A Novel Method for Thermal Image Based Electrical-Equipment Detection

Chapter · October 2019

DOI: 10.1007/978-3-030-31654-9_21

Xiao Wang
Anhui University

A novel method for thermal image based electrical-equipment detection

Futian Wang[1,2], Songjian Hua[1], Xiao Wang[1], Zhengzheng Tu[1], Cheng Zhang[1], and Jin Tang[1]⋆

[1]School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[2]Key Lab of Industrial Image Processing & Analysis of Anhui Province, Hefei, 230601, China
wft@ahu.edu.cn, hsj928@foxmail.com, wangxiaocvpr@foxmail.com, zhengzhengahu@163.com, cheng.zhang@ahu.edu.cn, tj@ahu.edu.cn

Abstract. An accurate and reliable thermal image based electrical-equipment detection is critical in smart power grids such as automatic defect diagnosis. However, few works have provided solutions to the task. To solve the problem, in this paper, we propose a new task named thermal image based electrical-equipment detection which includes two contributions. First, we have created a large-scale thermal electrical-equipment benchmark from 5558 thermal images which were taken during electrical-equipment inspection in reality. Second, we used the self-attention mechanism to get better detection performance. We have made some improvements based on Dual Attention Network (DANet) and applied it to further improve feature representation, we named our method Channel-Position Dual Attention Network (CPDANet). The experiment results show that our novel method can improve the mean Average Precision (mAP) from Faster R-CNN's 89.9% to 91.4%. The project page of this paper can be found at: https://sites.google.com/view/electrical-equipment-detection/.

Keywords: Object detection · Benchmark dataset · Self-attention

1 Introduction

In this paper, we propose a new visual task named thermal image based electrical-equipment detection. Object detection is a fundamental and challenging problem in the computer vision community, whose goal is to identify all objects of interest in an image and determine their location and size. Thermal image based electrical-equipment detection is a important sub-task of object detection, it also is a critical component of smart power grids. Typically, at present, in the process of defect diagnosis, a equipment need to be manually marked, and then judging whether the equipment is defective according to the temperature information. Therefore, accurate detection of each type of electrical-equipment is the

⋆ Corresponding Author

premise of automatic defect diagnosis. That is the equipment can be automatically marked and calculated relative temperature information, automatic defect diagnosis of the equipment can be achieved. Once a equipment is detected to be defective, the workers can timely repair the defect and avoid economic losses and changes in life, which significantly improve efficiency and safety of power grids. Sample detections over test set are illustrated in Fig. 1.

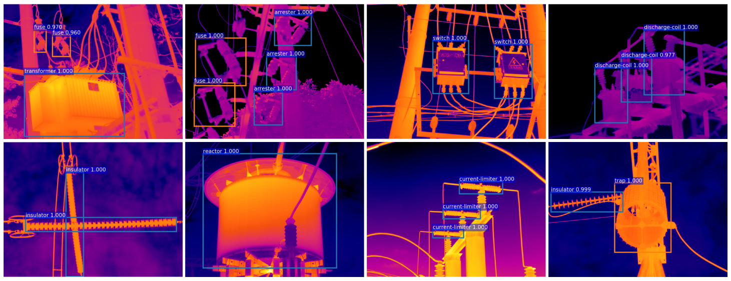

Fig. 1. Sample detections over the test set.

Traditional object detection methods are based on handcrafted features and easily stagnates their performance. Recently, with the development of deep learning, these problems existing in traditonal methods have been solved. Deep learning methods have shown superior performance for many tasks including object detection. Convolutional neural networks (CNNs) as one particular variant of deep neural networks, have shown their superiorities for many tasks including detection. PASCAL VOC [2] and ImageNet ILSVRC [20] are widely used to evaluate detection performance.

Object detection research has reached a very high level, such as vehicle detection [34], traffic-sign detection [35], pedestrian detection [15], face deteciton [24], person head detection [27], and so on. However, there is little research on thermal image based electrical-equipment detection. To solve this problem, we have created a realistic thermal electrical-equipment benchmark. Images in benchmark cover large variations in defect types (such as normal and defective, there are three criterias to judge whether the equipment is defective, temperature difference, relative temperature difference and hot spot temperature), season conditions (the change of season will cause the change of air temperature and also affect temperature of equipments, which will affect visual effect of images) and diurnal conditions (diurnal variations also will cause the change of air temperature and affect temperature of equipments).

Particularly, we analyzed our dataset and found that the dataset has one characteristic: there is a strong correlation between different equipments, i.e. the appearance of a equipment is often accompanied by the appearance of one or more other specific kinds of equipments. So, capturing contextual dependencies and integrating local features with their global dependencies by self-attention mechanism is a good way to get better detection performance. However, most methods do not focus on this.

To address above problem, we have made some improvements based on DANet [3], and applied it to integrate local features with their global dependencies, this have resulted in the better detection performance. The framework is illustrated in Fig. 4.

The primary contributions of this paper are as follows.

- We have created a new, realistic thermal electrical-equipment benchmark. The electrical-equipments in our benchmark cover real-world conditions which are large variations in such aspects as defect types, season conditions and diurnal conditions, examples of occlusion and incomplete equipment are also included. Our benchmark is annotated with a bounding box for each electrical-equipment, as well as giving it's class label. We call this benchmark TEED001. This benchmark will be open to public later.
- We have made some improvements based on DANet [3] and proposed CP-DANet and applied it to further improve feature representation. The experiment results show that improvements can improve detection performance and robust.

2 Related Work

According to the relevance to our work, we review related works following two research lines, i.e., Object Detection by CNNs and Attention modules.

2.1 Object Detection by CNNs

The CNNs was initially rekindled by the use of image classification in [10], and adapted to object detection quickly. It is observed in OverFeat [21] that the use of a convolutional network in the sliding window fashion is inherently efficient in nature by Sermanet et al., because many calculation can be reused in the overlapping region.

For using CNNs to object detection, another widely used strategy is to calculate some generic object proposals firstly and then classify only on these candidates. The first to use this strategy is R-CNN [5], but the following two reasons lead to the slow speed of it. First, it is costly to generate object proposals which is category-independent. Generating 1000 proposals for the Pascal VOC 2007 images, Selective search [25] need about 3 s, EdgeBoxes approach [36] which is more efficient still need about 0.3 s. Second, since each candidate proposal applies the deep convolution network,which increases the time cost. To improve efficiency, Kaiming He et al. proposed spatial pyramid pooling network (SPP-Net) [6], which increases the speed of the R-CNN by about 100 times.

Then, based on R-CNN, Girshick et al. proposed Fast R-CNN [4], which didn't uses the SVM classifier used in R-CNN, but uses a softmax layer above the network instead. Because of ignoring object proposal time, Fast R-CNN processing one image takes 0.3 seconds. In order to overcome the bottleneck in the object proposal step, in Faster R-CNN [19], Ren et al. proposed region proposal networks (RPNs) which use convolutional feature maps to generate object proposals. This allows the object proposal generator to share full-image convolutional features with the detection network.

Compared with the two-stage detection framework mentioned above, the one-stage detection framework has more advantages in speed, such as SSD [16] and YOLO [17]. Wei Liu et al. proposed Single Shot MultiBox Detector (SSD) [16] in 2016. SSD is based on VGG-16 [23] backbone network, and it is ends with extra convolutional layers. In the same year, Joseph Redmon et al. proposed You Only Look Once: Unified, Real-Time Object Detection (YOLO) [17], which frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities.

2.2 Attention modules

Vaswani et al. proposed Attention is all you need [26] in 2017, which is the first work to propose the self-attention mechanism and apply it in machine translation. After this, attention mechanism have been widely applied in the field of Natural Language Processing (NLP). Attention modules can model long-range dependencies and directly draw global dependencies of inputs. Such a mechanism is dispensed with recurrence and convolutions entirely, which improves the parallelism and efficiency of model training. This attention mechanism has been extended in many NLP applications, such as natural language inference [22], text representation [12], sentence embedding [14] and so on.

Meanwhile, attention mechanism is also widely used in the field of computer vision. For example, the work relation networks for object detection [8] proposed by Huet et al. in 2018. They proposed an object relation module to model the relationships among a set of objects, which improves object recognition. Self-attention generative adversarial networks [33] was proposed by Zhang et al. in 2018 for better image generation, because it introduces self-attention modules, which can efficiently find global dependencies within internal representations. Jun Fu et al. proposed Dual Attention Network for Scene Segmentation [3] in 2018, this paper captures rich contextual dependencies based on the self-attention mechanism and proposes DANet adaptively integrate local features with their global dependencies and contributes to more precise segmentation results. Wang et al. also explore the attention weights in moving object detection [11], dual-modal saliency detection [30] and visual tracking [28] [32] [29].

3 Benchmark

This section will introduce the details of the newly created benchmark dataset, called TEED001 in this paper, including dataset collection, annotation, statistics and baseline approaches.

3.1 Data Collection

It is a popular method to generate image datasets by downloading Internet images retrieved by search engines using keywords, many widely used datasets have been generated in this way, such as ImageNet ILSVRC [20] and Microsoft COCO [13]. However, there are few images of real-world thermal electrical-equipment on the Internet, even if they exist, the electrical-equipments are incidental: such images will not be tagged with the names of any equipments. Such an approach can't be used here. So, ideal way to generate images dataset is to collect useful thermal electrical-equipment images from lots of real world images taken during equipment inspection in reality by FLIR, DALI and FLUKE thermal imager.

3.2 Data Annotation

After collecting images, the next step is to annotate these images by hand. Our image dataset contains three three regions of the power system: substation, transmission line and distribution line. A substation is an electrical system with high-voltage capacity. In order to transport the power from the power plant to a remote place, the voltage must be increased to become a high-voltage power, next the voltage should be lowered as the user requires. Normally, substations mainly include Step-up Type Substation and Step-down Substation, and so on. Transmission line uses a transformer to boost the electric energy generated by the generator, and then accesses the transmission line through a control device such as a circuit breaker. Distribution line refers to the line that sends power from the Step-down Substation to the Distribution Transformer or sends the power of the Distribution Substation to the power unit. During electrical-equipment annotation, we recorded the bounding box and class label. Equipment annotation case is similar to Fig. 1.

3.3 Data Statistics

After random selection, our new benchmark has 5558 images, These iamges contains 21 classes, 11180 electrical-equipment instances in total. There is an imbalance between different classes of electrical-equipment in our benchmark. Because some electrical-equipments are just rarely used. Instances per class are given in Fig. 2; most instances appear in relatively few classes. The image sizes (in pixels) of the electrical-equipments is given in Fig. 3; note that large electrical-equipments are most common, because in the actual shooting process, the lens will be zoomed.

In summary, the benchmark we created provides detailed annotation for each equipment: it's bounding box and class label. All images in this benchmark have resolution 640×480. And cover large variations in temperature conditions. It will hopefully provide a suitable basis for research into detecting thermal electrical-equipment. We created the benchmark for this purpose.

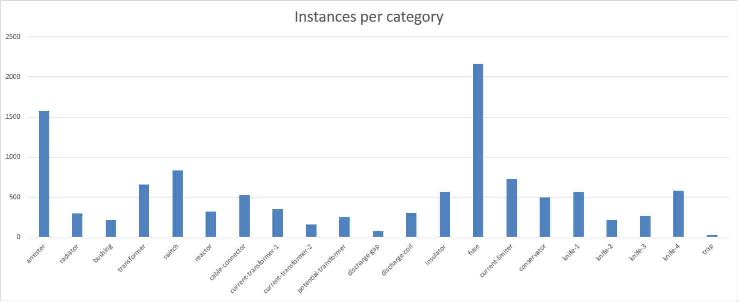

Fig. 2. Number of instances in each class.

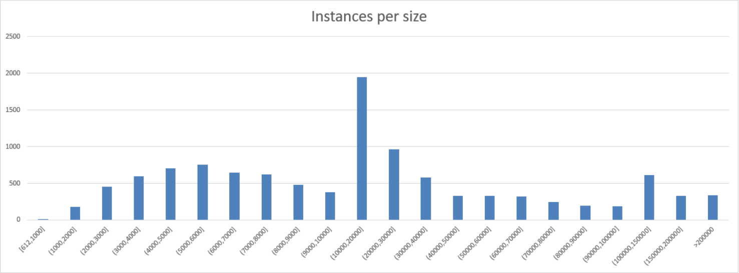

Fig. 3. Number of instances of each size.

3.4 Baseline Approaches

In order to verify the validity of our dataset, we have selected several advanced detectors available as baseline approaches for evaluation, which include Fast R-CNN [4], YOLOv2 [18], SSD [16], R-FCN [1], Faster R-CNN [19]. We experimented on these advanced baseline approaches and their mAP reached 79.9%, 81.7%, 89.7%, 89.1% and 89.9% respectively. The detail of results are shown in Table 1.

4 Proposed Approach

Both SE-Net [9] and CBAM [31] proposed channel attention and spatial attention, but they are implemented differently, and the way to get the final feature representation is also different. In addition, Dual Attention Network for Scene Segmentation [3] learned from two approaches above and proposed DANet. Based on the superiorities of the three approaches, we propose our own network architecture CPDANet. We use channel attention and position attention of DANet to get final feature representation. The architecture is shown in Fig. 4.

The implementation details of channel attention module and position attention module refer to Dual Attention Network for Scene Segmentation [3]. The structure of channel attention module is illustrated in Fig. 4(A). The structure

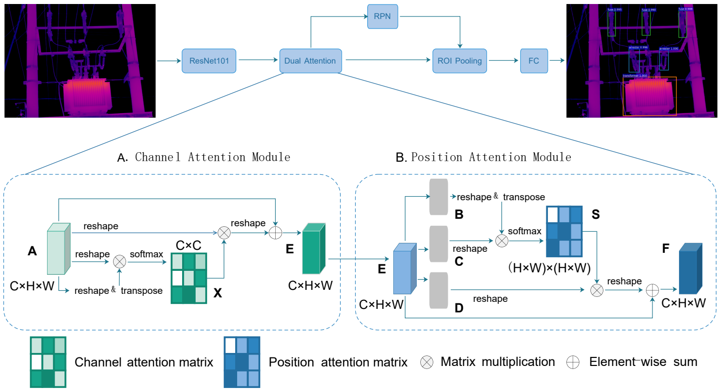

Fig. 4. Overall framework. And the details of Channel Attention Module and Position Attention Module are illustrated in (A) and (B).

of position attention module is illustrated in Fig. 4(B). In [3], channel attention module and position attention module are in parallel. Original features $A \in \mathbb{R}^{C\times H\times W}$ is used as input of channel attention and position attention, then fusing the outputs of the two attention branches in the way of element-wise summation to get the final feature maps.

In CPDANet, we don't fuse the outputs of the two attention branches in the way of element-wise summation used in Dual Attention Network for Scene Segmentation [3], we adapt a serial structure. We use CBAM [31] as reference. Firstly, we input the features $A \in \mathbb{R}^{C\times H\times W}$ obtained by ResNet-101 [7] into channel attention, and obtain the final output of channel attention module $E \in \mathbb{R}^{C\times H\times W}$. The final features E of each channel is a weighted sum of the features of all channels and original features, which models the long-range semantic dependencies between feature maps. It emphasizes class-dependent feature maps and helps to boost feature discriminability. Then we input E to position attention module, and obtain the final output $F \in \mathbb{R}^{C\times H\times W}$ of CPDANet as follows:

$$F_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + E_j \tag{1}$$

where s_{ji} measures the i^{th} position's impact on j^{th} position, α is initialized to 0 and gradually learned and assigned to a greater weight [33]. The Eq.(1) show that the resulting features F at each position is a weighted sum of the features at all positions and the output of channel attention module. Therefore, it has a global contextual view and selectively aggregates contexts according to the position attention map.

In summary, CPDANet can capture rich contextual dependencies which can get the better detection performance. The results of the two methods are shown in Table 1(Faster R-CNN+DANet and Faster R-CNN+CPDANet). We can get that our method is more advantageous.

Particularly, we reverse the order of channel attention and position attention and input the features obtained by ResNet-101 [7] into position attention and get the output $E' \in \mathbb{R}^{C\times H\times W}$ firstly, then use E' as the input of channel attention, and finally get the further improved feature representation through channel attention, it's named P-C Dual Attention Network (PCDANet). The final output $F' \in \mathbb{R}^{C\times H\times W}$ as follows:

$$F'_j = \beta \sum_{i=1}^{C} (x_{ji} E'_i) + E'_j \tag{2}$$

where x_{ji} measures the i^{th} channel's impact on the j^{th} channel, β is initialized to 0, and gradually learned and assigned to a greater weight. However, we got a very poor detection performance in this method. The experimental results are shown in Table 1(Faster R-CNN+PCDANet).

5 Experimental Results

We carry out comprehensive experiments on TEED001 dataset for evaluating the proposed method. Experimental results demonstrate that the proposed method achieves state-of-the-art performance on our dataset. Next, the detailed experimental process will be introduced.

5.1 Dataset And Evaluation Metrics

Dataset: Our algorithm is evaluated on our own dataset: TEED001. We separated TEED001 into a training set and a testing set. Training set contains 4448 images, and testing set contains 1110 images with about 4:1 ratio to give the deep learning methods plenty of training samples.

Evaluation Metrics: There are many evaluation metrics for object detection: accuracy, confusion matrix, precision and recall, Average Precision (AP), mAP etc. In this paper, we adopt mAP, which is the most widely used evaluation metric. And it is the average of multiple categories of AP, it's value must be in the [0,1] interval, the larger the better.

5.2 Implementation Details

We implemented our method based on Pytorch. We employed the most commonly used learning rate policy, i.e., every time reach a certain number of iterations, we multiplied the current learning rate by the learning rate decay. The learning rate is initialized to 0.001, learning rate decay is set to 0.1. In addition, momentum and weight decay coefficients are set to 0.9 and 0.0001 respectively, batchsize is set to 2 for our dataset. Furthermore, all the parameter settings are available in the source code to be released for accessible reproducible research. Particularly, in our experimental evaluations of our neural network, both training and testing were done on a Linux PC with an Intel(R) Core(TM) i7-7700K 4.2GHz CPU with 32GB RAM, one NVIDIA GeForcd Gtx 1080 GPU and 8GB memory.

Table 1. mAP on the TEED001 test dataset produced by all experiments. (All class names use abbreviations. Comparison of abbreviations and full names of all classes is illustrated in Table 2)

Methods	mAP	ar	ra	bu	tr	sw	re	cc	ct1	ct2	in
Fast R-CNN	79.9	69.9	99.1	72.4	89.0	90.6	75.2	70.7	80.4	73.1	61.6
Faster R-CNN	89.9	80.7	100.0	81.6	90.6	90.8	90.7	78.4	90.9	90.9	78.9
R-FCN	89.1	88.5	99.8	87.8	89.6	90.0	87.7	77.6	91.3	93.7	85.2
SSD300	89.7	84.6	100.0	74.6	90.8	90.3	92.8	67.3	90.4	89.7	87.9
YOLOv2	81.7	71.0	100.0	79.0	90.7	90.4	81.5	63.5	90.6	90.6	81.1
Faster R-CNN+DANet	90.7	80.8	100.0	80.9	89.2	90.8	90.9	81.0	90.8	90.6	80.3
Faster R-CNN+only Channel Attention	88.5	80.5	100.0	81.4	90.0	89.7	88.6	80.2	90.9	90.6	79.2
Faster R-CNN+only Position Attention	91.0	80.5	100.0	90.1	89.9	90.7	90.9	80.7	90.7	90.6	80.8
Faster R-CNN+PCDANet	86.9	78.7	100.0	80.8	89.5	89.9	90.2	77.9	90.6	90.3	80.4
Faster R-CNN+CPDANet(ours)	91.4	87.8	100.0	89.5	90.2	90.6	89.5	79.2	90.7	90.9	81.0
Methods	mAP	pt	dg	dc	fu	co	cl	k1	k2	k3	k4	ta
Fast R-CNN	79.9	100.0	44.9	42.7	80.3	90.7	81.2	90.0	90.9	80.9	81.4	72.7
Faster R-CNN	89.9	100.0	90.9	81.1	88.5	90.9	90.6	90.6	90.9	100.0	89.8	100.0
R-FCN	89.1	100.0	48.1	88.9	89.0	90.7	90.0	90.9	92.3	100.0	90.6	100.0
SSD300	89.7	100.0	98.2	82.9	87.3	90.4	85.3	90.6	90.9	99.8	90.6	100.0
YOLOv2	81.7	99.8	67.3	70.5	79.4	90.3	86.7	89.4	95.3	90.9	89.7	18.2
Faster R-CNN+DANet	90.7	100.0	100.0	90.2	88.1	89.9	90.6	90.4	90.2	100.0	90.0	100.0
Faster R-CNN+only Channel Attention	88.5	100.0	75.5	90.1	79.2	90.6	81.5	90.5	90.9	100.0	89.2	100.0
Faster R-CNN+only Position Attention	91.0	100.0	97.2	89.7	88.2	90.8	90.7	90.2	90.5	100.0	89.3	100.0
Faster R-CNN+PCDANet	86.9	100.0	53.2	80.8	80.2	81.7	90.8	90.5	90.9	100.0	89.6	100.0
Faster R-CNN+CPDANet(ours)	91.4	100.0	100.0	89.6	86.9	90.9	90.3	90.2	90.3	100.0	90.7	100.0

Table 2. Comparison of abbreviations and full names of all classes.

Abbreviations	ar	ra	bu	tr	sw	re	cc	ct1
Full names	arrester	radiator	bushing	transformer	switch	reactor	cable-connector	current-transformer-1
Abbreviations	ct2	in	pt	dg	dc			
Full names	current-transformer-2	insulator	potential-transformer	discharge-gap	discharge-coil			
Abbreviations	fu	co	cl	k1	k2	k3	k4	ta
Full names	fuse	conservator	current-limiter	knife-1	knife-2	knife-3	knife-4	trap

5.3 Ablation study

We employ our method on the top of the dilation network to capture long-range dependencies for getting better detection performance. In order to verify the effectiveness of our proposed method, we conducted the experiments as follow.

Firstly, we employed CPDANet after feature extraction to further improve feature representation. The experimental results are shown in Table 1. The data show that CPDANet can indeed further improve the detection performance compared to Faster R-CNN. Secondly, we used the DANet to further improve feature representation, the experimental results show that although mAP has also been improved to a certain extent, there is still a certain gap between the 0.8% improvement in the experiment used DANet and the 1.5% improvement in the experiment used CPDANet. The experiment proves that our improvements on DANet is positive. Then, we used only channel attention module, and only position attention module to further improve feature representation respectively. Only position attention module improved mAP by 1.1%, which is not as good as the experiment used CPDANet, even only channel attention module reduced mAP by 0.4%. Finally, we used PCDANet to get feature representation and reduced mAP by 3%. The experimental results prove that the dual attention mechanism is necessary. In summary, our method has got the best performance in all the comparative experiments.

5.4 Comparison with State-of-the-art

We further compare our method with selected advanced baseline approaches on the TEED001 dataset. We evaluated 5 detectors on our dataset, including Fast R-CNN [4], YOLOv2 [18], SSD [16], R-FCN [1], Faster R-CNN [19]. And all detectors achieved the best results as shown in Table 1. Compared with the above mentioned methods, our method has great advantage in mAP and improve mAP by 11.5%, 9.7%, 1.7%, 2.3% and 1.5% respectively. Apart from the advantages of the overall framework, the biggest advantage is feature representation in our method integrate local features with their global dependencies.

6 Conclusions and Future works

In this paper, we have created a new benchmark for thermal image based electrical-equipment detection. And this is a leading work to propose this type of benchmark. We tested multiple baseline approaches with good experimental results. In addition, we proposed CPDANet based on the Dual Attention Network and applied it to further improve feature representation. The further improved feature map has gotten better detection performance.

In the future, we plan to add more images to those classes that have fewer images and study pixel-level segmentation of electrical-equipment based on the results of object detection, both of which can improve the accuracy of electrical-equipment defect detection.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China (Grant Nos. 61602006, 61872005), Anhui Provincial Natural Science Foundation (No. 1908085MF206).

References

1. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems. pp. 379–387 (2016)
2. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
3. Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H.: Dual attention network for scene segmentation. arXiv preprint arXiv:1809.02983 (2018)
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European conference on computer vision. pp. 346–361. Springer (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. computer vision and pattern recognition pp. 770–778 (2016)
8. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3588–3597 (2018)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 **7** (2017)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
11. Li, C., Wang, X., Zhang, L., Tang, J., Wu, H., Lin, L.: Weighted low-rank decomposition for robust grayscale-thermal foreground detection. IEEE Transactions on Circuits and Systems for Video Technology **27**(4), 725–738 (2016)
12. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2016)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
14. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)
15. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)
16. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
18. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. arXiv preprint (2017)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence (6), 1137–1149 (2017)

20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
21. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
22. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: Disan: Directional self-attention network for rnn/cnn-free language understanding. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Song, G., Liu, Y., Jiang, M., Wang, Y., Yan, J., Leng, B.: Beyond trade-off: Accelerate fcn-based face detector with higher accuracy. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7756–7764 (2018)
25. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2), 154–171 (2013)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
27. Vu, T.H., Osokin, A., Laptev, I.: Context-aware cnns for person head detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2893–2901 (2015)
28. Wang, X., Li, C., Luo, B., Tang, J.: Sint++: robust visual tracking via adversarial positive instance generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4864–4873 (2018)
29. Wang, X., Li, C., Yang, R., Zhang, T., Tang, J., Luo, B.: Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. arXiv preprint arXiv:1811.10014 (2018)
30. Wang, X., Sun, T., Yang, R., Li, C., Luo, B., Tang, J.: Quality-aware dual-modal saliency detection via deep reinforcement learning. Signal Processing: Image Communication **75**, 158–167 (2019)
31. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
32. Xiao Wang, Tao Sun, C.L.Y.L.R.Y.J.T.B.L.: Learning target-aware attention for robust tracking with conditional adversarial network (2019)
33. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
34. Zhou, Y., Liu, L., Shao, L., Mellor, M.: Dave: a unified framework for fast vehicle detection and annotation. In: European Conference on Computer Vision. pp. 278–293. Springer (2016)
35. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2110–2118 (2016)
36. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European conference on computer vision. pp. 391–405. Springer (2014)