electronics

MDPI

Article

SOD-YOLO: A High-Precision Detection of Small Targets on High-Voltage Transmission Lines

Kaijun Wu, Yifu Chen, Yaolin Lu, Zhonghao Yang, Jiayu Yuan and Enhui Zheng *

School of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China; p22010855027@cjlu.edu.cn (K.W.); p22010854011@cjlu.edu.cn (Y.C.); s22010811023@cjlu.edu.cn (Y.L.); p23010854144@cjlu.edu.cn (Z.Y.); p23010854149@cjlu.edu.cn (J.Y.)
* Correspondence: ehzheng@cjlu.edu.cn

Abstract: Wire clamps and vibration-proof hammers are key components of high-voltage transmission lines. The wire clips and vibration-proof hammers detected in Unmanned Aerial Vehicle (UAV) power inspections suffer from small size, scarce edge information, and low recognition accuracy. To address these problems, this paper proposes a small object detection (SOD) model based on the YOLOv8n, called SOD-YOLO. Firstly, an extra small target detection layer was added to YOLOv8, which significantly improves the small target detection accuracy. In addition, in order to enhance the detection speed of the model, the RepVGG/RepConv ShuffleNet (RCS) and a OneShot Aggregation of the RCS (RCSOSA) module were introduced to replace the C2f module in the model backbone and neck shallow networks. Finally, to address the excessive focus on low-quality sample bounding boxes during model training, we introduced Wise-CIoU loss instead of CIoU loss, which improved the detection accuracy of the model. The experimental results indicate that SOD-YOLO achieved a mean average precision of 90.1%, surpassing the YOLOv8n baseline model by 7.5% while maintaining a model parameter count of 3.4 M; the inference speed reached 88.7 frames/s, which meets the requirement of real-time recognition.

Keywords: YOLOv8; small target detection; wire clamps; vibration-proof hammers

Citation: Wu, K.; Chen, Y.; Lu, Y.; Yang, Z.; Yuan, J.; Zheng, E. SOD-YOLO: A High-Precision Detection of Small Targets on High-Voltage Transmission Lines. *Electronics* **2024**, *13*, 1371. https://doi.org/10.3390/electronics13071371

Academic Editors: José Matas and Ahmed Abu-Siada

Received: 13 March 2024
Revised: 28 March 2024
Accepted: 3 April 2024
Published: 4 April 2024

1. Introduction

The power grid is an important public infrastructure for national economic and social development, and the safe operation of high-voltage transmission lines is essential for the stable supply and use of electricity. In the operation of transmission lines, due to the complex and changeable open-air environment, various abnormal situations may occur, such as loosening, fracturing, or falling off of components like overhanging wire clips, tension-resistant wire clips, and vibration-proof hammers. These anomalies not only affect the stable operation of the transmission line but may even cause major safety accidents in serious cases. These parts belong to the category of small targets, which occupy a small area in the image with low resolution, high localization requirements, scarce edge information, and serious misdetection and omission. Therefore, real-time and accurate identification of small targets on transmission lines is of great significance. The traditional detection methods mainly rely on manual inspection and simple auxiliary tools, with low inspection efficiency and high risk, unable to meet the requirements of intelligent inspection. In recent years, with the popularity of UAV technology in intelligent inspection, inspection efficiency has greatly improved [1].

The inspection images obtained from UAV photography need to be further subjected to target detection. Traditional image morphological detection usually uses algorithms such as the Hough transform [2], LBP texture features [3], Canny edge detection [4], and other algorithms. However, due to the complexity of the transmission line background and the blurring of the morphological features of small targets in the image, it is difficult for

traditional algorithms to recognize them by image shape and edge profile features. Deep learning-based object detection algorithms with strong feature extraction capability and good robustness can identify and localize targets more accurately [5].

Currently, deep learning-based object detection algorithms are generally categorized into one-stage and two-stage detection approaches. Two-stage object detection algorithms are mainly represented by Region-Based Convolution Neural Networks (R-CNN) [6], Fast R-CNN [7], Faster R-CNN [8], and so on. Earlier researchers tended to focus more on two-stage object detection algorithms in the pursuit of higher detection accuracy. Zhang et al. [9] used the Faster R-CNN deep learning framework to design a two-stage cascade Region Proposal Network (RPN) to improve the accuracy of defect detection of four types of anti-vibration hammers. Bao et al. [10] employed focal loss to enhance the classification performance of the RPN, which was introduced into the Cascade R-CNN, which solved the problem of data category imbalance. Zhai et al. [11] proposed a geometric characteristic learning (GCL) model and applied it to Faster R-CNN to generate artificial samples for 3D modeling. Secondly, to enhance the extraction of geometric features from the vibration-proof hammer, monochrome-backgrounded artificial samples were employed during the training stage. Zhou et al. [12] proposed a deep aggregation feature extraction network and an efficient weighted feature-fusion network to replace the original ResNet and Feature Pyramid Network (FPN) of Cascade R-CNN, which balances the inference speed and average accuracy during detection; while the two-stage target detection method offers improved accuracy, its detection speed is sluggish, failing to fulfill the real-time demands of UAV detection.

To increase the speed of detection, current research mainly uses one-stage target detection methods such as the Single Shot MultiBox Detector (SSD) [13], You Only Look Once v3 (YOLOv3) [14], YOLOv7 [15], etc. Tu et al. [16] used the K-means++ clustering method to calculate an anchor in the YOLOv3 model, resulting in better effect frames to improve detection accuracy. Jia et al. [17] proposed a YOLOv4 model-based shockproof hammer identification algorithm. Multi-scale cavity convolution was utilized to increase the receptive field, resulting in richer global information, which greatly improved the accuracy of recognition. Yuan et al. [18] introduced the Squeeze and Excitation (SE) attention mechanism into YOLOv5 in order to further improve the visual recognition of transmission lines by assigning different weights to the images from the channel dimension, thus obtaining feature information with different levels of importance. The recognition accuracy is higher through the improved YOLOv5 network, which is significant in reducing the workload of inspectors. Di et al. [19] introduced multilayer convolutional operations and feature pyramid structure into YOLOv5 and established a target detection model suitable for transmission lines. Lu et al. [20], based on the YOLOv5s algorithm, incorporated a lightweight Ghost convolution module into the backbone network, which reduces the feature map redundancy and improves the inference speed in the feature extraction part of the model. In addition, an attention mechanism based on coordinated attention (CA) was incorporated to effectively extract key feature information. Yu et al. [21], based on the YOLOv7 algorithm, achieved hyperparameter optimization using the Genetic Algorithm (GA) by replacing the convolutional and pooling layers of YOLOv7 with Space to Depth (SPD) spatial-depth-transformed convolution to improve the accuracy of foreign object identification for aerial UAVs. Shao et al. [22], based on the YOLOv7 Tiny algorithm, using the Focal-DIoU (Distance–Intersection over Union) loss function to effectively solve the problem of sample category imbalance and sample difficulty imbalance. In addition, they optimized and improved the SPPCSPC_S-F (Spatial Pyramid Pooling and Cross-Stage Partial Connections) module, which was optimized to reduce the cost of computational resources and time while ensuring model accuracy.

All in all, the existing models suffer from the problem that detection accuracy and speed cannot be balanced simultaneously. On the one hand, two-stage target detection methods cannot adequately capture small targets with shallow features, and their inference speed is still challenging. On the other hand, the poor effectiveness of multi-scale fusion in

the first-stage target detection methods leads to low accuracy of the model in recognizing small objects. In addition, the small object detection effect in UAV inspection images is also often affected by the complex background. To solve the above problems, an improved network SOD-YOLO, based on YOLOv8n, is introduced in this article.

The following are the main contributions of this article:

- To enhance the detection ability of the model for small targets, the small object detection layer (SODL) was incorporated into YOLOv8n, feature maps of different scales were acquired, and multi-scale feature extraction and fusion were performed. The detection head was designed after large-scale feature mapping to optimize the detection performance for small targets.
- By integrating the RCSOSA module into the backbone and neck shallow layers of the SOD-YOLO model, this approach significantly improved the accuracy and speed of model identification.
- To balance the strength of the bounding box regression and punishment for low-quality data during model training, we designed the Wise Intersection over Union–Complete Intersection over Union (WIoU-CIoU) loss as the bounding box regression loss function. It effectively reduces the harmful gradient of low-quality samples and improves the detection accuracy of the SOD-YOLO model with the same inference speed and number of model parameters.

2. Related Work

2.1. YOLOv8 Algorithm

YOLOv8 is a one-stage target detection algorithm open-sourced in January 2023 by Ultralytics. Compared to the previous YOLO model, this model has higher speed and accuracy and provides a unified framework for model training, including image classification, target detection, and instance segmentation. In this paper, YOLOv8n is used as a baseline for improvement, and the model structure is depicted in Figure 1.

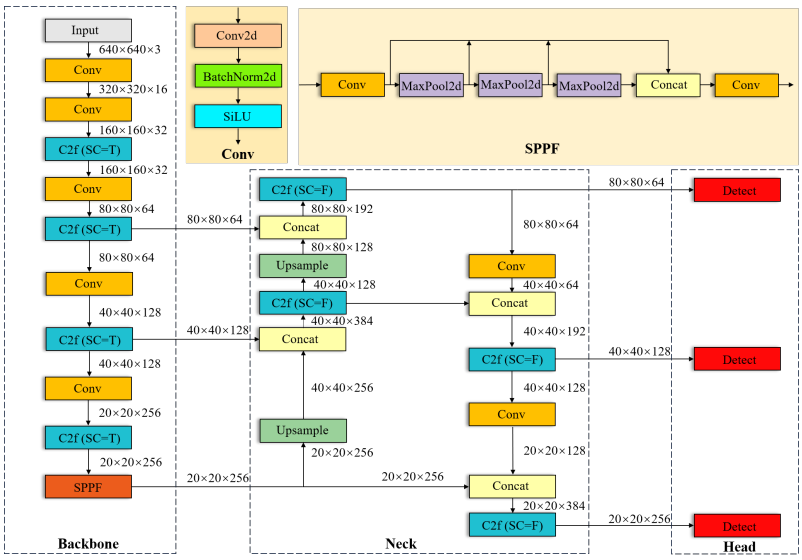

Figure 1. The structure diagram of YOLOv8. SC=T, SC=F indicates Shortcut=True, Shortcut=False, respectively.

The YOLOv8 model comprises three main parts: backbone, neck, and head. The backbone is mainly responsible for extracting key features from input images. The backbone consists of multiple Conv, Cross-Stage Partial Network Fusion (C2f), and spatial pyramid pooling fast (SPPF) modules. The Conv module consists of Conv2d, batch normalization, and the SiLU activation function. YOLOv8 refers to the C3 module in YOLOv3 and the Efficient Layer Aggregation Network (ELAN) idea in YOLOv7 to use the C2f structure,

which uses more gradient streams in parallel and adds the Split operation to the performance of the model in feature extraction; the structure of the module is shown in Figure 2. Meanwhile, the C2f structure reduces the number of module parameters and computational complexity compared to the C3 structure, which improves the running speed of the model. The SPPF module is an effective spatial pyramid pooling structure that enables the network to feature extract at different scales by dividing the image into pyramid layers of different sizes. The FPN [23] and Path Aggregation Network (PAN) [24] modules are used in the YOLOv8 neck architecture to enhance the characterization of features at different scales of the model through multilayer feature fusion and enhancement. The Decoupled Head structure is adopted to learn the target category information and location information through different network branches to avoid interference caused by the network in processing different tasks. In addition, to avoid the complexity and uncertainty of anchor matching, the anchor-free method is employed for generating the bounding box of the target object by making predictions directly at each position in the network.

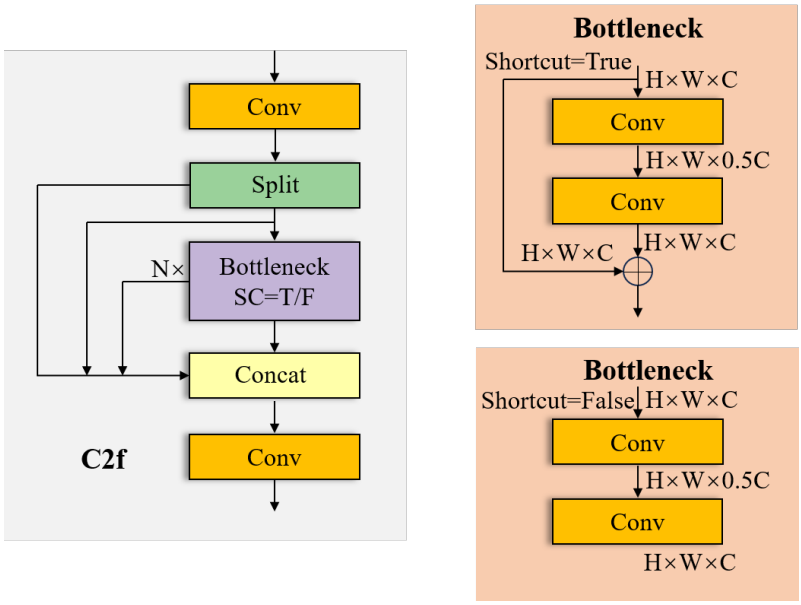

Figure 2. The structure diagram of the C2f module.

2.2. Small Object Detection Layer (SODL)

The UAV inspection images have high resolution, so there are problems such as small target areas relative to the background and inconspicuous target features. Having a large downsampling rate in the YOLOv8 baseline model generates a smaller feature map size. This results in small targets object occupying fewer pixels on the feature map, making it difficult to obtain feature information of the targets. In addition, multiple targets occluding each other cause the model to have difficulty distinguishing target categories, leading to the phenomenon of missed and false detection.

To solve the above problems, an SODL was incorporated into YOLOv8 to fuse the shallow and deep features of small targets to enhance the model's ability to learn about small targets. Five layers P1, P2, P3, P4, and P5 are included in the backbone network of the YOLOv8 network model. The resolution of the network input is 640 × 640, and five different multi-scale feature maps are generated after five downsamplings of the backbone network: 320 × 320 (P1), 160 × 160 (P2), 80 × 80 (P3), 40 × 40 (P4), and 20 × 20 (P5). The original model designs three detection heads on the P3, P4, and P5 layers with different sizes of receptive fields, which can cover the detection of different target scales. However, due to the specific scenario of UAV transmission line inspection in this paper, a large number of targets with a resolution of less than 32 × 32 exist in the data. Furthermore, the receptive fields of the P3 layer feature maps remain sizable, containing more background information

and potentially interfering with the detection of small targets. Therefore, we added an extra high-resolution detection head in the P2 layer of YOLOv8 and only two downsampling operations to obtain rich shallow feature information of small targets. Second, the deep network feature information was effectively fused in the neck part to further improve the model detection accuracy. However, adding an extra small target detection layer made the model computationally larger and the Frames Per Second (FPS) lower. However, it demonstrated a great improvement in small target detection accuracy.

2.3. RCSOSA

In UAV power inspection, both inference speed and detection accuracy play crucial roles. As deep learning continues to evolve rapidly, increasingly intricate network architectures, such as ResNet [25], Convnext [26], Swin Transformer [27], etc., have greatly improved the detection accuracy of vision tasks, but all of them make the inference of the models slower. On the other hand, some lightweight models such as MobileNetv3 [28] use deeply separable convolution and a linear bottleneck structure to reduce model computation and parameters. ShuffleNet improves the parallel capability of the model through channel shuffle and group convolution. Although the lightweight model can accelerate the model inference speed, the accuracy of small target detection cannot meet the detection requirements. To tackle the preceding issues, Kang et al. [29] proposed the RCSOSA module by considering the detection accuracy and inference speed together. Firstly, inspired by ShuffleNet [30], the authors designed a structured parameterized convolution based on channel shuffle called RCS, and the framework of the module is illustrated in Figure 3.

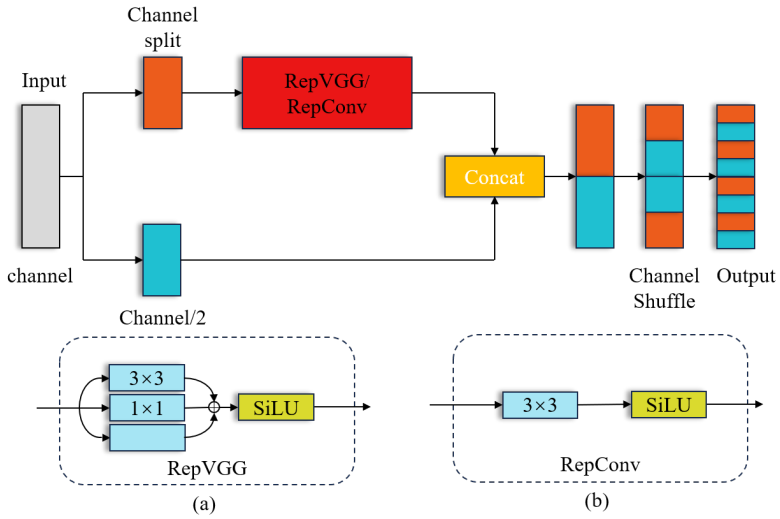

Figure 3. The structure diagram of the RCS module. Where (a) and (b) denote the RepVGG module and the RepConv module, respectively.

The RCS structure is trained using the RepVGG structure with 3 × 3 convolution, 1 × 1 convolution, and Identity branches. The multi-branch topology allows for richer feature information during training. In the inference stage, structure reparameterization is performed, using 3 × 3 RepConv instead of 3 × 3 convolution, 1 × 1 convolution, and Identity branches. The multi-branch topology allows for richer feature information during training. The simple single-branch structure reduces the memory and computation in the inference phase to increase the inference speed. Concat splicing is done in channel dimension by RepVGG with RepConv followed by Concat splicing with the Channel Split part of the tensor. Finally, the two-branch feature map channels are recombined through Channel Shuffle to promote the information exchange between different channels to extract richer feature information. To mitigate the computational burden associated with the RCS module, the RCSOSA module is proposed by combining RCS and One-Shot Aggregation (OSA), as shown in Figure 4. This module achieves feature reuse by repeatedly stacking

RCS modules. Three feature cascades are retained on the OSA pathway to reduce the computational effort and achieve fast inference with high accuracy.

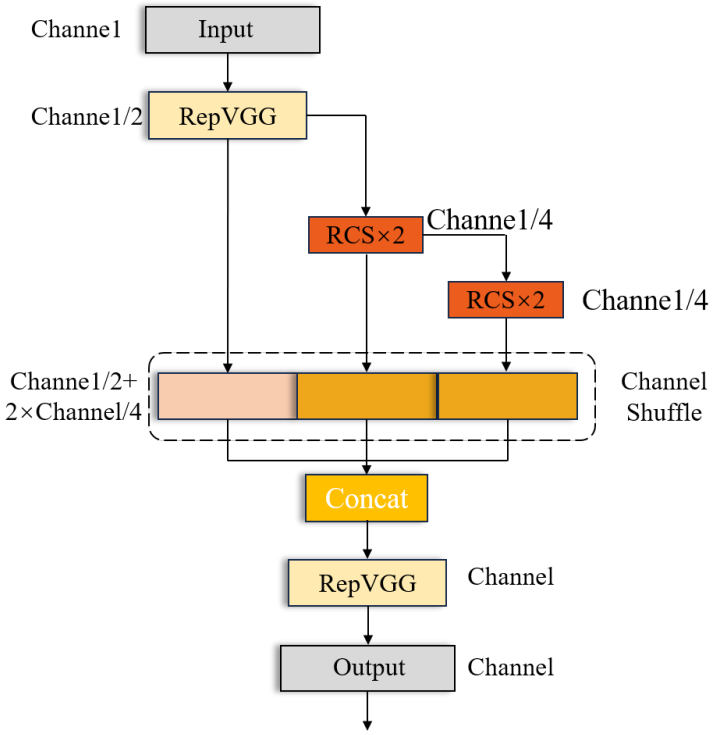

Figure 4. The structure diagram of the RCSOSA module.

3. Methods

To enhance the detection capabilities of tiny targets in inspection images of drones, we introduce the SOD-YOLO model, the architecture of which is depicted in Figure 5. Firstly, we incorporated a small object detection layer into the SOD-YOLO model. By fusing shallow and deep features and adding a detection head after the shallow feature map, the sensitivity to small targets was enhanced and the accuracy of the model's detection was significantly enhanced. Secondly, the RCSOSA module was added to the shallow network of the backbone and neck of the SOD-YOLO model, replacing the original C2f module, to improve the inference speed of the model and the improvement of small target recognition accuracy. In addition, most studies have not considered the issue of low-quality samples in the training dataset, and, if the model excessively regresses to the bounding box of low-quality samples, it will decrease the detection the model's precision. To tackle this issue, drawing inspiration from Wise-IoU [31], we devised the Wise-CIoU bounding box loss function as a replacement for the CIoU bounding box loss function employed in the baseline model. We added a dynamic non-monotonic focusing method to the default CIoU [32] of YOLOv8. Instead of the traditional IoU, this method uses "outlier" as the primary criterion for assessing the quality of anchor boxes. It introduces a strategic approach to gradient gain assignment, which aims to alleviate competition among top-performing anchor frames while minimizing the negative impact of low-quality samples on the gradients. To a certain extent, this approach helped to minimize the bounding box regression loss, enhanced convergence speed, and, ultimately, boosted the model's detection precision.

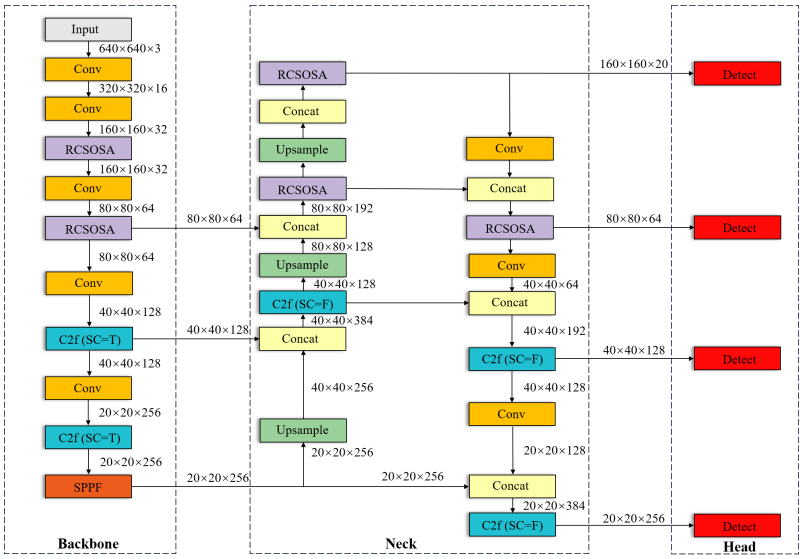

Figure 5. The framework diagram of SOD-YOLO.

The CIoU incorporates three important factors, namely the overlapping area of bounding boxes, the distance from the center point and the aspect ratio, and consists of the following components:

1. Penalty term R_{CIoU}. The penalty term R_{CIoU} is employed to facilitate the alignment of the predicted box with the real box. The formula for this is given below.

$$R_{CIoU} = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{1}$$

where b and b^{gt} are denoted as the centroids of the prediction box and the real box, respectively, and ρ denotes the computation of the Euclidean distance between the two centroids, and c denotes the diagonal distance of the smallest enclosing region that can contain both the prediction box and the real box. In addition, α and v denote the weight parameter and the bounding box aspect ratio similarity, respectively.

2. Weight parameter α. The formula of it is as follows:

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{2}$$

3. Prediction of the overlap ratio between bounding boxes and real bounding boxes IoU. The formula is shown below.

$$IoU = \frac{A \cap B}{A \cup B} \tag{3}$$

A denotes the prediction box and B denotes the ground-truth box.

4. Similarity of bounding box aspect ratio v. The formula of it is as follows:

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

where w and h represent the width and height of the predicted bounding box, while w^{gt} and h^{gt} correspond to the width and height of the ground-truth bounding box.

Combining the aforementioned four equations, the final L_{CIoU} loss function is calculated as detailed below.

$$L_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{5}$$

Based on Wise-IoU loss, we first used CIoU as a penalty term and designed L_{CIoUv1}. This is a bounding box loss function that includes a two-layer attention mechanism and is calculated using the following formula:

$$L_{CIoUv1} = R_{CIoU} L_{IoU} \tag{6}$$

In Formula (6), $L_{IoU} \in [0, 1]$. This considerably diminishes the R_{CIoU} of superior-quality anchor boxes.

Wise-IoU defines an outlier degree for dynamic non-monotonic focusing β, used to assess the quality of anchor frames. The larger the outlier value, the worse the quality of the anchor box. We allocated a smaller gradient gain to it, which effectively prevented low-quality samples from generating larger harmful gradients. The formula is as follows:

$$\beta = \frac{L^*_{IoU}}{\overline{L_{IoU}}} \in [0, +\infty) \tag{7}$$

where L^*_{IoU} represents the monotonic focusing coefficient, and $\overline{L_{IoU}}$ is the exponential running average with momentum m.

Combining L_{CIoUv1} and outlier degree β, we designed a Wise-CIoU and added outlier β to the loss function. The formula is presented below.

$$L_{Wise-CIoU} = r L_{CIoUv1}, r = \frac{\beta}{\gamma \alpha^{\beta - \gamma}} \tag{8}$$

Among them, α and γ are hyperparameters used to adjust the size of r. When β is not equal to γ, the anchor box obtains the highest gradient gain. Among them, $\overline{L_{IoU}}$ is a dynamic parameter, so $L_{Wise-CIoU}$ will choose the optimal gradient gain strategy at the current moment.

4. Experiments

4.1. Dataset

The experimental data in this article come from State Grid Zhejiang Company, and 3376 transmission line inspection images were collected by UAVs shooting high-voltage transmission towers. The dataset was annotated using the image annotation software labelimg, and randomly divided into the training set, validation set, and test set according to the ratio of 7:2:1. The label category information and the number of targets contained in it are shown in Table 1. Shockproof hammer has the highest proportion in the dataset, with a total of 9045 targets.

Table 1. Dataset label categories and quantities.

Category	Training Set	Validation Set	Test Set
suspension clamp	1401	419	188
strain clamp	3114	874	461
shockproof hammer	6265	1859	921

In this paper, small targets are defined according to the ratio of the target bounding box area to the squared area of the image, when it is less than 0.03. A total of 7227 targets in this dataset meet the definition of small targets, accounting for 46.62% of all targets. Among them, shockproof hammer has the highest percentage of the total number of small targets. Figure 6 depicts the distribution of target object sizes for each type of labels. From the figure, it is evident that the small targets are primarily located in the lower left position, indicating a large proportion of small targets in the dataset.

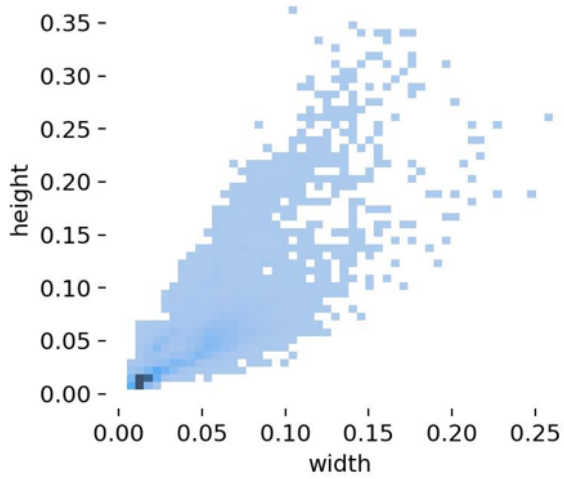

Figure 6. Dataset target size distribution. The colour shades in the graph indicate the density of the target size distribution. The darker the color, the denser the distribution of targets at that size. The lighter the color, the more sparsely distributed the target is at that size.

4.2. Experimental Platform and Hyperparameter Settings

This experiment was performed on the Ubuntu 18.04 operating system utilizing Python 3.9.18, Pytorch 1.12.1, and CUDA 11.4. Training, validation, as well as testing were done using NVIDIA GeForce RTX 3060.

In our experiments, the preprocessing images were resolved at 640 × 640 pixels, with 200 epochs and the batch size of 16. The model optimizer employed was Stochastic Gradient Descent (SGD), and the learning rate was fixed at 0.01. In addition, to save memory and speed up training, the cache was set to True. To improve the model generalization, the mosaic enhancement technique was used and mosaic enhancement was switched off in the last 10 rounds of epoch of model training.

4.3. Evaluation Metrics

Evaluation metrics are essential to assess the model's performance. To test the effectiveness of the proposed method, Precision (P), Recall (R), Average Precision (AP), Mean Average Precision (mAP), Frames Per Second (FPS), and Giga Floating Point Operations (GFLOPs) were used as the evaluation metrics for the performance of the algorithms in this experiment. P, R, AP, and mAP were computed as outlined below.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$AP = \int_0^1 P(r)dr \tag{11}$$

$$mAP = \frac{1}{k}\sum_{i=0}^{k} AP_i \tag{12}$$

In the above equation, Precision (P) denotes the proportion of the samples predicted by the model to be positive samples that are correctly identified as positive samples, whereas True Positive (TP) indicates the number of samples correctly identified by the model as belonging to the positive category, and which, in fact, do belong to the positive category, and False Positive (FP) refers to the number of samples that the model incorrectly predicts as belonging to the positive category, when they actually belong to the negative category.

Recall (R) indicates the percentage of positive samples correctly identified by the model out of all actual positive samples, whereas False Negative (FN) denotes the number of samples that are predicted to be negative but are positive. Average Precision (AP) denotes the average of the precision at different recall rates, and Mean Average Precision (mAP) denotes the average of the AP of each category, that is, the ability of the response model to compute the identified categories.

In addition, FPS is the quantity of images that the model can process per second, and GFLOPs can be utilized to measure the complexity of the model. The above two metrics can be utilized to comprehensively assess the detection speed and efficiency of the model.

4.4. Experimental Results

(1) To verify the superior accuracy and detection speed of the SOD-YOLO model, a comparison with other models on the same dataset was conducted, and the findings are presented in Table 2.

Table 2. Comparison of performance of various models.

Model	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	GFLOPs	FPS
YOLOv3 [14]	91.3	83.8	88.2	103.7	282.2	24.4
YOLOv3-tiny	88.9	68.4	74.2	12.1	18.9	120.1
YOLOv5n [33]	88.2	75.2	80.6	2.5	7.1	71.7
YOLOv6n [34]	90.3	64.6	80.7	4.3	11.8	81.1
YOLOv8n	90.0	76.1	82.6	3.0	8.1	93.2
YOLOv8s	88.9	80.1	84.8	11.1	28.4	78.6
SOD-YOLO	92.7	84.2	90.1	3.4	21.9	88.7

It is evident that the SOD-YOLO model demonstrates excellent performance. As far as the target detection metric mAP is concerned, our SOD-YOLO model has a large improvement. Compared with the YOLOv3 model, which has a model parameter count of 103.7 M, the mAP of SOD-YOLO increases by 1.9%. Compared to YOLOv3-tiny and YOLOv5n, SOD-YOLO has a 15.9% and 9.8% increase in mAP, respectively. In addition, SOD-YOLO achieves a 7.5% and 5.3% improvement compared to the baseline models YOLOv8n and YOLOv8s, respectively.

In addition, the inference speed is also an important index to evaluate the model performance. Among them, although YOLOv3 has a high detection precision, the large number of complex parameters in its model results in a detection speed of only 24.4 frames/s, which cannot reach the real-time detection requirements. YOLOv3-tiny FPS reaches 120.1, the fastest inference speed, but its detection accuracy cannot reach the detection requirements. SOD-YOLO's detection speed reaches 88.7 frames/s, compared with YOLOv5n, YOLOv6n, and YOLOv8s, which are improved by 17, 7.6, and 10.1, respectively. The baseline model YOLOv8n achieves an inference speed of 93.2 frames/s, but it has a large disadvantage in detection accuracy compared to SOD-YOLO.

Combining the two evaluation metrics, the results demonstrate the superior performance of our SOD-YOLO model.

(2) To verify the validity of SODL, RCSOSA, and Wise-CIoU in the SOD-YOLO model, we conducted ablation experiments utilizing the dataset presented in this paper. Where S-YOLO denotes the model with only the SODL module, SR-YOLO denotes the model with the SODL and RCSOSA modules, and SOD-YOLO denotes the model with SODL, RCSOSA, and Wise-CIoU. The experimental results are presented in Table 3.

Table 3. Comparison of performance of various models.

Model	SODL	RCSOSA	Wise-CIoU	mAP@0.5 (%)	Parameters (M)	GFLOPs	FPS
YOLO v8n	×	×	×	82.6	3.0	8.1	93.2
S-YOLO	✓	×	×	87.9	2.9	12.2	77.0
SR-YOLO	✓	✓	×	89.0	3.4	21.9	88.6
SOD-YOLO	✓	✓	✓	90.1	3.4	21.8	88.7

In Table 3, "✓" indicates that the module is used and "×" indicates that the module is not used.

The results show that the mAPs of S-YOLO, SR-YOLO, and SOD-YOLO are higher than that of the baseline model YOLOv8n, reaching 87.9, 89.0, and 90.1, respectively. Among the experiments, the S-YOLO model exhibited an improvement in mAP by 5.3%, which indicates that the addition of the SODL module increases the sensitivity to small targets in the dataset, which is favorable to the improvement of the precision of the small object detection. However, the inference speed of the S-YOLO module is 77.0 frames/s, marking a decrease of 16.2 compared to YOLOv8n. This substantial reduction in inference speed leads to a notable impairment in model efficiency. SR-YOLO adds the RCSOSA module based on S-YOLO, and the RCSOSA module can significantly reduce the memory occupation and computation in the inference stage, and its inference speed is reduced compared with that of YOLOv8n by only 4.6. Meanwhile, compared with S-YOLO, its average accuracy is improved by 1.1%. SOD-YOLO improves the loss function on the SR-YOLO model by using Wise-CIoU loss instead of the original CIoU loss, which improves the accuracy of the model detection without changing the quantity of model parameters, computation, and detection speed. Wise-CIoU can determine the loss function of the model in real time. CIoU can judge the quality of the sample bounding box in real time, to assign gradients. It can prevent the model from over-emphasizing low-quality samples, thereby enhancing the model's accuracy. Although the detection speed of the enhanced SOD-YOLO algorithm is slightly reduced, it demonstrates a notable improvement in the detection of small targets.

The training progress of various models is depicted through bounding box regression loss curves, which are presented in Figure 7. The figure reveals a consistent downward trend in all the losses. When SOD-YOLO is iterated 125 times, the loss is stable and the network converges. Compared with YOLOv8s, S-YOLO, and SR-YOLO, our SOD-YOLO model converges faster and has lower loss values.

Figure 8 shows the original image and the result map after YOLOv8 and SOD-YOLO model detection, where suspension_clamp, strain_clamp, and shockproof_hammer denote suspension clamp, strain clamp, and shockproof hammer, respectively. From Figure 8a,b, it can be seen that there are 12 and 11 targets in the original image, respectively, and the target distribution is concentrated, which makes detection difficult. There is one missed target shockproof hammer in YOLOv8n, while all targets are detected in SOD-YOLO. In the original image of Figure 8c, there are six targets, and two shockproof hammers are far away from the shooting position, with little pixel information and blurred targets. Five targets are detected in YOLOv8n, with two missed targets and one false detection target shockproof hammer, while the targets are all detected in SOD-YOLO. In addition, the bounding box of the detected targets in SOD-YOLO is more complete to include the whole target and the confidence level of the targets is also improved greatly.

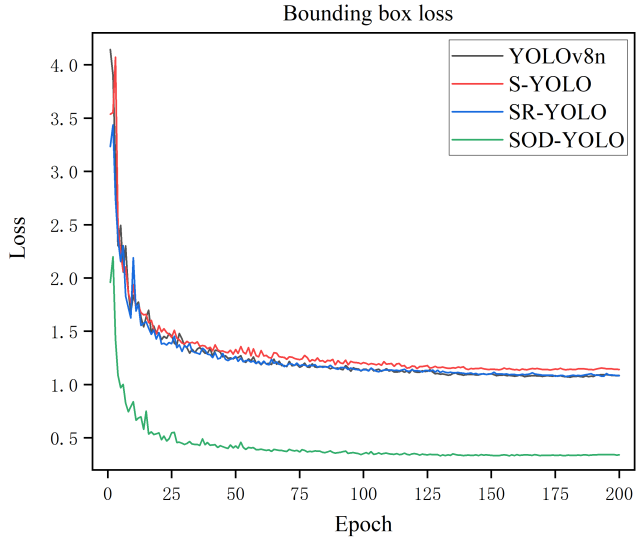

Figure 7. Boundary box regression loss for different model training.

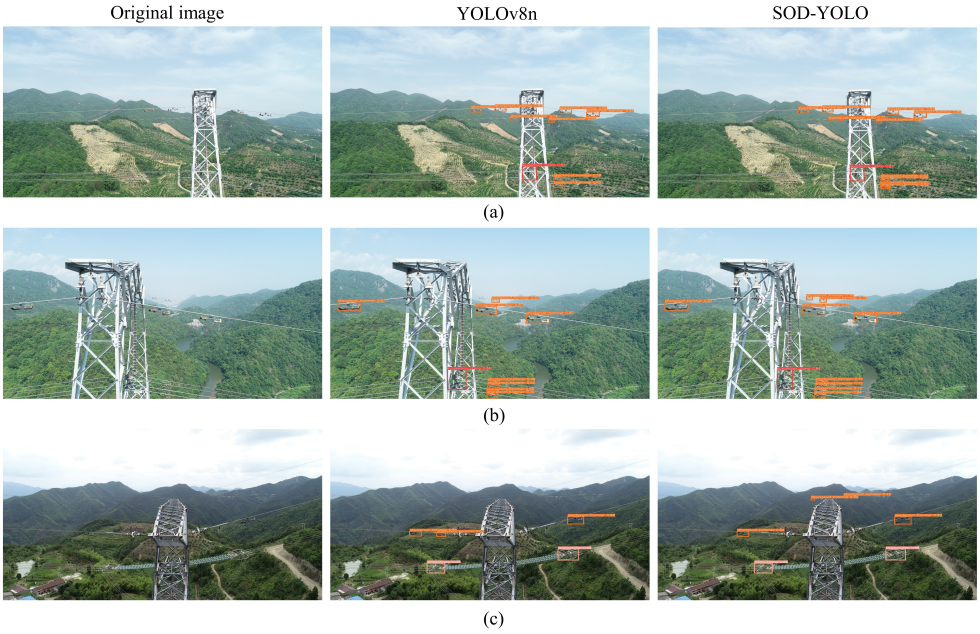

Figure 8. (**a**–**c**) represents three sets of 500 KV high-voltage transmission line inspection images. The first column is the three original images, the second column is the YOLOv8n model detection result image, and the third column is the SOD-YOLO model detection result image.

(3) To demonstrate the generalization ability and superior performance of the SOD-YOLO algorithm, the algorithm of this paper is compared with the current advanced algorithm on the VisDrone2019 dataset using the same training hyperparameters as above. The obtained experimental results of the target category and validation set are shown in Table 4. VisDrone2019 is a dataset of aerial drone photography collected by the AISKYEYE team [35]. The dataset contains 8629 images, of which 6471 are in the training set, 548 in the validation set, and 1610 in the test set. The dataset covers a wide range, and the background is complex and varied. The presence of numerous small targets and overlapping phenomena among them poses a challenge for detection, resulting in a lower mAP, thereby highlighting its significance for further research.

Table 4. Experimental results of different models on the VisDrone2019 validation set.

Model	Target Class (AP/%)										mAP @0.5(%)
	Pedestrian	Person	Bicycle	Car	Van	Truck	Tricycle	A-t	Bus	Motor	
Faster R-CNN [8]	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7	21.7
Cascade R-CNN [36]	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4	23.2
YOLO v3 [14]	18.1	9.9	2.0	56.6	17.5	17.6	6.7	2.9	32.4	17.0	17.6
YOLO v5s [33]	40.8	32.6	13.6	74.6	37.6	32.8	21.9	12.5	44.9	40.0	35.1
MSA- YOLO [37]	33.4	17.3	11.2	76.8	41.5	41.4	14.8	18.4	60.9	31.0	34.7
YOLO v7-tiny	39.6	36.2	9.6	77.5	38.3	30.3	19.4	10.2	49.6	44.5	35.5
YOLO v8n	34.4	27.3	7.2	75.8	38.8	28.1	21.2	11.1	46.6	35.6	32.6
SOD -YOLO	44.1	27.4	11.8	80.5	41.1	31.0	23.9	14.7	49.5	45.0	37.9

In Table 4, "A-t" represents awning tricycle.

Table 4 clearly shows that our SOD-YOLO model's mAP on the VisDrone validation set outperforms the other good models in all categories, reaching 37.9%. Compared to the baseline model YOLOv8n, it improves by 5.3%. The small target categories Pedestrian, Person, Bicycle, Tricycle, Awning tricycle, and Motor improved by 9.7%, 0.1%, 4.6%, 2.7%, 3.6%, and 9.4% compared to YOLOv8n, and the medium and large target categories Car, Van, Truck, and Bus improved by 4.7%, 2.3%, 2.9%, and 2.9% compared to YOLOv8n. The results illustrate that SOD-YOLO can significantly enhance the detection accuracy of small objects while improving the detection precision of medium and large models to some extent.

Based on the aforementioned experimental findings, it is evident that SOD-YOLO has superior detection performance in tiny target detection compared to other models. On the dataset of this paper, the mAP is 90.1% and the FPS reaches 88.7, which meets the demands of real-time detection of small objects.

5. Conclusions

In this paper, aiming for the safe operation of transmission lines, a dataset of aerial photography of UAV inspection under high-voltage towers was constructed, with a total of 3376 RGB images. A SOD-YOLO model for small target detection applicable to transmission lines is proposed. To enhance the model's ability to extract features from small targets, we initially introduced a dedicated small object detection layer into the YOLOv8n model, which, by combining shallow and deep network features, improved the detection precision of the model. Then, the RCSOSA module was incorporated to replace the C2f module in the shallow networks of the backbone and neck. This module uses a simple single-branch structure in the inference phase, which decreased the computational effort of the model and significantly improved the model recognition speed. Finally, to further improve the model detection precision, the CIoU loss was replaced with the Wise-CIoU loss. The loss function effectively addresses the issue of low-quality samples in the dataset by assigning appropriate gradient gains to different samples. This enhancement improved the model's capability in regressing bounding boxes. The experimental findings demonstrate the excellent performance of the SOD-YOLO model. Specifically, the mAP of SOD-YOLO attained a value of 90.1%, the inference speed was 88.7 frame/s, and the parameter count of the model was only 3.4 M, fulfilling the need for real-time detection during UAV inspections. In addition, the model's performance on the VisDrone2019 dataset significantly outperformed other superior models.

Only a relatively small number of targets in the high-voltage transmission line dataset constructed in this paper were defective, and it is not possible to construct a transmission line defect dataset. In future research, we will further take into account the loosening bolts and aging of wire clamps, as well as the missing or damaged vibration-proof hammers. Efficient feature extraction methods will also be further investigated to enhance the detection precision and speed of small objects.

Author Contributions: Conceptualization, K.W. and Y.C.; methodology, K.W.; software, K.W.; validation, K.W., Y.C., and Y.L.; formal analysis, Z.Y.; investigation, K.W.; resources, J.Y.; data curation, Y.L.; writing—original draft preparation, K.W.; writing—review and editing, Y.C.; visualization, Z.Y.; supervision, J.Y.; project administration, E.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Data Availability Statement: Upon reasonable request, the data used in this study can be provided.

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable suggestions on improving this article.

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

References

1. Luo, Y.; Yu, X.; Yang, D.; Zhou, B. A survey of intelligent transmission line inspection based on unmanned aerial vehicle. *Artif. Intell. Rev.* **2023**, *56*, 173–201. [CrossRef]
2. Ballard, D.H. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognit.* **1981**, *13*, 111–122. [CrossRef]
3. Satpathy, A.; Jiang, X.; Eng, H.L. LBP-based edge-texture features for object recognition. *IEEE Trans. Image Process.* **2014**, *23*, 1953–1964. [CrossRef] [PubMed]
4. Ding, L.; Goshtasby, A. On the Canny edge detector. *Pattern Recognit.* **2001**, *34*, 721–725. [CrossRef]
5. Bi, Z.; Jing, L.; Sun, C.; Shan, M. YOLOX++ for transmission line abnormal target detection. *IEEE Access* **2023**, *11*, 38157–38167. [CrossRef]
6. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for small object detection. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part V 13; Springer: Berlin/Heidelberg, Germany, 2017; pp. 214–230.
7. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef] [PubMed]
9. Zhang, K.; Huang, W. Defect detection of anti-vibration hammer based on improved faster R-CNN. In Proceedings of the 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 25–27 September 2020; pp. 889–893.
10. Wenxia, B.; Yangxun, R.; Dong, L.; Xianjun, Y.; Qiuju, X. Defect detection algorithm of anti-vibration hammer based on improved cascade R-CNN. In Proceedings of the 2020 International Conference on Intelligent Computing and Human–Computer Interaction (ICHCI), Sanya, China, 4–6 December 2020; pp. 294–297.
11. Zhai, Y.; Yang, K.; Zhao, Z.; Wang, Q.; Bai, K. Geometric characteristic learning R-CNN for shockproof hammer defect detection. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105429. [CrossRef]
12. Zhou, F.; Wen, G.; Qian, G.; Ma, Y.; Pan, H.; Liu, J.; Li, J. A high-efficiency deep-learning-based antivibration hammer defect detection model for energy-efficient transmission line inspection systems. *Int. J. Antennas Propag.* **2022**, *2022*, 3867581. [CrossRef]
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
14. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
15. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
16. Renwei, T.; Zhongjie, Z.; Yongqiang, B.; Ming, G.; Zhifeng, G. Key parts of transmission line detection using improved YOLO v3. *Int. Arab J. Inf. Technol.* **2021**, *18*, 747–754. [CrossRef]
17. Guo, J.; Xie, J.; Yuan, J.; Jiang, Y.; Lu, S. Fault Identification of Transmission Line Shockproof Hammer Based on Improved YOLO V4. In Proceedings of the 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA), Nanjing, China, 25–27 June 2021; pp. 826–833.
18. Yuan, J.; Zheng, X.; Peng, L.; Qu, K.; Luo, H.; Wei, L.; Jin, J.; Tan, F. Identification method of typical defects in transmission lines based on YOLOv5 object detection algorithm. *Energy Rep.* **2023**, *9*, 323–332. [CrossRef]
19. Di, T.; Feng, L.; Guo, H. Research on Real-Time Power Line Damage Detection Method Based on YOLO Algorithm. In Proceedings of the 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 26–28 May 2023; pp. 671–676.
20. Ding, L.; Rao, Z.Q.; Ding, B.; Li, S.J. Research on defect detection method of railway transmission line insulators based on GC-YOLO. *IEEE Access* **2023**, *11*, 102635–102642. [CrossRef]
21. Yu, C.; Liu, Y.; Zhang, W.; Zhang, X.; Zhang, Y.; Jiang, X. Foreign Objects Identification of Transmission Line Based on Improved YOLOv7. *IEEE Access* **2023**, *11*, 51997–52008. [CrossRef]

22. Li, M.; Ding, L. DF-YOLO: Highly Accurate Transmission Line Foreign Object Detection Algorithm. *IEEE Access* **2023**, *11*, 108398–108406. [CrossRef]
23. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8440–8449.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
28. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
29. Kang, M.; Ting, C.M.; Ting, F.F.; Phan, R.C.W. RCS-YOLO: A fast and high-accuracy object detector for brain tumor detection. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 600–610.
30. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
31. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
34. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
35. Aktaş, M.; Ateş, H.F. Small object detection and tracking from aerial imagery. In Proceedings of the 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 15–17 September 2021; pp. 688–693.
36. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
37. Su, Z.; Yu, J.; Tan, H.; Wan, X.; Qi, K. MSA-YOLO: A Remote Sensing Object Detection Model Based on Multi-Scale Strip Attention. *Sensors* **2023**, *23*, 6811. [CrossRef]