scientific reports

OPEN

A bird target detection model designed for substation scenarios

Chunxue Shao[1], Jiabin Huang[2], Xinnan Fan[3], Xiaotian Wang[1], Yuanxue Xin[3] & Pengfei Shi[4]✉

Substations are critical to the power grid, but they are often disturbed by bird activity, which can lead to power failures and outages. Conventional bird-detection methods are costly and lack long-term effectiveness. To address these issues, this paper proposes a bird target detection method, YOLO-birds, designed explicitly for substation scenarios. It utilizes the Faster-BiFPN module to optimize feature extraction and fusion by combining low-level and high-level features, thereby enhancing detection accuracy. In addition, the SPPBiF attention mechanism is introduced to address the challenge of detecting targets at different scales and small objects, such as birds. To further improve robustness, the Focal-EIoU loss function is also utilized to mitigate the effect of low-quality samples. To support this research and improve the detection performance in real-world scenarios, a self-constructed dataset of bird images focusing on security threats at substations was created. Experimental results show that YOLO-birds achieves a mAP50 of 90.2%, which validates the effectiveness of the proposed method compared to other methods. The method can efficiently detect birds inhabiting substations, which can help to differentiate the prevention of bird-caused accidents in power grids.

Keywords Substation, Deep learning, Object detection, Attention mechanisms

The safe and stable operation of the power system is the cornerstone of modern society development, and the substation, as the center of the power system, is directly related to the reliability of the entire power grid[1]. However, disturbance and damage caused by bird activities to substation equipment are becoming increasingly prominent and have become one of the key challenges facing the power industry. Birds may pose safety hazards, such as short circuiting of the equipment, contamination of the insulator, and line tripping, which not only threaten the stable operation of the power grid but may also cause significant economic losses.With the rapid development of computer vision technology, intelligent bird monitoring methods based on image recognition are becoming an effective path to prevent and control bird damage in substations. The key to bird-target detection lies in the quick and accurate identification of bird targets in images within complex scenes. This problem is highly relevant to the core task in computer vision and also presents a realistic demand for the application of deep learning in power scenarios. Therefore, the development of an efficient and accurate bird target detection method for substations is of great significance for the timely identification of potential bird threats, the implementation of preventive measures, and the enhancement of substation safety and operational efficiency.

In recent years, bird detection has emerged as a hot research topic at the intersection of computer vision and ecological monitoring, attracting significant attention. Traditional methods primarily rely on manual feature extraction and classifier design, but are limited by complex backgrounds, small bird sizes, and easy obstruction, resulting in low detection accuracy and robustness. With the development of deep learning, object detection methods based on convolutional neural networks have been introduced into the field of bird detection, significantly improving detection efficiency and accuracy. In object detection methods, mainstream models can be divided into two categories: single-stage models, such as SSD[2], YOLO[3–5], and EfficientDet[6]; and two-stage models, such as Mask R-CNN[7], Faster R-CNN[8], and Cascade R-CNN[9]. To address the challenges posed by diverse bird postures and complex environments in real-world applications, researchers have built upon existing object detection methods, incorporating multi-scale feature fusion, attention mechanisms, and lightweight network structures to enhance detection performance for target birds further.

Although existing bird target detection methods have made significant progress in various scenarios, these methods still suffer from high false positive rates when applied to substation scenarios, and it is difficult to identify small bird targets in complex substation environments accurately. Therefore, we propose a network for bird target detection in substation scenarios: YOLO-birds, aimed at improving the performance of bird target

[1]College of Information Science and Engineering, Hohai University, Changzhou 213200, China. [2]State Grid Hangzhou Power Supply Company, Hangzhou 310000, China. [3]Key Laboratory of Power Transmission Distribution Equipment Technology, Hohai University, Changzhou 213200, China. [4]College of Artificial Intelligence and Automation, Hohai University, Changzhou 213200, China. ✉email: shipf@hhu.edu.cn

detection in substations by addressing the limitations of existing algorithms. The main contributions are as follows:

- A plug-and-play Faster-BiFPN module is designed to replace the original RepNCSPELAN4, introducing a spatial pyramid pooling module to optimize feature screening and improve accuracy through multi-level feature pyramids and bidirectional information transfer. This not only ensures computational efficiency but also strengthens the model's ability to identify bird targets of different sizes.
- An attention mechanism, SPPBiF, is proposed to improve computational performance through content-aware sparsity.SPPBiF reduces computational complexity by capturing global and local features through global and local self-attention mechanisms, and at the same time enhances the expressive ability and adaptability of the model. The attention distribution is dynamically adjusted according to the content of the input image to better adapt to targets of different scales and complexity, and to improve the detection of small-target birds in complex backgrounds.
- The Focal-EIoU loss function is used instead of the original CIoU loss function to address the issue of dramatic loss value oscillation caused by low-quality samples and to provide stronger robustness for small datasets and noisy Boxes, thereby improving the network's prediction accuracy.
- A large-scale bird dataset for substations, encompassing various scenarios and lighting conditions, is constructed to provide a reliable benchmark for related research. Extensive experiments validate the effectiveness of the proposed method, which outperforms existing approaches in detection accuracy, speed, and robustness, offering a practical solution for substation bird monitoring.

Related work

Small target detection and substation bird-repellent algorithms are rapidly evolving in the field of multi-target detection and security monitoring to enhance the accuracy of small target detection and its effectiveness in complex scenarios. Research on small target detection algorithms[10] focuses on multi-scale feature extraction, feature aggregation, and attention mechanisms to improve the model's accuracy in handling small targets. In the field of substation bird target detection, researchers have explored various aspects of this topic. Zhang et al.[11] proposed a method based on an improved Faster R-CNN, which enhanced detection accuracy by introducing the attention mechanism. However, the computational complexity was high, making it challenging to satisfy real-time monitoring requirements. Li et al.[12] applied YOLOv3 to the bird detection in substations, which made a breakthrough in detection speed, but there are still difficulties in detecting small targets under complex backgrounds. Wang et al.[13] improved YOLOv5 by designing a lightweight network structure to enhance the feature extraction ability, but the generalization ability under different lighting conditions needs to be improved. In exploring another innovative approach for bird target detection, Shakeri et al.[14] proposed a background removal technique based on Gaussian mixture models. This method can quickly separate foreground targets from background elements. However, it may over-screen, resulting in the removal of specific bird targets that are not easily detectable or have a high similarity to the background, which are then misclassified as background. Bird target detection methods are shown in Table 1.

In recent years, multi-scale feature fusion methods, such as the Generalized Efficient Layer Aggregation Network (GELAN)[15] used in YOLOv9, have demonstrated significant advantages. The overall network structure of YOLOv9 is shown in Fig. 1. The Yolov9 network model is mainly composed of BackBone, Neck, and Head. In the backbone network, RepNCSPELAN4 is the core structure of the Yolov9 backbone network, and the CSP[16] module and ELAN[17] module are also introduced for feature extraction. Additionally, YOLOv9's backbone network employs residual blocks to construct a deeper network, thereby mitigating the gradient vanishing problem. In the neck layer, the network utilizes Feature Pyramid Networks (FPN)[14] to convolve further high-dimensional features, which are upsampled and fused with the shallow features to ensure the network's detection accuracy for targets at different scales. The decoding prediction side of YOLOv9 mainly consists of a prediction layer, classification, and regression. The core idea of YOLOv9 is to simultaneously predict the class, location, and confidence of a target on different scales of feature maps through a single forward propagation. Deep learning has made significant progress in several areas, and the main research directions include system architectures, learning methods, and more general objective functions. However, most research ignores the problem of information loss that may occur in the data during the feed-forward process, a problem that can lead to biased gradient flow during model updating and ultimately affect the model's prediction accuracy. To address this problem, the YOLOv9 algorithm designs a new lightweight network architecture based on gradient path planning, specifically the Generalised Efficient Layer Aggregation Network (GELAN)[15], in the design of the network architecture. The GELAN optimizes the feature representation using the CSPNet block and RepConv. Reduces the gradient loss in the information flow, making it a good balance of accuracy and speed for small

Method	Advantages	Limitations
Improved Faster R-CNN[11]	High detection accuracy	High computational complexity
Improved YOLOv3[12]	High detection speed	Low accuracy in detecting small targets in complex backgrounds
Improved YOLOv5[13]	Strong feature extraction capabilities	Low generalisation capabilities
Getting the bugs out of AI[14]	Separating foreground objects and background elements	High false positive rate

Table 1. Comparison of bird detection methods.

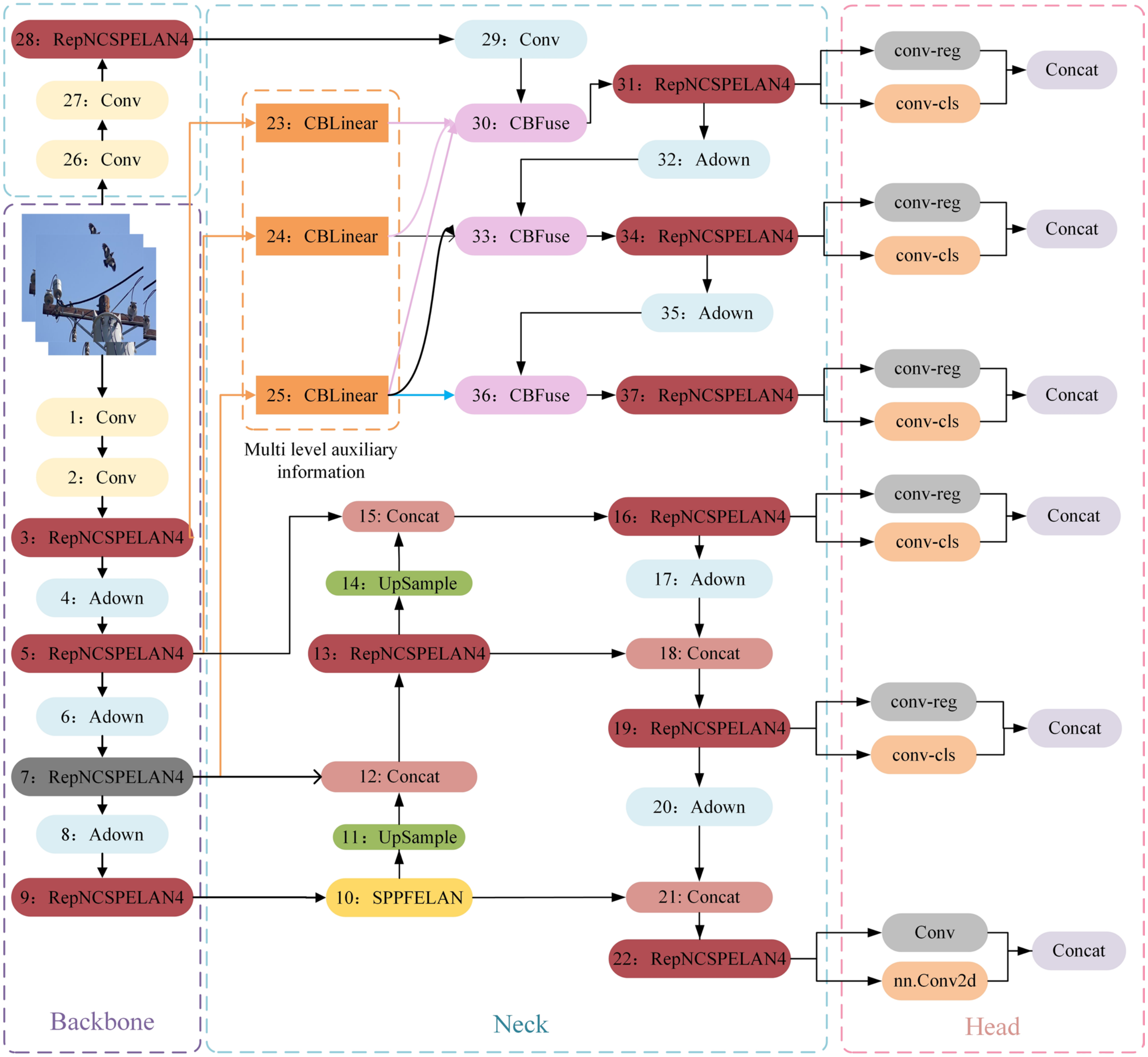


Fig. 1. YOLOv9 target detection network framework.

target detection. In addition, Anchor-Free models such as FCOS and CenterNet avoid positional bias by directly regressing to the center point of the target. They are suitable for detecting small targets at different scales[18]. YOLOv9's GELAN architecture and Programmable Gradient Information (PGI)[15] enable detection in substation environments with higher robustness and speed through accurate gradient updates, resulting in real-time performance. However, YOLOv9 also suffers from some adaptability issues that make it challenging to meet the needs of specific scenarios. The substation environment presents a variety of disturbing factors, including dense power equipment and significant variations in lighting. These factors may weaken the robustness of YOLOv9 in distinguishing backgrounds, especially in complex backgrounds such as substations, and there is still room for improvement in the performance of small object detection.

Methods

To adapt to the specific requirements of bird detection in substations, YOLOv9 is enhanced, and a deep learning-based bird target detection method, YOLO-birds, is proposed to improve the model's detection performance further. To optimize the feature screening process and enhance the model's ability to detect small targets in complex backgrounds, we adopt the Faster-BiFPN network structure as a replacement for the original backbone feature extraction network in YOLOv9. On this basis, we further introduce the SPPBiF attention mechanism into the feature fusion network. This innovation enables the model to dynamically adjust the attention distribution according to the content of the input image, thus capturing target features more accurately. Considering that target detection algorithms are often affected by factors such as lighting conditions and background complexity, we optimize the loss function. Specifically, we replace the original CIoU loss function with the Focal-EIoU loss function. This improvement effectively addresses the issue of dramatic oscillations in loss values caused by low-quality samples while also providing enhanced robustness for small-scale datasets and scenes with noise. These comprehensive optimization measures not only improve the model's detection accuracy but also enhance its adaptability in various complex environments. The structure of the improved model, YOLO-birds, is shown in Fig. 2.

Faster-BiFPN network

To optimize computational efficiency, enhance spatial feature extraction capabilities, and improve the network's ability to perceive critical information from small targets in complex backgrounds, we introduced innovative modifications to the core modules of FasterNet. Specifically, we designed a lightweight and efficient Weighted Bidirectional Feature Pyramid Network (BiFPN)[19] to construct a plug-and-play Faster-BiFPN network. This structure replaces the ResNCSPELAN4 network in YOLOv9, aiming to simplify the architecture while significantly reducing computational redundancy and memory access overhead. These improvements preserve channel-wise information and achieve more efficient feature extraction.As illustrated in Fig. 3, the Faster-BiFPN network comprises a Partial Convolution (PConv)[20], two Point-Wise Convolutions (PWConv), and the BiFPN feature processing mechanism.

To optimize computational efficiency, enhance spatial feature extraction capabilities, and increase the network's ability to perceive key information about small targets in complex contexts, this study introduces innovative improvements to the core module of FasterNet. We introduce a lightweight and efficient weighted bidirectional

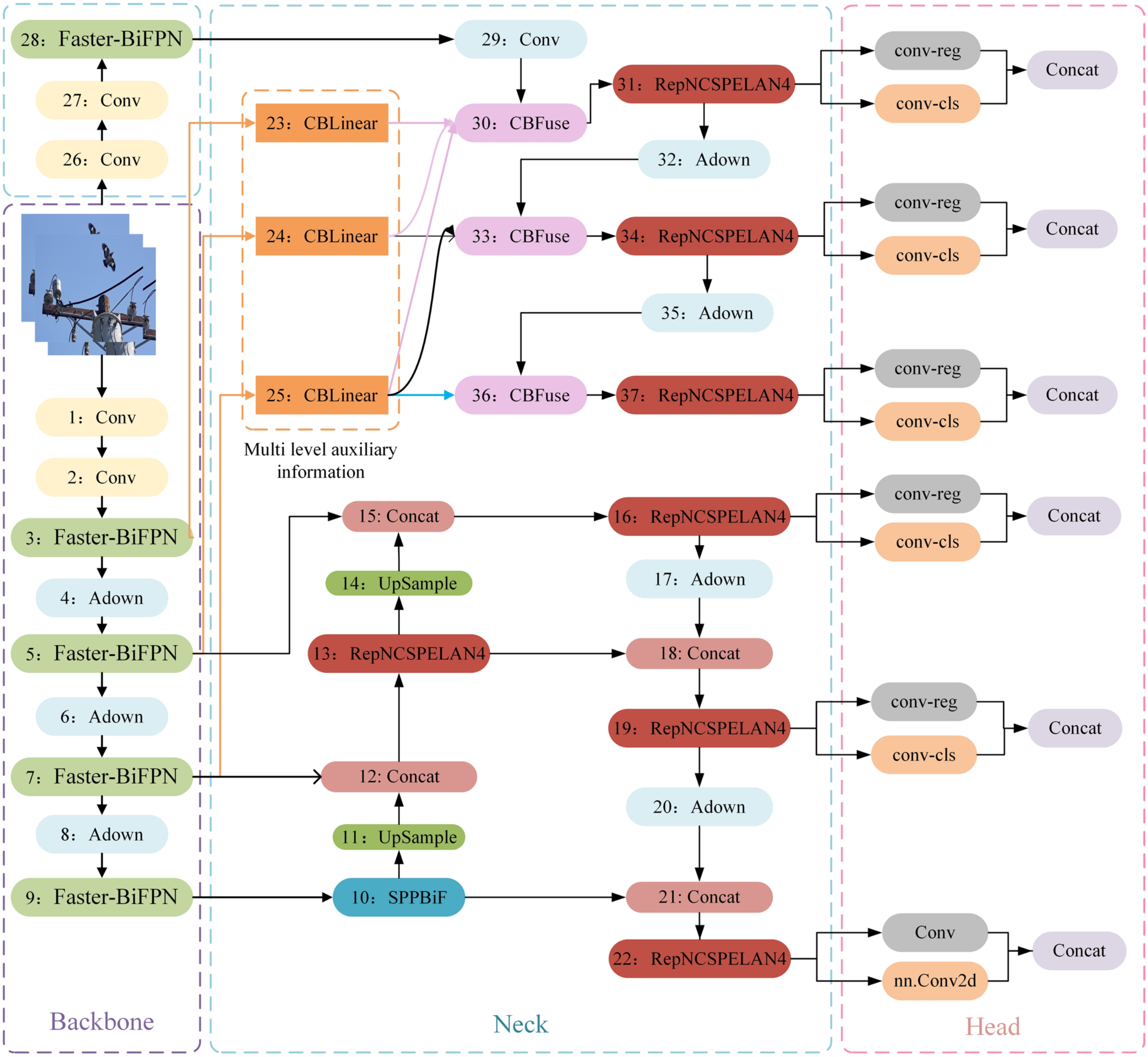
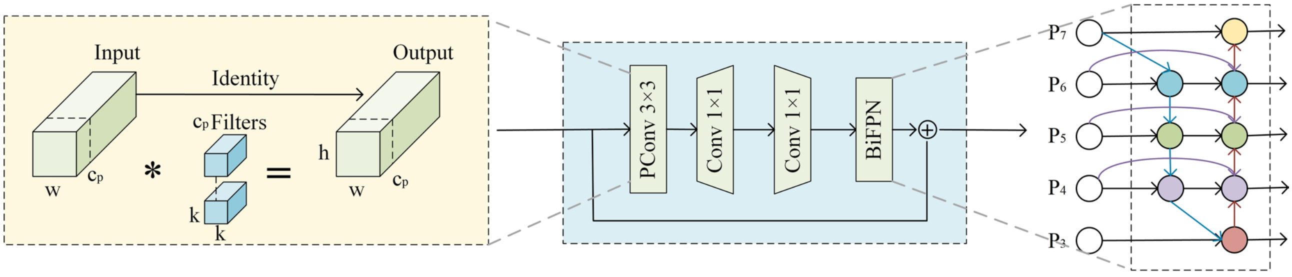


Fig. 2. YOLO-birds target detection network framework.



Partial Convolution(Pconv)

Faster-BiFPN

BiFPN

Fig. 3. Structure of the faster-BiFPN network.

feature pyramid network (BiFPN)[19] to construct a plug-and-play Faster-BiFPN network architecture, replacing the ResNCSPELAN4 network in YOLOv9. This improvement aims to simplify the network architecture while significantly reducing computational redundancy and memory accesses while preserving the information of each channel for more efficient feature extraction. As shown in Fig. 3, the Faster-BiFPN network consists of a Partial Convolution PConv (Partial Convolution)[20], two Point-by-Point Convolution PWConv (Point-wise Convolution), and the feature processing mechanism BiFPN.

As shown in the figure, the method operates by applying regular convolution for spatial feature extraction to only some of the input channels while maintaining the size of the other channels constant. In order to optimize memory access to consecutive channels, the first or last consecutive channel is computed as if it were representative of the entire feature map. Without compromising generality, it is assumed that the input and output feature maps have the same number of channels. The computation amount is:

$$h \times w \times k^2 \times c_p^2 \tag{1}$$

The computational amount of PConv is only 1/16 of the conventional convolution when the number of channels for the convolution operation is 1/4. In addition, the memory access of Pconv is smaller, i.e.

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx \; h \times w \times 2c_p \tag{2}$$

where h and w are the height and width of the input feature map in pixels, k denotes the size of the convolution kernel, and c_p denotes the number of channels of the input feature map.

In this study, we refer to the jump connection of the BiFPN feature fusion network and fuse the point-by-point convolution of the FasterNet core feature extraction network with BiFPN, which deletes some nodes that have only one input edge and, at the same time, adds jump connections between the input image nodes and the output image nodes to retain their original features, preventing the loss of contextual relationships in feature fusion, and enhancing the performance in the detection of small targets in the process of substation bird identification.

Different from the traditional feature fusion, BiFPN also distinguishes the fusion of different input features in order to learn the importance of different input features, which is a weighted fusion mechanism and for this, BiFPN uses a fast normalization method, with accuracy similar to softmax-based fusion but faster, and the fast normalization method is expressed as follows:

$$Out = \sum_i \frac{w_i}{\epsilon + \sum w_i} \times In_i \tag{3}$$

where w_i stands for the weights, $w_i \geq 0$ can be guaranteed with the activation function ReLU, ε is used to avoid numerical instability, is a minimal value, In_i represents the input features, *Out* represents the result of weighted feature fusion.

SPPBiF attention mechanism

The attention mechanism demonstrates significant advantages in solving the bird detection problem in substation images. It not only effectively mitigates the omission and misdetection problems caused by bird occlusion but also substantially improves detection accuracy in complex scenes, such as substations[21]. To enhance the feature extraction capability of YOLOv9 for small targets in complex background environments, an SPPBiF attention mechanism is designed in this study. This mechanism aims to address the challenges faced by traditional target detection algorithms when dealing with complex scenes and small targets. The overall structure of the SPPBiF attention mechanism and the internal structure of the BiFormer module are shown in Fig. 4.On the left is the overall architecture of SPPBiF, and on the right is the internal structure of the BiFormer module.

While traditional attention mechanisms face severe challenges when scaling to larger models or dealing with more complex tasks, the SPPBiF Attention Mechanism module is improved by the pooling part of the spatial pyramid, which extends the model's sensory field, enriches the feature information of different sizes of birds, and further strengthens the model's ability in extracting and fusing practical feature information of different

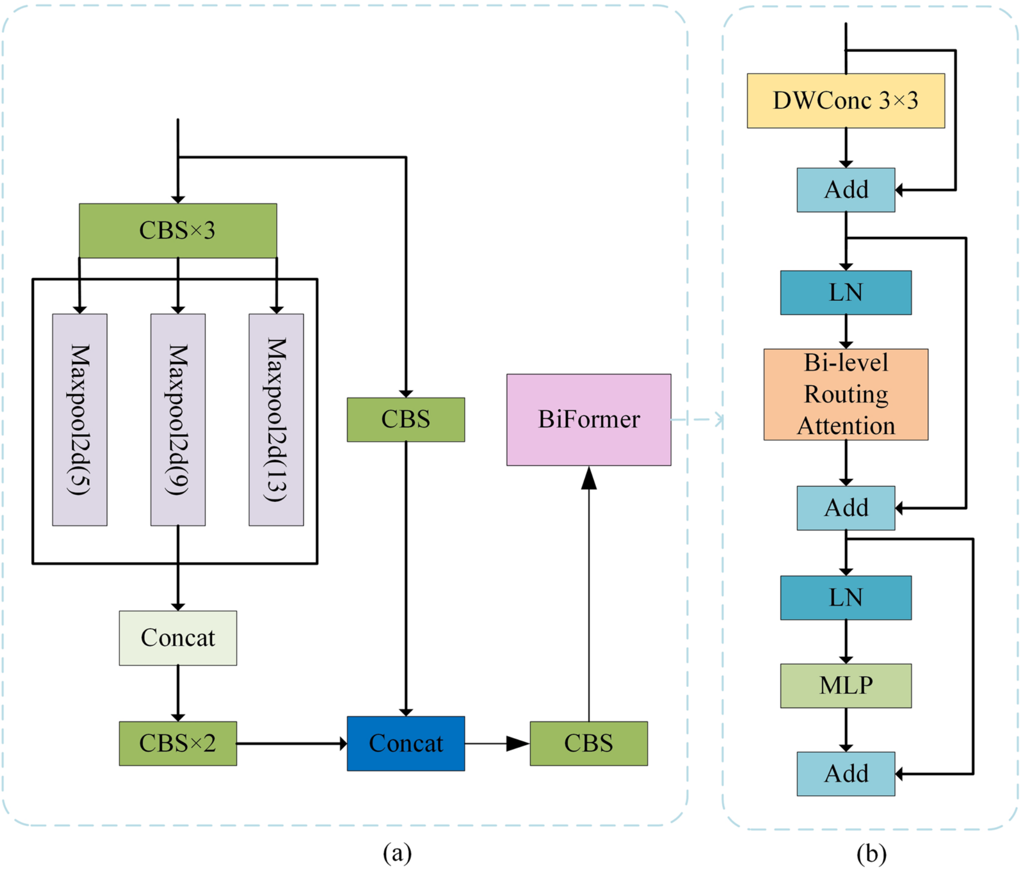


Fig. 4. Illustration of the SPPBiF attention mechanism.

sizes. Moreover, the BiFormer[22] is utilized to use two-level routing attention as the basic building block, which possesses dynamic, query-aware sparsity. Therefore, the SPPBiF attention mechanism module can flexibly implement computational allocation and content perception, which is more helpful for fine-grained target recognition in target detection tasks and detection in complex scenes.

The SPPBiF attention mechanism first divides the features into 2 branches for processing in the feature processing stage, where one branch directly performs the convolution operation, and the other branch enters the SPP structure, which uses the maximum pooling operation to obtain different sensory fields to adapt to images with different resolutions. Within the SPP structure, it first undergoes three convolutional operations, which further enhance the expression of features and help improve the accuracy of target detection. Then, after four maximal pooling operations, using convolution kernels of 1×1, 5×5, 9×9, and 13×13 sizes, respectively, the local and global contextual information of targets of different sizes is captured by performing pooling operations at different scales, generating multi-scale feature maps. Then, these feature maps at different scales are spliced to form a richer feature representation. Finally, the spliced features are fed into the BiFormer module, which adaptively assigns attention weights to different locations and channels in the feature map through its Bi-level Routing Attention mechanism to enhance the feature representation of important regions. This is partly to enhance the model's global dependencies and help it better understand the important information and targets in the scene.

The core module of BiFormer is Bi-Level Routing Attention (BRA); the core idea of BRA is to filter out most of the irrelevant key-value pairs at the rough region level, only retain a small portion of the routing region, and then apply fine-grained token-to-token attention to the union of these routing regions. It can reduce the amount of computation and improve efficiency.

The specific steps for implementing the BRA mechanism are as follows: The specific steps for implementing the BRA mechanism are as follows:

- Given a 2D input feature X∈ $R^{H\times W\times C}$, it is divided into S*times* S non-overlapping regions, with each region containing HW/S^2 feature vectors. The input X is then reshaped into X^r. Subsequently, the query Q, key K, and value V are obtained through linear transformations as defined in Equation (4):

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v \tag{4}$$

here, $W^q, W^k, W^v \in R^{C\times C}$ represent the projection weights for the query, key, and value, respectively.

- By constructing a directed graph, find out the regions that should be attended to for each given region. This step begins by averaging regions over Q and K to obtain region-level Q^r and K^r. Then, by matrix multiplication between the transpositions of Q^r and K^r, an adjacency matrix of region-to-region affinities, A^r, is obtained, and the elements in A^r measure the degree to which the 2 regions are semantically related. From this, only the top k most relevant connections are retained for each region, and a routing index matrix I^r is generated by obtaining the top k indices of each row to retain the top k most relevant connections for each region:

$$A^r = Q^r (K^r)^T \tag{5}$$

$$I^r = topkIndex(A^r) \tag{6}$$

- With the region-to-region routing index matrix I^r, fine-grained Token-to-token attention can be applied. First, the key and value tensor is collected, and then attention can be applied to the collected key-value pairs, i.e.:

$$K^g = gather(K, I^r), V^g = gather(V, I^r) \tag{7}$$

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{8}$$

Where $Attention(Q, K^g, V^g)$ is the attention operation applied to the collected key-value pairs, LCE (V)[23] is a local context augmentation term that is parameterized by a deep convolution to augment the local information. As shown in Fig. 4, BiFormer Block is used to implicitly encode the relative position information using 3×3 deep convolution. Then, the BRA module and the multilayer perceptron (MLP)[24] module are applied sequentially for cross-position relation modeling and per-position embedding, respectively.SPPBiF is the BiFormer concatenated in tandem and spliced together with the other half of the original feature map to perform the feature fusion, which enables the SPPBiF module to extract and integrate feature information from multiple scales to obtain richer gradient flow information. As a result, this paper replaces the SPPCSPC module of the backbone network with the SPPBiF module. It integrates the BiFormer feature of the Transformer domain to enhance the feature extraction capability of the backbone network, thereby improving the network detection accuracy.

Focal-EIoU loss function

The original CIoU (Complete Intersection over Union) loss function[25] of YOLOv9 is a commonly used bounding box regression loss in the target detection task, which can effectively improve the fitting effect of the target box by taking into account the IoU, the distance from the centroid of the bounding box, and the difference in aspect ratio. It performs better in regular scenes and can help the model predict the target frame more accurately. However, CIoU may encounter some drawbacks in special scenarios such as substations. For small target birds in substations, due to their small area and the positional occlusion problem of the target, it is easy to miss, and the change in IoU (intersection and concurrency ratio)[26]will be small, and even if the bounding box has a significant deviation from the real box, the change in the IoU value will be more limited. This leads to the CIoU loss not being sensitive enough in optimizing small targets, which in turn affects the accuracy of the regression. In small target detection, the centroid of the target frame may be shifted more. The difference in aspect ratio is also larger, and relying solely on the loss of IoU versus center distance may not be sufficient to capture the location of the small target accurately. Therefore, the CIoU loss function may lead to a decrease in the accuracy of bounding box regression when dealing with targets in complex scenarios (e.g., substations).

The prediction process of target bounding box regression is unbalanced in the model training samples, and the number of high-quality anchor frames with minor regression errors is much smaller than the number of low-quality anchor frames with significant errors. Poor quality anchor frames produce excessive gradients that interfere with the training process and, therefore, are less effective when using EIoU Loss, which is defined as follows:

$$L_{EIoU} = 1 - IOU(A, B) + \frac{\rho^2(b, b^{gt})}{(h^c)^2 + (W^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} \tag{9}$$

where $IOU = |a \cap b| / |a \cup b|$, w^c and, h^c are the width and height of the minimum enclosing frame, b and b^{gt} denote the coordinates of the centroids of the predicted frame and the real frame, respectively, $\rho()$ denotes the Euclidean distance of the centroid coordinates, c is the diagonal distance of the predicted frame from the minimum external enclosing hold of the real frame, w^{gt} and h^{gt} are the width and height of the real frame. W and h are the width and height of the predicted frame.

In order to improve the detection of small, occluded, or dense targets, reduce the false detection rate, and improve the model's recognition accuracy for specific categories, the Focal-EIoU (Focal Enhanced IoU)[26] loss function is used instead of the original CIoU loss function. The Focal-EIoU Loss combines the category balancing advantage and the location accuracy advantages, making the model more focused on detecting small, occluded targets and matching the position of the predicted frame to the real frame[27]. Its main advantages and effects are as follows:

- Optimisation for the target: Focal-EIoU enhances optimization for the target in the training process by introducing the Focal mechanism, which assigns more weight to targets that are more difficult to detect. The principle of Focal loss is to give a smaller weight to those samples that are easy to predict and a larger weight to those that are difficult to predict, which makes the model focus more on the complex samples.

Operating System	20.04.6
CPU	i5-10400F CPU
GPU	NVIDIA GeForce RTX 3090
RAM	64GB
Deep Learning Framework	PyTorch
Development Languages	Python
Code Editor	VSCode

Table 2. Experimental environment settings.

Batchsize	8
Image Size	640×640
Optimiser	Adam
Epoch	200

Table 3. Model parameter settings.

- EIoU's improvement of the target box: EIoU (Enhanced IoU) further strengthens the constraints on the shape and size of the bounding box based on CIoU and takes into account the geometrical difference between the target box and the real box. EIoU's calculation is more comprehensive, allowing for better optimization of the target's bounding box prediction. By combining the distance, IoU, and size difference of the bounding box, Focal-EIoU can provide more fine-grained bounding box regression.
- Loss amplification: Due to the introduction of the Focal mechanism, Focal-EIoU amplifies the loss of those samples with lower IoU, making the model more 'sensitive' when the target error is significant, thus better guiding the model to optimize these targets.

The calculation steps of the Focal-EIoU loss function are roughly as follows:

- Calculate the IoU values of the predicted bounding box and the real bounding box.
- According to the size of the IoU value, the samples are divided into high-quality samples and low-quality samples. The specific division criteria can be adjusted according to the actual situation.
- Introduce an adjustment factor to balance the contribution of high-quality samples and low-quality samples to the loss. The value of the adjustment factor can be adjusted according to the actual situation, typically ranging from 0 to 1.
- For high-quality samples, the original IoU loss function is used directly; for low-quality samples, a smaller weight is assigned to them according to the size of the IoU value.
- The losses of all samples are weighted and summed to obtain the final Focal-EIoU loss function. Focal-EIoU Loss separates high-quality anchor frames from low-quality anchor frames and improves the detection performance for small targets while maintaining the accuracy for large targets. The Focal-EIoU Loss is calculated as follows:

$$L_{Focal-EIoU} = (IOU)^{\gamma} L_{EIoU} \tag{10}$$

Where γ is the hyperparameter that controls the curvature of the curve, by adjusting γ, it is possible to control how much attention the model pays to the problematic and easy samples during the training process. A smaller γ tends to balance the contribution of all samples, while a larger γ will make the model pay more attention to complex samples and improve the detection of small targets.

Experiment

The experimental environment setup is shown in Table 2. The deep learning framework used in this paper is PyTorch 1.12.0. The software programming environment is Python 3.8.16. The operating system is Ubuntu 20.04.6. All programs are run on an NVIDIA GeForce RTX 3090 graphics card with 64 GB of onboard memory.

The model parameters are set as shown in Table 3. During the training process, the input image size was set to 640 *times* 640 pixels, the batch size was 8, and the total number of training rounds was 200. These parameters were kept constant in the subsequent ablation experiments and comparison experiments to ensure comparable and fair results. Additionally, all network models were trained in the same hardware environment, and their detection performance was subsequently verified on the test set.

Dataset and evaluation metrics

Given the lack of datasets dedicated to substation bird identification on public platforms, we decided to adopt a self-constructed approach by collecting data in the field through web crawlers and drone cameras (State Grid Hangzhou Power Supply Company provided the data collected in the field, and all data does not contain private information such as faces, and will not be used for commercial purposes with proper permission.)This paper integrates the acquired images, removes duplicate images, and preliminarily constructs a raw dataset containing over 1,700 images. To further enhance the model's generalization ability, we apply various data augmentation techniques to the images in the dataset. Specifically, we first use rotation processing to randomly rotate the images by angles ranging from -15 ° to 15 ° , thereby simulating the appearance of birds under different shooting angles and enabling the model to better adapt to bird images from various angles in real-world scenarios. Next, we applied Gaussian blurring using a Gaussian blur algorithm, with the blur radius randomly selected between 0.5 and 2 pixels, to simulate image blurring caused by factors such as camera shake and air pollution under different shooting conditions, thereby improving the model's ability to recognize blurred images. Finally, we added Gaussian noise to the images, with the noise intensity randomly determined between 0.01 and 0.05, to enhance the model's robustness against noisy images. Combining the above data augmentation methods expanded the dataset size from the initial 1,700 images to over 2,800 images. Figure 5 shows some sample images. This paper used the Labellmg tool for data annotation to ensure standardized labeling. To ensure the effectiveness of model training and evaluation, the dataset was randomly divided into training, validation, and test sets in an 8:1:1 ratio.

To comprehensively evaluate the model's performance, we adopted the precision-recall (PR) curve and mean average precision (mAP) as the primary evaluation metrics, which are centered on the calculation of precision and recall. The precision rate reflects the proportion of truly relevant instances in the model prediction results. In contrast, the recall rate indicates the proportion of relevant instances successfully identified by the model to the total relevant instances. The PR curve visually demonstrates the dynamic equilibrium relationship between the precision rate and the recall rate. The larger the area under the curve, the higher the average precision (AP) of the model and the better the performance. Specifically, the mathematical definitions of these evaluation metrics are given by equations (11) to (14). Through these quantitative metrics, we can objectively and systematically analyze the model's performance in the bird recognition task, providing a clear direction for subsequent optimization. The experimental results are evaluated based on the precision rate, recall rate, average precision (mAP@0.5, mAP@0.5:0.95), the number of model parameters, and model size.

Precision and recall are calculated from the TP, FP, and FN of the confusion matrix.

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

In the above, TP refers to the number of correctly detected targets in the results, FP represents the number of incorrectly detected targets, and FN denotes the number of missed targets in the actual targets.

The area under the Precision-Recall (PR) curve represents the Average Precision (AP), and its calculation formula is as follows:

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

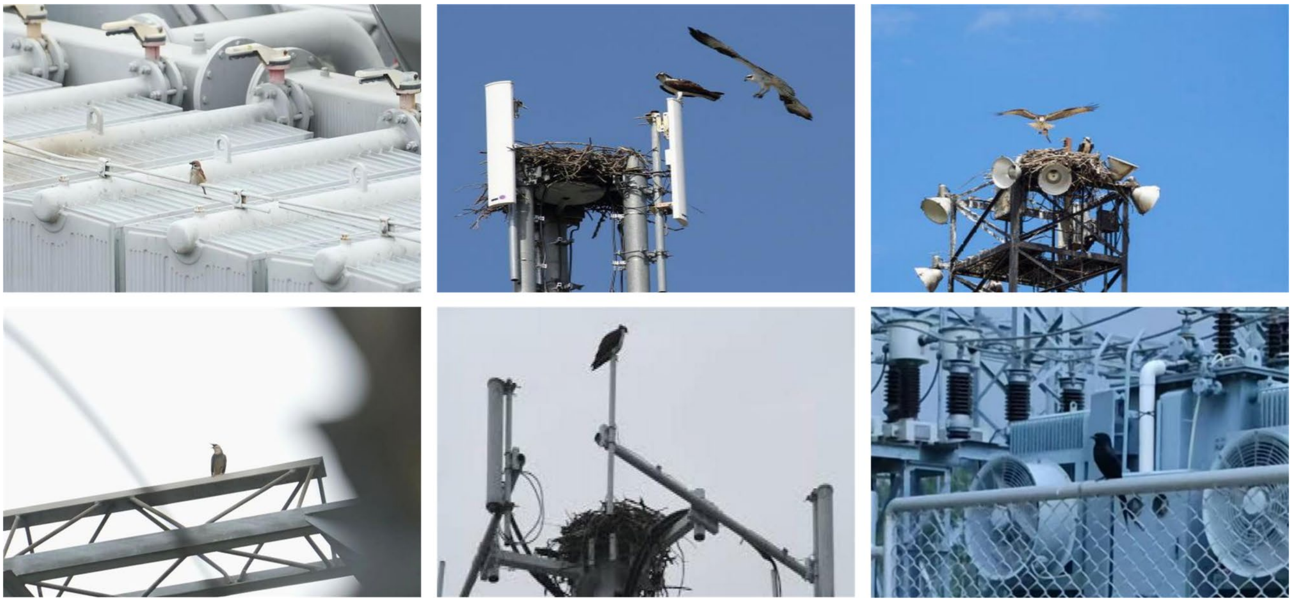


Fig. 5. Sample data sets.

mAP is the average of the AP values across all classes, where mAP@0.5 refers to the mean detection precision for all target classes when the IoU threshold is set to 0.5. mAP@0.5:0.95 represents the average detection precision calculated across all 10 IoU thresholds (ranging from 0.5 to 0.95) with a step size of 0.05.

$$mAP = \frac{\sum_{j=1}^{N} AP(j)}{N} \tag{14}$$

Here, N is the number of classes in the dataset, with higher mAP indicating better object detection performance. Denote the model's parameter count, reflecting computational memory usage.

Experimental results and analysis

Loss function comparison experiment

The loss function used in the YOLO-Birds model of this paper is Focal-EIoU. In order to verify the superiority of the introduced loss function for small target detection, comparison experiments are conducted using Focal-EIoU and some mainstream loss functions applied to the same baseline model, and the experimental results are shown in Table 4.

As can be seen from the comparison of the results in Table 4, it is evident that Focal-EIoU significantly outperforms other mainstream loss functions in detecting small objects. Specifically, the mAP@0.5 reached 85.4%, which is 4.0% higher than that of CIoU, while the mAP@0.5:0.95 reached 59.4%, an improvement of 3.4% over CIoU. Focal-EIoU applies weighted treatment to difficult-to-detect samples, focusing more on challenging bounding boxes, which helps the model better handle small objects with uneven difficulty. The experimental results further validate the effectiveness of Focal-EIoU in improving the performance of the model for small target bird detection in substations. The change curves of different loss functions are shown in Fig. 6.

Comparative experiments

To further validate the effectiveness and superiority of the YOLO-Birds network in bird detection at substations, the YOLO-Birds model was compared with currently popular object detection models. The experimental models include Faster RCNN, SSD, YOLOv5n, YOLOv8n, YOLOv9c, YOLOv10s, YOLOv11n, and YOLOv12n. All experiments were conducted under the same dataset and training environment, adhering to the principle of controlling variables. Table 5 presents the performance results of the different networks.

The experimental results indicate that the algorithm proposed in this paper achieves the best mAP50 performance when compared to standard single-stage object detection algorithms. The improved YOLOv9 model outperforms the baseline model, with an 8.8% increase in mAP50 and a 7.7% improvement in mAP50-95. Compared to other single-stage algorithms, such as YOLOv11n and YOLOv12n, the YOLO-birds model also shows improved mAP50, demonstrating higher accuracy and average precision. The detailed data in Table 5 indicate that the YOLO-birds model proposed in this paper exhibits excellent detection performance on the substation bird dataset, enabling more accurate detection of bird targets in substations, particularly in complex scenarios dominated by small objects. Figure 7 shows the performance curve comparison between YOLO-birds and other models.

To ensure the objectivity of the evaluation results and the model's generalization performance, this study constructed a completely independent test dataset (provided by State Grid Hangzhou Power Supply Company). The test set is strictly decoupled from the training set regarding time dimension, data collection devices, and application scenarios, covering diverse substation operational environments, different weather conditions, and sample data from multiple bird species. This effectively avoids potential bias in model evaluation caused by training data. This study adopted a standardized evaluation metric system on the independently constructed test set to comprehensively quantify the model's performance, including mean average precision (mAP), precision, and recall. This metric system not only accurately evaluates the model's detection accuracy but also deeply measures its generalization adaptability in complex and variable environments. The evaluation results of the independent test set are detailed in Table 6. A comparative analysis of the evaluation results from the training set and test set found that the model demonstrated good performance stability on the test set, fully proving the authenticity and effectiveness of its detection capabilities.

Ablation experiment

In order to validate the effectiveness of each module, we conducted an ablation experiment on the substation bird dataset for comparative analysis. The ablation experiment is designed to explore the impact of different improvements on the model's detection performance, providing important guidance for the model's design

Loss function	mAP50	mAP50-95
CIoU	81.4	56.0
EIoU	78.3	52.3
GIoU	76.4	48.2
SIoU	75.2	46.9
WISE-IoU	83.8	55.6
Focal-EIoU	**85.4**	**59.4**

Table 4. Performance comparison of various loss functions.

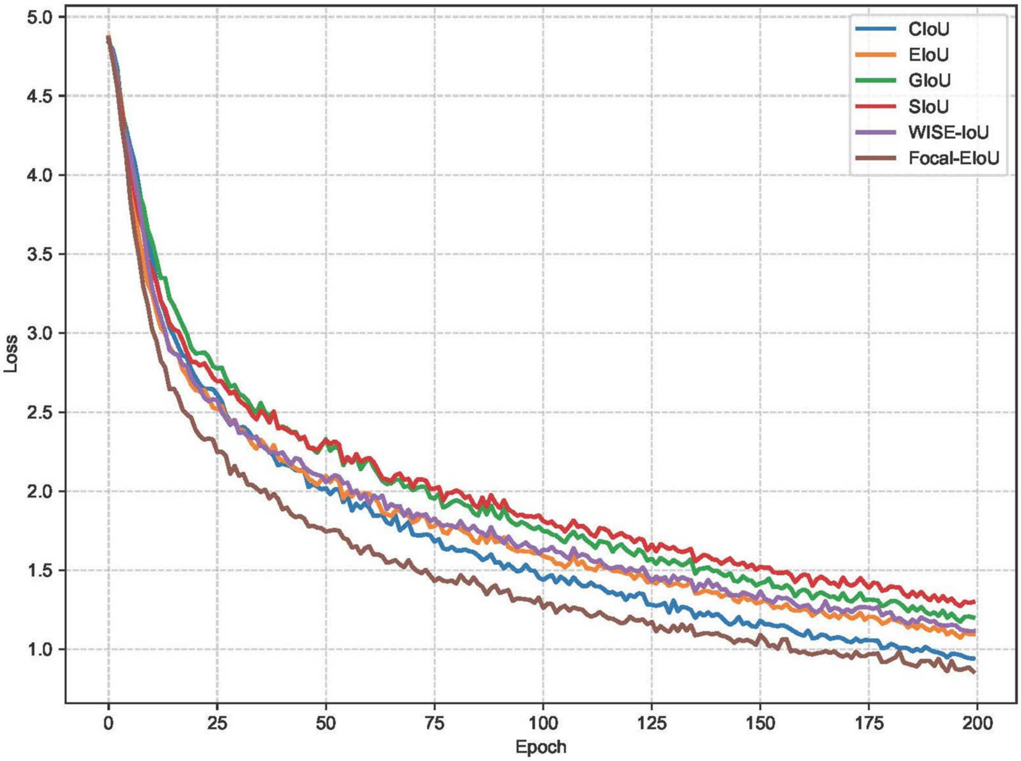


Fig. 6. Comparison of loss function.

Model	Precision	Recall	mAP50	mAP50-95
Faster RCNN	74.5	77.2	72.4	44.5
SSD	76.3	78.5	74.3	44.7
YOLOv5n	82.1	73.9	79.2	54.4
YOLOv8n	83.9	69.7	79.6	56.4
YOLOv9c	85.1	70.9	81.4	56.0
YOLOv10s	85.3	83.1	86.5	56.1
YOLOv11n	82.6	75.4	83.7	54.2
YOLOv12n	82.5	80.3	84.6	53.9
Ours	**84.9**	**86.8**	**90.2**	**63.7**

Table 5. Comparative experiments–different models.

and enhancement. During the experiment, the same parameter configuration was used, and the results of the ablation experiment are presented in Table 7. Where "✓" indicates the use of the corresponding improvement method, and "×" indicates the absence of the corresponding improvement method. To verify the effectiveness of each module, this paper designs ablation experiments on the substation bird dataset for comparative analysis. The ablation experiments are used to explore the impact of different improvement methods on the model's detection performance, which is of great significance as a guide for designing and improving the model. The same parameter configurations are used in the experimental process, and the results of the ablation experiments are shown in Table 7, where '✓' represents the use of the corresponding improvement method, and '×' represents the non-use of the corresponding improvement method.

As shown in Table 7, the comprehensive performance of the model has been significantly improved by introducing innovative network structures and modules. Based on the YOLOv9 model, by introducing the Faster-BiFPN module and the SPPBiF module, both mAP50 and mAP50-95 are improved, effectively enhancing the model's ability in feature extraction and improving the average detection accuracy. To further optimize performance, the Focal-EIoU loss function is used to enhance the model's detection accuracy for small targets, re-emphasizing the key role of the loss function in optimizing detection performance. When the three modules are fused, mAP50 and mAP50-95 both reach their highest values, which improves the original network by 8.8% and 7.7%, respectively, verifying that the proposed improvement strategy in this paper effectively enhances the detection and identification accuracy of birds in substations.

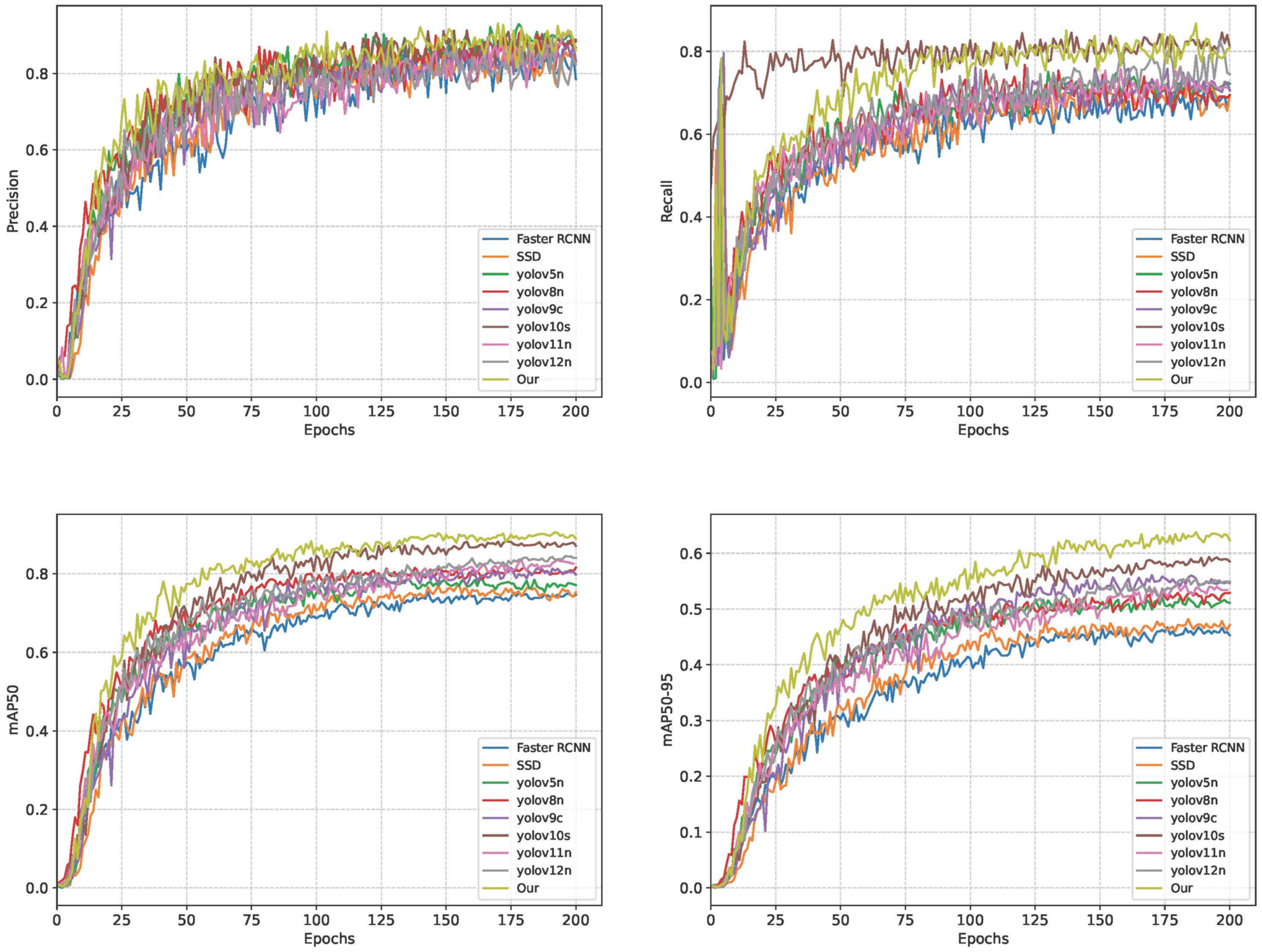


Fig. 7. Curves of YOLO-birds and other models.

Precision	Recall	mAP50	mAP50-95
84.7	86.8	90.4	63.6

Table 6. Model performance on an independent test set.

Baseline model	Faster-BiFPN	SPPBIF	Focal-EIoU	mAP50	mAp50-95
YOLOv9	×	×	×	81.4	56.0
Improved1	✓	×	×	84.1	58.2
Improved2	×	✓	×	83.8	56.1
Improved3	×	×	✓	85.4	59.4
Ours	✓	✓	✓	**90.2**	**63.7**

Table 7. Ablation experiment results of YOLOv9-birds.

Figure 8 illustrates the variation in mAP@0.5 throughout the training process for the YOLO-birds and YOLOv9 models. In this figure, the orange solid line corresponds to the YOLO-birds model, while the blue solid line represents the YOLOv9 model. The x-axis indicates the number of training epochs, whereas the y-axis displays the mAP@0.5 values, which reflect the models' detection accuracy at a confidence threshold of 0.5. As depicted in Fig. 6, the YOLO-birds model demonstrates a consistently higher mAP@0.5 compared to the YOLOv9 model once the training process reaches convergence. This observation underscores the superior performance of the YOLO-birds model in achieving higher accuracy in target detection tasks.

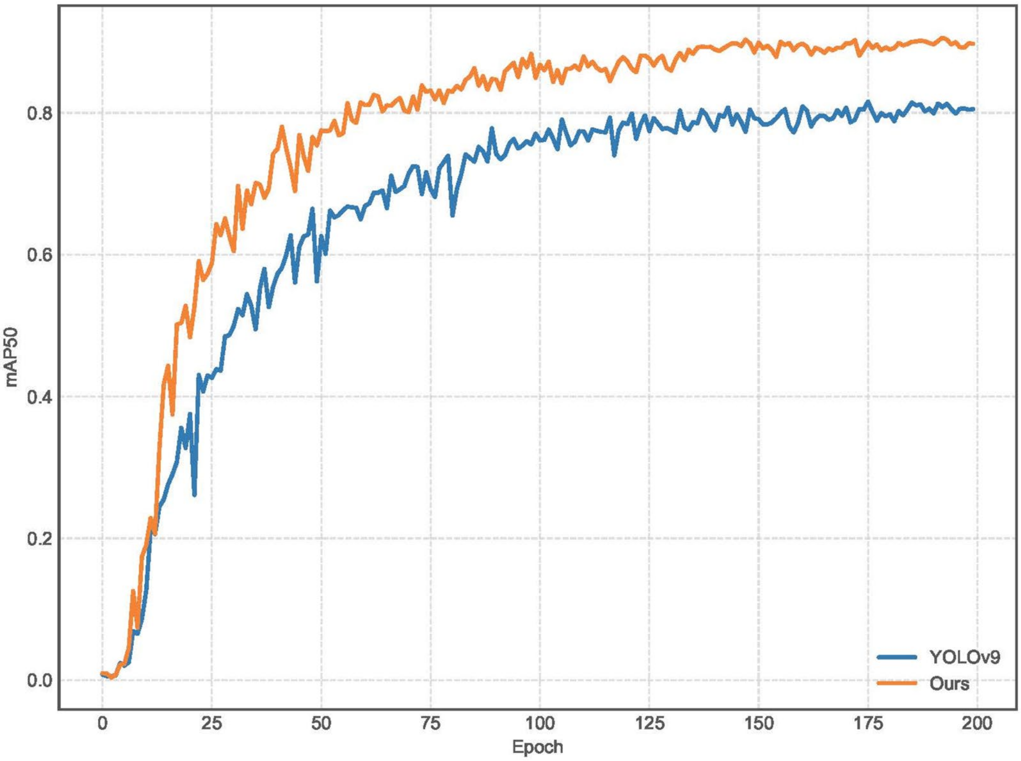


Fig. 8. Comparison of mAP50 between the YOLOv9 and YOLO-bird.

Model	FLOPs	mAP50	mAp50-95
YOLOv9	38.7	81.4	56.0
YOLOv9 + SE	39.0	76.1	50.9
YOLOv9 + EMA	40.2	71.7	46.1
YOLOv9 + CBAM	41.5	79.6	54.3
YOLOv9 + SPPBiF	**35.8**	**83.8**	**56.1**

Table 8. Comparison of computational performance and accuracy.

To comprehensively validate the effectiveness and superiority of the SPPBiF attention mechanism, this study conducted an in-depth experimental analysis from two dimensions: computational performance and feature representation. First, we conducted a detailed assessment of the computational overhead of the SPPBiF mechanism. Under the same training and testing configurations, we compared SPPBiF with classic attention mechanisms such as SE, CBAM, and EMA. The primary evaluation metrics included the number of computational complexity (FLOPs)and detection accuracy (mAP). The experimental results are shown in Table 8. As can be seen from the table, compared with the baseline YOLOv9, the model integrating SPPBiF reduces FLOPs by 7.5% , fully demonstrating the superiority of this mechanism in terms of efficiency-accuracy trade-off.

To investigate the working mechanism of the SPPBiF mechanism further, this paper conducted feature representation analysis experiments. Figure 9 presents the visualization comparison results of the SPPBiF attention mechanism: the original input image (Fig. 9a), the attention heatmap of the baseline model (Fig. 9b), and the attention heatmap after integrating the SPPBiF module (Fig. 9c). Visualization analysis indicates that the SPPBiF mechanism demonstrates exceptional performance in target localization accuracy, particularly in small-scale bird target detection tasks in complex background environments. Compared to the baseline model, the SPPBiF mechanism significantly enhances the response intensity to bird target regions while effectively reducing background noise interference, achieving more focused and precise feature learning.

Experimental results visualization analysis

Figure 10 presents a performance comparison of this paper's method with mainstream target detection algorithms, such as YOLOv9, for substation bird detection. According to the figure, the algorithm presented in this paper achieves a higher detection accuracy for substation birds, thereby reducing the false detection rate. Due to the occlusion of the target birds by the power equipment, the other methods suffer from false detections and omissions in their detection results. However, the algorithm proposed in this paper can accurately locate

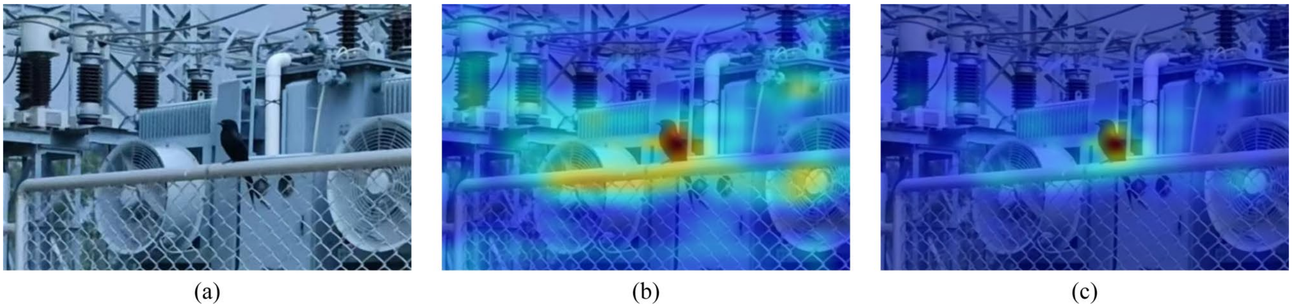


Fig. 9. Comparison of visualization results after adding the SPPBiF attention mechanism.

the position of small target birds and does not produce the phenomenon of false detection or missed detection. This reflects the excellent performance of the algorithm presented in this paper in mitigating interference from complex substation scenes. The experimental results show that YOLO-birds can well detect the objects in the image from different fields and viewpoints, including the situation of complex backgrounds, random direction, noise interference, etc. It can accurately extract key features in the image and completely and accurately detect objects in the image, thereby improving detection accuracy and reducing the rates of misdetection and false detection.

Discussion

In this paper, we propose the YOLO-birds algorithm to address the challenge of detecting small bird targets in substations with complex backgrounds. Compared to previous studies, our method achieves significant improvements in data, methodology, and detection performance. In terms of data, we constructed a dedicated dataset based on real substation scenarios. This ensures greater diversity and representativeness of training samples, covering multiple perspectives and bird scales. In terms of methodology, we build upon YOLOv9 and integrate enhanced feature extraction and fusion mechanisms by introducing the Faster-BiFPN module. Additionally, we optimize the loss function to place greater emphasis on boundary regression for small targets, an aspect often overlooked in previous studies. In terms of results, experimental comparisons with state-of-the-art networks such as YOLOv11 and YOLOv12 demonstrate that our YOLO-birds model achieves superior detection performance. Specifically, the mAP@50 metric improved by 8.8% on the experimental dataset. Furthermore, the method outperforms other models in terms of accuracy, recall, and small object detection metrics, demonstrating its practical application value in bird monitoring at substations.

Although this study has achieved good results in detection accuracy, there is still room for optimisation in model parameter quantity and inference speed (FPS). Future work will focus on further compressing and optimising the model structure to reduce computational complexity and parameter scale effectively. The aim is to improve the model's inference speed and deployment adaptability while maintaining accuracy, providing more practical detection solutions for real-world applications in mobile devices or edge computing scenarios.

Conclusion

Addressing the issues of low detection accuracy and frequent false positives and false negatives in small-object bird detection within substation scenarios, this paper proposes an improved YOLOv9-based bird detection method for substations, named YOLO-birds. The main conclusions are as follows:

This paper introduces the Faster-BiFPN module to replace the original RepNCSPELAN4, thereby reducing computational redundancy and memory access and effectively achieving efficient feature extraction, which improves the model's detection accuracy. This paper introduces the SPPBiF attention mechanism to optimize the backbone network of the original model, effectively enhancing the model's ability to extract and fuse feature information of different sizes, thereby improving the model's detection accuracy in complex scenarios. At the same time, by integrating the attention mechanism, the model's feature extraction capability for output feature maps is achieved, thereby improving the overall detection performance of the model. This paper replaces CIoU Loss with Focal-EIoU Loss, effectively suppressing the fluctuation in loss values caused by low-quality samples and significantly enhancing the robustness of the object detection model. A series of experiments demonstrates that the improved algorithm proposed in this paper achieves an 8.8% improvement in detection accuracy on mAP50 and a 7.7% improvement on mAP95 compared to the original algorithm.

In summary, the improved YOLO-birds substation bird detection method proposed in this paper achieves higher detection efficiency and accuracy compared to the original YOLOv9 model. The model demonstrates good overall performance and high feasibility, aiding in the timely detection of birds within substations and ensuring the safe and stable operation of power systems.

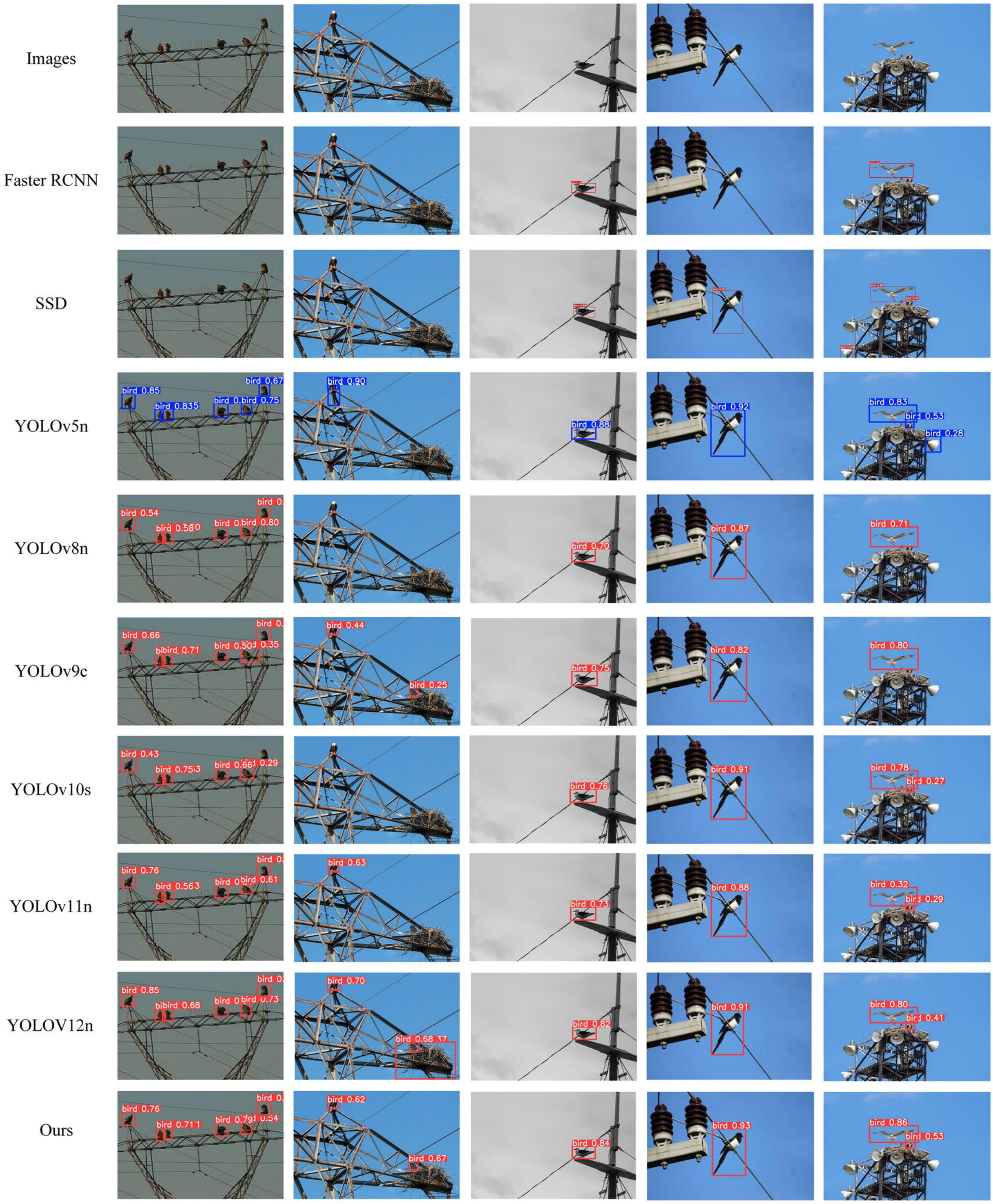


Fig. 10. Comparison of visual images detected by different methods.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 17 December 2024; Accepted: 1 October 2025
Published online: 06 November 2025

References

1. Jenssen, R. et al. Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning. *Int. J. Electr. Power Energy Syst.* **99**, 107–120 (2018).
2. Cai, Z. & Vasconcelos, N. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1483–1498 (2019).
3. Girshick, R. Fast r-cnn. *arXiv preprint* arXiv:1504.08083 (2015).
4. 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.* **39**, 1137–1149 (2016).
5. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1904–1916 (2015).
6. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In: *Proc. IEEE international conference on computer vision*, 2961–2969 (2017).
7. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4–24 (2020).
8. Mao, Q.-C., Sun, H.-M., Liu, Y.-B. & Jia, R.-S. Mini-yolov3: real-time object detector for embedded applications. *Ieee Access* **7**, 133529–133538 (2019).
9. Redmon, J. You only look once: Unified, real-time object detection. In: *Proc. IEEE conference on computer vision and pattern recognition* (2016).
10. Zhang, Z. et al. Research on bird nest image recognition and detection technology of transmission lines based on improved faster-rcnn algorithm. In *2023 5th Asia Energy and Electrical Engineering Symposium (AEEES)* (ed. Zhang, Z.) 218–222 (IEEE, 2023).
11. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint* arXiv:2004.10934 (2020).
12. Redmon, J. & Farhadi, A. Yolo9000: better, faster, stronger. In: *Proc. IEEE conference on computer vision and pattern recognition*, 7263–7271 (2017).
13. Liu, W. et al. Ssd: Single shot multibox detector. In *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (ed. Liu, W.) 21–37 (Springer, 2016).
14. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In: *Proc. IEEE conference on computer vision and pattern recognition*, 2117–2125 (2017).
15. Jiang, T., Zhao, J. & Wang, M. Bird detection on power transmission lines based on improved yolov7. *Appl. Sci.* **13**, 11940 (2023).
16. Wang, C.-Y. *et al.* Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 390–391 (2020).
17. Wang, C.-Y., Liao, H.-Y. M. & Yeh, I.-H. Designing network design strategies through gradient path analysis. *arXiv preprint* arXiv:2211.04800 (2022).
18. Vo, H.-T., Thien, N.N. & Mui, K.C. Bird detection and species classification: using yolov5 and deep transfer learning models. *Int. J. Adv. Comput. Sci. Appl.***14** (2023).
19. Tan, M., Pang, R. & Le, Q.V. Efficientdet: Scalable and efficient object detection. In: *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790 (2020).
20. Ma, X. et al. Pconv: The missing but desirable sparsity in dnn weight pruning for real-time execution on mobile devices. *Proc. AAAI Conf. Artif. Intell.* **34**, 5117–5124 (2020).
21. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021).
22. Zhu, L., Wang, X., Ke, Z., Zhang, W. & Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In: *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 10323–10333 (2023).
23. Ren, S., Zhou, D., He, S., Feng, J. & Wang, X. Shunted self-attention via multi-scale token aggregation. In: *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 10853–10862 (2022).
24. Guo, M.-H., Liu, Z.-N., Mu, T.-J. & Hu, S.-M. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 5436–5447 (2022).
25. Zheng, Z. et al. Distance-iou loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **34**, 12993–13000 (2020).
26. Zhang, Y.-F. et al. Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing* **506**, 146–157 (2022).
27. Tong, Z., Chen, Y., Xu, Z. & Yu, R. Wise-iou: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint* arXiv:2301.10051 (2023).

Acknowledgements

This work was supported in part by Key Project of Jiangsu Provincial Key Laboratory of Power Transmission and Distribution Equipment Technology Team (2023JSSPD01), and in part by Key Laboratory about Maritime Intelligent Network Information Technology of the Ministry of Education (EKLMIC202405).

Author contributions

X.F. and X.W. wrote the main manuscript text and prepared figures, C.S. verified the experimental data, P.S. supervised the writing of the manuscript, X.F. and P.S. were responsible for the funding of the project, C.S. and Y.X. edited the article, J.H. provided the test data. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.