The detection of vibration dampers based on optimized RetinaNet

Mingsheng Ma
Beijing University of Civil Engineering and Architecture
Beijing, China
hmaymms@126.com

De Zhang
Beijing University of Civil Engineering and Architecture
Beijing, China
zhangde@bucea.edu.cn

Feng Liu
Nanjing University of Posts and Telecommunications
Nanjing, China
liuf@njupt.edu.cn

***Abstract*—The vibration damper is an important part that can reduce the vibration of transmission lines. When it loosens, falls off, or deforms, it needs to be replaced in time. Nowadays, unmanned aerial vehicle (UAV) is often utilized to help detect the faults of vibration dampers equipped on high-overhead power lines for the sake of convenience and safety. Moreover, with the continuous development of deep learning technique, convolutional neural network (CNN) based methods has achieved a significant progress in the fields of object detection and localization, defect recognition, etc. For the vibration damper object detection, however, there is no large-scale public dataset available for training. In addition, features such as the varying pattern of vibration dampers, complex backgrounds, and mutual occlusion among key components, limit the performance and actual application of CNN-based detection methods for vibration dampers. Therefore, in this paper we build a large dataset containing 38,295 images collected with UAV and 22,252 annotation information files. In order to detect the vibration dampers accurately, we propose a RetinaNet optimization model based on the FreeAnchor method. The optimization process is implemented from two aspects, including data enhancement and model structure improvement. The experimental results demonstrate that the optimized model can outperform the original RetinaNet model and other classical detection methods, reaching a relatively-high accuracy of 83.1%.**

***Keywords*—vibration damper, object detection, RetinaNet, deep learning, convolutional neural network**

I. Introduction

UAV inspection of vibration dampers gradually replaces the traditional manual inspection with the advantages of high efficiency and safety. The large amount of inspection image data generated by UAV inspections urgently needs to be analyzed and processed. Relying on manual analysis of image data is not only time-consuming and labor-intensive but also prone to omission and misdetection. Therefore, the automated analysis of image data and the use of deep learning methods to improve the effect of fault identification has become one of the hot spots in the research and application of smart power grid operation and maintenance.

In several disciplines, deep learning technology has produced amazing outcomes [1]. Compared with the manually designed features, the features extracted by using deep neural networks to automatically learn high-level features from a large amount of data are richer and more expressive, with accuracy and robustness that traditional detection approaches cannot reach [2-3].

Deep learning-based object detection algorithms are generally classified into two categories: the one-stage algorithms and the two-stage algorithms. The representative algorithms of the one-stage are SSD [4] and YOLO series [5], and the representative algorithm of the two-stage is Faster RCNN [6]. The main difference between them is that the one-stage algorithm extracts features directly in the network and predicts the categories and locations of all objects at once. The two-stage algorithm, on the other hand, generates all the candidate anchors that may contain the object and screens them according to certain rules. After that, detection is performed on the filtered candidate anchors to obtain finer classification and localization. To further enhance the feasibility of deep learning in industrial applications, especially vibration dampers detection, researchers have done many studies. Their work has shown that object detection using the convolutional neural network(CNN) can greatly improve accuracy [7–10]. CNN has three structural properties: local connectivity, weight sharing, and spatial or temporal sampling. They allow CNN to have some degree of translation, scaling, and warping invariance. In addition to a better representation of the extracted features, CNN can complete the extraction, selection, and classification of features in only one model, enhancing the divisibility of the features.

While existing work can reduce the inspection cost and improve the inspection accuracy to a certain extent, they have the following limitations:

- **Limited Labeling Data.** CNN training requires a large amount of data to obtain better results, however, there is no large-scale public dataset for training in the field of transmission line inspection, and the amount of data about vibration dampers in the current mainstream public dataset is relatively small.
- **Complex Background Interference.** The background includes the sky, clouds, mountains, forests, grass, houses, rivers, etc. The transmission line itself also interferes with the extraction of vibration dampers information. The complex background increases the difficulty of feature extraction.

In light of these shortcomings, the main contributions of this paper are as follows:

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00249

- **Vibration Damper Dataset Construction.** The dataset contains 38,295 images and 22,252 annotation information files.
- **Model Selection.** By evaluating the detection speed, accuracy, loss function, and other indexes, the RetinaNet model that is most suitable for the vibration dampers detection task is selected to realize the object detection of vibration dampers, with an average accuracy of 77.2%.
- **Model Optimization and Data Enhancement.** To address the problem of insufficient detection accuracy, optimization is carried out in terms of data and model structure. The second cleaning of all data is completed, and the FreeAnchor method is introduced into the model, optimizing the model accuracy to 83.1%.

II. PROPOSED METHODOLOGY

A. Datasets

Vibration Dampers are mainly used to prevent wire breakage problems due to transmission line vibration, which is vital to ensure safe operation. They are generally suspended below the transmission line. They are small in size, with the dimensions of length, width, and height generally ranging from 300mm × 40mm × 40mm to 500mm × 60mm × 60mm, and their main constituent material is silver-white cast iron. They are available in a variety of different forms, as shown in Fig. 1.

Fig. 1. Different shapes of vibration dampers.

The vibration damper dataset constructed includes aerial photographs of the vibration damper and its labeling information. To ensure the richness of the data, the UAV selected diverse shooting backgrounds under different weather conditions and took images of different sizes and occlusion conditions. When making the dataset, the location and category of the object need to be labeled and the corresponding XML file needs to be generated. We use LabelImg software to complete the labeling. The size of the occluded area should be estimated, if the object is occluded by more than 2/3 of the area, it will not be labeled, and the rest of the cases of occlusion will still be labeled, as shown in Fig.

90% of the dataset images are randomly selected as the training set and the remaining 10% as the test set. There are a total of 38,295 images, divided into 35,000 for the training set and 3,295 for the test set.

B. Modeling

The structure is shown in Fig. 2. RetinaNet consists of the Residual Network (ResNet) [11], the Feature Pyramid Network (FPN), and two sub-networks for bounding box regression and classification, and utilizes Focal Loss to solve the problem of imbalance of foreground-background categories in traditional one-stage detectors.

ResNet introduces the residual structure, it allows the convolutional network to learn the residual mapping. In forward propagation, the feature mapping contains less image information layer by layer, and the addition of direct mapping ensures that the network in layer ***i+1*** will contain more image information than layer ***i***. A comparison of ordinary and residual convolutional layers is shown in Fig. 3.

FPN is a feature extraction method capable of fusing multilayer features to improve the recognition of objects at different scales [12]. The feature map constitutes a hierarchical relationship from shallow to deep, with shallow features reflecting details such as light, dark, edges, and so on, and deep features reflecting the overall structure. RetinaNet, based on the original hierarchical structure, merges the deeper features into the shallower layers one by one to form a new feature pyramid, so that each layer combines the information of the details and the whole, reflecting comprehensive, as shown in Fig. 4.

In the object detection algorithm, the anchor-selected object is called the positive sample, and vice versa is called the negative sample. Positive samples have core values, while too many negative samples will have a negative impact on the detection results. The object that is easy to distinguish from the background feature information is called the easy sample, and vice versa is called the negative sample. When there are too many easy samples and very few difficult samples, it will be unfavorable for training. Focal Loss is proposed to dynamically adjust the weights in the RetinaNet algorithm [13], solving the problem of serious imbalance between the proportion of positive and negative samples and difficult and easy samples in the one-stage detection algorithm. Focal Loss is defined as:

$$\boldsymbol{FL}\left(p_t\right) = -\alpha_t(1-p_t)^{\gamma}\log\left(p_t\right) \tag{1}$$

Where $\boldsymbol{p}$ is the predicted probability of the sample in the category, $\boldsymbol{\alpha_t}$ is the positive and negative sample weights, and $(1-p_t)^{\gamma}$ is the difficult and easy sample weights. γ is a modulation factor, and the larger γ is, the lower the contribution of the easy sample loss is.

FreeAnchor is a learning-based matching method that can break the limitation of matching in IoU [14]. In the anchor-based object detection algorithm, IoU is used to measure the overlap between the anchor and GT. When IoU exceeds a certain threshold, the anchor is considered to be used to predict the object, otherwise, the anchor is considered as background. However, in the case of very dense objects, the use of IoU does not guarantee that the anchor points cover enough object features.

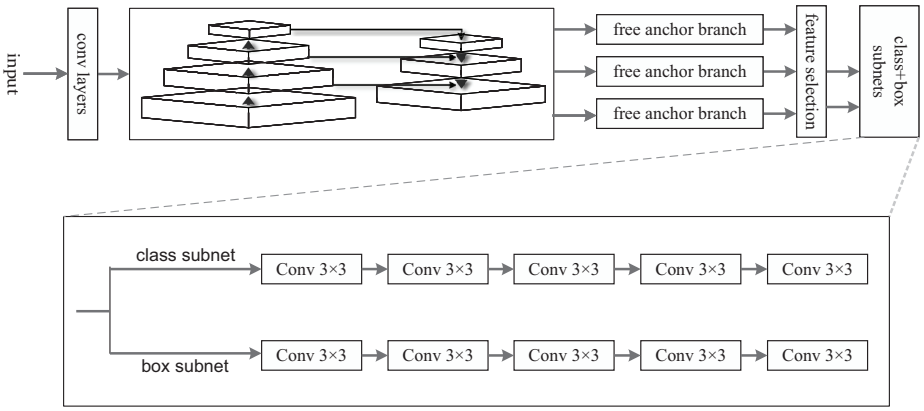

Fig. 2. The structure of our model.

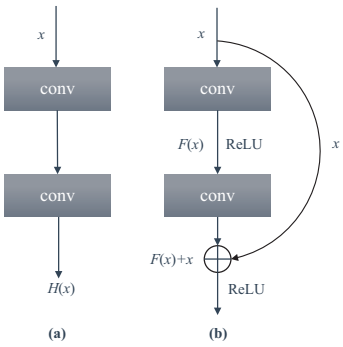

Fig. 3. Comparison of **(a)** Ordinary Convolutional Layer and **(b)** Residual Convolutional Layer.

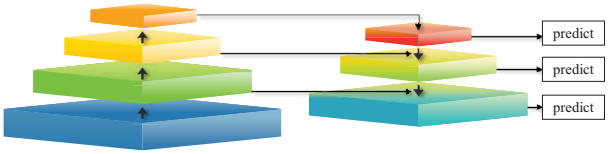

Fig. 4. Feature Pyramid Network.

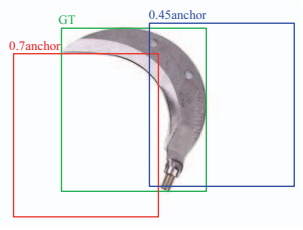

Fig. 5. Schematic of IoU matching error.

Fig. 5 shows that after IoU matching, the algorithm believes that the red box expresses more information about the object due to its better score, however, in reality, the blue box expresses more features than the red box even though it has a lower score. FreeAnchor changes the manual setting of matching rules to free anchor matching so that the network can independently learn to select the anchor that best reflects the characteristics of the object to match with GT, thus making the matching effect optimal. FreeAnchor is defined as:

$$\begin{aligned} L(\theta) = & -w_1 \sum_i \log(Mean\_max\,(X_i)) \\ & + w_2 \sum_j FL_-\left(P\{a_j \in A_-\}\left(1 - P\,(\theta)_j^{bg}\right)\right) \end{aligned} \tag{2}$$

where $X_i = \{P(\theta)_{ij}^{cls} P(\theta)_{ij}^{loc} | a_j \in A_i\}$, $P(\theta)_{ij}^{cls}$ is classification confidence, $P(\theta)_{ij}^{loc}$ is localization confidence, A_i is all candidate anchor sets, $FL_-(p)$ is Focal Loss, the $Mean\_max$ function is used to select the best matching anchor for each object, and w_1 and w_2 are the weighting factors.

C. Data Enhancement

Rich and high-quality data play a key role in improving modeling results, especially in avoiding model overfitting. Data enhancement methods include data cleaning, scale transformation, contrast transformation, noise interference, and so on. Data cleaning is to re-check the labeling information of the image to ensure that the labeling frame is close to the object and to reduce the occurrence of omission and mislabeling. Scale transformation is to zoom in or out of the image using different ratios. Contrast transformation uses histogram equalization to make the gray scale of the image uniformly distributed within a certain interval. Noise interference means noise superimposition on each pixel RGB channel of the image. We enhance the existing data in terms of both data cleaning and scale transformation.

III. RESULTS

In the field of object detection, a commonly used evaluation metric is mean average precision (mAP), which is used to measure whether the model predicts the box category and location accurately. The mAP is defined as:

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{3}$$

where Q is the set of categories for object detection, and $AveP(q)$ is the average accuracy of the objects under the computed categories.

A. Comparison of RetinaNet and other models

To better evaluate the performance of the RetinaNet model, we use Faster RCNN, SSD, YOLO3 model, and RetinaNet model to conduct comparison experiments.

TABLE I
COMPARISON OF MODEL TRAINING EFFECT

Model	Category	Backbone Network	Network Depth	Average Speed/s	mAP/%
Faster RCNN	two-stages	ResNet50	50	0.2994	73.7
SSD	one-stage	SSDVGG	16	0.2652	43.1
YOLO3	one-stage	DarkNet	53	0.0648	62.1
RetinaNet	one-stage	ResNet50	50	0.2725	77.2

Table 1 provides the experimental results. The RetinaNet model has the highest accuracy, the Faster RCNN model is second, and the performance of the SSD model is much lower than the other three models. SSD is poor due to its weak semantic information of shallow feature maps and extreme imbalance in the proportion of positive and negative samples. Faster RCNN uses the region proposal network to generate the candidate anchors. RetinaNet utilizes FPN. RetinaNet's special structure, Focal Loss, solves the problem of the severe imbalance in the proportions of positive and negative samples, as well as the proportion of difficult and easy samples. Although it is slower than SSD and YOLO3 in detection speed, it has the highest detection accuracy of 77.2%.

B. Comparison of the optimized and original RetinaNet model

To better evaluate the performance of the optimized model, comparative experiments are conducted between the optimized model and the original RetinaNet model, and the results are shown in Table 2. After introducing the FreeAnchor method, compared with the original RetinaNet model, the optimized model improved the accuracy by 3.1% and 4.9% before and after data enhancement, respectively. With the combined use of data enhancement and the FreeAnchor method, the optimized model achieved an accuracy of 83.1%.

Fig. 6 compares the average accuracy of the bounding boxes(bbox mAP) of the original RetinaNet model and the optimized model runs before and after data enhancement. The bbox mAP of the optimized model is higher than that of the original RetinaNet model, showing greater advantages.

TABLE II
OPTIMIZATION EFFECT OF THE MODEL

Model	Backbone Network	Network Depth	Data Enhancemen	Average Dpeed/s	mAP/%
original model	ResNet	50	No	0.2725	77.2
original model	ResNet	50	Yes	0.3832	78.2
optimized model	ResNet	50	No	0.2923	80.3
optimized model	ResNet	50	Yes	0.3479	83.1

FreeAnchor allows the anchor and the GT to be freely matched according to the model performance during the training process to represent more features of the object. Therefore the optimized model based on the FreeAnchor method performs better in the vibration damper detection task.

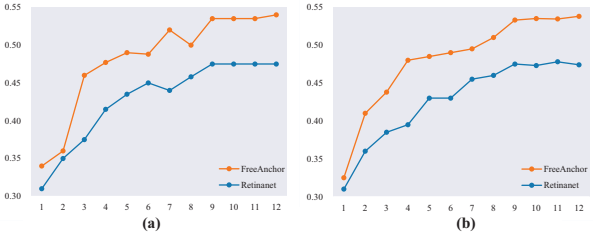

Fig. 6. bbox mAP of original and optimized models, **(a)** based on original data and **(b)** based on data enhancement.

IV. CONCLUSION

In this paper, we construct the vibration damper dataset and complete the labeling, segmentation, and enhancement of 38,295 images. We propose two aspects of optimization ideas:

- **Dataset Construction and Data Enhancement.** We carry out secondary cleaning on all 38,295 image data to ensure that the labeling frame is close to the object and reduce the occurrence of omission and mislabeling. At the same time, the data is enhanced by scale transformation, zooming in and out of the image.
- **Model Optimization.** The FreeAnchor method is introduced to allow the anchor and the GT to be flexibly matched based on model performance throughout the training process, to better describe the object's features.

The method proposed in this paper can achieve good detection results. In comparison with other algorithms, the RetinaNet optimized model based on the FreeAnchor has significantly improved the detection results on the vibration damper dataset we constructed.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No. 62271035), Beijing Municipal Natural Science Foundation (No. 4232021), the Open Project of Anhui Province Key Laboratory of Intelligent Building & Building Energy Saving (No. IBES2022KF08) and the Scientific Research Project of Beijing University of Civil Engineering and Architecture (No. ZF17072).

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 527–1554, 2006.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2014.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2015.

[5] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2015.

[6] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[7] W. Bao, Y. Ren, N. Wang, G. Hu, and X. Yang, "Detection of abnormal vibration dampers on transmission lines in uav remote sensing images with pma-yolo," *Remote. Sens.*, vol. 13, p. 4134, 2021.

[8] L. Xinyu, L. Yating, H. Jiang, M. Xiren, and J. Chen, "Slippage fault diagnosis of dampers for transmission lines based on faster r-cnn and distance constraint," *Electric Power Systems Research*, vol. 199, p. 107449, 2021.

[9] W. Bao, X. Du, N. Wang, M. Yuan, and X. Yang, "A defect detection method based on bc-yolo for transmission line components in uav remote sensing images," *Remote. Sens.*, vol. 14, p. 5176, 2022.

[10] Z. Zhao, G. Guo, L. Zhang, and Y. Li, "A new anti-vibration hammer rust detection algorithm based on improved yolov7," *Energy Reports*, 2023.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2015.

[12] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936–944, 2016.

[13] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017.

[14] X. Zhang, F. Wan, C. Liu, X. Ji, and Q. Ye, "Learning to match anchors for visual object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3096–3109, 2019.