$

where A_s and A_c denote the spatial and channel attention maps generated by the corresponding attention branches in Fig. 2, respectively. Specifically, the attention weights are obtained through convolutional transformations followed by GeLU and Sigmoid activations. Subsequently, F_{att} is sent to the Multi-aperture Operation (MonaOP) sub-module. As shown in Fig. 2, MonaOP adopts three depthwise convolution branches with kernel sizes of 3×3 , 5×5 and 7×7 to aggregate spatial information under different receptive fields. The multi-aperture aggregation and subsequent 1×1 convolution transformation are formulated as shown in Eq. (6):

$$F_{op} = F_{att} + \frac{1}{3} \sum_{k \in \{3,5,7\}} \text{DW}_{k \times k}(F_{att}), \quad F'_{op} = F_{op} + \text{Conv}_{1 \times 1}(F_{op}) \quad (6)$$

where $\text{DW}_{k \times k}(\cdot)$ denotes the depthwise convolution branch with kernel size $k \times k$, corresponding to the $\text{DW}_{3 \times 3}$, $\text{DW}_{5 \times 5}$, and $\text{DW}_{7 \times 7}$ operations in Fig. 2. $\text{Conv}_{1 \times 1}(\cdot)$ denotes the 1×1 convolution operation used for feature transformation. Finally, the output of MCA is obtained through GeLU activation, upper projection, and residual connection with the original input. The process is shown in Eq. (7):

$$y = x_0 + U^l \left(\mu \left(F'_{op} \right) \right) \quad (7)$$

Here, U^l represents the upper projection layer used to restore the channel dimension, and $\mu(\cdot)$ denotes the GeLU activation function. Through the above hierarchical design, the C2MSA module enhances multi-scale contextual perception, spatial-channel feature screening, and fine-grained feature discrimination, thereby improving the robustness of defect detection in complex substation scenarios.

3.2 EMFPN Feature Pyramid Structure

To balance multi-scale feature fusion capability and model complexity, this paper proposes an Enhanced Multi-scale Bidirectional Feature Pyramid Network (EMFPN). Conventional feature pyramid structures usually rely on dense connections or additional attention modules to enhance cross-layer information interaction, but these designs tend to introduce a considerable number of parameters and computational overhead. Therefore, EMFPN improves the interaction efficiency between deep semantic features and shallow detail features by reconstructing key bidirectional transmission paths and jointly optimizing the upsampling and downsampling processes. During the upsampling stage, an Enhanced Upsampling Convolution Block (EUCB) is designed to address the detail loss of small defect targets while restoring spatial resolution. During the downsampling stage, an Enhanced Downsampling Convolution Block (GSConvD) is proposed to alleviate semantic information loss and background interference during feature compression by integrating multi-scale receptive fields and attention mechanisms. By collaboratively introducing EUCB and GSConvD into the bidirectional fusion path, EMFPN achieves cross-layer feature alignment and complementary fusion with relatively low computational cost, thereby obtaining more discriminative multi-scale feature representations. The core module of EMFPN is shown in Fig. 3.

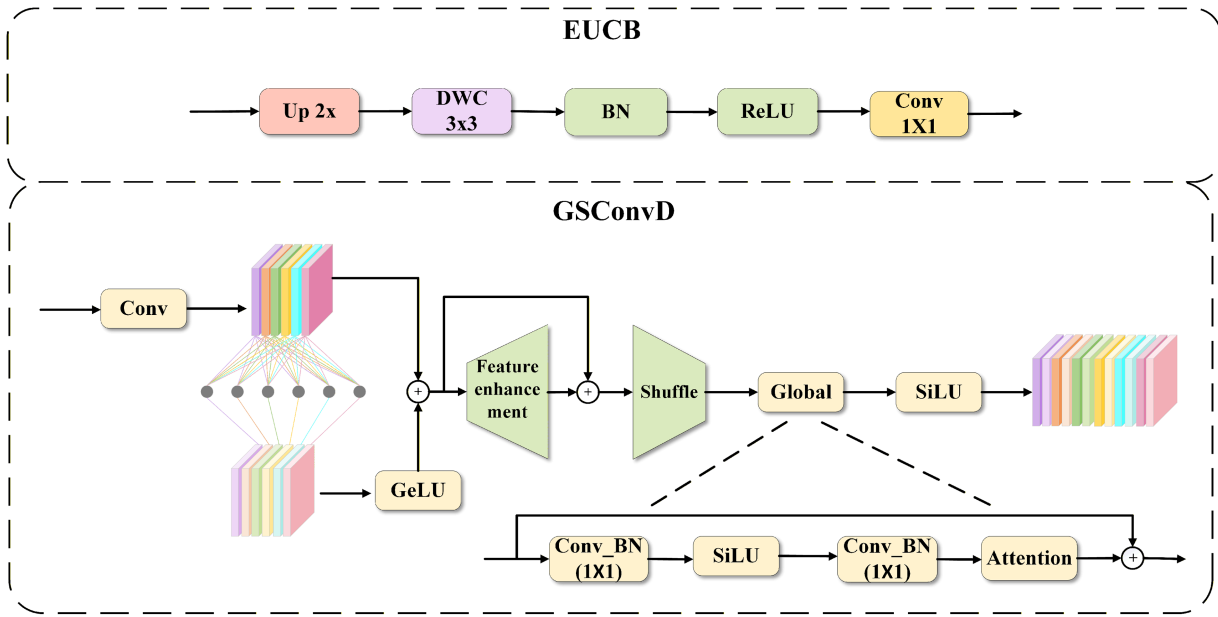


Figure 3: The core module of EMFPN.

3.2.1 EUCB Module

This module aims to restore the resolution of feature maps in a learnable manner and enhance details. This is achieved through a sequential three-step design: First, the feature map is upsampled by a factor of 2 via a standard upsampling layer. Next, the upsampled features are refined through a 3×3 depthwise convolution (DWC), followed by batch normalization (BN) and a Rectified Linear Unit (ReLU) activation. Finally, a 1×1 convolution adjusts the channel dimension to meet the input requirements of the subsequent stage. This design effectively alleviates the problem of detail loss in the traditional upsampling process by introducing learnable convolution operations instead of simple interpolation.

3.2.2 GSConvD Module

The structure of the GSConvD module is shown in Fig. 3. Given an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, where B , C , H , and W denote the batch size, number of channels, height, and width, respectively, GSConvD first applies a convolution operation to obtain the initial feature representation, as shown in Eq. (8):

$$Y_1 = X * W_1 \quad (8)$$

where $*$ denotes the convolution operation, and W_1 denotes the convolution kernel. Then, Y_1 is further transformed by another convolution and GeLU activation, and an extended feature is obtained through residual connection, as expressed in Eq. (9):

$$Y_2 = Y_1 + \mu(Y_1 * W_2) \quad (9)$$

where W_2 denotes the convolution kernel of the second convolution operation, and $\mu(\cdot)$ denotes the GeLU activation function. The original feature Y_1 and the extended feature Y_2 are concatenated along the channel dimension and refined by the feature enhancement block, as defined in Eq. (10):

$$Y_c = \text{Cat}[Y_1, Y_2], \quad Y_f = Y_c + F_e(Y_c) \quad (10)$$

where $\text{Cat}[\cdot]$ denotes channel-wise concatenation, $F_e(\cdot)$ denotes the feature enhancement block and Y_f denotes the enhanced feature. Finally, channel shuffle and global context aggregation are applied to the enhanced feature, followed by SiLU activation to obtain the output feature, as shown in Eq. (11):

$$Y = \text{SiLU} \left(G_c \left(\text{Shuffle} \left(Y_f \right) \right) \right) \quad (11)$$

where $\text{Shuffle}(\cdot)$ denotes the channel shuffle operation, $G_c(\cdot)$ denotes the global context aggregation path, including the Global operation and the subsequent attention refinement, and $\text{SiLU}(\cdot)$ denotes the SiLU activation function. Through the above process, GSConvD enhances feature discrimination during downsampling while maintaining a lightweight structure.

3.3 HAFusion Module

Current multi-scale feature fusion methods often rely on fixed operations such as addition or concatenation to merge cross-layer features. This static approach lacks the ability to dynamically assess feature importance and may introduce noise when fusing deep semantic information with shallow spatial details. To overcome these limitations, we propose the Hybrid Attention Reparameterization Fusion (HAFusion) module. HAFusion employs a dual-branch structure that combines dynamic and static processing pathways. The first branch, Local-Global Attention (LG-Attention), dynamically focuses on semantically salient regions based on feature content. The second branch, RepConv, leverages structural re-parameterization to provide robust, static spatial transformations. Together, these branches complement each other, balancing adaptive feature selection with structured representation enhancement. The architecture of the HAFusion module is depicted in Fig. 4.

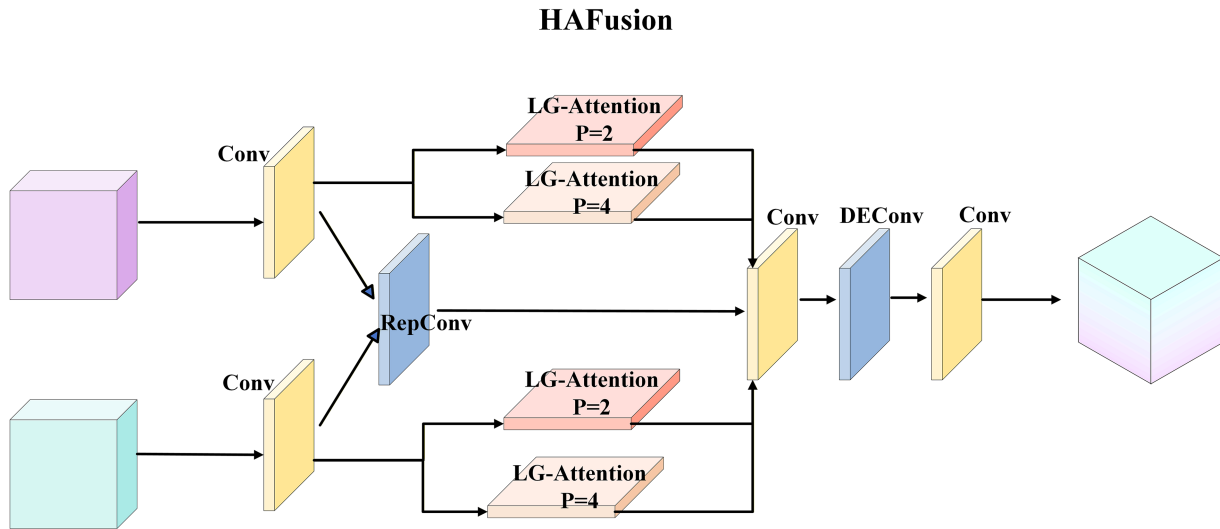


Figure 4: HAFusion module structure diagram.

The HAFusion module aims to fuse two input feature maps $X_1 \in \mathbb{R}^{B \times C_1 \times H \times W}$ and $X_2 \in \mathbb{R}^{B \times C_2 \times H \times W}$ from different levels, where B , H , and W represent the batch size, height, and width, respectively, and C_1 and C_2 are the channel numbers of the two input features. The two inputs are first compressed by 1×1 convolution to obtain $F_1, F_2 \in \mathbb{R}^{B \times C_h \times H \times W}$, where $C_h = C_{\text{out}}/2$. Then, LG-Attention branches with patch sizes $p = 2$ and $p = 4$ generate attention masks $M_{k,p} \in [0, 1]^{B \times C_h \times H \times W}$, where $k \in \{1, 2\}$ denotes the input feature branch and $p \in \{2, 4\}$ denotes the receptive-field scale. These masks are used to dynamically select effective features, which are then fused with the RepConv branch to obtain the output Y .

To stabilize attention learning and suppress redundant background responses, entropy-based and sparse regularization terms are introduced for the attention masks $M_{k,p}$, which directly affect cross-layer feature selection. The entropy-based regularization term prevents the attention masks from prematurely concentrating on only a few channels or spatial regions, as shown in Eq. (12):

$$L_{\text{ent}} = -\frac{1}{4N_m} \sum_{k \in \{1,2\}, p \in \{2,4\}} \sum_i M_{k,p,i} \log(M_{k,p,i} + \varepsilon) \quad (12)$$

where $N_m = B \times C_h \times H \times W$ denotes the number of elements in each attention mask, $M_{k,p,i}$ denotes the i -th element of $M_{k,p}$, and ε is a numerical stability constant. This term is optimized with a negative coefficient in the total loss to maintain mask diversity during training.

The sparse regularization term encourages the masks to focus on defect-related discriminative regions, as defined in Eq. (13):

$$L_{\text{sparse}} = \frac{1}{4N_m} \sum_{k \in \{1,2\}, p \in \{2,4\}} \|M_{k,p}\|_1 \quad (13)$$

where $\|\cdot\|_1$ denotes the L_1 -norm. The above two terms are used as auxiliary regularization constraints for the attention masks during training, so that HAFusion can enhance effective cross-layer features while suppressing redundant background responses.

3.3.1 LG-Attention Module

As the core component of HAFusion, LG-Attention realizes the collaborative modeling of local and global context, and its structure is shown in Fig. 5. Given an input feature $F \in R^{B \times C_h \times H \times W}$, it is first unfolded into non-overlapping local patches with patch size p , as shown in Eq. (14):

$$Z_p = \text{Unfold}(F, p) \in R^{B \times N_p \times p^2 \times C_h} \quad (14)$$

where $p \in \{2, 4\}$, $N_p = HW/p^2$ denotes the number of non-overlapping patches, $\text{Unfold}(\cdot)$ denotes the patch extraction operation, and Z_p denotes the patch feature obtained with patch size p . Then, Z_p is averaged along the channel dimension and mapped into token embeddings through two MLP layers and layer normalization, as expressed in Eq. (15):

$$T_p = \text{MLP}_2(\text{LN}(\text{MLP}_1(\text{Mean}_C(Z_p)))) \quad (15)$$

where, $\text{Mean}_C(\cdot)$ denotes channel-wise averaging, $\text{MLP}_1(\cdot)$ and $\text{MLP}_2(\cdot)$ denote linear mappings, $\text{LN}(\cdot)$ is the layer normalization operation, $T_p \in R^{B \times N_p \times p^2 \times D}$, and D is embedding dimension. To highlight the key information in the local features, the embedded representations are subjected to Softmax normalization and element-wise reweighting is performed, as shown in Eq. (16):

$$A_p = \text{Softmax}(T_p), \quad T_{\text{att}} = T_p \otimes A_p \quad (16)$$

Here, A_p denotes the attention weight, and $\otimes$ denotes element-wise multiplication, with broadcasting used for dimension alignment when necessary. A learnable prompt vector $q_p \in R^D$ is further introduced, and the attention mask is generated by cosine similarity, as shown in Eq. (17):

$$M_p = \text{Clamp}(\cos(T_{\text{att}}, q_p), 0, 1) \quad (17)$$

where, $\cos(\cdot, \cdot)$ represents cosine similarity, $\text{Clamp}(\cdot, 0, 1)$ indicates that the results are restricted within the range of $[0, 1]$. Finally, the mask-fused representation is linearly mapped and restored to the spatial feature format, as shown in Eqs. (18) and (19):

$$T_{\text{proj}} = (T_{\text{att}} \otimes M_p) W_p \quad (18)$$

$$F_p = \text{Interpolate}(\text{Reshape}(T_{\text{proj}}), \text{size} = (H, W)) \quad (19)$$

Among them, W_p denotes the linear projection matrix. In Eq. (19), when M_p is multiplied with T_{att} , broadcasting is used for dimension alignment when necessary. $\text{Reshape}(\cdot)$ restores tokens to the spatial feature map, and $\text{Interpolate}(\cdot)$ resizes the feature to the input resolution. $F_{p=2}$ and $F_{p=4}$ correspond to local-context and global-context enhanced features, respectively, and are used for subsequent fusion.

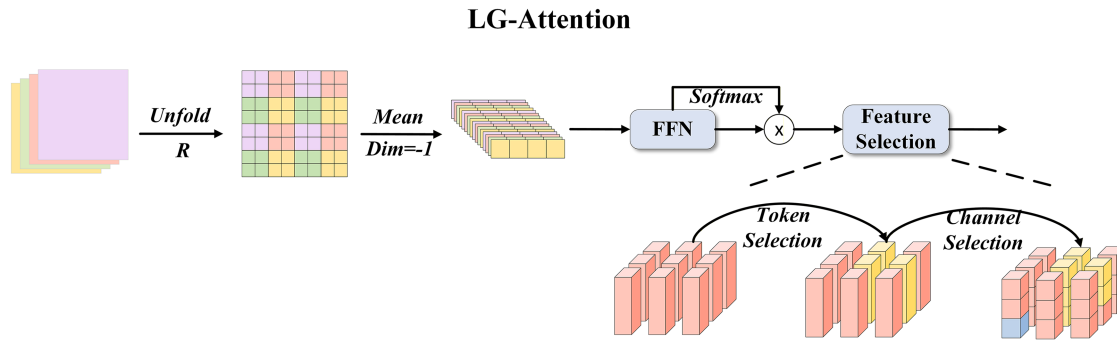


Figure 5: LG-Attention module structure diagram.

3.3.2 RepConv Module

As another core branch of HAFusion, RepConv delivers strong structured representation while maintaining high inference efficiency, which is essential for real-time applications. Unlike conventional convolutional layers, which maintain a fixed structure and a single receptive field across both training and inference, RepConv adopts a decoupled design that separates training from inference. During training, it employs a multi-branch structure consisting of 3×3 and 1×1 convolutions, residual connections, and batch normalization (BN) layers. During inference, these branches are merged via structural reparameterization into a single 3×3 convolution. Fig. 6 illustrates the reparameterization process from two distinct perspectives. Fig. 6a presents the transformation from a multi-branch training architecture to a single-branch inference architecture from a structural viewpoint. Fig. 6b illustrates the equivalent fusion of the corresponding convolution kernels from a parametric viewpoint.

3.4 GDFLHead Module

The existing decoupled detection heads often encounter the problem of weakened feature interaction in the defect detection of substations. Especially when detecting subtle or vaguely visible defects (such as silicone color change, instrument reading deviation), it is difficult to optimize the classification confidence and positioning accuracy simultaneously, resulting in missed detections or positioning drift. To address this issue, this paper proposes GDFLHead. The core innovation of this method lies in modeling the bounding box positioning as a discrete probability distribution and achieving explicit collaborative learning of classification and regression tasks at the feature level through distribution-focused loss.

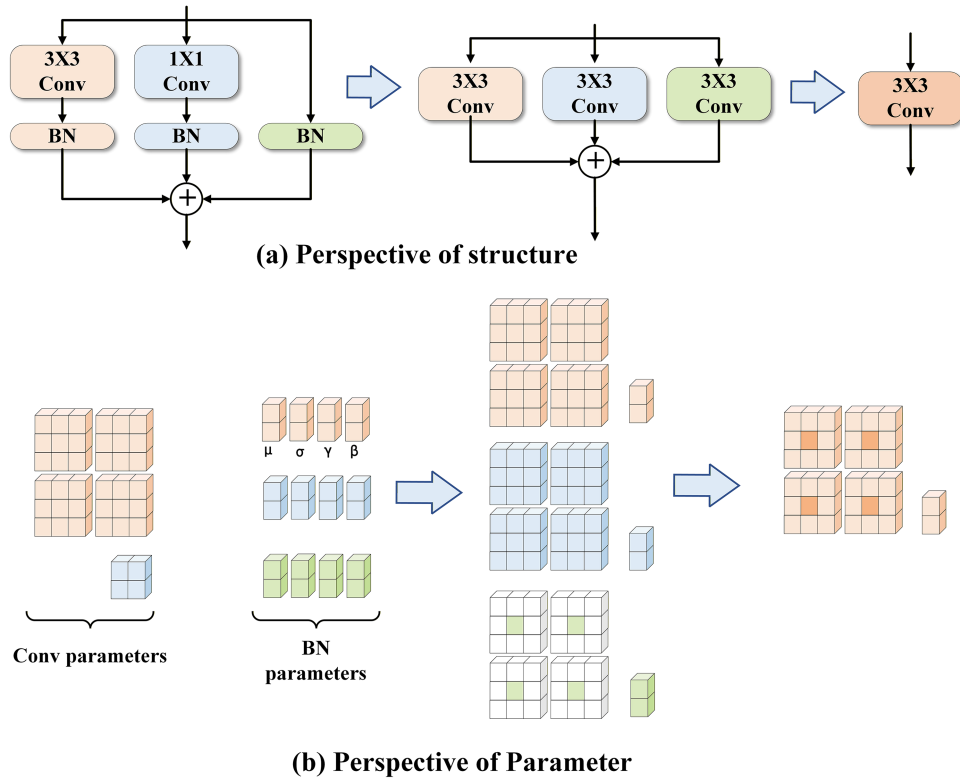


Figure 6: RepConv module structure.

As shown in Fig. 7, the GDFLHead mainly consists of a prediction branch and a decoding branch: in the prediction part, the feature map is successively enhanced by two groups of 3×3 grouped convolutions and outputs a $4 \times \text{reg_max}$ -dimensional regression distribution and an nc -dimensional classification score through the Cv2 and Cv3 convolution layers, respectively. In the decoding part, through dynamic anchor generation, distribution focusing transformation, and coordinate decoding, the final detection results that integrate position and category information are output.

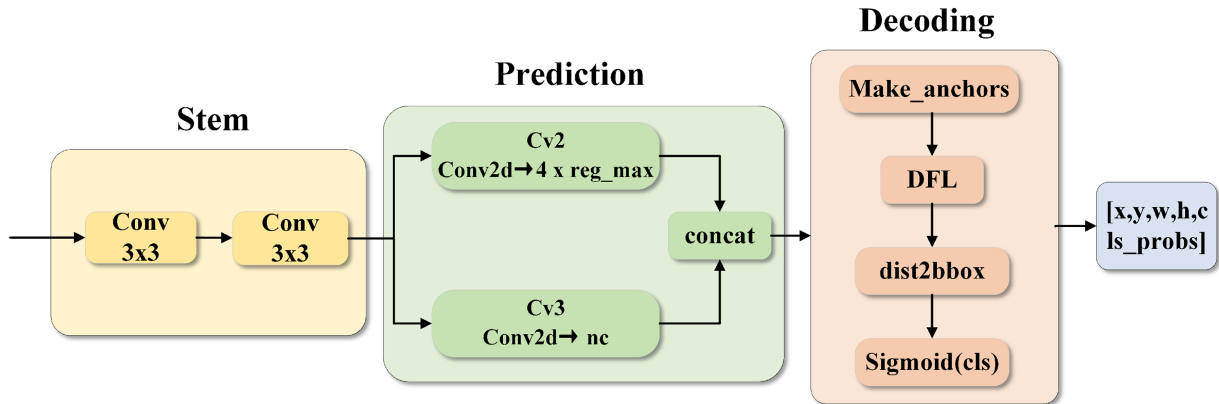


Figure 7: GDFLHead module structure diagram.

4 Experiments and Discussion

4.1 Experimental Dataset

High-quality datasets are the foundation for achieving intelligent defect detection in substations. This study takes a 220 kV substation in Jilin Province as the research object, collects on-site inspection data and historical data, and builds a dedicated dataset for intelligent substation inspection, obtaining a total of 6640 images. The data are mainly collected in summer and autumn and cover different lighting conditions. Most of these images contain a single abnormal target, while a few images contain multiple abnormal targets in the same environment, with multi-target co-occurrence samples accounting for approximately 29% of the total samples. The dataset was constructed for practical intelligent substation inspection scenarios, in which the detection categories include equipment defects, abnormalities around substation equipment, and abnormal events. This setting is consistent with the practical requirement of intelligent inspection systems to detect both equipment defects and abnormalities. Table 2 shows the corresponding label IDs and sample quantities for each detection category.

Table 2: Detection categories in the substation inspection dataset.

Numbering	Detection Category	Label ID	Number of Samples
1	Blurred dials	bj_bpmh	869
2	Broken dials	bj_bpps	723
3	Abnormal meter readings	bjdsyc	789
4	Bird's nest	yw_nc	883
5	Cover damage	gbps	654
6	Suspended object	yw_gkxfw	729
7	Discoloration of respirator silicone	hxq_gjbs	1174
8	Oil on the ground	sly_dmyw	833
9	Not wearing work clothes	wcgz	815

In terms of data annotation, this paper uses the LabelImg tool to generate XML format annotation files and converts them to the TXT format required by the YOLO algorithm. The dataset is divided into training, validation, and test sets in a 7:1:2 ratio. During the division process, equipment instances and scenes are used as the basic units to avoid highly similar images of the same equipment appearing in different data subsets simultaneously, enhancing the independence of the test set and the objectivity of experimental evaluation. To further quantify the data division scheme, after manual review and statistics, the training set, validation set, and test set contain 4648, 664, and 1328 images, respectively, with corresponding target instance numbers of 5228, 747, and 1494, and the number of multi-target co-occurrence samples being 1348, 193, and 385, respectively. They cover 137, 20, and 39 equipment instances and 62, 9, and 18 scene clusters, respectively. Due to the current dataset being sourced from a single substation and mainly collected in summer and autumn, this paper has not yet conducted cross-site validation, and the generalization ability of the model in other substations, different geographical regions, and complex weather conditions still needs further evaluation.

4.2 Experimental Environment and Parameters

The hardware and software environment used in this study is shown in Table 3. To ensure the consistency and fairness of the comparative experiments, all models involved in the comparative experiments were trained and evaluated on the same dataset split under the same experimental settings.

Table 3: Hardware and software configuration.

Platform	Configuration Information
System	Ubuntu20.04
GPU	NVIDIA GeForce RTX 3090(24 G)
CPU	15 vCPU Intel Xeon Platinum 8358P
Language	Python 3.8.0
GPU calculate platform	CUDA 11.8 cuDNN 8.2.0
Deep learning framework	PyTorch 1.10.0

During training, all models used an input image size of 640×640 , with the number of training epochs set to 200, the batch size set to 16, and the number of workers set to 16. Stochastic gradient descent was used to optimize the loss function, with an initial learning rate of 0.01, a final learning rate of 0.001, momentum of 0.937, and weight decay of 0.0005. To ensure a fair comparison, no pretrained weights were loaded for any model. In addition, the same data augmentation strategy, learning rate schedule, confidence threshold, and non-maximum suppression (NMS) threshold were adopted for all comparison models.

4.3 Evaluation Metrics

To comprehensively evaluate the performance of the model in different substation equipment defect detection tasks, this paper uses precision (P), recall (R), F1-score, mean average precision (mAP), number of parameters, billion floating-point operations per second (GFLOPS), and frames per second (FPS) for quantitative analysis. The specific formulations for these evaluation metrics are presented as follows:

$$P = \frac{TP}{TP + FP} \quad (20)$$

$$R = \frac{TP}{TP + FN} \quad (21)$$

$$F_1 = \frac{2PR}{P + R} \quad (22)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N \int_0^1 P_i(R_i) dR \quad (23)$$

$$FPS = \frac{1}{T_p + T_i + T_n} \quad (24)$$

Here, TP represents the number of defects correctly detected, FP represents the number of negative samples incorrectly predicted as defects, FN represents the number of true defects missed by the model, and N is the total number of detection categories in the dataset. In addition, mAP50 denotes the mean average precision calculated at an IoU threshold of 0.50, while mAP50–95 denotes the average of AP values calculated over multiple IoU thresholds from 0.50 to 0.95 with a step size of 0.05. T_p denotes the time required for image preprocessing, T_i denotes the inference time, and T_n denotes the time for post-processing.

4.4 Ablation Experiment of Different Modules

To verify the contribution of each improved module to the model's performance, this paper conducted systematic ablation experiments on the proposed C2MSA backbone module (A), HAFusion cross-layer fusion module (B), EMFPN feature pyramid (C), and GDFLHead detection head (D). The quantitative results are shown in Table 4, and the evolution trend of their performance is visualized in Fig. 8.

Table 4: Results of the ablation experiment.

Module	A	B	C	D	mAP50 (%)	mAP50-95 (%)	P (%)	R (%)	F1 (%)	Params (M)	FPS
M1					72.2	47.5	76.1	67.8	71.7	2.58	469
M2	✓				73.4	47.8	79	68	73.1	2.17	526
M3		✓			73.7	48.7	78.2	67.7	72.6	3.08	377
M4			✓		73.8	48	77.3	68.3	72.5	2.06	467
M5				✓	73.1	48.3	79.8	67.2	73	1.97	537
M6	✓	✓			76.5	50.9	80.3	70.5	75.1	3.32	461
M7	✓	✓	✓		77.8	51.5	81.5	73.2	77.1	3.39	478
GRID-YOLO	✓	✓	✓	✓	78.3	52.3	82	73.6	77.6	3.18	506

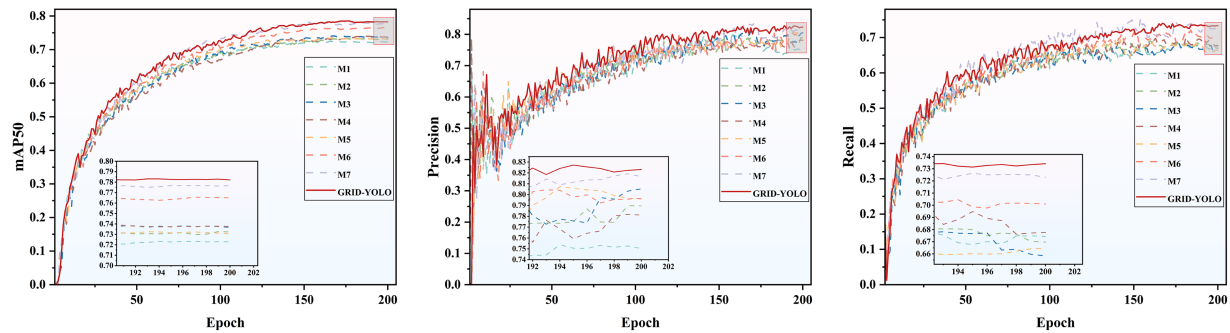


Figure 8: Performance training curves of different module configurations in ablation experiments.

From the perspective of the effect of single-module introduction, experiments M2 and M5 stand out particularly in terms of lightweight and real-time performance. C2MSA increased mAP50 from 72.2% to 73.4%, reduced the number of parameters, and increased the inference speed to 526 FPS, indicating that this module can enhance the ability of semantic feature extraction while taking into account parameter efficiency and real-time performance. GDFLHead improved both Precision and mAP50, reducing the number of parameters to the lowest 1.97 M and increasing the inference speed to 537 FPS, demonstrating a good balance between accuracy and efficiency. In contrast, experiments M3 and M4 focused more on performance enhancement. HAFusion increased mAP50-95 from 47.5% to 48.7%, but the number of parameters increased and the inference speed dropped to 377 FPS, suggesting that while it strengthened cross-layer feature interaction, it also brought a certain computational cost. EMFPN increased the recall rate from 67.8% to 68.3%, and the number of parameters slightly decreased, indicating that its bidirectional feature alignment mechanism can improve the multi-scale feature fusion effect while controlling the model complexity.

In terms of module combination, GRID-YOLO achieved the most significant performance improvement compared to the baseline M1, with mAP50 increasing by 6.1 percentage points. Notably, this gain was not at the expense of significant efficiency, with the complete model having 3.18 M parameters and an inference speed increase of approximately 8%. Compared to M7, GRID-YOLO achieved better detection performance with fewer parameters, indicating that the modules can form effective collaborative optimization, thereby achieving simultaneous improvement in detection accuracy and model efficiency.

Table 5 presents the detection accuracy of the nine detection categories under different module combinations, which is used to analyze the contribution of each module to different categories. The experimental results reveal the differentiated improvement characteristics of each module.

Table 5: The comparison results of nine categories evaluated on dataset.

Module	bpmh	bpps	bjdsyc	yw_nc	gbps	yw_gkxfw	hxq_gjbs	sly_dmyw	wcgz	mAP50 (%)
M1	82.5	70.4	79.7	61.3	66.6	58.4	95.1	47.4	88.5	72.2
M2	80.8	68.5	83.4	63.6	70.5	57.7	94.9	52	88.9	73.4
M3	83.6	69.1	80.1	65.5	69.6	57.6	95.8	53.1	88.6	73.7
M4	83	73.4	82.5	67.4	69	59.2	95.5	47.2	86.9	73.8
M5	81.6	69.8	81.4	69	65.5	58.5	94.4	49	88.6	73.1
M6	85.3	73.9	85.4	67.7	71.6	62	97.2	55.2	90.9	76.5
M7	88.6	76.1	85.6	68.1	71.2	63.1	98.3	54.6	94	77.8
GRID-YOLO	89.4	77.3	86.9	67.3	73.5	64	98.7	56	91.7	78.3

As shown in Table 5, the proposed improvement modules enhance the detection performance of different detection categories to varying degrees. Among the single-module configurations, M3 achieves a detection accuracy of 95.8% for “Discoloration of respirator silicone”, which is the best result among the single-module settings, indicating that HAFusion helps screen effective cross-layer features and enhance key defect information. With the combination of multiple modules, the overall performance is further improved. M7 increases the accuracy of “Blurred dials” to 88.6%, while GRID-YOLO further achieves the best results for “Blurred dials” (89.4%), “Oil on the ground” (56%), and “Suspended object” (64%). In particular, “Oil on the ground” is improved by 8.6 percentage points compared with the baseline. These results indicate that, after systematic integration, the modules complement each other in feature extraction, multi-scale fusion, and localization optimization, significantly improving the detection capability for small defects under complex backgrounds.

4.4.1 Ablation Experiment of Backbone Network Module

This paper further compares the proposed C2MSA with several representative improved backbone structures through ablation experiments to analyze the influence of different module designs on the feature extraction capability of the model. As shown in Table 6, C2MSA achieves the best performance in terms of accuracy ($P = 79\%$) and mAP50 (73.4%), with improvements of 2.9 and 1.2 percentage points respectively compared to the benchmark model. Its mAP50-95 also reaches 47.8%, the highest among all comparison structures. At the same time, C2MSA outperforms most comparison methods in terms of inference speed ($FPS = 526$), demonstrating better comprehensive performance. Fig. 9 further presents the mAP50 convergence curves of each structure during the training process. It can be observed that C2MSA has a more

stable performance improvement trend and a better final convergence result. The above results indicate that the proposed C2MSA can effectively enhance the semantic representation ability of complex target features without significantly sacrificing inference efficiency, thereby improving the feature discrimination ability of the model in multi-scale object detection.

Table 6: Ablation results of different backbone modules.

Model	mAP50 (%)	mAP50-95 (%)	P (%)	R (%)	Params (M)	FPS
Base	72.2	47.5	76.1	67.8	2.58	469
C2BRA	71.9	47.1	75.9	66.6	2.6	514
C2CGA	72.3	47	78.5	63.5	2.56	474
C2DA	72.7	47.4	74.9	67.9	2.6	427
C2DPB	71.9	46.6	76.1	66.2	2.59	506
C2Pola	71.3	46.4	75.6	66.6	2.67	510
C2MSA	73.4	47.8	79	68	2.17	526

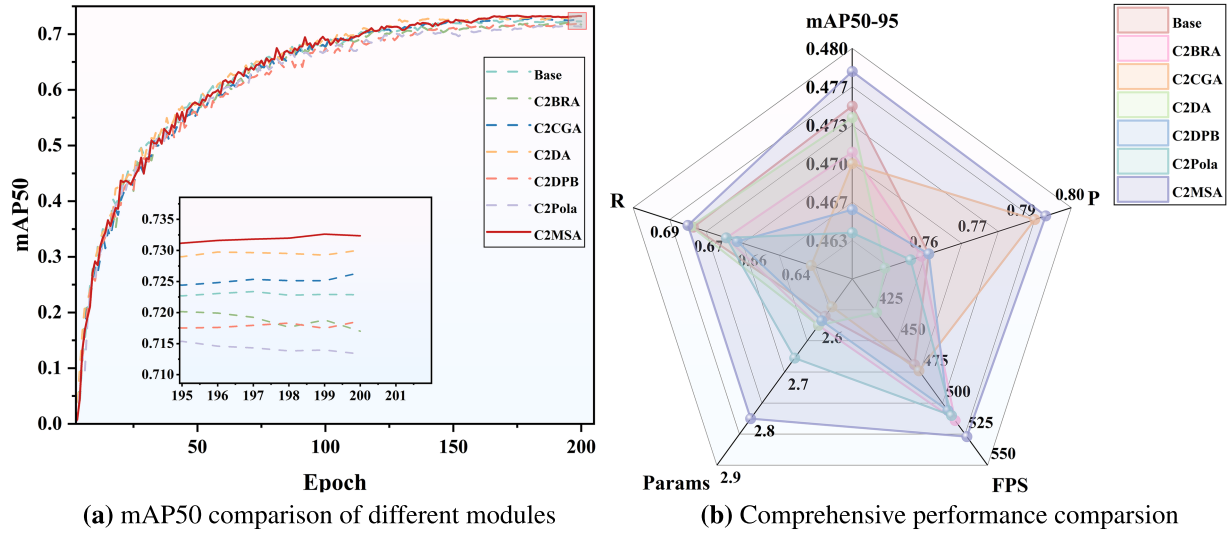


Figure 9: Comparison of different modules in the number of parameters, mAP and FPS.

4.4.2 Ablation Experiment of Feature Pyramid Module

To verify the effectiveness of the proposed EMFPN in multi-scale feature fusion, this paper conducted comparative experiments with several commonly used feature pyramid structures and comprehensively evaluated their performance in the task of substation equipment defect detection. The experimental results are shown in Table 7. EMFPN achieved 73.8% in mAP50, outperforming all comparison methods, with a 1.2 percentage point improvement over the Global Feature Perception Network (GFPN) (72.6%). Meanwhile, it has only 2.06M parameters, significantly lower than GFPN's 3.66M, demonstrating an excellent balance between accuracy and parameters. As shown in the training curve in Fig. 10, EMFPN converges more stably and has a higher performance ceiling. The above results indicate that EMFPN can more effectively enhance multi-scale feature interaction and fusion, and is more conducive to retaining key defect information under

conditions of significant scale variation, complex background, and target occlusion in substation scenarios, thereby improving the detection robustness of the model.

Table 7: Comparison of different feature pyramid networks (FPNs).

Model	mAP50 (%)	mAP50-95 (%)	P (%)	R (%)	Params (M)
Base	72.2	47.5	76.1	67.8	2.58
GFPN	72.6	47.8	78	66.3	3.66
BiFPN [22]	71	46.2	75.7	67.2	1.98
RFPN [23]	72.9	47.4	76	68.1	2.53
MAFPN [24]	72.7	47.1	77	67.7	2.69
HSFPN	69.8	44.7	73.8	64.9	1.92
EMFPN	73.8	48	77.3	68.3	2.06

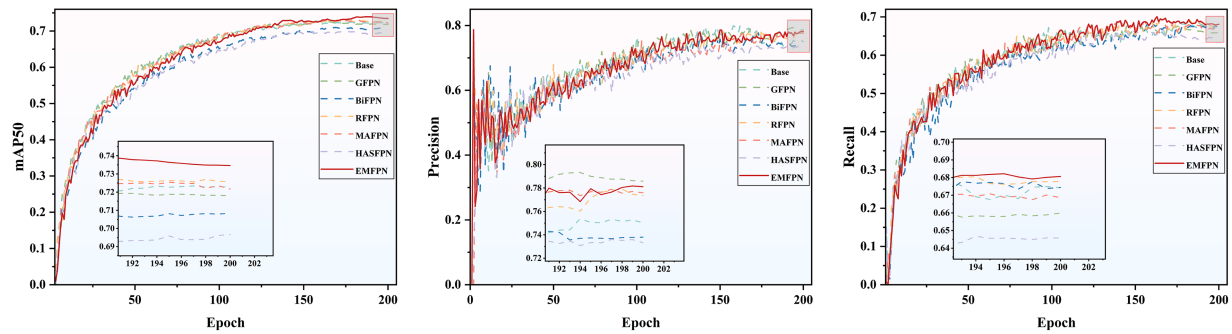


Figure 10: Comparison of different FPNs in precision, recall and mAP50.

4.4.3 Comparative Analysis of Different Detection Heads

According to the comparative results in Table 8, the GDFLHead proposed in this paper performs best in comprehensive detection performance. Fig. 11 further visually compares the computational efficiency and inference speed of each detection head. The left graph plots mAP50 against GFLOPS, showing that GDFLHead resides in the low-GFLOPS region while attaining the highest mAP50, confirming its ability to deliver high accuracy with low computational overhead. The right graph compares FPS, where GDFLHead again leads with an inference speed of 537 FPS. Combining Table 8 and Fig. 11, it can be seen that GDFLHead stands out in parameter efficiency, computational load, and inference speed, better meeting the engineering requirements of real-time, high accuracy, and computational efficiency for defect detection of substation equipment.

Table 8: Comparison of different detection heads.

Model	mAP50 (%)	mAP50-95 (%)	P (%)	R (%)	Params (M)	GFLOPS	FPS
Base	72.2	47.5	76.1	67.8	2.58	6.3	469
Localization Quality Estimator (LQE)	71.7	47.2	75.2	67.7	2.59	6.3	401
Re-parameterized Shared Convolution Detection Head (RSCD)	71.5	45.9	75.7	67.6	2.42	5.6	484
Lightweight Asymmetric Detection Head (LADH)	70.7	46.1	74.3	67	2.28	5.2	501
LSCD [25]	72.6	46.7	79	67.1	2.42	5.6	504
LSCSBD [26]	72	46.8	79.6	65.2	2.46	6.2	431
GDFLHead	73.1	48.3	79.8	67.2	1.97	4.3	537

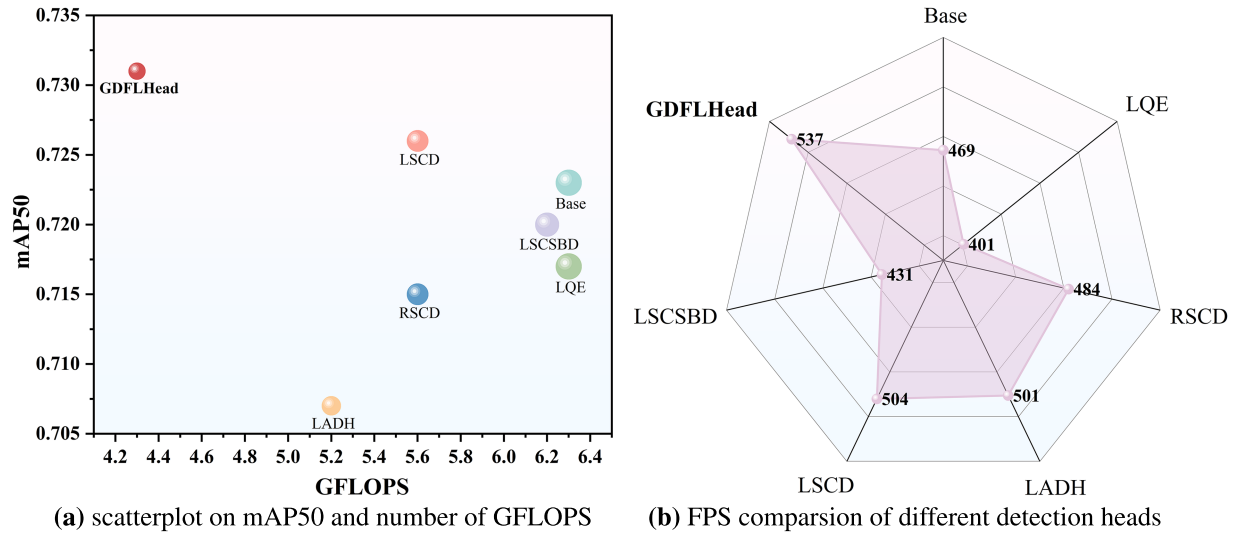


Figure 11: Comparison of different detection heads in mAP, GFLOPS and FPS.

4.5 Performance Comparison of Different Object Detection Algorithms

This paper conducts comparative experiments on different types of representative object detection models, including two-stage detectors, lightweight real-time detection models, and the recent representative improved methods of the YOLO series. To ensure the fairness of the comparison, all models in Table 9 were retrained and evaluated on the same dataset split under the unified experimental settings described in Section 4.2. The results are shown in Table 9.

Table 9: Performance comparison results of different models.

Model	mAP50 (%)	mAP50-95 (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPS	FPS
Faster R-CNN	74.8	47.6	78	68.9	73.2	41.39	182	29
RetinaNet	76.4	49.1	81.4	70.2	75.4	36.5	181	31
RTMDet-tiny	61.3	39.8	64	59.4	61.6	4.87	8.1	125.7
TOOD	75	49.5	73.2	69.5	71.3	32.04	173	22.9
YOLOv5s	72.5	44.5	79	67.4	72.7	7.03	16	476
YOLOv7-Tiny	68.3	40.5	70	64.7	67.2	6.03	13.2	667
YOLOv8	72.5	47.8	78.7	66.3	72	3.01	8.1	526
YOLOv9t	68.3	44.9	74.7	63.4	68.6	2.62	11.1	333
YOLOv10n	70.4	46.5	79.1	64.1	70.8	2.27	8.4	625
YOLOv11n	72.2	47.5	76.1	67.8	71.7	2.58	6.3	469
YOLOv12	70.3	46	72.8	66.5	69.5	2.51	5.8	412
YOLOv13	70.7	45.4	74.1	67.7	70.8	2.45	6.1	365
Hyper-YOLO	73.3	49.3	79.6	69	73.9	3.62	9.5	393
GOLD-YOLO	70.8	45.7	77.8	61.2	68.5	5.89	9.2	397
DAMO-YOLO	71.8	47.4	76.9	67.7	72	5.76	11.2	629
LiteYOLO-ID [27]	71.8	41.4	80.6	66.1	72.6	3.71	9.1	407
MID-YOLO [28]	74.7	49.9	83.5	71	76.7	2.76	5.7	419
TOL-YOLO [14]	75.1	50.2	81	70.3	75.3	5.83	16.5	322
DM-YOLO [15]	73.9	49.4	79	69.3	73.8	3.6	8.3	418
GRID-YOLO	78.3	52.3	82	73.6	77.6	3.18	6.58	506

From Table 9, it can be seen that different types of object detection models present a clear trade-off relationship among detection performance, model complexity, and inference efficiency. Two-stage detectors represented by Faster R-CNN have certain competitiveness in some precision indicators, but their parameter quantity and computational cost are relatively large, making it difficult to meet the application requirements of real-time inspection scenarios. Lightweight models represented by YOLOv7-Tiny have relatively high inference speed, but their performance in terms of mAP, Recall, and F1-score is relatively limited. In contrast, some improved models proposed in recent years, such as MID-YOLO, show strong performance in terms of accuracy, but still have certain gaps in comprehensive performance compared with the proposed method. On this basis, GRID-YOLO demonstrates stronger comprehensive competitiveness in the main detection performance indicators while maintaining a low model complexity and high inference efficiency. Compared with the benchmark model YOLOv11n, its mAP50 and mAP50–95 have increased by 6.1 and 4.8 percentage points, respectively, and the inference speed has increased by approximately 8.0%. Considering the significant scale differences and strong background interference of detection targets in substation inspection scenarios, GRID-YOLO can better balance detection performance and model efficiency, and has strong application potential in this task.

4.6 Visual Analysis of Test Results

This paper uses the gradient-weighted class activation mapping (Grad-CAM) to conduct a heatmap visualization comparison of the baseline model YOLOv11n, the recent mainstream detection model YOLOv13, and GRID-YOLO. This is done to visually analyze the feature-focusing ability of the proposed algorithm on the defect areas of substation equipment. The results are shown in Fig. 12.

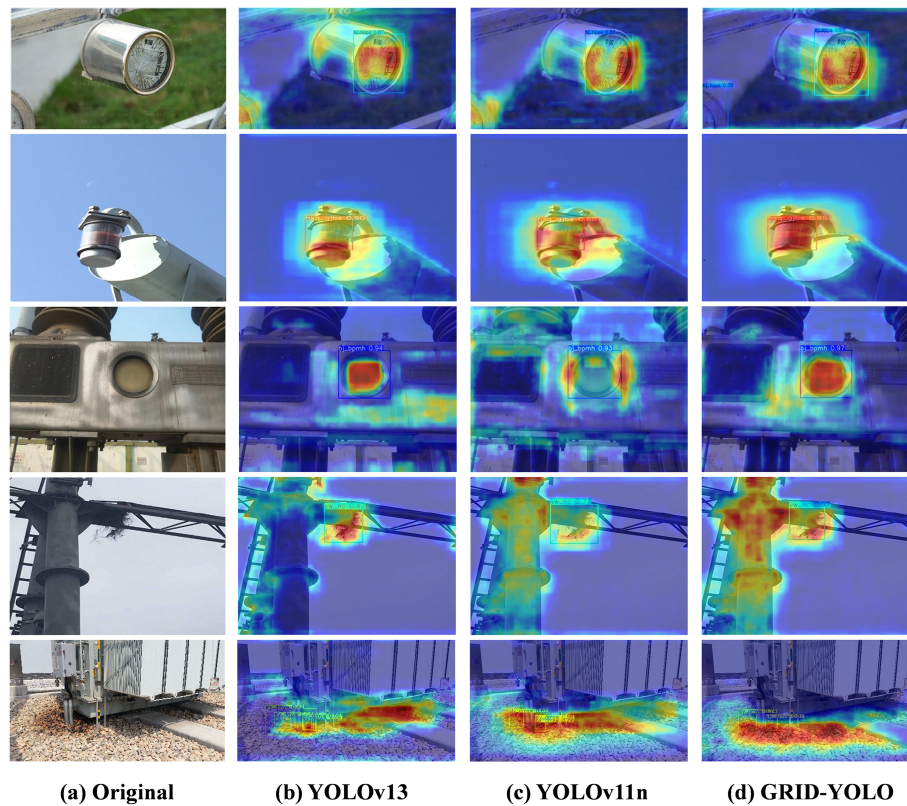


Figure 12: Visualization of heatmap.

Fig. 12a shows the original substation inspection image, which contains typical fine defects such as blurred meter faces and discolored silica gel in the breather. Fig. 12b is the heatmap of YOLOv13, where the activation areas are relatively diffuse, and some high-intensity responses deviate from the actual defect locations and spread to the background areas, indicating that the model's semantic localization ability for defect targets in complex scenes is limited. Fig. 12c is the heatmap of YOLOv11n. Compared with YOLOv13, its attention has somewhat converged, but there is still a certain degree of background noise response, especially insufficient attention to small-scale defects. In contrast, the heatmap of GRID-YOLO shown in Fig. 12d can more effectively focus on the real defect areas and simultaneously suppress background interference, maintaining a relatively clear activation distribution even in complex scenarios with low contrast and occlusion overlap. This indicates that the proposed method can more effectively guide the model to focus on key defect areas, thereby enhancing the detection stability in complex scenes.

4.7 Comprehensive Comparison with Advanced Methods

To further evaluate the comprehensive performance of the proposed method against recent substation equipment defect detection models, several representative methods proposed in recent years are selected for comparison, and all comparison models are retrained under a unified training protocol. The comparison results are shown in Table 10.

Table 10: Comparison of GRID-YOLO with advanced defect detection models for power equipment.

Model	mAP50 (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPS	FPS
YOLO-SS-tiny [29]	72.1	80.6	69.4	74.6	2.51	5.8	474
YOLO-CFA [30]	73.3	78.5	64	70.5	3.05	8.7	339
CB_YOLO [16]	72.8	78	67.8	72.5	3.73	8.25	297
GC-YOLO [17]	71.3	73.4	66.7	69.9	9.61	27.3	126
ES_YOLO [3]	74	78.1	72.1	75	4.03	9.3	247
PBZGNet-n [31]	71.4	73.2	69	71	2.87	7.51	409
SA-LOLO [32]	76.7	80.7	74.2	77.3	2.9	6.7	418
GRID-YOLO	78.3	82	73.6	77.6	3.18	6.58	506

From Table 10, it can be seen that GRID-YOLO demonstrates strong competitiveness in terms of comprehensive detection performance and computational efficiency. Compared with lightweight methods such as YOLO-SS-tiny, YOLO-CFA, and CB_YOLO, the proposed method performs better in metrics such as mAP50 and F1-score, while maintaining a high inference efficiency with a FPS of 506. Although high-complexity models represented by GC-YOLO have certain detection capabilities, their computational costs are significantly higher. In contrast, GRID-YOLO achieves better comprehensive results with lower computational complexity. Notably, SA-LOLO has a slightly higher Recall than GRID-YOLO, but the proposed method still has advantages in comprehensive metrics such as mAP50, F1-score, and FPS. Overall, GRID-YOLO achieves a good balance between detection performance and model efficiency in the substation equipment defect detection task, providing a feasible basis for subsequent engineering applications.

5 Conclusion and Future Work

This paper proposes a lightweight model, GRID-YOLO, for defect detection of substation equipment based on YOLOv11n. The model has been specifically optimized for defect representation in complex

backgrounds, multi-scale feature interaction, and lightweight detection output, thereby enhancing the ability to identify and locate subtle defects. Compared with the base model, GRID-YOLO maintains high inference efficiency and lightweight characteristics while further improving detection accuracy, recall rate, and the ability to recognize small-scale and blurred defects, achieving a better balance between detection performance and model efficiency. The above results indicate that the proposed method can provide effective support for the equipment defect detection task in smart inspection of substations, and offer a feasible basis for achieving efficient and accurate defect identification under resource-constrained conditions. However, due to the current data mainly coming from a single substation and collected mainly in summer and autumn, this paper has not yet conducted external cross-site validation, and its generalization ability in different substations, different geographical regions, and complex weather conditions still needs further evaluation.

The next step of research will focus on exploring the integrated application of GRID-YOLO in smart grid and Internet of Things systems to achieve real-time monitoring of equipment status and automatic alarm, and to enhance the model's robustness in complex environments by integrating multi-source data such as infrared and ultraviolet. At the same time, the model will be extended to other power scenarios such as transmission lines to verify its cross-scenario generalization ability.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: data curation and conceptualization, Hongliang Tian; methodology, experimental implementation, and writing—original draft preparation, Xiaoke Liu; results analysis and interpretation, Bolin Song; manuscript review and formatting, Chenying Pei. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is named the SDD Dataset. Dataset information is available at: <https://github.com/lxk829/Substation-Defect-Detection>.

Ethics Approval: Not applicable.

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

References

1. Li Y, Xu Y, Xu M, Wang S, Xie Z, Li Z, et al. Automatic infrared image recognition method for substation equipment based on a deep self-attention network and multi-factor similarity calculation. *Glob Energy Interconnect*. 2022;5(4):397–408. doi:10.1016/j.gloi.2022.08.006.
2. Tan Y, Zhou L, Xue X, Duan B. Exploration of key technologies for equipment operation and maintenance based on new power systems. *Int J Thermofluids*. 2023;20(2):100482. doi:10.1016/j.ijft.2023.100482.
3. Sun Y, Sun X, Lin Y, Yang Y, Li Z, Du L, et al. Substation equipment defect detection based on improved YOLOv8. *Sensors*. 2025;25(11):3410. doi:10.3390/s25113410.
4. Li Z, Qin Q, Yang Y, Mai X, Ieiri Y, Yoshie O. An enhanced substation equipment detection method based on distributed federated learning. *Int J Electr Power Energy Syst*. 2025;166(1–2):110547. doi:10.1016/j.ijepes.2025.110547.
5. Jabid T, Ahsan T. Insulator detection and defect classification using rotation invariant local directional pattern. *Int J Adv Comput Sci Appl*. 2018;9(2):37. doi:10.14569/IJACSA.2018.090237.
6. Yu Y, Cao H, Wang Z, Li Y, Li K, Xie S. Texture-and-shape based active contour model for insulator segmentation. *IEEE Access*. 2019;7:78706–14. doi:10.1109/ACCESS.2019.2922257.
7. Tsirtsakis P, Zacharis G, Maraslidis GS, Fragulis GF. Deep learning for object recognition: a comprehensive review of models and algorithms. *Int J Cogn Comput Eng*. 2025;6(8):298–312. doi:10.1016/j.ijcce.2025.01.004.

8. Wang J, Sun Y, Liu Y, Zhang K. Lightweight substation equipment defect detection algorithm for small targets. *Sensors*. 2024;24(18):5914. doi:10.3390/s24185914.
9. Dong H, Yuan M, Wang S, Zhang L, Bao W, Liu Y, et al. PHAM-YOLO: a parallel hybrid attention mechanism network for defect detection of meter in substation. *Sensors*. 2023;23(13):6052. doi:10.3390/s23136052.
10. Wang Q, Yang L, Zhou B, Luan Z, Zhang J. YOLO-SS-Large: a lightweight and high-performance model for defect detection in substations. *Sensors*. 2023;23(19):8080. doi:10.3390/s23198080.
11. Qian X, Wang X, Yang S, Lei J. LFF-YOLO: a YOLO algorithm with lightweight feature fusion network for multi-scale defect detection. *IEEE Access*. 2022;10:130339–49. doi:10.1109/ACCESS.2022.3227205.
12. You X, Zhao X. A insulator defect detection network based on improved YOLOv7 for UAV aerial images. *Measurement*. 2025;253(Pt. A):117410. doi:10.1016/j.measurement.2025.117410.
13. Wang Q, Hu Z, Li E, Wu G, Yang W, Hu Y, et al. YOLOLS: a lightweight and high-precision power insulator defect detection network for real-time edge deployment. *Energies*. 2025;18(7):1668. doi:10.3390/en18071668.
14. Zhou C, Fu Y, Mi K. TOL-YOLO: a lightweight transformer oil leakage detection algorithm based on YOLOv8. *Signal Process Image Commun*. 2026;144(12):117533. doi:10.1016/j.image.2026.117533.
15. Ji Y, Zhang D, He Y, Zhao J, Duan X, Zhang T. Improved YOLO11 algorithm for insulator defect detection in power distribution lines. *Electronics*. 2025;14(6):1201. doi:10.3390/electronics14061201.
16. Ding L, Rao Z, Ding B, Li S. Research on defect detection method of railway transmission line insulators based on GC-YOLO. *IEEE Access*. 2023;11:102635–42. doi:10.1109/ACCESS.2023.3316266.
17. Mei Z, Xu H, Yan L, Wang K. IALF-YOLO: insulator defect detection method combining improved attention mechanism and lightweight feature fusion network. *Measurement*. 2025;253(Part C):117701. doi:10.1016/j.measurement.2025.117701.
18. Liu J, Cui G, Xiao C. A real-time and efficient surface defect detection method based on YOLOv4. *J Real Time Image Proc*. 2023;20(4):77. doi:10.1007/s11554-023-01333-4.
19. Hui M, Yao J, Fu Z, Hai T, Zhang M, Pan T. YOLO-CSS: a lightweight defect detection model for complex substation scenarios. *Meas Sci Technol*. 2025;36(8):086003. doi:10.1088/1361-6501/adf65d.
20. Zhou G, Liu M, Wang H, Zheng Y. Power equipment image enhancement processing based on YOLO-v8 target detection model under MSRCR algorithm. *Int J Low Carbon Technol*. 2024;19:1717–24. doi:10.1093/ijlct/ctae122.
21. Wang S, Xu H, Shao M, Zhao L. Insulator defect detection based on improved you-only-look-once v4 in complex scenarios. *J Electron Imaging*. 2023;32(2):023029. doi:10.1117/1.JEI.32.2.023029.
22. Qin Y, Zeng Y, Wang X. Insulator defect detection in complex environments based on improved YOLOv8. *Entropy*. 2025;27(6):633. doi:10.3390/e27060633.
23. Wang X, Zhang Q, Chen C. Dual-branch information extraction and local attention anchor-free network for defect detection. *Sci Rep*. 2024;14(1):10886. doi:10.1038/s41598-024-61324-8.
24. Hu S, Tao X, Zhao X. MCANet: feature pyramid network with multi-scale convolutional attention and aggregation mechanisms for semantic segmentation. *J Vis Commun Image Represent*. 2025;110(10):104466. doi:10.1016/j.jvcir.2025.104466.
25. Xu M, Cai Z, Cai K, Li D, Miao Y, Xu C. IPE-YOLO: a multi-scale defect detection method for power equipment inspection. *Electronics*. 2025;14(19):3767. doi:10.3390/electronics14193767.
26. Jin L, Ding W, Han S, Wang J. A real-time edge inference method for insulator contamination detection with YOLOv11-ssL. *IEEE Trans Instrum Meas*. 2025;74:3534315. doi:10.1109/TIM.2025.3565254.
27. Li D, Lu Y, Gao Q, Li X, Yu X, Song Y. LiteYOLO-ID: a lightweight object detection network for insulator defect detection. *IEEE Trans Instrum Meas*. 2024;73:3418082. doi:10.1109/TIM.2024.3418082.
28. Zhu L, Mao Y, Huang X, Chen C, Yang W, Lei B. MID-YOLO: an enhanced YOLOv8-based method for multi-type insulator defect detection. *Neurocomputing*. 2026;676:132959. doi:10.1016/j.neucom.2026.132959.
29. Wang Q, Liu R, Qin S, Pu J, Shi R, Wang Y. Optimizing power system edge computing with a high-performance and light-weight YOLO-based substation equipment defect detection network. *Int J Electr Power Energy Syst*. 2025;172(1):111088. doi:10.1016/j.ijepes.2025.111088.
30. Zhou Z, Zeng B, Xie Y, Zeng H, Li T, Wei B, et al. Research on infrared image recognition method for substation equipment based on YOLO-CFA. *Results Eng*. 2026;29(10):109503. doi:10.1016/j.rineng.2026.109503.

31. Hu M, Zhuang Y, Wang J, Hu Y, Sun D, Xu D, et al. PBZGNet: a novel defect detection network for substation equipment based on gradual parallel branch architecture. *Sensors*. 2026;26(1):300. doi:10.3390/s26010300.
32. 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(9):112354. doi:10.1016/j.epsr.2025.112354.