Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085854

ARTICLE

A Cost-Sensitive Transformer-Based Network for Small Defect Detection in Power Transmission Line

Congcong Ma[1,2], **Wenqing Zhao**[1,3], **Feifei Fu**[1,2] **and Jiaqi Mi**[4,*]

[1]Department of Computer, North China Electric Power University, Baoding, China
[2]Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding, China
[3]Hebei Key Laboratory of Knowledge Computing for Energy & Power, Baoding, China
[4]College of Artificial Intelligence, Nankai University, Tianjin, China

*Corresponding Author: Jiaqi Mi. Email: mjq937794326@163.com

Received: 19 May 2026; Accepted: 06 August 2026; Published: 15 September 2026

ABSTRACT: Small object detection is a common and challenging task in transmission line inspection scenarios. Existing one-stage and two-stage object detection methods in this scenario are still constrained by extremely small object scale and limited feature representation capability, resulting in suboptimal performance in small object detection. To address these issues, this paper proposes a Cost-Sensitive Transformer-based Network for small object detection. First, a Transformer-based backbone is designed, coupled with a neck that integrates Feature Pyramid and Path Aggregation structures, enabling enhanced multi-scale feature extraction and fusion. Second, a hybrid-domain attention-based decoupled detection head is proposed, where an independent localization confidence branch enhances bounding box precision for small objects and reduces missed detections. Third, a cost-sensitive loss is designed to mitigate multi-dimensional imbalance in small object detection tasks. Extensive experiments are conducted on a proprietary transmission line inspection dataset and the COCO benchmark. The proposed method achieves a mean Average Precision (mAP) of 92.7% on the proprietary dataset, outperforming the state-of-the-art by 0.7%, and attains an APS of 35.9% on COCO, showing a significant improvement over the baseline. These results demonstrate the effectiveness and generalization capability of the proposed method for small object detection in transmission line inspection tasks.

KEYWORDS: Small object detection; transformer; cost-sensitive learning; transmission line inspection

1 Introduction

Object detection is a fundamental task in computer vision, aiming to accurately recognize and localize objects in images or videos [1]. Despite significant advances in deep learning, small object detection remains challenging due to weak feature representation, strict localization requirements, and severe class imbalance, resulting in substantial performance degradation [2]. Nevertheless, small object detection plays an indispensable role in critical applications such as public security surveillance [3], industrial precision inspection [4], medical image analysis [5], and defense-related remote sensing [6], highlighting its significant research value.

In transmission line inspection scenarios, the challenge of small object detection is particularly prominent. Common defects such as insulator damage, missing vibration dampers, and bird nests may lead to severe consequences, including line faults and even large-scale power outages. However, these defects

typically occupy only a very small number of pixels in inspection images. Therefore, accurate detection of such small defects is of critical importance for ensuring the safety and reliability of power infrastructure.

Existing methods for small defects detection in power transmission line can be broadly categorized into one-stage, two-stage, and emerging Transformer-based methods. One-stage detectors prioritize efficiency, while two-stage detectors improve localization via region proposals; however, both remain limited in small object scenarios. Transformer-based methods provide an effective alternative by leveraging self-attention mechanisms to capture global representations and cross-region dependencies that are difficult to obtain using conventional local feature extraction schemes. However, their direct application to transmission line inspection is still suboptimal, as small defects often exhibit weak feature responses and are easily overwhelmed by dominant background clutter in highly imbalanced scenes.

To overcome the aforementioned challenges, we develop a Cost-Sensitive Transformer-based Network for Small Defect Detection (CSTD). The major contributions of this study are summarized as follows:

(1) To address the challenge of limited information in small defects, a Transformer-based backbone is introduced to enhance discriminative feature extraction. Meanwhile, the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) are employed for multi-scale feature fusion, improving the interaction between fine-grained details and semantic context.

(2) To alleviate the sensitivity of small objects to localization errors, a decoupled detection head based on a hybrid-domain attention mechanism is designed. In addition, an independent localization confidence prediction branch is introduced to improve bounding box localization accuracy and reduce missed and false detections of small objects.

(3) To address the imbalance between foreground and background samples as well as the uneven distribution of objects at different scales, a cost-sensitive loss function is proposed to guide the model to focus more on small-object regions, thereby improving the detection performance of small-scale defect targets.

(4) Extensive experiments are conducted on a proprietary transmission line inspection dataset and the public COCO dataset. Specifically, the proposed method achieves 92.7% mAP on the inspection dataset, outperforming existing state-of-the-art methods by 0.7%. On the COCO dataset, the proposed method achieves an APS of 35.9%, showing a significant improvement over the baseline model.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents the experimental settings and result analysis. Section 6 concludes the paper with discussion and conclusions.

2 Related Works

This section reviews related work on general object detection methods, covering conventional two-stage and one-stage detectors, followed by a discussion of recent developments in transmission line defect detection.

2.1 Two-Stage Object Detection Methods

Two-stage detectors typically adopt a proposal-based detection paradigm, where potential object regions are first generated and subsequently refined through region-wise feature extraction and prediction. Based on this framework, the final object categories and bounding boxes are obtained through classification and regression operations [7]. Representative methods include Faster R-CNN [8], Mask R-CNN [9], Cascade R-CNN [10], ViT-FRCNN [11], and Dynamic R-CNN [12]. Sparse R-CNN [13] further introduces a sparse detection paradigm with learnable proposals and dynamic convolution, improving both efficiency and accuracy.

By explicitly separating foreground and background through region proposals, two-stage methods mitigate class imbalance to some extent [14]. However, the decoupled proposal generation and per-region processing lead to redundant feature computation, resulting in higher model complexity and slower inference.

2.2 One-Stage Object Detection Methods

One-stage detectors directly perform object localization and classification from feature representations in a unified manner, avoiding the intermediate proposal generation process. Representative methods include the YOLO series [15], SSD [16], RetinaNet [14], and FCOS [17], as well as recent variants such as Focus DETR [18] and Salience DETR [19].

Recent advances in Transformer-based architectures have further expanded one-stage detection paradigms. DETR [20] introduces an end-to-end framework that eliminates post-processing steps such as NMS, thereby simplifying the detection pipeline. Swin Transformer [21] enhances representation capacity while maintaining efficiency through a hierarchical design and shifted window mechanism. DINO [22] improves DETR by incorporating denoising training, hybrid query initialization, and a two-stage optimization strategy, significantly boosting performance. DRFNet [23] introduces a dynamic receptive field module that mimics adaptive visual perception, integrating global and local information to improve small object detection. Recent studies [24] have demonstrated that hybrid CNN-Transformer architectures can effectively combine local representation and global dependency modeling. Tian et al. [25] verified its effectiveness in image denoising by integrating CNN-based features with Transformer-based long-range dependencies. In general, one-stage detectors provide advantages in terms of computational efficiency and deployment simplicity. Nevertheless, their detection performance may still degrade when facing challenging scenarios involving small targets, complex backgrounds, and limited visual information.

2.3 Defect Detection for Transmission Line Inspection

With the increasing adoption of UAV inspection and intelligent maintenance technologies in power systems, vision-based transmission line defect detection has received growing research interest. Existing approaches mainly focus on addressing challenges such as complex environmental interference, small defect scales, and real-time deployment requirements.

For instance, Xu and Tang proposed MAP-YOLOv8 [26] for insulator defect detection, which improves feature fusion in the neck network and introduces attention mechanisms, achieving enhanced accuracy while maintaining real-time performance and model compactness. Yang et al. proposed STDISNet [27] based on Swin Transformer for UAV-based transmission line inspection, by incorporating hierarchical multi-scale modeling and a lightweight feature fusion module, the method achieves strong robustness under occlusion, cluttered backgrounds, and low-contrast conditions. To further improve real-time performance and deployment efficiency, More and Bansode proposed FCN-YOLOS [28], which integrates the strong feature representation capability of Faster R-CNN with the high-speed inference of YOLOv8, and employs neural architecture search (NAS) for balanced hyperparameter optimization, achieving a trade-off between accuracy and efficiency. In addition, Conv-DDLSTM [29] introduces model compression strategies including pruning, quantization, and knowledge distillation, significantly reducing computational cost and energy consumption while maintaining competitive detection performance, making it suitable for resource-constrained edge devices.

3 Method

3.1 Overall Architecture

Small defect detection in transmission line inspection remains challenging due to the extremely small object scale, limited discriminative features, and severe data imbalance. Small defects usually occupy only a few pixels in inspection images, making their feature representations vulnerable to information loss during deep feature extraction. Meanwhile, significant imbalance among positive and negative samples as well as objects of different scales prevents conventional loss functions from adequately focusing on difficult small-object samples. To address these challenges, this paper proposes a Cost-Sensitive Transformer-based Network for Small Defect Detection. The proposed framework improves small object detection from three complementary perspectives: feature representation, detection head design, and optimization objective.

Fig. 1 illustrates the overall architecture of the proposed CSTD model. Swin-T is adopted as the backbone network for feature extraction. The window attention mechanism enhances local detail representation for small objects, while the shifted window strategy facilitates cross-window contextual interaction, thereby improving small-object feature modeling capability. In the neck, an FPN and PAN-based structure is designed for multi-scale feature fusion. Its fundamental building blocks include several key convolutional modules which are derived from the YOLO series [30]. These modules facilitate hierarchical feature fusion and improve the representation capability of small defects. Specifically, Convolution-Batch Normalization-SiLU (CBS) serves as the basic convolutional unit for feature extraction and nonlinear representation; Efficient Layer Aggregation Network (ELAN) enhances multi-level feature aggregation; Max Pooling (MP) enlarges the receptive field by downsampling; and Spatial Pyramid Pooling (SPP) captures multi-scale contextual information through spatial pyramid pooling. Finally, three detection heads are constructed on multi-scale feature maps for objects of different sizes. High-resolution feature maps preserve richer detail information and help reduce missed detections of small objects.

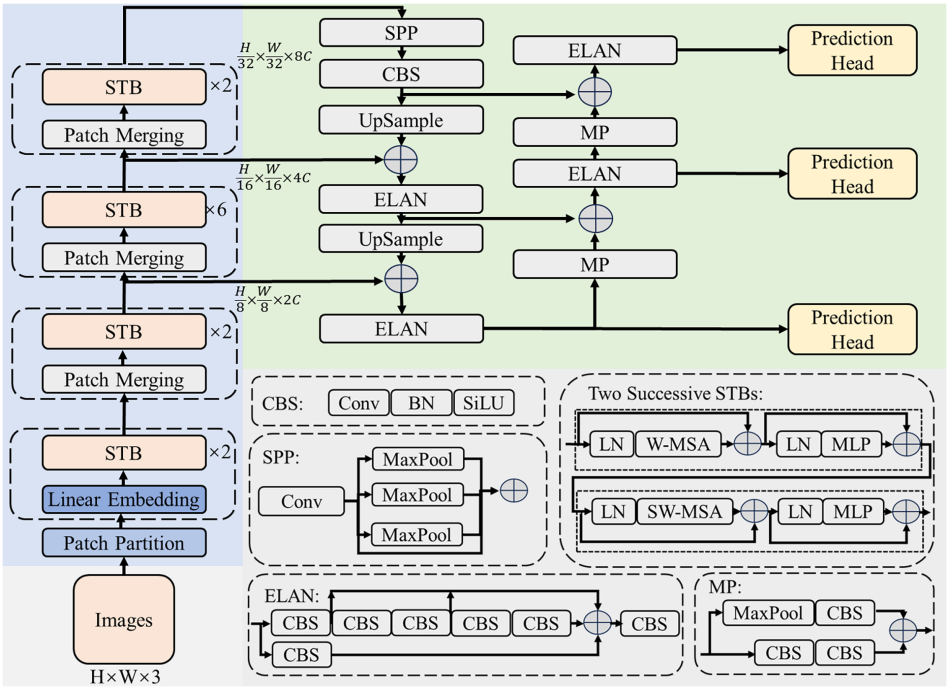

Figure 1: Overall architecture of the CSTD model. The Swin Transformer Block (STB), including Layer Normalization (LN), Multilayer Perceptron (MLP), Window-based Multi-Head Self-Attention (W-MSA), and Shifted Window-based Multi-Head Self-Attention (SW-MSA), is adopted from the Swin Transformer architecture [21]. Copyright 2021 IEEE.

The data flow of the proposed framework can be summarized as follows. Given an input image, multi-scale features are first extracted by the Transformer backbone and subsequently fused through the neck network. The fused features are then fed into the decoupled detection head for classification, localization, and localization confidence prediction. Finally, the proposed cost-sensitive loss function jointly optimizes the entire network for accurate small defect detection in transmission line inspection scenarios.

It should be noted that although Transformer-based backbones exhibit superior feature representation capability due to the self-attention mechanism, the proposed method differs from DETR-style frameworks that formulate object detection as a sequence-to-sequence learning task. Instead, traditional detection components, including anchor-based mechanisms, one-to-many label assignment, multi-scale feature map detection, and proposal generation, are retained. These strategies effectively enhance localization and classification performance, particularly for small-object detection. The main contributions of this work lie in the proposed Hybrid-Domain Attention Decoupled Detection Head and the Cost-Sensitive Loss Function, which are described in detail in Sections 3.2 and 3.3, respectively.

3.2 Decoupled Detection Head with Hybrid-Domain Attention

A decoupled detection head based on a hybrid-domain attention mechanism is designed, as illustrated in Fig. 2. Each detection head consists of three decoupled branches for object classification, bounding box regression, and localization confidence prediction, respectively. An additional localization confidence branch is introduced to improve bounding box localization accuracy and reduce missed detections of small objects [31]. Furthermore, a hybrid-domain attention mechanism is incorporated into the detection head to enhance contextual feature representation around target regions, enabling the model to focus adaptively on informative regions and discriminative features.

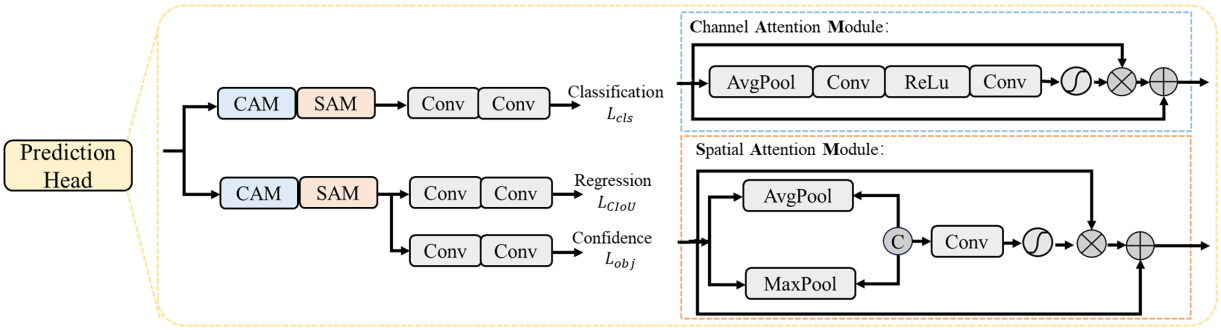

Figure 2: The decoupled detection head.

Specifically, the Channel Attention Module (CAM) first applies global average pooling to the input feature map to generate a 1 × 1 × C channel descriptor. Two convolution layers are then employed to learn channel attention weights, followed by Sigmoid normalization. The resulting attention features are fused with the input features through a long skip connection to produce the CAM output. The Spatial Attention Module (SAM) performs global average pooling and global max pooling on the input feature map to generate two H × W × 1 spatial descriptors. The spatial attention weights are learned through convolution operations and normalized using a Sigmoid function. The generated spatial attention features are then fused with the input features via a long skip connection to obtain the SAM output. The hybrid-domain attention mechanism effectively optimizes attention allocation and guides the network to focus on more informative features, thereby improving small-object detection performance. The CAM and SAM modules are adopted from standard attention designs and are used as feature refinement components rather than novel architectural contributions.

3.3 Cost-Sensitive Loss Function

A cost-sensitive loss function is designed for the proposed CSTD model to improve localization accuracy and alleviate multi-dimensional imbalance in small-object detection. In addition, an enhanced localization confidence loss is introduced to reduce missed detections of small objects. The overall loss function consists of three components: regression loss L_{reg}, classification loss L_{cls}, and confidence loss L_{conf}.

$$Loss = \lambda_{box} L_{reg} + \lambda_{cls} L_{cls} + \lambda_{obj} L_{conf} \tag{1}$$

Here, λ_{box} denotes the weight of the regression loss, λ_{cls} denotes the weight of the classification loss, and λ_{obj} denotes the weight of the confidence loss.

3.3.1 Regression Loss

The bounding box regression loss adopts the CIoU loss [32], which jointly considers overlap ratio, center distance, and aspect ratio consistency between predicted and ground-truth boxes. It provides comprehensive bounding box optimization and is particularly effective for small-object localization.

$$L_{\text{reg}}(p, g) = 1 - \text{IoU}(p, g) + \frac{\rho^2(p, g)}{c^2} + \alpha\vartheta \tag{2}$$

where $\text{IoU}(p, g)$ denotes the intersection-over-union between the predicted box and ground-truth box, $\rho(p, g)$ represents the Euclidean distance between their center points, and c is the diagonal length of the minimum enclosing box. The aspect ratio penalty term is defined as

$$\vartheta = \frac{4}{\pi^2}\left(\arctan\frac{w_g}{h_g} - \arctan\frac{w_p}{h_p}\right)^2 \tag{3}$$

and the balancing coefficient is

$$\alpha = \frac{\vartheta}{1 - \text{IoU}(p, g) + \vartheta} \tag{4}$$

where $\frac{w_g}{h_g}$ and $\frac{w_p}{h_p}$ denote the aspect ratios of the ground-truth box and predicted box, respectively.

3.3.2 Classification Loss

Based on Focal Loss [14], a cost-sensitive factor is introduced to increase the contribution of small objects during optimization. The classification loss is formulated as

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} -\varphi_i (1 - p_{y_i})^{\gamma} \log(p_{y_i}) \tag{5}$$

where γ is the focusing parameter of Focal Loss (typically set to 2), y_i denotes the ground-truth class index, (p_{y_i}) is the predicted probability of class y_i, and N is the number of objects.

The cost-sensitive factor is defined as:

$$\varphi_i = 1 + \left[1 + \exp\left(\beta - \frac{C_{Image}}{\mu C_{g_i}}\right)\right]^{-1} \tag{6}$$

where C_{Image} denotes the diagonal length of the input image, and C_{g_i} represents the diagonal length of the ground-truth box for object i. μ and β are hyperparameters, typically set to 5 and 4, respectively. The value of φ_i lies in (1, 2).

By design, the cost-sensitive factor assigns larger weights to smaller objects, thereby increasing the penalty for misclassification of small targets and improving detection performance in small-object scenarios.

3.3.3 Confidence Loss

To further improve the confidence prediction branch, a cost-sensitive confidence loss is introduced to enhance the model's sensitivity to small-scale objects, reduce missed detections, and improve overall detection performance. Based on the standard binary cross-entropy confidence loss, a cost-sensitive factor is incorporated to increase the response of the model to small-object predictions, thereby mitigating false negatives for small targets.

The confidence loss is formulated as:

$$L_{conf} = -\phi_i \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right] \tag{7}$$

where $y_i \in \{0, 1\}$ denotes the ground-truth label, with $y_i = 1$ indicating the presence of an object in the predicted bounding box and $y_i = 0$ otherwise. $p_i \in (0, 1)$ represents the predicted objectness confidence score, and ϕ_i is the cost-sensitive factor.

4 Experiment Setting

4.1 Dataset

To validate the effectiveness and generalization capability of the proposed CSTD for small-object defect detection in transmission line inspection scenarios, experiments were conducted on both a transmission line inspection dataset and the public COCO [33] dataset. The transmission line inspection dataset contains 800 images covering three representative defect categories: insulator self-shattering defects, bird-nest foreign-object defects, and vibration damper detachment defects. Following the object-scale definition strategy of the COCO dataset and considering the characteristics of transmission line inspection images, objects with pixel areas smaller than 0.5% of the image area are defined as small objects, objects larger than 2% are defined as large objects, and the remaining objects are categorized as medium objects. The datasat distribution is presented in Table 1. Among all annotated objects, 973 are small-scale targets, accounting for approximately 75% of the dataset, indicating a significant small-object distribution characteristic. The dataset is divided into training and testing sets with a ratio of 8:2.

Table 1: Distribution of the transmission line inspection dataset.

Defect Categories	Insulator	Bird-Nest	Vibration Damper	Total
Small objects	42	132	799	973
Medium objects	49	82	85	216
Large objects	17	84	0	101
Total	108	298	884	1290

4.2 Evaluation Metrics

Considering the differences between the transmission line inspection dataset and the large-scale generic COCO dataset in terms of dataset scale, task characteristics, and evaluation objectives, different evaluation protocols are adopted in this study. For the COCO dataset, the official evaluation protocol of the dataset is adopted, including metrics such as AP, $AP50$, $AP75$, APS, APM, and APL [33]. For the transmission line inspection dataset, due to its relatively small scale and task-specific defect detection characteristics, the average precision (AP) under a fixed IoU threshold of 0.5 is employed as the primary evaluation metric for different categories and object scales, thereby providing a more intuitive assessment of the model's detection performance and small-object recognition capability in practical engineering scenarios. AP_{bir}, AP_{vib}, and AP_{ins} denote the detection performance for bird-nest, vibration damper, and insulator classes, respectively. Mean Average Precision (mAP) is used to evaluate the overall detection performance across all categories. Furthermore, considering the dominance of small objects in the transmission line inspection dataset, scale-specific metrics AP_S, AP_M, and AP_L are introduced to evaluate detection performance for small, medium, and large objects, respectively.

All reported results are obtained using 5-fold cross-validation, and the final performance is calculated as the average over all folds.

4.3 Experimental Environment

All experiments were conducted on an NVIDIA GeForce RTX 3060 Ti GPU with 8 GB of memory. The GPU features a memory clock of 14 GHz and a 256-bit memory interface. The system is equipped with an Intel Core i7-12700K CPU and 32 GB of RAM, running on Windows 10. The model was implemented using the PyTorch framework. PyCharm was adopted as the integrated development environment (IDE), and Python was used as the primary programming language.

4.4 Hyperparameter Settings

In the experimental setup, the loss function weights λ_{box}, λ_{cls}, and λ_{obj} are set to 2, 1, and 1, respectively. The model is trained for 36 epochs using the AdamW optimizer with an initial learning rate of 0.0001, a weight decay of 0.05, and a batch size of 16. Furthermore, the hyperparameters μ and β are optimized through parameter searches, with the results presented in Fig. 3a, b, respectively.

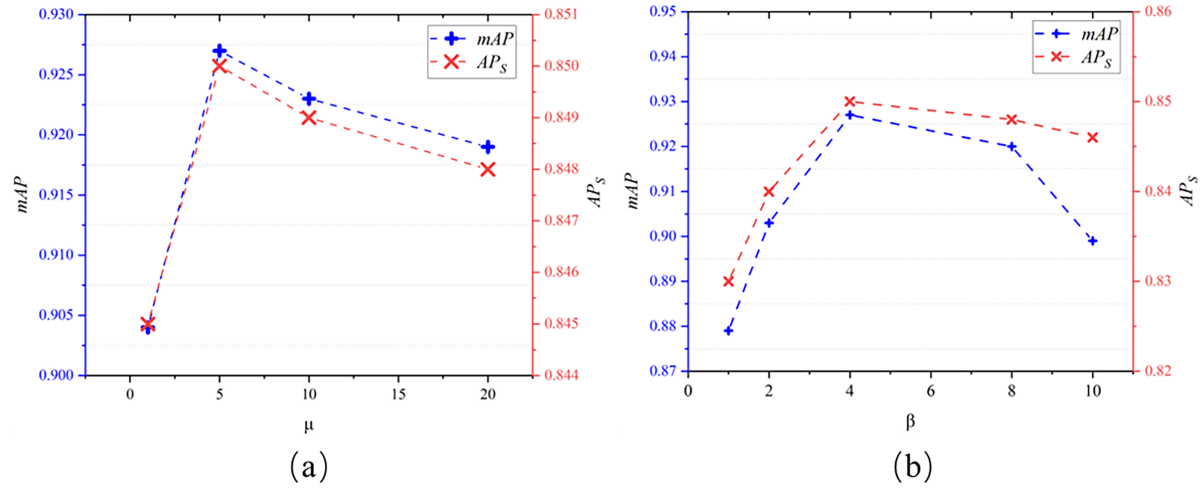

Figure 3: Results of hyperparameter tuning.

Hyperparameter μ. To investigate the effect of the hyperparameter μ on model performance, μ is set to 0, 5, 10, 15, and 20, respectively, and evaluated using the mAP and AP_S metrics. The results are shown in Fig. 3a. As μ increases, the detection performance first improves and then declines. When $\mu = 5$, both mAP and AP_S achieve their optimal values, indicating the best balance between overall detection performance and small-object detection capability. Therefore, μ is finally set to 5.

Hyperparameter β. To investigate the effect of the hyperparameter β on model performance, β is set to 0, 2, 4, 6, 8, and 10, respectively, and evaluated using the mAP and AP_S metrics. The results are shown in Fig. 3b. As β increases, the detection performance also exhibits a trend of first increasing and then decreasing. When $\beta = 4$, both mAP and AP_S reach their maximum values, demonstrating the best trade-off between overall detection performance and small-object detection capability. Therefore, β is finally set to 4.

4.5 Training and Implementation Details

Data Augmentation Strategy. To mitigate overfitting caused by the limited size of the training dataset, an online data augmentation strategy is adopted in this study. Specifically, the proposed augmentation pipeline includes random horizontal flipping (with a probability of 0.5), color jittering (including random variations in brightness, contrast, and saturation), random scaling, and Mosaic augmentation (with a probability of 0.5). Furthermore, all augmented images are resized to 640 × 640 using the Letterbox strategy to ensure consistent input resolution and stable matching between multi-scale detection heads and anchor boxes.

Transfer Learning Strategy. Specifically, the proposed CSTD model is first pre-trained on the PASCAL VOC2012 dataset. During pre-training, the input resolution follows the same resizing strategy as data augmentation, while maintaining a uniform batch-wise scale with a batch size of 16. The model is trained for 20 epochs with an initial learning rate of 1×10^{-3}. After pre-training, the detection head is modified to match the number of categories in the power line defect dataset. The backbone network is frozen, while the classification layers are randomly initialized. The model is then fine-tuned on the target dataset for 36 epochs with a reduced learning rate of 1×10^{-4}.

Label Assignment Strategy. An anchor-based one-to-many label assignment strategy is adopted in this work. All input images are resized to 640 × 640 using the Letterbox strategy to ensure geometric consistency. The detection head operates on three feature scales with strides of 8, 16, and 32, respectively. Each scale is assigned three anchor boxes, resulting in a total of nine anchors. The anchor configurations are defined as: (10, 13), (16, 32), (17, 42), (38, 62), (61, 33), (57, 101), (116, 101), (122, 137), (178, 185), which are obtained via K-means clustering on the annotated bounding boxes of the power line defect dataset. During the first five epochs, a fixed IoU threshold of 0.5 is used for positive sample assignment. After the warm-up stage, an ATSS-inspired dynamic matching strategy is introduced.

5 Experiments

5.1 Ablation Experiments

Three groups of ablation experiments are conducted to evaluate the effectiveness of different components in the proposed CSTD model.

To investigate the influence of backbone networks, ResNet50, VGG, and Darknet53 are adopted as feature extraction backbones within the CSTD framework and compared with the complete CSTD model. The results are presented in Table 2. Compared with the baseline backbone networks, the adopted backbone improves mAP, AP_S, AP_M, and AP_L by 7.7%, 2.6%, 3.7%, and 4.9%, respectively.

Table 2: Ablation study of different backbone networks (%).

Methods	AP_{bir}	AP_{vib}	AP_{ins}	mAP	AP_S	AP_M	AP_L
CSTD-ResNet50	74.0	86.9	98.4	86.4	83.0	88.8	96.3
CSTD-VGG	72.6	84.2	98.2	85.0	82.4	88.0	94.0
CSTD-Darknet53	<u>80.6</u>	<u>90.2</u>	<u>98.6</u>	<u>89.8</u>	<u>83.9</u>	<u>90.1</u>	<u>98.0</u>
CSTD (Ours)	**83.3**	**95.9**	**98.8**	**92.7**	**85.0**	**91.7**	**98.9**

Note: The best results are highlighted in bold, and the second-best results are underlined.

To analyze the influence of different detection heads, a hybrid-attention decoupled head, a conventional convolution-based decoupled head, and a non-decoupled head are evaluated, as shown in Table 3. Compared with the convolution-based decoupled head, the proposed detection head improves mAP, AP_S, AP_M, and AP_L by 1.7%, 1.9%, 2.0%, and 0.6%, respectively. Compared with the non-decoupled head, the corresponding improvements are 1.0%, 1.1%, 1.2%, and 0.5%, respectively.

Table 3: Ablation study of different detection heads (%).

Methods	AP_{bir}	AP_{vib}	AP_{ins}	mAP	AP_S	AP_M	AP_L
CSTD-ConvHead	83.2	92.7	97.0	91.0	83.1	89.7	98.3
CSTD-NDHead	**83.7**	<u>93.7</u>	<u>97.8</u>	<u>91.7</u>	<u>83.9</u>	<u>90.5</u>	<u>98.4</u>
CSTD (Ours)	<u>83.3</u>	**95.9**	**98.8**	**92.7**	**85.0**	**91.7**	**98.9**

Note: The best results are highlighted in bold, and the second-best results are underlined.

The influence of different loss functions on model performance is reported in Table 4. Specifically, $CSTD - loss_{Focal}$ denotes the model using the standard Focal Loss as the classification loss, while $CSTD - loss_{BCE}$ denotes the model using the standard binary cross-entropy (BCE) loss as the confidence loss. Compared with the standard Focal Loss, the proposed cost-sensitive classification loss improves mAP, AP_S, AP_M, and AP_L by 2.1%, 1.0%, 1.5%, and 0.8%, respectively. Compared with the standard BCE loss, the proposed confidence loss improves these metrics by 1.0%, 1.3%, 1.3%, and 0.4%, respectively. These results demonstrate that the proposed loss functions effectively enhance the overall detection performance, particularly for small- and medium-scale objects.

Table 4: Ablation study of different loss functions (%).

Methods	AP_{bir}	AP_{vib}	AP_{ins}	mAP	AP_S	AP_M	AP_L
$CSTD - loss_{Focal}$	82.8	90.4	<u>98.5</u>	90.6	<u>84.0</u>	90.2	<u>98.1</u>
$CSTD - loss_{BCE}$	<u>82.9</u>	<u>94.9</u>	97.4	<u>91.7</u>	83.7	<u>90.4</u>	98.5
CSTD (Ours)	**83.3**	**95.9**	**98.8**	**92.7**	**85.0**	**91.7**	**98.9**

Note: The best results are highlighted in bold, and the second-best results are underlined.

5.2 Comparative Experiments

The proposed CSTD model is compared with seven representative object detection methods on the transmission line inspection dataset, and the results are presented in Table 5. As shown in Table 5, the proposed CSTD achieves the best overall detection performance among all compared methods. Compared with the second-best YOLOv8 model, CSTD improves mAP, AP_S, and AP_L by 0.7%, 0.8%, and 1.3%,

respectively. Moreover, CSTD achieves the best performance in terms of overall accuracy and small- and large-scale object detection, as indicated by the highest values of mAP, AP_S, and AP_L. These results demonstrate that the proposed CSTD model exhibits superior overall detection performance and strong multi-scale object detection capability for transmission line defect detection tasks.

Table 5: Comparison of different object detection models on the transmission tower inspection dataset (%).

Methods	AP_{bir}	AP_{vib}	AP_{ins}	mAP	AP_S	AP_M	AP_L
Faster-RCNN	26.1	19.7	95.7	47.2	63.6	73.1	34.8
YOLOv3	80.0	82.9	96.3	86.4	79.8	86.1	95.2
FCOS	56.2	62.2	83.0	67.1	56.2	68.7	76.4
DETR	73.2	88.3	98.1	86.5	82.9	88.9	97.1
Swin	83.3	92.0	98.3	91.2	84.1	90.5	97.2
DINO	67.5	84.1	<u>98.4</u>	83.3	81.3	87.8	94.5
DRFNet	80.0	<u>92.9</u>	98.3	90.4	84.0	90.4	<u>98.0</u>
YOLOv5	67.4	81.1	97.9	82.1	81.9	82.7	90.7
YOLOv8	**85.1**	92.9	97.9	<u>92.0</u>	<u>84.2</u>	**92.1**	97.6
CSTD (Ours)	<u>83.3</u>	**95.9**	**98.8**	**92.7**	**85.0**	<u>91.7</u>	**98.9**

Note: The best results are highlighted in bold, and the second-best results are underlined.

Fig. 4 further illustrates the localization accuracy on an insulator instance. The top-left image shows the original image with ground truth annotations, while the remaining subfigures present zoomed-in comparisons between different methods and the proposed CSTD. In the visualization, green boxes denote ground truth, red boxes denote predictions from CSTD, and blue boxes denote predictions from comparison methods. The corresponding confidence scores are also reported in the figure.

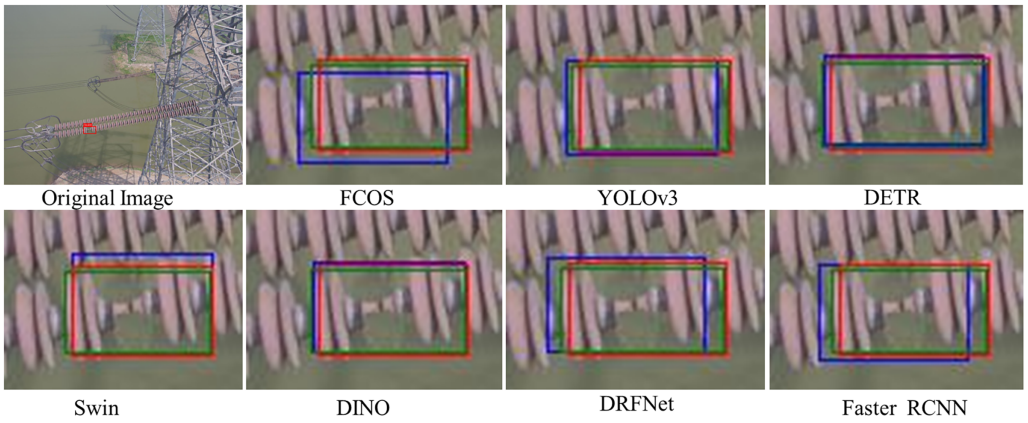

Figure 4: Visualization of localization accuracy of different models for insulator defects in transmission line. The confidence scores of the predicted bounding boxes generated by different methods are as follows: CSTD = 0.82, FCOS = 0.72, YOLOv3 = 0.87, DETR = 0.78, Swin = 0.74, DINO = 0.80, DRFNet = 0.74, Faster RCNN = 0.77.

Fig. 5 presents a qualitative comparison between the proposed CSTD and several representative methods on the transmission line inspection dataset, providing visual evidence of the model's superiority in reducing missed detections of small objects. According to the ground truth (GT) annotations, the original image contains 12 vibration damper instances. As shown in the results, both CSTD and Swin achieve the best performance on small-object detection, each missing only one object. DINO misses two objects, DETR and

DRFNet each miss three objects, while Faster R-CNN and YOLOv3 exhibit the poorest performance, each missing five objects. These results demonstrate the superior small-object detection capability of the proposed CSTD model. Despite the strong performance of CSTD, failure cases still occur for extremely small and low-contrast objects due to weak feature representation and background interference, suggesting the need for more robust fine-grained feature enhancement in future work.

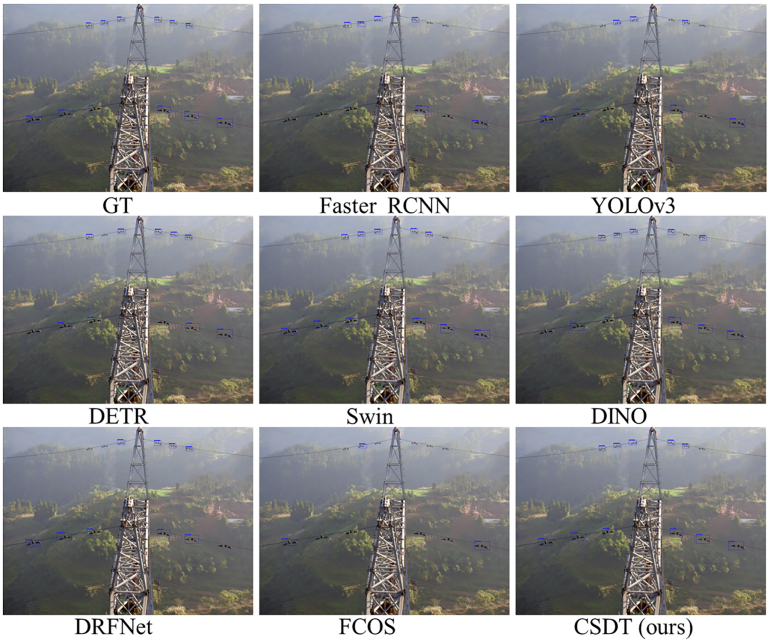

Figure 5: Visualization of small-object detection results of different models.

Overall, the visualization results and confidence analysis indicate that CSTD achieves higher localization accuracy and confidence levels on insulator detection, outperforming competing methods and further validating its effectiveness.

Table 6 shows that CSTD achieves a good balance between accuracy and efficiency. It maintains moderate parameter size and computational cost while achieving real-time inference speed (30 FPS), outperforming most Transformer-based detectors in efficiency. Overall, the proposed method provides a favorable trade-off between detection performance and computational complexity.

Table 6: Model complexity analysis.

Metrics	Faster-RCNN	YOLOv3	FCOS	DETR	Swin	DINO	DRFNet	CSTD (Ours)
Params	134M	62M	68M	137M	**48M**	234M	74M	<u>55M</u>
FPS	12	**32**	25	15	28	5	28	<u>30</u>
FLOPs	215G	<u>155G</u>	**132G**	210G	267G	284G	180G	215G

Note: The best results are highlighted in bold, and the second-best results are underlined.

5.3 Generalization Experiments

The generalization performance of the proposed method is evaluated on the COCO benchmark, and the results are reported in Table 7.

Table 7: Performance comparison of different object detection models on the COCO dataset (%).

Methods	Epochs	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	36	ResNet-101	42.0	62.5	45.9	25.2	45.6	54.6
YOLOv3	–	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
RetinaNet	36	ResNet-101	40.4	60.2	43.2	24.0	44.3	52.2
FCOS	–	ResNeXt-101	44.7	64.1	48.4	27.6	47.5	55.6
DETR	500	ResNet-50	43.3	63.1	45.9	22.5	47.3	61.1
DINO	12	ResNet-50	<u>49.0</u>	66.6	<u>53.5</u>	32.0	<u>52.3</u>	<u>63.0</u>
Focus DETR	12	ResNet-50	48.8	<u>66.8</u>	52.8	31.7	52.1	63.0
Salience DETR	12	ResNet-50	**49.2**	**67.1**	**53.8**	<u>32.7</u>	**53.0**	**63.1**
CSTD (Ours)	12	Swin-T	47.8	63.1	51.2	**35.9**	52.1	58.5

Note: The best results are highlighted in bold, and the second-best results are underlined.

As shown in the results, Salience DETR achieves the best overall performance on the COCO dataset, with an AP of 49.2%. The proposed CSTD model yields a slightly lower overall AP, but still ranks among the top-performing methods, demonstrating its strong generalization capability.

Notably, CSTD achieves the best performance on small-object detection, reaching the highest AP_S, which improves by 3.2% over the second-best method, Salience DETR. This indicates that CSTD possesses stronger representation and detection capability for small objects, making it particularly suitable for small-object-dense scenarios such as transmission line inspection.

However, the overall performance of CSTD on the COCO dataset still leaves room for improvement. The performance degradation on medium and large objects may be attributed to the limited model capacity or its design bias toward small-object modeling, which restricts its ability to represent large-scale objects. Future work will focus on enhancing multi-scale feature modeling to further improve the general detection performance and adaptability of the proposed model.

6 Discussion and Conclusion

Small-object defect detection in transmission line inspection remains a challenging problem due to weak feature representation, scale variation, and complex backgrounds. To address this issue, we propose a Cost-Sensitive Transformer-based Detector (CSTD) with the hybrid-domain attention decoupled head and a cost-sensitive optimization strategy. Extensive experiments on a self-built dataset and the COCO benchmark demonstrate that the proposed method achieves consistently superior performance in small-object detection, particularly in terms of AP_S. These results indicate that the proposed framework effectively enhances fine-grained feature representation for small-scale defects in real-world inspection scenarios. However, failure cases still occur for extremely small and low-contrast defects, where limited visual information and similarity with dark backgrounds weaken discriminative feature extraction. Future work will focus on developing more effective fine-grained feature enhancement and contextual modeling strategies to improve robustness under complex inspection conditions. Overall, this work provides an effective solution for small-object defect detection in transmission line inspection and offers insights for multi-scale object detection in complex real-world scenarios.

Acknowledgement: Not applicable.

Funding Statement: This research was supported by the Fundamental Research Funds for the Central Universities (Grant No. 2026MS183), the Open Fund of the Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education (Grant No. ESIC202503), and the National Natural Science Foundation of China (Grant No. 62371188). The authors gratefully acknowledge the financial support from these funding agencies.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, methodology, supervision, data collection, dataset construction, model design and implementation, Congcong Ma and Jiaqi Mi; experimental validation and result analysis, Feifei Fu; visualization, Wenqing Zhao; writing—original draft preparation, Congcong Ma; writing—review and editing, Jiaqi Mi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The COCO dataset used in this study is publicly available and can be accessed from the official COCO dataset website. The transmission line inspection dataset is not publicly available due to data privacy and security restrictions associated with industrial inspection scenarios. The data were collected and annotated for research purposes and are not permitted for public release. Reasonable requests for further information may be directed to the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nikouei M, Baroutian B, Nabavi S, Taraghi F, Aghaei A, Sajedi A, et al. Small object detection: a comprehensive survey on challenges, techniques and real-world applications. Intell Syst Appl. 2025;27(1):200561. doi:10.1016/j.iswa.2025.200561.
2. Mukundan A, Karmakar R, Gupta D, Wang HC. Deep learning-based toolkit inspection: object detection and segmentation in assembly lines. Comput Mater Contin. 2026;86(1):1–23. doi:10.32604/cmc.2025.069646.
3. Yuan X, Chakravarty A, Lichtenberg EM, Gu L, Wei Z, Chen T. An empirical analysis of deep learning methods for small object detection from satellite imagery. Expert Syst Appl. 2026;307(11):131061. doi:10.1016/j.eswa.2025.131061.
4. Wang G, Hao Z, Cao W, Mei H. YOLO-EPDS: a small object detection algorithm for power transmission line nut spacing looseness. IET Image Process. 2026;20(1):e70279. doi:10.1049/ipr2.70279.
5. He W, Zhang Y, Xu T, An T, Liang Y, Zhang B. Object detection for medical image analysis: insights from the RT-DETR model. In: Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence; 2025 Feb 14–16; Kuala Lumpur, Malaysia. p. 415–20. doi:10.1145/3730436.3730506.
6. Song J, Zhou M, Luo J, Pu H, Feng Y, Wei X, et al. Boundary-aware feature fusion with dual-stream attention for remote sensing small object detection. IEEE Trans Geosci Remote Sens. 2025;63:5600213. doi:10.1109/TGRS.2024.3514376.
7. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. p. 580–7. doi:10.1109/CVPR.2014.81.
8. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
9. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/iccv.2017.322.
10. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 6154–62. doi:10.1109/CVPR.2018.00644.

11. Beal J, Kim E, Tzeng E, Park DH, Zhai A, Kislyuk D. Toward transformer-based object detection. arXiv:2012.09958. 2020.
12. Zhang H, Chang H, Ma B, Wang N, Chen X. Dynamic R-CNN: towards high quality object detection via dynamic training. In: Computer vision–ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 260–75. doi:10.1007/978-3-030-58555-6_16.
13. Sun P, Zhang R, Jiang Y, Kong T, Xu C, Zhan W, et al. Sparse R-CNN: end-to-end object detection with learnable proposals. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 14449–58. doi:10.1109/cvpr46437.2021.01422.
14. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2999–3007. doi:10.1109/ICCV.2017.324.
15. Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018.
16. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Computer vision–ECCV 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
17. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9626–35. doi:10.1109/iccv.2019.00972.
18. Zheng D, Dong W, Hu H, Chen X, Wang Y. Less is more: focus attention for efficient DETR. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 6651–60. doi:10.1109/ICCV51070.2023.00614.
19. Hou X, Liu M, Zhang S, Wei P, Chen B. Salience DETR: enhancing detection transformer with hierarchical salience filtering refinement. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 17574–83. doi:10.1109/CVPR52733.2024.01664.
20. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer vision–ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 213–29. doi:10.1007/978-3-030-58452-8_13.
21. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
22. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, et al. *DINO*: DETR with improved DeNoising anchor boxes for end-to-end object detection. arXiv:2203.03605. 2022.
23. Tan M, Yuan X, Liang B, Han S. DRFnet: dynamic receptive field network for object detection and image recognition. Front Neurorobot. 2023;16:1100697. doi:10.3389/fnbot.2022.1100697.
24. Miri Rekavandi A, Rashidi S, Boussaid F, Hoefs S, Akbas E, Bennamoun M. Transformers in small object detection: a benchmark and survey of state-of-the-art. ACM Comput Surv. 2026;58(3):1–33. doi:10.1145/3758090.
25. Tian C, Cheng T, Ma Y, Zhu Q, Zhang B, Zhang D. A heterogeneous CNN with transformers for image denoising. IEEE Trans Consum Electron. 2026;72(2):2658–67. doi:10.1109/TCE.2026.3665312.
26. Xu ZY, Tang X. Transmission line insulator defect detection algorithm based on MAP-YOLOv8. Sci Rep. 2025;15(1):10288. doi:10.1038/s41598-025-92445-3.
27. Yang X, Liu K, Guan R, Yu J, Liu Y. Transformer-based detection of hidden hazards in power transmission lines for UAV inspection. IEEE Access. 2025;13(1):150992–1000. doi:10.1109/ACCESS.2025.3603163.
28. More SS, Bansode R. FCN-YOLOS: an effective deep-learning model for real-time object detection. J Field Robot. 2025;42(8):4053–74. doi:10.1002/rob.70001.
29. More SS, Bansode R, Upadhyay GM, Akashe S. Optimized duo directed LSTM for efficient object detection on resource constrained edge devices. Discov Internet Things. 2026;6(1):66. doi:10.1007/s43926-026-00318-6.
30. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/CVPR52729.2023.00721.

31. Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. Adv Neural Inf Process Syst. 2020;33:21002–21012.
32. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 2020;34(7):12993–3000. doi:10.1609/aaai.v34i07.6999.
33. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer vision–ECCV 2014. Cham, Switzerland: Springer International Publishing; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.