Journal of Intelligent Learning Systems and Applications, 2024, 16, 1-11
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

A CNN-Based Single-Stage Occlusion Real-Time Target Detection Method

Liang Liu[1], Nan Yang[1], Saifei Liu[1], Yuanyuan Cao[1], Shuowen Tian[1], Tiancheng Liu[1], Xun Zhao[2]

[1]Rizhao Power Supply Company, State Grid Shandong Electric Power Company, Rizhao, China
[2]School of Automation, Beijing Information Science and Technology University, Beijing, China
Email: Liuliang@163.com, Yannang@163.com, Caoyuanyuan@163.com

How to cite this paper: Liu, L., Yang, N., Liu, S.F., Cao, Y.Y., Tian, S.W., Liu, T.C. and Zhao, X. (2024) A CNN-Based Single-Stage Occlusion Real-Time Target Detection Method. *Journal of Intelligent Learning Systems and Applications*, **16**, 1-11.
https://doi.org/10.4236/jilsa.2023.161001

Received: May 10, 2023
Accepted: February 4, 2024
Published: February 7, 2024

Abstract

Aiming at the problem of low accuracy of traditional target detection methods for target detection in endoscopes in substation environments, a CNN-based real-time detection method for masked targets is proposed. The method adopts the overall design of backbone network, detection network and algorithmic parameter optimisation method, completes the model training on the self-constructed occlusion target dataset, and adopts the multi-scale perception method for target detection. The HNM algorithm is used to screen positive and negative samples during the training process, and the NMS algorithm is used to post-process the prediction results during the detection process to improve the detection efficiency. After experimental validation, the obtained model has the multi-class average predicted value (*mAP*) of the dataset. It has general advantages over traditional target detection methods. The detection time of a single target on FDDB dataset is 39 ms, which can meet the need of real-time target detection. In addition, the project team has successfully deployed the method into substations and put it into use in many places in Beijing, which is important for achieving the anomaly of occlusion target detection.

Keywords

Real-Time Mask Target, CNN (Convolutional Neural Network), Single-Stage Detection, Multi-Scale Feature Perception

1. Introduction

In recent years, with the rapid development of neural networks, deep learning has made great progress as an important component of neural networks, and has been widely used in speech recognition, image classification, natural language processing and other fields. From the genre, deep learning can be divided into

three main directions: supervised learning, unsupervised learning, and reinforcement learning. With the rapid development of deep learning technology, deep learning has become increasingly prominent in the field of target detection. Many new excellent deep learning methods and target detection frameworks have been proposed.

The current advanced target detection framework can be divided into two categories: two-stage and single-stage. Most of the most advanced detection methods are implemented by a two-stage detection framework, such as Faster R-CNN [1], R-FCN [2], FPN [3] and Cascade R-CNN. Compared with the two-stage target detection framework, the single-stage detection framework has the advantages of simple structure and fast speed. Typical representatives are SSD [4] and Yolo [5] [6], they have achieved a balance between speed and accuracy. The latest single-stage detection framework Yolov5 achieves an accuracy comparable to the two-stage target detection framework. In general, the accuracy of the current single-stage detection framework is still difficult to surpass the two-stage detection framework.

Target detection is a specific task in the field of target detection. The development of general target detection technology [7]-[12] has greatly promoted the development of target detection [13] [14] [15] [16]. Specifically, different from generic object detection, target detection features smaller ratio variations (from 1:1 to 1:1.5) but much larger scale variations (from several pixels to a thousand pixels).

By drawing on the ideas of Stefanos *et al.*, the research on target detection was mainly divided into traditional methods based on hand-designed features and modern methods based on Convolutional Neural Net (CNN) to extract features.

With the continuous progress of single-stage target detection technology, the latest target detection methods [17] [18] mainly focus on the realization of single-stage target detection, Since single-stage target detection does not use the suggestion box for the "rough detection + fine-tuning" strategy, only one forward propagation calculation is performed, so the detection speed has been greatly improved compared with the two-stage target detection method.

As a fine detection technique for non-contact substation targets, target detection is critical for its real-time and accuracy. However, in device injury detection, the target of the subject is mostly occluded by objects. The lack of features and the noise aliasing caused by the occluders can seriously reduce the accuracy of traditional target detection methods, or even fail to detect the target. In addition, in public places with high traffic flow, the real-time nature of target detection under high traffic flow is also a major challenge. How to shorten the detection time of the target under the premise of guaranteeing the detection accuracy becomes the key to non-contact substation target detection.

To summarize, our key contributions are:

• The backbone of the algorithm refers to the idea of multi-scale feature perception in SSD [4], and builds a lightweight multi-scale convolutional network

according to the task characteristics of target detection, so as to solve the problem of poor detection effect of the original SSD network on the small target.

• The single-stage target detection model generally has the problem of imbalance between simple samples and difficult samples during training. We improve the loss function used in the training of the original SSD algorithm and use the improved cross-entropy loss function to solve the problem between simple samples and difficult samples and improve the comprehensive detection performance of the trained model.

• In view of the characteristics of target detection tasks in substation, we constructed a mask occlusion target dataset for algorithm model training to improve the accuracy of the detection model for targets wearing masks.

2. Model and Materials

2.1. Algorithm's Overall Structure

The algorithm draws on the use of feature pyramid prediction in SSD, and integrates the detection results of multiple convolutional layers to realize the detection for targets of different sizes. The overall structure of the algorithm is shown in Figure 1, including the backbone network, the FPN feature fusion network, and the detection network. The detection network is composed of a positioning sub-network and a classification sub-network, which respectively complete the target bounding box positioning and classification confidence prediction.

2.2. Backbone Network

The backbone network is composed of a basic network and a high-level network. The convolutional structure parameters of the network are shown in Table 1. The backbone network obtains a larger reception and richer semantic information by expanding the down-sampling multiple layer by layer.

Table 1. Convolution structure of the backbone network.

Convolution layer	Parameters (kernel size, channel)	Feature output (high, wide)
Conv1(C1)	{(3 × 3, 32), (1 × 1, 16)}	h = H/2, w = W/2 (C1_2)
Max_pooling1	2 × 2	h = H/4, w = W/4
Conv2(C2)	{(3 × 3, 32), (1 × 1, 32)}	h = H/4, w = W/4 (C2_2)
Max_pooling2	2 × 2	h = H/8, w = W/8
Conv3(C3)	{(3 × 3, 64), (1 × 1, 32), (3 × 3, 64)}	h = H/8, w = W/8 (C3_3)
Max_pooling3	2 × 2	h = H/16, w = W/16
Conv4(C4)	{(3 × 3, 128), (1 × 1, 64), (3 × 3, 128)}	h = H/16, w = W/16 (C4_3)
Max_pooling4	2 × 2	h = H/32, w = W/32
Conv5(C5)	{(3 × 3, 256), (1 × 1, 128), (3 × 3, 256)}	h = H/32, w = W/32 (C5_3)
Max_pooling5	2 × 2	h = H/64, w = W/64
Conv6(C6)	{(3 × 3, 256), (1 × 1, 256), (3 × 3, 256)}	h = H/64, w = W/64 (C6_3)

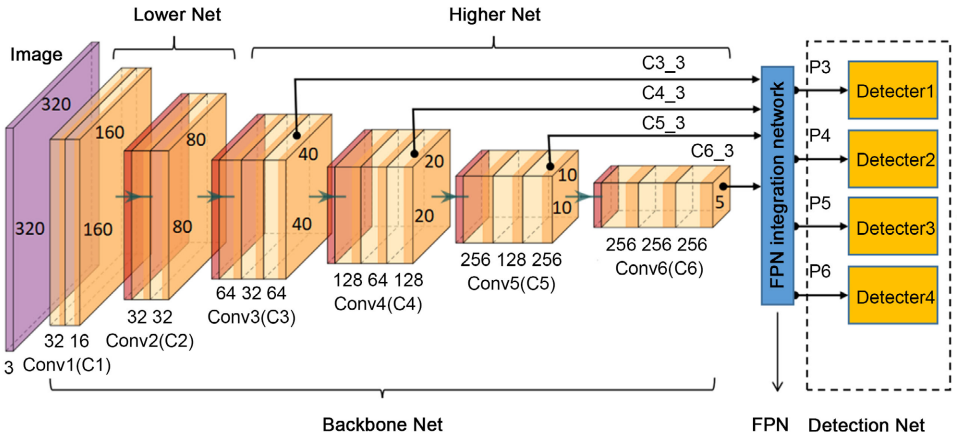


Figure 1. Algorithm's overall structure framework.

There is only one actual target detection classification type for the basic network design. Compared with the multi-target detection of SSD, the classification difficulty of basic network is greatly reduced. Therefore, the VGG [5] deep classification network is discarded and two shallow convolutional networks C1 and C2 are used to extract the shallow semantic information in the input image. The high-level network is composed of four convolutional networks C3, C4, C5, and C6, which output four different scale feature maps to enhance the algorithm's ability to detect targets of different sizes. In order to reduce the amount of parameter calculations and speed up the detection speed, the algorithm uses the SSH [19] method to cancel the last fully connected layer of the convolutional network. In addition, the cancellation of the fully connected layer makes the size of the input image no longer limited during testing.

The downsampling factor of the backbone network is 64, that is, the target with 64 pixels is output as a point on the last layer of feature map (C6_3).

2.3. Detection Network

To deepen the semantic information of the input feature maps, the detector of the detection network uses two different 3 × 3 convolution kernels to convolve the four fusion feature maps of P3_3-P6_3. There are four detectors with the same structure in the detection network. The connection structure of the first detector is shown in **Figure 2**. The detector consists of a positioning sub-network (box net) and a classification sub-network (class net) that are used to do target bounding box location and classification confidence prediction for fusion feature maps of different scales. The detailed process is: the positioning sub-network convolves the fusion feature map through a 3 × 3 convolution kernel, and outputs a list of coordinate values of the target prediction frame [[x1, y1, w1, h1], [x2, y2, h2, w2], …], the classification sub-network outputs the classification confidence degree list [S1, S2, S3, …] corresponding to each target frame in the fusion feature map through a 3 × 3 convolution kernel. After sorting the classification confidence degree list, the pre-diction box with a score greater than 0.3 is selected as the candidate box, and then the candidate box is subjected to non-maximum

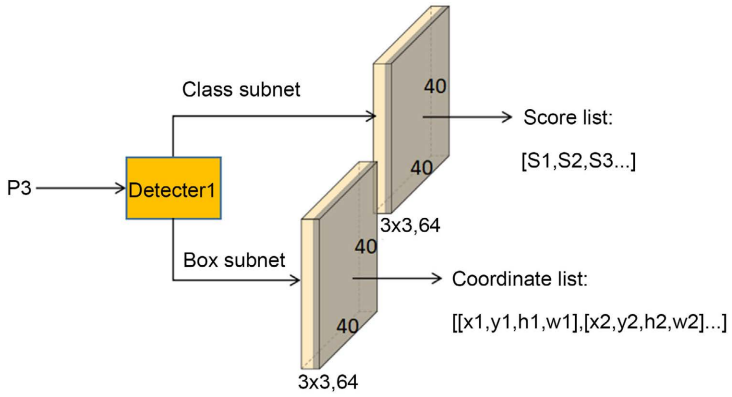


Figure 2. Connection structure of detection network's first detector.

suppression processing, and the candidate box with an IOU greater than the threshold 0.3 is selected as the final target bounding box.

2.4. IOU Calculation

The positive and negative samples need to be divided before model training. The method is to match the default box and the ground truth box, and calculate IOU between the default box and the ground truth box through the Jaccard function. Assuming that A represents the area of the default box, B represents the area of the ground truth box, and threshold is the set threshold. When the IOU is greater than the set threshold, the A default box is classified as a positive sample, otherwise it's a negative sample, as shown in Equation (1):

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} > \text{threshold} \tag{1}$$

In general, the number of positive samples matched by Jaccard calculation is far less than the number of negative samples that cannot be matched, resulting in an imbalance in the number of positive and negative samples, which makes it difficult to do model training. If the number of negative samples is not suppressed, and the samples are directly trained, the training direction will be dominated by the negative samples, and the detection performance of the resulting model will be poor. In order to solve the problem of serious imbalance in the ratio of positive samples to negative samples, the Hard Negative Mining (HNM) method is used in training process. The default boxes are sorted according to the classification confidence degree, the default boxes with high confidence are selected to be trained, and the ratio of positive samples to negative samples are controlled to nearly 1:3.

3. Experimental and Analysis

The target detection model in the article is built and trained in the PyTorch environment. The training is based on NVIDIA GeForce GTX1050 GPU and Intel CORE i7 CPU. The training process uses Sto-chastic Gradient Descent (SGD), the learn rate is set to 0.01, the weight_decay is set to 0.0005, the momentum is set to 0.9, the batch_size is set to 16, the max_epoch is set to 500 iterations, and

the input image size is 320 × 320 pixels during training.

3.1. Detection Network

The main evaluation indexes of the experiment use Precision and Recall, namely P-R curve, and use multi-category mean Average Precision (*mAP*) to comprehensively evaluate the detection effect of the model, as follows:

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

where P is the accuracy rate, TP is the number of correctly predicted targets in the samples, FP is the number of backgrounds that are incorrectly predicted as targets in the samples, and the accuracy rate refers to the proportion of the number of targets correctly predicted by the model to the total number of targets. In Equation (3), R is the recall rate, FN is the number of targets that are incorrectly predicted as the background in the samples, and the recall rate refers to the ratio of the number of targets correctly predicted by the model to the total number of truly labeled targets in the samples. *mAP* refers to the average value of multiple categories of *AP*, and the calculation method of *AP* is the area enclosed by the *P-R* curve and the coordinate axis.

3.2. Experimental Results

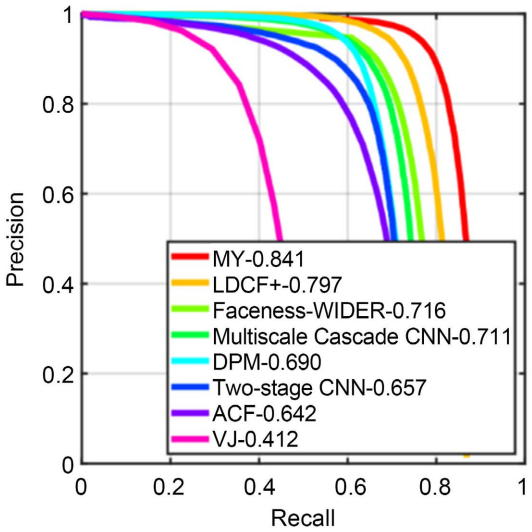
In order to verify the performance of the model, the above target detection methods are compared with traditional target detection methods based on hand-designed features (VJ [8], DPM, ACF and LDCF+) and modern target detection methods based on CNN for accuracy testing on the WIDER TARGET test set. The IOU threshold during the test is set to 0.3, and the P-R curves on the three subsets of Easy, Medium, and Hard are shown in **Figure 3**. The red curve in **Figure 3** represents our target detection method. According to the difficulty of target detection, the three subsets are divided into three difficulty levels: easy, medium, and difficult.

Comparing the experimental results of the methods in **Figure 5**, it can be concluded that the MY method achieved the best detection results on the three subsets of the WIDER TARGET test set, and obtained the *mAP* of Easy-0.841, Medium-0.802, and Hard-0.600. In addition, it can be seen from **Figure 5** that the MY method has a better recall rate under the same precision rate of different methods.

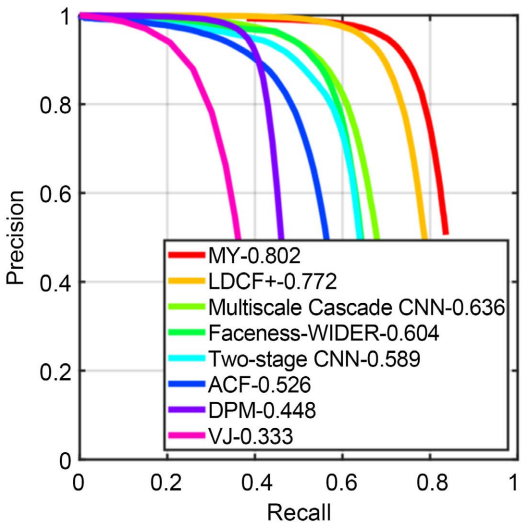
Table 2 shows the *mAP* performance of MY method on three subsets. Compared with traditional target detection methods based on hand-designed features and several CNN-based target detection methods, the MY method has a significant improvement in detection accuracy, and achieved the highest *mAP* on the three test subsets. Among them, the detection accuracy of the MY method on the Easy and Medium subsets is more than 2 times higher than that of the VJ method, and the detection accuracy of the Hard subset is 4.38 times higher than

Table 2. Comparison of *mAP* of different methods on different difficulty subsets of the WIDER TARGET test.

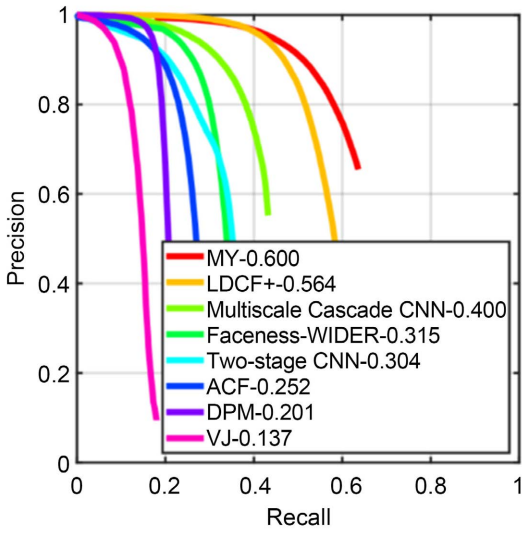
Method	*mAP*		
	Easy	Medium	Hard
MY (ours)	0.841	0.802	0.600
LDCF+	0.797	0.772	0.564
Targetness	0.716	0.604	0.315
Multi-scale Cascade CNN	0.711	0.636	0.400
DPM	0.690	0.448	0.201
Two-stage CNN	0.657	0.589	0.304
ACF	0.642	0.526	0.252
VJ	0.412	0.333	0.137



(a)



(b)



(c)

Figure 3. P-R curves of different methods on different difficulty subsets of the WIDER TARGET test set: (a) Result on easy subset; (b) Result on medium subset; (c) Result on hard subset.

that of the VJ method. Compared with the classic CNN-based target detection method targetness, the MY method achieves the highest scores on all three subsets. The experimental results can fully illustrate that the CNN-based target detection method generally has more advantages than the traditional target detection method based on hand-designed features in terms of detection accuracy.

In order to further verify the performance of the model, the MY method and the classic target detection methods based on deep learning (Faster R-CNN, CNN Cascade, Targetness) are supplemented with experiments on the FDDB data set. The IOU threshold is set to 0.5 during the experiment. The detection speed of each method and the true positives rate corresponding to different false positives are obtained. The experimental results are shown in **Table 3**.

From the experimental data in **Table 3**, it can be concluded that the detection speed of the MY method on the FDDB target data set is 14 fps, and the detection time of a single target is about 39 ms after conversion, which can basically achieve the effect of real-time target detection. In addition, the true rate of this method is greater than 0.9 in different stages of false positives on FDDB, indicating that this method has a good positive detection rate for various types of targets.

Figure 4 shows the effect of real-time target detection on the picture in the substation. As can be seen from the resultant figure, the MY method can successfully detect multiple substation equipment targets in the current substation. Compared with DaSiamRPN and SSD, our method has good detection performance for real-time targets, high traffic targets and occlusion targets.

3.3. Alarm System

The project team has actively used its strengths to develop a fully automated

deep learning-based target detection system for cooperative substations, enabling the detection of electrical substation equipment in a masked environment. The system uses our target detection model to detect foreign object targets inside the transformer, and can be run directly on an embedded board. The test results are shown in Figure 5.

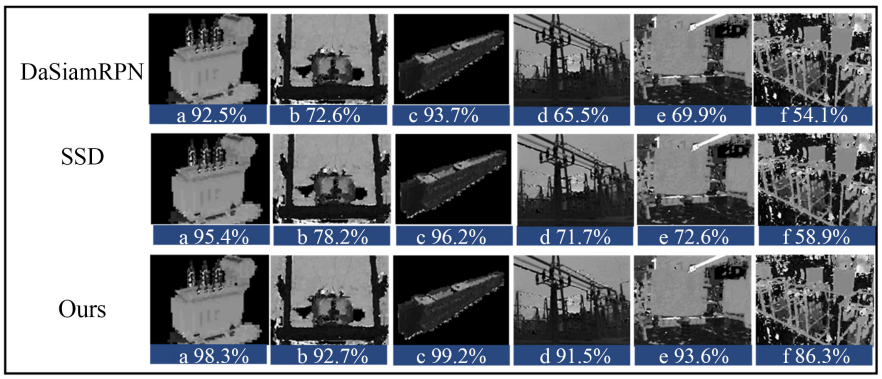


Figure 4. Results of the testing of the electrical substation equipment in the substation using the MY method.

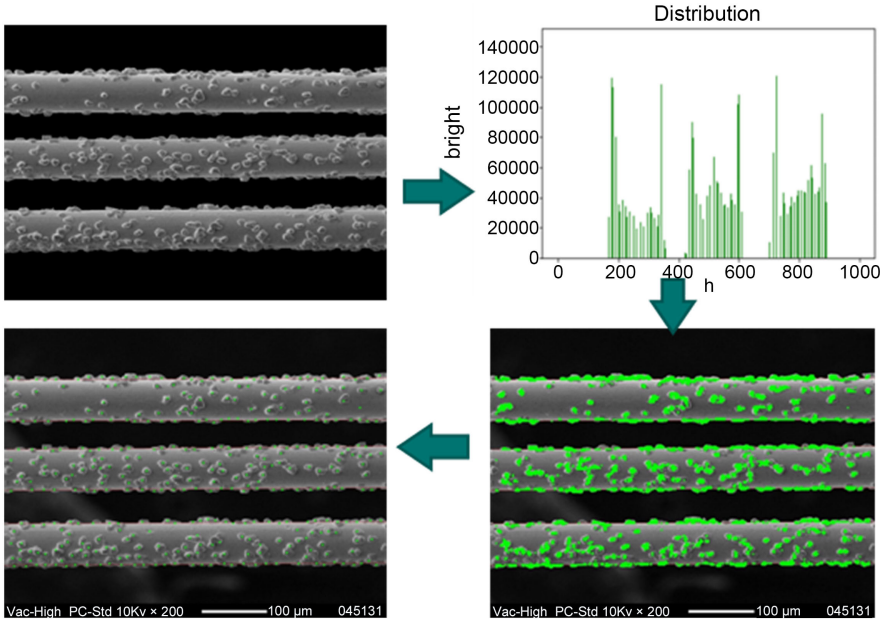


Figure 5. Results of target detection of foreign objects in transformers.

Table 3. False positives correspond to the true positives rate.

Method	False positives corresponds to the true positives rate				Speed	Time of single
	100	200	500	1000		
CVPR2016 FR-CNN	0.90	0.93	0.95	0.96	3 fps	185 ms
MY (ours)	0.90	0.92	0.94	0.94	14 fps	39 ms
ICCV2015 Targetness	0.87	0.8	0.89	0.90	20 fps	28 ms
CVPR2015 CNN Cascade	0.85	0.85	-	-	14 fps	40 ms

4. Conclusion

A CNN-based single-stage masked real-time target detection method is proposed for the practical needs of substation target anomaly detection. The model is trained on a self-built masked occlusion target dataset through various tests on the algorithm backbone network structure and parameter optimisation methods. Experiments show that the target detection method proposed in this paper has good detection performance for masked targets. Compared with the traditional target detection methods, the detection accuracy has a general advantage and achieves the best detection *mAP*. Supplementary experiments compare the detection speed of the method with the modern classical methods, and the true rates corresponding to different numbers of false positives, which show that the method has high real-time performance and accuracy, and can meet the needs of real-time target detection, but its speed still needs to be further improved. In conclusion, our proposed target detection method has good detection performance in terms of accuracy, real-time, and occlusion target detection, which is of great significance for safeguarding the safety of substations.

Acknowledgements

This research was funded by State Grid Limited Science and Technology Project, Grant No. 520617230001.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Xia, Y.Q., Gao, R.Z. and Lin, M. (2020) Green Energy Complementary Based on Intelligent Power Plant Cloud Control System. *Acta Automatica Sinica*, **46**, 1844-1868.

[2] Chen, Y.J., Zhu, X.T. and Yu, Y.R. (2022) Empirical Analysis of Lightning Network: Topology, Evolution, and Fees. *Ruan Jian Xue Bao/Journal of Software*, **33**, 3858-3873.

[3] Zheng, H.B., Sun, Y.H., Liu, X.H., *et al.* (2021) Infrared Image Detection of Substation Insulators Using an Improved Fusion Single Shot Multibox Detector. *IEEE Transactions on Power Delivery*, **36**, 3351-3359. https://doi.org/10.1109/TPWRD.2020.3038880

[4] Ala, G., Favuzza, S. and Mitolo, M., Musca, R. and Zizzo, G. (2022) Forensic Analysis of Fire in a Substation of a Commercial Center. *IEEE Transactions on Industry Applications*, **56**, 3218-3223. https://doi.org/10.1109/TIA.2020.2971675

[5] Nassu, B.T., Marchesi, B., Wagner, R., *et al.* (2022) A Computer Vision System for Monitoring Disconnect Switches Distribution Substations. *IEEE Transactions on Power Delivery*, **37**, 833-841. https://doi.org/10.1109/TPWRD.2021.3071971

[6] Balouji, E., Bäckström, K., McKelvey, T. and Salor, Ö. (2020) Deep-Learning-Based Harmonics and Interharmonics Predetection Designed for Compensating Significantly Time-Varying EAF Currents. *IEEE Transactions on Industry Applications*, **56**, 3250-3260. https://doi.org/10.1109/TIA.2020.2976722

[7] Guan, X., Gao, W., Peng, H., Shu, N. and Gao, D.W. (2022) Image-Based Incipient

Fault Classification of Electrical Substation Equipment by Transfer Learning of Deep Convolutional Neural Network. *IEEE Canadian Journal of Electrical and Computer Engineering*, **45**, 1-8. https://doi.org/10.1109/ICJECE.2021.3109293

[8] Zheng, H.B., Cui, Y.H., Yang, W.Q., *et al.* (2022) An Infrared Image Detection Method of Substation Equipment Combining Iresgroup Structure and CenterNet. *IEEE Transactions on Power Delivery*, **37**, 4757-4765. https://doi.org/10.1109/TPWRD.2022.3158818

[9] Ou, J.H., Wang, J.G., Xue, J., Zhou, X., *et al.* (2023) Infrared Image Target Detection of Substation Electrical Equipment Using an Improved Faster R-CNN. *IEEE Transactions on Power Delivery*, **38**, 387-396. https://doi.org/10.1109/TPWRD.2022.3191694

[10] Han, S., Yang, F., Jiang, H., *et al.* (2021) A Smart Thermography Camera and Application in the Diagnosis of Electrical Equipment. *IEEE Transactions on Instrumentation and Measurement*, **70**, 1-8. https://doi.org/10.1109/TIM.2021.3094235

[11] Li, J., Xu, Y., Nie, K., *et al.* (2023) PEDNet: A Lightweight Detection Network of Power Equipment in Infrared Image Based on YOLOv4-Tiny. *IEEE Transactions on Instrumentation and Measurement*, **72**, 1-12. https://doi.org/10.1109/TIM.2023.3235416

[12] Zhou, N., Luo, L.E., Sheng, G.H. and Jiang, X.C. (2019) High Accuracy Insulation Fault Diagnosis Method of Power Equipment Based on Power Maximum Likelihood Estimation. *IEEE Transactions on Power Delivery*, **34**, 1291-1299. https://doi.org/10.1109/TPWRD.2018.2882230

[13] Fan, Z., Shi, L., Xi, C., *et al.* (2022) Real Time Power Equipment Meter Recognition Based on Deep Learning. *IEEE Transactions on Instrumentation and Measurement*, **71**, 1-15. https://doi.org/10.1109/TIM.2022.3191709

[14] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324

[15] Yang, S., Luo, P., Loy, C.-C. and Tang, X. (2016) From Facial Parts Responses to Target Detection: A Deep Learning Approach. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 3676-3684. https://doi.org/10.1109/ICCV.2015.419

[16] Kayhan, O.S. and van Gemert, J.C. (2020) On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, 13-19 June 2020, 14262-14273. https://doi.org/10.1109/CVPR42600.2020.01428

[17] Bai, Z.Q., Cui, Z.P., Rahim, J.A., Liu, X. and Tan, P. (2020) Deep Facial Non-Rigid Multi-View Stereo. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, 13-19 June 2020, 5850-5860. https://doi.org/10.1109/CVPR42600.2020.00589

[18] Hang, R., Liu, Q., Hong, D. and Ghamisi, P. (2019) Cascaded Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience & Remote Sensing*, **57**, 5384-5394. https://doi.org/10.1109/TGRS.2019.2899129

[19] Mou, L.C., Lu, X.Q., Li, X.L. and Zhu, X.X. (2020) Nonlocal Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoence and Remote Sensing*, **58**, 8246-8257. https://doi.org/10.1109/TGRS.2020.2973363