IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 7, 2018, accepted July 11, 2018, date of publication July 23, 2018, date of current version August 20, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2859048

A Deep Learning Approach for Oriented Electrical Equipment Detection in Thermal Images

XIAOJIN GONG[1], (Member, IEEE), QI YAO[1], MENGLIN WANG[1], AND YING LIN[2]
[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[2]State Grid Shandong Electric Power Research Institute, Jinan 250002, China

Corresponding author: Xiaojin Gong (gongxj@zju.edu.cn)

This work was supported by the Natural Science Foundation of Zhejiang Province, China, under Grant LY17F010007.

ABSTRACT Due to the high precision and non-contact characteristics, infrared thermography has been widely used in equipment inspection to ensure the safety of electric power systems. A fundamental step toward automatic inspection and diagnosis is the detection of equipment in thermal images. Therefore, this paper presents a deep learning approach to detect equipment parts in real-time. Specifically, we propose a deep convolutional neural network that predicts the coordinates, orientation angle, and class type of each equipment part. A prior concerning orientation consistency between parts is also incorporated into our model to improve the prediction results. For evaluation, we construct a large image set containing various kinds of scenarios. Experiments on the data set show that our method is robust to noise, achieving 93.7% mean average precision when the intersection over union threshold is 0.5, and running at 20 fps on GPU. We believe that our high accurate detection results can benefit the subsequent diagnosis.

INDEX TERMS Electrical equipment detection, oriented object detection, automatic diagnosis.

I. INTRODUCTION

Equipment inspection plays an important role in ensuring the safety of electric power systems. By monitoring electrical devices, we can detect the degradation in time and prevent from unplanned power outage, fire hazards and other potential risks. For this purpose, thermal imaging cameras are extensively used. They provide a non-contacting way to sense the infrared energy emitted from equipment surface, so that the inspection can be conducted without shutting down any system. The collected thermal images reveal temperature distributions, from which we are able to diagnose equipment status. However, traditional diagnosis is mainly performed by experienced electricians. With the dramatic increase of sensing data, nowadays, it becomes more and more desirable to make diagnosis automatic.

A fundamental step towards automatic diagnosis is the detection of electrical equipment in thermal images, via either segmenting equipment regions or localizing equipment with bounding boxes [1]–[3]. Considering that different parts of an electrical device may demonstrate different temperature patterns, instead of detecting each equipment as a whole, this work aims to detect equipment parts for better diagnosis. In contrast to general object detection in color images, our task has the follows characteristics: 1) Due to background distraction or improper settings of cameras, thermal images may present in over-centralized temperature distribution, resulting in low intensity contrast. 2) Most images are captured by hand-held cameras, in which equipment is not well aligned to be upright. Thus, equipment in images may be tilted slightly or even severely. 3) Different equipment may contain a various number of parts, but the parts of an equipment should share the same orientation angle regarding to the rigid-body property. These phenomena, together with cluttered background and variations in appearance, shape and scale, make our task highly challenging. Typical examples that contain the above-mentioned phenomena are demonstrated in Figure 1.

In this paper, we propose a deep convolutional neural network (CNN) based on YOLO [4], [5] to predict the coordinates, orientation angle, and class type of each equipment part. More specifically, we state our contributions as follows:

- To the best of our knowledge, the proposed method is the first one applying the deep learning technique for oriented electrical equipment detection. Our approach can provide location, orientation, and class type of each equipment part, thus making it convenient for the subsequent status diagnosis.
- We propose a way to integrate the orientation consistency prior into our model, by which the detection results are improved, especially for small-size equipment parts.

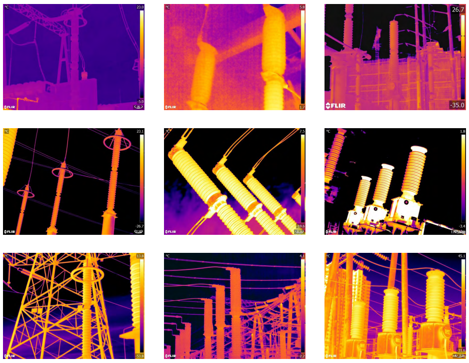


FIGURE 1. Different scenarios of equipment in thermal images. The first row shows images in low contrast; the second row presents tilted equipment; and the third row includes examples of cluttered background. Moreover, from the leftmost column to right, the electrical equipment types are surge arresters, current transformers, breakers, and potential transformers, respectively.

- We construct a large thermal image set containing four major types of electrical equipment, which makes it possible to train a deep learning based model and perform a thorough evaluation.

II. RELATED WORK

A. AUTOMATIC DIAGNOSE TECHNIQUES

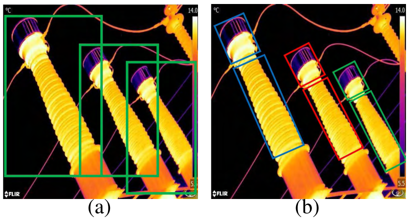
Automatic diagnosis is a goal that we have been pursuing for years. To this end, various techniques [6]–[12] have been developed in the last decades. Most of them share a common framework containing three stages: equipment detection, feature extraction, and status classification. For instance, Almeida *et al.* [9] applied the Watershed transformation to segment out surge arresters and designed a neuro-fuzzy network to classify conditions into faulty, normal, light, and suspicious categories. Zou and Huang [11] used the K-means clustering algorithm to segment electrical equipment, extracted statistical features, and used SVM for classification. These works presented prototype systems validated only in small data sets containing hundreds of simple images or even less. There is still a large gap to fill between them and real applications. Recently, deep learning approaches [13] are applied to fault diagnosis. These techniques can detect fault in an end-to-end manner so that they will gradually become the major research trend with the increase of sensing data.

B. ELECTRICAL EQUIPMENT DETECTION

As pointed out in [9]–[11], the success of diagnosis highly depends on the correct detection or segmentation of equipment. Therefore, a variety of research [1]–[3], [14]–[16] have been conducted to solve the detection problem. For instance, Jadin *et al.* [1] detected region of interests via extracting, matching, and clustering sparse feature points. Chen *et al.* [14] designed a local definition cluster complexity measurement to extract regions. Albalooshi *et al.* [2] employed the active contour model for segmentation. Zhao *et al.* [3], [16] used binary shape prior and feature pooling to detect insulators. Jadin and Taib [15] applied the Otsu method to threshold equipment regions. Most of these methods use hand-crafted features which are sensitive to variance and noise. Moreover, they detect the equipment as a whole and are validated only on a couple of simple images. In contrast, we detect each part of equipment. Our approach is validated in a large number of images and demonstrated a high robustness.

C. ORIENTED OBJECT DETECTION

Object detection is a key problem in computer vision. With the overwhelming success of deep learning, state-of-the-art detection methods such as R-CNN [17] and its variants [18], YOLO [4], [5], and SSD [19] are all based on deep neural networks. These methods have achieved high performance on detecting upright objects. However, as pointed out in [20], orientation changes may result in large variation in appearances and suffer from severe background distraction, leading to the failure of detection for oriented objects. Moreover, using upright bounding box to localize equipments may result in redundant background noise and unnecessary overlap, as shown in Figure 2 (a).



(a) (b)

FIGURE 2. (a) is an example using a traditional object detection method that predicts upright bounding boxes (marked in green). (b) demonstrates our part detection method predicting oriented bounding boxes. Boxes in different colors indicate different groups they are in.

Nowadays, more and more images are collected by hand-held cameras. Objects in these images are often not well aligned. Therefore, oriented object detection began to attract some research interest in recent years. For instances, He *et al.* [21] and Shi *et al.* [22] proposed different methods in detecting oriented text. In contrast to detect tilted text in color images, our equipment part detection is conducted in infrared thermal images which have lower contrast in intensity, thus making the predication more difficult. Moreover, almost all the methods in detecting oriented text treat a text line as a whole, while a piece of electrical equipment can be semantically divided into several parts. How to exploit the structural information between parts plays an essential role in our problem, which is not discussed in oriented text detection.

III. THE PROPOSED METHOD

When an image is given, we detect equipment parts via predicting a set of oriented bounding boxes, each of which is parameterized by its center coordinates, width, height, and orientation angle. To this end, we design a regression-based detection framework, which is built upon the powerful deep convolutional neural network similar to YOLO [4], YOLO9000 [5], and SSD [19]. In contrast to these existing

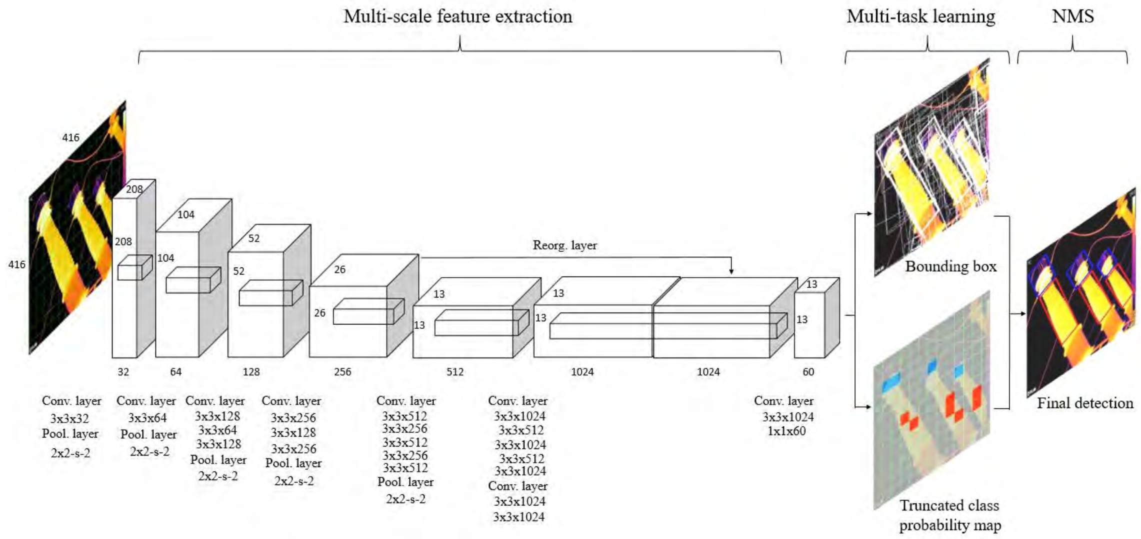


FIGURE 3. The overview of our framework. A deep convolutional neural network takes an iron-colored thermal image as input and outputs both oriented bounding boxes and associated class probabilities, followed by a non-maximum suppression (NMS) step to obtain final detection results. In the figure, Conv. layer refers to convolutional layer, Pool. layer denotes max-pooling layer, and Reorg. Layer is reorganization layer. Each layer stack contains one or multiple connected convolutional layers, followed with a pooling layer.

methods that predict upright bounding boxes, our approach additionally predicts the orientation angle and takes into account an orientation constraint between parts.

Figure 3 depicts the entire framework. A thermal image colored in the iron palette, which is the format collected during routine inspection, is first fed into a stack of convolutional layers and max-pooling layers for feature extraction. Then, the features obtained on different layers are concatenated through a reorganization layer and fed forward into two more convolutional layers. The output of the last layer produces both the oriented bounding boxes and associated class probabilities. Finally, a Non-Maximum Suppression (NMS) process is adopted to localize the parts of the highest probabilities. More details are introduced below.

A. THE NETWORK ARCHITECTURE

Our network architecture consists of six layer stacks for feature extraction. As listed in Figure 3, each of the first two stacks contains a convolutional layer and a max-pooling layer; the third or the fourth stack contains three convolutional layers followed by one max-pooling layer; the fifth is of five convolutional layers followed by one max-pooling layer; and the sixth is of seven convolutional layers. Each convolutional layer with a $k \times k$ kernel size and c channels is represented by $k \times k \times c$ in Figure 3; analogously, a pooling layer of $k \times k - s - n$ refers to the layer using the $k \times k$ kernel size and a n stride. The features extracted by the fourth and the sixth layer stacks are concatenated through a reorganization layer and fed forward into two more convolutional layers for parameter prediction. The size of each layer is described in Figure 3. The network takes a fixed-size 416 × 416 × 3 thermal image as the input. The image is divided into $S \times S$ grids and each grid cell predicts B bounding boxes. A box has five parameters delineating the coordinates, one confidence score reflecting how confident the box predictor is, together with K class probability values. Therefore, the output is a $S \times S \times (B \cdot (5 + 1 + K))$-dimensional tensor.

When predicting B bounding boxes in each cell, an anchor box scheme is adopted as in [18], [5], and [19]. That is, instead of predicting the coordinates directly, we predict five parameters $(t_x, t_y, t_w, t_h, t_\theta)$ that are related to an anchor box, as shown in Figure 4. Let us assume that a cell is offset from the top left corner of the image by (c_x, c_y) and the anchor box is of width w_a and height h_a. Then, the predicted bounding box (x, y, w, h, θ) in this cell can be calculated from the predicted parameters by

$$\begin{aligned} x &= \sigma(t_x) + c_x \\ y &= \sigma(t_y) + c_y \\ w &= w_a \exp(t_w) \\ h &= h_a \exp(t_h) \\ \theta &= t_\theta, \end{aligned} \tag{1}$$

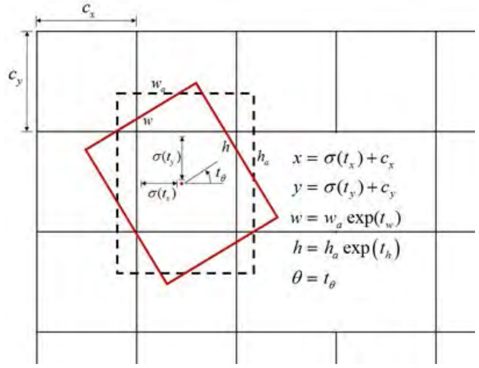


FIGURE 4. A bounding box (marked in red) that is delineated related to an anchor box (marked in dash line).

in which $\sigma()$ is a logistic function scaling t_x or t_y into [0, 1]. We use the K-means algorithm to cluster B anchor boxes, that is obtaining B different box sizes, from training samples. By this means, we are able to incorporate the size prior of equipment parts into our model, making the network stable for training and achieving better localization results.

B. LOSS FUNCTION

When training our network, a set of thermal images are given as training samples. Each equipment part in the images is annotated with an oriented bounding box and a class label as ground truth. The network is trained to predict both localization and classification results, and thus it is a multi-task learning problem. Therefore, we design a multi-task loss function $\mathcal{L}$ that includes a localization loss $\mathcal{L}_{loc}$, a classification loss $\mathcal{L}_{cls}$, together with a loss $\mathcal{L}_{ort}$ constraining the orientation angle between parts to be consistent. That is,

$$\mathcal{L} = \mathcal{L}_{loc} + \mathcal{L}_{cls} + \mathcal{L}_{ort}. \tag{2}$$

1) LOCALIZATION LOSS

Let's denote the location of a predicted bounding box by $\mathbf{t} = (t_x, t_y, t_w, t_h, t_\theta)$. Although multiple bounding boxes are predicted, we expect only one that has the highest Intersection over Union (IOU) with the ground truth to be 'responsible' for predicting an equipment part. Thus, we use an indicator $\mathbf{1}_{ij}^{obj}$ to denote if the j-th predicted bounding box in the i-th cell is responsible for a part and $\mathbf{1}_{ij}^{noobj}$ to denote if not. Only those responsible boxes matter in localization learning. Meanwhile, a confidence score s_{ij} is also provided to measure how confident this box predictor is. The score is expected to be high if the box is responsible, otherwise it is of a low confidence.

To this end, the localization loss is designed as follows:

$$\begin{aligned}\mathcal{L}_{loc} = &\sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{1}_{ij}^{obj}(s_{ij} - \hat{s}_{ij})^2 \\ &+ \lambda_{noobj}\sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{1}_{ij}^{noobj}(s_{ij} - \hat{s}_{ij})^2 \\ &+ \sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{1}_{ij}^{obj}||\mathbf{t}_{ij} - \hat{\mathbf{t}}_{ij}||_2^2,\end{aligned} \tag{3}$$

where the values with '$\hat{}$' denote the ground truth. The ground-truth confidence score $\hat{s}_{ij}$ is set to be 1 if $\mathbf{1}_{ij}^{obj} = 1$, otherwise it is 0. Moreover, S^2 is the total number of grid cells, B is the number of boxes predicted in each cell, and $||\cdot||_2^2$ is a L_2 norm. λ_{noobj} is a parameter used to decrease the confidence loss from those 'irresponsible' boxes.

2) CLASSIFICATION LOSS

In our task, we have K types of electrical equipment parts. The constructed network predicts a K-dimensional vector $\mathbf{p}$ for each bounding box. Each entry of $\mathbf{p}$ represents the probability for a predicted bounding box belonging to a certain class. The classification loss takes into account only those 'responsible' bounding boxes as well. Therefore, it is defined as follows:

$$\mathcal{L}_{cls} = \sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{1}_{ij}^{obj}||\mathbf{p}_{ij} - \hat{\mathbf{p}}_{ij}||_2^2, \tag{4}$$

where $\hat{\mathbf{p}}$ is a K-dimensional binary vector denoting the classification ground truth. It is of all 0 entries except the one corresponding to the labeled class, which is set to be 1.

3) ORIENTATION CONSISTENCY LOSS

When detecting different parts of one equipment such as current transformer or potential transformer, these parts should share the same orientation angle because they belong to a rigid body. This prior can be adopted to improve the prediction of bounding boxes. Thus, we propose a consistency loss to constrain the orientation angle between parts.

Before placing this constraint, we first need to determine which parts are on the same equipment. Thus, we first divide the annotated bounding boxes into different groups and each group indicates a single equipment instance, as shown in Figure 2(b). This can be done by checking the orientation angles of parts according to the following condition. When given two annotated bounding boxes a and b, they belong to the same group if

$$|t_{\theta_a} - t_{\theta_b}| < \delta_1 \quad \text{and } |t_{\bar{\theta}} - t_{\check{\theta}}| < \delta_2, \tag{5}$$

where δ_1 and δ_2 are thresholds manually set. t_{θ_a} and t_{θ_b} denote the angle of a and b, respectively. $t_{\bar{\theta}}$ denotes the average orientation angle of a and b, and $t_{\check{\theta}}$ denotes the orientation of the line concatenating the center of part a and part b.

Now, we have all parts divided into different groups. Assuming that an image contains G groups of equipment parts, and each group may contain a various number of parts. We denote the g-th group as Ω_g. Then, the orientation consistency loss is formulated in the following:

$$\mathcal{L}_{ort} = \sum_{g=1}^{G}\sum_{i=1}^{S^2}\sum_{j=1}^{B} \mathbf{1}_{ij}^{obj}\mathbf{1}_{ij}^{\Omega_g}(t_{\theta_{ij}} - t_{\bar{\theta}_{\Omega_g}})^2. \tag{6}$$

Here, $\mathbf{1}_{ij}^{\Omega_g}$ indicates if the j-th predicted bounding box in cell i is responsible for a part in Ω_g, and $t_{\bar{\theta}_{\Omega_g}}$ denotes the average orientation angle of all parts in the group.

C. TRAINING AND TESTING

1) TRAINING

We use aforementioned loss function to train our network in an end-to-end manner. Training images and their annotations are fed into the network. The loss is optimized by Stochastic Gradient Descent (SGD) with a batch size of 16, momentum of 0.9, and weight decay of 0.0005. We adopt the multi-step strategy in Caffe [23] to adjust our learning rate. For the first 35000 iterations, the learning rate is fixed to 0.01; it is reduced to 0.001 afterwards. Since there is no existing model

TABLE 1. The number of different parts, equipments, and images in the dataset.

Equipment Part	Current transformer	Potential transformer	Arrester	Breaker	Total No. of parts
Bushing	4621	3284	2993	3205	14103
Bellows	4621	0	0	0	4664
Grading ring	0	956	2993	0	3949
Bushing coupler	0	668	0	0	668
Flange	0	3476	3657	0	7151
Arc-extinguishing chamber	0	0	0	3672	3672
Total No. of equipment (No. of images)	4621(2209)	3284(2324)	2993(2264)	3205(1158)	14104(7955)

pre-trained on electrical device dataset, we randomly initialize our model and train it from scratch.

2) TESTING

In testing, an image alone is fed into the trained network. The output produced in the last layer predicts the location, the confidence, and the class probabilities. We multiply the class probabilities with the confidence to indicate the box confidence prediction. The boxes with the highest confidence prediction are selected by NMS as the final results.

IV. EXPERIMENTS

In this section, we first introduce a thermal image dataset that we have constructed. Then a series of experiments are conducted to validate the performance of the proposed approach.

A. DATASET

As far as we know, there is no thermal image dataset publicly available for electrical equipment detection and diagnosis. In order to evaluate the proposed method, we construct a dataset by ourselves. Images are captured by hand-held thermal cameras such as FLIR T640, FLIR T660, and FLIR P660 [24] when electricians perform routine inspection in a number of transformer substations in Shandong province, China. This dataset focuses on four major types of transformation equipments, including Current Transformer (CT), Potential Transformer (PT), Surge Arrester (SA), and Circuit Breaker (CB). Typical examples are illustrated in Figure 1, from which we can observe that each type presents large variations in appearance, viewpoint, and background.

In order to train our network and evaluate testing results, we manually annotate 7955 images. Each equipment part in the images is annotated with an oriented bounding box and a class label. According to the composition of these equipments, we categorize all parts into six classes: bushing, bellows, grading ring, bushing coupler, flange, and arc-extinguishing chamber. For instance, a current transformer is composed of a bellows and a bushing, but a potential transformer may consist of various number of bushing coupler, flange, grading ring and bushings. Table 1 lists the number of parts and equipments existing in all images. The distribution of the orientation angle for all parts is also illustrated in Figure 5. Since we take the positive x-axis as the reference 0° angle, the orientation distributes from 65° to 115° and the majority is upright.

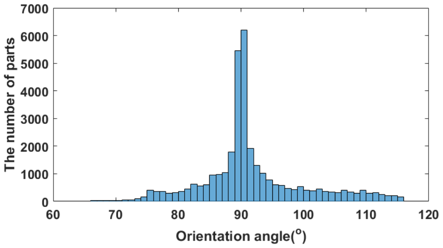


FIGURE 5. The orientation distribution of all equipment parts in our dataset.

B. EXPERIMENTAL SETUP

In the dataset, we randomly select 60% images for training and the rest for testing. The original resolution of an image is 480 × 640 or 640 × 480. Both are scaled into 480 × 480 to fit the input size. Meanwhile, we augment training data by the following means: each training image is randomly cropped into a 416 × 416 patch and horizontally flipped with probability of 0.5, in addition to applying some shift in hue, saturation and exposure.

Throughout all experiments, we adopt the default parameter settings in [5], i.e. $S = 13$, $\lambda_{noobj} = 0.2$, and we set $K = 6$ to represent the six equipment part classes. K-means algorithm is employed to cluster $B = 5$ anchor boxes, which are obtained in size of 42×195, 92×66, 81×120, 38×299, and 38×36.

The experiments are conducted on a desktop running with Ubuntu 14.04.3 with one Nvidia GeForce GTX 1080 GPU. It takes 10 hours or so to train the model. For testing, it reaches 20 frames per second.

C. EXPERIMENTAL RESULTS

1) COMPARISON TO OTHER METHODS

We first carry out an experiment using the proposed full model and compare it with the state-of-the-art methods that were developed for upright object detection, which include YOLO9000 [5], SSD [19], and Faster R-CNN [18]. The results are evaluated with respect to Average Precision (AP) as defined in [25], which is a criterion measuring the area under the precision-recall curve for a class. When computing the precision and recall, we judge a predicted bounding box to be true positive if the Intersection over Union (IoU) with ground truth is greater than a threshold. The threshold is commonly set to be 0.5 as in most object detection work.

TABLE 2. Experimental results compared with other methods regarding to AP(%) and mAP(%).

Part Method	Bushing	Bellows	Bushing coupler	Grading ring	Flange	Arc-extinguishing chamber	mAP(%)	FPS
FRCNN [18]	**97.2**	**98.3**	**95.4**	89.7	44.1	97.2	87.0	11
SSD [19]	87.6	90.1	88.0	84.2	80.2	90.0	86.7	46
YOLO9000 [5]	91.1	96.8	80.9	89.2	82.7	92.6	88.9	59
Proposed	95.9	96.1	87.7	**93.5**	**91.8**	**97.5**	**93.7**	20

TABLE 3. Ablation experimental results compared regarding to mAP(%).

IoU_thresh Model	0.5	0.55	0.6	0.65	0.7	0.75	0.8
Proposed w/o	93.2	91.3	87.6	82.4	74.5	61.1	44.8
Proposed	**93.7**	**91.8**	**89**	**84.4**	**76.3**	**64.2**	**47.9**

TABLE 4. Ablation experimental results compared regarding to AP(%) and mAP(%).

Part Method@(IoU)	Bushing	Bellows	Bushing coupler	Grading ring	Flange	Arc-extinguishing chamber	mAP(%)
Proposed w/o@0.5	95.6	96.5	84.6	93.7	92.1	96.9	93.2
Proposed@0.5	95.9	96.1	87.7	93.5	91.8	97.5	93.7
Proposed w/o @0.7	86.9	87.7	52.2	80.2	53.3	86.4	74.5
Proposed@0.7	87.6	88.9	58.0	82.3	53.9	86.6	76.3

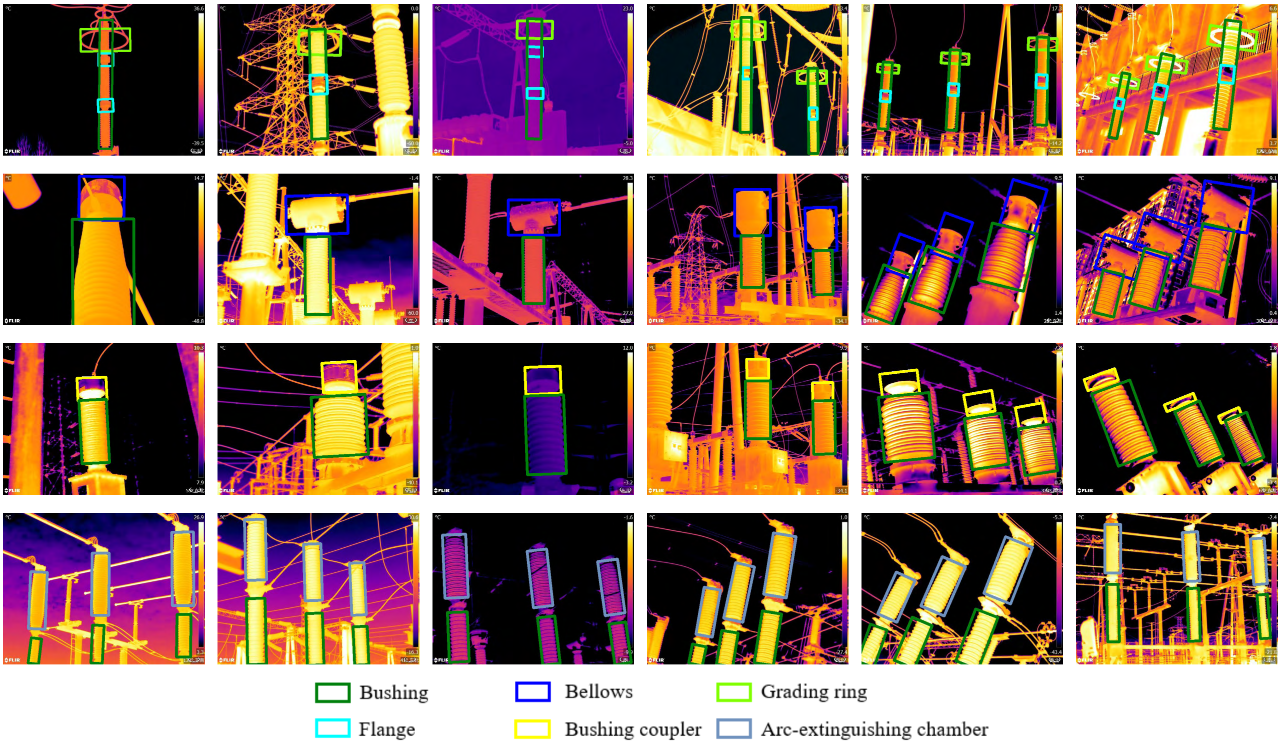


FIGURE 6. The testing results obtained by our proposed full model. From top to bottom, the equipment is SA, CT, PT, and breaker, respectively.

Table 2 presents the AP for each class, together with the mean Average Precision (mAP) for all classes. The results show that our proposed method achieves the highest mAP, outperforming all other upright object detection methods. This improvement is mainly benefited from the consideration of rotation in our model, by which our model is more robust to the appearance variance caused by rotation and suffered less from background noise.

Table 2 also lists the time comparison in terms of frames per second (FPS). Since non-maximum suppression of oriented bounding-boxes is a little different from that of upright ones when computing IoU, we also take the time of non-maximum suppression into consideration when estimating FPS for fair comparison. Our full model can reach 20 FPS which is faster than Faster R-CNN. Due to the complexity in non-maximum suppression of oriented bounding-boxes, our full model is slower than YOLO9000 and SSD.

2) ABLATION EXPERIMENTS

We further conduct an experiment to investigate the effectiveness of the proposed orientation consistency loss. To this end, the proposed model without the orientation loss is trained and tested as well (this model is denoted as "Proposed w/o"). The results are also evaluated with respect to mAP. In order to evaluate the localization performance better, we report the results with an IoU threshold varied from 0.5 to 0.8. Table 3 list the mAP under different thresholds for the full model and the one without the orientation constraint. Table 4 also presents the AP for each class with an IoU of 0.5 and 0.7, respectively. From these results we can make the following observations:

- The mean average precision gradually drops when the IoU threshold is raised from 0.5 to 0.8 because higher IoU requires more precise localization results. When comparing the proposed full model to the one without orientation loss, we observe that the full model consistently outperforms the other model. The mAP is increased by 0.5% when IoU is 0.5 and increased to 3.1% along with the increase of IoU. It implies that the orientation constraint between parts offers better localization performance.
- By comparing APs for each part class, we observe that AP is improved more for those equipment parts such as bellows, bushing coupler, and grading ring that have a small ratio of height to width. The reason is that the prediction of orientation for these parts is more sensitive to noise than other types. With our orientation consistency loss, we achieve more robust prediction results.

Figure 6 presents the typical results detected on the test set using our proposed full model. It demonstrates that our approach can detect parts precisely, robust to variations in size, temperature, and orientation angles. Promising detection results are achieved even when the background is cluttered.

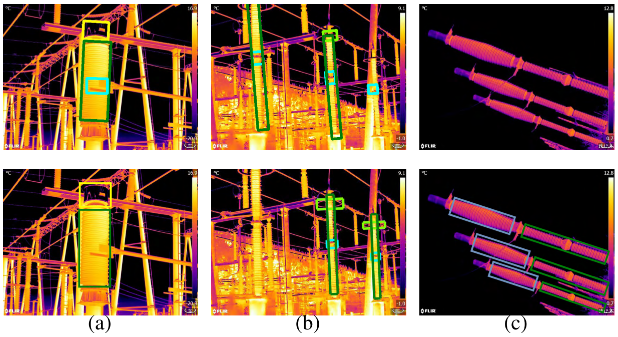


FIGURE 7. Failed cases. The top row presents the detection results and the bottom shows the ground truth.

However, there are still a small amount of failed cases. Typical failures are presented in Figure 7. (a) and (b) are the incorrect or missing detection caused by background noise, and (c) is the missing detection due to sever orientation and the lack of similar training data. The first two cases can be improved by considering the co-occurrence between parts and the latter might be improved by increasing the similar training samples.

V. CONCLUSION

Equipment detection is a fundamental step towards automatic inspection and diagnosis. Inspired by the recent success of deep learning techniques, we have proposed an approach to detect fine-grained equipments in thermal images. Our approach is able to detect equipment parts no matter they are upright or tilted, by predicting boxes tightly bounded. The approach has been validated in a large dataset that we constructed. The experiments show that our approach is promising. We believe that our work can benefit the subsequent diagnosis, which is also the task we will carry out in the future.

REFERENCES

[1] M. S. Jadin, K. H. Ghazali, and S. Taib, "Detecting ROIs in the thermal image of electrical installations," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng.*, Nov. 2014, pp. 496–501.

[2] F. A. Albalooshi, E. Krieger, P. Sidike, and V. K. Asari, "Efficient thermal image segmentation through integration of nonlinear enhancement with unsupervised active contour model," *Proc. SPIE*, vol. 9477, p. 94770C, Apr. 2015.

[3] Z. Zhao, N. Liu, and L. Wang, "Localization of multiple insulators by orientation angle detection and binary shape prior knowledge," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 6, pp. 3421–3428, Dec. 2015.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[5] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6517–6525.

[6] S. A. Merryman and R. M. Nelms, "Diagnostic technique for power systems utilizing infrared thermal imaging," *IEEE Trans. Ind. Electron.*, vol. 42, no. 6, pp. 615–628, Dec. 1995.

[7] A. T. P. So, W. L. Chan, C. T. Tse, and K. K. Lee, "Fuzzy logic based automatic diagnosis of power apparatus by infrared imaging," in *Proc. 2nd Int. Forum Appl. Neural Netw. Power Syst. (ANNPS)*, 1993, pp. 187–192.

[8] Y. Chou and L. Yao, "Automatic diagnosis system of electrical equipment using infrared thermography," in *Proc. Int. Conf. Soft Comput. Pattern Recognit.*, 2009, pp. 155–160.

[9] C. A. L. Almeida *et al.*, "Intelligent thermographic diagnostic applied to surge arresters: A new approach," *IEEE Trans. Power Del.*, vol. 24, no. 2, pp. 751–757, Apr. 2009.

[10] M. S. Jadin and S. Taib, "Recent progress in diagnosing the reliability of electrical equipment by using infrared thermography," *Infrared Phys. Technol.*, vol. 55, no. 4, pp. 236–245, 2012.

[11] H. Zou and F. Huang, "A novel intelligent fault diagnosis method for electrical equipment using infrared thermography," *Infr. Phys. Technol.*, vol. 73, pp. 29–35, Nov. 2015.

[12] A. Rahmani, J. Haddadnia, and O. Seryasat, "Intelligent fault detection of electrical equipment in ground substations using thermo vision technique," in *Proc. 2nd Int. Conf. Mech. Electron. Eng.*, vol. 2, Aug. 2010, pp. V2-150–V2-154.

[13] G. Zhao, G. Zhang, Q. Ge, and X. Liu, "Research advances in fault diagnosis and prognostic based on deep learning," in *Proc. Prognostics Syst. Health Manage. Conf.*, 2016, pp. 1–6.

[14] C. Chen, W. Qin, Z. Fang, and Y. Zhang, "Infrared image transition region extraction and segmentation based on local definition cluster complexity," in *Proc. Int. Conf. Comput. Appl. Syst. Modeling*, vol. 3, Oct. 2010, pp. V3-50–V3-54.

[15] M. S. Jadin and S. Taib, "Infrared image enhancement and segmentation for extracting the thermal anomalies in electrical equipment," *Elektronika Elektrotechnika*, vol. 120, no. 4, pp. 107–112, 2012.

[16] Z. Zhao, G. Xu, and Y. Qi, "Representation of binary feature pooling for detection of insulator strings in infrared images," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 5, pp. 2858–2866, Oct. 2016.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[19] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[20] S. He and R. W. Lau, "Oriented object proposals," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 280–288.

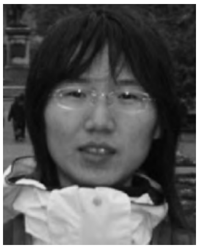
[21] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu. (2017). "Deep direct regression for multi-oriented scene text detection." [Online]. Available: https://arxiv.org/abs/1703.08289

[22] B. Shi, X. Bai, and S. Belongie. (2017). "Detecting oriented text in natural images by linking segments." [Online]. Available: https://arxiv.org/abs/1703.06520

[23] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[24] *FLIR Systems: Thermal Imaging Guidebook for Industrial Applications*, FLIR Syst., Inc., Wilsonville, OR, USA, 2011.

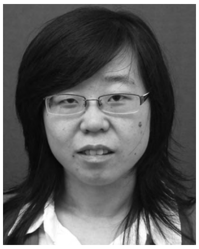
[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.



XIAOJIN GONG received the B.A. and M.A. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from Virginia Tech, USA, in 2009. She is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Her research interests include computer vision, image processing, and artificial intelligence.

QI YAO was born in Taihe, Jiangxi, China, in 1995. He received the B.E. degree in information science and electronic engineering from Zhejiang University, Hangzhou, Zhejiang, in 2017, where he is currently pursuing the master's degree in information and communication engineering. His main research interests include computer vision and deep learning.

MENGLIN WANG received the B.E. degree from Xidian University, Xi'an, China. She is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Her research interests include computer vision and machine learning.



YING LIN received the B.E. and Ph.D. degrees in information engineering from Zhejiang University, Hangzhou, China. Her research interests include computer vision, image processing, and artificial intelligence.