Neural Computing and Applications (2025) 37:7067–7087
https://doi.org/10.1007/s00521-024-10957-x

REVIEW

An adaptive dual-weighted feature network for insulator detection in transmission lines

Jie Zhang[1] · Xiabing Wang[1] · Yinhua Li[1] · Dailin Li[1] · Fengxian Wang[1] · Linwei Li[1] · Huanlong Zhang[1] · Xiaoping Shi[2]

Received: 26 November 2023 / Accepted: 13 December 2024 / Published online: 11 February 2025

Abstract
In the field of electrical power applications, high-voltage insulators necessitate routine inspection to assure the security and stability of the whole electric power system operation. Accurately positioning the insulator is extremely crucial for proceeding to the insulator defect detection. However, during UAV electrical line inspection, the presence of the electric power line magnetic field engenders a reduction in the pixel representation of the insulator within the image data, thereby diminishing the accuracy of insulator detection. In response to the prevailing issues, we present the creation of the adaptive dual-weighted feature network in this paper. Simultaneously, we create an insulator dataset to substantiate the effectiveness of enhanced model in detecting small insulators. Firstly, the integration of context fusion network is employed to capture comprehensive contextual features for each effective feature map. In addition, a cross-scale residual perception network is incorporated into the neck prior to three concatenation modules, facilitating the collection of diverse information across levels. Finally, a Dual-Weighted Feature Fusion module is designed to replace the conventional concatenation pattern within the neck, thus achieving a more precise representation of object features. Experiments are conducted on the insulator dataset, the RSOD dataset and the NWPU VHR-10 dataset to evaluate the designed model, resulting in mAP values that were 3.92%, 1.55% and 2.39% higher than the YOLOv7, respectively.

Keywords Contextual features · Cross-scale residual perception network · Small objects · Insulator detection · ADFNet

1 Introduction

Insulators are particularly essential components in electric power transmission and distribution system, designed to prevent the flow of electrical current from the conductors to the supporting structures. The integrity and stability of the electric power network are maintained by them through providing electrical isolation and mechanical support. The reliable operation of transmission lines and the stability of electric power systems can be seriously compromised by a

✉ Yinhua Li
2019028@zzuli.edu.cn

Jie Zhang
2018007@zzuli.edu.cn

Xiabing Wang
332201060104@email.zzuli.edu.cn

Dailin Li
332201060067@email.zzuli.edu.cn

Fengxian Wang
2019031@zzuli.edu.cn

[1] College of Electric and Information Engineering, Zhengzhou University of Light Industry, No.5 Dongfeng Road, Jinshui District, Zhengzhou 450002, Henan Province, People's Republic of China

[2] Control and Simulation Center, Harbin Institute of Technology, No.2 yikuang street, Nangang District, Harbin 150008, Heilongjiang Province, People's Republic of China

defective insulator [1]. Thus, as a preliminary task to the power system fault detection, insulator detection is a pretty crucial assignment.

This paper mainly studies how to design a high-performance object detection model. The methods of object detection in recent years are summarized as follows:

(1) Traditional methods: Insulator detection with the conventional approaches in aerial images heavily depends on local and shape-based characteristics. The traditional insulator detection methods [2–4] employ a model constructed by traditional characterization techniques to capture and detect insulator characteristics. They propose a multi-feature and multi-scale descriptor, Otsu algorithm and adaptive morphology, respectively, to improve the accuracy of insulator detection. Zhai and his colleagues explored the potential of morphology by integrating the algorithm into the drone detection system [5]. These algorithms are capable of automatic detection, real-time monitoring and adaptive detection, but lack the ability to effectively learn defect features, which is challenging for the accurate identification of insulator defect areas. It can be seen that the performance of the model constructed using the traditional feature technology is not satisfactory in the complex background.

(2) Convolutional Neural Network methods: DCNNs [6], YOLOv5-s with DenseNet201 network [7] and SSD with two-stage fine-tuning [8] are the models for object detection through cascading mode. They enhance the accuracy and robustness significantly by fine-tune and elevating the expression ability of blocked features, respectively. Siamese ID-YOLO model [9] is a one-stage model for insulator detection, which utilizes the canny-based edge detection operator to enhance insulator edges and capture more semantic object features, thus improving detection accuracy and detection speed.

(3) Transformer [10] methods: Carion introduced DETR [11], a novel object detection model based on Transformer. DETR serves as an end-to-end detector, simplifying the training process. In these papers [12–14], the DETR was improved by adding different forms of attention, making the model more focused on object features rather than background information and ultimately improving the detection accuracy and convergence speed. Through the comparison of self-attention methods, it can be seen that these methods have excellent performance in detecting objects with obvious mutations in the image, but the overall detection accuracy is still not satisfactory, especially for objects with few pixels.

Through the above analysis, it can be seen that it is still a challenging problem to extend the existing small object detection and insulator detection to large-scale satellite images. Achieving accurate and robust object detection in satellite images is an extremely challenging task due to the following reasons:

(1) Lack of high-quality and well-annotated public datasets [15] and comprehensive benchmarks. For multi-object detecting in satellite images, it is meaningful to evaluate the performance of different algorithms fairly and comprehensively.
(2) The background of satellite images is more complex, including all elements in urban areas such as rivers, buildings, vegetation and houses.
(3) The proposed model is not versatile enough, and its ability to extract comprehensive and complete features [16] is weak. Therefore, these problems pose significant challenges to existing general object detection methods.

To address these problems, we propose an Adaptive Dual-Weighted Feature Network for Insulator Detection in Transmission Lines. In addition, we build an insulator dataset to prove the excellent performance of this model on insulator detection.

In summary, the contributions made by this article are as follows:

(1) The Context Fusion Network (CFN) is proposed and applied to all valid feature maps obtained from the backbone network of the model. Semantic weights are calculated through the deeper-level feature map, and then, the acquired weights are used to guide the features.
(2) The Cross-Scale Residual Perception (CSRP) network was combined with the valid feature layer to collect more information from different levels, which not only helped to increase the convergence speed, but also addressed the problem of accuracy degradation due to gradient drop as the number of network layers increased.
(3) A Dual-Weighted Feature Fusion (DWFF) network is designed to replace the common concatenation pattern in the neck to fully represent the characteristics of the object. The accuracy and performance of the model can be improved by effective feature extraction and feature weighting.
(4) An insulator dataset is produced to demonstrate that the excellent performance of the improved model in small objects detection.

2 Related work

2.1 YOLOv7 network structure

YOLOv7 [17] is a lately published deep learning model that follows on from its predecessor, YOLOv6 [18]. YOLOv7 provides significantly improved and more accurate object detection performance without increasing computational complexity and cost. This object detector surpasses other well-known detectors [19] by reducing approximately 40% of the parameters and 50% of the computation required for state-of-the-art real-time object detection. This enables it to perform inferences more quickly with higher detection accuracy.

The network architecture of YOLOv7 includes two primary components: the backbone and the predict head. Unlike YOLOv5, it refers to the neck and head together as the predict head. Extended-ELAN and MPConv [20] structures have been added to the backbone network, which is applied to extract the feature information from the input image and output the effective features. The neck aims to combine low-level spatial information with high-level semantic features by adopting the multi-scale feature fusion mode to preserve more details for feature enhancement, so that the module improves the detection accuracy of small objects [21]. Three detection heads each predicts the features of the three sizes after feature concatenation. YOLOv7 has the ability to compute the optimal anchor by using K-means [22] on different training datasets to obtain frame values. YOLOv7 tries a variety of activation function compositions, for example, leakyReLU and SiLU. There are also six derived models for YOLOv7, including YOLOv7-tiny, YOLOv7-x, YOLOv7-D6, YOLOv7-E6, YOLOv7-E6E and YOLOv7-W6. It utilizes a cascade-based model scaling methodology that can produce a model of the appropriate scale for the actual task to meet the detection requirements.

The YOLOv7 model has a faster and more robust network architecture that employs a more efficient feature integration approach, more precise object recognition performance, a more stable loss function and an optimized assignment of labels [23] and model training efficiency. Accordingly, compared to other deep learning models, YOLOv7 uses far less expensive computational hardware and can be trained much faster on small datasets without the use of pre-trained weights. The authors of YOLOv7 introduced several architectural changes, including compound scaling, the extended efficient layer aggregation network (E-ELAN), a bag of freebies with planned and reparameterized convolution, coarseness for auxiliary loss and fineness for lead loss.

2.2 Improved network models based on YOLO

The following are some papers based on the first-stage YOLO series to improve object detection papers, they, respectively, put forward novel ideas and modules, which can obtain better detection accuracy and other excellent performance. Kumar et al. [24] proposed this study is to demonstrate the applicability of the latest object identification algorithms (YOLOv5 and YOLOv7) in real-world scenarios for object recognition. The study focuses on utilizing surveillance footage obtained from UAVs/drones to identify targets in a real-world environment. Wang et al. [25] introduced YOLO-G2S, a military object detection algorithm that achieves a smaller model size by sacrificing a small amount of accuracy. In YOLOv5, they replaced the three modules with C3G2 modules and modify the RELU activation function, resulting in reduced model size and improved detection speed in military object detection. Yang et al. [26] introduced an innovative deep convolutional network architecture called TS-YOLO for multi-scale object detection, which incorporates three spatial pyramid pooling (SPP) modules into YOLOv4. This integration enables the extraction of enhanced semantic information in intricate environments. In 2021, Zhu [27] introduced TPH-YOLOv5, which enhances object detection in aerial images by incorporating a prediction head capable of detecting objects at various scales and replacing the original prediction head with a Transformer Prediction Head (TPH) combined with a Transformer to explore the potential of self-attention. While this method effectively improves performance, it has limited theoretical innovation and is constrained by object occlusion environments. In 2022, Li [28] introduced Acam-YOLO, an object detection algorithm for UAV aerial imagery. This method incorporates the Adaptive Co-Attention Module (ACAM) into the backbone and feature enhancement networks, enabling adaptive cooperative attention. By extracting spatial and channel attention features from sliced input features and weighting them synergistically, the algorithm enhances detection accuracy. However, these approaches face challenges in effectively capturing the features of tiny targets against complicated backdrops and achieving high detection accuracy.

The latest advancements in object detection models have been developed into the DEtection TRansformer (DETR) model, which has gained popularity due to its innovative characteristics. However, it falls short in terms of accuracy. Consequently, in the pursuit of a model with excellent detection performance, opting for the YOLOv7 to modify remains the preferable choice. Through a comprehensive series of empirical tests, we have substantiated that

our enhanced model possesses outstanding detection efficiency.

2.3 The attention mechanism

The human perceptual process relies heavily on attention, which plays a crucial role in how individuals selectively process and allocate their cognitive resources to different stimuli. The concept of the attention mechanism stems from research on how human attention selectively processes image information. It enables neural networks to possess an adaptive perceptual capability for computer vision tasks, in particular by focusing the model's attention on the crucial parts of the input and extracting essential features. The attention mechanism has been implemented in various ways in practical applications. The Recurrent Attention Model (RAM) [29] pioneered the combination of the attention mechanism with deep neural networks. Recurrent Neural Networks (RNN) were fundamental tools for the attention mechanism in its early stages. To incorporate spatial attention into CNN, the Spatial Transformer Network (STN) [30] was proposed, which automatically selects the features of the region of interest and performs spatial data transformations with different deformations. In contrast with spatial attention, the Squeeze-and-Excitation Network (SENet) [31] demonstrates a unique channel attention network that adaptively predicts potentially critical features.

Through the above summary, we get inspiration from the Channel Attention of [31] and put forward the Context Fusion Network module. Different from the above attention, this attention can focus on channel feature more effectively and has higher detection accuracy when combined with CFE module. Dual-Weighted Feature Fusion is proposed, which combines the channel and spatial attention in [32]. By selecting the more important feature details, the DWFF network can more precisely calculate the features of the entire image, thus weighting the important information, making it also provide directional information for shallow and deep features in a novel way. It is subsequently proved that these two attention-related modules do promote the model to locate and identify objects of interest more accurately, especially small objects, thus achieving high performance.

3 Theoretical model

3.1 Adaptive dual-weighted feature network

The architecture of our proposed ADFNet for transmission line insulator detection is depicted in Fig. 1. To address the existing problems, we have designed three modules in the ADFNet paper: the Context Fusion Network, the Cross-Scale Residual Perception (CSRP) network and the Dual-Weighted Feature Fusion (DWFF) module.

The model comprises three components, namely, backbone layers, neck and detection layers, as depicted in Fig. 1. The backbone layers consist of CSPDarknet53 [33], which gradually extracts deep-level features from input insulator images through a series of convolutional layers, SPPCSPC module, extended-ELAN module, MPConv module and residual blocks. It serves as a feature extractor in insulator detection tasks, crucial for extracting high-level semantic features at different levels from input insulator images. The ELAN module introduces a novel attention mechanism, which enables the network to focus more on crucial features, thereby enhancing the detection performance of the model. The extended-ELAN module, as an enhancement to the original ELAN, preserves the transition dynamics of the original ELAN while modifying the computational block's layer structure and utilizing expand, shuffle and merge operations to augment cardinality. This approach aims to bolster network learning capabilities without compromising the integrity of the original gradient pathways. This is particularly beneficial for addressing challenges in detecting small or occluded targets where performance may be suboptimal. Within the Neck module, ADFNet integrates not just the conventional PAFPN structure, but also incorporates designed Context Fusion Network, the Cross-Scale Residual Perception (CSRP) network and the Dual-Weighted Feature Fusion (DWFF) module to optimize the process of feature extraction. This holistic strategy is employed to achieve a more comprehensive extraction of insulator features. FPN enhances feature extraction networks, fusing three effective feature layers obtained in the backbone section to combine insulator features from different scales. In the ADFNet, Panet's structure is utilized, performing both upsampling and downsampling to achieve feature fusion. The detection layers employ detection heads representing large, medium and small object sizes, with the RepConv module exhibiting structural differences during training and inference. As the classifier and regressor of YOLOv7, the Yolo Head treats feature maps as collections of individual feature points, each featuring three anchor boxes with several channel features. The Yolo Head's role is to assess these feature points, determining if there are objects corresponding to the anchor boxes. The employed decoupled head simultaneously handles classification and regression through a 1x1 convolution.

The loss function used by the ADFNet model is as shown in Equation (1):

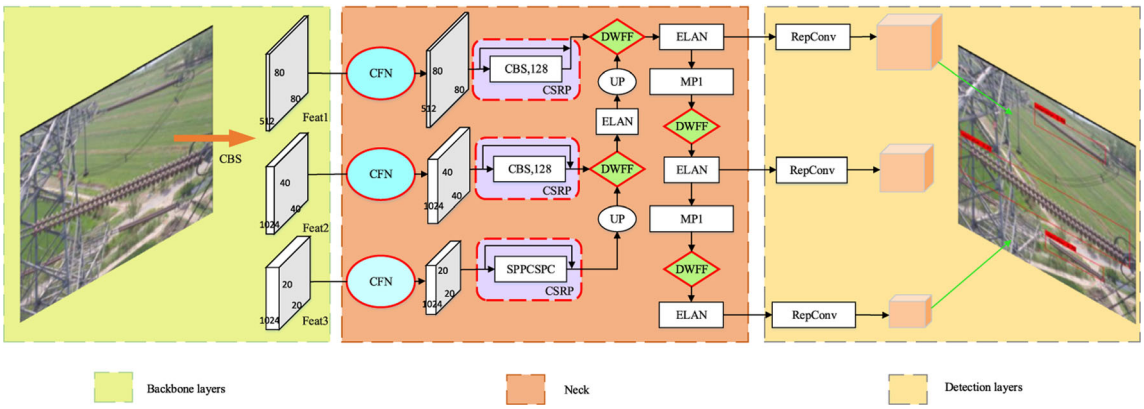


Fig. 1 The architecture of ADFNet. The boxes with red border represent the designed Context Fusion Network (CFN), Cross-Scale Residual Perception (CSRP) network, and Dual-Weighted Feature Fusion (DWFF) layers compared to the origin model, respectively. We also add three Context Fusion Networks, three CSRP networks and four DWFF modules to improve performance

$$L_{\text{object, Loss}} = L_{\text{class, Loss}} + L_{\text{loc, Loss}} + L_{\text{conf, Loss}} \tag{1}$$

where L_{class} and L_{loc} and L_{conf} denote classification loss, localization loss and object confidence loss, respectively. *BCE* cross-entropy loss is used for both classification and confidence loss, and $S(x_i)$ represents the Sigmoid function.

$$S(x_i) = \frac{1}{1 + e^{-x}} \tag{2}$$

The calculation of the Binary Cross-Entropy (BCE) loss is as follows: w_i is used to average the results, and y_i represents the true sample label:

$$L_i = -\mathrm{w}_i[\mathrm{y}_i \cdot \log \mathrm{S}(\mathrm{x}_i) + (1 - y_i) \cdot \log(1 - \mathrm{S}(x_i))] \tag{3}$$

The localization loss function utilizes the *CIoU* loss, which incorporates various factors such as aspect ratio, overlap area and center distance. This results in improved detection accuracy, especially when dealing with non-overlapping detection boxes. The formula for calculating the *CIoU* loss is as follows:

$$\mathcal{L}_{\text{CIoU}} = 1 - I_{\text{IoU}} + \frac{\rho^2(b, b_{\text{gt}})}{c^2} + \alpha v \tag{4}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{\text{gt}}}{h_{\text{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{5}$$

$$\alpha = \frac{v}{(1 - I_{\text{IoU}}) + v} \tag{6}$$

where *IoU* represents the intersection region between the predicted and true boxes; b, b_{gt}, v and α represent the predicted box, the true box, a measure of the consistency of the aspect ratios and a balancing parameter to give the overlap area factor a higher regression priority, respectively. Figure 2 exhibits the loss diagram for the bounding box regression. The distance between the centers of the two bounding boxes is represented by $d = \rho^2(b, b_{gt})$, while c signifies the diagonal length of the minimal enclosing area which can contain both the predicted and true boxes.

3.2 Context fusion network

We propose a Context Fusion Network dubbed CFN, whose structure is shown in Fig. 3. It is significant for object detection to pay attention to the visual context surrounding subjects. Current CNN networks [34] learn object characteristics by overlaying multiple convolutional layers and pooling layers. However, some objects vary greatly in size, shape and location, and previous approaches often directly use bottom-up convolution and pooling layers, which may not be able to efficiently process such complex variations. Inspired by CFE (Context Feature Extraction), we have attempted to devise a fancy model, CFN, to extract size [35], shape [36] and position invariant features. The *CFN* is applied to every valid feature map, as displayed in Fig. 1. Firstly, the *CFN* is performed on the input feature map E to generate information of contextual feature [37]. Additionally, the feature is forwarded to the average pooling layer and the channel attention module to acquire a selected feature map. *CA* [31] and Avgpool denote the

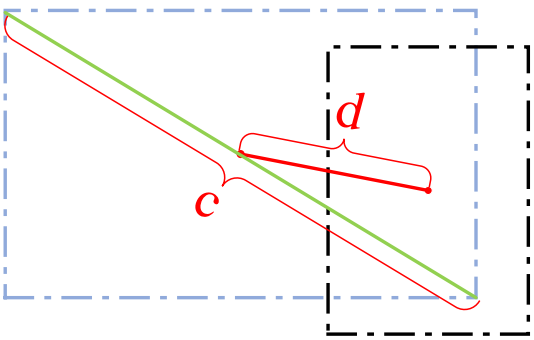


Fig. 2 Illustration of CIoU loss formula

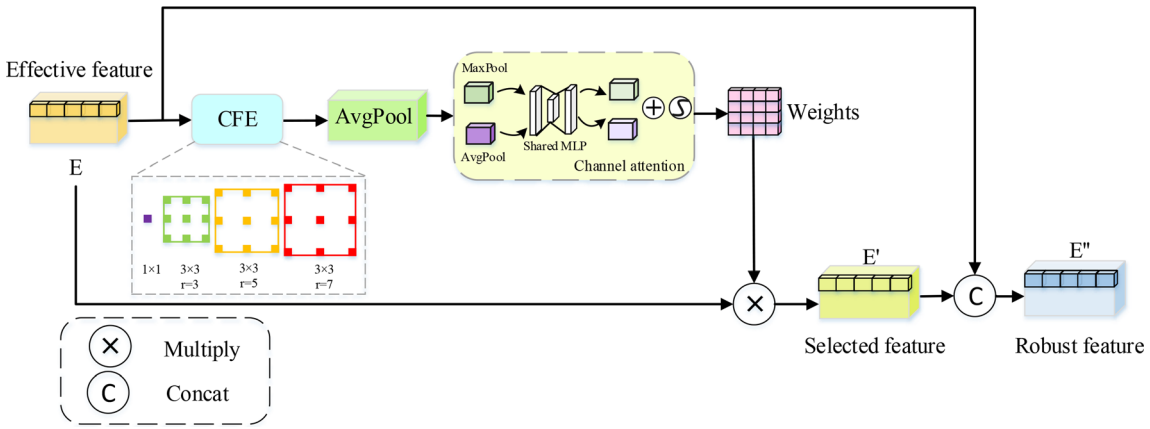


Fig. 3 The Context Fusion Network. The CFN is composed of three branches: The first branch applies residual learning to fuse the selected features, the second branch extracts attention from contextual features through average pooling and channel attention to obtain weights and the third branch multiplies the effective feature layer with the weights output by branch 2 to obtain the selected features. A context-aware feature extraction (CFE) module utilizes a feature from a side output of the network as input. It consists of three 3x3 convolutional layers with varying dilation rates and a 1x1 convolutional layer

channel attention and average pooling, respectively. The semantic weight matrix $M(E)$ is developed by average pooling and channel attention, which is computed as follows:

$$M(E) = CA(Avgpool(CFE(E))) \tag{7}$$

The weight matrix $M(E)$ generated by the above operations including CFE, average pooling and channel attention is multiplied with the input feature map to obtain the selected feature map E'. Then, it is joined with the input valid feature map to acquire the robust feature map E''. In brief, E' and E'' can be summarized as follows:

$$E' = M(E) \times E, \tag{8}$$

$$E'' = E' \otimes E \tag{9}$$

To significantly enhance the feature extraction [38] capability of the detection module, the Context Fusion Network guides key features using the semantic matrix generated by the second branch to enhance the saliency of the object. Additionally, fusing shallow features with deep feature information helps enhance feature representation and improve the model's detection accuracy.

The context fusion network can extract spatial relationship features between insulators and their surrounding environments. Insulators may be challenging to distinguish from complex backgrounds, and the context fusion network aids in extracting features that suppress or ignore irrelevant background information, focusing on the unique characteristics of the insulator itself. This contributes to an improved signal-to-noise ratio in feature extraction. By considering the context fusion network, it can enhance the contrast between the insulator and its surrounding environment. This involves extracting features that emphasize differences in appearance, texture or color, making the object insulator more prominent within the overall scene. This is instrumental in elevating the signal-to-noise ratio in feature extraction, and thus, the accuracy of insulator detection can be improved.

3.3 Cross-scale residual perception network

YOLOv7 adopts FPN [39] with PAN [40] to extract and fuse features at the neck. Despite its ability to tackle the problem of different scale objects in diverse scene images, it is not the top-ranking method for feature integration.

To collect more information from different levels, we propose Cross-Scale Residual Perception (CSRP) network. This not only helps to boost the convergence speed, but also addresses the problem of accuracy degradation due to gradient drop when the number of network layers increases. It is a network composed of two modules in parallel, the convolutional module and the cross-scale connection. Specifically, a cross-scale residual module is added at both ends of the convolutional module at different levels in the network neck. This configuration achieves the Cross-Scale Residual Perception (CSRP) network. The objective is to facilitate the transfer of features and the flow of information across different scales, aiming to obtain a more comprehensive and enriched feature representation. The incorporation of cross-scale residual modules enables the fusion of features at non-adjacent layers, addressing the characteristics of insulator images with sparse pixels and limited features. This approach proves advantageous in enriching feature information.

The thought of trans-scale connectivity is integrated into YOLOv7, which links input and output nodes of the same level across layers. Therefore, the distance of the path from

shallow-level to deep-level information is reduced, and the deep-level rich semantic features are integrated with the shallow-level positional features. To boost the precision of prediction, we adopt contiguous layers using a series rather than simple addition. The location of the Cross-Scale Residual Perception (CSRP) network is the connection from the effective feature layer to the back of the feature fusion module. The CSRPNet is designed to integrate information across multiple scales. It can extract features from different levels of abstraction, allowing it to capture fine-grained details of the object insulator. This multi-scale feature fusion ensures a comprehensive representation of insulator characteristics. By incorporating cross-scale residual connections, the network can effectively propagate information across different scales. This consideration extends beyond the immediate surroundings of the insulator to encompass a broader context, aiding in the extraction of features that emphasize the unique and discri minative aspects of the insulator. This results in a more robust and discriminative feature set, crucial for accurate feature extraction, enhanced discriminability, adaptability to different sizes, handling scale variations and effectively utilizing hierarchical information for feature extraction, especially when dealing with small targets in complex environments.

3.4 Dual-weighted feature fusion

Inspired by DANet [32], we propose DWFF, which stands for dual attention-weighted feature fusion. The attention-weighted feature fusion network enables more precise computation of channel and spatial attention by selecting more significant feature details, but it also offers direction information for the shallow-level and deep-level features in a novel way.

Features extracted from deep layers usually have superior semantic information, such as the shape, texture and parts of the object, while shallow features often refer to low-level features such as color, texture and edges, which are more intuitive and can be used for the initial localization and detection of the object area. Therefore, the extraction of both low- and high-level features has its unique characteristics and roles. In object detection, fusing deep and shallow features can better combine their advantages and improve the performance and robustness of the model.

As illustrated in Fig. 4, DWFF module includes two branches: a shallower branch and a deeper branch. Firstly, the shallow features obtain spatial weight by using spatial attention (SA) in [32] and then multiply this weight with the low-level features to obtain a weighted feature map. Additionally, the deep features acquire a channel weight by using channel attention and then multiply this weight with the high-level features to obtain a weighted feature map. Finally, the weighted feature maps obtained by these two branches are fused to obtain the final attention-weighted feature fusion map.

DWFF employs dual weights to selectively integrate characteristics from various hierarchical sources. This allows the network to prioritize relevant information for insulator detection while minimizing the impact of irrelevant or noisy features. Selective integration contributes to more effective and distinctive feature extraction. The adaptive feature importance enhances the network's ability to capture salient features of insulators. This ensures that the network assigns more weight to features contextually important for different environments and scenarios, improving feature extraction. DWFF's ability to selectively fuse features based on importance enhances the network's robustness to such variations. This is particularly beneficial for feature extraction in dynamic or changing environments. By weighting features, DWFF contributes to an improved signal-to-noise ratio in the feature extraction process. It emphasizes information features relevant to insulator characteristics while attenuating irrelevant information, reducing the impact of background noise on detection accuracy. In enhancing insulator detection accuracy, DWFF makes a significant contribution. The position of DWFF is to replace the feature fusion module of the neck in the overall structure, which leads to better performance results than before.

4 Experiment

To demonstrate the validity of our designed model, we conducted experiments not only on the insulator dataset, but also on two widespread satellite image datasets, the RSOD dataset and the NWPU VHR-10 dataset. We selected commonly accepted evaluation metrics, including mAP, Recall, P and AP. Experimental results prove that the model outperformed other existing models.

4.1 Establishment of insulator dataset

To realize the potential of neural networks, high-quality and well-annotated datasets are required. An insulator dataset was created using images captured by a drone to validate the effectiveness of ADFNet for insulator detection tasks. This dataset incorporates several data augmentation techniques, including random flipping, image cropping, random adjustments of brightness, contrast, saturation and the addition of Gaussian noise. As a result, our dataset consists of 4000 images with a total of 7500 insulator targets. The number distribution of the test set, training set, validation set as well as training and validation

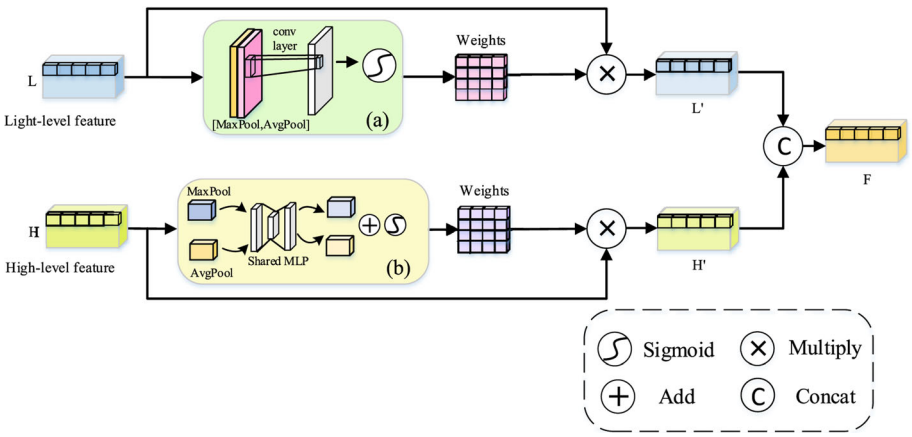


Fig. 4 Overview of Dual-Weighted Feature Fusion. DWFF module includes two branches: a shallower branch and a deeper branch. The two branches use spatial and channel attention to obtain spatial and channel weight for shallow and deep features, respectively. This weight is then multiplied with the underlying features to obtain the weighted feature map. Finally, the weighted feature maps from both branches are fused to obtain the final weighted feature fusion map. The structural block **a** represents spatial attention, and the structural block **b** represents channel attention

set is shown in Table 1. The purpose of dataset enhancement is to reduce the occurrence of network overfitting, so that the trained model can achieve more excellent generalization ability and detection performance. The distribution of insulators in these images includes uniform distribution, interweaving of bounding boxes and nesting of bounding boxes. Our dataset includes a variety of scenes, from simple to complicated backgrounds, and from single to multiple objects of varying sizes, which ensures the generalizability of the model.

This insulator dataset was derived from aerial images taken by transmission line drones in Henan Province to conduct experiments with insulator samples in different scenarios. The insulator images contain various challenging insulator targets, such as complex context, occlusion and small targets, to better evaluate the applicability and robustness of different approaches.

4.2 Evaluation metrics

We selected the mean average recall (R), precision (P), precision (mAP) and average precision (AP) as evaluation metrics. R and P are calculated using False Positive (FP), True Positive (TP), False Negative (FN) and True Negative (TN) as follows:

$$P = \frac{TP}{(TP + FP)'} \tag{10}$$

$$R = \frac{TP}{(TP + FN)}. \tag{11}$$

F_1 is calculated using precision and recall. With forming a precision–recall curve (P-R curve) from precision and recall, the AP can be computed using the area enclosed by the P-R curve.

$$F_1 = \frac{2Precision * Recall}{Precision + Recall}, \tag{12}$$

$$AP = \int_0^1 P(R)dR. \tag{13}$$

And mAP is the mean of all APs in N classes.

$$mAP = \frac{1}{N}\sum AP. \tag{14}$$

Based on the pixel size occupied by the targets in the images, the triple levels are as shown in Table 2:

In Sect. 4 of this chapter, visualized images of the dataset have been clearly presented, demonstrating that the method effectively enhances the performance of object detection, particularly for small objects.

Table 1 Distribution of the number of images in the insulator dataset

Image number	4000
Number of images in the test set	400
Number of images in the training set	3240
Number of images in the validation set	360
Number of images in the training and validation set	3600

4.3 Implementation details

We implement ADFNet in PyTorch 1.2 deep learning framework with CUDA 10.0. All improved models were trained and tested on an Intel NVIDIA TITAN RTX with 256 GB of memory and 8163 processor. In the training

Table 2 Definition of the object hierarchy

Object Level	Small Object	Medium Object	Large Object
Size Range (pixel)	0×0 ∼ 32×32	32×32 ∼ 96×96	≥96×96

stage, we froze the YOLOv7 backbone for the first 50 epochs of training, after which we skipped the freezing epochs to commence the real training phase of 150 epochs, which saved considerable training time. We use the Adam optimizer to train these models with an initial learning rate of 0.001 and a batch size of 8. The used optimizer is the Adam optimizer, whose internal momentum parameter is 0.937. Limited by the mAP calculation principle, the network needs to obtain almost all prediction boxes when calculating mAP. Therefore, the value of the confidence should be set small, and we will set the confidence level to 0.001 to get all possible prediction boxes. We set the value of confidence to 0.001. The three datasets used in the experiments were trained for 150 epochs under the same conditions. It is important to note that all images were automatically scaled to 640 × 640 by ADFNet to facilitate the detection of small objects.

4.4 Experimental results

4.4.1 Experiment on insulator dataset

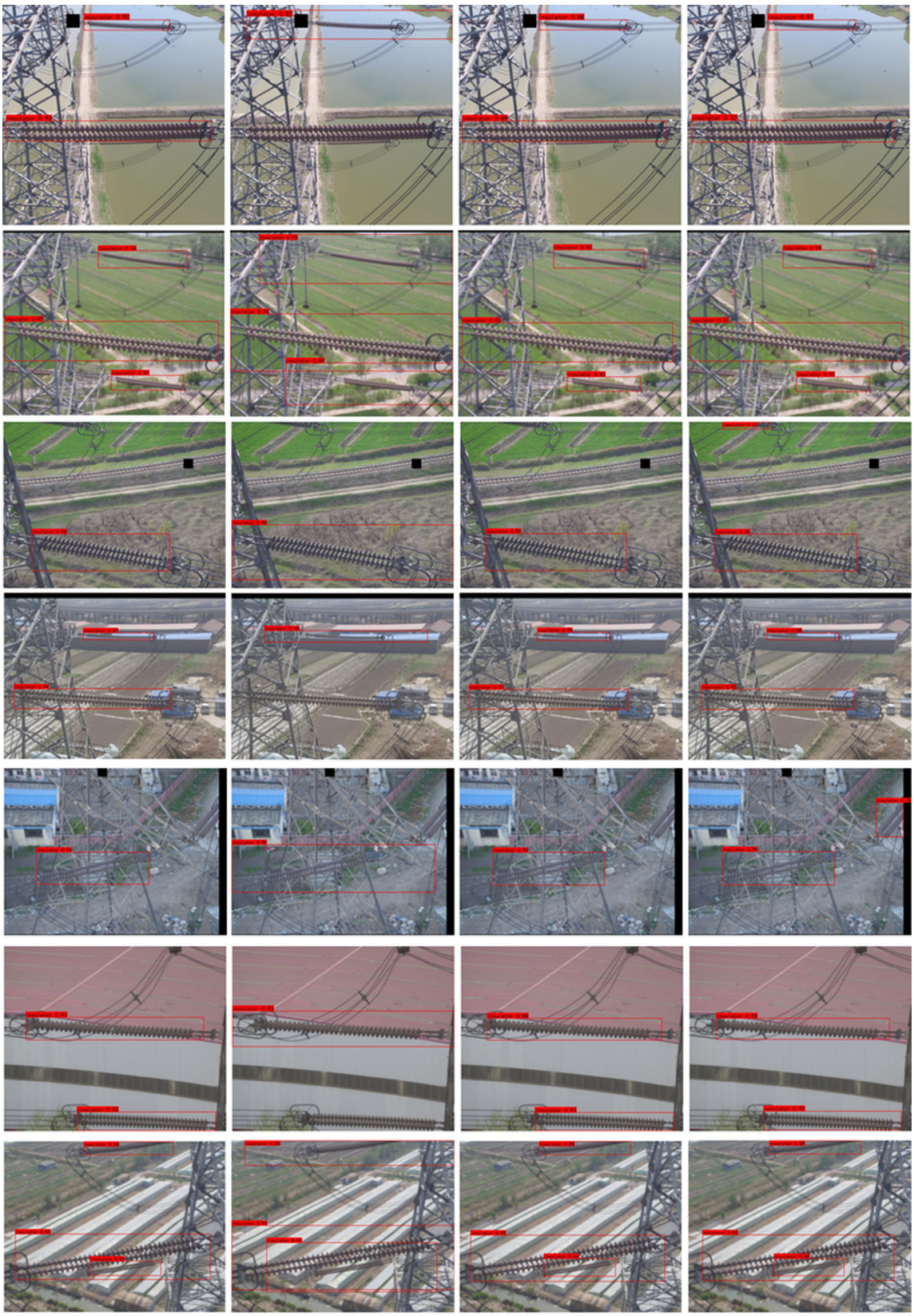
To validate the effectiveness of our approach for insulator detection on transmission lines, we first selected the images of insulators on transmission lines. The results of precision, recall, F1 and mAP achieved by distinct detection models are demonstrated in Table 3. The average accuracy of the YOLOv7 detection is 94.15%, while the mAP of the ADFNet is 98.19%. It is noteworthy that ADFNet outperforms other models, such as DETR (91.03%, 7.16% higher mAP), CenterNet (95.07%, 3.12% higher mAP), GhostNet-YOLOv4 (79.30%, 18.89% higher mAP), YOLOv7-l (94.15%, 4.04% higher mAP), YOLOv7-x (96.24%, 1.95% higher mAP), SSD (94.20%, 3.99% higher mAP), YOLOX (93.75%, 4.44% higher mAP), YOLOv8-s (97.72%, 0.47% higher mAP), YOLOv8-x (96.36%, 1.83% higher mAP) and YOLOv8-n (95.84%, 2.35% higher mAP). The significant improvements demonstrate the superiority of ADFNet over insulator detection.

Meanwhile, recall, F1 and mAP all have higher values than other algorithms on the insulator dataset except precision. Although ADFNet's ratio of true positives to false positives with accurate predictions, at 94.84%, is not as good as YOLOv8-s' value of 96.20%, it achieves an extremely high mAP. In addition, ADFNet has the best R, which is 0.92%, 3.07% and 4.16% higher than YOLOv8-s, YOLOv8-x and YOLOv8-n, respectively. In our model, the value of P is 2.06% lower than that of the highest model DETR and 12.35% higher than that of the lowest model DETR. The value of R is 54.71% higher than that of GhostNet-YOLOv4, the lowest model. The value of F1 is 0.37 higher than that of GhostNet-YOLOv4, the lowest model. It can be seen that in general, our model has a good performance in object detection, especially in the recall, F1 and mAP.

These detection results of distinct models including YOLOX [35], YOLOv7 [17], YOLOv8, ours and other models on insulator dataset are shown in Fig. 5. This shows that our model has excellent detection performance on insulator dataset. In the first row, it is noticeable that the original YOLOv7 code fails to detect the large insulator at the bottom. However, after our improvements, the insulator is successfully detected, and the confidence of the medium-sized objects above increases from 0.87 to 0.91, surpassing the other two models. In the second row, it is evident that all three insulators have exceptionally high accuracy, with confidences of 0.94, 0.97 and 0.93, respectively. In the third and fifth rows, despite the addition of Gaussian noise and only a small portion of the insulator visible in the top-left corner, our model is still able to detect it effectively, achieving a mAP of 0.51. While the other three models fail to recognize this positive sample, the performance of our model not only improves confidence, but also increases recall rate, indirectly improving the F1 score of the model.

Table 3 Precision, recall, F1 and mAP on insulator dataset for distinct detection models

Detector	P(%)	R(%)	F1	mAP(%)
DETR [11]	82.49	92.07	0.87	91.03
CenterNet [41]	96.67	90.23	0.93	95.07
GhostNet-YOLOv4 [42]	95.09	42.30	0.59	79.30
YOLOv7-l [17]	94.25	90.46	0.92	94.15
YOLOv7-x	94.41	93.53	0.94	96.24
SSD [43]	96.90	79.08	0.87	94.20
YOLOX-s [35]	92.97	92.75	0.93	93.75
YOLOX-x	93.28	92.36	0.92	94.84
YOLOv8-s [44]	96.20	96.09	0.96	97.72
YOLOv8-x	95.50	93.94	0.95	96.36
YOLOv8-n	94.06	92.85	0.94	95.84
ADFNet (Ours)	94.84	97.01	0.96	98.19


YOLOX
YOLOv7
YOLOv8
ADF-YOLOv7

◀**Fig. 5** Comparisons of the detection results on the insulator dataset. We observe that the results of the ADFNet are more accurate than other methods. Zoom in to see more details

Comparing the fourth and sixth rows, it is evident that our proposed model demonstrates superior accuracy in detecting the insulators at the top, with confidence of 0.88 and 0.95, respectively. In the seventh row, where nested bounding boxes are present, our model is still able to detect them with high confidence. Additionally, the insulator above, which is partially obscured, is detected by our designed model with the highest confidence of 0.89. The innovation we introduced in our model, referred to as the attention-weighted feature fusion network, grants these advantages to the model. Our proposed approach effectively leverages these advantages. This module allows more accurate computation of channel and spatial attention by selecting more significant feature details, but it also offers guiding information for the shallow- and deep-level features in a novel way. As a result, the targets that are partially obscured or only partially visible, such as the insulators, can be detected with high confidence despite the interference or occlusion.

To demonstrate the excellent performance of our model on small insulators, we present a comparison of images from different models for small insulator detection. As shown in the visualization results in Fig. 6, ADFNet outperforms other models in the case of pixels less than 32 $\times$ 32, strong occlusion, analog interference and intersection with other electric power equipment. The first and second row visualizations demonstrate that the model can detect small insulators with high accuracy even under the influence of strong interference. The superiority is due to our contextual module, which is able to recognize and extract more detailed features of small targets based on their contextual information. In the third, fourth and fifth rows, the model accurately identifies very small and heavily occluded insulators, with confidence levels of 75% and 81%, respectively. In the sixth and seventh rows, the model demonstrates its ability to detect small insulators with confidence levels of 83% and 56% even in the presence of analog interference. Additionally, the majority of the detection mean Average Precision (mAP) values for small targets are higher in ADFNet compared to the other models. There are two reasons for the excellent improvement results: First, the fusion module DWFF with weighted attention is adopted (instead of the common and simple feature fusion) to increase the expressiveness of the network; second, the Cross-Scale Residual Perception (CSRP) network is used to combine the lower-level output feature map with the higher level to retain more shallow information. Hence, by employing these innovative techniques and successfully reducing background interference, our model demonstrates enhanced accuracy in predicting small insulators, even when they are obstructed or partially obstructed. These findings unequivocally highlight the exceptional detection performance of ADFNet when it comes to small insulator targets.

Our proposed ADFNet also boasts a relatively fast detection speed. As evidenced in Table 4, when detecting the same image, it achieves a detection time of 0.0614. While not the fastest, it still outperforms other algorithms, being 0.0161 s faster than YOLOX, 0.0212 s faster than DETR and 0.0025 s faster than GhostNet-YOLOv4. ADFNet achieves an FPS of 16.27, which is 3.03 higher than the lowest FPS achieved by DETR. Although it falls short of surpassing CenterNet and YOLOv7, our model excels in other metrics, making the pursuit of perfection an unattainable goal. YOLOv7's FPS is only 0.04 lower than YOLOv8's FPS. However, this is the only advantage of YOLOv8, which does not compensate for poor object detection accuracy, especially for small objects. ADFNet's slower detection speed compared to YOLOv7 can be attributed to the three innovative ideas: the Context Fusion Network, the Cross-Scale Residual Perception (CSRP) network and double-weighted feature fusion (DWFF) module. These additions result in a more complex model structure, increased parameter size and higher computational requirements, consequently resulting in a relatively slower detection speed. Despite this drawback, ADFNet still outperforms YOLOX, DETR and GhostNet-YOLOv4 in terms of detection speed.

To support our assessment, we evaluate the insulator model using the metrics of recall, F1 and precision, which vary with the Score Threshold, as well as the PR curves, as shown in Fig. 7. A large number of tests have shown that when the Score Threshold value is 0.5, the effect is the best, so we finally use 0.5 as the standard for the experiment. The values presented in Table 3 within this article were computed with Score Threshold at 0.5. Specifically, we show the experimental results, including recall, F1 and precision scores in (a), (b) and (c) of Fig. 7, by varying Score Threhold from 0 to 1.

These line charts are shown in Fig. 7. From Fig. 7a, it can be seen that, except for DETR, the line of our model is closer to the top right than any other, indicating that these values of our model are very high. The recall rate of SSD is the worst. When Score Threshold is set to 0.6, our designed model achieves the highest recall rate, surpassing SSD by 0.16. Throughout the entire process, the largest difference between our proposed model and the lowest one is 0.29, which occurs when Score Threshold is equal to 0.8. As for DETR, it indeed has a good recall rate when Score Threshold is higher than 0.8, but when Score Threshold is less than 0.8, its recall rate is lower than YOLOv8 and our

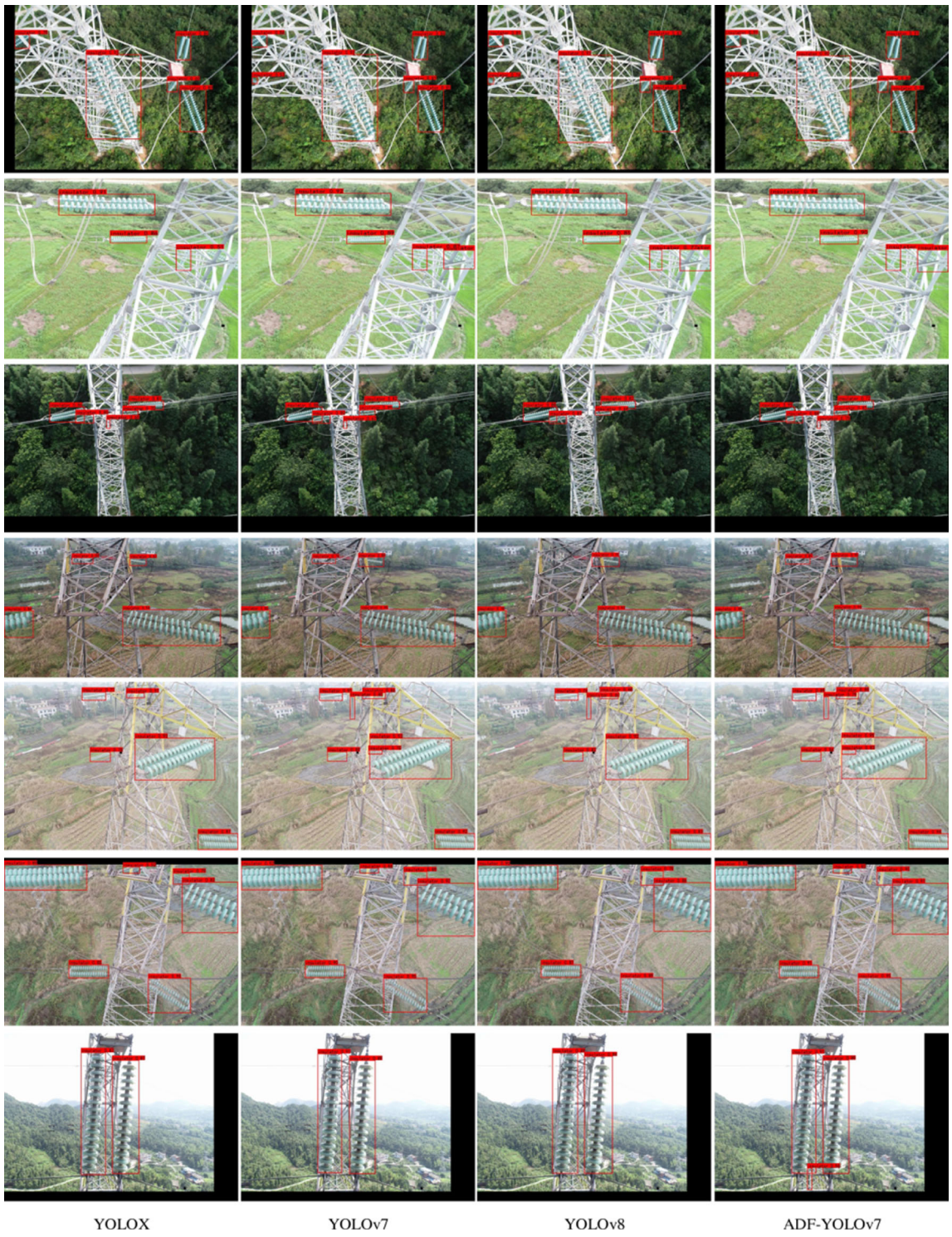


Fig. 6 Comparisons of the small insulator detection results on the insulator dataset. We observe that the results of the ADFNet are more accurate than other methods. Zoom in to see more details

proposed algorithm ADFNet. The high recall rate is due to the fact that we added the context module and adjusted the parameters of the model appropriately.

From Fig. 7b, it can be seen that, SSD and DETR both exhibit good F1 scores around 0.5. However, when Score Threshold is set to 0, the F1 score drops significantly, indicating the model's instability. However, ADFNet amazingly combines the strengths of both models while mitigating their shortcomings. This amalgamation presents a significant advantage of this model. The F1 score, being

Table 4 Inference time and speed on the insulator dataset for distinct detection models

Methods	Inference time per image (second)	Speed (FPS)
YOLOX [35]	0.0826	12.09
DETR [11]	0.0755	13.24
YOLOv8 [44]	0.0473	21.13
GhostNet-YOLOv4 [42]	0.0639	15.64
CenterNet [41]	0.0539	18.53
YOLOv7 [17]	0.0474	21.09
ADFNet (Ours)	0.0614	16.27

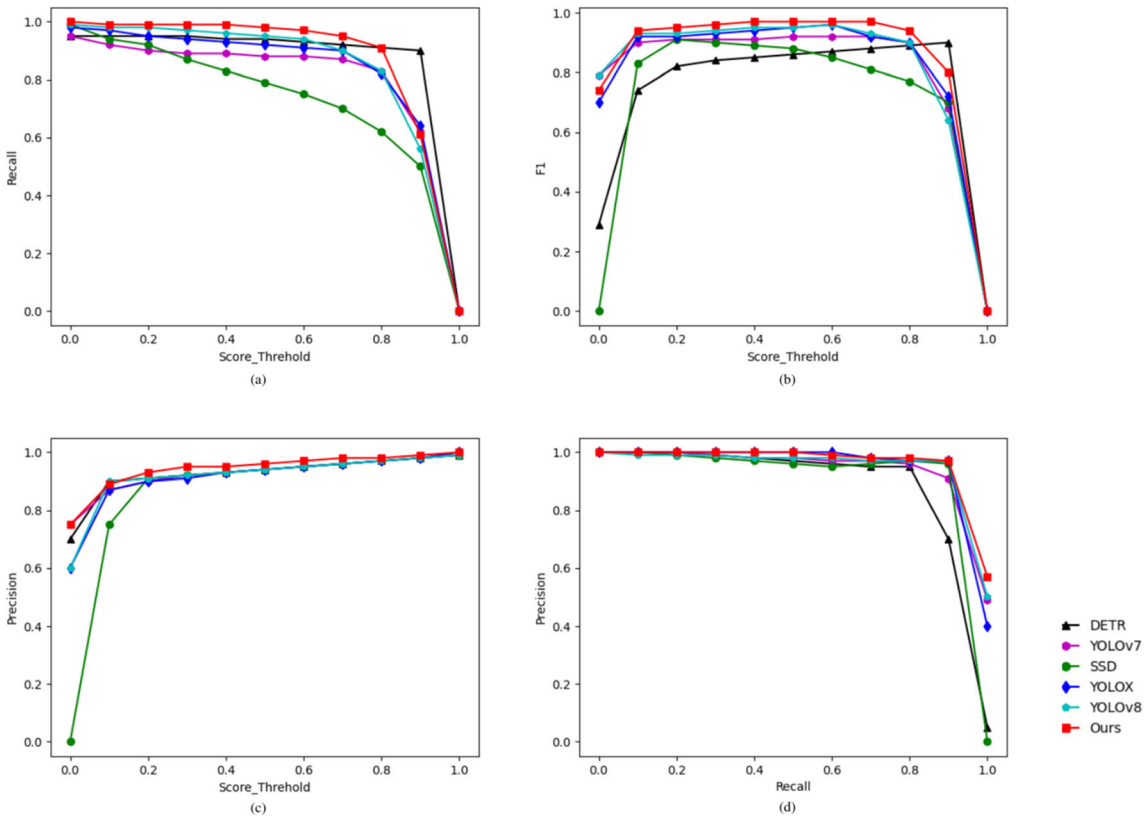


Fig. 7 Experimental results of F1, recall and precision, which vary with the Score Threshold, as well as the PR curves

the harmonic average of accuracy and recall rate, indirectly proves the model's good accuracy and recall rate. One key factor contributing to this success is the innovative Cross-Scale Residual Perception (CSRP) network proposed in our model. This mechanism gathers feature information from different levels and effectively fuses them, facilitating better object feature extraction and improved detection accuracy. Moreover, it addresses the problem of declining accuracy caused by gradient descent when the number of network layers increases. Ultimately, this approach enhances the likelihood of correctly predicting positive samples.

From Fig. 7c, it is evident that the trend of our method's line remains consistently at the top, highlighting its superior accuracy compared to other benchmark methods. Specifically, when Score Threshold is equal to 0, the minimum value of SSD is 0, while our ADFNet surpasses it by 0.75. Despite YOLOv8 being the latest YOLO algorithm introduced this year, our model outperforms it in all aspects, except for a minor 0.01 accuracy advantage of YOLOv8 at Score Threshold = 0.1. Despite DETR's recent popularity, it fails to demonstrate a significant advantage and appears unimpressive in this context.

The PR curve of our method and other baselines are presented in Fig. 7d. As can be seen, while their trends are remarkably similar, our method produces a PR curve on the insulator dataset that is positioned close to the upper right corner. This indicates that our method outperforms other benchmark methods in terms of both precision and recall.

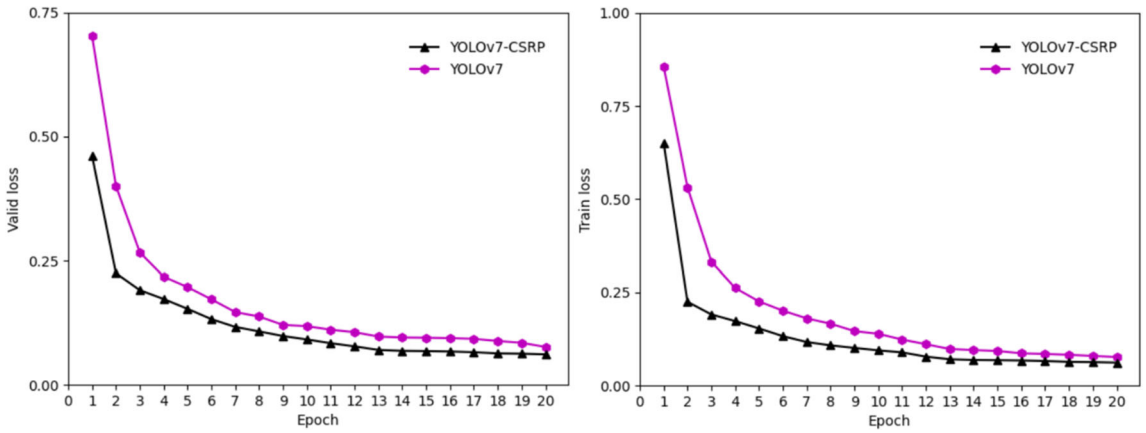


Fig. 8 The training and validation loss values before and after the incorporation of the CSRP network. ADFNet-CSRP represents a model that joins the Cross-Scale Residual Perception network based on YOLOv7

We present the training and validation loss values before and after the incorporation of the CSRP network for the first 20 epochs, as depicted in Fig. 8. It is evident from the initial 20 epochs that the training and validation loss values of ADFNet-CSRP decrease rapidly and converge quickly. Moreover, from the ablation experiments, it can be inferred that the ADFNet-CSRP model achieves an improvement in mAP within the same number of epochs. Therefore, this demonstrates that the proposed CSRP network effectively enhances the convergence speed.

To ensure fairness and impartiality, we controlled the parameters in the same scenario, where insulators were detected using different models. This allowed for a clearer comparison of the advantages and disadvantages of each model. The specific parameters are shown in Table 5, and we use FBS, UBS, GPU and P to, respectively, denote Frozen Batch Size, UnFrozen Batch Size, GPU Memory and Parameters.

It is clear from the table that it has approximately 39 MB more parameters than YOLOX-s, which has a minimum parameter count of 9 MB. The model we have designed has the highest parameter count, which is one of its disadvantages. The reason for the large number of parameters is the addition of three innovative ideas to YOLOv7: the Context Fusion Network, the Cross-Scale Residual Perception (CSRP) network and a Dual-Weighted Feature Fusion (DWFF) module, making the model a more complex structure. Despite having a larger number of parameters, the graphs and data above demonstrate that this model performs well in terms of detection accuracy and speed. Additionally, it has higher recall and F1 scores compared to other models. Therefore, these advantages largely compensate for the disadvantage of having a larger parameter count. If a higher-performance computer is used, the parameter count becomes less influential due to the excellent model, and it may also lead to even more outstanding detection results.

Based on the above tables and pictures, the performance of the model in insulator detection we designed is better than other models.

4.4.2 Experiment on public datasets

To verify the generality of this model, we used generic datasets such as RSOD dataset and NWPU VHR-10 dataset.

RSOD dataset is composed of 976 remote sensing images captured by the drone, which has a total of four typical scene categories, the labels are aircraft, oiltank,

Table 5 The parameter information for different detection models

Methods	Backbone	FBS	UBS	GPU (GB)	P(MB)
YOLOX [35]	CSPDarknet	8	4	24	9.0
SSD [43]	VGG-16	8	4	24	26.3
CenterNet [41]	ResNet50	8	4	24	32.7
GhostNet-YOLOv4 [42]	GhostNet	8	4	24	12.7
DETR [11]	ResNet50	8	4	24	36.8
YOLOv7 [17]	ELAN-Net	8	4	24	37.6
YOLOv8 [44]	CSPDarknet53	8	4	24	11.1
ADFNet (Ours)	ELAN-Net	8	4	24	47.9

Table 6 Test consequences on the RSOD dataset for distinct detection models

Detector	AP (%)				mAP (%)
	Playground	Overpass	Oiltank	Aircraft	
Deformable [11]	99.7	89.6	89.6	71.9	87.9
SSD [43]	100.0	93.5	98.9	57.1	81.5
CenterNet [41]	92.9	85.8	97.4	73.6	87.4
YOLOv7 [17]	99.0	91.0	100.0	98.0	96.8
YOLOv8 [44]	100.0	88.0	100.0	96.0	96.0
YOLOX [35]	99.0	87.0	100.0	97.0	95.8
ADFNet (Ours)	100.0	96.0	100.0	98.0	98.4

overpass and playground, respectively. ADFNet was tested on the RSOD dataset and compared with other typical models. The performance results are presented in Table 6. The RSOD dataset achieved the highest mAP value, with a value of 98.40% for the ADFNet method, which is 1.6% higher than the mAP value of YOLOv7. These results demonstrate the effectiveness of the approach for detecting small objects in satellite images. The findings indicate that the model improves the detection capability of small objects while maintaining the performance of detecting medium objects.

To confirm the effectiveness of Context Fusion Network, Cross-Scale Residual Perception (CSRP) network and DWFF modules for small object detection, an ablation experiment was conducted using the YOLOv7 model. The input image size was set to 640 × 640 pixels, the batch size was set to 8 and the training time for each network was set to 150 epochs. The experimental results are presented in Tables 6 and 7.

From Table 6, it can be observed that our model achieves the best detection performance on the RSOD dataset with a confidence level of 98.4%. Specifically, for the detection of playground and oiltank, the confidence level is 100%, while the lowest among other detection models are 92.9% and 85.8%, with the confidence level 7.1% and 14.2% higher, respectively. For the detection of overpass, the confidence level is 96%, while the lowest confidence level of CenterNet is 85.8%, resulting in a 10.2% higher confidence level. The confidence level for aircraft detection is 98%, which is 32.9% higher than the model with the lowest confidence level, SSD. In comparison, our method continues to demonstrate satisfactory performance on the RSOD dataset, further indicating its suitability for small targets and excellent detection speed for objects of various sizes. Therefore, the conclusion is consistent with the above, with excellent detection performance for multiple size targets.

Table 7 Precision, recall and F1 achieved by distinct detection models on the RSOD dataset

Detector	P(%)	R(%)	F1
YOLOX [35]	92.64	95.61	0.94
YOLOv7 [17]	94.82	93.65	0.94
YOLOv8 [44]	92.25	91.29	0.91
ADFNet (Ours)	96.39	92.53	0.94

The test results of precision, recall and F1 achieved by distinct detection models are demonstrated in Table 7. All values are higher than the other three models, except for recall, which is not particularly good on the RSOD dataset. Although recall is not as excellent as YOLOv7 and YOLOX, it achieves an extremely high mAP.

As shown in Fig. 9, the visualization results of distinct methods on the RSOD dataset demonstrate that the main advantage of our model is its ability to improve the detection accuracy of small targets without affecting the detection of large targets. This performance is further evidenced by the following visualization of the detection results, which effectively demonstrates this capability. In Fig. 9, the first and second lines showcase a plethora of aircraft images. Despite the partial visibility of the aircraft on the left side of the image, both ADFNet and YOLOX exhibit exceptional detection capabilities by successfully identifying the obscured targets. Moving on to the third and fourth lines, which display oiltank images, a minuscule pixel of an oil barrel can be found in the right and upper left corners, respectively. Remarkably, ADFNet demonstrates its superior detection prowess by successfully detecting these small objects with confidence scores of 0.66 and 0.56, respectively. On the contrary, the other three algorithms fail to recognize these objects. In the fifth row, the image showcases two different types of targets, namely, a playground and an aircraft. Although it is evident that an aircraft is present on the right side, only ADFNet demonstrates its exceptional detection capabilities by accurately predicting its presence with a confidence score of 0.58. Taken together, these observations highlight the strengths of our model, as it not only accurately detects small targets, but also shows improved detection capabilities for medium and large targets. These advantages are attributed to our

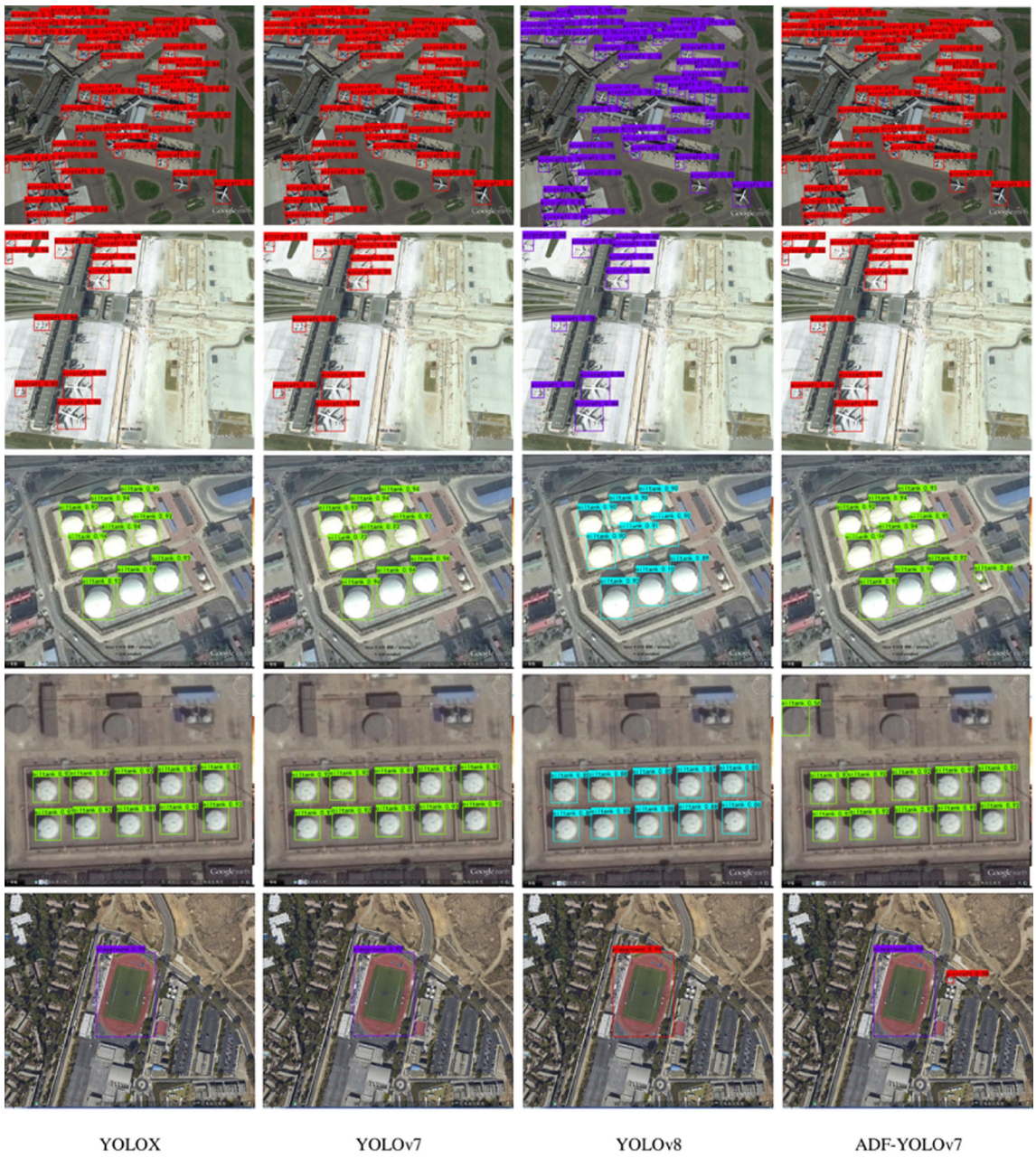


Fig. 9 Visualization results of distinct methods on the RSOD dataset. We observe that the results of the ADFNet are more accurate than other methods. Zoom in to see more details

innovation, which allows us to overcome the interference of background information by leveraging contextual information. It enables to fuse low-level information with high-level information across scales, and important information from different levels of feature maps to be extracted through weighted aggregation and then integrated. As a result, by leveraging the contextual information and effectively mitigating the impact of background interference, small objects that are occluded or partially blocked can be more accurately predicted. These results indicate that ADFNet exhibits excellent detection performance for varisized targets. There are two reasons for the excellent improvement results: First, the Context Fusion Network is designed to capture more contextual feature information, thereby improving the detection performance of small targets; second, the fusion module DWFF with weighted attention is adopted (instead of the common and simple feature fusion) to increase the expressiveness of the network.

NWPU VHR-10 dataset consists of 715 images with 2 m spatial resolution of and 85 images with 8 cm spatial resolution, which are 800 high-resolution RGB remote

Table 8 The comparative results of experiments on NWPU VHR-10 dataset

Models	P(%)	R(%)	F1(%)	mAP(%)
SSD [43]	78.35	78.35	0.79	81.26
YOLOv7-x [17]	92.82	92.82	0.92	95.70
YOLOv7-l	92.34	91.96	0.91	94.97
YOLOv8-x [44]	91.03	90.37	0.91	93.36
YOLOv8-s	88.64	89.31	0.89	92.88
YOLOv8-n	88.35	88.15	0.88	91.60
YOLOX-x [35]	82.41	83.02	0.82	85.91
YOLOX-s	81.65	82.14	0.81	85.03
ADFNet (Ours)	93.74	93.65	0.94	97.36

sensing images with sizes almost close to 1000 × 1000 based on satellite altogether. This dataset was clipped from the Google Earth and Vaihingen datasets, and then, experts annotate the images by hand. It is a geospatial object detection dataset of ten categories which are composed of airplane (PL), bridge (BR), ship (SP), baseball diamonds (BD), storage tanks (ST), basketball court (BC), tennis courts (TC), ground track field (GTF), vehicle (VH) and harbor (HB). The training set and the validation set account for 3/4 and 1/4 of the total images, respectively, which are chosen randomly.

Diverse models are tested in NWPU VHR-10 dataset, and the comparison results are shown in Table 8. This table comprises different versions of the YOLO series, SSD and our model. It is evident from observations that our proposed model achieves the best results across all metrics, with improvements to YOLOv7 increasing the mAP on this dataset by 2.39%. The mAP of ADFNet exceeds that of the worst-performing SSD by 16.10%, attributed to the proposed Context Fusion Network module, which enhances the model's attention to contextual information, thereby accurately distinguishing objects from background interference. ADFNet's precision, recall and F1 score are, respectively, 15.39%, 15.3% and 0.15 higher than those of the worst-performing models. These experimental results convincingly demonstrate the superiority of ADFNet, credited to the proposed Context Fusion Network, Cross-Scale Residual Perception network and Dual-Weighted Feature Fusion modules.

4.5 Ablation experiments

We conduct ablation experiments to validate the efficiency of ADFNet and the contribution of each network component. The ablation studies are based on insulator dataset and RSOD dataset, and results are presented in Table 9 and Table 10.

Adding the CFN, Cross-Scale Residual Perception (CSRP) network and DWFF modules increased the mAP of the ADFNet model on the insulator dataset by 3.97%, 3.87% and 3.94%, respectively. When all three modules were added to YOLOv7, the mAP increased by 4.04% compared to the YOLOv7 alone. Similarly, the detection

Table 9 The results of ablation experiments on insulator dataset

YOLOv7	CFN	CSRP	DWFF	F1	Precision (%)	Recall (%)	mAP (%)
√	–	–	–	0.92	94.25	90.46	94.15
√	√	–	–	0.97	94.81	97.01	98.12
√	–	√	–	0.95	93.79	96.78	98.02
√	–	–	√	0.96	95.03	97.25	98.09
√	√	√	–	0.95	95.01	95.11	98.06
√	√	-	√	0.96	96.15	95.42	98.13
√	–	√	√	0.95	93.63	96.32	98.05
√	√	√	√	0.95	93.78	97.13	98.19

Table 10 The results of ablation experiments on RSOD dataset

YOLOv7	CFN	CSRP	DWFF	F1	Precision (%)	Recall (%)	mAP (%)
√	–	–	–	0.94	94.81	93.65	96.85
√	√	–	–	0.94	96.06	92.48	98.15
√	–	√	–	0.95	96.75	93.52	98.06
√	–	–	√	0.94	95.41	92.70	98.01
√	√	√	√	0.94	96.39	94.53	98.40

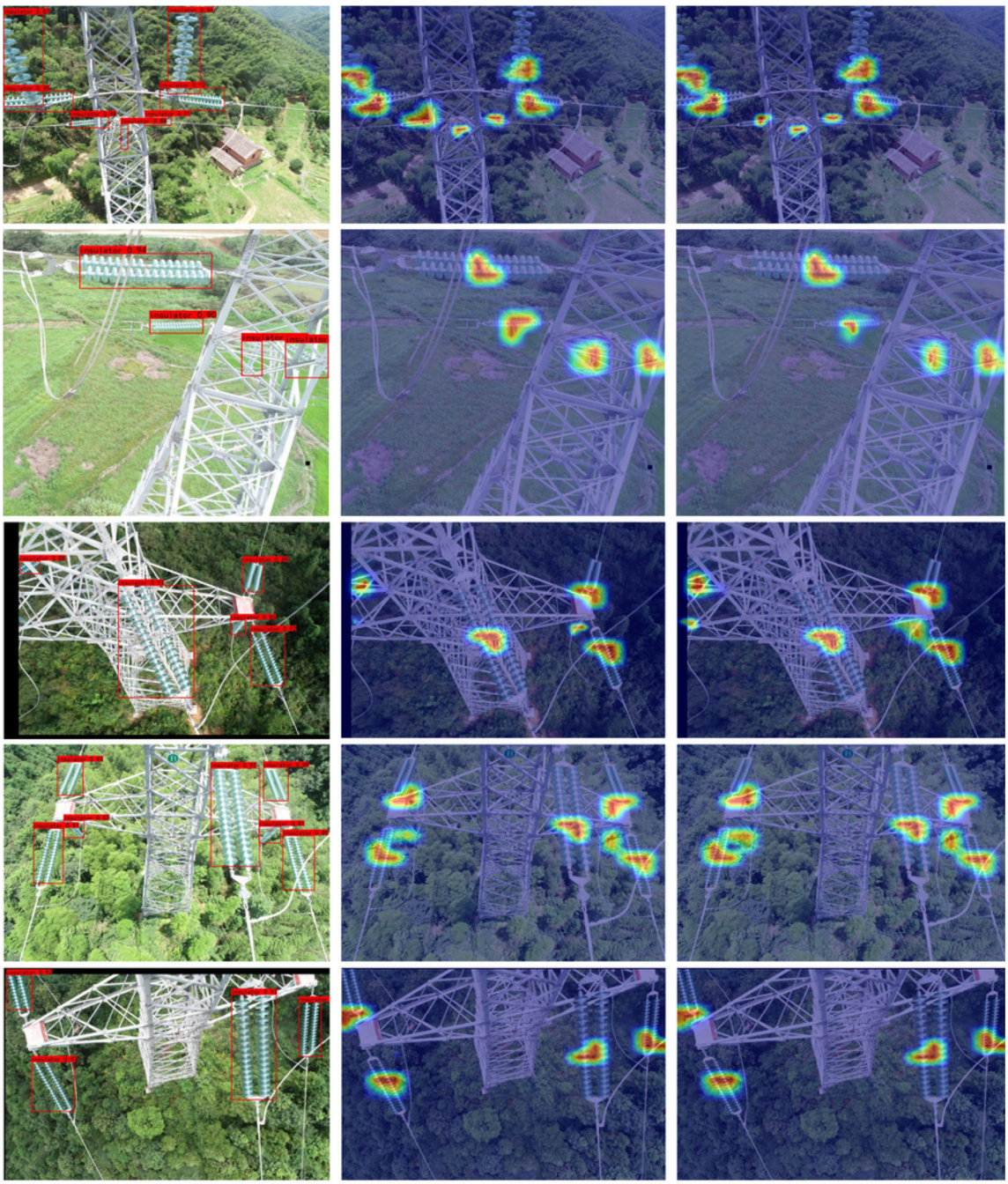


Fig. 10 Visualization results of saliency maps. For better visualization, the displayed activation maps were adjusted to the same spatial dimensions

results on the RSOD dataset rose by, respectively, 1.30%, 1.21% and 1.16%, respectively. The overall experimental result was an increase of 1.55% over the YOLOv7. Those tables reveal that F1, P and R are not entirely positively correlated with mAP. As shown in Table 9 for the insulator dataset, following the addition of each module, all metrics exhibit improvements compared to the source code. The CFN module achieves the highest F1 value, improving by 5.00% compared to the source code. The combination of CFN and DWFF modules achieves the highest precision, with a 6.80% increase compared to the source code. The DWFF module achieves the highest recall value, increasing by 6.79% compared to the source code. On the RSOD dataset, as indicated in Table 10, the CSRP module achieves the highest F1 and precision values, improving by 1.00% and 1.94%, respectively, compared to the source

code. ADFNet achieves the highest recall value, increasing by 0.88% compared to the source code and surpassing the worst value by 2.05%. These results suggest that the introduction of the CFN, Cross-Scale Residual Perception (CSRP) network and weighted feature fusion can effectively improve detection accuracy.

The graphs and tables above demonstrate the comparison between our model and other models, providing a comprehensive analysis that fully substantiates the universality of our model in all aspects. Specifically, our model exhibits exceptional performance in detecting small targets, highlighting its strengths in this area.

Through Grad-CAM, heatmap visualizations are generated for different prediction targets, as shown in Fig. 10. This indicates that network parameters assign varying weights to different parts of the object, resulting in different degrees of activation, facilitating a more intuitive interpretation. With the addition of three excellent modules compared to the original pre-trained features, object-aware features are more adept at separating the object from the background. The YOLOv7 and ADFNet columns display the input and output heatmaps of CFN, CSRP and DWFF.

The model integrates deep and shallow features through the CSRP module when processing text, thereby highlighting boundaries at the line or word level and enhancing semantic text regions at higher layers. Additionally, CFN employs pyramid dilated convolutions to capture high-level semantic features. The discriminative regions in the output primarily concentrate on crucial areas of the desired object. Consequently, the features learned by DWFF encompass both fundamental visual patterns and complementary high-level semantic information. We observe that highlighted regions in the image vary across spatial locations, including the background. Notably, features produced by ADFNet are more prominent and accurate, facilitating object detection.

5 Conclusions

This manuscript has proposed the ADFNet model to address the problem of poor detection accuracy of small insulators in electric power inspection. The ADFNet network structure is designed to optimize detection accuracy by extracting distinctive features between the critical features of insulators and the complicated background. In the meantime, though summarizing the most recent advancements and examine the pivotal characteristics of these research endeavors, we uncover the limitations of studies focused on detecting sheltered insulators and small insulators ground on deep learning techniques. To improve precision of electric power lines insulator detection, the method focuses on triple aspects: Firstly, the shallow-level features output from the backbone network were fed into Context Fusion Network and then introduced into Dual-Weighted Feature Fusion (DWFF) module, which effectively reduce the feature information loss of small targets and improve detection capabilities. Finally, these Cross-Scale Residual Perception (CSRP) networks are designed in order to collect more information from different levels, while also reducing the path distance from shallow to deep information. We implement the designed ADFNet on a self-made insulator dataset and the widely used RSOD dataset, obtaining mAPs of 98.19% and 98.4%, both of which outperformed other models. These demonstrate its robustness, and the ADFNet exhibited universality and excellent performance for detecting small targets. ADFNet has the ability to enhance small object detection by being integrated with different detectors and backbones.

It is proved that our method has great advantages in insulator detection, this implies that ADFNet can also be applied to other small components detection of electric power line (e.g., anti-vibration hammers, bird nests, etc.). Insulator detection in electric power line remains a burgeoning field of research, characterized by numerous challenges. In the future, should a more comprehensive dataset become accessible, the subsequent focus will be directed toward detecting various types of insulators and identifying insulator defects. Meanwhile, we would like to explore practical applications of ADFNet in more fields.

Author contributions Jie Zhang, Xiabing Wang and Dailin Li helped in conceptualization, methodology, writing—original draft preparation and reviewing; Yinhua Li, Fengxian Wang and Linwei Li worked in software and validation and Huanlong Zhang and Xiaoping Shi helped in examination of original draft and revisions.

Funding This work is supported by the grants from National Science Foundation of China (Nos. 62102373 and 62006213), Henan Province Key Research and Development Project (241111210400) and the Science and Technology Research Project of Henan province (No. 242102321034).

Data availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Not applicable.

Conflict of interest No potential conflict of interest was reported by the authors.

References

1. Liu J, Liu C, Wu Y, Xu H, Sun Z (2021) An improved method based on deep learning for insulator fault detection in diverse aerial images. Energies 14(14):4365
2. Liao S, An J (2014) A robust insulator detection algorithm based on local features and spatial orders for aerial images. IEEE Geosci Remote Sens Lett 12(5):963–967
3. Li B, Wu D, Cong Y, Xia Y, Tang Y (2012) A method of insulator detection from video sequence. In: 2012 Fourth international symposium on information science and engineering, pp. 386–389. IEEE
4. Zhai Y, Wang D, Zhang M, Wang J, Guo F (2017) Fault detection of insulator based on saliency and adaptive morphology. Multim Tools Appl 76:12051–12064
5. Zhai Y, Chen R, Yang Q, Li X, Zhao Z (2018) Insulator fault detection based on spatial morphological features of aerial images. IEEE Access 6:35316–35326
6. Chen J, Liu Z, Wang H, Núñez A, Han Z (2017) Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network. IEEE Trans Instrum Measurement 67(2):257–269
7. Huang Z, Hu S, Zhang L (2022) Fault detection of insulator in distribution network based on yolov5s neural network. In: 2022 international conference on artificial intelligence and computer information technology (AICIT), pp. 1–5. IEEE
8. Miao X, Liu X, Chen J, Zhuang S, Fan J, Jiang H (2019) Insulator detection in aerial images for transmission line inspection using single shot multibox detector. IEEE Access 7:9945–9956
9. Wang J, Li Y, Chen W (2022) Detection of glass insulators using deep neural networks based on optical imaging. Remote Sens 14(20):5153
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems **30**
11. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp. 213–229. Springer
12. Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159
13. Roh B, Shin J, Shin W, Kim S (2021) Sparse detr: Efficient end-to-end object detection with learnable sparsity. arXiv preprint arXiv:2111.14330
14. Gao P, Zheng M, Wang X, Dai J, Li H (2021) Fast convergence of detr with spatially modulated co-attention. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3621–3630
15. Zhang J, Wang F, Zhang H (2024) Compressive sensing spatially adaptive total variation method for high-noise astronomical image denoising. Vis Comput 40(2):1215–1227
16. Zhang J, Wang F, Zhang H (2023) A novel cs 2g-starlet denoising method for high noise astronomical image. Opt Laser Technol 163:109334
17. Wang C-Y, Bochkovskiy A, Liao H-YM (2023) Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7464–7475
18. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W, et al (2022) Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976
19. Dong R, Xu D, Zhao J, Jiao L, An J (2019) Sig-nms-based faster r-cnn combining transfer learning for small target detection in vhr optical remote sensing imagery. IEEE Tran Geosci Remote Sens 57(11):8534–8545
20. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
21. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 580–587
22. Liang J, Cui Y, Wang Q, Geng T, Wang W, Liu D (2024) Clusterfomer: Clustering as a universal visual learner. Advances in Neural Information Processing Systems **36**
23. Li S, He C, Li R, Zhang L (2022) A dual weighting label assignment scheme for object detection. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9387–9396
24. Kumar S, Kumar C (2023) Deep learning based target detection and recognition using yolo v5 algorithms from uavs surveillance feeds. In: 2023 international conference for advancement in technology (ICONAT), pp. 1–5. IEEE
25. Wang H, Han J (2023) Research on military target detection method based on yolo method. In: 2023 IEEE 3rd international conference on information technology, big data and artificial intelligence (ICIBA), vol. 3, pp. 1089–1093. IEEE
26. Yang W, Ding B, Tong LS (2022) Ts-yolo: An efficient yolo network for multi-scale object detection. In: 2022 IEEE 6th information technology and mechatronics engineering conference (ITOEC), vol. 6, pp. 656–660. IEEE
27. Zhu X, Lyu S, Wang X, Zhao Q (2021) Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In: proceedings of the IEEE/CVF international conference on computer vision, pp. 2778–2788
28. Li Z, Wang Z, He Y (2023) Aerial photography dense small target detection algorithm based on adaptive collaborative attention mechanism. J. Aeronaut 10:1–12
29. Mnih V, Heess N, Graves A, et al (2014) Recurrent models of visual attention. Adv Neural Inf Process Syst **27**
30. Jaderberg M, Simonyan K, Zisserman A, et al (2015) Spatial transformer networks. Adv Neural Inf Process Syst **28**
31. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7132–7141
32. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146–3154
33. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934
34. Ding X, Zhang X, Han J, Ding G (2022) Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11963–11975
35. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430
36. Ibrahim AWN, Ching PW, Seet GG, Lau WM, Czajewski W (2010) Moving objects detection and tracking framework for uav-based surveillance. In: 2010 Fourth Pacific-Rim Symposium on Image and Video Technology, pp. 456–461. IEEE
37. Li J, Wei Y, Liang X, Dong J, Xu T, Feng J, Yan S (2016) Attentive contexts for object detection. IEEE Trans Multim 19(5):944–954
38. Cheng Z, Choi H, Feng S, Liang JC, Tao G, Liu D, Zuzak M, Zhang X (2023) Fusion is not enough: Single modal attack on fusion models for 3d object detection. In: the Twelfth international conference on learning representations

39. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2117–2125
40. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8759–8768
41. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: Keypoint triplets for object detection. In: proceedings of the IEEE/CVF international conference on computer vision, pp. 6569–6578
42. Yuan X, Li D, Sun P, Wang G, Ma Y (2022) Real-time counting and height measurement of nursery seedlings based on ghostnet-yolov4 networkc and binocular vision technology. Forests 13(9):1459
43. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37. Springer
44. Jocher G, Chaurasia A, Qiu J (2023) Yolo by ultralytics. URL: https://github. com/ultralytics/ultralytics