

Insulator Defect Detection Method Based on Mixture of Experts Learning

Tianyi Zhang*, Ru Liu, Xiaolong Wang

State Grid Lanzhou Power Supply Company, Lanzhou 730030, China

Keywords: Computer vision; Insulator defect detection; Object detection; Mixture of experts learning

Citation: Zhang T.Y., Liu R., Wang X.L.: Insulator Defect Detection Method Based on Mixture of Experts Learning. Data Intelligence, 2026, Vol. 8, Art. No.: 20260010. DOI: <https://doi.org/10.3724/2096-7004.di.2026.0010>

ABSTRACT

Insulator defect detection is a critical technology for ensuring the safe operation of power systems. However, factors such as diverse defect types, small target scales, and complex backgrounds severely limit the performance of traditional detection methods. This paper proposes a multi-modal insulator defect detection method based on a mixture of experts learning. The method uses Grounding DINO as the base framework and introduces multiple lightweight LoRA-based expert modules in the shallow layers of the image encoder, enabling the model to learn specific features of different defect types. During training, a random expert sampling strategy is adopted to enhance generalization capability, and a two-stage training scheme is used to optimize overall performance; during inference, all expert outputs are fused to fully utilize complementary features. Experimental validation on a self-built dataset containing 2000 real-scene images shows that the proposed method significantly outperforms existing mainstream methods in detection accuracy, particularly excelling in detecting small-target defects while maintaining high inference efficiency. This research provides an effective technical solution for intelligent insulator defect detection in complex scenarios.

1. INTRODUCTION

The safe and stable operation of power systems is an important guarantee for the normal functioning of modern society. As key components in power transmission and transformation lines, insulators serve dual functions of mechanical support and electrical insulation [1]. However, insulators exposed to outdoor environments for extended periods are prone to various defects such as damage, contamination, flashover traces, and cracks. If these defects are not detected and addressed promptly, they may lead to

*Corresponding author: Tianyi Zhang (E-mail: mmdw_xz@163.com; ORCID: 0009-0004-1840-7551).

degraded insulation performance, flashover discharge, or even large-scale power outages [2]. Traditional manual inspection methods suffer from low efficiency, high costs, and significant safety risks, making them inadequate for modern power grid maintenance requirements [3].

In recent years, deep learning-based computer vision technology has made significant progress in object detection [4]. The emergence of Convolutional Neural Networks (CNNs) has greatly enhanced image feature extraction capabilities, driving the development of a series of classic detection algorithms such as Faster Region-based Convolutional Neural Network (Faster R-CNN) [5], You Only Look Once (YOLO) series [6], and Single Shot MultiBox Detector (SSD) [7]. These methods perform excellently in natural scene object detection tasks and have been widely applied in power equipment defect detection [8]. However, insulator defect detection still faces numerous challenges:

First, insulator defects exhibit significant small target characteristics. In high-resolution inspection images, defect regions often occupy only a small portion of the image, placing higher demands on the feature extraction capabilities of detection algorithms [9]. Second, complex background environments increase detection difficulty. Insulator image backgrounds contain various elements such as sky, towers, and conductors, presenting different features under different weather and lighting conditions, easily causing false positives and missed detections [10]. Third, the diversity of defect types requires models to possess strong feature learning capabilities. Different types of defects vary significantly in visual appearance, making it difficult for a single model to accurately detect all defect types simultaneously [11]. Finally, the high cost of acquiring training data. High-quality defect annotation requires professional personnel, and samples of certain severe defects are rare, leading to class imbalance issues in datasets [12]. We show some examples of insulator anomalies in Figure 1.

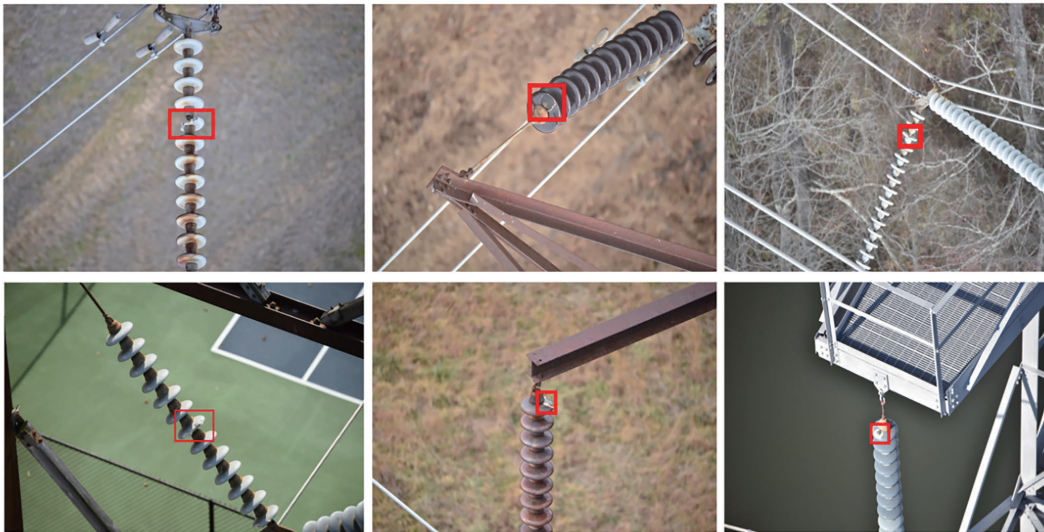


Figure 1. Example Images of Insulator Anomalies.

The introduction of Transformer architecture has brought new breakthroughs to object detection [13]. DETR (Detection Transformer) first applied Transformer to end-to-end object detection, modeling long-range dependencies in images through global attention mechanisms [14]. Subsequent improvements such as Deformable DETR [15] and DINO [16] further enhanced detection performance and convergence speed. More importantly, the rise of vision-language pre-training models has opened a new paradigm for multi-modal object detection [17]. CLIP [18] learned rich vision-language aligned representations through contrastive learning on large-scale image-text pairs. Based on this, a series of open-vocabulary detection methods have been proposed, such as GLIP [19], OWL-ViT [20], and Grounding DINO [21], which can detect objects of arbitrary categories based on text descriptions, demonstrating strong generalization capabilities.

As one of the most advanced multi-modal detection models, Grounding DINO achieves excellent performance in open-vocabulary detection tasks through deep fusion of visual and language features [21]. However, directly applying it to insulator defect detection still has limitations: the model has limited feature learning capability for different defect types, making it difficult to optimize detection performance for all defects simultaneously. This inspires us to consider: it is worthwhile to investigate whether introducing specialized learning mechanisms can enable the model to learn specific feature representations for different defect types.

Mixture of Experts (MoE) models provide an effective solution [22]. MoE processes different input patterns through multiple expert networks and dynamically selects and combines expert outputs through gating mechanisms, demonstrating superior performance in both natural language processing [23] and computer vision [24] tasks. However, traditional MoE methods typically require numerous parameters with high computational overhead. The emergence of Low-Rank Adaptation (LoRA) technology provides a new solution for parameter-efficient fine-tuning [25]. By decomposing weight updates into low-rank matrices, LoRA can achieve effective model adaptation with minimal parameters.

Based on the above analysis, this paper proposes a multi-modal insulator defect detection method based on mixture of experts learning. The main contributions are as follows:

- **Proposed a mixture of experts learning framework for insulator defect detection.** By introducing a multi-expert mechanism based on Grounding DINO, multiple expert modules learn specific features of different defect types, effectively improving the model's detection capability for complex defects.
- **Designed a parameter-efficient expert structure based on LoRA.** Using low-rank decomposition to construct lightweight expert modules significantly reduces parameter overhead while ensuring detection performance, making the method more suitable for practical deployment.
- **Proposed an effective multi-expert training strategy.** Through two-stage training and random expert sampling strategies, the balance between expert specialization and model generalization is achieved, improving training efficiency and final performance.

2. RELATED WORK

2.1 Traditional Object Detection Methods

Object detection is one of the fundamental tasks in computer vision, aiming to identify and locate target objects in images [26]. Early detection methods mainly relied on handcrafted features such as HOG (Histogram of Oriented Gradients) [27] and SIFT (Scale-Invariant Feature Transform) [28], combined with sliding windows and classifiers for detection. The rise of deep learning has completely transformed the object detection field.

Two-stage detectors are represented by the Region-based CNN (R-CNN) series. R-CNN [29] first applied CNN to object detection, generating candidate regions through selective search and extracting features with CNN. Fast R-CNN [30] introduced RoI Pooling to accelerate feature extraction, and Faster R-CNN [5] further proposed Region Proposal Networks (RPN), achieving end-to-end training. Subsequent improvements include FPN [31] introducing feature pyramids to enhance multi-scale detection capability, and Cascade R-CNN [32] improving detection accuracy through cascaded structures.

Single-stage detectors pursue a balance between speed and accuracy. YOLO (You Only Look Once) [6] transformed the detection task into a regression problem, achieving real-time detection. The YOLOv2-v8 series continuously improved network structures and training strategies [33-34]. SSD [7] predicted targets of different sizes through multi-scale feature maps. RetinaNet [35] proposed Focal Loss to address class imbalance issues. Recent FCOS [36] and CenterNet [37] adopted anchor-free designs, simplifying the detection process.

2.2 Transformer-based Detection Methods

The Transformer architecture initially achieved success in natural language processing [38] and was subsequently introduced to computer vision tasks. Vision Transformer (ViT) [39] first demonstrated the effectiveness of pure Transformer architecture in image classification. Swin Transformer [40] better adapted to visual tasks through hierarchical structures and shifted window mechanisms.

DETR (Detection Transformer) [14] was the first Transformer-based end-to-end detector, eliminating hand-designed components like NMS by treating detection as a set prediction problem. However, DETR suffered from slow convergence and poor small object detection performance. Deformable DETR [15] introduced deformable attention mechanisms, significantly improving convergence speed and detection accuracy. Conditional DETR [41] accelerated training through conditional spatial queries. DINO [16] combined contrastive denoising training and mixed query selection to achieve state-of-the-art performance. Recent works like DAB-DETR [42] and DN-DETR [43] continue to improve query design and training strategies.

2.3 Multi-modal Object Detection

The success of vision-language pre-training models has driven the development of multi-modal object detection. Contrastive Language-Image Pre-training (CLIP) [18] learned powerful vision-language aligned representations through contrastive learning on 400 million image-text pairs. ALIGN [44] and FILIP [45] further expanded training scale and improved alignment methods.

Open-vocabulary detection aims to detect categories unseen during training. OVR-CNN [46] achieved open-vocabulary detection through vision-language mapping. ViLD [47] improved detection performance using CLIP's knowledge distillation. RegionCLIP [48] enhanced localization capability through region-level contrastive learning. GLIP [19] unified detection and phrase grounding tasks, pretraining on large-scale data. OWL-ViT [20] implemented simple and efficient open-vocabulary detection based on Vision Transformer.

Grounding DINO [21] is one of the most advanced multi-modal detection models. It combines DINO's powerful detection capability with multi-modal feature fusion, achieving deep interaction between vision and language through cross-modal decoders. The model achieves excellent performance on multiple benchmark datasets and demonstrates strong zero-shot generalization capability. X-DETR [49] and GDINO-X [50] further extend Grounding DINO's capabilities.

2.4 Power Equipment Defect Detection

Deep learning has been widely applied in power equipment defect detection. Early works mainly used traditional CNNs for insulator detection and fault diagnosis [51]. Tao et al. [52] proposed an insulator defect detection method based on Faster R-CNN. Ling et al. [53] used YOLOv3 to detect multiple components of transmission lines.

Addressing the special characteristics of insulators, researchers have proposed various improvements. Zhao et al. [54] designed cascade detectors to improve small defect detection accuracy. Liu et al. [55] enhanced key features through attention mechanisms, and improved detection performance using multi-scale feature fusion. Recently, some works have begun exploring Transformer applications in power equipment detection [56-58], but multi-modal methods remain rare in this field.

2.5 Mixture of Experts Models

Mixture of Experts (MoE) models process different input patterns through multiple expert networks [22]. Sparsely-Gated MoE [59] reduces computational cost through sparse activation. Switch Transformer [23] extends MoE to trillion-parameter scales. Expert Choice [60] improves expert selection mechanisms.

In computer vision, Vision MoE [24] first applied MoE to image classification. Soft MoE [61] improves performance through soft routing mechanisms. Recent works explore MoE applications in detection and segmentation tasks [62-63].

Parameter-efficient fine-tuning methods provide new approaches for large model adaptation. LoRA [25] achieves efficient fine-tuning through low-rank decomposition. AdaLoRA [64] adaptively allocates parameter budgets. QLoRA [65] further reduces memory requirements through quantization. These methods provide technical foundations for constructing lightweight expert modules.

3. THE PROPOSED APPROACH

3.1 Overall Architecture

The proposed mixture of experts learning framework builds upon Grounding DINO, enhancing the model's detection capability for insulator defects through multi-expert mechanisms. As shown in Figure 1, the system architecture contains four core components: (1) dual-encoder structure for extracting visual and language features; (2) multi-expert modules learning defect-specific representations through multiple lightweight experts; (3) feature enhancement module for deep fusion of cross-modal features; (4) crossmodal decoder generating final detection results.

Given input image $I \in \mathbb{R}^{H \times W \times 3}$, where H , W , and 3 represent the height, width, and number of channels of the input image, respectively. And defect category text descriptions $T = \{t_1, t_2, \dots, t_c\}$, the system outputs detection result set $\{(b_i, c_i, s_i)\}_{i=1}^N$ where b_i, c_i, s_i represent bounding box coordinates, category labels, and confidence scores respectively.

3.2 Base Framework: Grounding DINO

Grounding DINO processes multi-modal inputs through a dual-encoder architecture, achieving effective fusion of visual and language information.

Visual Feature Extraction: Uses Swin Transformer as the image encoder to extract multi-scale hierarchical features:

$$F_v = \text{SwinTransformer}(I) = \{f_v^l\}_{l=1}^L$$

where $f_v^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ represents the feature map at layer l , and L is the number of layers. Swin Transformer effectively captures local and global context information through shifted window mechanisms.

Language Feature Encoding: Uses pre-trained BERT model to process text descriptions, generating semantic representations:

$$F_t = \text{BERT}(T) \in \mathbb{R}^{N_t \times D}$$

where N_t is the text sequence length and D is the feature dimension. In this paper, we use the "defect region of insulator" as the inputting text.

Cross-modal Feature Enhancement: Achieves deep interaction of vision-language features through bidirectional attention mechanisms:

$$\begin{aligned}\tilde{F}_v &= F_v + \text{MHA}(Q = F_v, K = F_t, V = F_t) \\ \tilde{F}_t &= F_t + \text{MHA}(Q = F_t, K = F_v, V = F_v)\end{aligned}$$

where MHA denotes multi-head attention. This bidirectional interaction enables visual features to perceive semantic information while language features can focus on relevant visual regions.

Detection Head Design: Adopts query-based decoder to generate detection results. Initialize a set of learnable queries $Q \in \mathbb{R}^{N_q \times D}$, which interact with enhanced features through multi-layer Transformer decoder:

$$\hat{Q} = \text{TransformerDecoder}(Q, \tilde{F}_v, \tilde{F}_t)$$

Finally, predict bounding boxes and categories through two independent MLP branches:

$$b_i = \text{MLP}_{\text{box}}(\hat{q}_i), \quad p_i = \text{MLP}_{\text{cls}}(\hat{q}_i)$$

3.3 Multi-Expert LoRA Mechanism

To enhance the model's discriminative capability for different defect types, we introduce multi-expert mechanisms at key layers of the image encoder. Considering practical deployment efficiency requirements, we construct parameter-efficient expert modules using Low-Rank Adaptation (LoRA) technology. The feature flow diagram for single-layer multi-expert training is shown in Figure 2.

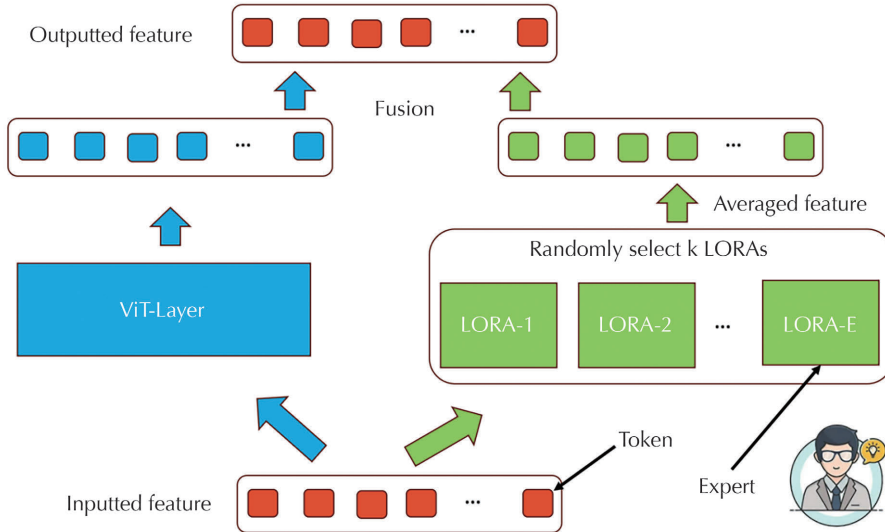


Figure 2. Feature Flow Diagram for Single-Layer Multi-Expert Training.

3.3.1 LoRA Expert Design

Traditional full fine-tuning requires updating all model parameters, incurring huge computational and storage overhead. LoRA achieves parameter-efficient model adaptation through low-rank decomposition. For pre-trained weights $W_0 \in \mathbb{R}^{d_{out} \times d_{in}}$, the e -th expert performs incremental updates through low-rank matrices:

$$W_e = W_0 + \Delta W_e = W_0 + \alpha B_e A_e$$

where $B_e \in \mathbb{R}^{d_{out} \times r}$ and $A_e \in \mathbb{R}^{r \times d_{in}}$ are low-rank decomposition matrices, $\text{rank } r \ll \min(d_{out}, d_{in})$, and α is a scaling factor. This design requires only $r \times (d_{out} + d_{in})$ parameters per expert, significantly reducing storage and computational costs.

The initialization strategy is crucial for LoRA performance. We adopt the following initialization scheme: A_e is initialized with Gaussian distribution $\mathcal{N}(0, \sigma^2)$, B_e is initialized as zero matrix, ensuring $\Delta W_e = 0$ at the beginning of training to maintain model stability. In our experiment, $r = 10$ achieves the best balance between accuracy and parameter efficiency.

3.3.2 Multi-Expert Deployment Strategy

Through experimental analysis, we deploy expert modules at layers 2, 3, and 4 of the image encoder. The selection of these relatively shallow layers is based on the following considerations: (1) shallow features contain more fine-grained information, beneficial for capturing detailed features of insulator defects; (2) introducing experts in shallow layers enables earlier feature differentiation, improving representation capability of subsequent layers; (3) avoiding excessive parameters in deep layers to maintain model efficiency.

For selected layer l , deploy E experts $\mathcal{E}^l = \{W_e^l\}_{e=1}^E$. Given input features x^l , each expert produces specific outputs:

$$h_e^l = (W_0^l + \alpha B_e^l A_e^l) \cdot x^l$$

3.3.3 Dynamic Fusion Mechanism

To balance expert specialization and model generalization, we design different fusion strategies:

Training Phase-Random Expert Sampling: Randomly activate k experts per forward pass ($k \sim \text{Uniform}\{2, 3\}$) and average their outputs:

$$h^l = \frac{1}{|\mathcal{S}|} \sum_{e \in \mathcal{S}} h_e^l, \mathcal{S} \sim \text{RandomSample}(\mathcal{E}^l, k)$$

This random sampling strategy brings multiple advantages: (1) prevents the model from over-relying on specific experts, enhancing robustness; (2) introduces randomness as regularization, reducing overfitting; (3) promotes collaborative learning among experts.

Inference Phase-Full Expert Ensemble: Fuse all expert outputs to fully utilize learned feature representations:

$$h^l = \lambda \cdot W_0^l x^l + (1 - \lambda) \cdot \frac{1}{E} \sum_{e=1}^E h_e^l$$

where $\lambda \in [0, 1]$ is a balance factor controlling the relative contribution of original and expert features. Through validation set search, we set λ to 0.5.

3.4 Optimization Strategy

3.4.1 Two-Stage Training Framework

To fully leverage the advantages of multi-expert mechanisms, we adopt a progressive two-stage training strategy:

Stage 1 - Rapid Expert Adaptation (20 epochs):

- Freeze all pre-trained parameters of Grounding DINO
- Only optimize newly added LoRA expert parameters $\Theta_{\text{LoRA}} = \{B_e^l, A_e^l\}$
- Use relatively large learning rate ($\eta = 10^{-3}$) to accelerate convergence
- Goal: Enable expert modules to quickly learn feature patterns of insulator defects

Stage 2 - Global Joint Optimization (15 epochs):

- Unfreeze all trainable parameters
- Use smaller learning rate ($\eta = 10^{-4}$) for fine-tuning
- Apply cosine annealing to gradually reduce learning rate to ($\eta_{\min} = 10^{-6}$)
- Goal: Optimize overall model performance and achieve collaborative work among components

3.4.2 Loss Function Design

Adopt multi-task learning framework with total loss function consisting of three parts:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{box}} \mathcal{L}_{\text{box}} + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}$$

Classification Loss: Use Focal Loss to alleviate class imbalance:

$$\mathcal{L}_{\text{cls}} = -\alpha_t (1 - p_t)^\gamma \log(p_t)$$

where p_t is predicted probability, α_t is class weight, and $\gamma = 2$ is focusing parameter.

Localization Loss: Combine L1 loss and GloU loss to optimize bounding box prediction:

$$\mathcal{L}_{\text{box}} = \|\hat{b} - b\|_1 + \mathcal{L}_{\text{GloU}}(\hat{b}, b)$$

IoU Loss: Directly optimize intersection over union between predicted and ground truth boxes:

$$\mathcal{L}_{\text{iou}} = 1 - \text{IoU}(\hat{b}, b)$$

Loss weights are set as $\lambda_{\text{cls}} = 2$, $\lambda_{\text{box}} = 5$, $\lambda_{\text{iou}} = 2$.

4. EXPERIMENTS AND ANALYSIS

4.1 Experimental Setup

Experiments are conducted on a self-built insulator defect detection dataset. The dataset is collected through field inspections at substations, containing insulator images from various real scenarios, totaling 2000 high-resolution images split equally into training and test sets of 1000 images each. The dataset covers major defect types including insulator damage, contamination, and cracks, fully reflecting the complexity and diversity of practical application scenarios. Over 59% of instances occupy less than 1% of the total image pixel area, highlighting the strong “small target” characteristics of our dataset. Additionally, a pie chart summarizing defect type distribution and three annotated image examples showing bounding boxes and defect labels have been added to improve reproducibility.

The multi-expert network configuration is as follows: 5 experts are deployed at layers 2, 3, and 4 of the model, each expert using LoRA structure with rank 10 to reduce parameters; during training, random sampling strategy is adopted where 2-3 experts are randomly selected for each forward pass, and fusion weight $\lambda = 0.5$ is used to balance expert outputs and original features. This configuration ensures model expressiveness while effectively controlling computational overhead.

We adopt the standard COCO evaluation system, including the following metrics:

- AP (Average Precision): Average precision with IoU thresholds from 0.5 to 0.95 (step 0.05), comprehensively reflecting detection performance.
- $\text{AP}_{50}/\text{AP}_{75}$: Precision at IoU thresholds of 0.5 and 0.75 respectively, evaluating performance under different localization accuracy requirements.
- AR (Average Recall): Average recall rate, reflecting detection completeness.
- FPS (Frames Per Second): Inference speed, evaluating real-time performance.

4.2 Main Results

To comprehensively evaluate our method's performance, we selected seven representative detection methods for comparison: including classic two-stage detector Faster R-CNN [5] (using ResNet-50-FPN backbone), latest single-stage detector YOLOv8-L [34] (achieving good balance between speed and accuracy), first end-to-end Transformer detector DETR [14] (eliminating hand-designed components), DINO [16] improving performance through contrastive denoising training, large-scale vision-language pre-training model GLIP [19] supporting open-vocabulary detection, Vision Transformer-based openvocabulary detector OWL-ViT [20], and current state-of-the-art multi-modal detector Grounding DINO [21] serving as our baseline.

Table 1 shows comprehensive performance comparison of different methods on the test set. Experimental results demonstrate that our method achieves best performance on all evaluation metrics. Compared to baseline Grounding DINO, AP improves by 4.3 percentage points to 51.8%, fully validating the effectiveness of multi-expert mechanisms. Although inference speed slightly decreases (from 8.7 FPS to 8.5 FPS), it still maintains real-time processing capability, meeting practical application requirements.

Table 1. Performance comparison of different methods on insulator defect detection dataset.

Method	AP	AP ₅₀	AP ₇₅	AR	FPS
Faster R-CNN	42.3	68.5	45.2	52.1	15.2
YOLOv8-L	45.7	71.2	48.9	55.8	28.6
DETR	41.8	67.3	44.1	51.3	12.5
DINO	44.2	70.1	47.3	54.2	11.8
GLIP	46.3	72.8	49.7	56.7	9.3
OWL-ViT	43.9	69.6	46.8	53.6	10.2
Grounding DINO (Baseline)	47.5	74.2	51.3	58.2	8.7
Proposed Method	51.8	78.6	55.9	62.5	8.5

4.3 Ablation Studies

Table 2 evaluates the influence of different multi-expert configurations on detection performance. The results reveal three key observations. First, introducing even a single expert yields a 0.9 AP improvement over the baseline, clearly demonstrating the effectiveness of the expert mechanism in enhancing feature representation. Second, gradually increasing the number of experts and stacking them across more layers leads to continuous gains, with the three-layer expert configuration achieving the highest accuracy. Third, positioning experts in the shallow layers significantly outperforms deployment in deeper layers. This is because shallow feature maps retain more detailed spatial and texture information, which is crucial for capturing the fine-grained characteristics of insulator defects and improving detection of small targets.

Table 2. Ablation Study of Multi-Expert Modules.

Configuration	AP	AP ₅₀
Baseline (no experts)	47.5	74.2
Single layer single expert (Layer 2)	48.3	75.1
Single layer 5 experts (Layer 2)	49.7	76.4
Two layers 5 experts (Layers 2, 3)	50.9	77.8
Three layers 5 experts (Layers 2, 3, 4)	51.8	78.6
Deep layers 5 experts (Layers 10, 11, 12)	49.2	75.9

Table 3 presents an analysis of the effect of LoRA rank on detection performance. Experimental results indicate that a rank of $r = 10$ achieves the optimal balance between accuracy and efficiency. When the rank is too low (e.g., $r = 5$), the model's capacity is restricted, leading to lower accuracy (AP = 50.2). Conversely, an excessively high rank ($r = 15, 20$) does not bring further gains and may introduce overfitting (AP = 51.6 and 51.4 respectively), degrading generalization to unseen data. These findings suggest that a rank of 10 is sufficient to effectively capture the key feature variation patterns of insulator defects while maintaining computational efficiency.

Table 3. Impact of LoRA rank r .

r value	AP	AP ₅₀
$r = 5$	50.2	76.9
$r = 10$	51.8	78.6
$r = 15$	51.6	78.4
$r = 20$	51.4	78.2

Table 4 shows the trade-off between expert numbers and performance-efficiency. 5 experts achieve the best performance-efficiency balance. Further increasing expert numbers brings limited performance improvement but significantly reduces inference speed. This may be because too many experts cause feature redundancy and increase optimization difficulty.

Table 4. Impact of expert numbers.

Number of Experts	AP	AP ₅₀	FPS
3 experts	50.1	76.8	8.9
5 experts	51.8	78.6	8.5
7 experts	51.7	78.5	8.2
10 experts	51.5	78.3	7.8

Table 5 compares the effects of deploying experts at different layers. Shallow layer deployment (Layers 2, 3, 4) performs best, consistent with our design philosophy: shallow features preserve more spatial detail information beneficial for detecting subtle defects; introducing experts in shallow layers enables earlier feature differentiation, providing better feature foundation for subsequent layers.

Table 5. Impact of expert layer positions.

Expert Layers	AP	AP ₅₀	Description
Layers 1, 2, 3	51.2	77.9	Very shallow
Layers 2, 3, 4	51.8	78.6	Shallow
Layers 5, 6, 7	50.6	77.3	Middle
Layers 10, 11, 12	49.2	75.9	Deep

4.4 Training Strategy Analysis

Table 6 compares the effects of three training strategies. Two-stage training performs best (AP = 51.8%, AP₅₀ = 78.6%) and converges in only 35 epochs. This strategy achieves rapid adaptation through Stage 1 (20 epochs) freezing backbone and training only expert modules, then achieves deep collaboration through Stage 2 (15 epochs) jointly optimizing all parameters, preserving pre-trained knowledge while introducing task-specific capabilities. In contrast, end-to-end training theoretically can achieve global optimum but performs poorly (AP = 49.3%) due to optimization instability and slow convergence (45 epochs); expert-only training converges quickly (30 epochs) but has lowest performance (AP = 48.7%) as it cannot optimize backbone features through backpropagation. Two-stage strategy achieves best balance between performance and efficiency, reducing training time by 22% while improving AP by 2.5 percentage points compared to end-to-end training, fully validating the effectiveness of progressive optimization strategy.

Table 6. Impact of different training strategies.

Training Strategy	AP	AP ₅₀	Convergence Epochs
End-to-end training	49.3	75.8	45
Expert-only training	48.7	75.2	30
Two-stage training	51.8	78.6	35

Table 7 shows the impact of different expert sampling strategies during training. Random selection of 2-3 experts achieves best performance (AP = 51.8%, AP₅₀ = 78.6%). This strategy randomly activates 2-3 experts each time, ensuring differentiated learning among experts to avoid homogenization while achieving implicit collaboration through partial overlap. Compared to fixed or random selection of single expert (AP of 49.5% and 50.3% respectively), multi-expert combination significantly improves feature expressiveness; while using all 5 experts has similar performance (AP = 51.4%). Random selection of 2-3 experts achieves best balance between performance and efficiency, validating the importance of moderate randomness and expert numbers for efficient learning.

Table 7. Impact of expert sampling strategies during training.

Sampling Strategy	AP	AP ₅₀
Fixed selection of 1 expert	49.5	76.2
Random selection of 1 expert	50.3	77.1
Random selection of 2-3 experts	51.8	78.6
Using all 5 experts	51.4	78.2

We conduct an ablation study comparing three inference fusion strategies in Table 8: equal-weight fusion of all experts, weighted fusion with $\lambda = 0.5$, and partial expert fusion using a random subset. The results show that while equal-weight fusion achieves slightly higher AP, the weighted fusion with $\lambda = 0.5$ delivers the best trade-off between accuracy and inference speed, aligning with the design adopted in our final model.

Table 8. Ablation on expert fusion strategies during inference.

Fusion Strategy	AP	AP ₅₀	FPS
All experts, equal weight	51.9	78.7	8.2
Weighted fusion ($\lambda = 0.5$)	51.8	78.6	8.5
Random subset of 3 experts	51.5	78.4	8.6

5. CONCLUSION

This paper proposes a multi-modal insulator defect detection method based on mixture of experts learning to address challenges such as small target detection difficulty, complex backgrounds, and diverse defect types in power system insulator defect detection. The method uses Grounding DINO as the base framework and introduces LoRA-based lightweight expert modules in shallow layers of the image encoder, achieving effective learning of specific features for different defect types. Core innovations include: designing parameter-efficient multi-expert network structure that ensures detection performance while controlling computational overhead; proposing two-stage training strategy and random expert sampling mechanism that effectively balance expert specialization and model generalization; adopting expert fusion mechanism to fully utilize complementary features and improve overall detection performance.

AUTHOR CONTRIBUTIONS

Tianyi Zhang: Conceptualization, Methodology, Writing–Original Draft.

Ru Liu: Data Curation, Experiments, Validation.

Xiaolong Wang: Supervision, Writing–Review & Editing, Project Administration.

ACKNOWLEDGEMENTS

This work is supported by State Grid Gansu Electric Power Company 2025 Research and Development Special Fund (522701250002).

REFERENCES

- [1] T. Jabid and T. Ahsan, “Insulator detection and defect classification using rotation invariant local directional pattern,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, 2018.
- [2] V. K. Kaliappan, F. G., D. K., et al, “Fault detection in polymer insulators for powerlines using image processing based on machine learning and deep learning algorithms,” *Turk. Online J. Qual. Inquiry*, vol. 12, no. 3, 2021.
- [3] B. Jalil, G. R. Leone, M. Martinelli, et al, “Fault detection in power equipment via an unmanned aerial system using multi modal data,” *Sensors*, vol. 19, no. 13, p. 3014, 2019.
- [4] L. Liu, W. Ouyang, X. Wang, et al, “Deep learning for generic object detection: A survey,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.
- [5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

- [7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.
- [8] S. Yang, X. Li, X. Jia, et al, "Deep learning-based intelligent defect detection of cutting wheels with industrial images in manufacturing," *Procedia Manuf.*, vol. 48, pp. 902–907, 2020.
- [9] C. Chen, Y. Zhang, Q. Lv, et al, "RRNet: A hybrid detector for object detection in drone-captured images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), Seoul, Korea (South), Oct. 2019.
- [10] W. Tang, Q. Yang, K. Xiong, et al, "Deep learning based automatic defect identification of photovoltaic module using electroluminescence images," *Sol. Energy*, vol. 201, pp. 453–460, May 2020.
- [11] I. Valavanis and D. Kosmopoulos, "Multiclass defect detection and classification in weld radiographic images using geometric and texture features," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7606–7614, Dec. 2010.
- [12] J. Johnson and T. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.
- [13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.
- [14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, pp. 213–229.
- [15] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. Int. Conf. Learn. Represent. (ICLR), Vienna, Austria, May 2021.
- [16] H. Zhang, F. Li, S. Liu, et al, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in Proc. Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, May 2023.
- [17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.
- [18] A. Radford, J. W. Kim, C. Hallacy, et al, "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), Virtual, Jul. 2021, pp. 8748–8763.
- [19] L. H. Li, P. Zhang, H. Zhang, et al, "Grounded language-image pre-training," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 10965–10975.
- [20] M. Minderer, A. Gritsenko, A. Stone, et al, "Simple open-vocabulary object detection with vision transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), Tel Aviv, Israel, Oct. 2022, pp. 728–745.
- [21] S. Liu, Z. Zeng, T. Ren, et al, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
- [22] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991.
- [23] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 120, pp. 1–39, 2022.
- [24] C. Riquelme, J. Puigcerver, B. Mustafa, et al, "Scaling vision with sparse mixture of experts," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, Virtual, Dec. 2021, pp. 8583–8595.
- [25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022.
- [26] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.
- [27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), San Diego, CA, USA, Jun. 2005, pp. 886–893.
- [28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

- [29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, Jun. 2014, pp. 580–587.
- [30] R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1440–1448.
- [31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.
- [32] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 6154–6162.
- [33] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
- [34] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, Jun. 2023, pp. 7464–7475.
- [35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 2980–2988.
- [36] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Korea (South), Oct. 2019, pp. 9627–9636.
- [37] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint arXiv:1904.07850, 2019.
- [38] A. Vaswani, N. Shazeer, N. Parmar, et al, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NIPS), vol. 30, Long Beach, CA, USA, Dec. 2017.
- [39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, May 2021.
- [40] Z. Liu, Y. Lin, Y. Cao, et al, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Virtual, Oct. 2021, pp. 10012–10022.
- [41] D. Meng, X. Chen, Z. Fan, et al, "Conditional DETR for fast training convergence," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Virtual, Oct. 2021, pp. 3651–3660.
- [42] S. Liu, F. Li, H. Zhang, et al, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022.
- [43] F. Li, H. Zhang, S. Liu, et al, "DN-DETR: Accelerate DETR training by introducing query denoising," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 13619–13627.
- [44] C. Jia, Y. Yang, Y. Xia, et al, "Scaling up visual and vision-language representation learning with noisy text supervision," in Proc. Int. Conf. Mach. Learn. (ICML), Virtual, Jul. 2021, pp. 4904–4916.
- [45] L. Yao, R. Huang, L. Hou, et al, "FILIP: Fine-grained interactive language-image pre-training," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022.
- [46] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Virtual, Jun. 2021, pp. 14393–14402.
- [47] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary detection via vision and language knowledge distillation," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022.
- [48] Y. Zhong, J. Yang, P. Zhang, et al, "RegionCLIP: Region-based language-image pretraining," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 16793–16803.
- [49] M. Cai, Y. Li, D. Yang, et al, "X-DETR: A versatile architecture for instance-wise vision-language tasks," in Proc. Eur. Conf. Comput. Vis. (ECCV), Tel Aviv, Israel, Oct. 2022, pp. 290–307.

- [50] T. Ren, Y. Chen, Q. Jiang, et al, "DINO-X: A unified vision model for open-world object detection and understanding," arXiv preprint arXiv:2411.14347, 2024.
- [51] S. F. Stefenon, L. O. Seman, G. Singh, et al, "Enhanced insulator fault detection using optimized ensemble of deep learning models based on weighted boxes fusion," *Int. J. Electr. Power Energy Syst.*, vol. 168, p. 110682, Jun. 2025. doi:10.1016/j.ijepes.2025.110682.
- [52] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 4, pp. 1486–1498, Apr. 2020.
- [53] Z. Ling, D. Zhang, R. C. Qiu, et al, "An accurate and real-time method of self-blast glass insulator location based on faster RCNN and U-net with aerial images," *CSEE J. Power Energy Syst.*, vol. 5, no. 4, pp. 474–482, Dec. 2019.
- [54] J. Xu, H. Liao, K. Li, et al, "Multi-scale feature fusion transformer with hybrid attention for insulator defect detection," *IEEE Trans. Instrum. Meas.*, vol. 74, 2025. (Note: Issue number and pages to be assigned).
- [55] S. Deng, L. Chen, and Y. He, "Insulator defect detection from aerial images in adverse weather conditions," *Appl. Intell.*, vol. 55, no. 6, p. 365, Apr. 2025.
- [56] J. Q. Zheng, H. Wen, M. Wang, et al, "Faster zero-shot multi-modal entity linking via visual-linguistic representation," *Data Intell.*, vol. 4, no. 3, pp. 493–508, Summer 2022.
- [57] W. Xu, X. Zhong, M. Luo, et al, "End-to-end insulator string defect detection in a complex background based on a deep learning model," *Front. Energy Res.*, vol. 10, p. 928162, Jul. 2022.
- [58] R. Wang, E. Dong, Z. Cheng, et al, "Transformer-based intelligent fault diagnosis methods of mechanical equipment: A survey," *Open Phys.*, vol. 22, no. 1, p. 20240015, Jan. 2024.
- [59] N. Shazeer, A. Mirhoseini, K. Maziarz, et al, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017.
- [60] Y. Zhou, T. Lei, H. Liu, et al, "Mixture-of-experts with expert choice routing," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, New Orleans, LA, USA, Nov. 2022, pp. 7103–7114.
- [61] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024.
- [62] Q. Zheng, H. Wen, M. Wang, et al, "Visual entity linking via multi-modal learning," *Data Intell.*, vol. 4, no. 1, pp. 1–19, Winter 2022.
- [63] J. Yang, K. Chen, K. Ding, et al, "Auto insurance fraud detection with multimodal learning," *Data Intell.*, vol. 5, no. 2, pp. 388–412, Summer 2023.
- [64] Q. Zhang, M. Chen, A. Bukharin, et al, "AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023.
- [65] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, New Orleans, LA, USA, Dec. 2023.

AUTHOR BIOGRAPHY

Tianyi Zhang, Master's degree, Engineer. His main research interests include intelligent operation and maintenance technology for power transmission and transformation, image recognition of transmission lines, and big data analysis in the power industry.

Ru Liu, Bachelor's degree, Senior Engineer. Her main research interests include power system automation, artificial intelligence, and big data analysis.

Xiaolong Wang, Master's degree, Senior Engineer. His main research interests include intelligent operation and maintenance of transmission lines, artificial intelligence, and big data analysis.