Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.058714

ARTICLE

GD-YOLO: A Network with Gather and Distribution Mechanism for Infrared Image Detection of Electrical Equipment

Junpeng Wu[1,2,*] **and Xingfan Jiang**[2]

[1]Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education, Northeast Electric Power University, Jilin, 132012, China

[2]School of Electrical Engineering, Northeast Electric Power University, Jilin, 132012, China

*Corresponding Author: Junpeng Wu. Email: junpengwu80@163.com

Received: 19 September 2024; Accepted: 16 January 2025; Published: 26 March 2025

ABSTRACT: As technologies related to power equipment fault diagnosis and infrared temperature measurement continue to advance, the classification and identification of infrared temperature measurement images have become crucial in effective intelligent fault diagnosis of various electrical equipment. In response to the increasing demand for sufficient feature fusion in current real-time detection and low detection accuracy in existing networks for Substation fault diagnosis, we introduce an innovative method known as Gather and Distribution Mechanism-You Only Look Once (GD-YOLO). Firstly, a partial convolution group is designed based on different convolution kernels. We combine the partial convolution group with deep convolution to propose a new Grouped Channel-wise Spatial Convolution (GCSConv) that compensates for the information loss caused by spatial channel convolution. Secondly, the Gather and Distribute Mechanism, which addresses the fusion problem of different dimensional features, has been implemented by aligning and sharing information through aggregation and distribution mechanisms. Thirdly, considering the limitations in current bounding box regression and the imbalance between complex and simple samples, Maximum Possible Distance Intersection over Union (MPDIoU) and Adaptive SlideLoss is incorporated into the loss function, allowing samples near the Intersection over Union (IoU) to receive more attention through the dynamic variation of the mean Intersection over Union. The GD-YOLO algorithm can surpass YOLOv5, YOLOv7, and YOLOv8 in infrared image detection for electrical equipment, achieving a mean Average Precision (mAP) of 88.9%, with accuracy improvements of 3.7%, 4.3%, and 3.1%, respectively. Additionally, the model delivers a frame rate of 48 FPS, which aligns with the precision and velocity criteria necessary for the detection of infrared images in power equipment.

KEYWORDS: Infrared image detection; aggregation and distribution mechanism; sample imbalance strategy; lightweight structure

1 Introduction

With the ongoing development and construction of substations, infrared imaging technology has become increasingly prevalent in the inspection of electrical equipment. This technology not only enhances the detection rate, efficiency, and reliability of identifying equipment anomalies but also mitigates potential risks. In contrast to visible light images, thermal images are more resilient to the impact of external weather conditions and fluctuations facilitating the timely collection of information about the electrical equipment's status and effectively preventing faults caused by overheating [1].

Recently, due to the ongoing development of artificial intelligence and information technology, deep learning-based methods for infrared image recognition and processing have successfully addressed various

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

challenges associated with traditional manual detection techniques. Li et al. [2] employed two parallel feature encoders in extracting both RGB and infrared image features, and then utilized a multi-modal feature fusion method in fusing the shallow feature output produced by the two encoders. Wang et al. [3] trained the Mask R-CNN algorithm by using migration learning and a dynamic learning rate for the improvement in infrared images of substation insulators. Yan et al. [4] firstly fused the visible and infrared images of substation equipment, trained the fused image dataset by the Mask R-CNN algorithm for the detection accuracy, and extracted semantic features of targets in dense target scenes while maintaining a streamlined model. Various solutions have previously been proposed for different application scenarios. To address this issue, researchers have proposed various solutions that have demonstrated their effectiveness in specific environments and conditions. Combined with the YOLOv5s model, Chen et al. [5] developed an efficient algorithm in detecting floating waste on aquatic surfaces, presenting its swift real-time capability for monitoring and water pollution management. For the precision of service robots in recognizing elevator buttons, Tang et al. [6] proposed the YOLOv5 algorithm in combination with robotic technology, and showed the improved efficiency and reliability of robots in providing assistance services. Wang et al. [7] utilized residual networks to enhance the model's detection on traffic signs in complex backgrounds, providing more reliable traffic information and road safety for drivers and autonomous vehicles. Ali et al. [8] integrated the Kalman filter to enhance the stability of the YOLO model when processing noisy data and forecasting the movement paths of vehicles, consequently improving the detection and tracking performance of vehicles in changing environments.

To enhance compatibility with the detection of electrical equipment in infrared imagery, a GD-YOLO model is proposed in this paper. Firstly, the Feature Pyramid Network (FPN) has been augmented with a Gather and Distribute mechanism, which encourages the enhanced to prioritize the effectiveness of feature consolidation. Secondly, the Grouped Channel-wise Spatial Convolutional with Cross Stage Partial Network (GCSCSP) architecture is engineered to diminish parameter count and computational demand, thereby accelerating the inference process. Finally, Adaptive SlideLoss is used to alleviate the difference between positive and negative samples, and the model utilizes the diagonal distance to ascertain the locations of the target and prediction frames, enabling it to more effectively distinguish between various objects and thereby enhance the precision of object detection in electrical equipment imagery.

The remainder of this paper is organized as follows: Section 2 reviews the literature related to Feature Pyramid Networks (FPNs) and the application of lightweight networks. Section 3 outlines the overall structure of the GD-YOLO detection model and explains the principles of its improved modules. Section 4 describes the creation of the dataset, the setting of environmental parameters, and the selection of relevant indicators for experimental analysis. Section 5 presents the comparative experiments and visualization analysis. Finally, Section 6 summarizes the research findings of this paper and offers insights into future research directions.

2 Literature Review

Deep learning-based object detection can be divided into two primary categories: one-stage detectors and two-stage detectors. Single-stage detectors, primarily employing the YOLO series [9] and SSD [10], directly process the input image to predict the category and position of the object. The R-CNN family [11] represents the two-stage detectors, which first generate a group of candidate regions before classifying and regressing these regions. Furthermore, the diversity in object dimensions within the image might lead to a loss of fine details during the feature extraction process at a particular scale. To tackle the challenge of varying scales, target detection models typically incorporate feature pyramid architecture. The traditional FPN [12] features a top-down pathway designed to integrate multi-scale features. However, this direct merging of distinctions between different feature layers can lead to a loss of information within the image. PAFPN [13]

enhances FPN by adding a bottom-up path to compensate for the low-level feature details in the high-level features. GraphFPN [14] incorporates a graph neural network to overcome the limitation of direct interaction between adjacent scale features. BiFPN [15] achieves efficient weighted feature fusion by utilizing a jump connection and two-way channel between different scales. AFPN [16] sequentially expands and fuses two adjacent features with varying resolutions, reducing disparities between features with different scales, and preserving valuable information for fusion. Gold-YOLO [17] employs the aggregation and distribution mechanism to inject globally fused multi-scale features into a higher-level feature layer, enabling efficient information exchange. To fulfill the immediate detection demands of electrical infrastructure, a hybrid approach that integrates conventional detection methods with deep learning is employed for the analysis of infrared imagery. Lianqiao et al. [18] conducted position recognition and classification on preprocessed infrared images, employing nonlinear least squares curve fitting to calculate the maximum temperature, and proposed a YOLO-based method for detecting infrared images of electrical equipment. Li et al. [19] enhanced the CSP structure by proposing YOLO-FIRI, which incorporates an attention mechanism into the residual block, improving the network's capacity to acquire robust. Yang et al. [20] proposed a detection method for infrared images of power equipment based on YOLO. This method uses Efficient-IoU (EIoU) for feature fusion, ultimately locating, identifying, and classifying objects, which better meets the requirements of power detection. Yu et al. [21] introduced the ES-Net, a network that effectively sequences and amalgamates features across various levels via a feature steering module and incorporates a multi-sensory field module to bolster the network's acquisition of robust features. Han et al. [22] introduced an approach for identifying regions of interest (ROI) predicated on the responsiveness of thermal image hot spots, and used the MobileNet detection network for the identification and detection of power equipment in infrared imagery, while utilizing a streamlined network architecture to enhance detection velocity. To enhance detection accuracy, Du et al. [23] implemented a mechanism that prioritizes negative samples by a combination of YOLO and attention mechanisms to reduce false positives.

In practical applications, low-resolution infrared images of substation equipment cause obstacles to image positioning, making it difficult to distinguish similar equipment in multiple substations in real-time. To address these challenges, this paper proposes an inference model that integrates aggregation and distribution mechanisms. The model uses global and locally aligned aggregation modules to splice and extract features and then locates the category and position information of infrared images by the diagonal distance between the prediction frame and the target frame. The major contributions of our proposed Infrared image target detection are as follows:

A Partial Convolution Group (PCG) is proposed based on SqueezeNet and depth separable convolution. This approach utilizes partial convolution to reduce the computational load while maintaining the same convolutional form. Building upon GSConv, the GCSConv and its lightweight module GCSCSP are developed by integrating the double branch PCG and depth separable convolution.

The feature alignment component and the information dissemination component are integrated into the neck of YOLOv8 to construct Gather and Distribution-Feature Pyramid Network (GD-FPN), which enables efficient information interaction among features at different levels and further enhance the synergistic information processing ability of the neck.

To address the imbalance between complex and simple samples in infrared images, an adaptive SlideLoss is developed to enhance classification. Additionally, when the prediction box shares the same aspect ratio as the ground truth box but has significantly different width and height values, MPDIoU is introduced to tackle where IoU may not accurately reflect the discrepancy. This new metric streamlines the calculation process and maintains the accuracy of bounding box regression.

3 Method

Given the low resolution and complex background of the infrared image, the YOLOv8 model remains inadequate in ensuring the accuracy of recognition. In this paper, the GD-FPN technique is employed in the neck to address the issue of information loss between feature fusion of different layers in the original FPN. Additionally, GCSCSP is utilized to replace the original C2f in order to reduce the number of parameters and computational load. Furthermore, the dual channel partial convolution DWConv is integrated with convolution to enhance the spatial correlation information of feature fusion. Moreover, Adaptive SlideLoss is adopted to mitigate the sample imbalance of infrared images, while MPDIoU is introduced to improve model convergence speed and detection performance. The GD-YOLO model is depicted in Fig. 1.

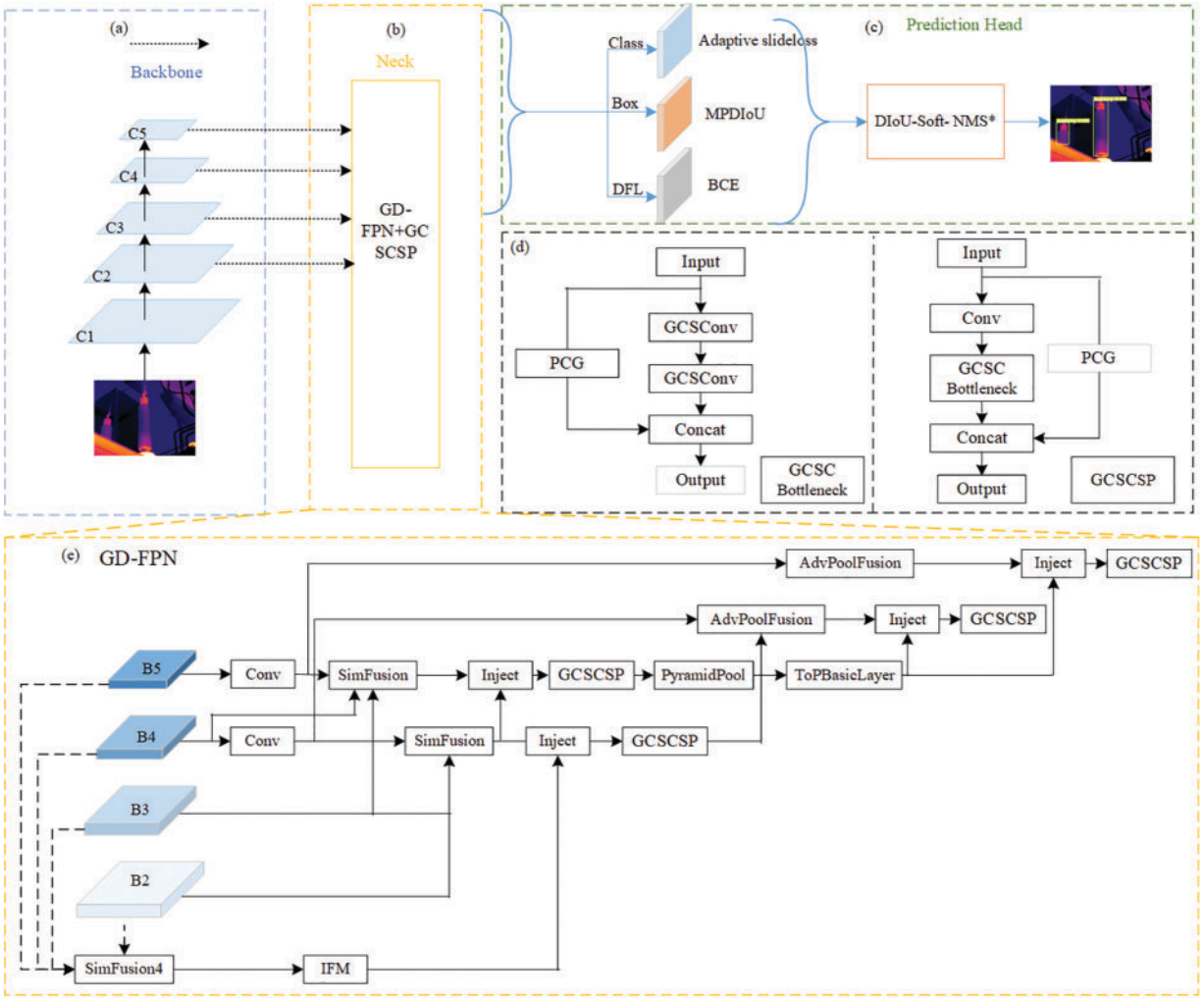

Figure 1: The GD-YOLO model is composed of three parts: (a) the backbone network, (b) the feature fusion network, and (c) the head network. In the feature fusion network, (d) is the constructed lightweight GC module, and (e) is the GD-FPN network, which enhances multi-scale feature fusion by adopting the gather and distribute mechanism

3.1 Network Based on Dual-Channel Spatial Correlation Information

In the context of infrared image detection, ensuring accuracy while maintaining low computational complexity has consistently been a focal point of interest. This objective is primarily accomplished through

the utilization of deep separable convolution operations, which serve to diminish the quantity of floating-point computations. Nevertheless, DWConv also exhibits certain limitations: the separation of channel information from the input image during the computation process results in the loss of some inter-channel related information. As suggested by GhostNet [24], it is argued that not all feature maps are necessary to be derived from convolutional operations. From the above, a partial convolution group module based on SqueezeNet and DWConv. The detailed computation procedure is depicted in Fig. 2.

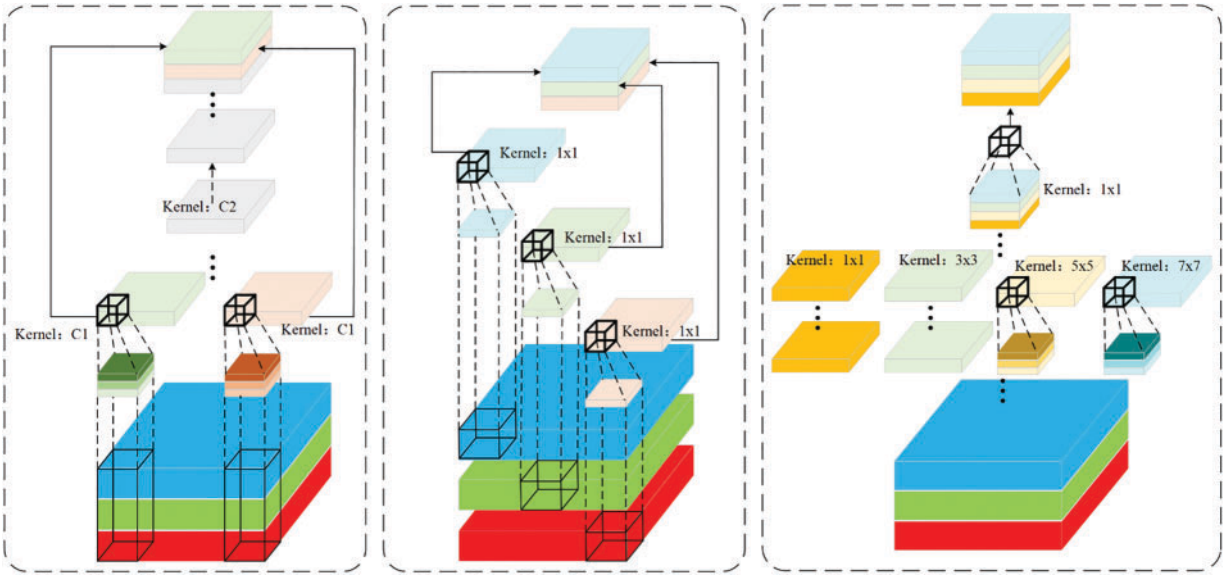

Figure 2: Schematic diagram of three types of convolutions

PCG incorporates the concept of grouping convolution from squeeze net, utilizing various convolution kernels and DWConv point-by-point convolution. Then, perform 1 × 1 pointwise convolution for the feature map without convolution operation and the feature map obtained through convolution operation to exchange information between channels. The concept of partial convolution is adopted to reduce the parameters of convolution. Finally, inspired by GSConv, and since PCG has the ability to capture different spatial features, it can be integrated into GSConv as a receptive field enhancement module. Through the collaboration of dual branches and DWConv, the limitation of the Depthwise Convolution stage of DWConv is alleviated, that is, it only emphasizes the feature changes in the channel dimension while ignoring the spatial position information. By integrating PCG, DWC, and Shuffle technologies, we introduce a dual-channel network module named GCSConv, aiming to improve the spatial correlation information. The feature map of GCSConv exhibits a greater number of contour features compared to those of DWConv and SqueezeNet. The GCSConv layer is utilized to replace the Conv layer in the backbone from a lightweight perspective. To fully utilize CNN features, the original C2f module is optimized in the neck. After studying DenseNet, VoVNet, and CSPNet, the streamlined structure, GCSCSP, for the detection of infrared images leveraging GCSConv is shown in Fig. 3.

3.2 Enhancement of Feature Fusion in GD-FPN

Currently, the conventional pyramid network still suffers from information loss during amalgamating features across disparate scales. The method of Gold-YOLO based on global information fusion, is incorporated into the architecture of YOLOv8 to create GD-FPN. As shown in Fig. 4, the GD-FPN structure is displayed. The network includes the feature alignment module SimFusion, the information fusion module

(IFM), and the information injection module Inject. The system employs the mechanism of aggregation and distribution and utilizes SimFusion to gather and merge information from multiple layers. This information is then distributed to various layers through the IFM and ultimately injected into different detection heads through the process of Injection.

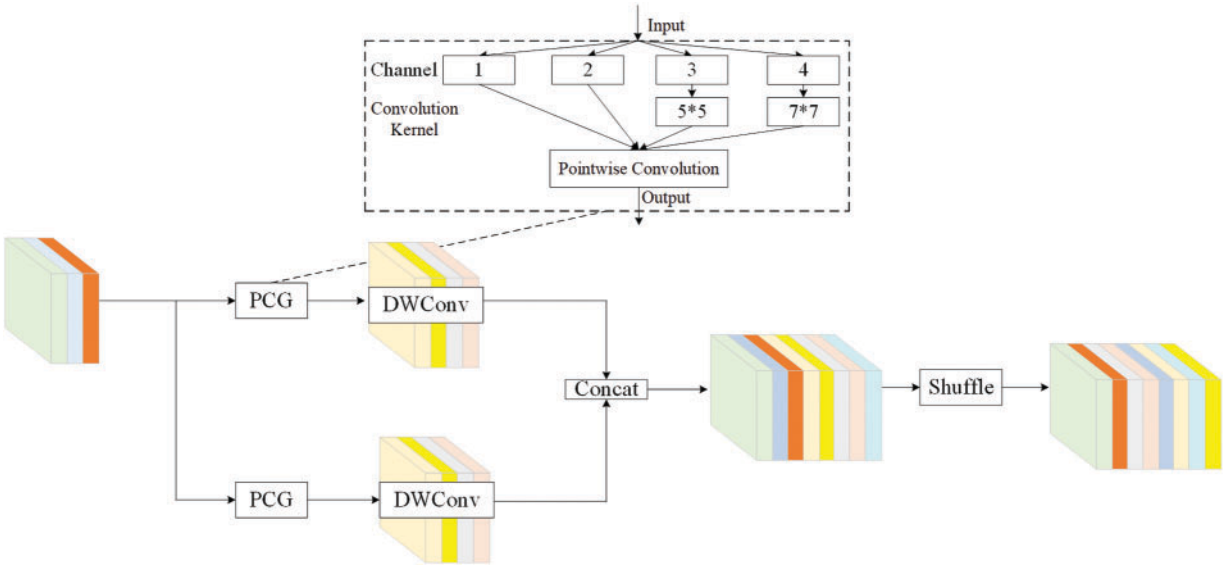

Figure 3: The network module of GCSConv

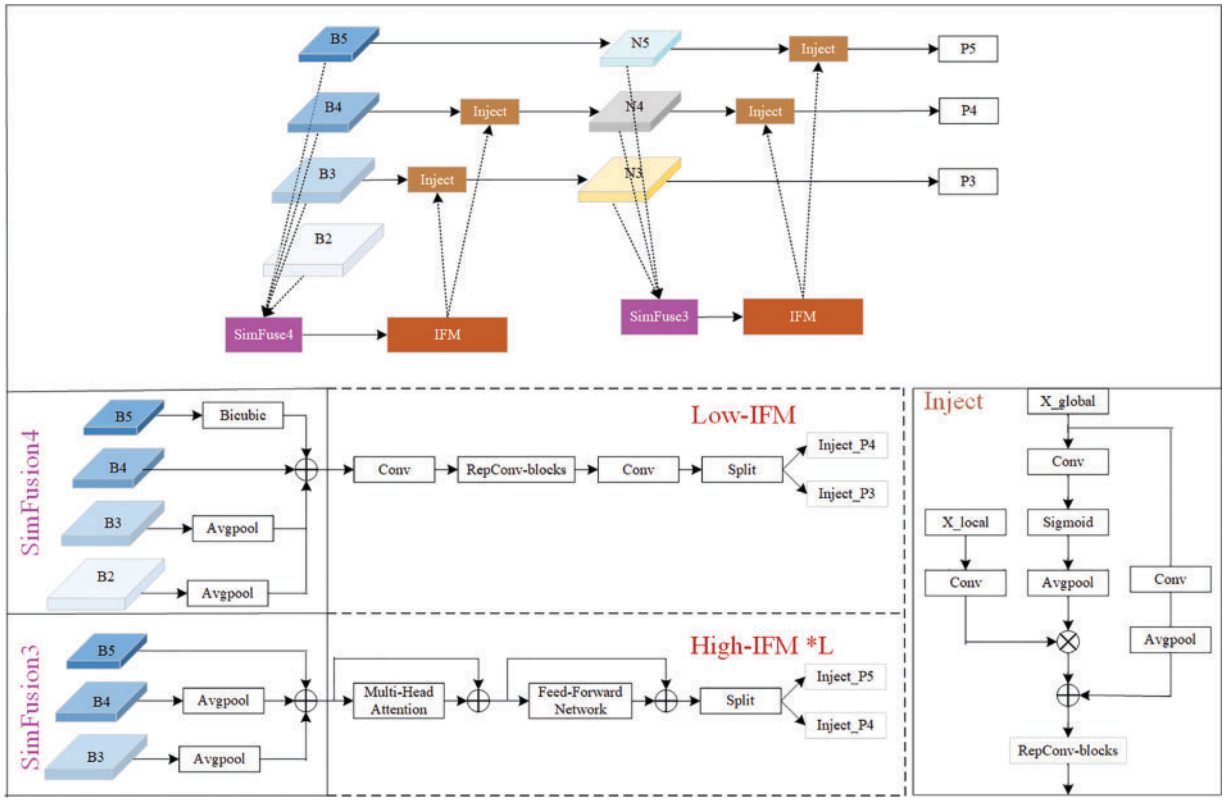

Figure 4: The network structure of the GD-FPN

In the Neck module, the outputs of B2, B3, B4, and B5 are initially combined by using SimFusion4, and the alignment of low-level features is achieved by utilizing B4 to facilitate effective information aggregation while preserving a greater amount of low-level information. By employing multi-layer re-parameterized convolution blocks with RepBlock, the transfer of information between different layers is facilitated, leading to an increased network capacity without significantly adding parameters or computational burden. This achieves the aggregation of low-order information of Low-IFM. Selective pooling operations are used to perform global feature down-sampling, and the local features and global features generated by IFM are cascaded across various tiers of feature layers. The alignment of the overall features, the fusion of Rep Block information, and the formula of information injection are as follows:

Finally, to more effectively capture contextual information and extract essential details, High-IFM employs a Transformer structure for high-dimensional mapping and complex transformation. This approach enhances the model's expressive capabilities and further improves its performance. The formulas for local feature alignment, transformer feature fusion, and information injection are as follows:

$$F_a1 = simFusion3\left(N_3 + N_4 + N_5\right) \tag{1}$$

$$F_b1 = RepBlock\left(F_a1\right) \tag{2}$$

$$F_{injec}P_5, F_{injec}P_4 = Split\left(F_b1\right) \tag{3}$$

$$F_a = simFusion4\left(B_2 + B_3 + B_4 + B_5\right) \tag{4}$$

$$F_b = RepBlock\left(F_a\right) \tag{5}$$

$$F_{injec}P_3, F_{injec}P_4 = Split\left(F_b\right) \tag{6}$$

3.3 Loss Function

Within the realm of object detection, samples are categorized as straightforward or intricate based on the IoU measure of the predicted bounding box and the actual bounding box. The simple sample's target box may slightly overlap with the real box due to optimization, suggesting that the sample is easily accommodated. In contrast, the difficult sample exhibits a low degree of overlap with the real box, resulting in a consistently low IoU even after extensive training. In the context of infrared image detection, there is a high prevalence of easy samples and a relatively low occurrence of difficult samples, leading to potential issues related to sample imbalance. YOLOv8 employs the Task Alignment Learning (TAL) to ascertain the assignment and pairing of positive and negative samples. The method is highly sensitive to the choice of IoU threshold, as it determines the allocation of positive and negative samples. This situation can lead to the determined target category being affected by excessive negative samples during the training process, which may impact the model's ability to effectively learn from these categories and subsequently hinder its capacity to handle challenging targets. To address the disparity between complex and simple samples in infrared images, SlideLoss is introduced within scope of this research. SlideLoss utilizes the average value of IoU of all bounding boxes as the threshold μ, and categorizes the samples into positive and negative samples based on μ. The sample of the boundary is highlighted by the slide weighting function, as shown in Fig. 5. The enhanced Adaptive SlideLoss is outlined below. The weighting function for the Slide can be expressed as follows:

$$y = \begin{cases} 1, x \le \mu - 0.1 \\ e^{\alpha(1-\mu)}, \mu - 0.1 \le x \le \mu \\ e^{\alpha(1-x)}, x \ge \mu \end{cases} \tag{7}$$

where μ represents the mean value of IoU, and α denotes the decay rate. When $\alpha = 0.96$, the optimal result can be obtained. The adaptive SlideLoss method involves the adaptive learning of the threshold parameter

μ for positive and negative samples. Placing higher weights in the vicinity of μ will amplify the relative loss associated with the classification of challenging samples, thereby directing greater emphasis toward the classification of such samples. This methodology seeks to boost the precision and reliability of the object detection model by reducing the direct disparity between the estimated box and the ground-truth box.

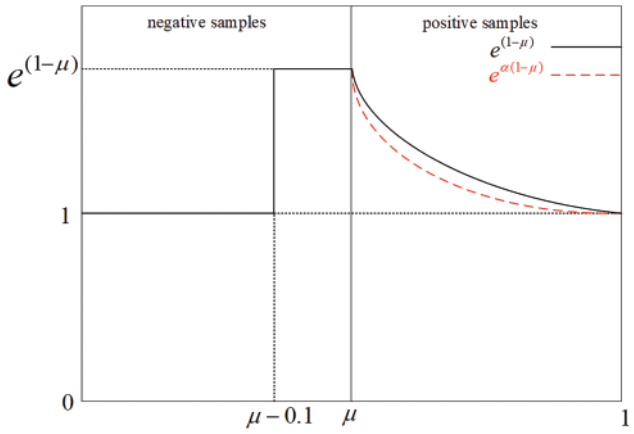

Figure 5: Adaptive SlideLoss

3.3.1 MPDIoU

The bounding box regression loss is calculated using a combination of Distribution Focal Loss (DFL) and Complete Intersection over Union (CIoU). CIoU considers both the distance between the center point of the prediction box and the real box and the aspect ratio. The formulation of CIoU can be written as follows:

$$L_{CIoU} = 1 - CIoU \tag{8}$$

$$CIoU = IoU - \left(\frac{d^2\left(b, b^{gt}\right)}{c^2} + av\right) \tag{9}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{10}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{11}$$

The formula involves the coordinate parameters of the center of the predicted box, the coordinate parameters of the center of the actual box, and the weight coefficient. It is used to assess the consistency of the parameters of the two rectangular boxes, specifically the width and height of the actual frame. When the height value differs significantly, it can limit the convergence speed and accuracy, rendering CIoU ineffective. An IoU loss, MPDIoU, based on the minimum point distance is utilized. It simplifies the comparison of the two bounding boxes by minimizing the diagonal distance between the predicted bounding box and the actual bounding box, allowing for adaptation to overlapping or non-overlapping bounding box regression.

4 Experiment

4.1 Datasets

The original images of the power equipment studied in this paper are derived from actual on scene electric power substation. The images in the dataset are saved in the default JPG format. The initial dataset contains 600 images, and the collected infrared images are enhanced through Mosaic-9 and Real-ESRGAN methods. The constructed dataset, comprising 5900 infrared images, is detailed in Table 1. This dataset includes infrared temperature measurement images of five types of power equipment: 220 kV lightning arresters, 220 kV current transformers, 220 kV voltage transformers, insulation bushings, and insulators. Among them, the voltage transformer, arrester, and insulation bushings are chosen as representatives of similar devices to enhance the model's universality.

Table 1: Composition of data sets

ID	Class name	Training sets	Val sets	Total
0	Arrester	800	200	1000
1	Current-transformer	960	240	1200
2	Potential-transformer	960	240	1200
3	Bushing	1000	250	1250
4	Insulator	1000	250	1250

4.2 Experimental Metrics

Common object detection algorithm metrics can be divided into two categories: detection accuracy evaluation metrics and model complexity evaluation metrics. In terms of detection accuracy, this paper selects mAP as the evaluation metric for model detection accuracy. In terms of model complexity, this paper selects the model's parameter count, computational load, and Frames Per Second (FPS) as the evaluation metrics for model complexity.

The mAP is the mean of the Average Precision (AP) across different object categories for a network model, with its calculation detailed in Formulas (15) and (14). In this context, N denotes the total number of object categories that the network model is capable of recognizing; AP signifies the average accuracy of the network model in identifying a particular category, which is calculated as the area under the curve plotting Precision (P) against Recall (R) for that category.

$$P = \frac{TP}{(TP + FP)} \tag{12}$$

$$R = \frac{TP}{(TP + FN)} \tag{13}$$

$$AP = \int_0^1 P(R)\, dR \tag{14}$$

$$mAP = \frac{\sum_{i=0}^{n} AP(i)}{N} \tag{15}$$

Precision P, also known as the positive predictive value, is used to measure the proportion of samples correctly predicted as positive among all samples predicted as positive by the model, and its calculation method is shown in Formula (12). Recall R, also known as sensitivity, is used to measure the proportion of

positive samples correctly predicted by the model out of all positive samples in the test set, and its calculation method is shown in Formula (13). In Formulas (12) and (13): *TP* represents true positives; *FP* represents false positives; *FN* represents false negatives.

4.3 Experimental Environment

The experiments are conducted via the PyTorch 1.8.1 deep learning framework, with Python 3.10 and a 64-bit Windows 10 operating system. The experimental hardware is a CPU Ryzen 5600X with 4.0 GHz, and the GPU is an NVIDIA RTX 3060Ti with 8 GB video memory. The GPU accelerators are CUDA 11.1 and CUDNN 8.1. During the training, we used an SGD optimizer at a learning rate of 0.01 and a batch size of 16, and then the model was trained for 300 epochs.

5 Experimental Analysis

5.1 Validation of GCSConv

In order to comprehensively assess the efficacy of GCSConv, a comparison of the speed and parameters of these modules with common convolution modules was proposed in Table 2.

Table 2: Comparison of convolution parameters

Name	All-Time (ms)	Mean-Time	FPS	FLOPs	Params
DWConv	27.16662	0.00906	110.430	872.415 M	1.408 K
Depth-Conv	66.931	0.02231	44.822	9.731 G	18.304 K
LightConv	45.96909	0.01532	65.261	9.731 G	18.048 K
PConv	37.91922	0.01264	79.116	5.100 G	9.472 K
Ghost-Conv	41.99972	0.01400	71.429	4.865 G	9.152 K
DCNV2	430.94253	0.14365	6.961	16.576 G	31.387 K
GSConv	85.44381	0.02848	35.111	39.762 G	75.712 K
GCSConv	73.29446	0.02443	40.931	40.837 G	73.883 K

DWConv primarily concentrated on the characteristic transformations of channel dimensions during the deep convolutional stage, neglecting the contextual interactions. GCSConv addressed the low frames per second and prolonged reaction times in the PCG by leveraging the synergistic effects of the dual-branch PCG and DWConv. Experimental outcomes demonstrated that the proposed GCSConv enhanced reaction time by 12% and accelerated reasoning time, with minimal to no increase in parameter count and computational load. By refining the spatial structure's relevant information, the model's robustness and efficacy were significantly bolstered.

5.2 Consumption of Inference Time and Computational Memory

To conduct a thorough evaluation of the GD-YOLO model, we assessed its time consumption by examining both the reasoning time and the speed of inference. Additionally, we evaluated the spatial consumption by determining the memory space and the size of the parameter set during the training phase. The time consumption was quantified through an analysis of the model's inference times across various image resolutions and on different hardware platforms. Meanwhile, the space consumption was gauged by measuring the memory usage during both the training and inference stages.

As could be seen from the time consumption in Table 3, the input image with a lower resolution had a faster detection speed. However, due to its fewer pixels, the computational amount required for feature fusion was relatively low, so it was not easy to extract the most obvious features during feature extraction. When the input image was set to the standard resolution of 640 × 640, although the model's parameters and inference time both increased, the accuracy rate was 1.6% higher compared to the resolution of 1080% and 3.1% higher compared to the resolution of 416. The detection speed could meet the real-time requirements.

Table 3: Time consumption of inference

Resolution	Inference-time (ms)	All-time (ms)	FPS	mAP50 (%)
320 × 320	2.7	5.28	189	83.7
416 × 416	4.56	8.91	112	85.6
640 × 640	11.32	23.1	48.9	88.9
1080 × 1080	31	60	16.4	87.3

As the batch size increases, the GPU memory consumption (GPU-Memory) generally increases upward. Nonetheless, compared to the baseline model, GD-YOLO has successfully trimmed its parameters by 0.3% and lessened the computational burden by 10%, which helped to mitigate memory consumption to some extent. This advantage was leveraged in resource-constrained environments by adjusting model parameters or adopting more efficient memory management strategies to reduce memory requirements. In large-scale environments, as evidenced by the data in Table 4, GD-YOLO demonstrated potential for deployment. Although GPU memory consumption increased with larger batch sizes, the relatively stable processing time and acceptable latency indicated that it could handle large volumes of data. The needs of large-scale deployment were met by utilizing high-performance GPUs.

Table 4: Evaluate the model's memory consumption

Batch	Inference-time (ms)	GPU-memory (MB)	GPU-util (%)	Computation latency (ms)	FPS
1	10.54	1895	10	105.43	29.86
2	9.08	2173	9	100.89	34.68
3	8.55	2453	13	65.77	36.84
4	7.75	2737	16	48.44	40.62
8	6.87	3897	28	24.54	45.82
16	6.44	5717	53	12.153	48.94
32	7.83	11,795	100	7.83	40.25

5.3 Ablation Experiment

To verify the effectiveness of the proposed GD-FPN pyramid network, lightweight GC module, SlideLoss, and MPDIoU, ablation experiments were conducted and presented in Table 5. This table outlines eight distinct scenarios, each showcasing a unique combination of the four strategies. While the baseline method foregoes these strategic implementations, the proposed method integrates a full suite of these advanced strategies: GD-FPN to bolster the feature pyramid network, GC to reduce computational complexity, SlideLoss to refine sliding window detection, and MPDIoU to enhance mean point detection intersection over union calculations.

In the ablation experiment, we adopted a method of sequentially adding different modules to evaluate the impact of each module on the entire GD-YOLO model. Initially, the GD-FPN pyramid network was introduced into the neck network, through which large-scale feature maps were extracted via low-dimensional branches, preserving as much feature information of small targets as possible. The aligned features were then distributed across different levels in a self-attention-like manner, further enhancing feature fusion. The high-dimensional branch down-sampled the low-dimensional fused features, better-handling targets of various sizes and improving the model's detection capabilities for objects of all sizes, leading to the most significant enhancement in mean average precision (mAP50). Compared to the base model, mAP50 increased by approximately 1.3%. Subsequently, the GC lightweight module was introduced, which extracts features using partial convolution, focusing on the parts of the features that contribute significantly to object detection. Additionally, the Depthwise Separable Convolution further reduced the number of parameters and computational load, cutting the total parameters by 10% compared to the base model. Building upon the GD-FPN framework, the SlideLoss function was added to alleviate the imbalanced infrared image samples by changing negative samples near the IoU mean to positive ones for training, thereby reducing the sensitivity of IoU and presenting an increase of 0.4% in mAP50 when adding GD-FPN alone. Based on GD and SlideLoss, a GC module was added to reduce the computational load caused by the complex structure of GD-FPN, leading to an increase of mAP50 about 0.3%. Furthermore, considering the complex background of infrared images and the overlap of different objects, traditional localization methods may lead to confusion or inaccuracy. To improve the localization accuracy, the diagonal distance of MPDIoU was adopted to more accurately measure the similarity between the target box and the predicted box, thus reducing it effectively. Finally, through the four combined effects in Table 5, the model's ability to extract multi-scale features was enhanced, mitigating the disparity between positive and negative samples, and allowing predicted boxes to fit more closely to the true boundaries of the target objects. Compared to the baseline model, mAP50 increased by 3.1%, and computational load decreased by 10%, indicating that the model effectively identifies detection tasks in infrared images of electrical equipment. The improved GD-YOLO model, as shown in Fig. 6, demonstrates substantial improvements, indicating a high degree of confidence in accurately extracting target objects under complex background interference.

Table 5: Ablation experiment

Baseline	GD	GC	Slideloss	MPDIoU	mAP50 (%)	mAP75 (%)	FPS	GFLOPs
√	×	×	×	×	85.8	55.1	50.3	27.4
√	√	×	×	×	87.1	55.4	45.8	30.3
√	×	√	×	×	86.7	55.9	66	24.1
√	√	×	√	×	87.5	55.1	46.1	30.4
√	√	√	×	×	87.8	55.8	48.0	24.8
√	√	√	√	×	88.2	56.8	48.4	24.8
√	√	√	×	√	88.3	56.2	48.5	24.8
√	√	√	√	√	88.9	57.6	48.9	24.8

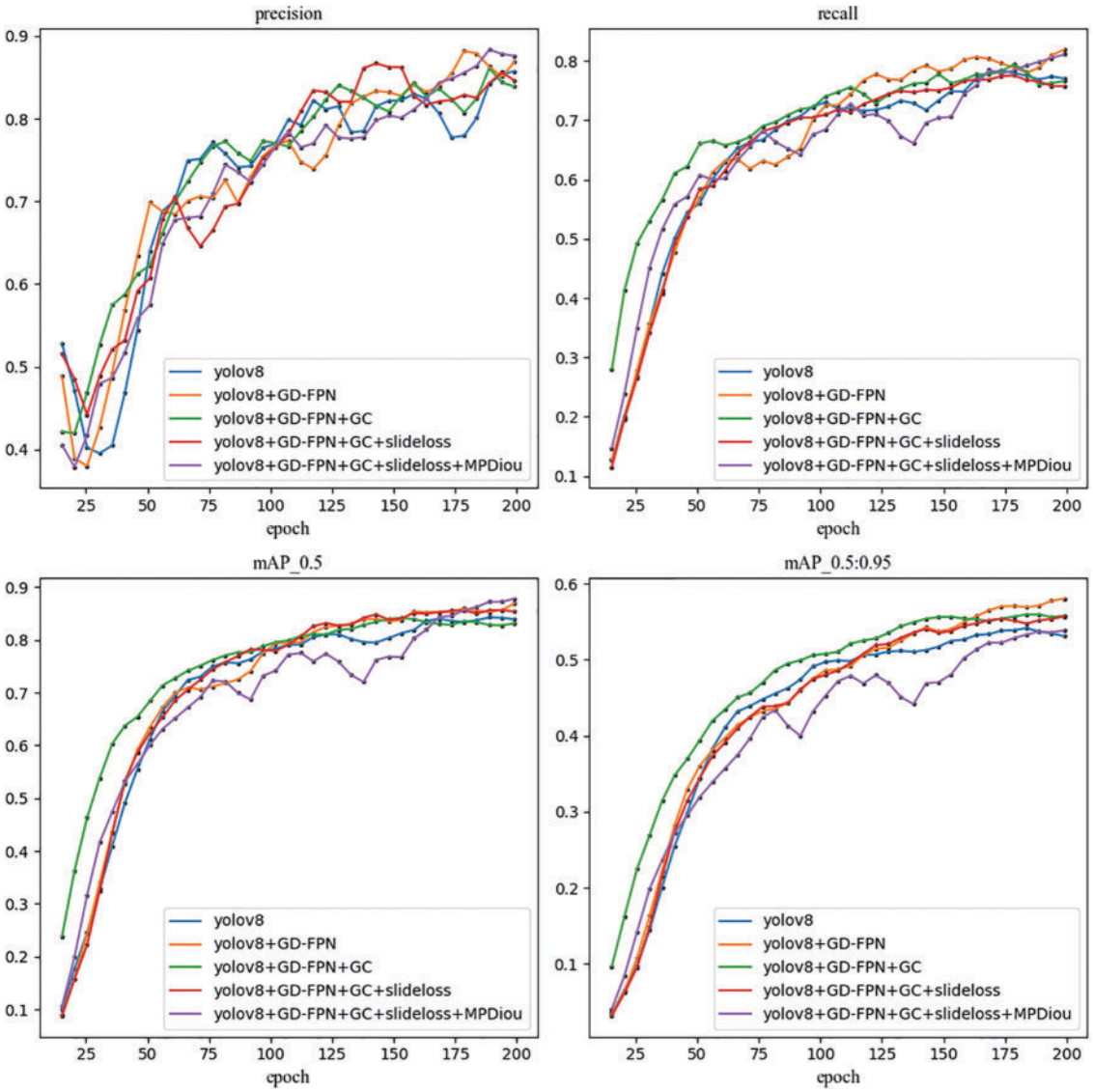

Figure 6: Evaluation index of the GD-YOLO

5.4 YOLO Series Models Comparison Experiment

The GD-YOLO model introduced in this study was benchmarked against the YOLOv5, YOLOv7, and YOLOv8 models. The focus of this paper is exclusively on evaluating these models in terms of their performance across the N, S, M, and L scales.

From Table 6, it was evident that the GD-YOLO models demonstrate exceptional capability in detecting infrared images of electrical equipment, achieving mAP of 88.9%. This performance was notably superior to its predecessors, surpassing YOLOv5 by 3.7%, YOLOv7 by 4.3%, and YOLOv8 by 3.1%. These improvements underscored GD-YOLO method can enhance the accuracy of object detection in infrared images of electrical equipment and meet real-time requirements.

Table 6: Comparison between different models

Models	Precision (%)	Recall (%)	mAP50 (%)	mAP75 (%)	Parameters (M)	GFLOPs
YOLOv8 n	85.3	83.7	84.4	54.6	2.9	8.1
YOLOv8 s	85.4	86.3	85.8	55.8	11.1	27.6
YOLOv8 m	86.9	87.5	87.2	56.0	25.3	77.9
YOLOv8 l	86.6	87.3	87.0	56.4	43.1	165.8
YOLOv7 n	85.9	84.2	85.1	54.8	1.8	4.5
YOLOv7 s	85.2	84.1	84.6	55.3	9.5	26.5
YOLOv7 m	86.0	84.5	85.3	57.2	21.9	60.3
YOLOv7 l	86.1	84.7	85.9	57.6	37.6	109
YOLOv5 n	85.1	84.5	84.5	56.7	2.0	4.3
YOLOv5 s	85.3	85.2	85.2	55.5	7.1	15.8
YOLOv5 m	86.8	86.3	86.5	56.6	20.4	49.4
YOLOv5 l	87.2	87.1	87.9	56.9	46.9	108.8
GD-YOLO	89.1	88.5	88.9	57.6	10.1	24.8

5.5 Performance Comparison

In order to validate the enhanced algorithm presented in this paper, we carried out comparative tests on the subsequent models, including GD-YOLO and YOLO models, the fault diagnosis models of RetinaNet and CenterNet, and the Improved models for detection of electrical equipment in infrared imagery. As depicted in Table 7, the GD-YOLO algorithm outperformed algorithms like YOLOv4, YOLOv8, and YOLOX in metrics such as mAP50, frames per second (FPS), and model size. Algorithms including Faster R-CNN, RetinaNet, and CenterNet all employed ResNet 50 for their base feature extraction. However, the feature maps produced by ResNet50 were of a single layer and comparatively low resolution, which meant they couldn't capture the intricate details of small objects effectively, resulting in a considerable number of detection failures. Compared to recent improvements in electrical equipment infrared image detection models such as ECA-Net, CBAF-FOCS, BASNet, FINet, and ISNet, GD-YOLO employed partial convolution to extract features, followed by Depthwise Separable Convolution to reduce the number of parameters and computational load, thereby achieving a reduction in model size and computational costs while maintaining accuracy, with a total parameter reduction of 10%. Furthermore, GD-YOLO adopted a local and global feature alignment mode, enhancing the model's neck's ability to integrate information. By using shallow and deep convergence modules along with attention-based modules to extract and fuse feature information, it enhanced the model's ability to detect objects across a range of sizes. This resulted in improvements in mAP50 of 1.4% over ECA-Net, 0.3% over CBAF-FOCS, 0.4% over BASNet, 0.9% over FINet, and 0.8% over ISNet. Additionally, it achieved a recognition speed of 48 frames per second, which was more than twice that of RetinaNet and CenterNet.

Table 7: Comparison with advanced models

Modules	mAP50 (%)	mAP75 (%)	GFLOPs	FPS
Faster-RCNN	83.2	54.1	29.1	21
YOLOv4	84.2	54.2	30.2	43
YOLOX-tiny	86.7	57.3	33.1	45

(Continued)

Table 7 (continued)

Modules	mAP50 (%)	mAP75 (%)	GFLOPs	FPS
YOLOv8	85.8	55.8	27.6	50
RetinaNet [25]	87.3	56.2	29.5	22
CenterNet [26]	86.5	57.1	28.8	24
ECA-Net [27]	87.3	56.8	29.4	47
CBAF-FOCS [28]	88.4	56.6	45.7	51
BASNet [29]	88.3	57.2	33.4	42
FINet [30]	87.8	56.8	34.2	41
ISNet [31]	87.9	55.8	29.8	44
GD-YOLO	88.7	57.9	27.6	48

5.6 Robustness Experiment

Adversarial perturbations refer to slight noise added to clean images, which was almost imperceptible but could be used and created samples by malicious-intentioned users to mislead models in fault recognition, accordingly posing a serious threat to the security of artificial intelligence in infrared image target detection in power systems.

Adversarial perturbations that may be encountered in practical applications were simulated by introducing Gaussian noise at different peak signal-to-noise ratio (PSNR) levels, aiming to verify the robustness of the GD-YOLO model against these perturbations. Specifically, Gaussian noise with PSNR values of 40, 35, 30, 25, and 20 dB was added during the model's training and testing phases to simulate adversarial attack scenarios ranging from mild to severe.

In Fig. 7, it can be concluded that at a PSNR value of 40 dB, due to the minimal addition of Gaussian noise, the training and inference datasets were very close to the original images, hence the mAP values of all models were relatively high. Nevertheless, the GD-YOLO model outperformed BASNet, FINet, and the standard YOLOv8 model in terms of accuracy. When the PSNR value was between 30 and 40 dB, YOLOv8 experienced the most significant performance decline against accumulated noise. In contrast, GD-YOLO, BASNet, and FINet, which employed deeper convolutional networks, could suppress noise interference through multi-scale feature extraction and contrast enhancement and demonstrated better robustness. When the PSNR value dropped to between 25 and 30 dB, noise interference significantly increased, and the image recognition accuracy of all models showed a downward trend. Specifically, the accuracy of YOLOv8 decreased by 12%, while the accuracy of GD-YOLO, BASNet, and FINet decreased by 7.1%, 8.3%, and 8.9%, respectively. This indicated that under higher noise interference, the deep network structure and feature extraction capabilities of GD-YOLO, BASNet, and FINet were more effective in maintaining performance. When the PSNR value further dropped to 20 dB, the GD-YOLO model, through residual connections in its GC module, helped the model learn the differences between samples, enhancing its resistance to noise. The multi-head attention mechanism in GD-FPN focused the model more on the target areas in the image, allowing it to maintain a classification accuracy of 55% even in high-interference environments. Compared to YOLOv8, the accuracy of GD-YOLO increased by 10%, by 4.5% compared to BASNet, and by 6.7% compared to FINet. These results showed that when the amplitude of disturbance is greater, the effect of the GD-YOLO model was more significant, proving its clear advantage in reducing the mAP loss caused by noise interference.

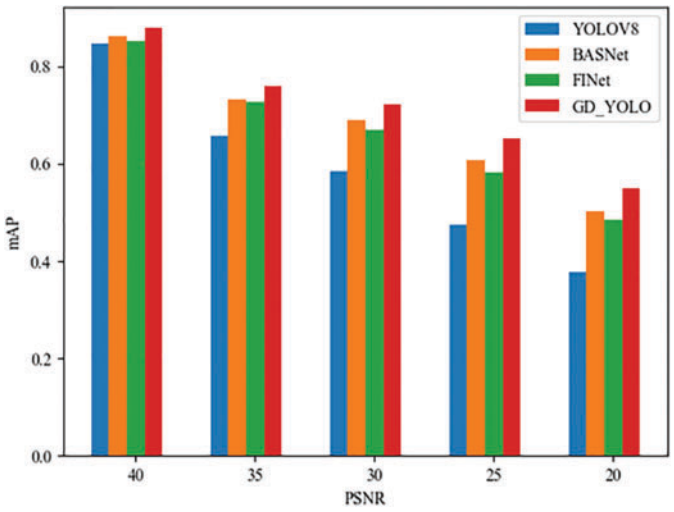

Figure 7: Comparison of accuracy of different PSNR

5.7 Detection Results

To facilitate and directly showcase the model's detection capabilities, the model's inference outcomes and heatmaps were employed for a comparative analysis of detection performance. Due to the complexity and low resolution of the background, infrared images of electrical equipment often contained a high proportion of objects that were misclassified as background, resulting in a greater possibility of missing class. In this study, GD-YOLO was used for inference experiments. Fig. 8 shows the inference results of the GD-YOLO model for several similar objects such as arrester, potential-transformer, bushing, and current-transformer, which ensures that the algorithm more comprehensively adapts to electrical equipment in infrared image scenes.

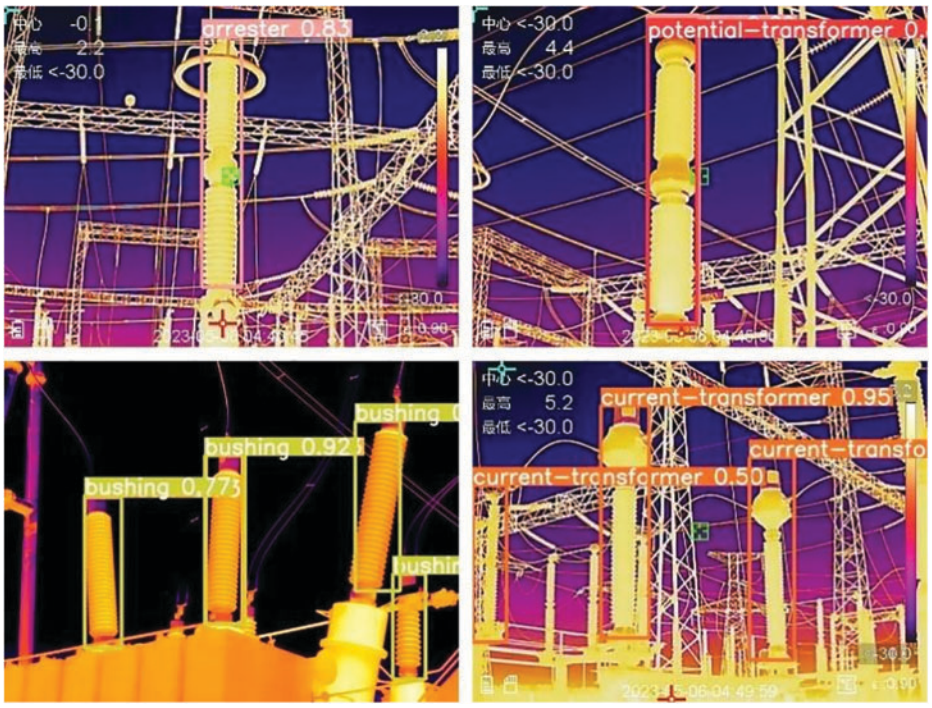

Figure 8: GD-YOLO inference results

The heatmap directly presents the regions of the feature map that capture the model's attention. The gradient values are derived from the backpropagation of the model's predicted class confidence using Gradient-weighted Class Activation Mapping (Grad-CAM). In the heatmap, pixels with higher gradients are indicated by more intense red hues, while those with lower gradients are shown with darker blue shades. The results of the experiment are illustrated in Fig. 9. As depicted in Fig. 9, YOLOv8 falls short in focusing on small objects and lacks sensitivity to distant targets. MPDIoU directs the model's attention primarily to the center of the target, enabling more precise localization of objects. This enhancement allows GD-YOLO to concentrate on targets of interest against an infrared background, thereby improving the overall detection performance of the model.

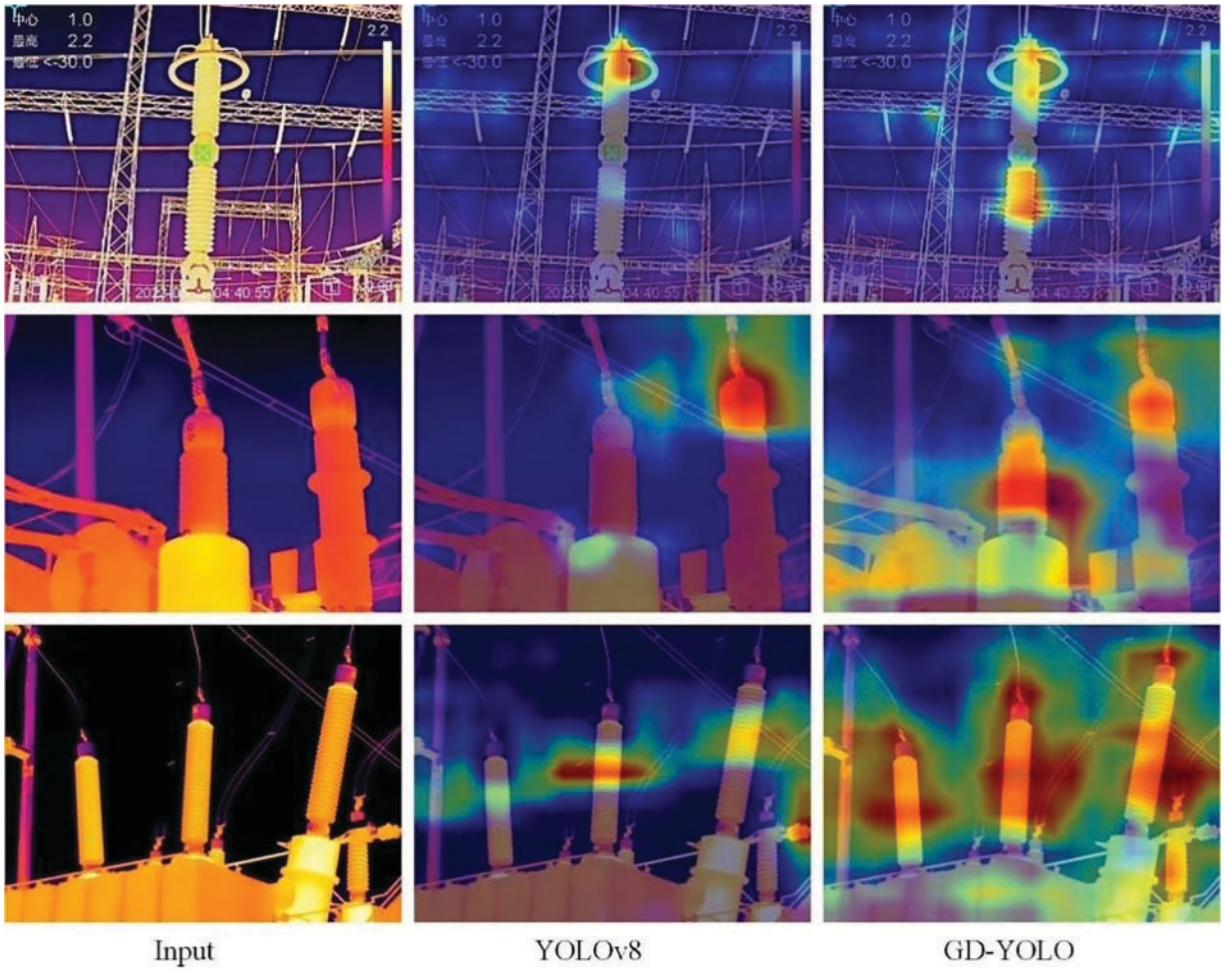

Figure 9: Visualization results of a heat map

6 Conclusion

The problem of low accuracy in infrared image detection by introducing the GD-YOLO mode is emphasized and solved in this paper. Firstly, Initially, the GD-FPN architecture is crafted to facilitate more effective communication and merging of information by holistically amalgamating features across various tiers and reintroducing the amalgamated global data into these distinct tiers. The multi-head attention mechanism focuses the model more on the target areas in the images, thus maintaining high classification accuracy. Then, a lightweight GC module is designed to replace C2f, reducing the overall number of model

parameters and computational load while maintaining model accuracy. Finally, the implemented adaptive SlideLoss function adjusts the loss dynamically in response to the intricacy of the samples, transforming nearby negative samples into positive ones for training by varying the IoU threshold, making the model focus more on challenging classifications during training and enabling it to learn more discriminative features.

Acknowledgement: We appreciate the valuable feedback provided by the reviewers.

Funding Statement: Science and Technology Department of Jilin Province (No. 20200403075SF), Education Department of Jilin Province (No. JJKH20240148KJ).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Junpeng Wu, Xingfan Jiang; data collection: Junpeng Wu; analysis and interpretation of results: Xingfan Jiang; draft manuscript preparation: Junpeng Wu, Xingfan Jiang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Fang H, Ding L, Wang L, Chang Y, Yan L, Han J. Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion. IEEE Trans Instrum Meas. 2022;71(2):1–20. doi:10.1109/TIM.2022.3198490.
2. Li B, Wang T, Zhai Y, Yuan J. RFIENet: RGB-thermal feature interactive enhancement network for semantic segmentation of insulator in backlight scenes. Measurement. 2022;205(8):112177. doi:10.1016/j.measurement.2022.112177.
3. Wang B, Dong M, Ren M, Wu Z, Guo C, Zhuang T, et al. Automatic fault diagnosis of infrared insulator images based on image instance segmentation and temperature analysis. IEEE Trans Instrum Meas. 2020;69(8):5345–55. doi:10.1109/TIM.2020.2965635.
4. Yan N, Zhou T, Gu C, Jiang A, Lu W. Bimodal-based object detection and instance segmentation models for substation equipments. In: IECON, 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society; 2020 Oct 18–21; Singapore. p. 428–34.
5. Chen Z, Huang C, Duan L, Tan B. Lightweight surface litter detection algorithm based on improved YOLOV5s. Comput Mater Contin. 2023;76(1):1085–102. doi:10.32604/cmc.2023.039451.
6. Tang X, Wang C, Su J, Taylor C. An elevator button recognition method combining YOLOV5 and OCR. CMC-Comput Mater Contin. 2023;75(1):117–31. doi:10.32604/cmc.2023.033327.
7. Wang X, Tian Y, Zheng K, Liu C. C2net-yolov5: a bidirectional res2net-based traffic sign detection algorithm. [cited 2024 Sep 10]. Available from: https://ssrn.com/abstract=4406700.
8. Ali S, Jalal A, Alatiyyah MH, Alnowaiser K, Park J. Vehicle detection and tracking in UAV imagery via YOLOV3 and kalman filter. Comput Mater Contin. 2023;76(1):1249–65. doi:10.32604/cmc.2023.038114.
9. Redmon J. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA.
10. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference; 2016 Oct 11–14; The Amsterdam, Berlin/Heidelberg, Germany: Springer; 2016. p. 21–37.

11. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. p. 580–7.
12. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 2117–25.
13. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 8759–68.
14. Zhao G, Ge W, Yu Y. Graphfpn: graph feature pyramid network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 2763–72.
15. Tan M, Pang R, Le QV. Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jue 13–9; Seattle, WA, USA. p. 10781–90.
16. Yang G, Lei J, Zhu Z, Cheng S, Feng Z, Liang R. Afpn: asymptotic feature pyramid network for object detection. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2023 Oct 1–4; Honolulu, Oahu, HI, USA. p. 2184–9.
17. Wang C, He W, Nie Y, Guo J, Liu C, Wang Y, et al. Gold-yolo: efficient object detector via gather-and-distribute mechanism. Adv Neural Inf Process Syst. 2024;36:51094–112.
18. Lianqiao L, Xiai C, Huili Z, Ling W. Recognition and application of infrared thermal image among power facilities based on yolo. In: 2019 Chinese Control And Decision Conference (CCDC); 2019 Jun 3–5; Nanchang, China. p. 5939–43.
19. Li S, Li Y, Li Y, Li M, Xu X. YOLO-FIRI: improved yolov5 for infrared image object detection. IEEE Access. 2021;9:141861–75. doi:10.1109/ACCESS.2021.3120870.
20. Yang Z, Xu Z, Wang Y. Bidirection-fusion-yolov3: an improved method for insulator defect detection using uav image. IEEE Trans Instrum Meas. 2022;71:1–8. doi:10.1109/TIM.2022.3220285.
21. Yu X, Lyu W, Zhou D, Wang C, Xu W. Es-net: efficient scale-aware network for tiny defect detection. IEEE Trans Instrum Meas. 2022;71:1–14. doi:10.1109/TIM.2022.3168897.
22. Han S, Yang F, Yang G, Gao B, Zhang N, Wang D. Electrical equipment identification in infrared images based on ROI-selected CNN method. Electr Power Syst Res. 2020;188(2):106534. doi:10.1016/j.epsr.2020.106534.
23. Du S, Zhang B, Zhang P, Xiang P, Xue H. Fa-yolo: an improved yolo model for infrared occlusion object detection under confusing background. Wirel Commun Mob Comput. 2021;2021(1):1896029. doi:10.1155/2021/1896029.
24. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C. Ghostnet: more features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 1580–9.
25. Cheng X, Yu J. Retinanet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. IEEE Trans Instrum Meas. 2020;70:1–11. doi:10.1109/TIM.2020.3040485.
26. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. Centernet: keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 02; Seoul, South Korea. p. 6569–78.
27. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 11534–42.
28. Yu J, Cheng X, Li Q. Surface defect detection of steel strips based on anchor-free network with channel attention and bidirectional feature fusion. IEEE Trans Instrum Meas. 2021;71:1–10. doi:10.1109/TIM.2021.3136183.
29. Bo W, Liu J, Fan X, Tjahjadi T, Ye Q, Fu L. Basnet: burned area segmentation network for real-time detection of damage maps in remote sensing images. IEEE Trans Geosci Remote Sens. 2022;60:1–13. doi:10.1109/TGRS.2022.3197647.
30. Zhang ZD, Zhang B, Lan ZC, Liu HC, Li DY, Pei L, et al. Finet: an insulator dataset and detection benchmark based on synthetic fog and improved yolov5. IEEE Trans Instrum Meas. 2022;71(8):1–8. doi:10.1109/TIM.2022.3194909.
31. Zhang M, Zhang R, Yang Y, Bai H, Zhang J, Guo J. Isnet: shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 877–86.