A Robot Foreign Object Inspection Algorithm for Transmission Line Based on Improved YOLOv5

Zhenzhou Wang[1], Xiaoyue Xie[1], Xiang Wang[1](✉), Yijin Zhao[2], Lifang Ma[1], and Pingping Yu[1]

[1] College of School of Information Science and Engineering,
Hebei University of Science and Technology, Shijiazhuang, Hebei, China
wangxiang@hebust.edu.cn

[2] College of Feduni Information Engineering Institute,
Hebei University of Science and Technology, Shijiazhuang, Hebei, China

Abstract. Aiming at the problems of slow detection rate and low accuracy of traditional transmission line inspection methods, a transmission line target detection model based on improved YOLOv5 is proposed in this paper. Firstly, the Bottleneck module in the Backbone network is replaced to improve the lightweight of the model; then the coordinate attention (CA) module is introduced to design the Backbone network to improve the performance of model detection; finally, the frame regression loss function is changed to improve the accuracy of detection. After the transmission line images are further expanded, the foreign object data sets of transmission line are constructed. Experiments on the above data sets show that: Compared with YOLOv5, the detection accuracy of the optimized model is improved by 6.7%, the mean average precision (mAP) reaches 87.0%, and the detection speed is improved by 16.0%. The YOLOv5 lightweight model proposed in this paper reduces the power consumption of the platform and improves the model detection speed and accuracy. It is more conducive to the deployment of the target detection model in the mobile terminal.

Keywords: Transmission line inspection · Target detection · YOLOv5 · The lightweight of the model · The coordinate attention

1 Introduction

Transmission lines are characterized by large capacity, long transmission distance, wide coverage and large demand. In the open-air environment for a long time, foreign bodies like kites, balloons and falling plastic films will hang on the transmission lines. This has a great impact on the normal transmission of power, so it is very necessary to inspect the transmission line. The traditional manual inspection method has low efficiency and complicated working process, and the result is easily affected by the external environment. It has high requirements for the professional ability of the staff and it is unable to meet the daily inspection needs of the transmission line. With people's attention to the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
Y. Xu et al. (Eds.): ML4CS 2022, LNCS 13657, pp. 133–147, 2023.
https://doi.org/10.1007/978-3-031-20102-8_11

safety of high voltage transmission lines and the development of science and technology, robot inspection is widely used to inspect high-voltage transmission lines.

In order to realize the detection of transmission lines, many scholars have tried many ways, in the early days, mainly through the traditional image processing methods. Reference [1] showed a fusion algorithm based on contour feature and gray similarity matching, which realized high-precision insulator contour extraction and accurate separation of insulator pieces, and established an insulator defect detection model based on insulator piece spacing and gray similarity. Reference [2] used the moving edge technology, designed the edge data processing layer, extracted the image features of transmission line inspection by image projection, and abstracted the obtained features into a two-dimensional plane to realize the image recognition of power grid transmission line inspection. However, the image feature extraction process of the above methods was very complex and slow, and the algorithm performance was easily affected by the geographical background difference of transmission lines and various weather, so the generalization ability was not high.

With the improvement of computer hardware level, deep learning algorithm also has rapid development. At present, the representative target detection algorithms are roughly divided into two categories: one is the two-stage algorithm based on candidate regions. The main representative algorithms are Region Convolutional Neural Network (RCNN) [3], Faster-RCNN [4], etc. After generating candidate regions that may contain detection targets, Convolutional Neural Networks (CNN) [5] is used to classify and regress the candidate blocks, and then the detection frame is obtained. Reference [6] showed a new layered architecture of a deep convolution neural network to locate and detect insulator defects. The cascaded network used CNN which based on a layered network to transform the defect detection problem into a two-level target detection problem. Reference [7] firstly used Faster-RCNN to quickly locate the insulator and then classified the location area, finally semantic segmentation judges whether the insulator is burst. Although the detection accuracy of the two-stage algorithm is high, this method firstly extracts the potential location of the target through the candidate area generation network, then carries out detection and recognition. As a result, the detection speed is slow and the real-time effect cannot be achieved. Therefore, the performance of target detection algorithm in embedded devices is greatly limited.

The other is a single-stage algorithm based on regression calculation. Representative algorithms include Single Shot MultiBox Detector (SSD) [8], You Only Look Once (YOLO) [9], YOLOv3 [10] and YOLOv5 [11]. This kind of algorithm uses the idea of regression to directly return the position of the target frame and the target category at multiple positions of the image. In recent years, scholars have applied it to the detection of different objects. Reference [12] designed a fault method based on single-stage SSD detection algorithm for insulator with multi-level perception in aerial images. Reference [13] completed the training of the insulator database by using the YOLO network model and achieved a good recognition effect. Reference [14] proposed an improved YOLOv3 model, which used SPP network and multi-scale prediction network to improve the accuracy of insulator fault detection.

To sum up, in this paper, the YOLOv5 recognition model is improved and applied to the field of transmission line inspection. Firstly, Ghost Bottleneck module is replaced with the Bottleneck module in the YOLOv5 network [15]; then the coordinate attention (CA) module [16] is introduced; and finally, efficient intersection over union loss (EIOU Loss) is used as the loss function. On the basis of ensuring the lightweight of the detection model, the high detection speed is considered. The inspection robot will take a large number of real-time pictures of high-voltage lines, and accurately identify and analyze these pictures by using the deep convolutional neural network in the training stage. Then, the deep semantic features of line foreign objects are automatically extracted in the test stage, which can reduce the rate of false detection and missed detection.

2 YOLOv5 Network Structure

In June 2016, Joseph Redmon et al. [13] proposed the YOLO deep learning framework. The YOLOv5 network structure inherits the extremely high detection speed of previous versions of YOLO, by introducing the concept of residual block in the residual neural network, the hierarchical relationship of the network and the logic of existing prediction modules are dynamically optimized to improve the performance for detecting small targets. YOLOv5 has four network structures, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x [5]. Due to the differences in the depth and width of the network structure, the size and precision of the four versions of the model increase successively. The research content of this paper is transmission line inspection. Compared with the detection accuracy, it has higher requirements for the detection speed of network structure, therefore, this paper selects YOLOv5s as the network structure. The YOLOv5 network structure is mainly divided into four parts: Input, Backbone, Neck, and Prediction [17], as shown in Fig. 1.

The Input network mainly includes data enhancement, adaptive picture scaling, adaptive anchor box calculation and other functions. Mosaic data enhancement is mainly carried out by randomly selecting four images for random scaling, distribution and splicing. The detection data set is enriched, the robustness of the algorithm is strengthened, and the model can better detect small target objects in the image. A small number of black edges are added to the original image, which are uniformly scaled to a standard size of $608 \times 608 \times 3$ and then sent to the neural network. Adaptive anchor frame calculation is trained on the initial anchor frame to get the prediction frame, calculate the gap between the prediction frame and the real frame, the gap between the prediction frame and the real frame can be calculated, then the network update parameters in reverse, which can make the pre-diction result more reasonable.

The Backbone network is mainly used to extract network features, containing Focus, CSP [18], SPP [19], etc. The Focus module performs four slice operations and a convolution operation of thirty-two convolution kernels. The original image of $608 \times 608 \times 3$ becomes the characteristic diagram of $304 \times 304 \times 32$, which reduces the loss of feature information caused by convolution. The CSP module solves the problem of large computation of network structure. It is mainly used to extract the network features in the samples and fuse the feature information of different layers to form a richer feature map. The SPP structure is a spatial feature pyramid pool structure, which is mainly used to expand the receptive field, integrate local and global features, and enrich the information of the feature map.

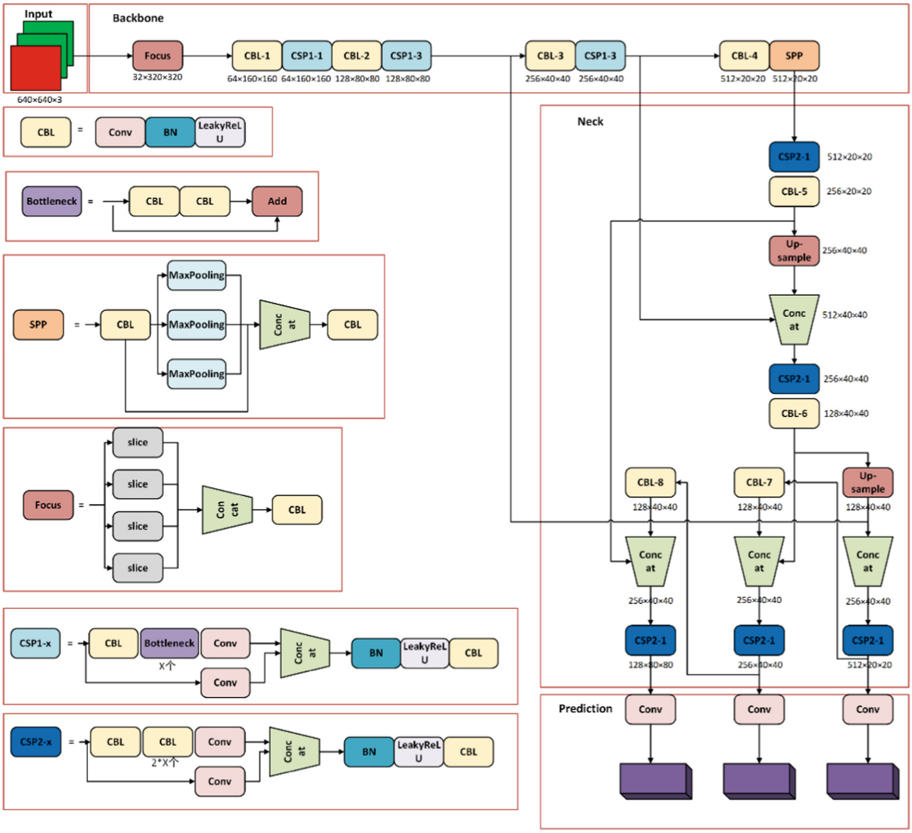


Fig. 1. YOLOv5 network structure.

The Neck network is mainly used to fuse different size feature maps and extracts high-level semantic features. It adopts the structure of FPN+PAN [20]. The FPN structure transmits the high-level feature information of the detected target from top to bottom through down-sampling. Then, the feature pyramids of two PAN structures are used for up-sampling, the shallow features are transmitted from bottom to top, and the information is transmitted to the prediction layer.

The Prediction network mainly includes loss function and Non Maximum Suppression (NMS) [21]. GIOU Loss [22] is used as a loss function in YOLOv5, which increases the measurement of the intersection scale and solves the problem of disjoint boundary boxes that IOU Loss [23] cannot handle. The NMS is used to filter the optimal target box, which solves the problem that a target has multiple candidate boxes. The NMS performs non-maximum suppression on the last detection box of the target, selects the prediction box with high score for retention, and removes the corresponding candidate box with low score.

3 Improvement of YOLOv5 Network Structure

The backbone network in YOLOv5 algorithm has more Bottleneck structures, and the convolution kernel in convolution operation contains a large number of parameters, which increases the deployment cost of the model. In this paper, the lightweight Ghost Bottleneck module is used to replace some standard volume layers in the YOLOv5 backbone network. Since the YOLOv5 network structure uses the same weighting method to extract feature information of different importance, there is a problem of no attention preference in the extraction process of the position coordinates and categories of the regression target frame in the output layer. By establishing the interdependence between channels, The CA module can adaptively calibrate the corresponding characteristics between channels. It is difficult for GIOU Loss to optimize the prediction frame in the horizontal or vertical direction, resulting in slow convergence. Therefore, EIOU Loss with better performance is adopted in this paper, the improved network model is called YOLOv5-GCE.

The improved process is divided into three steps and the steps are as follows. The improved YOLOv5 network structure is shown in Fig. 2 and the algorithm flow is shown in Fig. 3.

1. The residual components of the first CSP1-1 structure in the Backbone part of YOLOv5 are replaced with one Ghost Bottleneck module. The second and third CSP1-3 residual components in the Backbone of YOLOv5 are replaced with three Ghost Bottleneck components, which can reduce the network scale and improve the calculation speed;
2. On the basis of step 1, the attention mechanism is embedded into the CSP module of the Backbone network to help the model to better extract the features of the target of interest, which well obtains the global receptive field and encodes accurate location information, greatly enhancing the accuracy of model training.
3. On the basis of step 2, EIOU Loss is used to improve the accuracy of detection.

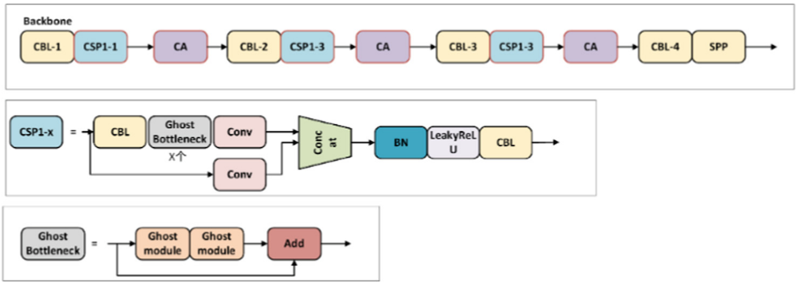


Fig. 2. Improved network structure diagram.

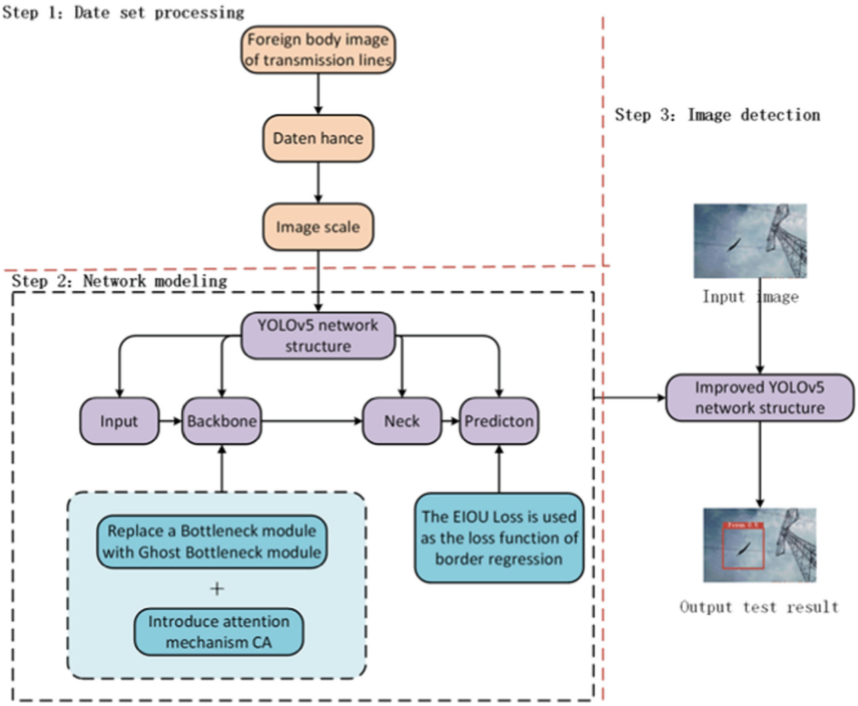

Step 1: Date set processing
Foreign body image of transmission lines
Daten hance
Image scale
Step 2: Network modeling
YOLOv5 network structure
Input
Backbone
Neck
Predicton
Replace a Bottleneck module with Ghost Bottleneck module
+
Introduce attention mechanism CA
The EIOU Loss is used as the loss function of border regression
Step 3: Image detection
Input image
Improved YOLOv5 network structure
Output test result

Fig. 3. Flow chart of improved YOLOv5 model.

3.1 Ghost Bottleneck

The original Bottleneck module in YOLOv5 generates too many redundant feature maps in the process of feature extraction, which not only occupies hardware memory, but affects the running speed of the network. In order to further reduce the demand for hardware resources, this paper uses the idea of the Ghost Net [16] structure for reference and replaces the heavy Bottleneck module in YOLOv5 with the Ghost Bottleneck module. Compared with the direct conventional convolution, the computation of Ghost convolution is greatly reduced, and only a simple linear transformation can produce most of the feature information. The module is mainly composed of two Ghost modules stacked. The Ghost module uses fewer parameters and low-cost linear operation to generate rich feature diagrams. Its principle is shown in Fig. 4 below.

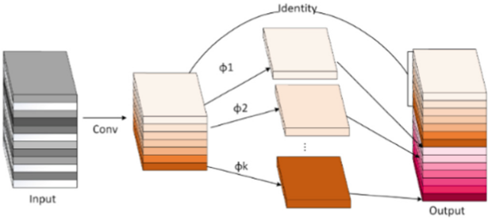


Fig. 4. Ghost module.

As can be seen from Fig. 4 above, the implementation of the module is divided into two parts. One part is obtained by ordinary convolution, the other part is generated by linear operation, and finally the two groups of feature maps are spliced together in the specified dimension. In this case, obtaining the same number of feature maps is only half of the computation. Ghost structure operation can be expressed as:

$$Y' = X * f' + b \tag{1}$$

This equation is a traditional convolution layer that outputs a small number of characteristic maps. Where, X is the input characteristic diagram, * is the convolution operation, f' is the convolution kernel of the convolution operation, Y′ is the characteristic diagram of channels, and b is the offset term.

$$y_{ij} = \Phi_{i,j}\left(y_i'\right), \forall i = 1, \cdots, m, j = 1, \cdots, s \tag{2}$$

This equation is a linear transformation operation for generating redundant features, where, y_i is the ith channel feature of Y′, $\Phi_{i,j}$ is the linear operation of the jth Ghost feature map generated by y_i'. It can be seen that each y_i' can generates s Ghost feature maps. Therefore, m*s characteristic graphs can be obtained by linear operation of m feature graphs.

3.2 Coordinate Attention

Due to the irregular shape of the kite and the thin film on the transmission line, the detection rate of foreign matters on the transmission line is not accurate. In recent years, the attention mechanism module has been widely used in the deep neural network in order to improve the performance of the model. Therefore, this paper introduces the CA module to improve the accuracy of the network. By splitting the two-dimensional global pool operation into two one-dimensional coding processes, the attentional mechanism can capture not only the cross-channel information but the direction perception and position perception information, so that the module can locate and recognize the target of interest more accurately. The structure of the CA module is shown in Fig. 5.

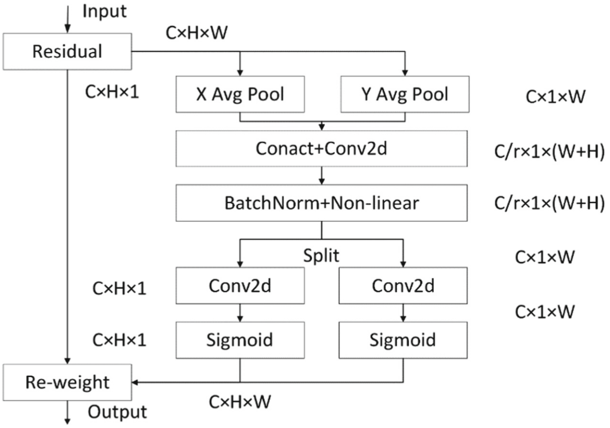


Fig. 5. Coordinate attention.

As can be seen from the above Fig. 5, the CA module firstly divides the input feature map into the horizontal direction and vertical direction. The average pooling of dimensions (H, 1) and (1, W) are used to encode each channel along with horizontal and vertical coordinates respectively. That is the output of the c channel with height H and the c channel with width W. The equation is as follows:

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} x_c(h, i) \tag{3}$$

$$Z_c^{\mathrm{w}}(\mathrm{w}) = \frac{1}{H} \sum_{0 \le i \le H} x_c(\mathrm{j,w}) \tag{4}$$

Then, the two transformations in the above formula are combined with two spatial directions, and two feature maps Z^h and Z^{w} are generated in a cascade. The feature maps in the two directions which obtain the global receptive field are spliced together, the 1 × 1 convolution module is used to transform them:

$$f = \delta(F_1([Z^h, Z^w])) \tag{5}$$

In this equation, [,] is the splicing operation along with the spatial dimension, δ is a nonlinear activation function, and f is the intermediate feature mapping of spatial information encoded in horizontal and vertical directions.

Along the dimension of space, it will decompose into two separate tensors f^{h} and f^{w}. The feature graph f is transformed by 1 × 1 convolution which obtains the characteristic graphs F_h and F_{w} with the same number of channels as the original. The attention weights g^{h} and g^{w} in height and width of feature images are obtained by Sigmoid activation function σ, and the equation is as follows:

$$g^h = \sigma(F_h(f^h)) \tag{6}$$

$$g^w = \sigma(F_w(f^w)) \tag{7}$$

Finally, the final feature map with attention weight in the width and height direction is obtained through multiplication weighting calculation on the original feature map, and its final output is shown in Eq. (8):

$$Yc = (i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

3.3 Loss Function

As the loss function of YOLOv5 network structure, GIOU Loss improves the existing problems when the real frame does not intersect with the prediction frame, but when the prediction frame is in the real frame of the target, its loss value will not change and its relative position relationship cannot be distinguished. It is also difficult to optimize the prediction frame in the horizontal or vertical direction, resulting in slow convergence. Therefore, EIOU Loss is introduced which clearly measures the differences of three geometric factors in the boundary box, namely overlapping area, center point and side length. EIOU Loss disassembles influence factors of aspect ratio on the basis of CIOU Loss and calculates the length and width of the target frame and anchor frame respectively. The loss function includes three parts: overlap loss, center distance loss and width height loss. The width and height loss minimizes the difference between the width and height of the target box and the anchor box, which make the convergence faster. EIOU Loss is shown in Eq. (9):

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \tag{9}$$

In this equation, c_w and c_h are the width and height of the smallest external frame covering the prediction frame and the real frame. Therefore, EIOU Loss frame regression loss function with better performance was adopted in this paper.

4 Experimental Results and Comparative Analysis

4.1 Experimental Data Set

In this paper, 340 transmission line images are taken in real-time by camera. It includes single target images and multi-target images, close-up images, long-range images and complex sample data in different weather, different time and different places. Due to a large amount of background information and noise in the sample images, it is a great challenge to the target detection network. In order to enhance the generalization ability of the model and make the model learn deeper feature information, firstly the data set is expanded. By rotating, clipping or converting brightness and contrast based on the original image data, the target detection image has different forms and scales. Finally, a total of 1500 data sets are obtained, and the training sets, test sets and verification sets are distinguished according to the ratio of 8:1:1. Some data samples are shown in Fig. 6.

Fig. 6. Partial data samples.

LabelImg, an annotation tool for image data set, is used to select the target line image, including the category and position of the target frame, and generates the corresponding XML file format for training and testing. After labeling, the sample data sets are made into the standard PASCAL VOC2012 format according to the data sets format of YOLO series algorithm. There are 2000 instances in 1500 images.

4.2 Experimental Environment

The experimental environment in this paper is shown in Table 1 below.

Table 1. Experimental environment.

Operating system	CPU	GPU	Memory	CUDNN	CUDA	Frame platform	Compiling language
Windows x64	i5-12500H	RTX3050	16 GB	7.6	10.1	Pytorch 1.6	Python 3.8

During all training and testing, the model parameters are configured as follows: the initial learning rate is 0.01 and the weight-decay is 0.0001, the momentum is 0.9, and the batch size is set to 32. According to the multiple training result of the model, it is found that the effect is the best when the epoch setting is 270.

4.3 Evaluation Criterion

The performance and detection effect of the experimental model in this paper is evaluated by the following five aspects: the recall(R), the precision (P), and mAP @0.5 [24], model size and detection speed frames per second (FPS) [9]. The specific calculation process is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TP}{N} \tag{10}$$

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

$$\mathrm{mAP} = \frac{1}{c}\sum\nolimits_{i=1}^{c}\int_{0}^{1} p(r)dr \tag{12}$$

4.4 Comparative Experiment

Data Set Experimental Results. In deep learning, the loss function value can reflect the error between the final prediction result of the target detection model and the actual real value, and it also can be used to analyze and judge the advantages and disadvantages of the training process, the convergence degree of the model and whether it is over fitted. The original model and the improved model are trained with epochs on the same data set for comparative analysis to verify whether the improved YOLOv5-GCE model can improve the performance of the network model. The transformation curves of loss function value with epoch are shown in Fig. 7.

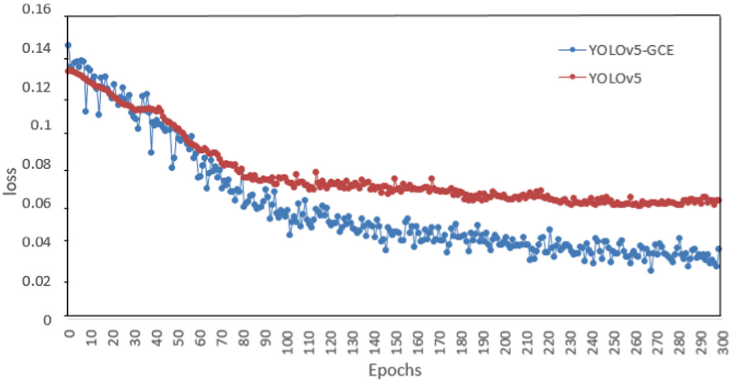


Fig. 7. Variation curve of loss function value with the number of training rounds.

As can be seen from Fig. 7, during the training process, both of them decline rapidly in the early stage and eventually level off, but the loss function of the YOLOv5-GCE model declines faster than the original YOLOv5 model. The GIOU Loss curve of the network model before and after the improvement tends to be stable when they train to 270 epochs, and the model converges when the loss value reaches 0.02. The improved model has a lower loss function value under the same number of training rounds, indicating that the improved model has less loss of details and stronger feature learning ability.

Training and evaluation parameters usually reflect the model training process and the effect of target detection. In this paper, the value of mAP@0.5 is used to judge the model performance, and the performance changes are shown in Fig. 8 below.

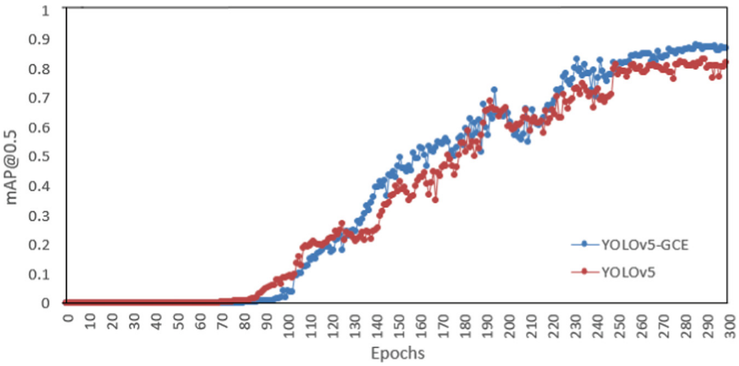


Fig. 8. mAP change curve.

As can be seen from Fig. 8, the mAP parameters of YOLOv5 and the improved model YOLOv5-GCE fluctuate in different degrees, and the overall smooth rise gradually converges at 270 epoch. After the parameter values of the YOLOv5-GCE model are relatively stable, they are higher than YOLOv5. It can be seen that the improved optimization model is significantly better than the original model YOLOv5.

Comparative Experimental Analysis. For the evaluation of model performance, ablation experiments are carried out on different models to verify the performance of different network structures by comprehensively considering the three aspects of mAP, FPS and model size. The ablation Experiment of the four network models is shown in Table 2.

Table 2. Ablation experiment.

Experiment	Ghost	CA	EIOU	R	P	mAP	FPS	Model size
1				82.4	80.8	81.5	36.3	12.5
2	√			81.5	79.4	80.2	43.5	8.8
3	√	√		86.0	85.6	85.8	40.5	9.1
4	√	√	√	87.2	86.9	87.0	42.1	9.7

It can be seen from the experimental data in Table 2 that after replacing the Bottleneck module in YOLOv5 with the Ghost Bottleneck module, the size of the model is only 70.4% of the original model, and the detection speed is 43.5 f/s. Compared with the original model, it is improved obviously. But at the same time, the accuracy of the model is flawed. In experiment 3, after adding the CA module to the network module, the loss of the data sets in the training process are reduced and more learning features are retained. The mAP is improved by 7.0% based on the model of the experiment 2, but it has no significant impact on the detection speed. In experiment 4, although the detection speed decreased slightly, the mAP was improved after replacing the frame regression loss function. To sum up, compared with other structures, the lightweight structure in this paper can achieve higher accuracy, better reasoning speed and better model performance.

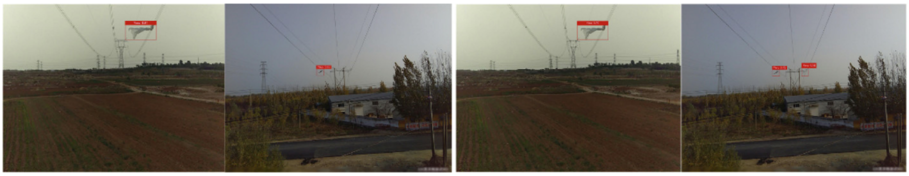
In order to further verify the efficiency of the network proposed in this paper, the classical network structures of YOLOv3 and YOLOv4 are added for multiple comparisons. They are trained with the same amount of epochs, and the results are shown in Table 3 below:

Table 3. Comparison of test results of different network structures.

Model	R	P	mAP	FPS	Model size
Faster RCNN	84.6	81.7	83.5	14.7	48
SSD	73.3	70.9	72.6	30.3	16.1
YOLOv3	74.3	72.4	73.8	31.5	16.6
YOLOv4	80.0	78.9	79.1	34.3	14.7
YOLOv5	82.4	80.8	81.5	36.3	12.5
YOLOv5-GCE	87.2	86.9	87.0	42.1	9.7

By analyzing the dates in Table 3, the detection accuracy of the YOLOv5-GCE model has absolute advantages over SSD, YOLOv3 and YOLOv4, and has been improved compared with the YOLOv5 network. The SSD algorithm has a detection speed of 30.3 FPS, but its average accuracy is low. The two-stage detection algorithm Faster R-CNN has a higher average accuracy of 83.5%, but the detection speed is the lowest. Not only the mAP is improved by 5.4%, but the model size is only 77.6% of the original model. The detection speed is also greatly improved. FPS is 5.8 higher than YOLOv5, 10.6 and 7.8 higher than YOLOv3 and YOLOv4 respectively.

The actual operation effect of YOLOv5s and YOLOv5-GCE is shown in Fig. 9 below.



(a) YOLOv5

(b) YOLOv5-GCE

Fig. 9. Comparison of detection results between YOLOv5 and improved YOLOv5.

By comparing the detection results of different scenes in Fig. 9 (a) and (b), it can be seen that the original YOLOv5 model misses the detection of small targets, while the improved YOLOV5-GCE model misses less. Some edge targets and fuzzy targets are more likely to be detected. At the same time, the improved YOLOV5-GCE model is easier to detect some edge and fuzzy tar-gets. For larger and clearer targets, both the original model and the improved network model can be detected accurately, but the improved network is slightly accurate. This shows that the improved network reduces the loss of image information, obtains more information and improves the integrity of effective information.

5 Conclusion

Aiming at the existing problems in the current transmission line inspection task, this paper proposes an improved line detection algorithm based on the YOLOv5 network structure. The computation is reduced, the model is compressed and the feature information extraction ability of the model is enhanced. Experimental results show that the new model is more simplified and its complexity is significantly reduced, which accelerates the detection speed of the model and improves the adaptability of the model to mobile devices. The method presented in this paper has some limitations in foreign body detection, further research is still needed to improve the detection accuracy on the premise of ensuring the detection efficiency of the algorithm. The YOLOv7 network model will be considered for classification and detection of all kinds of foreign objects to improve the detection accuracy.

Acknowledgments. We acknowledge funding from the Special Project of Cultivating Scientific and Technological Innovation Ability of University and Middle School Students (Grant No. 2021H011404), the Special Project of Cultivating Scientific and Technological Innovation Ability of University and Middle School Students (Grant No. 2021H010203), the Hebei College and Middle School Students Science and Technology Innovation Ability Cultivation Special Project (Grant No. 22E50075D) and the Sub Project of National Key R&D Plan Covid-19 Patient Rehabilitation Training Posture Monitoring Bracelet Based on 4G Network (Grant No. 2021YFC0863200-6).

References

1. Tan, P., Li, X.F., Xu, J.M., Ma, J.E., Wang, F.J., Ding, J., et al.: Catenary insulator defect detection based on contour features and gray similarity matching. J. Zhejiang Univ. – Sci. A: Appl. Phys. Eng. **21**(1), 64–73 (2020)
2. Jalil, B., Moroni, D., Pascali, M., Salvetti, O.: Multimodal image analysis for power line inspection. In: International Conference on Pattern Recognition and Artificial Intelligence, Beijing, pp. 13–17 (2018)
3. Jubayer, F., et al.: Detection of mold on the food surface using YOLOv5. Curr. Res. Food Sci. **4**, 724–728 (2021)
4. Yan, B., Fan, P., Lei, X.Y., Liu, Z.J., Yang, F.Z.: A real-time apple targets detection method for picking robot based on improved YOLOv5. Remote Sens. **13**(9), 1619 (2021)
5. Jalil, B., Leone, G.R., Martinelli, M., Moroni, D., Berton, A.: Fault detection in power equipment via an unmanned aerial system using multi modal data. Sensors **19**(13), 3014 (2019)
6. Tao, X., Zhang, D., Wang, Z., Liu, X., Zhang, H., Xu, D.: Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. Trans. Syst. Man Cybern.: Syst. **5**(4), 1486–1498 (2020)

7. Wang, Y., Wang, J., Gao, F., Hu, P., Li, J.: Detection and recognition for fault insulator based on deep learning. In: 2018 11th International Congress on Image and Signal Processing. Bio Medical Engineering and Informatics, Beijing, pp. 1–6 (2018)
8. Zhao, J.Q., Zhang, X.H., Yan, J.W., Qiu, X.L., Yao, X., Tian, Y.C., et al.: A wheat spike detection method in UAV images based on improved YOLOv5. Remote Sens. **13**(16), 3095 (2021)
9. Perera, R., Guzzetti, D., Agrawal, V.: Optimized and autonomous machine learning framework for characterizing pores, particles, grains and grain boundaries in microstructural images. Comput. Mater. Sci. **196**, 110524 (2021)
10. Chowdhury, P.N., Shivakumara, P., Nandanwar, L., Samiron, F., Pal, U., Lu, T.: Oil palm tree counting in drone images. J. Pre-proof **153**, 1–9 (2021)
11. Ning, Z.X., Wu, X.J., Yang, J., Yang, Y.Q.: MT-YOLOv5: mobile terminal table detection model based on YOLOv5. Conf. Ser. **1978**(1), 012010 (2021)
12. Jiang, H., Qiu, X.J., Chen, J., Liu, X., Zhuang, S.: Insulator fault detection in aerial images based on ensemble learning with multi-level perception. IEEE Access **7**, 61797–61810 (2019)
13. Wang, J.H., Xiao, T., Gu, Q.Y., Chen, Q.: YOLOv5_CSL_F: YOLOv5's loss improvement and attention mechanism application for remote sensing image object detection. In: 2021 International Conference on Wireless Communications and Smart Grid, pp. 197–203 (2021)
14. Liu, J.J., Liu, C.Y., Wu, Y.Q., Xu, H.J., Sun, Z.: An improved method based on deep learning for insulator fault detection in diverse aerial images. Energies **14**(14), 4365 (2021)
15. Han, K., Wang, Y.H., Tian, Q., Guo, J., Xu, C.: Ghost net: more features from cheap operations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, pp. 1580–1589 (2020)
16. Zha, M.F., Qian, W.B., Yi, W.L., Hua, J.: A lightweight YOLOv4-based forestry pest detection method using coordinate attention and feature fusion. Entropy **23**(12), 1587 (2021)
17. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. IEICE Transactions on Fundamentals of Electronics. Communications and Computer Sciences (2019)
18. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Yeh, I.H.: CSP net: a new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Washington, pp. 390–391 (2020)
19. He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **379**(9), 1904–1920 (2014)
20. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Utah, pp. 8759–8761 (2018)
21. Tang, J.L., Liu, S.B., Zheng, B., Zhang, J., Wang, B., Yang, M.K.: Smoking behavior detection based on improved YOLOv5s algorithm. In: The 9th IEEE International Symposium on Next-Generation Electronics, Changsha, pp. 1–4 (2021)
22. Rezatofighi, H., Gwak, N., Gwak, J.Y., Sadeghian, A., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, pp. 658–666 (2019)
23. 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, New York, pp. 779–788 (2016)
24. Wang, X.Z., Wei, J.Y., Liu, Y., Li, J.H., Zhang, Z., Chen, J.Y., et al.: Research on morphological detection of FR I and FR II radio galaxies based on improved YOLOv5. Universe **7**(7), 211 (2021)