Robust Tiny Insulator Defect Detection under Complex UAV Backgrounds with State Space Modeling

Chuanxing Geng, Heng Li, Hui Liu, Xiaojun Xue*
Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China
Email: 1240397058@qq.com, xuexiaojun35678@kust.edu.cn

***Abstract*—Insulator defect detection in UAV inspection is challenging due to tiny defect scales, complex backgrounds, multi-view variations, and long-chain structural dependencies. To address these issues, we propose PRD-DFINE, a lightweight and robust end-to-end detection framework based on DFINE. The framework introduces three key components: (1) a Progressive Injection Feature Pyramid (PI-FPN) to enhance fine-grained representation for tiny defects, (2) a Context-Driven Robustness Enhancement module (CRFE) to suppress background interference during multi-scale fusion, and (3) a Decoding-Enhanced Non-Causal Visual State Space Model (DSSM) for efficient global context modeling. Experiments on IFDD and InsPLAD show that PRD-DFINE outperforms mainstream detectors while remaining lightweight, achieving strong robustness and generalization under complex UAV scenarios. With only 9.1M parameters and 29.67 GFLOPs, it is suitable for real-world deployment. Furthermore, qualitative heatmap visualizations confirm that our method produces more compact and discriminative responses on defect regions, indicating improved resistance to background distractions.Code is available at https://github.com/GcxMaxx/PRD-DFINE.**

***Index Terms*—Defect detection; UAV inspection; insulator; state space model; end-to-end object detection.**

I. Introduction

With the expansion of power grids, insulator defects such as contamination, cracks, and damage have become more frequent, posing risks to system reliability [1]. UAV-based inspection provides an efficient solution, yet accurate detection remains challenging due to tiny defects, complex backgrounds, and multi-view variations. Moreover, the elongated structure of insulator strings and sparse defect distribution require effective global context modeling for stable localization [2].

Recent end-to-end detectors such as DETR [3] remove hand-crafted components through set prediction, but they still suffer from weak shallow-detail preservation, semantic conflicts in multi-scale fusion, and high self-attention cost on high-resolution features in UAV insulator inspection [4], [5]. To address these issues, we propose PRD-DFINE, a robust and efficient DFINE-based framework for tiny defect perception, background-robust fusion, and efficient global context modeling. The main contributions of this paper are summarized as follows:

- We propose PI-FPN, a progressive cross-level injection strategy that enhances fine-grained representation with low multi-scale fusion overhead.
- We design CRFE to suppress background interference and alleviate semantic conflicts via context-guided fusion and spatial-frequency enhancement.
- We develop DSSM, a non-causal visual state space module for efficient global context modeling with reduced complexity.

* Corresponding author: Xiaojun Xue (xuexiaojun35678@kust.edu.cn).
This work was supported in part by the National Natural Science Foundation of China under Grant 62263016, and in part by the Yunnan Provincial University Service Key Industry Science and Technology Program under Grant FWCY-QYCT2024003.

II. Related Work

A. Insulator Defect Detection

Insulator defect detection has become an important task in UAV-based transmission line inspection. Existing methods mainly rely on deep detectors to exploit texture and context cues, but small defects under complex backgrounds still lead to false positives and missed detections, especially under viewpoint, scale, and illumination variations [1], [2].

B. End-to-End Detectors for UAV Inspection

DETR formulates detection as set prediction with Hungarian matching [3]. Subsequent variants, including Deformable DETR [6], Conditional DETR [7], DAB-DETR [8], DN-DETR [9], and DINO [10], improve convergence and localization. For real-time scenarios, RT-DETR improves efficiency through a hybrid encoder [11], while DFINE further enhances localization via distribution refinement [12]. However, complex UAV backgrounds, tiny defects, and long-chain structures still pose challenges to fine-grained representation, robust fusion, and efficient global modeling [13], [14].

III. Method

As shown in Fig. 1, we propose PRD-DFINE, a robust and efficient detection framework built upon DFINE. The framework enhances tiny defect detection, robustness under complex backgrounds, and global context modeling through three key components: PI-FPN, CRFE, and DSSM.

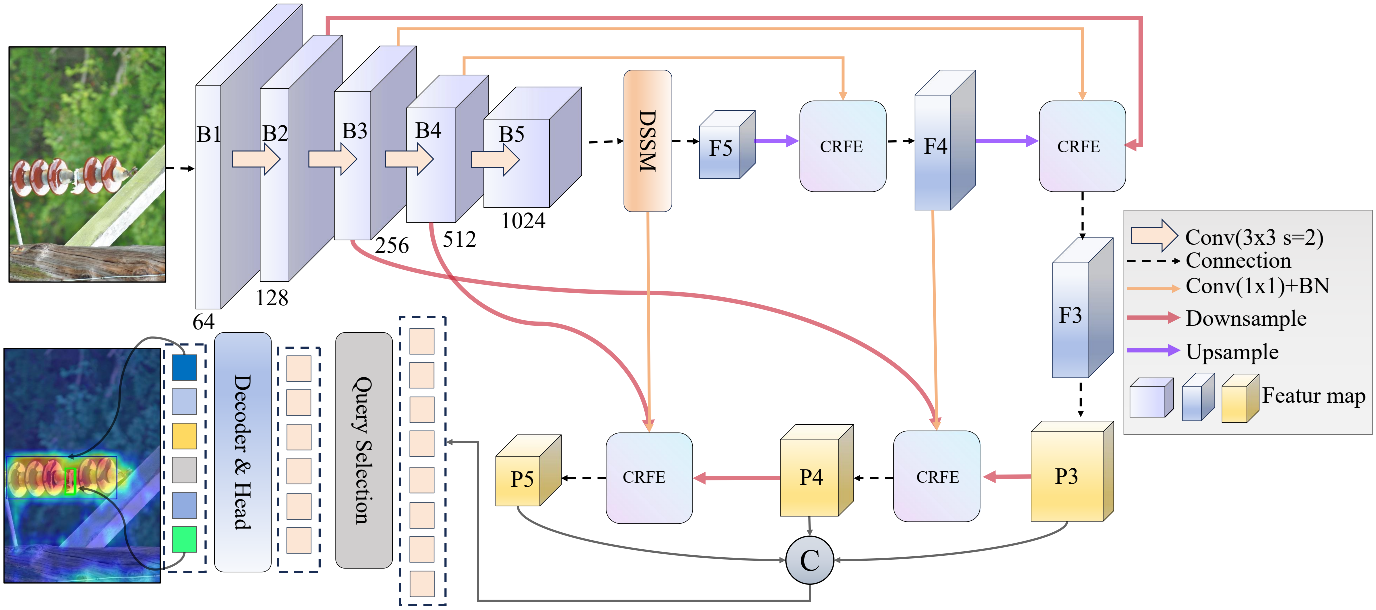


Fig. 1: Overall architecture of PRD-DFINE.

A. *Progressive Injection Feature Pyramid (PI-FPN)*

In UAV-based transmission line inspection scenarios, insulator defects typically exhibit small scale, weak texture, and blurred boundaries, making them easily affected by complex backgrounds. During hierarchical feature extraction, repeated downsampling inevitably weakens fine-grained structural details, leading to missed detections of tiny defects. In contrast, shallow features preserve rich spatial and edge information, which is critical for small object detection.

Feature Pyramid Network (FPN) enhances shallow features via top-down pathways and lateral connections [15]. PANet further strengthens feature propagation through bottom-up aggregation [16], while BiFPN introduces bidirectional fusion with learnable weights for adaptive multi-scale interaction [17]. As illustrated in Fig. 2a, these methods improve multi-scale representation but mainly rely on repeated fusion operations, which may introduce redundant computation and insufficient preservation of fine-grained details, especially in high-resolution UAV inspection scenarios. Moreover, directly incorporating shallow features (e.g., B_2) into multiple pyramid levels significantly increases computational cost and may lead to semantic inconsistency [18].

To address these issues, we propose a Progressive Injection Feature Pyramid Network (PI-FPN), which introduces a multi-stage progressive injection strategy to gradually propagate shallow structural information into deeper semantic representations. Instead of directly fusing high-resolution features across all pyramid levels, PI-FPN performs hierarchical cross-level feature injection in a stepwise manner:

$$B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_5, \tag{1}$$

where shallow features are progressively aligned and injected into higher-level representations. At each stage, injected features are first transformed to match the target resolution and then fused with the corresponding higher-level features. This progressive propagation enables fine-grained details to be continuously preserved and refined across multiple semantic levels.

Compared with conventional multi-scale fusion, the proposed progressive injection offers two advantages: (1) it avoids repeated processing of high-resolution feature maps, thereby reducing computational overhead, and (2) it mitigates semantic conflicts by gradually integrating shallow details into deeper representations, improving feature consistency and localization stability for tiny defects.

To efficiently support cross-scale feature alignment in this process, we further design SCFConv (Shuffle-based Context-preserving Factorized Convolution), as illustrated in Fig. 2b. SCFConv is a lightweight information-preserving downsampling and feature realignment module. Unlike conventional downsampling operations (e.g., pooling or strided convolution) that may discard spatial details, SCFConv explicitly transfers local spatial information into the channel dimension while enhancing inter-channel interaction.

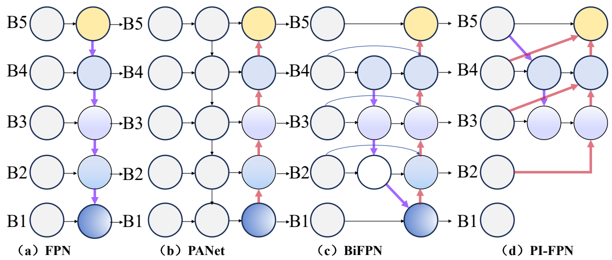
Given input feature $X \in \mathbb{R}^{B\times C_1\times H\times W}$, SCFConv first applies a channel shuffle operation to enhance inter-channel interaction:

$$X_s = \text{Shuffle}(X; G), \tag{2}$$

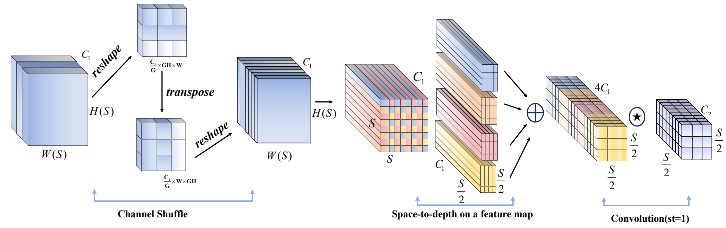
$$\text{Shuffle}(X; G) = \text{Reshape}\big(\text{Transpose}(\text{Reshape}(X))\big), \tag{3}$$

where G denotes the number of channel groups. Subsequently, a spatial-to-channel rearrangement is performed using a 2×2 sampling pattern, which decomposes spatial information into multiple sub-features:

$$\begin{aligned}
X_{00} &= X_s(:, :, 0{:}2{:}H,\ 0{:}2{:}W),\\
X_{01} &= X_s(:, :, 0{:}2{:}H,\ 1{:}2{:}W),\\
X_{10} &= X_s(:, :, 1{:}2{:}H,\ 0{:}2{:}W),\\
X_{11} &= X_s(:, :, 1{:}2{:}H,\ 1{:}2{:}W),
\end{aligned} \tag{4}$$



(a) Comparison of feature pyramid network architectures.



(b) Architecture of SCFConv.

Fig. 2: Comparison of feature pyramid architectures and SCFConv structure.

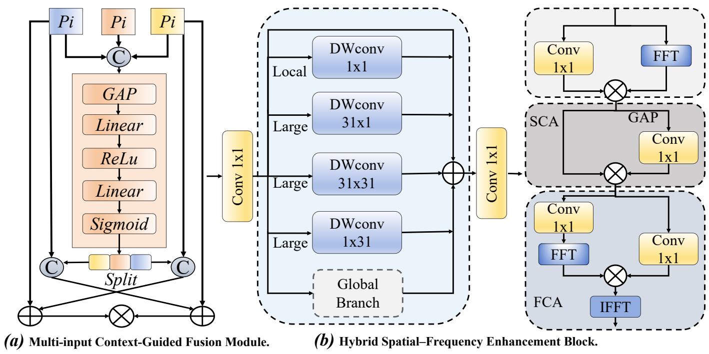


Fig. 3: Structure of the CRFE module, consisting of (a) CGFM and (b) HSFEB.

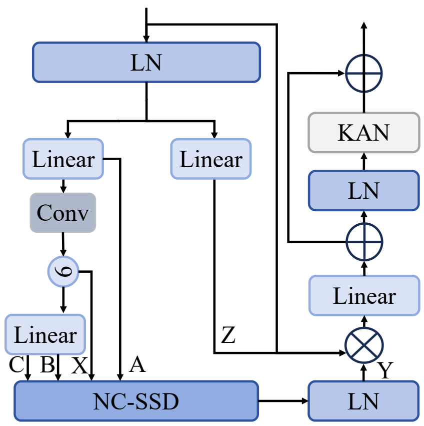


Fig. 4: Structure of the DSSM module.

$$X' = \text{Concat}(X_{00}, X_{01}, X_{10}, X_{11}) \in \mathbb{R}^{B\times C_2 \times \frac{H}{2} \times \frac{W}{2}}. \quad (5)$$

Finally, a $k \times k$ convolution is applied for feature fusion and channel compression:

$$Y = \text{Conv}_{k\times k}(X'). \quad (6)$$

Through this design, SCFConv jointly performs spatial compression and channel enrichment, enabling efficient cross-scale alignment while preserving fine-grained structural information. This makes it particularly suitable for progressive feature injection in high-resolution detection tasks.

Overall, PI-FPN transforms shallow feature utilization from a one-step fusion process into a progressive multi-level refinement mechanism. This design enhances tiny defect sensitivity, improves localization stability, and maintains computational efficiency under complex UAV inspection environments.

B. Context-Driven Robustness Enhancement Module (CRFE)

In UAV inspection scenarios, defect regions are often easily confused with complex background patterns, leading to unstable predictions. To address this issue, we propose CRFE to enhance feature robustness during multi-scale fusion.

As shown in Fig. 3, CRFE consists of two components: (1) a Context-Guided Fusion Module (CGFM) for adaptive multi-scale aggregation, and (2) a Hybrid Spatial–Frequency Enhancement Block (HSFEB) for feature refinement.

1) Context-Guided Fusion Module (CGFM): CGFM performs unified fusion of multi-scale features with context-aware weighting. Given concatenated features F, global context is extracted via:

$$s = \sigma(W_2\delta(W_1\text{GAP}(F))), \quad (7)$$

where GAP denotes global average pooling. The resulting weights are applied to each branch:

$$\tilde{P}_k = P_k \odot s_k, \quad (8)$$

where s_k corresponds to the weight of the k-th branch. This design enables cross-branch interaction and alleviates semantic conflicts during fusion.

2) Hybrid Spatial–Frequency Enhancement Block (HSFEB): To further enhance feature robustness, HSFEB integrates channel-wise and spatial frequency modeling to capture both local details and global structural information.

Channel Frequency Modeling (FCA). Given input feature X, global channel context is first extracted via global average pooling, and transformed into the frequency domain:

$$X_{FCA} = \mathcal{F}^{-1}\left(\mathcal{F}(X_G) \otimes W_c\right), \quad (9)$$

where $X_G = \text{GAP}(X)$ and W_c denotes learnable channel-wise modulation parameters.

Channel Feature Enhancement (CFE). Based on the frequency-aware representation, channel responses are recalibrated:

$$X_{CFE} = X \odot \sigma(W_s(X_{FCA})), \quad (10)$$

where $W_s(\cdot)$ denotes a learnable transformation and σ is the activation function.

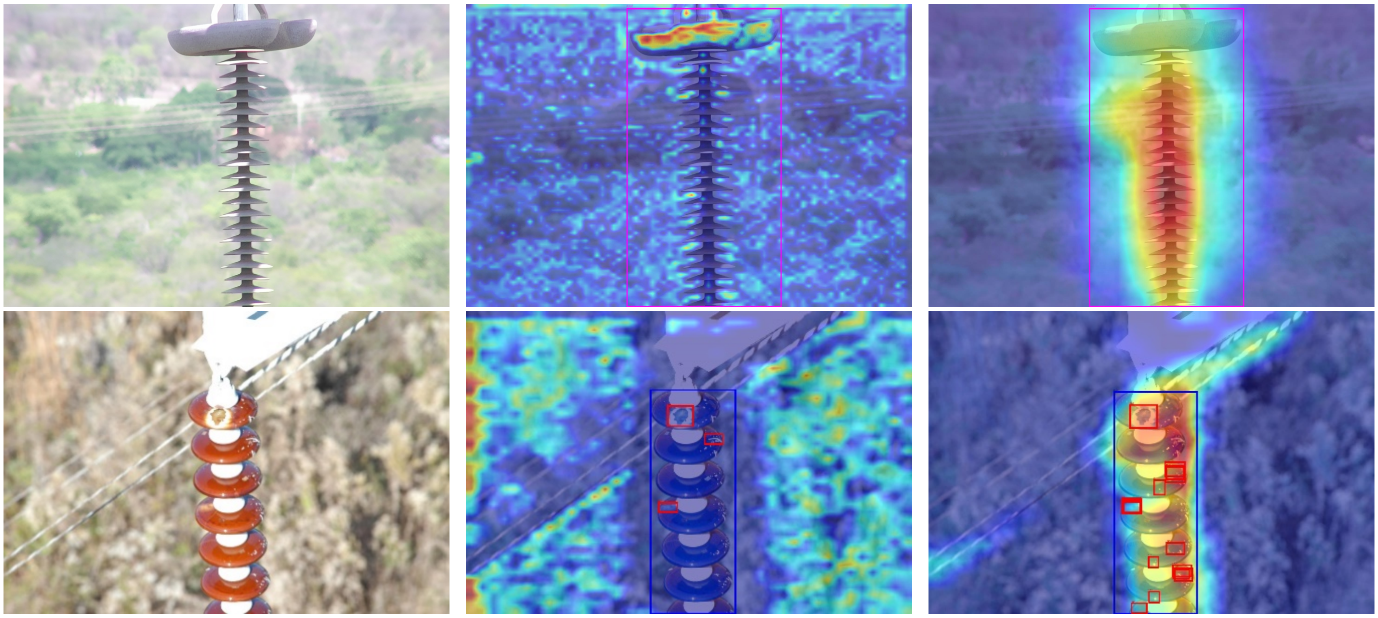
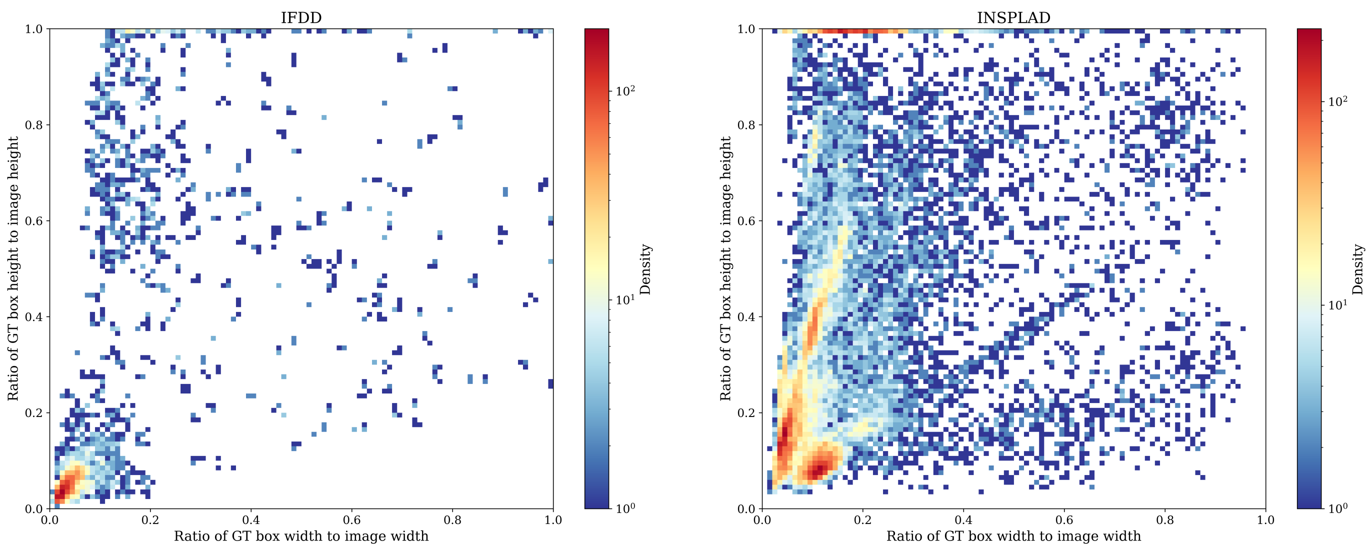


Fig. 5: Comparison of visualization results: (a) original image, (b) baseline, (c) ours.



(a) IFDD

(b) InsPLAD

Fig. 6: Density distribution of ground-truth bounding box sizes on IFDD and InsPLAD datasets.

Spatial Frequency Enhancement (SFE). To capture global spatial dependencies, the feature is further transformed into the frequency domain:

$$X_{SFE} = \mathcal{F}^{-1}\left(\mathcal{F}(X_{CFE}) \otimes W_f\right), \tag{11}$$

where W_f represents learnable spatial-frequency modulation parameters.

The final output combines spatial and frequency-domain enhancement to improve discriminability under complex backgrounds.

C. DSSM: Decoding-Enhanced Non-Causal Visual State Space Model

As shown in Fig. 4, DSSM is a decoding-enhanced non-causal visual state space module for efficient long-range dependency modeling in high-resolution features. Unlike Transformers with quadratic complexity, DSSM adopts a non-causal state space formulation to enable global context modeling with reduced computational cost.

Given input feature X, DSSM performs state evolution as:

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Ch_t + Dx_t, \tag{12}$$

which enables bidirectional information propagation across spatial tokens.

To enhance representation capability, a learnable nonlinear decoding function is introduced:

$$X_{\text{out}} = X + \text{KAN}(\text{LN}(X)), \tag{13}$$

$$\text{KAN}(z) = W_o\phi(W_i z), \tag{14}$$

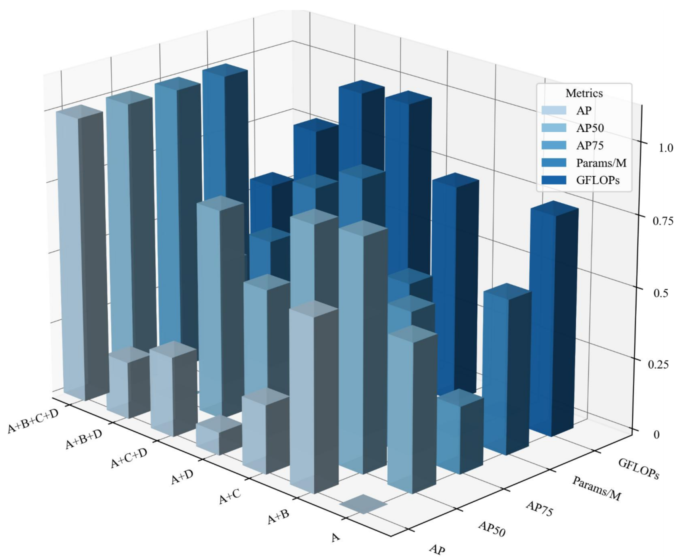


Fig. 7: Ablation results on IFDD.

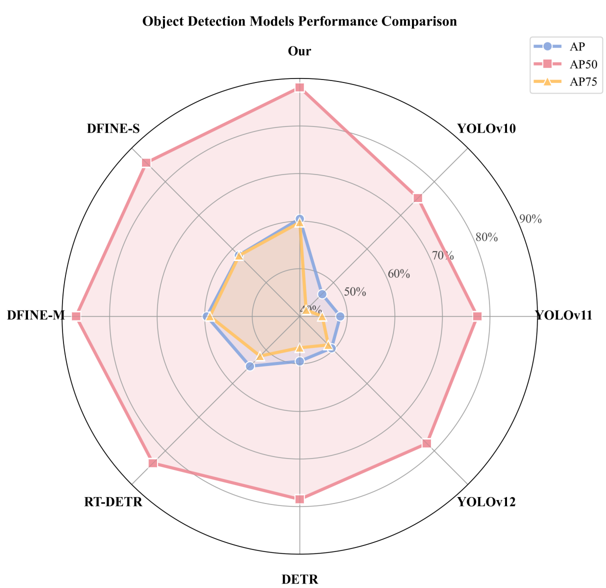


Fig. 8: Performance comparison on IFDD.

where $\mathrm{KAN}(\cdot)$ denotes a nonlinear mapping. By decoupling global dependency modeling and feature reconstruction, DSSM provides an efficient solution for global context aggregation in long-chain and high-resolution detection scenarios.

IV. EXPERIMENTS AND RESULTS

A. Datasets

We evaluate the proposed method on IFDD and InsPLAD to assess both defect-specific performance and general detection capability in UAV inspection scenarios. IFDD contains 1,600 high-resolution insulator images with two defect categories (pollution flashovers and physical damage) [19]. The dataset is dominated by small and medium-sized targets, making it suitable for fine-grained defect detection. InsPLAD is a UAV-based dataset with 10,607 images and annotations for 17 asset categories [20]. We use the full 17-class setting without filtering, following the official split (7,935 training / 2,626 test images).

Note that the two datasets differ in task formulation: IFDD focuses on two-class defect detection, while InsPLAD involves multi-class object detection. Thus, InsPLAD serves as a complementary benchmark to evaluate robustness and generalization under complex UAV backgrounds, viewpoint changes, and scale variations. As shown in Fig. 6, both datasets are dominated by small objects.

TABLE I: Ablation study results on IFDD and InsPLAD (A: baseline, B: PI-FPN, C: CRFE, D: DSSM).

Dataset	A	B	C	D	AP	AP_{50}	AP_{75}	Params (M)	GFLOPs
IFDD	✓				0.612	0.898	0.603	10.2	24.82
	✓	✓			0.627	0.917	0.612	11.5	42.02
	✓		✓		0.618	0.916	0.609	10.4	25.02
	✓			✓	0.614	0.898	0.604	9.6	19.70
	✓		✓	✓	0.619	0.912	0.595	9.8	19.95
	✓	✓		✓	0.617	0.865	0.610	10.4	24.10
	✓	✓	✓	✓	**0.637**	**0.929**	**0.629**	**9.1**	29.67
InsPLAD	✓				0.687	0.829	0.718	10.2	24.82
	✓	✓			0.701	0.842	0.733	11.5	42.02
	✓		✓		0.695	0.844	0.714	10.3	25.02
	✓			✓	0.688	0.818	0.711	9.6	19.70
	✓		✓	✓	0.690	0.844	0.703	9.8	19.95
	✓	✓		✓	0.705	0.837	0.738	10.4	24.10
	✓	✓	✓	✓	**0.712**	**0.856**	**0.744**	**9.1**	29.67

TABLE II: Comparison with mainstream detectors on IFDD and InsPLAD.

Dataset	Model	AP	AP_{50}	AP_{75}	Params (M)	FLOPs (G)
IFDD	YOLOv10	0.490	0.792	0.441	9.5	23.6
	YOLOv11	0.512	0.815	0.470	9.4	21.8
	YOLOv12	0.525	0.820	0.515	9.2	**21.2**
	DETR	0.522	0.827	0.490	41.0	86.2
	RT-DETR	0.577	0.882	0.545	20.0	58.3
	DFINE-M	0.627	0.918	0.620	20.0	56.33
	DFINE-S	0.612	0.903	0.611	10.2	24.8
	Ours	**0.637**	**0.929**	**0.629**	**9.1**	29.67
InsPLAD	YOLOv10	0.612	0.742	0.651	9.5	23.6
	YOLOv11	0.591	0.718	0.566	9.4	21.8
	YOLOv12	0.615	0.691	0.647	9.2	**21.2**
	DETR	0.644	0.726	0.615	41.0	86.2
	RT-DETR	0.661	0.764	0.653	20.0	58.3
	DFINE-M	0.688	0.834	0.699	20.0	56.33
	DFINE-S	0.687	0.829	0.718	10.2	24.82
	Ours	**0.712**	**0.856**	**0.744**	**9.1**	29.67

B. Experimental Setup

Models are implemented in PyTorch 2.3.0 and trained on an NVIDIA RTX 4090 GPU using AdamW (lr=1e-4, batch size=8) with a cosine scheduler for 300 epochs. Input images are resized to 640×640.

C. Ablation Studies

To evaluate the effectiveness of each component, we conduct ablation studies on both IFDD and InsPLAD based on the DFINE-S baseline. Note that IFDD evaluates two-category defect detection, while InsPLAD follows a 17-class object detection setting, allowing us to assess both defect-specific performance and generalization ability.

As shown in Table I, each proposed module contributes consistently to performance improvement. Incorporating PI-FPN improves AP on IFDD from 0.612 to 0.627, demonstrating its effectiveness in enhancing tiny defect perception through progressive injection of shallow features. CRFE further improves robustness under complex UAV backgrounds by suppressing background interference and alleviating semantic conflicts during multi-scale fusion. DSSM enhances global contextual modeling, leading to more stable detection for long-chain structures. Fig. 7 further illustrates the performance differences among different module combinations on IFDD.

The combination of modules yields complementary gains. In particular, integrating CRFE and DSSM (A+C+D) further improves performance, indicating that robust feature fusion and global modeling are mutually beneficial. The full model (A+B+C+D) achieves the best results on both datasets, reaching an AP of 0.637 on IFDD and an AP of 0.712 on InsPLAD, validating the effectiveness and synergy of all components.

Fig. 5 provides qualitative comparisons. The baseline produces scattered activations and is easily distracted by background patterns, whereas PRD-DFINE generates more compact and discriminative responses concentrated on defect regions, demonstrating improved robustness and localization stability.

D. Comparative Experiments

We compare PRD-DFINE with mainstream detectors on IFDD and InsPLAD, including YOLOv10/YOLOv11/YOLOv12, DETR, RT-DETR, and DFINE. Note that the two datasets differ in task formulation (two-class defect detection vs. 17-class object detection), thus providing complementary evaluations of accuracy and robustness.

As shown in Table II, PRD-DFINE achieves the best overall performance on both datasets. Specifically, it achieves an AP of 0.637 on IFDD and an AP of 0.712 on InsPLAD, outperforming both YOLO-series and transformer-based baselines. Compared with DFINE-S, our method provides consistent accuracy improvements while reducing model parameters to 9.1M, demonstrating a better accuracy-efficiency trade-off.

Fig. 8 further illustrates the performance comparison, where PRD-DFINE shows superior detection accuracy, highlighting its robustness under complex UAV inspection conditions and its effectiveness in handling small-scale and challenging targets.

V. CONCLUSIONS

This paper presented PRD-DFINE, a lightweight and robust end-to-end framework for insulator defect detection under complex UAV backgrounds. The proposed PI-FPN improves fine-grained representation for tiny defects, CRFE enhances robustness during multi-scale fusion, and DSSM enables efficient global context modeling with low computational overhead. Experiments on IFDD and InsPLAD demonstrate that PRD-DFINE achieves better accuracy-efficiency trade-offs than mainstream detectors while maintaining strong robustness and generalization. These results indicate its potential for practical UAV-based transmission line inspection.

REFERENCES

[1] Y. Liu, D. Liu, X. Huang, and C. Li, "Insulator defect detection with deep learning: A survey," *IET Generation, Transmission & Distribution*, vol. 17, no. 16, pp. 3541–3558, 2023.

[2] N. Al-Qubaydhi, A. Alenezi, T. Alanazi, A. Senyor, N. Alanezi, B. Alotaibi, M. Alotaibi, A. Razaque, and S. Hariri, "Deep learning for unmanned aerial vehicles detection: A review," *Computer Science Review*, vol. 51, p. 100614, 2024.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[4] A. V. et al., "Attention is all you need," in *NeurIPS*, 2017.

[5] A. D. et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[6] X. Z. et al., "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[7] D. M. et al., "Conditional detr for fast training convergence," in *ICCV*, 2021.

[8] S. L. et al., "Dab-detr: Dynamic anchor boxes are better queries for detr," in *ICLR*, 2022.

[9] F. L. et al., "Dn-detr: Accelerate detr training by introducing query denoising," in *CVPR*, 2022.

[10] H. Z. et al., "Dino: Detr with improved denoising anchor boxes," in *ICLR*, 2023.

[11] Y. Z. et al., "Detrs beat yolos on real-time object detection (rt-detr)," *arXiv*, 2023.

[12] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task in detrs as fine-grained distribution refinement," *arXiv*, 2024.

[13] D. Li, P. Yang, and Y. Zou, "Optimizing insulator defect detection with improved detr models," *Mathematics*, vol. 12, no. 10, p. 1507, 2024.

[14] W. Chen, S. Li, and X. Han, "Idd-detr: Insulator defect detection model based on bidirectional cross-scale fusion," *Sensors*, vol. 25, no. 18, 2025.

[15] T.-Y. L. et al., "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018, pp. 8759–8768.

[17] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *CVPR*, 2020, pp. 10 781–10 790.

[18] Q. C. et al., "You only look one-level feature," in *CVPR*, 2021, pp. 13 039–13 048.

[19] T. Wang, Y. Wang, P. Zong, S. Xu, W. Wei, and H. Liu, "Bmf-sdformer: Multi-scale feature fusion transformer," *IET Image Processing*, 2026.

[20] A. L. B. V. e Silva et al., "Insplad: A dataset and benchmark for power line inspection," *arXiv*, 2023.