Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Minor Surface Defect Detection for Power Line Insulators UsingFrequency-Space Fusion and Focused Diffusion Pyramids

xiwei lin
Changchun University of Technology
xiu ji
jixiu523@ccit.edu.cn
Changchun Institute of Technology
sijia liu
Changchun University of Technology
yiqi liu
Changchun Institute of Technology
huan cheng
Changchun Institute of Technology

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11033446/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Minor Surface Defect Detection for Power Line Insulators Using Frequency-Space Fusion and Focused Diffusion Pyramids

Xiwei Lin[1], Xiu Ji[2*], Sijia Liu[1], Yiqi Liu[2], Huan Cheng[2]

[1]School of Electrical and Electronic Engineering, Changchun University of Technology, Changchun, 130000, China.
[2]Future Industry Technology Innovation Research Institute, Changchun Institute of Technology, Changchun, 130000, China.

*Corresponding author(s). E-mail(s): jixiu523@ccit.edu.cn;

Abstract

To address the challenges of insulator defect detection in complex backgrounds and resource-constrained edge scenarios, this paper proposes a lightweight and high-precision detection method based on the RT-DETR framework. An efficient feature extraction network, FSNet, is designed to enhance feature representation through dynamic routing and adaptive fusion of spatial- and frequency-domain information. Furthermore, a Feature-Focused Diffusion Pyramid Network is introduced to improve multi-scale contextual interaction, thereby strengthening the localization and classification capability for defects of varying sizes. Experimental results on a self-built insulator dataset demonstrate that the proposed method achieves a 4.8 percentage point improvement in mAP@0.5 over the baseline model, while reducing computational cost by 2.8%. In addition, extensive experiments on a public dataset further validate the strong generalization ability of the proposed method. Compared with several representative CNN-based and Transformer-based detectors, the proposed method consistently achieves a superior trade-off between detection accuracy and efficiency. Finally, edge deployment on the NVIDIA Jetson Nano platform confirms its feasibility and practical value for real-time UAV-based transmission line inspection.

Keywords: Transmission line insulators, Fault detection, Transformer architecture, Frequency-spatial feature fusion, Feature diffusion, Self-built dataset, Edge computing

1 Introduction

Transmission lines form the arterial network of modern power systems, and their reliable operation is crucial for ensuring grid safety and stability [1]. As key components providing electrical insulation and mechanical support, insulators are inevitably exposed to harsh outdoor environments over extended periods. Consequently, they are highly susceptible to degradation caused by ultraviolet radiation, temperature fluctuations, and mechanical stresses, leading to surface contamination, cracks, and damage [2, 3]. If not promptly detected and repaired, minor surface defects may escalate into catastrophic failures—such as insulator flashover and conductor breakage—resulting in substantial economic losses and societal disruption. Consequently, implementing regular and efficient insulator inspection mechanisms is imperative for achieving predictive maintenance.

Traditionally, transmission line inspections have heavily relied on manual visual inspections. This process is not only labor-intensive and inefficient but also carries high safety risks in complex terrains like mountains and forests [4]. In recent years, with the rapid advancement of unmanned aerial vehicle (UAV) technology, aerial inspections using drones have emerged as the mainstream solution due to their high maneuverability and extensive coverage capabilities [5, 6]. However, the massive volume of high-resolution images captured by drones poses significant challenges for manual interpretation, creating an urgent need for automated intelligent detection algorithms based on computer vision [7].

In the current landscape, the domain of object detection is principally governed by deep learning methodologies, which can be broadly bifurcated into Convolutional Neural Network (CNN) frameworks (such as the YOLO family) and the recently burgeoning Transformer-based models [8]."While CNN-based methods offer advantages in inference speed [9–13], they are inherently constrained by the local receptive fields of convolutional operations. Such an intrinsic restriction impedes the extraction of long-distance correlations, frequently leading to suboptimal detection results in environments with intricate backgrounds. Conversely, Transformer paradigms — notably DETR and its real-time evolution, RT-DETR — harness multi-head self-attention to capture holistic contextual information, thereby demonstrating superior robustness in complex inspection scenarios.Despite their outstanding accuracy, Transformer-based models face significant barriers to deployment on resource-constrained edge devices due to their massive parameter counts and high computational costs (FLOPs). In practical drone inspection scenarios, onboard computing platforms like the NVIDIA Jetson series operate under strict power constraints. Existing Transformer models often struggle to achieve real-time inference on such edge hardware, forcing researchers into difficult trade-offs between detection accuracy and model efficiency [14]. Furthermore, aerial images of power transmission lines exhibit extreme scale variations and diverse background features [15–17]. Lightweight models typically struggle to maintain semantic consistency across multi-scale features, leading to missed detections of small or blurred defects. Furthermore, the scarcity of high-quality public insulator defect datasets limits the validation of algorithm robustness in real-world, variable environments.

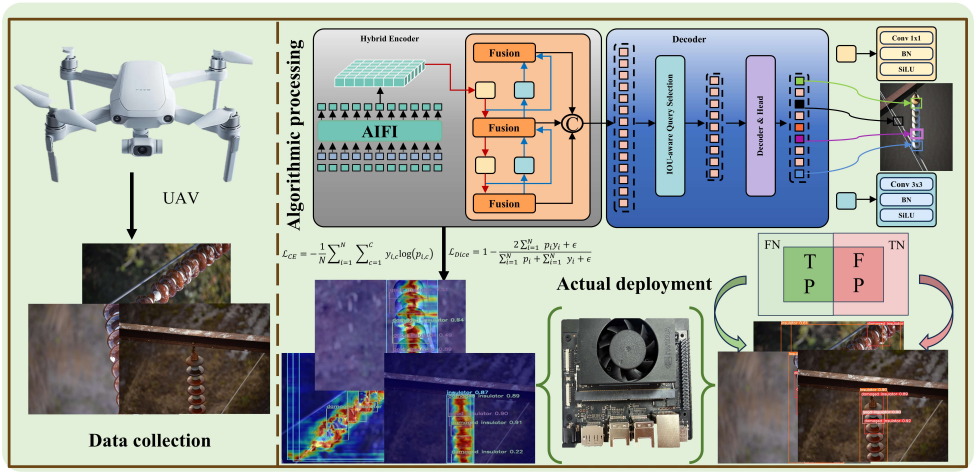

Fig. 1 Overall Workflow of This Paper.

A computationally efficient strategy for high-precision insulator inspection is proposed herein to bridge these gaps, with a focus on edge computing compatibility. The overall architecture, depicted in Figure 1, comprises data gathering, computational analysis, and final deployment on the edge. The main contributions are summarized as follows:

1) Designing the lightweight feature extraction backbone network FSNet based on novel Spatial-Frequency Splitting (SFS) convolutions. By dynamically splitting features between spatial and frequency domains, this architecture enables fine-grained perception and adaptive selection, significantly enhancing feature expression capabilities while drastically reducing parameter redundancy.

2) We propose the Focusing Diffusion Pyramid Network to address the challenge of multi-scale defect detection. Through a customized feature focusing module and diffusion mechanism, this network endows features at all scales with rich contextual description capabilities, significantly improving the detection performance of small and blurred defects in complex backgrounds.

3) A comprehensive insulator defect dataset was constructed, and the proposed model was successfully deployed on the NVIDIA Jetson Nano edge computing platform. Experimental results demonstrate that compared to state-of-the-art baseline models, our approach achieves an optimal trade-off between accuracy and speed, validating its efficiency and engineering value for real-time drone inspection tasks.

The subsequent structure of this paper is as follows: Section 2 reviews prior work on insulator defect detection and lightweight network design. Section 3 details the proposed method, including the specifics of FSNet and the Feature-Focused Diffusion Pyramid Network. Section 4 presents the experimental setup, comparative analysis, and ablation studies. Finally, Section 5 summarizes the paper and discusses future research directions.

2 Related Work

2.1 Transmission Line Insulator Inspection

Early methods for detecting insulator defects primarily relied on physical sensing technologies. For instance, Mei et al. [18] employed terahertz imaging combined with edge detection algorithms to identify internal and external defects in composite insulators, effectively overcoming limitations imposed by optical illumination conditions. Similarly, Wu et al. [19] proposed a multi-information fusion approach to assess the condition of epoxy insulators in gas-insulated switchgear (GIS), integrating partial discharge signals with visual data. While these non-optical methods offer high reliability, they typically require expensive and bulky equipment, making them unsuitable for flexible drone inspections. With the rise of computer vision, deep learning has become the standard approach for analyzing aerial images of power components [20]. Convolutional neural networks (CNNs) are widely adopted due to their robust local feature extraction capabilities. Qingyi and Bo [21] developed a transfer learning model for concrete crack detection, demonstrating the effectiveness of pre-trained networks in infrastructure monitoring. To address complex background challenges, Liu et al. [22] enhanced the YOLOv8 model with SLF-YOLO for metal surface defect detection, improving feature fusion efficiency. Recently, Ji et al. [23] applied an improved YOLOv10 algorithm to infrared image classification of power equipment, achieving a balance between speed and accuracy. However, despite continuous evolution in CNN architectures, their inherent inability to model long-range global dependencies limits their performance in detecting defects requiring understanding of the overall context of insulator strings..

2.2 Real-Time Transformer Object Detection

The emergence of visual Transformers revolutionized object detection by introducing self-attention mechanisms. Notably, Zhao et al. [24] and Lv et al. [25] successfully addressed the high computational latency of traditional DETR models, demonstrating that Transformers can outperform YOLO variants in both speed and accuracy. Following this breakthrough, numerous lightweight variants tailored for specific domains have emerged. In the field of drone aerial photography, Wang et al. [26] proposed PHSI-RTDETR for infrared small object detection, incorporating a patch-based high-frequency suppression module to reduce noise. Similarly, for agricultural monitoring, Wang et al. [27] adapted RT-DETR for tomato ripeness detection, achieving lightweight deployment through backbone network optimization. Ma et al. [28] introduced OWRT-DETR for open-water search and rescue, enhancing small target recall through a novel query selection mechanism. To refine feature representations, Zhou et al. [29] integrated spatial residuals into the detection framework, significantly improving underwater target localization. In autonomous driving, Zhang et al. [30] designed TSD-DETR for long-range traffic sign detection, emphasizing the importance of multi-scale feature interactions. To tackle the challenge of extreme scale inconsistency in UAV imagery, Wang et al. [31] proposed Amfef-DETR, which integrates an end-to-end adaptive strategy for multi-level feature extraction. In the power sector, Ji et al. [32]

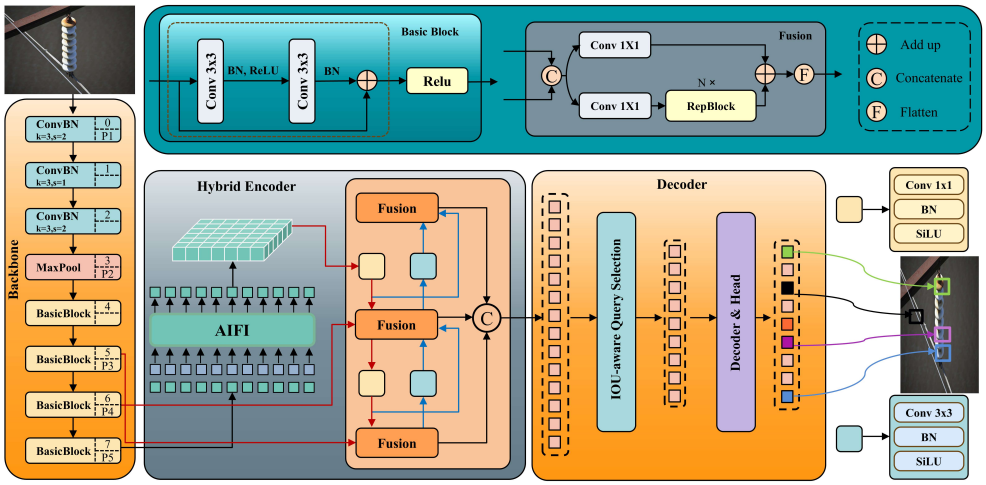

Fig. 2 Schematic diagram of the standard RT-DETR architecture. This architecture comprises a ResNet backbone based on BasicBlocks for feature extraction, a hybrid encoder combining AIFI and a fusion module based on RepBlocks, and a Transformer decoder for end-to-end prediction.

proposed SCAD-DETR for switchgear equipment, validating Transformers' potential in industrial inspections. Despite these advances, existing RT-DETR variants typically retain high computational overhead or fail to adequately preserve high-frequency edge details of insulators during subsampling. Consequently, developing a lightweight Transformer architecture sensitive to multi-scale spatio-temporal-frequency features remains a critical research gap.

3 Methods

3.1 RT-DETR Model

The Real-Time Detection Transformer (RT-DETR) marks a pivotal advancement in object detection , effectively reconciling the trade-off between the inference velocity of CNN-based YOLO models and the superior precision of Transformer paradigms. As depicted in Figure 2, the framework is functionally divided into three core modules: a backbone (e.g., ResNet or HGNet) for extracting primary features, a hybrid encoder for integrating multi-scale features, and a Transformer decoder equipped with auxiliary heads.

Distinct from conventional DETR models, which are often plagued by sluggish convergence and heavy computational burdens, RT-DETR incorporates a streamlined hybrid encoder. By disentangling intra-scale interactions from cross-scale fusion, this design facilitates efficient multi-scale feature processing while preserving global context awareness. Nevertheless, the substantial parameters and floating-point operations (FLOPs) inherent in the standard backbone remain prohibitive for direct deployment on resource-limited edge platforms, such as drones used for insulator inspection.

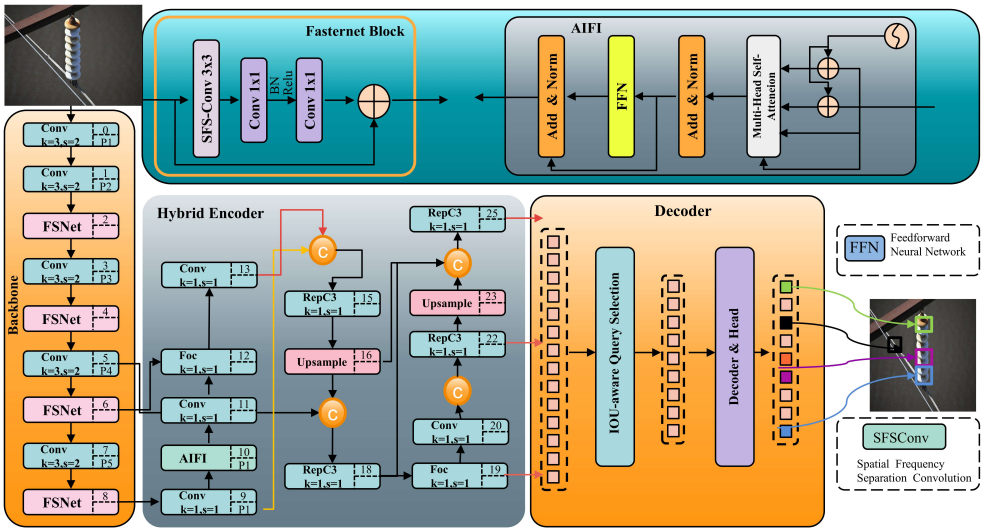

Fig. 3 Proposed detection network architecture. Includes feature extraction, encoding, decoding, and detection head architecture.

3.2 Overview of the Proposed Framework

Although RT-DETR achieves a good balance between real-time performance and accuracy, it still faces challenges in extracting subtle defect features and meeting lightweight model requirements when applied to the specific scenario of transmission line insulators. In response to these constraints, we present a refined object detection framework built upon the RT-DETR architecture. This proposed model synergizes spatial-frequency separation convolution with a reparameterization strategy, as depicted in the schematic in Figure 3. In the backbone network section, to more efficiently extract insulator texture and edge features, this paper replaces the original backbone structure with FSNet. As shown in the left and upper-left details of Figure 3, FSNet is composed of multiple stacked Fasternet Blocks. The core innovation of this module lies in introducing SFS-Conv (Spatial Frequency Separation Convolution). SFS-Conv spatially separates feature maps into frequency components, processing high-frequency details (e.g., insulator cracks) and low-frequency backgrounds independently. Feature reconstruction is then achieved through 1x1 convolutions and residual connections. This configuration streamlines the computational process while amplifying the network's sensitivity to subtle anomalies, resulting in the generation of multi-scale feature maps (P3, P4, P5).

Whereas the Hybrid Encoder preserves the AIFI module to ensure global context modeling for high-level features (P5), the multi-scale fusion component undergoes a substantial transformation. As depicted in the central region of Figure 3, we integrate the RepC3 and Focus (Foc) modules to expedite inference speeds without compromising feature representational power.The RepC3 module employs structural reparameterization techniques, utilizing a multi-branch architecture during training

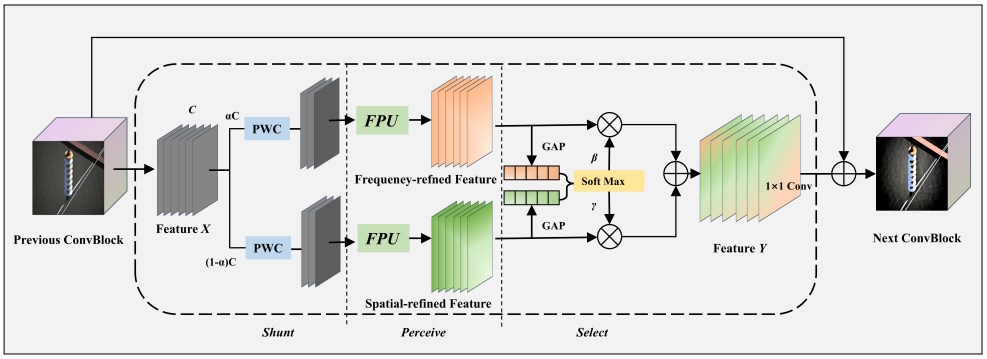

Fig. 4 Detailed schematic diagram of the key component SFS-Conv. This module achieves spatial-frequency feature decoupling and adaptive fusion through a "Shunt-Perceive-Select" mechanism.

to enrich the feature space, then converging to a single-path convolution during inference for computational acceleration. Concurrently, the Focus module is applied during downsampling, reducing information loss through slicing operations. Deep features processed by AIFI are cross-scale fused (Concatenate) with shallow features (P3, P4) through upsampling, Focus operations, and the RepC3 module, forming a more expressive multi-scale feature pyramid. Finally, the fused features are fed into the decoder. As shown on the right side of Figure 3, the decoder continues the IOU-aware Query Selection mechanism to filter high-confidence query vectors, which are then input into the detection head (Decoder & Head). This component combines the enhanced features extracted by preceding modules with a multi-layer perceptron and attention mechanism to ultimately output the precise location and category information of insulators and their defects, achieving high-precision real-time detection in complex backgrounds.

3.3 Efficient Feature Extraction Backbone Network

To curtail computational overhead without compromising detection precision, we engineer the FSNet Block and embed it within the backbone architecture.The core of this module lies in its Spatial Frequency Separation mechanism, which effectively decouples high-frequency detail features (e.g., cracks, defect edges) from low-frequency semantic features (e.g., background, insulator body) in insulator images. Its detailed structure and internal operational mechanism are illustrated in Figure 4.

The overall design of the FSNet Block follows the concept of residual architecture. As shown on the left side of Fig. 4, the input features first undergo core feature extraction via SFS-Conv (Spatial Frequency Separation Convolution). Subsequently, they sequentially pass through a 1×1 convolution for channel fusion, batch normalization (BN), and ReLU activation. Finally, another 1×1 convolution is applied and added to the input through a residual connection.

SFS-Conv represents the core innovation of the module, employing a three-stage "Shunt–Perceive–Select" strategy internally, as illustrated in the right panel of Fig. 4.

During the *Shunt* phase, assuming the input feature map is $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, SFS-Conv first divides it proportionally along the channel dimension by a ratio α, yielding

two sub-feature groups:

$$\mathbf{X}_{\text{split}} = [\mathbf{X}_{\text{freq}}, \mathbf{X}_{\text{spa}}], \tag{1}$$

where $\mathbf{X}_{\text{freq}} \in \mathbb{R}^{\alpha C \times H \times W}$ captures frequency-domain details, while $\mathbf{X}_{\text{spa}} \in \mathbb{R}^{(1-\alpha)C \times H \times W}$ preserves spatial structural information.

Subsequently, in the *Perceive* phase, each branch is processed by a dedicated Feature Processing Unit (FPU) to generate the frequency-refined features $\mathbf{F}_{\text{freq}}$ and spatial-refined features $\mathbf{F}_{\text{spa}}$. To adaptively aggregate these two types of features, the *Select* phase introduces an attention mechanism. First, the spatial information is compressed by Global Average Pooling (GAP) to produce a global context descriptor $\mathbf{S}$:

$$\begin{aligned} \mathbf{S} =& \text{GAP}([\mathbf{F}_{\text{freq}}, \mathbf{F}_{\text{spa}}]) \\ =& \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} [\mathbf{F}_{\text{freq}}(i,j), \mathbf{F}_{\text{spa}}(i,j)]. \end{aligned} \tag{2}$$

Then, a Softmax function is applied to generate adaptive weight vectors β and γ corresponding to the frequency and spatial branches, respectively, satisfying $\beta + \gamma = 1$. The final output feature $\mathbf{Y}$ is obtained by applying a 1×1 convolution to the weighted fusion:

$$\mathbf{Y} = \text{Conv}_{1\times1}(\beta \cdot \mathbf{F}_{\text{freq}} + \gamma \cdot \mathbf{F}_{\text{spa}}). \tag{3}$$

Through this mechanism, the FSNet Block not only effectively reduces redundant computations but also significantly enhances the model's sensitivity to minute defects (high-frequency components) in insulators, achieving a favorable balance between lightweight design and high performance.

3.4 Feature Focusing Diffusion Pyramid Network

We introduce the Feature Focusing Module as the integral pillar of the FocusingDiffusion Pyramid Network (visualized in Figure 5). Its primary objective isto bolster multi-scale feature representation, thereby effectively counteracting the loss of contextual data often encountered in clutteredenvironments.

The design of the Feature Focusing Module (FFM) aims to achieve deep "focusing" and "diffusion" of features through forced alignment of multi-scale features and parallel extraction of multi-receptive fields. First, to address the spatial inconsistency of features across different levels, the module receives three-scale features $\mathbf{X} = \{\mathbf{X}_{P3}, \mathbf{X}_{P4}, \mathbf{X}_{P5}\}$ from the backbone network. To fuse this information within a unified feature space, the module downsamples the high-resolution feature $\mathbf{X}_{P3}$ using the ADown operator and upsamples the low-resolution semantic feature $\mathbf{X}_{P5}$. Both are then uniformly mapped to the intermediate scale of $\mathbf{X}_{P4}$. The aligned features are concatenated along the channel dimension to form the aggregated feature $\mathbf{X}_{\text{align}}$, which is rich in hierarchical information. The computation is expressed as:

$$\begin{aligned} \mathbf{X}_{\text{align}} =& \text{Concat}\big(\Phi_{\text{down}}(\mathbf{X}_{P3}),\ \Phi_{\text{id}}(\mathbf{X}_{P4}), \\ & \Phi_{\text{up}}(\mathbf{X}_{P5})\big), \end{aligned} \tag{4}$$

where Φ_{down}, Φ_{id}, and Φ_{up} denote the downsampling, identity mapping, and upsampling operations, respectively.

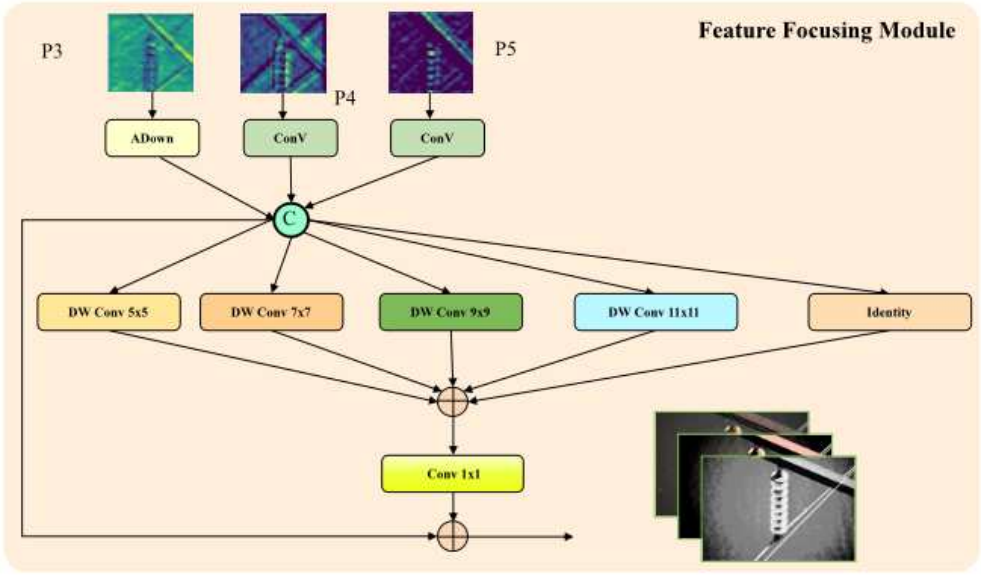

Fig. 5 Schematic diagram of the Feature Focusing Module. This module effectively captures and aggregates contextual information from different receptive fields through multi-scale feature alignment, stitching, and multi-core parallel deep convolutions.

Building upon the feature alignment, to simulate the human visual "focus" mechanism and capture contextual information across varying scales, the module incorporates a parallel depthwise convolutional block analogous to the Inception architecture. Specifically, a kernel set $\mathbf{K} = \{5, 7, 9, 11\}$ is defined, and depthwise convolutions (DWConv) are employed to extract features in parallel. Notably, the module adopts a comprehensive aggregation strategy, in which the features extracted from each branch are stacked and summed together with the original aligned feature $\mathbf{X}_{\text{align}}$. This design constructs an implicit multi-residual structure that preserves the original information while integrating multi-scale contextual features. The aggregated feature $\mathbf{F}_{\text{agg}}$ is computed as:

$$\mathbf{F}_{\text{agg}} = \mathbf{X}_{\text{align}} + \sum_{k \in \mathbf{K}} \text{DWConv}_{k \times k}(\mathbf{X}_{\text{align}}). \tag{5}$$

Finally, the aggregated feature $\mathbf{F}_{\text{agg}}$ is fed into a 1×1 pointwise convolution (PWConv) layer to enable inter-channel information exchange and feature dimensionality reduction. To further facilitate gradient backpropagation and prevent feature degradation in deep networks, the module introduces a final residual connection, adding the transformed feature back to the initial aligned feature $\mathbf{X}_{\text{align}}$. The final enhanced feature $\mathbf{Y}$ is obtained as:

$$\mathbf{Y} = \mathbf{X}_{\text{align}} + \text{PWConv}(\mathbf{F}_{\text{agg}}). \tag{6}$$

Through this series of transformations, the Feature Focusing Module effectively propagates cross-scale contextual information to all levels required for detection, significantly enhancing the model's capability to characterize insulators and their defects. The specific forward propagation process of the Feature Focusing Module is summarized in Algorithm 1.

Algorithm 1 Feature Focusing Module (FFM) Process

Require: Multi-scale features $\mathbf{X} = \{X_{P3}, X_{P4}, X_{P5}\}$; Kernel set $\mathbf{K} = \{5, 7, 9, 11\}$
Ensure: Enhanced feature $\mathbf{Y}$

Multi-Scale Alignment
$X_{\text{align}} \leftarrow \text{Concat}(\text{ADown}(X_{P3}), \text{Conv}(X_{P4}), \text{Upsample}(X_{P5}))$
Focusing & Aggregation
$F_{\text{agg}} \leftarrow X_{\text{align}} + \sum_{k \in K} \text{DWConv}_{k \times k}(X_{\text{align}})$
Refinement
$\mathbf{Y} \leftarrow X_{\text{align}} + \text{PWConv}(F_{\text{agg}})$
return $\mathbf{Y}$

4 Experiments

4.1 Experimental Environment and Datasets

The experimental hardware specifications are listed in Table 1 below. Both training and validation curves have been smoothed to enhance visual clarity. To evaluate

Table 1 Experimental Equipment Models and Parameters

Lab environment	Environment configuration
Operating system	Windows 10
CPU version	Intel(R) Core(TM) i7-7800X
GPU version	NVIDIA GeForce RX 2080 × 3
RAM	32G
Deep learning framework	Pytorch
Python version	3.10
CUDA version	11.1

the effectiveness and generalization capability of the proposed method in practical inspection scenarios, experiments were conducted on two insulator datasets: a self-constructed aerial inspection dataset and a publicly available dataset obtained from Roboflow Universe [33].

The self-constructed dataset was developed using real aerial images collected by DJI unmanned aerial vehicles along multiple transmission-line corridors. These images cover diverse backgrounds, viewing angles, object scales, illumination conditions, and operating environments commonly encountered in power-line inspection tasks. The

dataset contains 3,069 high-resolution images, all of which were manually annotated in YOLO format. According to the inspection targets and defect conditions, the annotated objects were divided into three categories: insulator strings, intact insulators, and damaged insulators. Insulator strings refer to complete insulator assemblies, intact insulators represent individual components without visible defects, and damaged insulators correspond to defective units requiring inspection and maintenance. The dataset was randomly divided into training, validation, and test sets at a ratio of 8:1:1. This division provides sufficient samples for model optimization while retaining independent subsets for hyperparameter tuning and final performance evaluation.

To further evaluate the robustness and generalization performance of the proposed method, a publicly available insulator dataset from Roboflow Universe was also employed. This dataset contains 8,692 annotated images acquired under diverse scenes and imaging conditions. According to its predefined partition, 8,289 images were assigned to the training set, 275 images to the validation set, and 128 images to the test set. Compared with the self-constructed dataset, the public dataset provides a larger number of samples and greater visual diversity. Therefore, experiments on this dataset can further examine the adaptability of the proposed method to variations in background complexity, insulator appearance, object scale, viewpoint, and image quality. The use of both the self-constructed and public datasets enables a more comprehensive evaluation of the model in terms of detection accuracy, robustness, and cross-scene generalization.

4.2 Parameter Settings and Evaluation Metrics

To ensure model convergence and optimal performance, specific hyperparameters were tuned for our dataset. The batch size was set to 16, with 300 epochs using the AdamW optimizer. Initial parameters included a learning rate of 0.0001 and weight decay of 0.0001. To stabilize training, a cosine annealing strategy was employed with linear warmup during the first 2,000 iterations. The backbone network's learning rate was set to 0.1 times that of the head network to preserve pre-trained features. Additionally, an exponential moving average (EMA) with a decay rate of 0.9999 was applied to smooth model weights and enhance generalization. To visually demonstrate training stability and convergence speed, we recorded loss changes during training. The efficacy of the employed optimization strategy is substantiated by the comparisons in Figure 6, where our method demonstrates accelerated convergence and reduced terminal loss relative to the baseline model.

To comprehensively evaluate the performance of the proposed model in insulator defect detection, four core metrics are adopted: Precision, Recall, F1-Score, and Mean Average Precision (mAP).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{8}$$

$$F1_{\text{score}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

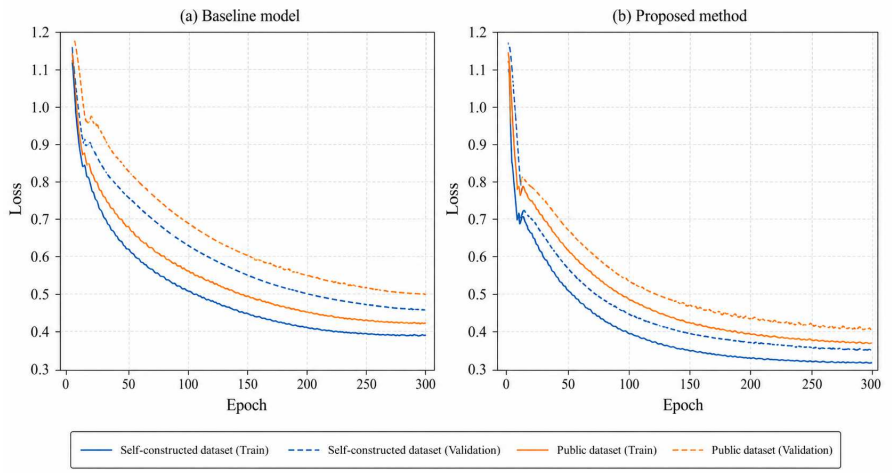

Fig. 6 (a) Baseline model and (b) Proposed method. The proposed method achieves faster convergence and lower final loss compared with the baseline model, which exhibits slower convergence and higher loss values.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} (\text{Precision}_i \times \text{Recall}_i) \tag{10}$$

Precision (Eq. 7) measures the accuracy of predictions, representing the proportion of correctly detected positive samples among all predicted positives. Recall (Eq. 8) reflects the model's ability to detect all true positives, representing the proportion of correctly detected positives among all actual positives. Given that these two metrics are often mutually constraining, we compute the F1 score (Eq. 9) as their harmonic mean to provide a comprehensive evaluation of model robustness. To quantify the holistic detection efficacy spanning multiple classes, this study leverages mAP (Eq. 10)—calculated as the arithmetic mean of Average Precision (AP) across the total N categories—as the cardinal metric for performance assessment.

4.3 Analysis of Experimental Results

We carried out thorough ablation analyses on an in-house dataset to quantitatively assess the impact of the introduced optimization strategy.Using the standard RT-DETR-r18 as the baseline model, we progressively integrated key components—namely the FSNet Block for lightweight feature extraction and the Feature Focusing Module (FFM) for multi-scale context enhancement. This stepwise analysis enabled us to quantify each module's specific contribution toward balancing detection accuracy and inference efficiency. Detailed comparative results are presented in Table 3.

4.3.1 Ablation Experiments

Table 2 presents the quantitative results of the ablation study. The baseline RT-DETR-r18 provides a competitive starting point, but its feature representation capability and efficiency are still constrained when applied to insulator defect detection under complex backgrounds. After introducing the FSNet Block, the model complexity is reduced while the inference efficiency is noticeably improved, indicating that the redesigned backbone can extract more discriminative features with a more compact architecture. This result demonstrates that the improvement does not rely on increasing model scale, but rather on a more efficient feature extraction strategy.

On this basis, the introduction of the FFM further enhances the detection performance, especially in terms of overall prediction consistency. This suggests that the module effectively strengthens contextual interaction and multi-scale feature aggregation, enabling the detector to better distinguish subtle defect regions from visually similar background interference. In other words, the FFM complements the lightweight backbone by compensating for the contextual information that is often weakened in compact models. When the FSNet Block and FFM are jointly integrated, the proposed method achieves the best overall performance among all ablation settings. The final model not only improves detection accuracy relative to the baseline, but also maintains lower computational cost and stronger real-time capability. This confirms that the two proposed components are not isolated improvements; instead, they form a mutually reinforcing design, in which the backbone enhances feature extraction efficiency and the neck strengthens feature diffusion and contextual perception. As a result, the proposed method achieves a more favorable balance between accuracy, complexity, and inference speed, which is particularly important for practical UAV-based insulator inspection tasks.

Table 2 Ablation Experiment Results

Model	Size	Param. (M)	F1	GFLOPs	mAP@0.5	FPS (F/s)
RT-DETR-r18	640	19.9	0.79	57.0	0.864	20.6
+ FSNet Block	640	19.0	0.85	56.3	0.885	26.0
+ FFM	640	19.0	0.86	56.9	0.890	27.1
Proposed Method	**640**	**18.6**	**0.88**	**55.4**	**0.912**	**31.2**

4.3.2 Comparative Experiments

To further evaluate the effectiveness of the proposed method, we conduct comparative experiments against several representative object detection models on the test set, including both CNN-based and Transformer-based detectors. The quantitative results are summarized in Table 3, which additionally reports performance on a public benchmark dataset to further validate the generalization capability of different methods.

Overall, the proposed method achieves the most favorable trade-off among detection accuracy, inference efficiency, and model complexity, demonstrating its strong

suitability for practical defect inspection tasks. From the perspective of detector families, CNN-based methods generally exhibit relatively low computational overhead and compact model size, making them attractive for deployment on resource-constrained platforms. However, their detection performance is still limited when dealing with defect patterns characterized by large appearance variations, subtle local textures, and complex backgrounds. This indicates that conventional convolutional architectures, while efficient, remain insufficient for capturing long-range dependencies and discriminative contextual information that are crucial for challenging insulator defect detection scenarios.

Table 3 Comparison of different detection models on two datasets

Model	Type	Dataset	Param. (M)	GFLOPs	mAP@0.5	FPS
SSD [34]	CNN	Self-constructed	26.3	32.5	0.802	28.5
		Public			0.752	19.9
YOLOv8s [35]	CNN	Self-constructed	11.2	28.5	0.848	25.0
		Public			0.763	18.0
DINO-4scale [36]	Transformer	Self-constructed	47.0	185.0	0.881	30.0
		Public			0.793	23.0
Co-DETR [37]	Transformer	Self-constructed	108.0	220.5	0.875	28.6
		Public			0.760	22.1
RT-DETR-R34 [33]	Transformer	Self-constructed	31.8	84.0	0.879	21.0
		Public			0.780	19.7
RT-DETR-r18 [33]	Transformer	Self-constructed	19.9	57.0	0.864	20.6
		Public			0.796	24.0
Proposed Method	**Transformer**	Self-constructed	**18.6**	**55.4**	**0.912**	**31.2**
		Public			**0.816**	**28.0**

4.3.3 Deployment Experiment

In contrast, Transformer-based detectors demonstrate stronger global modeling capability and generally achieve higher accuracy. Nevertheless, this often comes at the cost of increased computational complexity and reduced real-time performance. Notably, the proposed method effectively mitigates this trade-off by maintaining high detection accuracy while keeping the model lightweight and efficient. It achieves the best performance in terms of both mAP and FPS across the two datasets, highlighting its robustness and practical applicability. For UAV-based inspection scenarios, where both response speed and onboard computing resources are critical, such heavy architectures are not always the most practical choice, even if their accuracy is competitive. Compared with these methods, the proposed approach exhibits a more balanced design philosophy. On the one hand, it preserves the strong feature modeling capability of Transformer-based detection, allowing it to achieve the best overall detection performance among all compared methods. On the other hand, it avoids the excessive complexity commonly associated with heavyweight Transformer detectors, thereby maintaining favorable inference efficiency and a relatively lightweight model size. This indicates that the proposed improvements do not merely increase model capacity, but

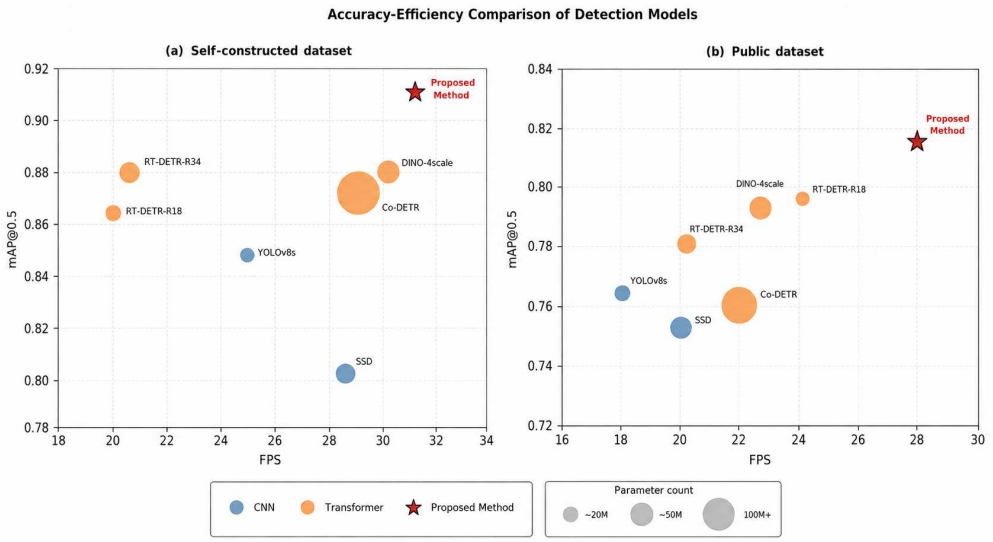

Fig. 7 Comparison of inference speed (FPS) and accuracy (mAP) across different models. Bubble size corresponds to model parameter count. Compared to other SOTA detectors, the proposed method (red asterisk) achieves the optimal trade-off, delivering outstanding accuracy and real-time performance($>$ 30 FPS)on Jetson Nano.

instead enhance the effectiveness of feature representation and information utilization in a more targeted and efficient manner. A particularly important observation is that the proposed method outperforms both baseline RT-DETR variants. Compared with the lighter baseline, the proposed model achieves better detection performance while retaining real-time inference capability, indicating that the introduced modifications effectively improve feature quality and localization precision without sacrificing efficiency. Meanwhile, compared with the larger baseline, the proposed method still delivers superior overall performance despite using fewer parameters, which further demonstrates that performance gains are not simply derived from increasing model scale, but from a more efficient architectural design.

Table 4 Model Performance Comparison on Jetson Nano

Model	**Latency (ms)** $\downarrow$	**Power (W)** $\downarrow$	**Temp. (°C)** $\downarrow$	**Efficiency (FPS/W)** $\uparrow$
RT-DETR-r18	48.5	8.2	64.5	2.51
YOLOv8s	23.5	7.5	61.0	3.66
Proposed Method	**32.1**	**6.4**	**58.2**	**4.88**

To rigorously evaluate the feasibility of the proposed method in battery-powered UAV missions, we conducted onboard deployment tests on a NVIDIA Jetson Nano (8GB) operating in MAXN (10W) mode, monitoring real-time performance metrics

using the jtop tool. As shown in Table 4, the experimental results demonstrate the model's outstanding engineering value. Specifically, our method achieves an average inference latency of just 32.1 ms, 16.4 ms lower than the baseline RT-DETR-r18 (48.5 ms). This low latency ensures the UAV can process data and respond to detected defects within consecutive video frames, meeting the stringent demands of high-speed aerial inspections. Furthermore, thanks to FSNet's efficient frequency-spatial splitting mechanism, GPU computational load is significantly reduced, lowering average power consumption to 6.4 W—1.8 W less than the baseline. To quantify the model's "green" performance, we introduce the energy efficiency ratio (FPS/W). Experiments demonstrate that our method achieves an energy efficiency ratio of 4.88 FPS/W, significantly outperforming both the baseline (2.51 FPS/W) and the widely used YOLOv8s. This confirms that our lightweight design strategy successfully maximizes performance per unit energy, making it an ideal solution for resource-constrained edge computing scenarios. Table 4 Performance Comparison on NVIDIA Jetson Nano Platform.

4.4 Visualization of Experimental Results

Aiming to provide visual evidence of the model's robustness against challenging backdrops, we illustrated detection examples and employed Grad-CAM to visualize the regions of interest (feature attention maps).Figure 8 presents a qualitative comparison between the baseline model (RT-DETR-r18) and our method under various challenging conditions, including blurred targets, occlusions, and complex background interference.

The detection results (rows 2 and 4) reveal that the baseline model suffers from missed detections and low confidence scores when encountering minute defects or strong background noise. In contrast, our method accurately localizes these challenging targets with tighter bounding boxes and significantly higher confidence scores. Furthermore, the heatmaps (Rows 3 and 5) provide deeper insights into the model's attention mechanism. The baseline model's attention tends to be dispersed, showing high activation on irrelevant background elements such as transmission towers and vegetation. In contrast, thanks to the Feature Focused Diffusion Pyramid Network (FFDPN), our method exhibits highly concentrated attention. The activation regions are strictly aligned with insulator defects, effectively suppressing background interference. This visual evidence conclusively validates that the proposed "focusing" and "diffusion" mechanisms successfully enhance the model's ability to characterize key targets.

5 Conclusion

This paper proposes a lightweight and high-precision insulator defect detection method based on the RT-DETR framework to address the challenges of complex backgrounds and resource-constrained edge deployment in UAV-based transmission line inspection. To enhance feature representation capability, a Frequency-Space Split Network (FSNet) is designed to adaptively integrate spatial-domain and frequency-domain information through dynamic routing and feature fusion. In addition, a Feature-Focused Diffusion Pyramid Network is introduced to strengthen multi-scale contextual

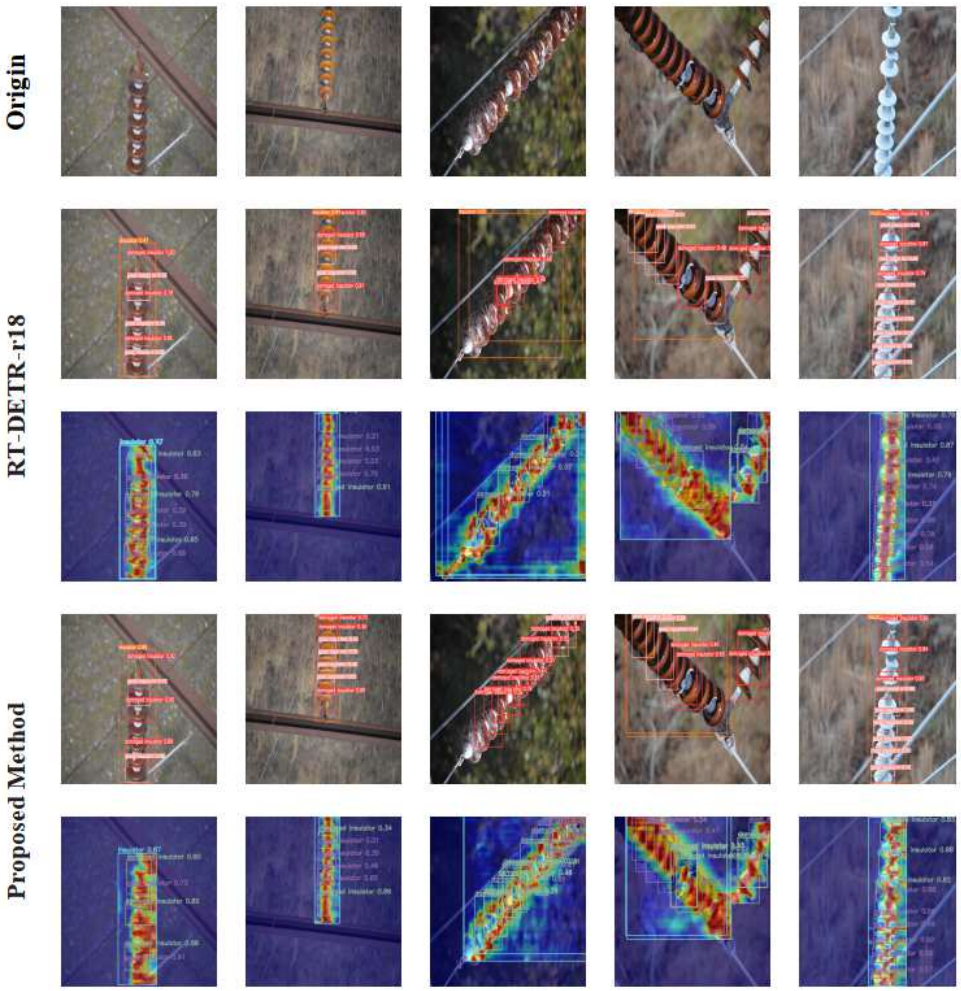

Fig. 8 Visual comparison of detection results with Grad-CAM heatmaps. Row 1: Original image; Rows 2-3: Detection and heatmap of RT-DETR-r18; Rows 4-5: Detection and heatmap of the proposed method. Warm colors indicate high attention.

interaction, thereby improving the model's ability to localize and classify defects of different sizes in complex scenes.

Extensive experiments conducted on a self-built insulator dataset demonstrate that the proposed method achieves a 4.8 percentage point improvement in mAP@0.5 compared with the baseline model, while reducing computational cost by 2.8%. Furthermore, experiments on a public dataset further verify the strong generalization capability of the proposed method across different scenarios. Compared with several representative CNN-based and Transformer-based detectors, the proposed method consistently achieves a superior trade-off between detection accuracy and efficiency.

Finally, deployment results on the NVIDIA Jetson Nano platform confirm that the proposed method can achieve real-time inference under edge computing constraints, demonstrating its feasibility and practical value for UAV-based intelligent inspection tasks.

Conflicts of interest

The authors declare that they have no competing interests.

Funding

This work was supported by the Science and Technology Project of State Grid Jilin Electric Power Co., LTD. (Grant No. SGJLCG00YJJS2400152).

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author contributions statement

X.W.L. and X.J. conceived the methodology and designed the framework. X.W.L. and S.J.L. constructed the dataset and conducted the experiments. Y.Q.L. and H.C. performed data curation and validation. X.W.L. analyzed the results and wrote the original draft. X.J. reviewed and edited the manuscript and supervised the project. All authors read and approved the final manuscript.

Disclosures

The authors declare that they have no financial or personal relationships with other people or organizations that could inappropriately influence (bias) their work. There are no professional or commercial interests of any nature or kind in any product, service, and/or company that could be construed as a potential conflict of interest regarding the manuscript entitled "Minor Surface Defect Detection for Power Line Insulators Using Frequency-Space Fusion and Focused Diffusion Pyramids".

Author Contributions

Xiwei Lin: Conceptualization, Methodology, Software, Validation, Writing - original draft. Xiu Ji: Formal analysis, Investigation, Resources, Data curation, Writing - review & editing. Sijia Liu: Visualization, Software, Validation. Yiqi Liu: Investigation, Resources, Data curation. Huan Cheng: Project administration, Supervision, Funding acquisition, Writing - review & editing.

References

[1] Hu, Y., Liu, K.: Inspection and Monitoring Technologies of Transmission Lines with Remote Sensing, (2017)

[2] Liu, Y., Guo, Y., Fan, Y., *et al.*: Optical imaging technology application in transmission line insulator monitoring: A review. IEEE Transactions on Dielectrics and Electrical Insulation **31**(6), 3120–3132 (2024)

[3] Maraaba, L., Al-Soufi, K., Ssennoga, T., *et al.*: Contamination level monitoring techniques for high-voltage insulators: a review. Energies **15**(20), 7656 (2022)

[4] Tao, X., Zhang, D., Wang, Z., *et al.*: Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. IEEE Transactions on Systems, Man, and Cybernetics: Systems **50**(4), 1486–1498 (2018)

[5] Yang, L., Fan, J., Song, S., *et al.*: A light defect detection algorithm of power insulators from aerial images for power inspection. Neural Computing and Applications **34**(20), 17951–17961 (2022)

[6] Panigrahy, S., Karmakar, S.: Real-time condition monitoring of transmission line insulators using the yolo object detection model with a uav. IEEE Transactions on Instrumentation and Measurement **73**, 1–9 (2024)

[7] Ji, X., Yue, Z., Yang, H., Li, M., Han, H.: Scad-detr: An infrared image detection method for switchgear equipment. IEEE Internet of Things Journal **12**(15), 30676–30697 (2025) https://doi.org/10.1109/JIOT.2025.3571499

[8] Liu, Y., Liu, D., Huang, X., *et al.*: Insulator defect detection with deep learning: A survey. IET Generation, Transmission & Distribution **17**(16), 3541–3558 (2023)

[9] Qiu, Z., Zhu, X., Liao, C., *et al.*: Detection of transmission line insulator defects based on an improved lightweight yolov4 model. Applied Sciences **12**(3), 1207 (2022)

[10] Wei, L., Jin, J., Deng, K., *et al.*: Insulator defect detection in transmission line based on an improved lightweight yolov5s algorithm. Electric Power Systems Research **233**, 110464 (2024)

[11] Zhou, M., Li, B., Wang, J., *et al.*: Fault detection method of glass insulator aerial image based on the improved yolov5. IEEE Transactions on Instrumentation and Measurement **72**, 1–10 (2023)

[12] Han, G., He, M., Zhao, F., *et al.*: Insulator detection and damage identification based on improved lightweight yolov4 network. Energy Reports **7**, 187–197 (2021)

[13] Zheng, J., Wu, H., Zhang, H., *et al.*: Insulator-defect detection algorithm based on improved yolov7. Sensors **22**(22), 8801 (2022)

[14] Li, D., Lu, Y., Gao, Q., *et al.*: Liteyolo-id: A lightweight object detection network for insulator defect detection. IEEE Transactions on Instrumentation and Measurement **73**, 1–12 (2024)

[15] Liu, C., Wu, Y., Liu, J., *et al.*: Mti-yolo: A light-weight and real-time deep neural network for insulator detection in complex aerial images. Energies **14**(5), 1426 (2021)

[16] Alam, M.N., Bhuiyan, R.H., Dougal, R.A., *et al.*: Design and application of surface wave sensors for nonintrusive power line fault detection. IEEE Sensors Journal **13**(1), 339–347 (2012)

[17] Wang, S., Liu, Y., Qing, Y., *et al.*: Detection of insulator defects with improved resnest and region proposal network. IEEE Access **8**, 184841–184850 (2020)

[18] Mei, H., Jiang, H., Yin, F., *et al.*: Terahertz imaging method for composite insulator defects based on edge detection algorithm. IEEE Transactions on Instrumentation and Measurement **70**, 1–10 (2021)

[19] Wu, Y., Ding, D., Wang, Y., *et al.*: Defect recognition and condition assessment of epoxy insulators in gas insulated switchgear based on multi-information fusion. Measurement **190**, 110701 (2022)

[20] Siddiqui, Z.A., Park, U.: A drone based transmission line components inspection system with deep learning technique. Energies **13**(13), 3348 (2020)

[21] Wang, Q., Bo, C.: A novel transfer learning model for the real-time concrete crack detection. Knowledge-Based Systems **301**, 112313 (2024)

[22] Liu, Y., Liu, Y., Guo, X., *et al.*: Metal surface defect detection using slf-yolo enhanced yolov8 model. Scientific Reports **15**(1), 11105 (2025)

[23] Ji, X., Yue, Z., Yang, H., Zhang, Z.: Infrared image classification and detection algorithm for power equipment based on improved yolov10. IEEE Access **12**, 184976–184988 (2024)

[24] Zhao, Y., Lv, W., Xu, S., *et al.*: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974 (2024)

[25] Lv, W., Xu, S., Zhao, Y., et al.: Detrs beat yolos on real-time object detection. CoRR (2023)

[26] Wang, S., Jiang, H., Li, Z., *et al.*: Phsi-rtdetr: A lightweight infrared small target detection algorithm based on uav aerial photography. Drones **8**(6), 240 (2024)

[27] Wang, S., Jiang, H., Yang, J., *et al.*: Lightweight tomato ripeness detection algorithm based on the improved rt-detr. Frontiers in Plant Science **15**, 1415297

(2024)

[28] Ma, S., Zhang, Y., Peng, L., et al.: Owrt-detr: A novel real-time transformer network for small object detection in open water search and rescue from uav aerial imagery. IEEE Transactions on Geoscience and Remote Sensing (2025)

[29] Zhou, J., He, Z., Zhang, D., et al.: Spatial residual for underwater object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[30] Zhang, L., Yang, K., Han, Y., *et al.*: Tsd-detr: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving. Engineering Applications of Artificial Intelligence **139**, 109536 (2025)

[31] Wang, S., Jiang, H., Yang, J., *et al.*: Amfef-detr: An end-to-end adaptive multi-scale feature extraction and fusion object detection network based on uav aerial images. Drones **8**(10), 523 (2024)

[32] Zhao, Y., *et al.*: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[33] Insulators break fault: insulator and insulator binding dataset. In: Roboflow Universe (Open Source Dataset) (2025)

[34] Liu, W., Anguelov, D., Erhan, D., *et al.*: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016). Springer

[35] Talaat, F.M., ZainEldin, H.: An improved fire detection approach based on yolo-v8 for smart cities. Neural Computing and Applications **35**(28), 20939–20954 (2023)

[36] Zhang, H., Li, F., Liu, S., *et al.*: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2023)

[37] Zhu, Z., Chen, S., Dong, Z., *et al.*: Co-detr: Towards real-time and better multi-task object detection with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6545–6554 (2023)