Research on Robustness of Related Images for Power Equipment Inspection

Chunlei Zhou 1,a, Xiangzhou Chen 1,b, Wenli Liu 1,c, Tianyu Dong 2,d, Huang Yun 2,e

1Big data center of State Grid Corporation of China, Financial Street China Re Center, No.18, Luomashi Street, Hufangqiao, Xicheng District, Beijing, China
2State Grid information & Telecommunication Group Anhui Jiyuan Inspection and Testing Technology Co., HeFei, AnHui, China
aemail: clzhou@sgcc.com.cn ,bemail: xiangzhou-chen@sgcc.com.cn ,cemail: wenli-liu@sgcc.com.cn ,demail: dongtianyu@sgitg.sgcc.com.cn ,
eemail: dongtianyu@sgitg.sgcc.com.cn ,
Corresponding author’s e-mail: aemail: clzhou@sgcc.com.cn

Abstract. With the increase in the number of traction substations year by year, manual inspections are gradually being replaced by unattended inspections. Target detection algorithms based on deep learning are more widely used in intelligent inspections of power equipment. However, in practical applications, it is found that due to the small target to be detected, the accuracy of the deep learning model will decrease when the shooting angle is inclined and the light conditions are poor. This is because the algorithm’s robustness is low, and the detection ability of the model will be seriously affected when the angle or illumination difference with the sample is large. Based on this, the feature fusion part of the YOLOv3 algorithm and the selection of the loss function and the size of the anchor frame are improved, and the improved ASFF fusion method is used to classify various images in the power equipment. Actual measurement and repeated experiments show that the proposed method can be effectively applied to image recognition of various power equipment, optimize robustness, and greatly improve the image recognition efficiency of power equipment.

1. Introduction
In recent years, the number of substations has increased year by year, and the number of equipment in substations has also gradually increased. Due to the large number of power equipment and the remote location, it requires more manpower, financial and material resources to maintain these equipment, so it is urgent to realize the intelligent inspection of power equipment. Intelligent inspection uses digital image processing technology, shallow machine learning or deep learning methods to classify objects on power equipment, and improves the efficiency of inspection fault discovery, which has important engineering practical value.

At present, computer vision methods based on deep learning theory have become a research hotspot in image target detection, which can be divided into two types: one-step method and two-step method. The two-step method includes Area Convolutional Neural Network R-CNN and its optimization algorithms Fast R-CNN[1] and Faster R-CNN[2]. The difference between the two-step method and the one-step method is that the two-step method first generates a number of target
candidate domains through a region suggestion strategy, performs feature extraction, uses a support vector machine to identify the target category, and finally outputs the target type in the image and marks its location.

Aiming at the adaptability, generalization and accuracy of digital image processing, shallow machine learning, one-step and two-step methods. At the same time, considering the high requirements for real-time performance in power equipment detection, this paper proposes an improved YOLOv3 algorithm to optimize robustness and improve the ability of power equipment inspection to image detection and positioning.

2. YOLOv3 algorithm
YOLOv3 is a target detection algorithm improved by Redmon on the basis of YOLOv2. In terms of network structure, YOLOv3 uses a new network structure Darknet53 as shown in Figure 1, instead of Darknet19 in YOLOv2. Since there are 53 convolutional layers, it is named Darknet53.

![Figure 1. Darknet53 structure diagram](image)

In addition, drawing on the idea of residuals [3] uses a large number of residual blocks in the network, thus avoiding the phenomenon of gradient explosion in the training process of too deep networks;In terms of loss function, YOLOv3 uses multiple independent logistic regression classifiers. Because when objects appearing in a target box belong to multiple categories at the same time, the softmax loss function in YOLOv2 can only consider them as one category.Therefore, multiple independent logistic regression classifiers are used in YOLOv3 to achieve multi-label classification;In terms of feature fusion, YOLOv3 draws on the idea of FPN[4] and predicts on 3 different sizes, each size predicts 3 target borders.For a picture, if it is initially divided into a K×K grid, and C categories need to be predicted, then the final tensor obtained for each scale is K×K×[3×(4+1+C)],It includes the 4 offset coordinates of the target frame and the confidence score.Because of the fusion of the feature maps of the first two layers, the model can obtain more low-level and high-level image semantic information.

3. Related work

3.1. Data set
In the detection process of this article, the accuracy of detection also depends on the quality of the data set, and the number of data sets also affects the detection effect.At present, the open source data set does not contain images in the power equipment scene, and only similar instrument images can be filtered from the COCO data set [5], as shown in Figure 2, and a total of 2076 images have been screened out.In addition, in addition to the open source data set, I collected, screened and marked the targets to be detected in the power equipment scene, and the collected images included images obtained from multiple angles. Cut out 1000 images from the video collected on site.In addition, in order to avoid the over-fitting problem caused by insufficient data of the network, the sample was augmented as shown in Figure 3, including scaling, mirroring, cropping and Gaussian noise. Through
augmentation, the number of images was increased to 5000. Finally, this experiment uses 3 data sets with a total of 15,380 images for target detection model training.

3.2. K-means combined with a priori to set the anchor box

K-means sets the anchor box in combination with a priori. The Anchor box is a number of boxes of different sizes obtained by statistics or clustering from the ground truth of the training set, constraining the range of the predicted object, and adding a priori experience of size, So as to achieve the purpose of multi-scale learning, and avoid blindly finding the model during training, which helps the model to converge quickly. The k-means clustering method is used in the original YOLOv3 to obtain an anchor box that matches the size of the object in the original data set. After the size of the anchor frame is determined, the number of anchor frames needs to be determined, but the specific number is obtained after the original author's experiment, and the author uses various numbers of anchor frames in the model. Finally, by comparing the recall rate between different numbers, it is found that the recall rate is the highest when 9 anchor boxes are selected, and the final 9 anchor boxes are the best as shown in Figure 4. They are [10, 13], [16, 30], [33, 23], [30, 61], [62, 45], [59, 119], [116, 90], [156, 198], [373, 326].

3.3. Adaptive Feature Fusion (ASFF)

Previously, because the number of layers of the target detection network was relatively shallow, the features after feature extraction were not obvious, but the specific location of the object was more accurate. With the development of deep learning, the number of network layers has gradually increased, and the development of network fusion has also improved, among which multi-scale fusion has developed rapidly. The FPN feature fusion method used in YOLOv3 [4] is shown in Figure 5. The
difference from other feature fusion is that its prediction is performed independently at different feature layers. However, this fusion method mostly uses concatenation direct connection, which does not make full use of the characteristics of different scales. Therefore, this paper uses adaptive feature fusion fusion [6] to replace concatenation. Adaptive Spatial Feature Fusion (ASFF), most of the previous fusions rely on element feature fusion methods and different levels of feature fusion methods that rely on cascading each other. It is a new pyramidal feature fusion that relies on the relationship between different data. The idea is to adapt and learn the map fusion of different scales of spatial weights. It is composed of two steps: the same rescaling and adaptive fusion, and function adjustment. Because the resolution and the number of channels between the three convolutional layers that need to be fused in YOLOv3 are different, the up-sampling and down-sampling methods of different layers should be corrected. For layers that need to be upsampled, for example, if you want to get P3, you need to adjust C5 to the same size as C3. The method used is to first adjust the number of channels to be consistent with level3 through 1×1 convolution, and then use interpolation to resize to the same size. For the layers that need to be downsampled, for example, if you want to get P5, you only need to use a 3×3 convolution with stride=2 for C4 to C5. If it is C3 to C5, you need to add a maximum pooling of stride=2 on the basis of 3×3 convolution, so that the sizes of C3 and C5 can be adjusted to be the same. Through the adaptive spatial feature fusion method, we can see that for small objects, we need more fine particles in the underlying features.

3.4. TSE loss function
As a basis for the deep neural network to punish misdetected samples, the loss function can greatly affect the effect of model convergence. In the final output of the YOLOv3 network, the (x, y) object confidence and category confidence are all activated by the Sigmoid function, and then SSE is used to calculate the final loss. In the network needed in this article, due to the relatively high similarity of object features in power equipment, there is a situation where the positive and negative samples are very unbalanced. It is easy to appear that the output of the network is 0 at the beginning of training. At this time, the negative sample means that the real value is 0 and the error is close to 0, and the real value is 1 has a large error. For predx and predy, the difference is (-1, 1) instead of (0, 1). Directly use the formula grad=tx−predx to calculate the gradient, which is exactly the derivative form of the squared loss function. Following the same idea, this article adopts the calculation formula of the derivative form of TSE (Tan-Squared Error) [3]. The derivative form is as Formula 1:

\[
\text{Loss Grad} = \tan(t - \sigma(z))/\tan(1)
\]

In YOLOv3, \(t - \sigma(z) \in (-1, 1)\), since \(t\) is also an indeterminate term, this difference has nothing to do with the value of \(\sigma(z)\) itself. Let \(x = t - \sigma(z)\) and make the function graph as shown in Figure 6.
Figure 6. TSE and SSE derivative image

From the graph above, it can be found that the value range of the derivatives of the two functions are both (-1,1), and the absolute value of the derivative of the sum squared loss function must be greater than the absolute value of the derivative of the Tangent square loss function. When there is a big difference between the actual value and the predicted value, the error can be reduced. When the result passes the sigmoid function, the partial derivative of w and the partial derivative of b can be gradually increased. This not only increases the speed of convergence at the beginning of the experiment, at the same time, the impact of the disappearance of the gradient is minimized. Because the value of the sum square loss function is slightly larger than the value of the Tangent square loss function, for the case where the error between the true value and the predicted value is close to 0, in order to get better convergence of the model, the adjustment range of the weight of the output layer can be made smaller. This can better solve the situation that the training model does not converge and the results diverge during the training process, and the adjustment range can be controlled within a small range. The derivative images of the two functions are shown in Figure 7.

Figure 7. The gradient of the derivative of TSE and SSE

4. Experimental results
The experiment was conducted on a server with Ubuntu 16.04 operating system, and the graphics card used was NVIDIA GTX 1080Ti. Using 2076 pictures filtered from the COCO data set and a self-made data set, a total of 15380 pictures are formed. In order to compare whether to use clustering to select the initial candidate frame on the level of model detection performance, first train according to the original anchor box parameters of YOLOv3, and then train according to the anchor box obtained by clustering. The training sets use the same data set. The performance of the final model on the test set is shown in Table 1.

| Initial candidate frame | mAP   | FPS |
|-------------------------|-------|-----|
| YOLOv3                  | 72.77 | 33  |
| Improved candidate border | 82.30 | 39  |
The results in Table 1 show that using the original YOLOv3 anchor box parameters compared to the model trained using the clustered anchor box parameters, the latter has a significant improvement in detection accuracy or detection speed. Compared with the original YOLOv3, its mAP has increased by 9.53 percentage points, and it can detect 6 more pictures per second. This experiment uses two different loss functions, SSE and TSE, and tests the same samples in different environments (including normal, inclined, reflective, and dim). It is concluded that using the TSE loss function can get better accuracy in different environments.

This article modified the original YOLOv3 anchor box, FPN feature fusion and SSE loss function, and obtained a higher accuracy rate through comparative experiments. The experimental results are shown in Table 2.

Table 2 Comparative experiment

| Yolov3 | Yolov3 or anchor | Yolov3+ anchor+ASSF | Yolov3+anchor+ASSF+TSE |
|--------|-----------------|---------------------|------------------------|
| mAP    | 71.9            | 74.1                | 81.7                   | 82.2                   |
| Pointer table | 84.5            | 86.0                | 88.1                   | 89.3                   |
| Counter | 69.3            | 80.6                | 81.4                   | 83.2                   |
| Digital meter | 71.2            | 70.1                | 78.8                   | 80.1                   |
| Round table | 75.2            | 57.3                | 80.8                   | 84.3                   |
| Indicator light | 76.6            | 82.9                | 84.7                   | 86.7                   |
| Square lamp | 65.3            | 73.2                | 77.6                   | 79.3                   |
| Rotary switch | 75.7            | 76.1                | 80.1                   | 81.3                   |
| Air switch | 78.2            | 80.4                | 84.7                   | 84.9                   |
| Fixed switch 1 | 82.0            | 82.0                | 84.1                   | 85.4                   |
| Fixed switch 2 | 40.7            | 52.2                | 76.7                   | 78.2                   |

5. Conclusion

Based on the improvement of the YOLOv3 network, this paper aims at the problem of low detection accuracy and low robustness caused by the small target to be detected in the intelligent inspection of power equipment, when the shooting angle is inclined and the light conditions are poor. When the K-means algorithm is used to cluster the initial target frame, the TSE loss function and adaptive feature fusion method improve the network detection accuracy, and solve the current common problem of low network detection accuracy. On a data set combining the same self-made data set and open source data set, the modified network in this article and the original network are trained and compared. The experimental results show that:

The improved method proposed in this paper has achieved a greater degree of improvement in image accuracy, optimized robustness, and verified the effectiveness of the network framework.

References

[1] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.
[2] Ren SQ, He KM, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montréal, Canada. 2015:91–99.
[3] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV,USA. 2016:770–778.
[4] Lin TY, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017:936-944.
[5] Chen XL, Fang H, Lin TY, et al. Microsoft COCO captions: Data collection and evaluation server[C]. https://arxiv.org/abs/1504.00325, 2015.

[6] Liu S, Huang D, Wang Y. Learning Spatial Fusion for Single-Shot Object Detection[J/OL]. https://arxiv.org/abs/1911.09516, 2019.