Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 7192814, 9 pages
https://doi.org/10.1155/2023/7192814

Research Article

Detection of the Pin Defects of Power Transmission Lines Based on Improved TPH-MobileNetv3

Mengxuan Li, Jingshan Han, Zhi Yang, Bin Zhao, and Peng Liu

China Electric Power Research Institute (CEPRI), Beijing, China

Correspondence should be addressed to Mengxuan Li; formlmx@126.com

Received 20 February 2023; Revised 13 April 2023; Accepted 24 April 2023; Published 5 May 2023

Academic Editor: Gongping Yang

Pins are essential connecting components in power transmission lines. Their extensive use yet leads to frequent defects. Given the small size of a pin and many similar components, the detection of such defects is not ideal, which is a technological problem in the identification and diagnosis of power defects. In response to the large size, complex background, and on-site requirements, such as real-time detection, of power transmission lines, this paper proposes a method to detect pin defects based on TPH-MobileNetv3 (Transformer prediction Head Mobilenetv3). This paper modifies and adds a self-attention layer to MobilNetV3-Small to improve the feature extraction capability of small targets after downsampling. A feature fusion structure with layers of self-attention and a convolutional block attention module (CBAM) is added to the neck network, and a transformer prediction head are added to the head network so that different scale characteristics can be fused and focused from space and channels to strengthen the detection of small targets. Compared with the traditional MobileNetV3, the detection accuracy of the algorithm in this paper has been raised by 24%, as shown in the detection results of measured data. Moreover, compared with the mainstream algorithms with the same detection accuracy, this algorithm not only reduces the model size and significantly enhances detection efficiency but also satisfies the requirement of edge image processing of power inspection.

1. Introduction

Power transmission lines are a vital channel of energy transportation in China. Transmission towers are the main carriers, whose structural safety is crucial for line operation [1]. Pins are an important fitting of power transmission lines, which fix nuts and prevent them from loosening. However, due to long-term exposure and mechanical vibration, pins may fall off or come out, followed by risks, such as loosened nuts and structures. Currently, pin defects are mainly inspected manually and through drones or helicopters. Particularly, drones are widely applied because of the close-up shots of risks, high intelligence, and edge computing. Pins account for a small proportion of images and are typically small targets due to multiple factors. For instance, pins are small. Most pictures taken by drones are high-definition. Additionally, there is a certain safe distance between drones and power transmission lines. Targets whose pixels are smaller than 32×32 or those represent less than 1% of the area are collectively referred to as small targets in the field of image recognition [2]. The diagnosis and identification of such targets can be tricky. In addition, the picture background of power transmission lines is mostly a complex mountainous environment with great environmental interferences and light influences. As a result, there are few research achievements in the detection of pin defects at home and abroad. The paper [3], based on the residual network, ResNet101 [4], adopted the feature pyramid and integrated multiscale features to improve the detection precision of small targets. Last, the K-means algorithm was used to optimize the anchor box and detect pin defects. A Faster-RCNN two-stage detection model was employed in the paper [5]. Regional candidates containing targets were screened out through the region extraction algorithm. Then, the candidates were further selected to obtain bounding boxes and conduct target classification. Higher detection accuracy was achieved at the expense of a lower detection speed. In terms of the paper [6], a deeper network was designed to expand the receptive field based on the residual network. Meanwhile, more shallow semantic information was retained to strengthen the detection of small targets. The detection speed was raised compared to the aforementioned Faster-RCNN model, while the global information fusion was inadequate. The conventional convolutional neural network (CNN) extracts the features of targets through multi-layer downsampling, which cannot fully integrate and utilize such features. Transformer [7] has emerged and has become extensively used in computer vision in recent years. Its attention mechanism boasts a strong feature extraction capability based on its globality and capture of the long-distance dependency of the feature map [8]. Therefore, Transformer makes up for the inadequate feature extraction of the traditional CNN. Thanks to the increase in intelligent applications, lightweight networks that can be deployed to the edge have recently become a popular research direction. Through reduced computation, model pruning, quantization, and knowledge distillation, such networks become lightweight, resulting in faster reasoning. A lightweight network is particularly important in deployment. MobileNetv3 [9] is a lightweight network for mobile devices proposed by Google, featuring a depthwise separable convolution that can dramatically reduce computation [10].

In order to realize the high-efficiency and high-precision detection of pin defects of power transmission lines, this paper introduces an algorithm for the detection of pin defects of power transmission lines, based on TPH-MobileNetv3, by combining the advantages of Transformer and MobileNet. Specifically, a MobileNetv3-based backbone network with the attention mechanism designed to extract target features. The Transformer Encoder is added to the end of the backbone network to focus on targets and reduce computation. Concurrently, a CBAM [11] module with channels and spatial attention is employed to enhance the feature fusion and target focus of MobileNetv3.

2. The Network Structure

The TPH-Mobilenetv3 structure of the network in this paper is shown in Figure 1. MobileNetv3 is classified into MobileNetv3-Large and MobileNetv3-Small. The former is mainly used for high computing resources, while the latter is low. MobileNetv3-Small is adopted as the backbone network in this paper, in that, the algorithm should, subsequently, be deployed on the edge side. The last four layers of the backbone network are replaced with a trans layer. Computing resources can be effectively saved by reducing the size of the feature map of the latter layer. The backbone network takes advantage of the translation invariance of CNN and the global correlation between feature maps of the Transformer.

Three feature maps of different scales are fused in the neck of the network. Because an excessive downsampling ratio can lead to the disappearance of features of small targets and raise the difficulty of localization and classification, bilinear interpolation for upsampling is adopted for the feature maps of layers 7, 9, and 13 to obtain larger feature maps. Context information is obtained through the Transformer structure before the network outputs features, so as to reinforce the detection of indistinguishable targets. Next, the CBAM module is used to intensify space and channel information to better distinguish dense and small targets.

For the head of the network, head of YOLO is used for reference. Target classification and location regression are conducted. The number of channels predicted is $b \times (4 + 1 + c)$, wherein b represents the number of prediction boxes in each feature grid, which is generally 3. 4 stands for the position of the prediction box, (x, y, h, w); 1 the background; c the number of target categories.

Network losses consist of the following three parts: the confidence loss of whether there is a target in the predictive frame, the classification loss of target categories in the predictive frame, and the localization loss of the bounding boxes of the predictive frame and the real box. The overall loss function is shown in the following equation:

$$L = \lambda_1 L_{\mathrm{conf}} + \lambda_2 L_{cls} + \lambda_3 L_{loc}, \tag{1}$$

where L is the overall loss; λ the weight; L_{conf} the confidence loss; L_{cls} the classification loss; and L_{loc} the localization loss.

In terms of confidence loss, binary cross entropy (BCE) is used as the loss function, as shown in the following equation:

$$L_{\mathrm{conf}} = -\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - y^{(i)}\right), \tag{2}$$

where $\hat{y}$ stands for the predicted probability of the i-th sample to be a certain category and $y^{(i)}$ the label of the i-th sample.

The target category loss means the difference between the predicted category and the real label. BCE is used as the loss function, as shown in the following equation:

$$\begin{aligned} L_{cls}(O, C) = -\sum_{i \in pos} \sum_{j \in cls} \left(O_{ij}\right) \ln\left(\mathrm{sigmoid}\left(C_{ij}\right)\right) \\ + \left(1 - O_{ij}\right) \ln\left(1 - \mathrm{sigmoid}\left(C_{ij}\right)\right), \end{aligned} \tag{3}$$

where O_{ij} is the prediction of whether the actual object in the target bounding box i is the current category and C_{ij} the predicted value.

The positioning loss is calculated by Complete-IoU (CIoU), as shown in the following equation:

$$L_{loc} = 1 - CIOU = 1 - \left(IOU - \frac{distance2^2}{distance\_C^2} - \frac{\nu^2}{(1 - IOU) + \nu} \right), \tag{4}$$

where $distance_2$ is the Euclidean distance between the center points of the prediction and real boxes. $distance_C$ stands for the diagonal distance between the smallest circumscribed rectangles between the prediction and real boxes. ν is a parameter to measure the consistency of the length-width ratio. The equation of definition is shown in the following equation:

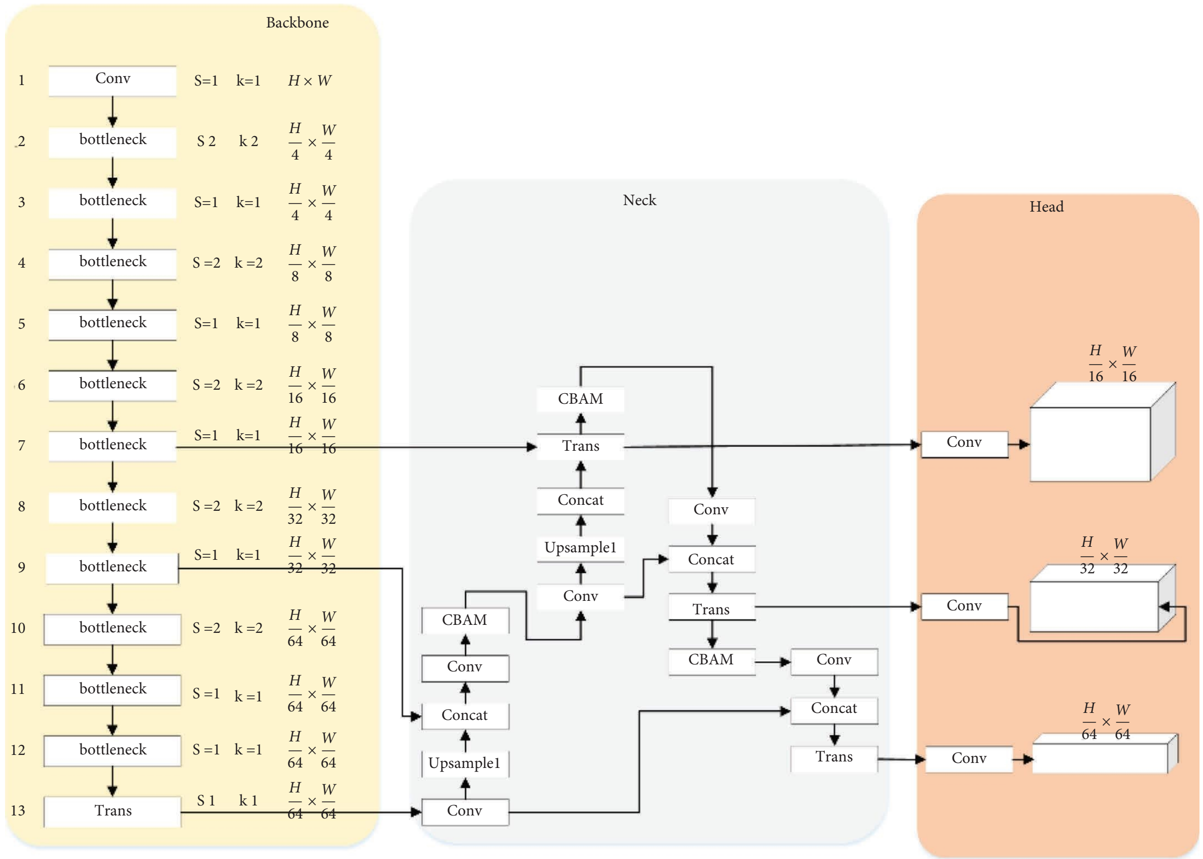

Backbone
1 Conv S=1 k=1 $H \times W$
2 bottleneck S 2 k 2 $\frac{H}{4} \times \frac{W}{4}$
3 bottleneck S=1 k=1 $\frac{H}{4} \times \frac{W}{4}$
4 bottleneck S =2 k =2 $\frac{H}{8} \times \frac{W}{8}$
5 bottleneck S=1 k=1 $\frac{H}{8} \times \frac{W}{8}$
6 bottleneck S =2 k =2 $\frac{H}{16} \times \frac{W}{16}$
7 bottleneck S=1 k=1 $\frac{H}{16} \times \frac{W}{16}$
8 bottleneck S =2 k =2 $\frac{H}{32} \times \frac{W}{32}$
9 bottleneck S=1 k=1 $\frac{H}{32} \times \frac{W}{32}$
10 bottleneck S =2 k =2 $\frac{H}{64} \times \frac{W}{64}$
11 bottleneck S =1 k =1 $\frac{H}{64} \times \frac{W}{64}$
12 bottleneck S =1 k =1 $\frac{H}{64} \times \frac{W}{64}$
13 Trans S 1 k 1 $\frac{H}{64} \times \frac{W}{64}$
Neck
CBAM
Trans
Concat
Upsample1
Conv
CBAM
Conv
Concat
Upsample1
Conv
Conv
Concat
Trans
CBAM
Conv
Concat
Trans
Head
$\frac{H}{16} \times \frac{W}{16}$
Conv
$\frac{H}{32} \times \frac{W}{32}$
Conv
$\frac{H}{64} \times \frac{W}{64}$
Conv

Figure 1: Structure of TPH-MobileNetv3.

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^{p}}{h^{p}}\right)^2, \tag{5}$$

where w^{gt} and h^{gt} refer to the width and height of the real box, while w^p and h^p stand for the width and height of the prediction box.

2.1. Transformer Encoder. Inspired by the excellent performance of Transformers in visual detection, this paper adds a self-attention-based Transformer coding module to the last layer of the MobileNetv3-Small backbone network and the whole network's neck. There are somewhat target feature losses in the backbone network after multiple downsampling. However, the multi-head attention mechanism in the Transformer module can calculate all the correlations between the features in the entire feature map to obtain the global view. Compared with the convolution module, the Transformer module can better extract features and fuse the features of different layers in the feature fusion structure of the neck of the network to capture more target information. The Transformer Encoder module consists of a multi-head attention layer and a multi-layer perceptron (MLP). Each layer is connected by a residual structure. The structure of the Transformer module is shown in Figure 2, and the calculation of attention is shown in the following equation:

$$\text{Attention}(Q, K, V) = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$

where Q, K, and V represent the query, key, and value generated for the input sequence X, respectively. All their dimensions are d_k. The product of Q and K is divided by $\sqrt{d_k}$. The result, after being processed with Softmax, multiplies V to obtain a weighted feature map.

2.2. CBAM Module. A CBAM module with space and channel attention is added to the feature fusion module, which sequentially computes attention maps along two independent dimensions, channel, and space and multiplies the attention maps to optimize adaptive features. CBAM is a lightweight module, whose overhead can be ignored. The structure of the CBAM module in this paper is shown in Figure 3.

3. Experimental Results and Analysis

3.1. The Experimental Environment and Data. The environmental configuration of this experiment included CPU model: Intel Xeon 6240R; GPU model: NVIDIA GeForce RTX 3090; operating system: Ubuntu16.04, CUDA10.0, and Cudnn7.6.5 and a deep learning framework based on PyTorch1.4 and Python3.7.

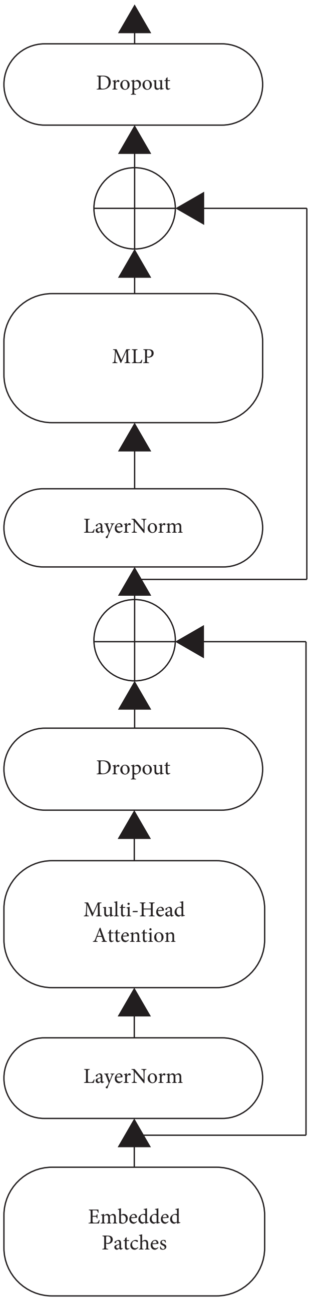


Figure 2: Structure of the Transformer module.

The experimental dataset consisted of 834 aerial images taken with a DJI drone. The pin defects defined in this paper include the following four types: bolt missing pin, bolt off pin, pin installation irregularities, and bolt not tightened. In power maintenance operations, if the above four types of defects are present, they all need to be overhauled, so these four types of images are grouped into the pin defect category, which is distinguished from the normal pin category. There were 1,573 target nuts with pin defects marked as lsqxz. The data were expanded, and then the dataset was randomly divided into the training set, the validation set, and the test set at the ratio of 8 : 1 : 1.

The statistics of the width-height ratios and pixel ratios of targets in the dataset are shown in Figure 4. The statistical graph of widths and heights indicates that 85.3% of the target width-height ratios were concentrated within the range of 1.0 to 1.5. The target detection anchor in this paper was set to (1 : 1, 2 : 3, 3 : 2), respectively. The targets accounted for a small proportion of the whole image, as revealed in the statistical graph of the pixel ratios of targets. In order to obtain a better detection effect, samples in the central point were randomly cut. Reasoning results were postprocessed during the testing and validation phases to obtain the detection result of targets on the original image.

3.2. Model Training. In this paper, Adam, a gradient descent method of adaptive learning rates, was used. The momentum was set to 0.9, and the weight attenuation coefficient was 0.0005. The size of the input image was the original size without conversion. The batch size was set to 16. Epoch was set to 100, and the initial learning rate 0.001. Learning rates were warmed up for initialization and adjusted through cosine annealing.

3.3. Assessment Method of Experimental Results. For the target detection of samples of a single category recall (R), precision (P), and average precision (AP) are commonly used to assess model performance. For the target detection of samples of multiple categories, mean average precision (mAP) and AP50 are generally used to assess model performance.

3.4. Comparison of the Results of Different Algorithms. In order to better demonstrate the superiority of TPH-Mobilenetv3 in the detection of pin defects, the experimental results of other mainstream target detection algorithms currently used in pin defects were compared and tested on the dataset of this paper, which are not necessarily the latest, but they have many on-site applications that can make the comparison results more valuable. For multi-stage networks, Cascade R-CNN [12] was selected. For two-stage networks, Faster-RCNN was selected. For single-stage network, SSD [13], MobileNetv3, YOLOV3 [14], YOLOV4 [15], and RetinaNet [16] were selected. The backbone network of each model, the recall, precision, and average precision of testing, floating point operations per second (FLOPs) used to measure the computation of the complexity of the model, and model parameters, Params, are shown in Table 1.

Table 1 reveals that SSD was poor in detecting small targets due to blurred information fuzziness after multiple downsampling and a lack of feature fusion. Thanks to its feature fusion structure, FPN, Faster-RCNN was significantly better than SSD in the precision of target detection, yet the effect was still not ideal. Similarly, because of FPN, RetinaNet and YOLOV3 had high recall rates. Nevertheless, the precision of RetinaNet was poor. With respect to two-stage networks, Cascade R-CNN had high precision but low average precision. YOLOV4 outperformed Faster-RCNN in average precision because of its data enhancement method and improved PANet feature fusion method. The performance indicators of MobileNetV3 were not prominent, while its computation was advantageous. The recall and

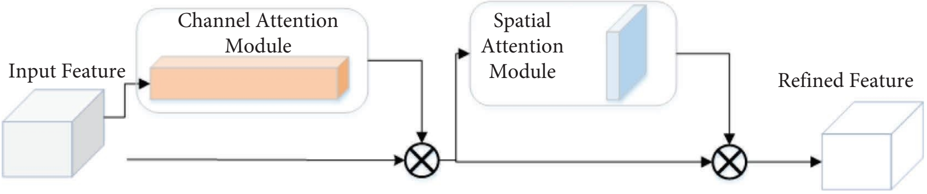
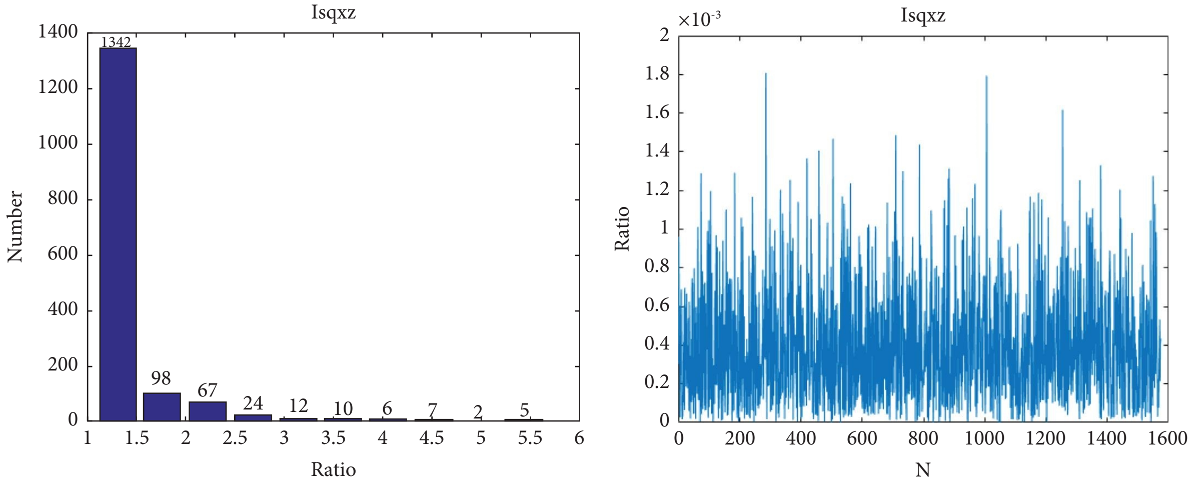


Figure 3: Structure of the CBAM module.



(a)

(b)

Figure 4: Statistics of the width-height ratios and pixel ratios of targets: (a) statistical graph of width-height ratios and (b) statistical graph of the pixel ratios of targets.

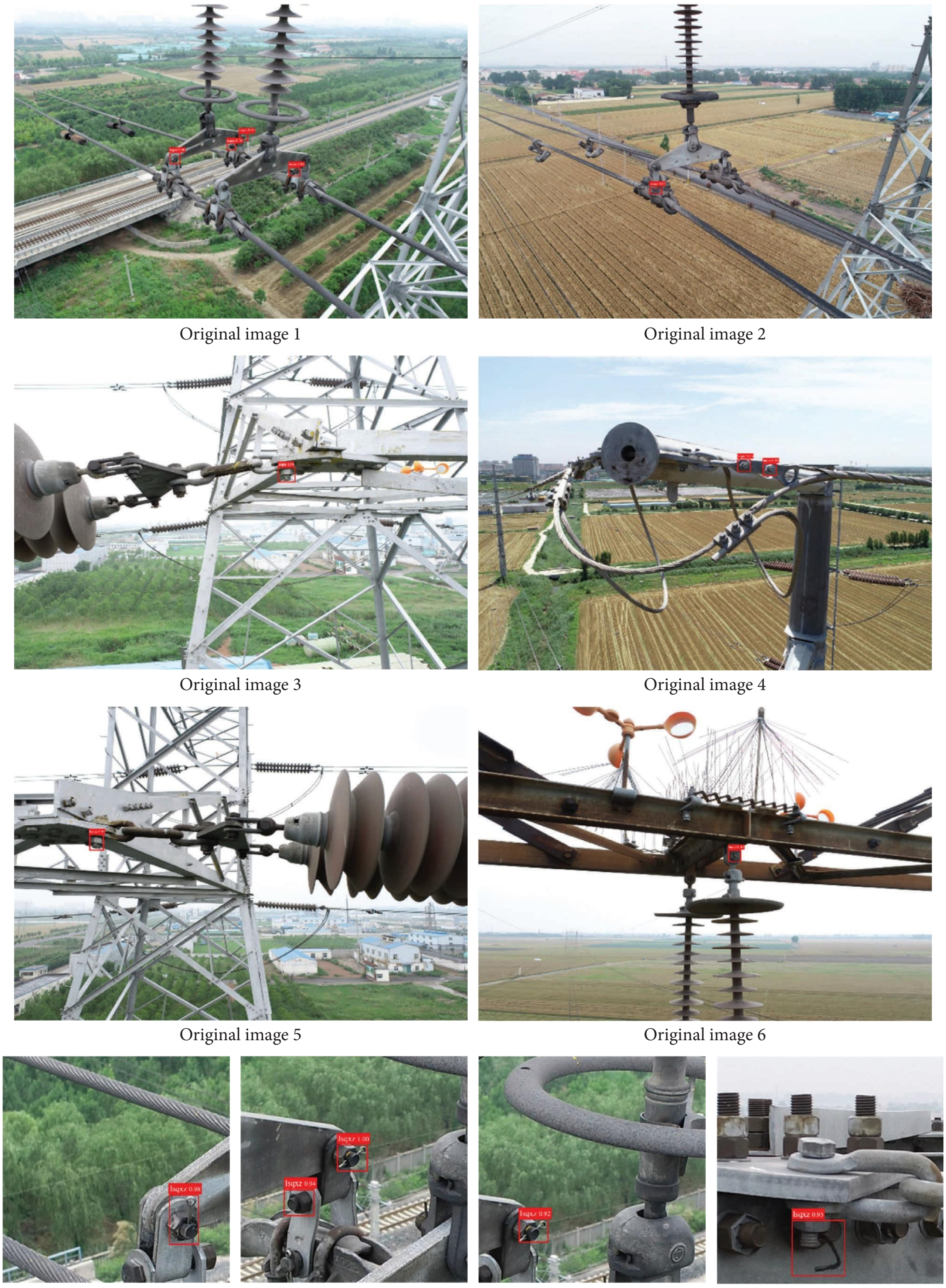
Table 1: Comparison of different algorithms by precision.

Model	Backbone network	Image size	R	P	AP	FLOPs (G)	Params (M)	FPS
SSD	VGG16	512^2	0.743	0.009	0.445	98.81	36.04	6.8
Faster-RCNN	ResNet-101-FPN	512^2	0.740	0.327	0.650	83.13	60.52	4.5
RetinaNet	ResNet-101-FPN	512^2	0.878	0.084	0.634	80.70	56.74	5.3
Cascade R-CNN	ResNet-101-FPN	512^2	0.780	0.481	0.668	110.77	88.16	5.4
YOLOV3	DarkNet-53	512^2	0.699	0.546	0.623	50.06	61.95	6.1
YOLOV4	CSPDarkNet53	512^2	0.823	0.236	0.681	39.85	27.60	5.8
MobileNetV3-small	MobileNetV3	512^2	0.645	0.016	0.472	0.32	2.9	30.1
Ours	MobileNetV3-small	512^2	0.820	0.402	0.715	28.56	18.67	21.3

average precision of the TPH-MobileNetV3 network in this paper were drastically improved as compared with the traditional MobileNetV3 network. Additionally, compared with YOLOv4, model parameters were reduced and FPS performance was greatly intensified.

For the power sector, timely detection of defects in power transmission lines and maintenance are top priorities. Thus, high recall is more important than high precision in that defects can be detected more effectively. In addition, in order to allow the terminal to quickly get detect results, network size is a key indicator. Table 1 implies that the model proposed in this paper balanced recall and precision and had a small size, making it advantageous over other models. Figure 5 shows some original and partially enlarged images of the detection results of TPH-MobileNetV3, from which we can see that the algorithm in this paper could practically detect pin defects from different angles.

In order to analyze the influence of Transformer Encoder and CBAM used in this paper on the detection effect of the model, traditional MobileNetV3-Small, MobileNet-Small + Transformer, and MobileNet-Small + Transformer + CBAM were used to test the same dataset. The test results are shown in Table 2. Due to its feature fusion structure, MobileNetV3-Small had low precision. Because of the Transformer layer added to its target feature extraction structure, MobileNet-Small + Transformer had a better recall, as compared with MobileNetV3-Small. Yet, the effect was not ideal.



Original image 1

Original image 2

Original image 3

Original image 4

Original image 5

Original image 6

Partially enlarged image 1

Partially enlarged image 2

Partially enlarged image 3

Partially enlarged image 4

Figure 5: Original and partially enlarged images of the detection results of TPH-MobileNetV3.

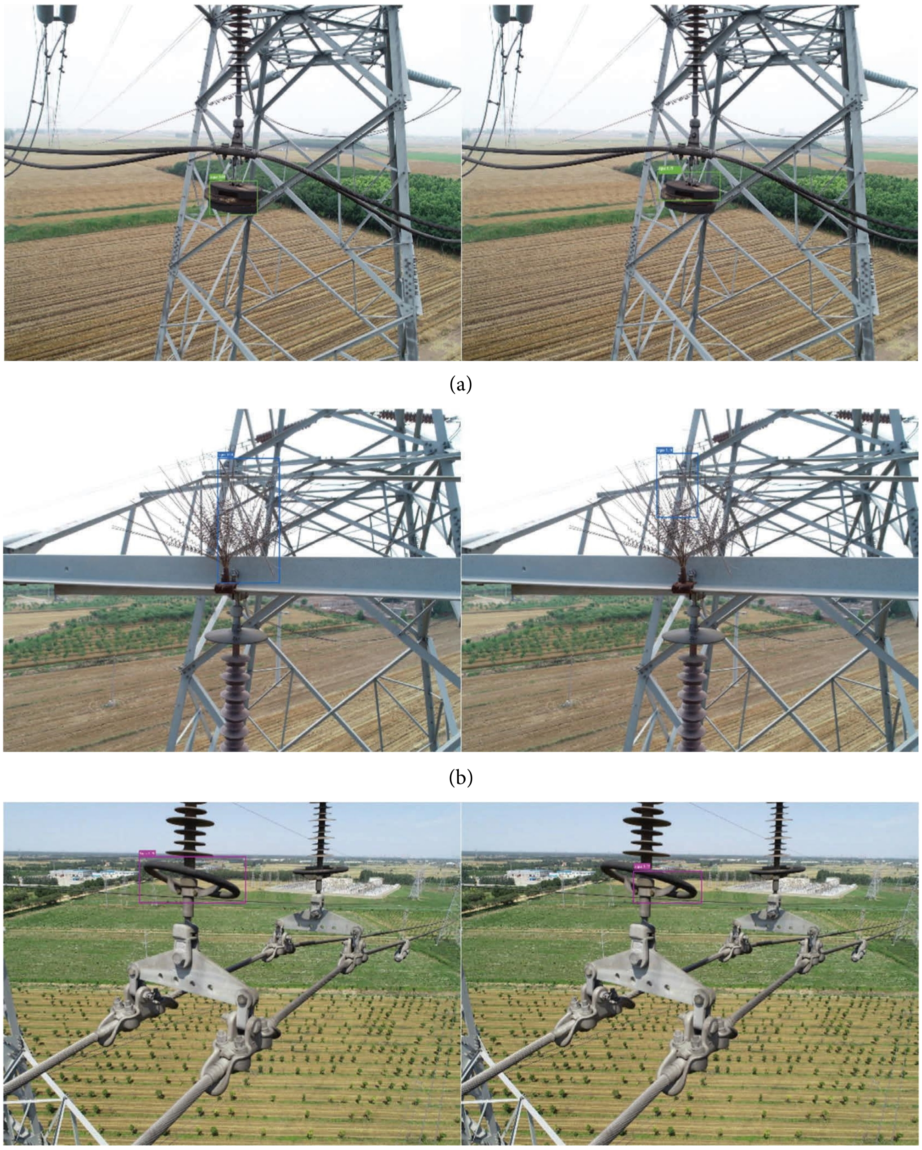
Comparatively, MobileNet-Small + Transformer + CBAM had significantly higher recall and precision, due to the improved backbone network for feature extraction and the attention-based feature fusion pyramid. Moreover, the improved structure was compared with the other structures in terms of the influence on the algorithm reasoning

Table 2: Comparative experiment on the validation set in terms of each modification of the algorithms.

Method	R	P	FPS
MobileNetV3-small	0.637	0.014	30.1
MobileNet-small + transformer	0.765	0.035	18.7
MobileNet-small + transformer + CBAM	0.820	0.402	12.3

Table 3: Comparative test results under the same conditions of faster-RCNN and the algorithm in this paper.

Model	Faster-RCNN		TPH-MobileNetV3	
Category	R	AP	R	AP
Heavy hammer corrosion	0.851	0.813	0.872	0.834
Foreign bodies, such as bird's nests, on transmission towers	0.932	0.849	0.943	0.858
Deformed grading rings	0.632	0.594	0.670	0.653
Separated wire strands	0.750	0.675	0.784	0.694
Improperly installed number plates of transmission towers	0.784	0.702	0.895	0.832
mAP	0.727		0.774	

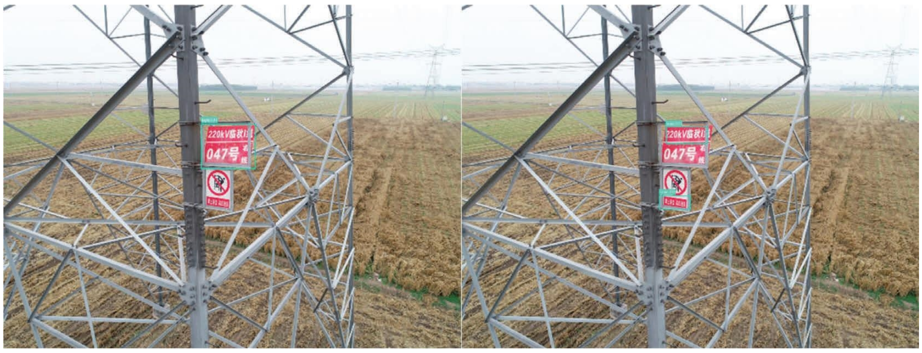


(a)

(b)

(c)

Figure 6: Continued.



(d)

Figure 6: Test results of the test set of TPH-MobileNetV3 and Faster-RCNN: (a) heavy hammer corrosion; (b) foreign bodies, such as bird's nests, on transmission towers; (c) deformed of grading rings; (d) improperly installed number plates of transmission towers.

time. FPS in the table means the frames reasoned per second regarding images with the width and height of 4,864 × 3,648 on RTX 3090. As modifications increased, the reasoning speed decreased and recall and precision were improved.

3.5. Defect Identification Results of Other Small Targets. Furthermore, other defects of power transmission lines were detected with the network proposed in this paper. Such defects fell into the following four categories: heavy hammer corrosion, foreign bodies, such as bird's nests, on transmission towers, deformed grading rings, and improperly installed number plates of transmission towers. A total of 1,576 images were taken, wherein 1,216 images were in the training set and 231 in the test set. The comparison of the test results of Faster-RCNN and the algorithm in this paper under the same experimental conditions are shown in Table 3. The mAP of Faster-RCNN was 0.727 and that of TPH-MobileNetV3 in this paper 0.774. The algorithm in this paper outperformed Faster-RCNN in terms of the detection of the above four types of defects. Moreover, compared with Faster-RCNN, the network model in this paper is smaller, and the processing frame rate is faster, which makes it more suitable for carrying on the edge side or drone side. Figure 6 demonstrates the test results on the test set of Faster-RCNN and TPH-MobileNetV3 in this paper. The first column is the test results of TPH-MobileNetV3, and the second one is Faster-RCNN. The model proposed in this paper had higher precision.

4. Conclusions

In response to the challenge of detecting the small targets of pin defects in the images taken during the inspection of high-voltage power transmission lines, this paper introduces a TPH-MobileNetV3 network with improved MobileNetV3-Small. The MobileNetV3-Small backbone network is modified into a self-attention layer with an attention mechanism so that small target features can still be precisely extracted upon multiple downsampling. Furthermore, the neck network of MobileNetV3 is optimized through the addition of a feature fusion structure with self-attention and CBAM layers to strengthen the detection of small targets. The experimental results reveal that the network proposed in this paper has well balanced recall, precision, and detection speed. Compared with the traditional MobileNetV3-Small, it has prominently raised detection precision despite somewhat increased computation. The recall was raised from 0.645 to 0.820, and precision from 0.472 to 0.715. Compared with network models with similar recall and precision, such as YOLOv4, the size of this paper's network fell from 27.06 M to 18.67 M, and computing power was reduced from 39.85 G to 28.56 G. FPS greatly rose from 5.8 to 21.3. In short, image detection on the edge side can be basically satisfied. In the subsequent research, we plan to combine a variety of defects of high-voltage power transmission lines in small target detection to reinforce the generalization of the network further.

Data Availability

The image data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is funded by the National Key R&D Program (2020YFB0905900).

References

[1] X. D. Gu, D. H. Tang, and X. H. Huang, "Deep learning-based defect detection and recognition of a power grid inspection image," *Power System Protection and Control*, vol. 49, no. 5, p. 7, 2021.

[2] G. Chen, H. Wang, K. Chen et al., "A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 2, pp. 936–953, Feb. 2022.

[3] C. Y. Gu, Z. Li, J. T. Shi, G. Shang, and X. Jiang, "Detection for pin defects of overhead lines by UAV patrol image based on improved faster-RCNN [J]," *High Voltage Engineering*, vol. 46, no. 9, p. 8, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[5] S. Ren, K. He, and R. Girshick, "Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[6] R. S. Li, Y. L. Zhang, D. H. Zhai et al., "Pin defect detection of transmission line based on improved SSD," *High Voltage Engineering*, vol. 47, pp. 3795–3802, 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention Is All You Need," 2017, https://arxiv.org/abs/1706.03762.

[8] X. K. Zhu and Q. W. Jiang, "Research on object detection for UAV images based on CNN and transformer [J]," *Journal of Wuhan University of Technology*, vol. 44, no. 2, p. 9, 2022.

[9] A. Howard, M. Sandler, G. Chu et al., "Searching for MobileNetv3," *Proceedings of the IEEE/CVF International Conference on Computer Vision.*, pp. 1314–1324, 2019.

[10] Y. Y. Wang, S. Luo, and Z. J. Wang, "Remote sensing target detection based on improved MobileNetV3 [J]," *Journal of Shaanxi University of Science and Technology*, vol. 40, no. 3, p. 8, 2022.

[11] S. Woo, J. Park, and J. Y. Lee, "CBAM: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, September 2018.

[12] Z. Cai, N. Vasconcelos, and R.-C. N. N. Cascade, "Cascade R-CNN: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, and A C. Berg, "SSD: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Springer, Amsterdam, The Netherlands, October 2016.

[14] J. Redmon and F. A. Yolov3, "An Incremental Improvement," 2018, https://arxiv.org/abs/1804.02767.

[15] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal Speed and Precision of Object Detection," 2020, https://arxiv.org/abs/2004.10934.

[16] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.