A Small-Sized Defect Detection Method for Overhead Transmission Lines Based on Convolutional Neural Networks

A Small-Sized Defect Detection Method for Overhead Transmission Lines Based on Convolutional Neural Networks

Qi Fu, Jiefeng Liu, *Senior Member, IEEE*, Xingtuo Zhang, Yiyi Zhang, *Member, IEEE*, Yang Ou, Runnong Jiao, Chuanyang Li, *Member, IEEE*, and Giovanni Mazzanti, *Fellow, IEEE*

***Abstract*—Insulator defect detection is essential to the reliable operation of overhead transmission lines. However, current automatic algorithms struggle to extract critical features due to the small size of insulator defects in inspection images, which may lead to potential failures. To address this issue, this paper proposes a novel and high-accuracy defect detection method based on deep learning technology, named insulator defect detection network (I2D-Net), which incorporates several innovative modules. First, we design a three-path feature fusion network (TFFN) to improve the network's ability to extract features from shallow layers. This hierarchical feature fusion mechanism across different network layers preserves spatial and semantic information, thereby maintaining the quality of features at different levels of the pyramid. Second, an enhanced receptive field attention (RFA+) block is incorporated to enable the network to adapt to different-scale defects and effectively distinguish them from the background. Finally, the context perception module (CPM) is introduced to better understand the surrounding features and their relationship with the defects. This improves defect localization capacity in the presence of interfering factors. Experimental results on the transmission line dataset demonstrate that the proposed method can accurately detect insulator defects and electrical components, even in challenging scenarios.**

***Index Terms*—Deep learning, feature fusion network, attention, defect detection, transmission line.**

NOMENCLATURE

I2D-Net	insulator defect detection network
TFFN	three-path feature fusion network
RFA+	enhanced receptive field attention
CPM	context perception module
HVOHL	high-voltage overhead transmission lines
CNN	convolutional neural network
(m) AP	(mean) average precision
RPN	region proposal network
FPN	feature pyramid network
CAM	channel attention module
SAM	spatial attention module
RoI	region of interest

This work was supported by the the Guangxi Natural Science Foundation Project (2023GXNSFFA026012) and Innovation Project of Guangxi Graduate Education（YCBZ2023008）. (*Corresponding author: Jiefeng Liu.*)

Qi Fu, Jiefeng Liu, Xingtuo Zhang, Yiyi Zhang, Yang Ou, and Runnong Jiao are with Guangxi Key Laboratory of Power System Optimization and Energy Technology, Guangxi University, Nanning 530004, China. (e-mail: 2112392021@st.gxu.edu.cn; jiefengliu2018@gxu.edu.cn; 1912401005@st.gxu.edu.cn; yiyizhang@gxu.edu.cn; 2212301040@st.gxu.edu.cn; 2212301025@st.gxu.edu.cn).

Chuanyang Li is with the Department of Electrical Engineering, Tsinghua University, Beijing 100084, China. (e-mail: lichuanyangsuper@163.com).

Giovanni Mazzanti is with the Dept. of Electrical, Electronic and Information Engineering, University of Bologna, Viale Risorgimento 2, Bologna 40136, Italy. (e-mail: giovanni.mazzanti@unibo.it).

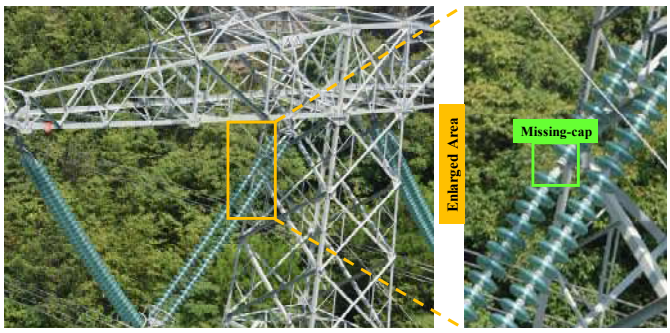


Fig. 1. A sample of missing-cap defects in overhead transmission line images captured by UAVs.

I. INTRODUCTION

Statistics as of 2020 show that the length of high-voltage (voltage levels of 220kV and above) overhead transmission lines (HVOHLs) in China has reached 79,000 kilometers. Insulators, as the critical component with electrical insulation and mechanical fixation function, are substantially applicated in HVOHLs. However, insulator failures occur frequently, mainly due to the cracks in the manufacturing process and deterioration caused by the harsh natural environment [1]. Once the insulator malfunctions, it may cause transmission line tripping and leakage, ultimately leading to power outage accidents and economic loss. Hence, the inspection of defective insulators to prevent further failures is essential for safe and stable power industry production [1-2].

Focusing on the insulator defects, there already have been some traditional methods to detect them. Utilities have historically dispatched helicopters to detect HVOHLs failures [3] (inaccurate, labor-intensive), or used other efficient means to analyze collected data. For instance, Palangar *et al.* [2] proposed an automatic detection device leveraging radio communication technology to diagnose insulator faults. Stefenon *et al.* [4] introduced an insulator conditions evaluation method based on ultrasound sensors and the ensemble extreme learning machine optimized by the particle swarm algorithm. They have completed efficient defect detection, but the external

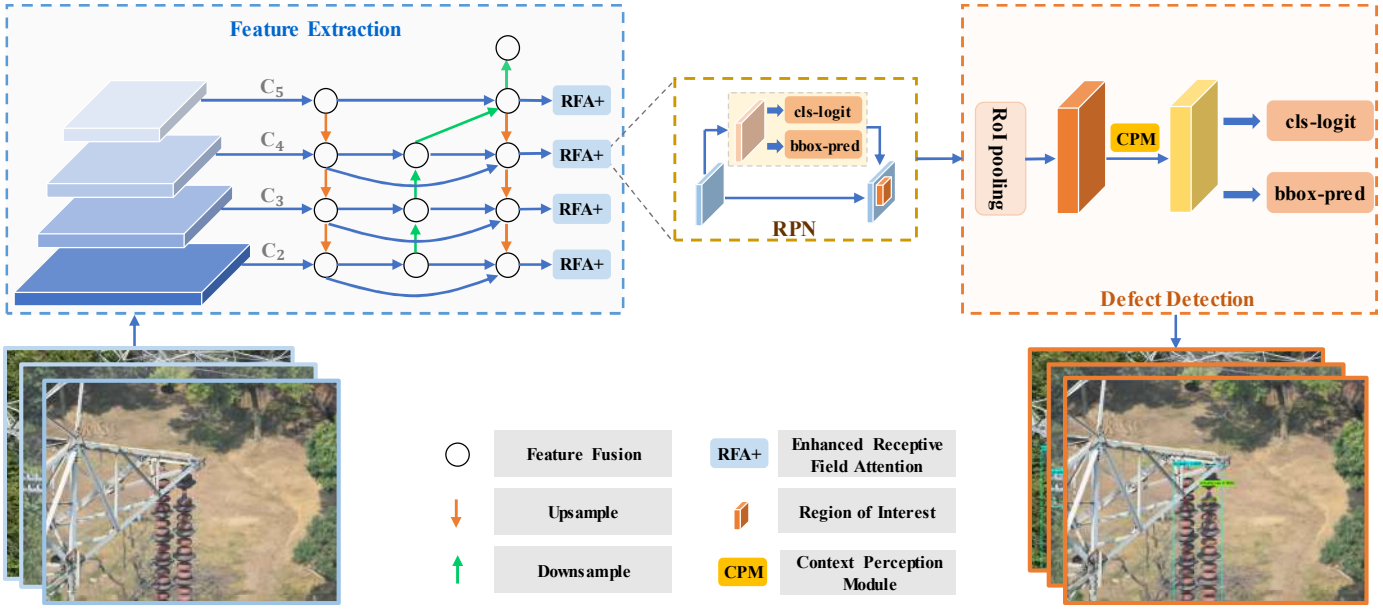


Fig. 2. The pipeline of the proposed I2D-Net.

noise can easily interfere with detection results. Nowadays, utility companies tend to use unmanned aerial vehicles (UAVs) to monitor overhead transmission lines. However, it is inefficient for utility labors to analyze the condition of insulators (normal or defective) from massive image data captured by UAVs. Thus, applying deep learning technology to automatically detect insulator defects is an alternative solution. Although insulators have many types of defects due to their long-term exposure, this paper focuses on the most common and frequently occurring defect: missing-cap [5].

Recently, deep learning technology, especially convolutional neural networks (CNNs), has made significant achievements in defect detection applications in power transmission industries [6-8]. Luo *et al*. [7] designed a defect detection model based on a deep convolutional neural network (DCNN) to enhance the feature representation of ultra-small bolts and achieved great performance in bolt defect detection. Zhao *et al*. [8] constructed an automated approach based on deep learning techniques for pin-missing defect detection for bolts in transmission lines. The CNN-based automatic defect detection paradigms are generally divided into segmentation-based approaches (e.g. [9]) and detection-based approaches (e.g. [10]). The segmentation-based method subdivides a digital image into collections of pixels and outputs the regions and corresponding categories of defects. While the detection-based method finds all the regions of interest in the image to generate detection boxes and then determines their category and localization. Since the task of insulator missing-cap defects detection faces huge challenges, such as complex backgrounds, diverse components, and small size relative to the original image (see Fig. 1), the segmentation-based approach is not applicable to this scenario. In the past decade, excellent object detection methods have been applied in insulator defects detection. Tao *et al.* [5] proposed a cascaded architecture based on a tandem of two-stage deep CNNs to detect missing-cap defects, where the feature extraction networks of two independent networks are not shared. Liu *et al.* [11] utilized you only look once v3 (YOLOv3) [12] to recognize insulators, and further diagnosed insulator faults with discharge phenomena by leveraging ultraviolet imaging technology. Wang *et al.* [13] utilized the typical single-shot detector (SSD) [14] algorithm to identify missing-cap defects. Lei *et al.* [15] proposed an insulator defect detection method based on faster region-based CNN (Faster R-CNN) [16] to recognize defective insulator strings with missing-cap. These methods have made progress in identifying normal insulators as well as detecting insulator defects. However, their detection accuracy for small-sized defects in various complex backgrounds is insufficient to meet the reliability for the operation and maintenance of overhead transmission lines.

To address the aforementioned problem, a high-accuracy approach for identifying small-sized missing-cap defects is proposed in this paper, named insulator defect detection network (I2D-Net). Our approach is designed specifically for offline scenarios where real-time constraints are not a primary concern, and aims to achieve state-of-the-art accuracy while maintaining reasonable computational costs. The key contributions are as follows:

(1) We first devise a novel three-path feature fusion network (TFFN) to alleviate the feature disappearance of missing-cap defects. Second, we propose an enhanced receptive field attention (RFA+) block to highlight the defect feature and suppress useless information. Third, we design a context perception module (CPM) to strengthen the defect localization ability.

(2) The I2D-Net is constructed for improving the detection accuracy of small-sized missing-cap defects in complex backgrounds by aggregating semantic information, spatial details, and contextual information of defects.

(3) The experimental results demonstrate that the proposed I2D-Net outperforms other existing works in detecting missing-cap defects with an average precision (AP) value of 89.4%. It reliably increases the detection accuracy of insulator missing-cap defects by a large margin.

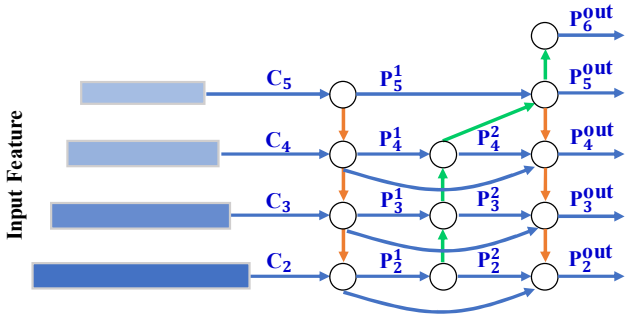


Fig.3. Architecture of TFFN. The leftmost rectangular box from bottom to top is the output feature map of progressive convolution in ResNet-50. The green arrow represents upsampling, the orange arrow represents downsampling, and the black circle represents the feature fusion process.

The rest of this paper is organized as follows. Section II introduces the details of our suggested method. Section III presents experimental results and corresponding explanations. Section IV concludes this paper.

II. Methodology

This paper proposes a novel I2D-Net for the detection of small-sized insulator defects in transmission line inspection images. The proposed model is based on the Faster R-CNN detector, which incorporates a feature pyramid network (FPN) to generate multi-scale feature maps. However, our model further enhances the feature fusion network by introducing the TFFN, which is specifically designed to improve the feature representation of small-sized insulator defects. Additionally, our model incorporates several novel components, including the RFA+ module and CPM, which are not present in the baseline model. These components are aimed at improving the accuracy of the model by focusing on important regions of the image and by refining the location and scale of detected objects.

The pipeline of our proposed I2D-Net is illustrated in Fig 2. Firstly, the input transmission line image is processed by the backbone convolution layers to generate multi-scale feature representations. Then, these features are forwarded through the TFFN and RFA+ block to strengthen their representation capacity. Next, the region proposal network (RPN) [16] is employed to generate candidate boxes and feature maps. Finally, the defect detection network produces the detection results.

A. Feature Extraction Network

1) Backbone: We adopt the ResNet-50 [17] pre-trained on the ImageNet dataset as the backbone network. The feature activations output by each stage's last residual block are used as the extracted multi-scale feature representations and denoted as $\{C_2, C_3, C_4, C_5\}$. Note that C_i (i = 2, 3, 4, 5) represents the feature level with a resolution of $1/2^i$ of the input images.

2) Three-path Feature Fusion Network: Since insulator missing-cap defects usually exist in the form of small objects [5], their features may vanish as the network deepens (pass through multiple convolutions and downsampling operation), thereby leading to poor performance of single-scale detectors. Simply detecting objects on multi-scale feature maps cannot reverse the situation, for the reason that the shallow layers of the network can better localize the object but with reduced recall performance.

FPN [18] is often used to solve above problems, however, it suffers from inefficient feature reuse that can hinder its performance in detection tasks. To be specific, FPN generates features at multiple scales, but only the features at the topmost level are used for object detection. This means that features generated at lower levels of the pyramid are not fully utilized. Therefore, inspired by [19-20], TFFN is designed to address the limitations of traditional FPN and provide a more effective way of combining features from multiple levels and scales in an image. The architecture of TFFN combines two top-down pathways and a bottom-up pathway in cross-scale connections, which allows for more accurate feature aggregation and stronger feature representations.

First, TFFN utilizes an additional bottom-up pathway to aggregate information from multiple levels of the pyramid. This operation helps mitigate the information loss that can occur in traditional FPN, where features are propagated in only one direction. By allowing information to flow along three paths, TFFN can further better preserve spatial information and maintain the quality of features at different levels of the pyramid. Secondly, TFFN enhances the semantic representation of small target features by adding an additional top-down pathway, while generating more flexible skip connections, effectively enabling multiple feature propagation along skip connections and thus richer feature aggregation at the nodes. Finally, TFFN incorporates skip connections, allowing the reuse of features from earlier layers in the network. This approach reduces the gradient disappearance problem that can occur in deep networks and helps to improve the overall accuracy of the model.

Specifically, the top-down pathway generates higher resolution features by upsampling high-level features with poor perception of details but stronger semantic information; then these features are fused with the corresponding features of the same size using the lateral connection. Thus, all the fused features will share rich semantic information. For encouraging the signal flow of localization details from shallow layers ($C_2 \rightarrow P_4^2$), the middle bottom-up pathway is built. In this path, features will be downsampled first and then fused with the corresponding left top-down features by lateral connection. In addition, each skip connection provides an additional information flow path, shortens the distance between the two paths, strengthens the feature representation, and compensates for the information loss.

Formally, the three feature fusion paths can be summarized in the following equations:

$$\begin{cases} P_n^1 = Conv^{1\times1}(C_n), & n = 5 \\ P_n^1 = Conv^{3\times3}(up(P_{n+1}^1) + Conv^{1\times1}(C_n)), & n = 4,3,2 \end{cases} \tag{1}$$

$$\begin{cases} P_n^2 = Conv^{3\times3}(down(P_{n-1}^2) + P_n^1) & n = 4,3 \\ P_n^2 = Conv^{3\times3}(P_n^1), & n = 2 \end{cases} \tag{2}$$

$$\begin{cases} P_n^{out} = down(P_{n-1}^{out}), & n = 6 \\ P_n^{out} = Conv^{3\times3}(down(P_{n-1}^2) + P_n^1), & n = 5 \\ P_n^{out} = Conv^{3\times3}(up(P_{n+1}^{out}) + P_n^2 + P_n^1) & n = 4,3,2 \end{cases} \tag{3}$$

where *up* is the nearest-neighbor interpolation upsampling operation, *down* is the max-pooling downsampling operation,

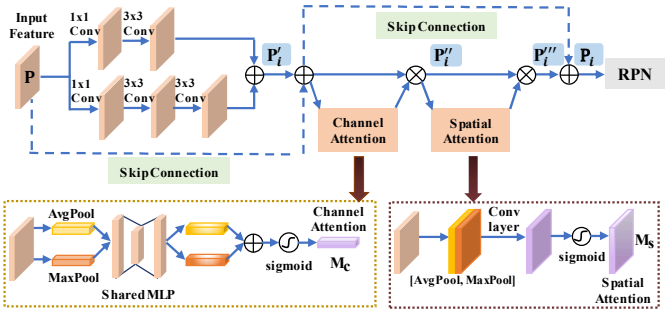


Fig. 4. Flowchart of RFA+ block. Note that ⊕ represents element-wise addition, and ⊗ represents element-wise product.

$Conv^{n\times n}$ represents the $n\times n$ convolution, and + denotes the element-wise addition.

The proposed TFFN accurately combines the spatial details stored in high-resolution features with semantically stronger information to achieve enhanced representation of small missing-cap defect features.

3) Enhanced Receptive Field Attention Block: Due to the limited proportion of missing-cap defects in overhead transmission line images, a substantial amount of non-critical information occupies a significant portion of the image. Given the constrained capacity of the network to process multiple sources of information, the attention mechanism is widely applied to selectively emphasize the valuable features. Moreover, the variation in object and scene sizes in input images necessitates the network's ability to extract features at multiple scales. However, the attention mechanism is constrained in its ability to gather information from a single receptive field. Consequently, the network is only capable of capturing features at a specific scale, rendering it inadequate for deployment in the aforementioned scenario. Thus, we first construct an enhanced receptive field structure that allows the network to linearly aggregate multi-scale information from different branches to improve detection accuracy and increase its adaptability to scale variance, eventually making accurate judgments about targets with different aspect ratios. Furthermore, two attention modules [21] are applied to strengthen highly correlated missing-cap defect features while weakening non-critical features or noises. As a result, this attention structure achieves a detailed distinction between defects and backgrounds. In short, we refer to the structure in Fig. 4 as the RFA+ block.

The forward process is presented below. Let one of the feature maps refined by TFFN is applied as input feature P of the RFA+ block. The receptive field enhancement process can be represented by the following equation:

$$P_i' = f(P) + P \tag{4}$$

where P_i' represents the output feature, and f represents the multiple receptive field convolution operation. Note that the parameters are shown in Table I. In detail, a 1×1 convolutional layer is first utilized to reduce the computation, and then a convolutional layer with a 3×3 receptive field is appended to extract features. Similarly, the 1×1 convolutional layer of branch II is followed by two stacked 3×3 convolutional layers, and such a structure is identical to the convolution process with a 5×5 receptive field but possesses fewer parameters.

TABLE I
THE CONVOLUTION KERNEL PARAMETERS OF THE TWO BRANCHES

Input ($H\times W\times C$)	Branch I	Branch II
P_2^{out}	1×1×64; 3×3×C	1×1×64; 3×3×C; 3×3×C
P_3^{out}	1×1×128; 3×3×C	1×1×128; 3×3×C; 3×3×C
P_4^{out}	1×1×160; 3×3×C	1×1×160; 3×3×C; 3×3×C
P_5^{out}	1×1×192; 3×3×C	1×1×192; 3×3×C; 3×3×C

The H, W, C represent the height, width, and channel, respectively.

Then, two kinds of attention mechanisms are acted on P_i' sequentially. Firstly, the channel attention module (CAM) enhances or suppresses different channels by modeling the importance of each feature channel. To be specific, two feature vectors $P_{max}^{c} \in R^{C\times1\times1}$ and $P_{avg}^{c} \in R^{C\times1\times1}$ generated by global max-pooling and global average-pooling, respectively, share a multi-layer perceptron (MLP) and further add together. Then, the sigmoid activation layer produces the channel attention weights $M_C \in R^{C\times1\times1}$. Finally, the weights are element-wisely multiplied by P_i' to obtain a new feature map $P_i'' \in R^{C\times H\times W}$. The calculation process is as follows:

$$P_i'' = P_i' \otimes \sigma(M_1 M_2(P_{max}^{c}) + M_1 M_2(P_{avg}^{c})) \tag{5}$$

where M_1 and M_2 are the weight parameters of two layers of the MLP, respectively. σ is the sigmoid activation function, ⊗ represents the element-wise product.

Unlike the CAM, the spatial attention module (SAM) mainly assigns weights to spatial positions. The feature maps $P_{avg}^{s} \in R^{1\times H\times W}$ and $P_{max}^{s} \in R^{1\times H\times W}$ are, respectively, generated after global average-pooling and global max-pooling along the channel dimension over the feature map P_i''. They are connected by using the concatenation method and then pass through a convolutional layer. The sigmoid activation layer applies to get the spatial attention weights $M_S \in R^{1\times H\times W}$. Finally, the weights are element-wisely multiplied by $P_i'' \in R^{C\times H\times W}$ to obtain a refined feature map $P_i''' \in R^{C\times H\times W}$, which is defined as:

$$P_i''' = P_i'' \otimes \sigma(Conv^{7\times7}([P_{avg}^{s}, P_{max}^{s}])) \tag{6}$$

where $Conv^{7\times7}$ represents the convolution operation with the kernel size of 7×7, [⊔] represents the concatenation method. At the end of the RFA+ block, the feature map $P_i \in R^{C\times H\times W}$ is produced by the skip connection. The following equation expresses this process:

$$P_i = P_i' + P_i''' \tag{7}$$

In summary, the feature maps $\{P_2^{out}, P_3^{out}, P_4^{out}, P_5^{out}, P_6^{out}\}$ output by TFFN, except for P_6^{out}, are input into the RFA+ block to get the refined feature map $\{P_2, P_3, P_4, P_5\}$. Next, these feature maps $\{P_2, P_3, P_4, P_5, P_6^{out}\}$ are fed into RPN.

B. Region Proposal Network

The RPN is implemented with a sliding window (a 3×3 convolutional layer) followed by two sibling 1×1 convolutional layers for classification and bounding box regression, respectively (more details in [16]). Generally, the RPN generates candidate boxes and then maps these candidate boxes to the input feature maps at different resolutions to obtain the region of interests (RoIs).

C. Defect Detection Network

1) Context Perception Module: The RoI pooling method only considers the information within the region of interest,

while neglecting the information from surrounding pixels. In this study，missing-cap defects may be located in unexpected areas or within groups of similar-looking objects in complex backgrounds. Capturing contextual information (the relationship between the target and its surroundings) is well adapted to distracting factors in the input image [22-23]. Ref. [22] and Ref. [23] have manually defined the context region as double the size of the RoI and as eight directions surrounding the original RoI, respectively. However, these methods are limited by the manual selection of the context region, which may constrain the generalization ability of the model by over-fitting to specific scenarios. Thus, inspired by Inception-ResNet V2 [24], a novel context perception module (CPM) utilizes different convolutional kernels to automatically determine the context region, with preserving the spatial resolution (see Fig. 5). By fusing the feature vector output from the RoI pooling layer with context information from different receptive fields, our proposed approach can more effectively leverage global image information to enhance the accuracy of missing-cap defect detection. This integration not only retains essential information from the RoI but also utilizes the context information for correction and supplementation. Compared with manual settings methods, the CPM can better exploit the learning capability of the model, and improve the model detection performance.

To be specific, the CPM can sense larger regions in the input space by n (n=1, 2) stacking convolution layers with the kernel size of 3×3 or max-pooling to integrate more contextual information. Then, after the element-wise addition operation, a convolutional layer is applied to ensure the discriminability of the detected features. Functionally, this module can help the defect detection network generate regression offsets closer to the ground-truth based on the relevance of the object to its surrounding region. Ultimately, the spatial localization of the target is more accurate, contributing to a reduction in the false detection rate of small targets.

2) The Architecture of Defect Detection Network: With the RoIs produced by RPN, we first need to convert RoIs of different sizes into fixed-resolution feature maps of n×n through the RoI pooling layer [16]. This paper sets n=14. Then the fixed-resolution feature maps are refined by the CPM and then mapped to a feature vector by two fully connected layers (FCs). Finally, an FC layer with an additional softmax layer is used to output the probability vector (cls-logit) over k object classes; simultaneously, another parallel FC layer outputs the predicted position offsets (bbox-pred) of each proposal.

D. Loss Function

In I2D-Net, the following multi-task loss function is utilized to train our network:

$$L(\{p_i\},\{u_i\}) = \frac{1}{N_{cls}} L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (8)$$

where i is the index of an anchor, L_{cls} is the cross-entropy loss, and λ is a factor used to balance the ratio of regression and classification loss. p_i and t_i respectively denote the predicted category probabilities and regression offsets. p_i^* and t_i^* respectively denote the ground truth, note that p_i^* before L_{reg}, limits only the positive samples to calculate box regression loss

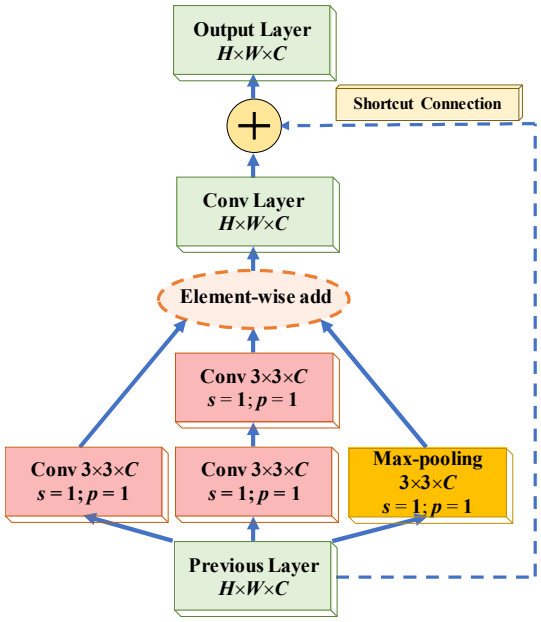


Fig. 5. Architecture of CPM. Each branch has a specific receptive field while corresponding to the semantics of each level. Note that s represents the stride, and p represents the padding.

(p_i^*=1) and is disabled otherwise (p_i^*=0). L_{reg} represents regression loss function, which uses robust loss (smooth L_1) in [16], and the expression is as follows:

$$smooth\ L_1(x) = \begin{cases} 0.5x^2 & if\ |x| < 1, \\ |x| - 0.5 & otherwise. \end{cases} \quad (9)$$

Moreover, we set N_{cls}=256 and N_{reg}=2400 to balance the two terms and then weight them by λ=10. To make the network easier to learn, we regress the offsets of the bounding box rather than its absolute coordinates, that is:

$$t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a \quad (10)$$

$$t_w = \log(w/w_a), t_h = \log(h/h_a) \quad (11)$$

$$t_x^* = (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a \quad (12)$$

$$t_w^* = \log(w^*/w_a), t_h^* = \log(h^*/h_a) \quad (13)$$

where x, y represent the center coordinates of the box, and w, h denote the width and height. x, x_a, and x^* correspond to the predicted box, anchor box, and ground-truth box, respectively (the same as y, w, h).

III. Experimental Results

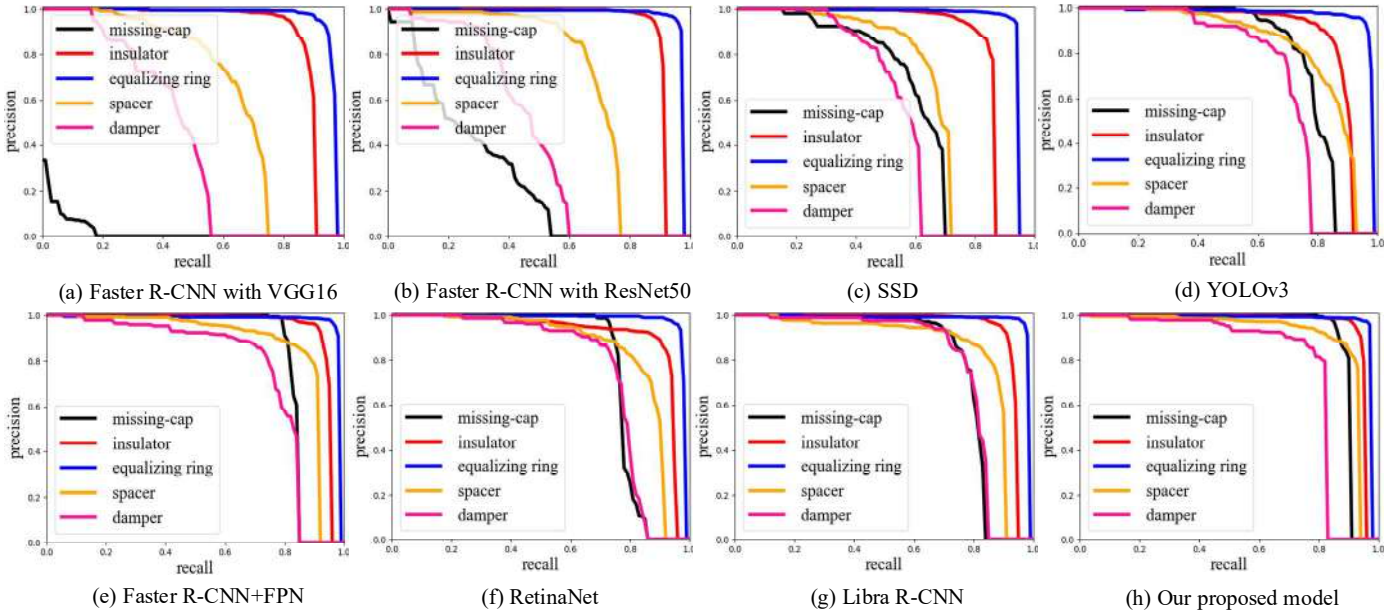
This section presents the experimental results, including dataset establishment, implementation details, evaluation metrics, comparison results, ablation studies, robustness discussion, and results on other datasets.

A. Dataset

Our dataset consists of two parts: the CPLID dataset [5] of 848 images, and the other part contains the dataset of 220 images of 220kV transmission lines captured by UAVs during daily inspections. There are five common object categories in transmission line images, including insulator, missing-cap defect, equalizing ring, spacer, and damper. To prevent

TABLE II
COMPARISON OF DIFFERENT MODELS WITH OUR SUGGESTED MODEL

Model	Backbone	mAP (%)	AP (%)					Speed (ms)	Params (M)
			missing-cap	insulator	equalizing ring	spacer	damper		
Faster R-CNN [16]	VGG-16 [26]	58.6	2.2	88.1	95.7	64.0	42.8	116.3	134.7
Faster R-CNN [16]	ResNet-50 [17]	65.5	26.9	90.3	96.3	69.2	44.9	119.0	165.1
SSD [14]	ResNet-50 [17]	71.3	59.8	84.4	93.4	65.1	53.6	64.5	**13.6**
YOLOv3 [12]	Darknet-53 [12]	83.2	81.3	89.5	90.3	81.9	73.3	58.8	61.6
Faster R-CNN+FPN [18]	ResNet-50 [17]	87.4	83.0	93.9	97.1	86.9	76.2	117.6	41.3
RetinaNet [27]	ResNet-50 [17]	85.3	77.9	91.1	97.1	84.5	75.8	103.2	32.3
Libra R-CNN [28]	ResNet-50 [17]	86.6	79.2	92.9	97.2	84.6	**78.9**	122.0	41.5
Camp-Net3 [22]	ResNet-50 [17]	87.7	84.9	94.2	97.2	88.6	73.5	133.3	41.7
BS-YOLO [29]	CSPNet [31]	86.9	85.4	93.3	97.0	87.7	71.2	**30.7**	63.9
FINet [30]	CSPNet [31]	88.5	86.0	**94.8**	**98.8**	87.3	75.6	90.1	86.3
I2D-Net	ResNet-50 [17]	**89.6**	**89.4**	94.4	96.4	**89.8**	78.2	156.2	87.5



(a) Faster R-CNN with VGG16 (b) Faster R-CNN with ResNet50 (c) SSD (d) YOLOv3

(e) Faster R-CNN+FPN (f) RetinaNet (g) Libra R-CNN (h) Our proposed model

Fig. 7. The P-R curves of different models.

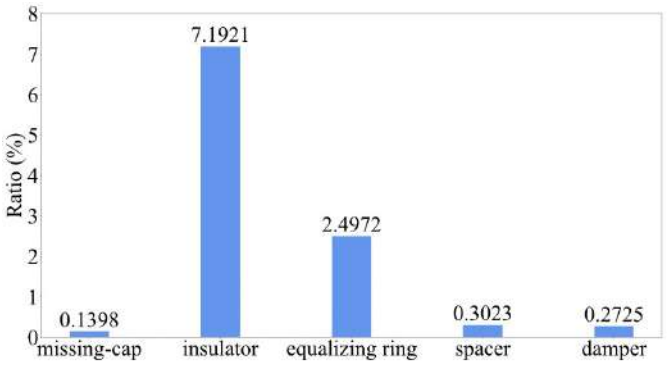


Fig. 6. The ratio of average area of target to original image (%).

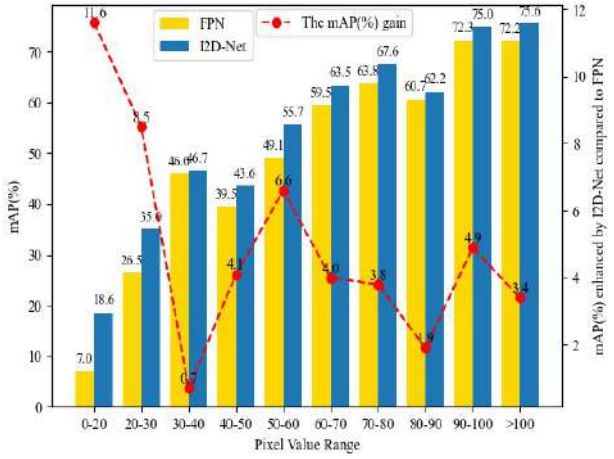


Fig. 8. The detection accuracy of targets with different sizes in the test set.

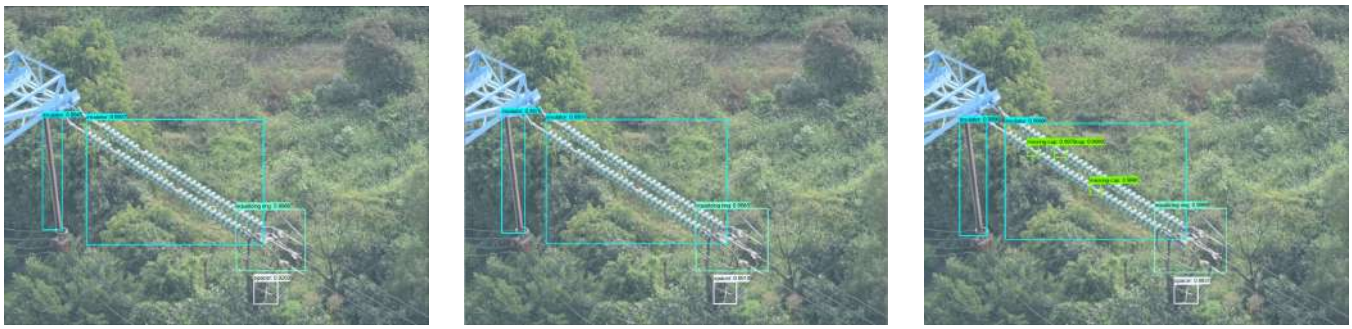
overfitting, we perform data augmentation, including brightness adjustment, rotation, and horizontal flip, which increases our dataset to a total of 1881 images. Then, we split the dataset into a training set (1278 images), a validation set (226 images), and a test set (377 images) according to the ratio of 75:15:20.

B. Implementation Details and Evaluation Metrics

The experimental hardware consists of a PC equipped with an Intel Core i9-9900K CPU and an NVIDIA GeForce RTX 2070 GPU, running on the Windows 10 operating system.

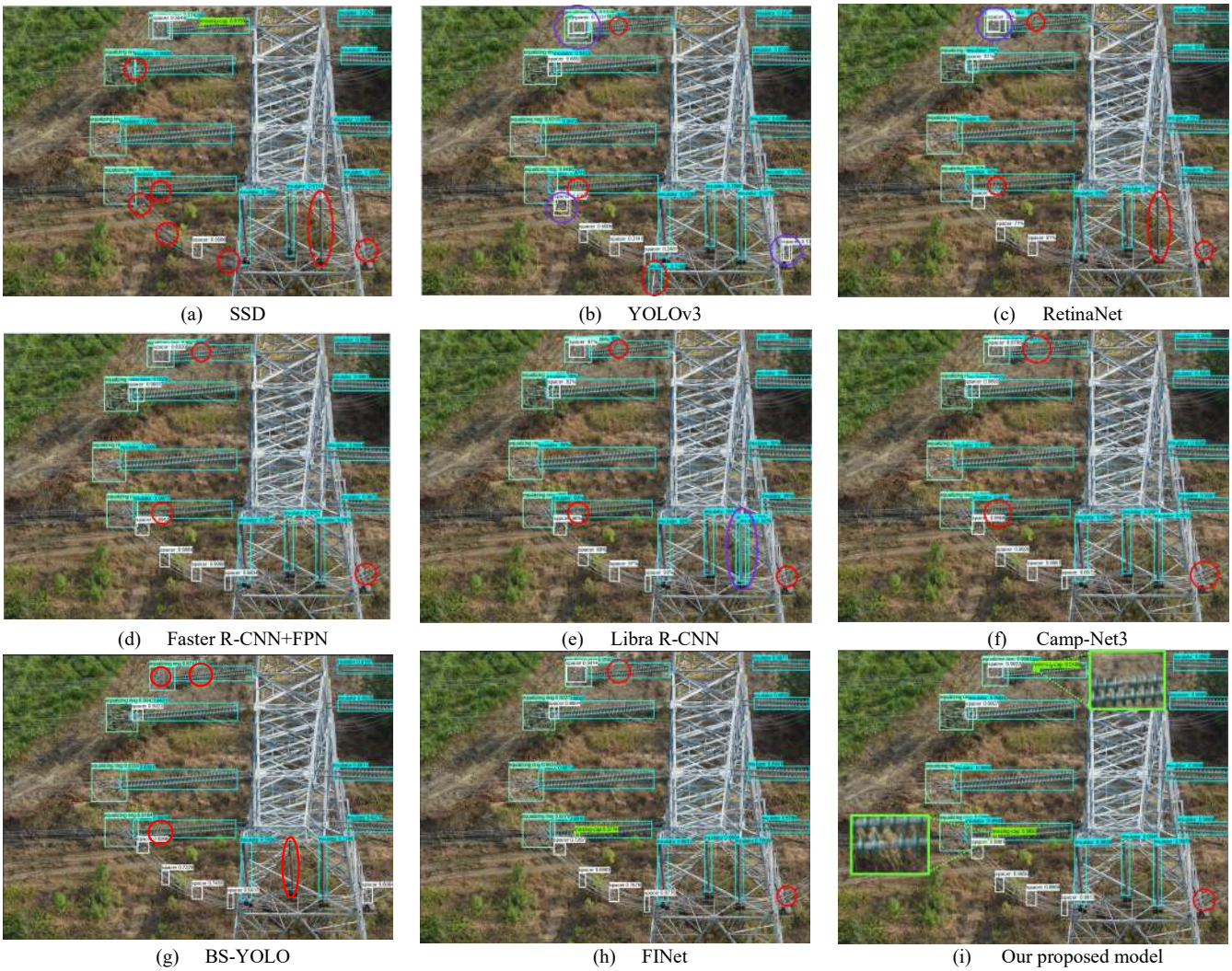
Our model is implemented using the PyTorch framework and trained for 15 epochs with a batch size of 1. Stochastic gradient descent (SGD) optimizer is employed with a momentum of 0.9 and a weight decay of 0.0001, with the learning rate set to 0.004 for the first 10 epochs and 0.0004 for the remaining 5 epochs. The scales of anchors are set to $\{32^2, 64^2, 128^2, 256^2, 512^2\}$, with three aspect ratios $\{0.5, 1, 2\}$.

In this paper, the mean average precision (mAP) [25] is utilized as the evaluation metric to quantify the accuracy of different detection networks.



(a) Faster R-CNN with VGG16 (b) Faster R-CNN with ResNet-50 (c) Our proposed model

Fig. 9. Comparison results with traditional detection models on the test set.



(a) SSD (b) YOLOv3 (c) RetinaNet

(d) Faster R-CNN+FPN (e) Libra R-CNN (f) Camp-Net3

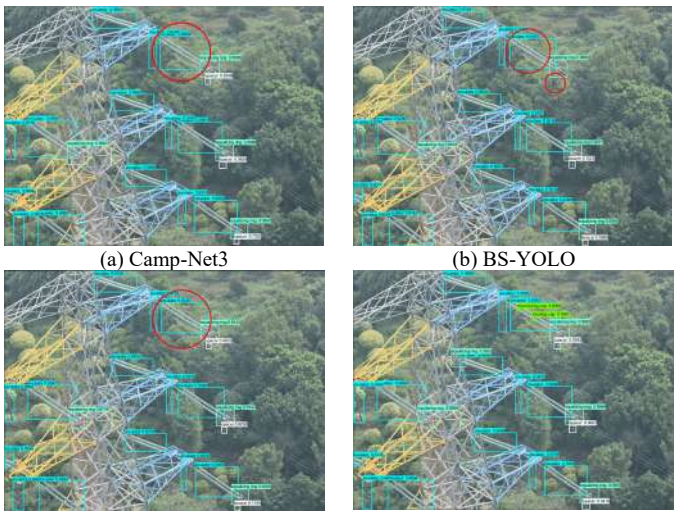
(g) BS-YOLO (h) FINet (i) Our proposed model

Fig. 10. Detection results on the test set using different models. The rectangle represents the predicted detection box; specifically, the green, blue, mint, and white rectangles represent missing-cap defect, insulator, equalizing ring, and spacer, respectively. The red circle represents the missed or false detection, and the purple circle represents duplicate overlapping detection boxes.

C. Comparison Results

1) Performance Comparison with Existing Models: Our suggested I2D-Net is compared with commonly used and advanced object detection models, such as Faster R-CNN [16], SSD [14], YOLOv3 [12], RetinaNet [27], Libra R-CNN [28], Camp-Net3 [22], BS-YOLO [29], and FINet [30].

The overall performance comparison results are shown in Table II, it can be noticed that our suggested model outperforms other models in terms of mAP. Note that Faster R-CNN with ResNet-50 has an mAP of 65.5%, and the AP of missing-cap defect is only 26.9%. In contrast, the mAP of the proposed model is 89.6%, and especially the AP of missing-cap defect is 89.4%, which is 62.5 points higher than that of Faster R-CNN with ResNet-50. Compared with the Faster R-CNN+FPN, our model improves the mAP by 2.2 points, and with an increase of 6.4, 0.5, 2.9 and 2.0 points for missing-cap defects, insulators, spacers and dampers, respectively. Note that the detection



(a) Camp-Net3 (b) BS-YOLO

(c) FINet (d) Our proposed model

Fig.11 Comparison results with various models in a highly complex background. The red circle represents the missed detection area.

results of the single-scale detectors (Faster R-CNN [16]) are not satisfactory for the missing-cap defects. The reason is that the downsampling operation will lead to the features of missing-cap defects disappearing, resulting in failure detection on the low-resolution feature map. With respect to the insulator and equalizing ring, the detection effects are excellent due to their larger sizes (see Fig. 6), and there is no significant difference between the various methods. In comparison, the AP values of missing-cap defect, spacer, and damper show a great contrast in different models. These objects occupying small sizes, especially the missing-cap defect, easily lose related features as the network deepens, and are mistaken as redundant features due to their high similarity to the background. Compared to the existing model, I2D-Net achieves the highest mAP with a value of 89.6%, in particular reaching 89.4% for the AP of missing-cap defects. This improvement validates the advantages of our proposed method.

Moreover, the precision and recall (P-R) curves of different models are displayed in Fig. 7. The larger the area enclosed by the P-R curve, the better the performance of the model. It can be seen in Fig. 7 that these AP curves in our proposed model are generally higher than other models, especially with a qualitative leap compared to the results in the second row.

In order to more fully validate the effectiveness of our method in identifying targets of different scales, we compute the corresponding mAP values for targets of different pixel sizes in the same test set. Comparison results from I2D-Net and Faster R-CNN+FPN are presented in Fig. 8. Obviously, our method I2D-Net achieves a remarkable improvement of 11.6 points at the mAP for targets in the 0-20 pixel range, Moreover, for targets in the 20-30 pixel range, the proposed method obtains a significant increase of 8.5 mAP points. In addition, for other targets in different pixel ranges, the I2D-Net generates higher mAPs than the baseline model. which confirms the advantage of our proposed method.

2) Comparison of Speed and Parameters: Table II presents the test speed and parameter count of different methods. Speed refers to the time required to test a single image. Our method takes an additional 0.0386s compared to Faster R-CNN+FPN, but significantly improves the AP for missing-cap by 7.7%. The Faster R-CNN with VGG16 and ResNet-50 models, with 134.7M and 165.1M parameters, respectively, have mAP of only 58.6% and 65.5%. In contrast, the proposed method has fewer parameters and achieves th e highest mAP of 89.6%.

The detection of transmission line inspection images is usually in the mode of offline detection [22]. Considering that parameters are stored in 4-byte format in a computer, the size of the proposed model is 350M (87.5×4). Although our model does not run fast enough, fortunately, nowadays computers are not hard to store a 350M file [22]. As the application scenario for this novel model is to be deployed on a computer for offline detection without requiring real-time performance, it is worthwhile to obtain a higher detection accuracy with an unremarkable increase in time.

3) Visualization Experiments: This section presents visualization experiments to further illustrate the performance of our suggested method.

In Fig. 9, both conventional target detection methods fail to detect the missing-caps, while our method successfully detects all defective targets. Subsequently, as shown in Fig. 10(a-i), except for the SSD model and FINet model, which detect only one of two missing-cap defects, none of the other models identifies any missing-cap defect. However, our model discovers all defective objects. Because it can adequately incorporate high-level features that capture global information with shallow features that preserve spatial details of defects. Thus, it effectively alleviates related feature loss of missing-cap defects and achieves accurate detection. Although the Faster R-CNN+FPN model also implements feature fusion, the detection effect of missing-cap defect is worse [see Fig. 10 (d)]. Concerning these object categories of equalizing ring and insulator, most of them have a larger size and distinguishable features. Therefore, all models can detect them in deep feature layers with a high confidence coefficient. Note that the YOLOv3, BS-YOLO, and FINet models (see Fig. 10) have a lower confidence coefficient in detecting targets compared to the proposed I2D-Net, which can affect the reliability of the predictions in practical engineering applications. Furthermore, the visualization of the detection comparison results in Fig.11 shows our method outperforms the other three methods by detecting missing-cap defects against a highly complex background. Notably, all three of the other methods miss these small-sized target defects.

To further explain the effectiveness of our model, we show the class activation maps of the specific feature map P_2 output by the RFA+ block in Fig. 12. As shown in Fig. 12(c), semantic aliasing occurs in the intersection of receptive fields between targets, and the class activation region is blurred. However, the class activation region in Fig. 12(d) shows that the ambiguous features on the edges are suppressed, contributing to more accurate localization of objects. Although both models adopt feature fusion, our I2D-Net embeds the RFA+ block which increases the observation precision of the target area to obtain more detailed information. Besides, it has the ability to precisely distinguish missing-cap defects from the background or other fittings.

TABLE III
ABLATION STUDIES OF THE PROPOSED I2D-NET ON OUR TRANSMISSION LINE INSPECTION IMAGE DATASET

Feature Fusion			RFA+	CPM	mAP (%)	AP (%)					Speed	Params
FPN	PANet	TFFN				missing-cap	insulator	equalizing ring	spacer	damper	(ms)	(M)
✓	-	-	-	-	87.4	83.0	93.9	97.1	86.9	76.2	117.6	41.3
-	✓	-	-	-	87.6	83.5	93.2	97.3	87.7	76.4	120.0	43.1
-	-	✓	-	-	88.1 (+0.7)	85.4	93.2	97.3	87.3	77.2	126.4	44.8
-	-	✓	✓	-	89.1 (+1.0)	88.0	94.4	96.2	89.2	77.8	151.5	85.5
-	-	✓	✓	✓	89.6 (+0.5)	89.4	94.4	96.4	89.8	78.2	156.2	87.5

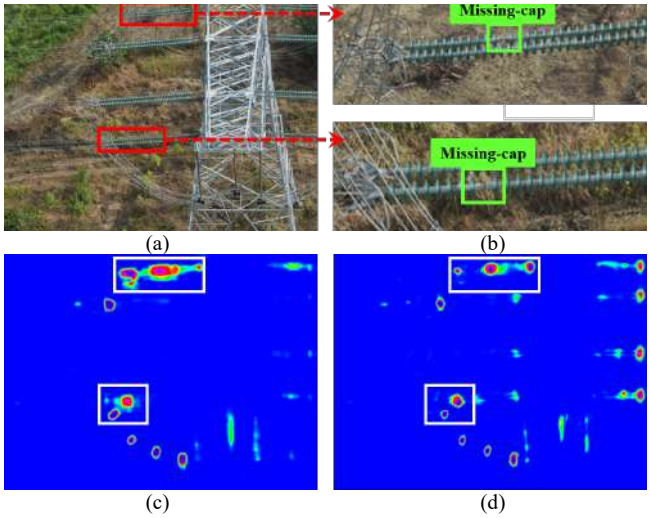


Fig. 12. Class Activation Maps. (a): The original image. (b): Enlarge images of defect area marked by the red boxes. (c)-(d): Class activation maps from Faster R-CNN+FPN (Left) and I2D-Net (Right). The white box represents the activated region of the missing-cap defect and its surroundings.

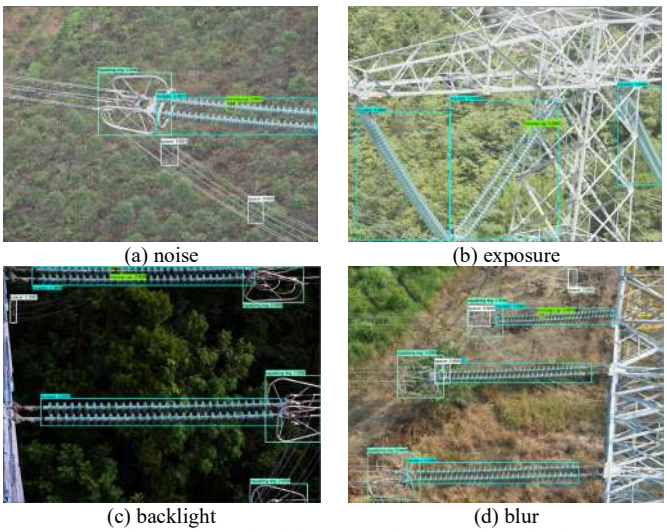


Fig. 13. Detection results under different interferences. The noise is a Gaussian noise with a mean of 0 and a variance of 0.1; the exposure and backlight are processed using the gamma function with exponential parameters set to 0.5 and 2.5 respectively, and the blur is the GaussianBlur method with a convolution kernel size of 3.

D. Ablation Studies

1) Effect of Three-path Feature Fusion Network: To verify the effect of the proposed TFFN, we conduct a comparative experiment with FPN and the path aggregation network (PANet) [19]. As shown in the first three rows of Table III, the TFFN model achieves 0.7-point and 0.5-point higher mAP than the FPN-based and PANet-based models, respectively, with a particularly significant increase of 2.4 and 1.9 points in AP for missing-cap defects. The designed TFFN overcomes the limitation of traditional FPN which suffers from information loss and fails to fully utilize the features at different levels. These experimental results illustrate that TFFN improves the accuracy for small-sized missing-cap defects with a slightly reduced speed.

TABLE IV
THE PERFORMANCE COMPARISON OF DISTORTION TEST IMAGES

Model	Backbone	mAP(%)				
		N.A.	noise	exposure	backlight	blur
FRCNN	VGG-16	58.6	50.7	51.7	51.6	55.1
FRCNN	ResNet-50	65.5	63.0	60.5	60.2	62.7
SSD	ResNet-50	71.3	70.3	70.6	67.1	70.5
YOLOv3	Darknet-53	83.2	82.7	82.0	79.9	81.3
FR-FPN	ResNet-50	87.4	82.8	84.4	84.4	84.1
Re-Net	ResNet-50	85.3	82.1	82.7	82.6	83.1
LRCNN	ResNet-50	86.6	84.3	82.3	83.2	84.6
Camp-Net3	ResNet-50	87.7	81.5	84.6	84.0	84.4
BS-YOLO	CSPNet	86.9	83.1	83.9	83.8	83.4
FINet	CSPNet	88.5	83.4	85.1	84.1	85.0
I2D-Net	ResNet-50	**89.6**	**84.6**	**86.3**	**84.9**	**85.4**

N.A. means that no degradation processing is performed on the test images. FRCNN represents the Faster RCNN model, FR-FPN denotes the Faster RCNN+FPN, Re-Net is the RetinaNet model, and LRCNN means the Libra R-CNN.

2) Effect of the Proposed I2D-Net: Ablation studies are utilized to understand the effectiveness of our proposed approach better. Table III presents the experimental results, which demonstrate how each submodule affects the performance of our method. It can be proved that each module positively affects the detection results. As can be seen from the last two rows in Table III, the addition of the RFA+ block and CPM increases the mAP of the entire model by 1.0 and 0.5 points, respectively.

E. Robustness Discussion

To analyze the effectiveness of our proposed method more rationally and comprehensively under many uncertainties, robustness experiments are conducted. Electromagnetic interference caused by drones, insufficient or excessive light, and shaky filming are considered in overhead transmission line inspection tasks, we simulate the above environment, thus adding four attacks to the image: noise, exposure, backlight, and blurring, for discussion.

As illustrated in Table IV, all methods exhibit performance degradation after various image distortion treatments. The impact of different interferences on the prediction of the proposed model is minimal. Compared to the other methods, I2D-Net maintains the best detection performance with the highest mAP value in the presence of disturbances.

The detection results from the proposed method for transmission line inspection images under different interferences are displayed in Fig. 13.

TABLE V
DOTAV1.5 TEST SET DETECTION RESULTS

Model	Backbone	mAP(%)
Faster R-CNN+FPN	ResNet-50	82.9
Libra R-CNN	ResNet-50	82.3
Camp-Net3	ResNet-50	83.2
I2D-Net	ResNet-50	**86.6**

TABLE VI
TT100K TEST SET DETECTION RESULTS

Model	Backbone	mAP(%)
Faster R-CNN+FPN	ResNet-50	57.8
Libra R-CNN	ResNet-50	60.4
Camp-Net3	ResNet-50	60.5
I2D-Net	ResNet-50	**71.6**

F. Results on DOTAV1.5 and TT100K Datasets

To further evaluate the proposed model, the experiments are implemented on the DOTAV1.5 and TT100K datasets. Tables V and VI show the comparison results with Faster R-CNN+FPN, Libra R-CNN and Camp-Net3 on the DOTAV1.5 and TT100K datasets respectively. As shown in Table V, I2D-Net attains the highest mAP value of 86.6% on the DOTAV1.5 dataset, which is 3.7 points higher than the baseline model. Moreover, from Table VI, I2D-Net obtains the highest mAP among all methods with 71.6% on the TT100K dataset. Therefore, the evaluation results demonstrate that the proposed model achieves better detection performance.

IV. CONCLUSION

In this paper, a high-accuracy CNN-based model (called I2D-Net) is proposed to address the problem of small-sized insulator missing-cap defect detection. The specific conclusions are as follows.

1) The proposed TFFN can provide a rich information flow to facilitate the appropriate fusion of shallow features and deep features. Compared to the baseline feature fusion network, our network has more powerful feature fusion capabilities, which makes the feature information of the small-sized missing-cap defect can be distributed in the multi-scale feature maps.

2) The presented RFA+ module can refine the fused feature. To be specific, the module provides a more diverse receptive field to better capture objects in different aspect ratios and emphasizes the features of objects by the attention module.

3) The CPM module can absorb contextual information without particularly deepening the network depth, which strengthens the performance of small missing-cap defects.

4) Experimental results prove that our I2D-Net outperforms other transmission line detection models in detection accuracy, which provides a practical application scheme for insulator missing-cap defect detection, as well as a research strategy for the intelligent inspection of overhead transmission lines.

REFERENCES

[1] M. Marzinotto, G. Mazzanti, E. A. Cherney and G. Pirovano, "An innovative procedure for testing RTV and composite insulators sampled from service in search of diagnostic quantities," *IEEE Elect. Insul. Mag.*, vol. 34, no. 5, pp. 27-38, Sep. 2018.

[2] M. F. Palangar, S. Mohseni, M. Mirzaie and A. Mahmoudi, "Designing an automatic detector device to diagnose insulator state on overhead distribution lines," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1072-1082, Feb. 2022.

[3] G. Jaensch, H. Hoffmann and A. Markees, "Locating defects in high voltage transmission lines," in *Proc. IEEE Int. Conf. Transm. Distrib. Constr. Live Line Maint. ESM.*, 1998, pp. 179-186.

[4] S. F. Stefenon, R. B. Grebogi, R. Z. Freire, A. Nied and L. H. Meyer, "Optimized ensemble extreme learning machine for classification of electrical insulators conditions," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5170-5178, Jun. 2020.

[5] 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.

[6] A. Ibrahim, A. Dalbah, A. Abualsaud, U. Tariq and A. El-Hag, "Application of machine learning to evaluate insulator surface erosion," *IEEE Trans. Instrum. Meas*., vol. 69, no. 2, pp. 314-316, Feb. 2020.

[7] P. Luo, B. Wang, H. Wang, F. Ma, H. Ma and L. Wang, "An ultrasmall bolt defect detection method for transmission line inspection," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1-12, 2023.

[8] Z. Zhao, H. Qi, Y. Qi, K. Zhang, Y. Zhai and W. Zhao, "Detection method based on automatic visual shape clustering for pin-missing defect in Transmission Lines," *IEEE Trans. Instrum. Meas*., vol. 69, no.9, pp. 6080-6091, Sep. 2020.

[9] W. Du, H. Shen and J. Fu, "Automatic defect segmentation in X-ray images based on deep learning," *IEEE Trans. Ind. Electron.*, vol. 68, no. 12, pp. 12912-12920, Dec. 2021.

[10] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas*., vol. 69, no. 4, pp. 1493-1504, Apr. 2020.

[11] Y. Liu, X. Ji, S. Pei, Z. Ma, G. Zhang, Y. Lin and Y. Chen, "Research on automatic location and recognition of insulators in substation based on YOLOv3," *High Volt.*, vol. 5, no. 1, pp. 62-68, 2020.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[13] W. Wang, Z. Wang, B. Liu, Y. Yang and X. Sun, "Typical defect detection technology of transmission line based on deep learning," in *Proc. Chin. Automat. Congr.*, Hangzhou, China, Nov. 2019, pp. 1185–1189.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Euro. Conf. Comput. Vis.*, 2016, pp. 21-37.

[15] X. Lei and Z. Sui, "Intelligent fault detection of high voltage line based on the Faster R-CNN," *Measurement*, vol. 138, pp. 379-385, May. 2019.

[16] 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.

[17] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.

[18] 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.*, 2017, pp. 936-944.

[19] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis*. *Pattern Recognit*., 2018, pp. 8759-8768.

[20] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10778-10787.

[21] S. Woo, J. Park, J.-Y. Lee *et al*., "CBAM: convolutional block attention module," in *Proc. Euro. Conf. Comput. Vis.*, 2018, pp. 3-19.

[22] R. Jiao, Y. Liu, H. He, X. Ma and Z. Li, "A deep learning model for small-size defective components detection in power transmission tower," *IEEE Trans. Power Deliv*., vol. 37, no. 4, pp. 2551-2561, Aug. 2022.

[23] P. Fang and Y. Shi, "Small object detection using context information fusion in faster R-CNN," in *Proc. IEEE Conf. Comput. Comput. Commun.*, 2018.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis*., vol. 88, no. 2, pp. 303-338, Jun. 2010.
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014.
[27] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318-327, Feb. 2020.
[28] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821-830.
[29] K. Hao, G. Chen, L. Zhao, Z. Li, Y. Liu and C. Wang, "An insulator defect detection model in aerial images based on multiscale feature pyramid network," *IEEE Trans. Instrum. Meas*., vol. 71, pp. 1-12, 2022.
[30] Z. -D. Zhang et al., "FINet: An insulator dataset and detection benchmark based on synthetic fog and improved YOLOv5," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-8, 2022.
[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal 730 speed and accuracy of object detection," 2020, *arXiv:2004.10934.*

Qi Fu was born in Hunan, China, in 2000. She is currently pursuing the M.S. degree in electrical engineering with the School of Electrical Engineering, Guangxi University, China.

Her major research interests include electric power equipment intelligent fault diagnosis.

Jiefeng Liu (SM'15) received the M.S. degree and Ph.D. degree in electrical engineering from Chongqing University, Chongqing, China, in 2011 and 2015, respectively. In 2018, he joined Guangxi University where he is an associate professor.

His major research interests include the field of condition assessment and fault diagnosis for high voltage equipment.

Xingtuo Zhang was born in Sichuan, China, in 1992. He is currently working toward the Ph.D. degree in electrical engineering with the School of Electrical Engineering, Guangxi University, China.

His major research interests include electric power equipment intelligent fault diagnosis.

Yiyi Zhang (M'14) received the Bachelor and Ph.D. degrees in electrical engineering in 2008 and 2014, respectively from Guangxi University, Nanning, China and Chongqing University, Chongqing, China. In 2014, he joined Guangxi University where he is an associate professor.

His current research interests include the intelligent diagnosis for transformers.

Yang Ou was born in Guangxi, China, in 1999. He is currently pursuing the M.S. degree in electrical engineering with the School of Electrical Engineering, Guangxi University, China.

His major research interests include electric power equipment intelligent fault diagnosis.

Runnong Jiao was born in Shandong, China, in 1999. He is currently pursuing the M.S. degree in electrical engineering with the School of Electrical Engineering, Guangxi University, China.

His major research interests include electric power equipment intelligent fault diagnosis.

Chuanyang Li (M) received the B.S. degrees and M.S degree of Electrical Engineering in Taiyuan University of Technology in 2011 and 2014. He received Ph.D. degree in Electrical Engineering in Tsinghua University in 2018. He works as a Postdoctoral Fellow at the Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" of the University of Bologna, Italy. His research interests include surface charge behavior, HVDC GIL, online monitoring of motor/cable insulation.

Giovanni Mazzanti (M'04–SM'15) is an Associate Professor of HV Engineering and Power Quality at the University of Bologna, Italy. His interests are life modeling, reliability and diagnostics of HV insulation, power quality, renewables and human exposure to electro-magnetic fields. Since 2009 he is consultant of TERNA (the Italian TSO) in the HVDC and HVAC cable systems area. He is author or coauthor of more than 200 published papers, and coauthor of the book HVDC Extruded Cable Systems: Advances in Research and Development, Wiley-IEEE Press July 2013. He is chairman of the IEEE DEIS TC "HVDC Cable Systems".