Front. Optoelectron. 2019, 12(3): 324–331
https://doi.org/10.1007/s12200-019-0853-1

RESEARCH ARTICLE

High accuracy object detection via bounding box regression network

Lipeng SUN[1], Shihua ZHAO[1], Gang LI[2], Binbing LIU (✉)[3]

1 State Grid Hunan Electric Power Corporation Limited Research Institute, Changsha 410007, China
2 State Grid Hunan Electric Power Corporation Limited, Changsha 410007, China
3 School of Optical and Electronics Information, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract As one of the primary computer vision problems, object detection aims to find and locate semantic objects in digital images. Different with object classification, which only recognizes an object to a certain class, object detection also needs to extract accurate locations of objects. In the state-of-the-art object detection algorithms, bounding box regression plays a critical role in order to achieve high localization accuracy. Almost all the popular deep learning based object detection algorithms have utilized bounding box regression for fine tuning of object locations. However, while bounding box regression is widely used, there is few study focused on the underlying rationale, performance dependencies, and performance evaluation. In this paper, we proposed a dedicated deep neural network for bounding box regression, and presented several methods to improve its performance. Some ad hoc experiments are conducted to prove the effectiveness of the network. Also, we apply the network as an auxiliary module to the faster R-CNN algorithm and test them on some real-world images. Experiment results show certain performance improvements on detection accuracy in term of mean IOU.

Keywords deep learning, object detection, bounding box regression, IOU distribution

1 Introduction

Object detection is one of the primary problems in computer vision. It aims to find and locate objects of interest from digital images or videos. Unlike image classification, object detection is more challenging because it deals with both "what it is" and "where it is" for objects in images. An example of object detection is shown in Fig. 1. Object detection not only needs to identify the class of the object ('car' in this case) in the image, but also has to locate the object in the image accurately. Usually the object location is represented by the rectangular box surrounding the object (that is, the bounding box), as annotated with the green frame in the image.

For a long time, object detection algorithms have followed a paradigm based on sliding window and pattern matching. To find the objects of interests, a scale variable, pixel-by-pixel sliding window is used to extract a set of candidate regions from the image, and the features of each candidate region are compared with the ground-truth features. Through the pattern matching the class of object and the location of object are determined at the same time. The features of candidate regions are usually extracted by handcrafted operators, such as Hessian points [1], HOG [2] and SIFT [3], and the pattern matching is usually achieved by means of Euclidean distance based classifier or more complex classifier model. Although these classical methods have been studied and used for a long time, they have great deficiencies. First, based on the spatial resolution of the image and the optional scale of the observation window, the amount of candidate regions is tremendous, which leads to huge computational complexity of the object detection algorithm. Secondly, the traditional classifier has poor performance because of the great variation of object appearance, geometrical deformation. This results in a poor performance of object detection.

In view of the great success of deep learning in the field

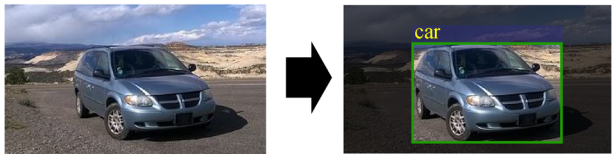


Fig. 1 Object detection: find and locate objects in images

Received August 28, 2018; accepted September 25, 2018

E-mail: liubinxp@163.com

of image classification [4,5], researchers try to tackle the object detection problems by deep learning [6,7]. In the past years, a series of object detection algorithms based on deep learning, such as R-CNN [8], faster R-CNN [9], SSD [10], YOLO [11], FPN [12] and mask R-CNN [13], are proposed. They keep refreshing new records of object detection performance [14]. After image classification, deep learning has achieved great success in the field of object detection.

Although the current popular object detection algorithms have quite different design paradigms, they indeed have some common points, that is, almost all object detection algorithms contain a candidate region proposal module and a boundary box regression (BBR) module. The candidate region proposal module is used to find some candidate regions which may (or may not) contain objects, while the BBR module is used to fine-tune the bounding boxes of the candidate regions to obtain more accurate locations. For example, the R-CNN algorithm uses selective search [15] to extract about 2000 candidate regions for each image; faster R-CNN uses region proposal network (RPN) to find candidate regions; YOLO and SSD use grids and anchors to generate a set of candidate regions, respectively. Candidate region proposal module is introduced to avoid the huge search space of sliding windows, which may cause huge computation cost. It should be noted that no matter which kind of candidate region proposal technique is used, all of these object detection algorithms include a BBR module in the final stage of the detection process to refine the position of object. BBR was first proposed and used by Felzenszwalb et al. [16], and then used in almost all object detection algorithms [17]. BBR module can work alone, or be integrated with other modules.

BBR has been proved to be effective in improving the mean average precision (mAP) [18] and localization accuracy of object detection algorithms. In some cases, especially in industrial occasions, such as autonomous driving and robotics [19], the accuracy of object position obtained by object detection algorithms is more important than general applications. Since the BBR module can significantly improve the localization accuracy of the bounding box, it is considered to be indispensable for the object detection algorithms.

Although BBR is essential for object detection algorithms, there is few study focused on the underlying rationale, performance dependencies, and performance evaluation of BBR. In this paper, we attempted to explore the underlying rationale of BBR, and proposed a dedicated deep neural network for BBR. Also, we presented several methods to improve the performance of BBR by analyzing the performance dependent factors of BBR. In Section 2, we discuss the theory of BBR and the factors that affect the performance of BBR, and propose the bounding box regression network. In Section 3, we verify the effectiveness of the BBR network and the proposed performance improvement methods by experiments. Finally, we discuss and conclude in Section 4.

2 Theory and method

The BBR modules proposed in existed literatures use grids or RPN to generate candidate regions, this implies that the samples used for model training is rare, and this leads to lower regression accuracy and worse generalization capability. If a dedicated bounding box regression network is built and trained, it would be easier to generate much more samples for training, and moreover, it is possible to afford us some deep insight into BBR.

In this paper, we try to improve the localization accuracy of object detection algorithm by training a dedicated bounding box regression network. Although the so called one-stage method is generally favored because of its integrity and simplicity, in some cases, in order to further improve the localization accuracy, building a dedicated, ancillary fine-tuning module is necessary.

2.1 Principle of bounding box regression

Boundary box regression tries to find a map from a candidate bounding box of object (called proposal bounding box) P, which is considered to be not inaccurate, to a new bounding box (called predicted bounding box) $\hat{G}$, which is subject to a higher consistency to the real bounding box (called ground-truth bounding box) G. Formally, for proposal bounding box $P = (P_x, P_y, P_w, P_h)$, bounding box regression refers to finding a mapping function f, that maps the features $\mathcal{P}$ of image region defined by P into a new bounding box $\hat{G} = (\hat{G}_x, \hat{G}_y, \hat{G}_w, \hat{G}_h)$, so that it is as close as possible to the ground-truth bounding box $G = (G_x, G_y, G_w, G_h)$, that is

$$f : \mathcal{P}^{(P)} \rightarrow \hat{G}.$$

It is important to note that the input of the regression model is $\mathcal{P}$, the features of the image region defined by the proposal bounding box P, instead of P itself. A schematic diagram of bounding box regression is shown in Fig. 2. The green frame represents the ground-truth bounding box of object annotated by hand, the blue frame represents the

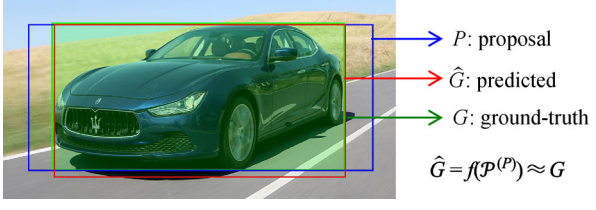


Fig. 2 Schematic diagram of bounding box regression

proposal bounding box, and the red frame represents the predicted bounding box inferenced by the BBR module.

Since $\hat{G}$ contains four variables, the regression model actually contains a set of 4 specific regression functions. That is

$$\begin{cases} G_x = f_x(\mathcal{P}), \\ G_y = f_y(\mathcal{P}), \\ G_w = f_w(\mathcal{P}), \\ G_h = f_h(\mathcal{P}). \end{cases} \tag{1}$$

In practice, the absolute coordinates are significantly dependent on the size of target image and they are difficult to train and converge. Therefore, the relative coordinates (to be exact, the relative positions of the proposal bounding boxes and the ground-truth bounding boxes) are actually used. The relative coordinates are usually represented by the bounding box transformation coefficients (BTC), a set of parameters that define how to transform the proposal bounding boxes to the ground-truth bounding boxes. To get larger value range of bounding box scales, it is useful to apply a logarithmic transformation for the scale parameters. It is also beneficial to the convergence of model training. Consequently, the boundary box regression formula is as follows:

$$\begin{cases} t_x = \dfrac{G_x - P_x}{P_w} = f_x(\mathcal{P}), \\ t_y = \dfrac{G_y - P_y}{P_h} = f_y(\mathcal{P}), \\ t_w = \log\dfrac{G_w}{P_w} = f_w(\mathcal{P}), \\ t_h = \log\dfrac{G_h}{G_w} = f_h(\mathcal{P}). \end{cases} \tag{2}$$

2.2 Flow of the proposed method

Based on the principles described above, the flow chart of the proposed method is shown in Fig. 3. First, we extract the ground-truth bounding boxes G from the training samples, and then generate a set of proposal bounding boxes P by random sampling. For each proposal bounding box P, we extract the features $\mathcal{P}$ of the image region defined by the proposal bounding box P by a pre-trained deep neural network, and calculate the BTC at the same time. Finally, the features $\mathcal{P}$ and the BTC are grouped as training samples, and fed into the bounding box regression network for training.

The calculation method of BTC is shown in Eq. (2). The generation process of proposal bounding box is discussed in Section 2.3. Image feature extraction is achieved through a deep neural network based on transfer learning (see Section 3).

2.3 IOU distribution adjustment

To study the effectiveness of the BBR network and the dependencies to model parameters, we generate proposal bounding boxes by random sampling. Note that deep neural networks have been proved to be more capable on memory and interpolation rather than reasoning and extrapolation, so it is helpful to improve the regression accuracy if the training samples have a more uniformed distribution. We are concerned about the distribution of IOU mainly from two inspirations. 1) It has been generally accepted that deep learning models, especially those models for regression purpose, are good at interpolation and memory rather than in extrapolation and inference. This means that a wide and flat probability distribution of the training samples is desired since it may cover as many new samples as possible in the Euclidean feature space. 2) By observing and analyzing the IOU distribution of faster R-CNN output, we draw the conclusion that the Laplace distribution of IOU is more advantageous to improve the regression accuracy.

Here, the proposal bounding box P is generated by jitter sampling near the ground-truth bounding box G.

The performance of object detection algorithm is usually measured by mAP [18]. This metric, which potentially contains another two metrics: recall and precision, embodies the comprehensive performance of object detection algorithm, such as whether to find as many objects as possible (recall) or find only the correct objects (precision). Unlike image classification, the bounding box of object predicted by the object detection algorithm may not completely overlap with the ground-truth bounding box. To determine whether a predicted bounding box identifies an object correctly, the Intersection over Union (IoU) [18] is usually computed. According to the calculation method of mAP, the predicted bounding

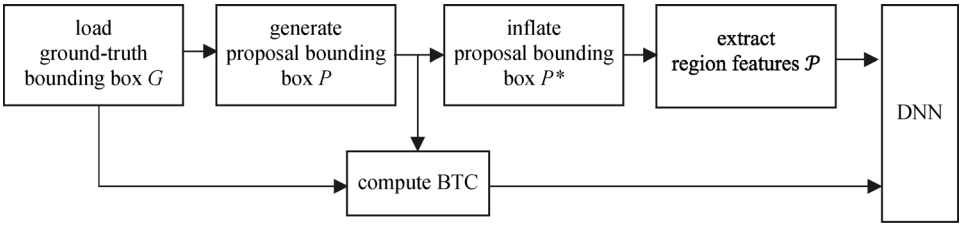


Fig. 3 Flow chart of the proposed method

boxes whose IOUs with the ground-truth bounding boxes exceed a certain threshold (e.g., 0.5) are considered to be the correct detection cases.

As mentioned above, deep neural networks are good at memory and interpolation rather than reasoning and extrapolation, so it is possible to improve the regression accuracy by training sample with more uniformed probability distribution. According to the generation algorithm of proposal bounding box described in the previous section, the translation and scaling are actually sampled with uniform distribution. However, the actual regression variables are BTC rather than translation and scaling. So we must take account into a proper distribution of BTC. In addition, the deep neural network module can actually perceive the overlap between the predicted bounding box and the ground-truth bounding box, so the probability distribution of IOU also should be considered elaborately.

The BTC and IOU distributions would be not uniformed if we generate bounding boxes by a uniformed distribution with regard to the translation (i.e., t_x and t_y in Fig. 4) and the scaling (i.e., t_w and t_h in Fig. 4), as described above. The results are shown in Fig. 4(a).

The two histograms of IOU are plotted with the same source data, but different parameter 'bins'. A histogram is a chart that plots the sample points which fall within various intervals. Because these intervals collect data, they are called bins. By a larger bin, histograms give a more rough and holistic sense of the density of the underlying distribution of the data. We plotted the distributions of IOU by two different bins so that we get better intuition from the charts.

It can be found that the IOU distribution in the range of [0.6,1.0] is not uniformed and appear as a Laplacian distribution with a peak value of 0.6. To correct this distribution, we use the Laplace sampling instead of uniform sampling for translation and scaling, and the final distribution is shown in Fig. 4(b). It can be seen that the new IOU distribution is an asymmetric bilateral Laplacian distribution with a peak value of about 0.9, which is more suitable for training the BBR network.

3 Experiments

As described in Section 2.1, the input of the BBR network is the features of the image region defined by object bounding box. The features can be extracted from image by handcrafted operators such as HOG [2], SIFT [3]. In deep learning-based object detection algorithms, they are usually generated through a deep convolution neural network. VGG [20] is a widely-used network architecture because of its simplicity and high-performance. In this paper, we use VGG16 as the backbone network for our BBR network.

The VGG16 network was originally designed for image classification. When it is used for regression, all the convolution layers are kept unchanged, while only replacing the output nodes of the fully connected layer from the image categories with the BTCs. That is to say, the input of the network is the original image data (a third order tensor), and the output is BTC (a first order tensor, including four bounding box transformation coefficients). Accordingly, we replace the cross-entropy loss function with the mean square error (MSE) loss function which is more suitable for regression problems.

Because the annotated images used for object detection is very scarce, it is impossible to train the VGG16 network from scratch. In this case, transfer learning [21,22] is an effective solution. In transfer learning, deep neural network is pre-trained from a large general annotated data set, and then fine-tuned on domain-specific data set. In this paper, we train the VGG16 network on ImageNet [23] classification data set, then retrain it on PASCAL VOC2007 classification data set, and finally fine-tune the network on PASCAL VOC2007 object detection data set.

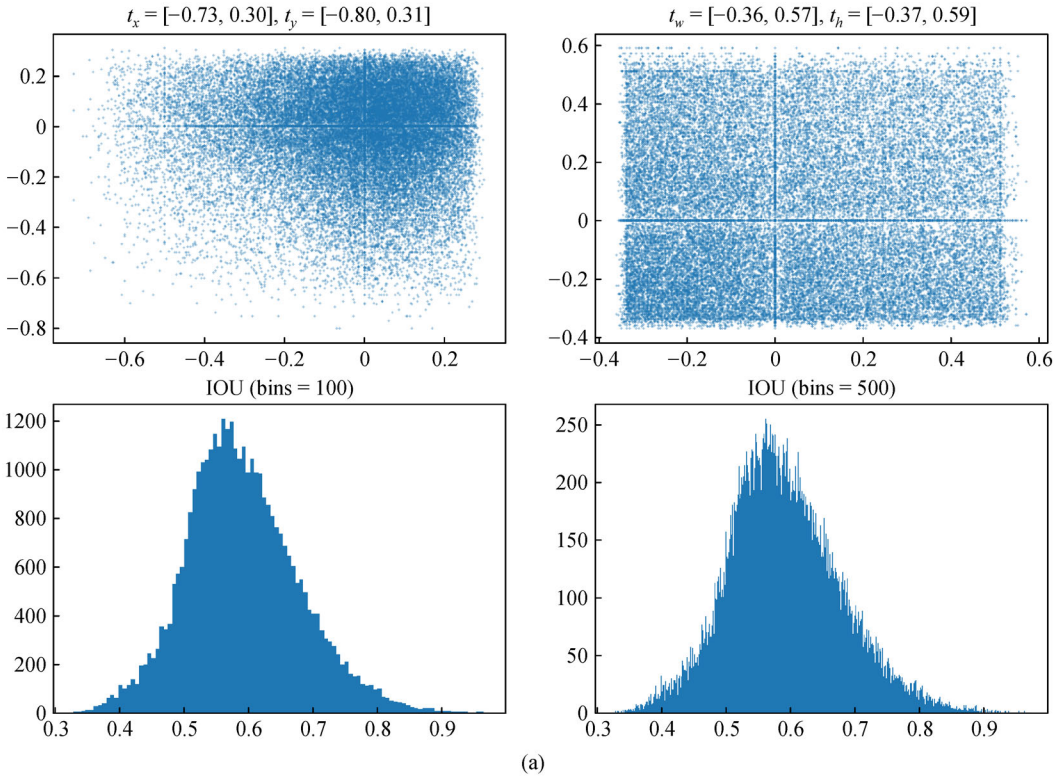
3.1 Evaluation for BBR network

By random sampling, we generate 100 proposal bounding boxes for each annotated object in VOC2007 object detection data set. For each proposal bounding box, we extract the features of the image region defined by the bounding box, and calculate the BTC. The features along with the BTC are grouped as training samples, and fed into the BBR network for training. The standard regression loss function MSE is used to optimize the BBR network by Gradient Descent. To visualize the regression accuracy of the BBR network, root mean square error (RMSE) is also recorder because it has the same data scale as BTC.

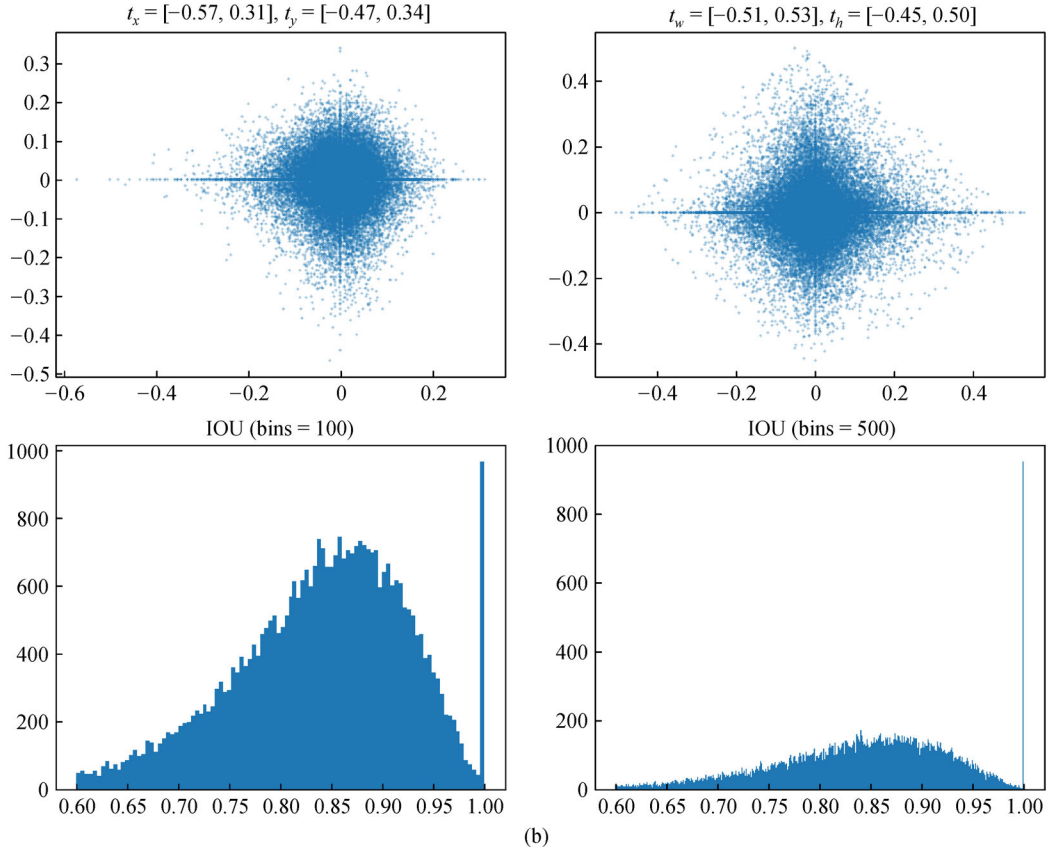
The convergence curve of RMSE is shown in Fig. 5. The optimization converged after four hours of training on NVIDIA GTX1080.

The performance of the BBR network was validated on VOC2007 test sub data set. We calculated the average IOU before regression, i.e., proposal IOU, and the average IOU after regression, i.e., predicted IOU. To verify the improvement of detection accuracy, three kinds of algorithm configurations are tested: 1) baseline: the basic algorithm configuration; 2) data_augmentation: algorithm configuration with augmentation; 3) IOU_uniform: algorithm configuration using Laplace sampling.

We used three kinds of method for data augmentation. The first is the location-sensitive method, including random cropping, random scaling, mirroring. The second is the location-independent method, including brightness adjustment, contrast adjustment and color adjustment. The third is the synthesis method. By using the segmentation annotations (mask information) included in data set, we extract the objects and combine them with some randomly selected image background to obtain the synthesized sample images.



(a)



(b)

Fig. 4 Comparison of adjusted distribution with original distribution of IOU. (a) IOU distribution on uniformly distributed BTC; (b) IOU distribution on Laplacian distributed BTC

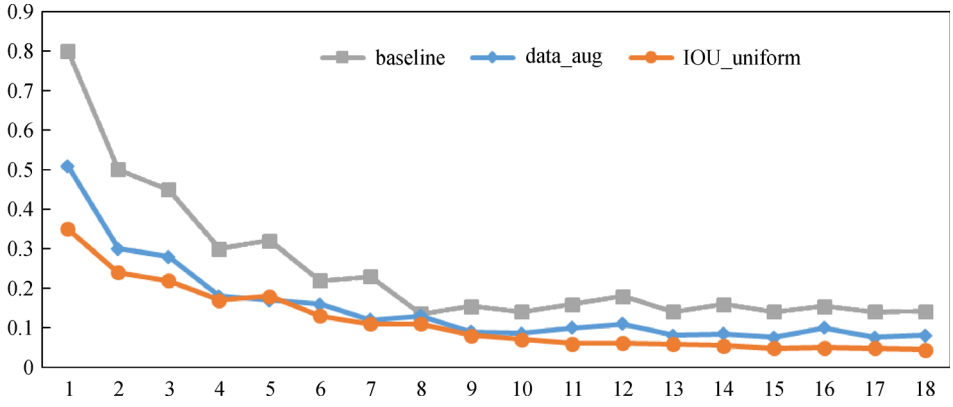


Fig. 5 Convergence curve of RMSE

Table 1 Improvement of IOU in three algorithm configurations

	baseline	data_aug	IOU_uniform
proposal IOU	0.7982	0.7982	0.7982
predicted IOU	0.8312	0.9076	0.9354

The test results are shown in Table 1. It can be seen that data enhancement and IOU distribution adjustment can effectively improve the localization accuracy of the BBR network.

3.2 Integration with faster R-CNN

To further test the effectiveness of BBR network in real-world applications, we integrate the proposed BBR network with faster R-CNN object detection algorithm. The test pipeline is as follows: first, faster R-CNN object detection algorithm is used to detect the objects in the images, then the obtained object bounding boxes are used as the proposal bounding boxes and are fine-tuned by the BBR network. The experimental results show that the IOU of the detection results can be improved from 0.8674 to 0.8892. Some test images are shown in Fig. 6.

The green frame is the ground-truth bounding box, the blue frame is the bounding box obtained by the faster R-CNN algorithm, and the red frame is the fine-tuned bounding box by the BBR network proposed in this paper. It can be found that the BBR network effectively improves the localization accuracy of object detection.

When building the proposed algorithm, we hope to obtain an end-to-end trainable object detection model by directly optimizing the faster R-CNN. However, since faster R-CNN is a well-designed model with elaborate hyperparameter tuning, it is hard to modify and optimize the algorithm framework directly. Actually, an additional, independent regression fine-tuning model may be quite useful in scenarios where precision is required, but speed performance is not critical. A practical example is that one can use classical computer vision algorithms (such as MSER) to detect screw holes in an industrial image, and then use BRN to fine-tune its position.

4 Conclusion

As an indispensable module, bounding box regression is used by many object detection frameworks to fine-tune the location of the predicted bounding box. To validate the effectiveness of bounding box regression, analyze its hyper-parameter sensitivity and explore performance improvement methods, we proposed a dedicated bounding box regression model based on deep neural network and

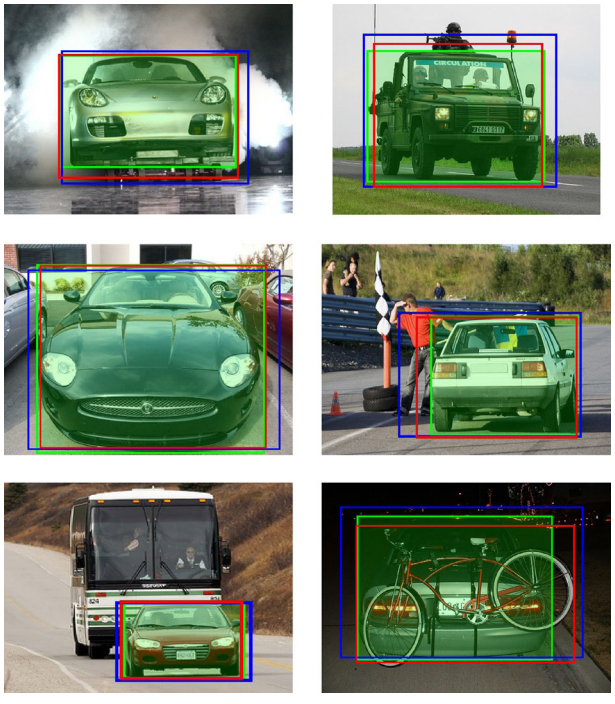


Fig. 6 Detection results of BBR network on some real-world images

presented several methods to improve its performance. Experimental data proved that the bounding box regression network can effectively improve the localization accuracy of object detection. The future works of our research include further improving the regression accuracy by incorporating coordinate-sensitive network structure to achieve a robust and general BBR network.

References

1. Mikolajczyk K, Schmid C. An affine invariant interest point detector. Vancouver, Canada. In: Proceedings of European Conference on Computer Vision. Beilin: Springer, 2002, 128–142
2. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005, 886–893
3. Lowe D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110
4. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M S, Berg A C, Li F F. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252
5. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2015, 770–778
6. Han J, Zhang D, Cheng G, Liu N, Xu D. Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Processing Magazine, 2018, 35(1): 84–100
7. Jiang H, Cheng M M, Li S J, Borji A, Wang J. Joint salient object detection and existence prediction. Frontiers of Computer Science, 2018, https://doi.org/10.1007/s11704-017-6613-8
8. Girshick R B, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014, 580–587
9. 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 and Machine Intelligence, 2017, 39(6): 1137–1149
10. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S E, Fu C, Berg A C. SSD: single shot multibox detector. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016, 21–37
11. Redmon J, Divvala S K, Girshick R B, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016, 779–788
12. Lin T Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017, 936–944
13. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, doi:10.1109/TPAMI.2018.2844175
14. Everingham M, Eslami S M A, Gool L V, Williams C K I, Winn J, Zisserman A. The pascal visual object classes challenge: A Retrospective. International Journal of Computer Vision, 2015, 111(1): 98–136
15. Uijlings J R R, van de Sande K E A, Gevers T, Smeulders A W M. Selective search for object recognition. International Journal of Computer Vision, 2013, 104(2): 154–171
16. Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645
17. Park H M, Cho D Y, Yoon K J. Greedy refinement of object proposals via boundary-aligned minimum bounding box search. IET Computer Vision, 2018, 12(3): 357–363
18. Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338
19. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012, 3354–3361
20. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv:1409.1556
21. Chen Z, Zhang T, Ouyang C. End-to-end airplane detection using transfer learning in remote sensing images. Remote Sensing, 2018, 10(1): 139
22. Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359
23. Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009, 248–255

Lipeng Sun completed the B.S. degree in 2005 from Chongqing University, Chong qing, China. He also received his master degree in 2008 from Chongqing University, Chongqing, China. His current research interest is high voltage and insulation technology.

Shihua Zhao graduated from Chongqing University and obtained the Ph.D. degree in 2013. He is now working in State Grid Hunan Electric Power Corporation Limited Research Institute. His main research interests include high voltage technology, power transformer fault detection and diagnosis. He is the author or the co-author of several technical papers.

Gang Li completed the B.S. degree in 1996 from Xi'an Jiaotong University, Xi'an, China. He has been engaged in operation and maintenance technology of transformer over 20 years. His current research interest is high voltage and insulation technology.

Binbing Liu received the bachelor degree in Optoelectronics from Huazhong University of Science and Technology, Wuhan, in 2000, and got the master degree and Ph.D. degree in Physical Electronics from the same university, in 2003 and 2013, respectively. Now, he is working on School of Optical and Electronics Information as a lecturer. His research interests include computer vision and machine learning.