Optimization of the Bounding Box Regression Process of SSD Model

Yuanzhou Yao[1, a], Yuhang Yang[1], Xinyue Su[3], Yihang Zhao[1],
Ao Feng[1], Yiting Huang[1] and Haibo Pu[1,2, b, *]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an, Sichuan 625000, China

[2] Sichuan Key Laboratory of Agricultural Information Engineering, Ya'an, Sichuan 625000, China

[3]College of Humanities, Sichuan Agricultural University, Ya'an, Sichuan 625000, China

[a] yaoyuanzhou@stu.sicau.edu.cn, [b, *] puhb@sicau.edu.cn

Abstract. Intersection over Union (IoU) has always been the most popular evaluation metric used in object detection benchmarks. However, IoU has a disadvantage that it is not feasible to optimize without overlapping bounding boxes. Therefore, proposed a generalized version as a new loss and a new indicator to address the weakness of IoU. Based on this, this paper innovatively incorporated this Generalized IoU (GIoU) as a loss function into the most advanced SSD object detection network model, and carried out experiments on the original model and the improved model respectively based on the standard detection data set PASCAL VOC. The experimental results proved that the improved model had higher accuracy and better effect.

Keywords: Intersection over Union (IoU); Generalized IoU (GIoU); SSD; VGG16.

1. Introduction

With the rapid development of the society and technology, the application of neural networks and deep learning to daily life has become a trend, for example, with the emergence of autonomous cars, intelligent surveillance cameras, facial recognition, and some other useful applications, the market of rapid and accurate object detection system is also booming. Many typical neural network structural models, like FASTER R-CNN, SSD, etc., have emerged in this process. No matter in which model, IoU is applied to the prediction of the bounding box. Intersection over Union (IoU) is the most popular evaluation metric used in object detection benchmarks. However, there is a gap between the parameters of the regression bounding box and maximizing the metric when optimizing the common distance loss. The best goal of the metric is the metric itself. In the case of an axis-aligned 2D bounding box, it shows that IoU can be used directly as a regression loss. However, IoU has a corresponding disadvantage such that optimization under non-overlapping bounding boxes is not feasible. So, when the borders are not overlapped, it is possible that the removal will not be optimized, and this will lose the effectiveness of learning.

In order to solve this problem, the evaluation metric of GioU was proposed not long ago. An analytical solution for calculating the GioU between two axis-aligned rectangles, allowing it to be used as a loss in this case. Incorporating GioU losses into the most advanced object detection algorithm, this method perfectly solves the problem that IoU can't optimize the non-overlapping bounding box and the removal. Compared to IoU, GIoU not only focuses on overlapping areas. When the prediction bounding box and the real bounding box are not well aligned with each other, the blank space between them increases in a closed figure. Therefore, the value of GioU not only better reflects how the two symmetric objects overlap, but also fits the overlap between the real and the prediction bounding box. This article will describe in detail to prove that the introduction of GioU into SSD will optimize the neural network structure through the experimental results and the improvement of accuracy.

2. Second Ssd Network Model Optimized by Giou

2.1 The Introduction to the Object Detection Field

Recently, the deep learning model has gradually replaced the traditional machine vision method as the mainstream algorithm in the field of object detection. The process of understanding a picture can be roughly divided into three parts: classification, detection and segmentation. There are also many deep learning models for detection, such as R-CNN series, Yolo, SPP-Net, SSD. By contrast, SSD training is better, and the network based on SSD structure model has higher recognition accuracy and faster recognition speed.

2.2 SSD Network Structure

The SSD algorithm is similar to Yolo in some places, for they both base on CNN single-stage object detection algorithm. For those object instances in the box, a fixed-size bounding box set and scores are produced by CNN, and non-maximum suppression is performed to produce the final detection box. The network structure of the SSD is in the shape of a pyramid as a whole. The first 5 layers of the SSD utilize the first 5 layers in the VGG16 network, and from the 6th layer, the SSD converts the fc6 and fc7 in the original VGG16 into 3×3 Convolutional layer conv6 and 1×1 Convolutional layer conv7. And at the same time, pool5 in VGG16 is changed from the original 2×2-S2 to 3×3-S1. In order to cope with this change, conv6 adopts Dilation Conv, which is to make the Receptive field grow exponentially without increasing the complexity of parameters and model. The dropout layer and the fc8 layer in VGG16 are removed from the SSD, and replaced by some convolutional layers (Conv8, Conv9, Conv10, Conv11). The size of these layers gradually decreases, thereby achieving the result of multi-scale prediction.

A key point in SSD is the introduction of the Prior Box, a pre-selected box for some objects which is similar to Anchor. With these pre-selected boxes, the speed of detection can be greatly improved. And then use Softmax function to classify. Finally, we can predict the position of ultimate real object through bounding box regression.

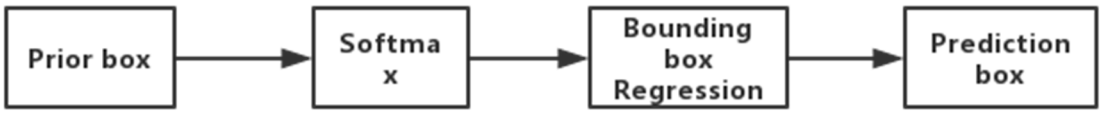

Figure 1. The process of predicting the real position

The rules for producing a prior box in an SSD are as follows:

(1) Center on the center point of each point on the feature map (offset=0.5), and generate a series of concentric prior box.

(2) The minimum side length of the square prior box is Min_size, and the maximum side length is $\sqrt{Min_{size} * Max\_size}$

(3) Each time an aspect ratio is set in prototxt; 2 rectangles are generated, the length is; *Min_size, and the width is: 1/*Min_size

The Min_size and Max_size of the priority box corresponding to the feature map are calculated by the following formula, where m is the quantity of feature map:

$$S_k = S_{Min} + \frac{S_{Max} - S_{Min}}{m-1}(k-1), k \in [1, m] \tag{1}$$

The width of the Prior box is calculated according to the proportion value . The formula for calculating the length and width of the default box is as follows:

$$w_k^a = S_k\sqrt{a_r}, h_k^a = s_k / \sqrt{a_r} \tag{2}$$

Where w is the width; h is the height; ar is the aspect ratio; and Sk is the size of default box

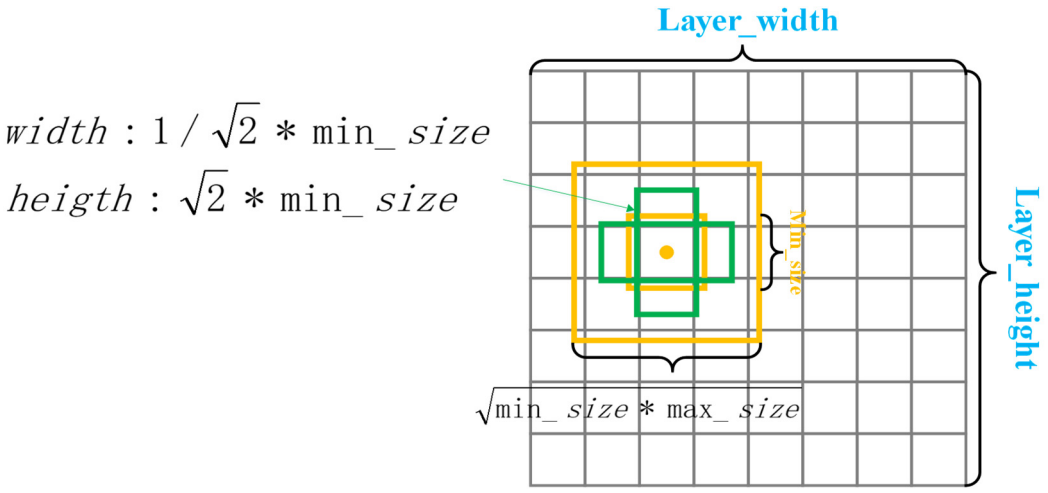

Figure 2. Aspect ratio of the Prior box

2.3 GIoU Optimization of the Model

Intersection over Union is a criterion for detecting a particular data set, which are used to compare the similarities of any two patterns. The IoU encodes the shape properties of the comparison object. (eg, the width, height, and position of the two bounding boxes) into the Region properties, and then calculates a normalized metric that focuses on its area (or volume). It is often used for the evaluation of object detection performance index because of this.

Generally, in SSD, there are more negative default boxes than positive default boxes. If the sample training is randomly selected, it will bias towards the negative samples (because the probability of extracting the negative samples is larger), which will make loss unstable. So, we have to add positive samples. What is positive sample? How to increase? The commonly used method is Hard Negative Mining, which compares the default box according to the confidence score, selects the box with high confidence for training, and controls the ratio of positive samples to negative samples at 1:3. This is an improvement. And without controlling it, it's very likely that all the samples that Sample will have are negative samples (that is, let the network find the correct object from these negative samples, which is obviously not possible), the reason why the ratio is 1:3 has already been explained clearly in other papers already.

First, we need to know what is IoU to understand its importance in this operation.

The denominator is the union of ground truth and default box. The numerator is the intersection of ground truth and default box. It reflects the overlapping ratio of these two to some extent. The ideal case is complete overlap when the ratio is 1. The formula is as follows:

$$IoU = \frac{area(C) \cap area(G)}{area(C) \cup area(G)} \tag{3}$$

This can be used as a criterion for the selection of positive samples. First, find the default box with the largest IoU for each ground truth, which will ensure that the ground truth has the default box to match at least. Then set a threshold for IoU, which will help us determine the ratio, 1:3. The default box that matches the ground truth is positive, and the default box without matching is negative.

IoU has various problems in the experiment. If the two objects do not overlap, the IoU value will be zero and will not reflect the distance between the two shapes. In the case of non-overlapping

objects, if IoU is used as a loss function (we won't discuss too much here), the gradient will be zero and cannot be optimized;

IoU cannot correctly distinguish the different alignments of two objects. More precisely, the two objects overlap in multiple different directions, and the intersection points are horizontally equal, and their IoU will be exactly equal. That is to say, it cannot reflect how the images overlap.

Improvements to the IoU problem: based on the basic mathematical principles of IoU, a new evaluation function is used to improve the selection of positive samples in the SSD. In the case that the ground truth does not overlap with the default box, it still can well reflect the degree of approximation degree of the prediction box and the real box.

IoU, the upgraded version of GIoU, is introduced as a new indicator to compare the approximation degree of any two figures. GIoU is designed in this way: if there are two arbitrary properties—A and B, find a minimum closed figure—C which is able to include both A and B, and then we calculate ratio of the area in C that does not overlap A to the total area of C, then the use the IoU ratio of A and B:

$$GIoU = IoU - \frac{|C \setminus (A \cup B)|}{|C|} \tag{4}$$

For IoU, when $0 \le IoU \le 1$, the closer the value of IoU to 1, the higher the approximation degree of the real box is.

For GioU, when $-1 \le GIoU \le 1$, the closer the value of GIoU to 1, the higher the approximation degree of the real box is.

When the two shapes completely coincidence, GIoU=IoU=1, and the operation process of GioU is as follows:

$$For\ the\ predicted\ box\ B^p\ ensuring\ x_2^p > x_1^p\ and\ y_2^p > y_1^p:$$

$$\bar{x}_1^p = \min(x_1^p, x_2^p), \bar{x}_2^p = \max(x_1^p, x_2^p),$$

$$\bar{y}_1^p = \min(y_1^p, y_2^p), \bar{y}_2^p = \max(y_1^p, y_2^p).$$

$$Calculating\ area\ of\ B^g: A^g = (x_2^g - x_1^g) \times (y_2^g - y_1^g)$$

$$Calculating\ area\ of\ B^P: A^p = (\bar{x}_2^p - \bar{x}_1^p) \times (\bar{y}_2^p - \bar{y}_1^p)$$

$$Calculating\ \text{intersection}\ I\ between\ B^p\ and\ B^g$$

$$x_1^L = \max(\bar{x}_1^p, \bar{x}_1^g), x_2^L = \min(\bar{x}_2^p, x_2^g),$$

$$y_1^L = \max(\bar{y}_1^p, \bar{y}_1^g), y_2^L = \min(\bar{y}_2^p, y_2^g),$$

$$I = \begin{cases} (x_2^L - x_1^L) \times (y_2^L - y_1^L), if\ (x_2^L > x_1^L, y_2^L > y_1^L) \\ 0, otherwise \end{cases}$$

$$Finding\ the\ coordinate\ of\ smallest\ enclosing\ box\ B^c:$$

$$x_1^c = \min(\bar{x}_1^p, x_1^g), x_2^c = \max(\bar{x}_2^p, x_2^g)$$

$$y_1^c = \min(\bar{y}_1^p, y_1^g), y_2^c = \max(\bar{y}_2^p, y_2^g)$$

$$Calculating\ area\ of\ B^c: A^c = (x_2^c - x_1^c) \times (y_2^c - y_1^c)$$

$$IoU = \frac{1}{U}, where\ U = A^p + A^g - I$$

$$GIoU = IoU - \frac{A^c - U}{A^c}$$

When the predicted value and Ground truth do not overlap, GIoU can also well evaluate the approximation degree of the prediction box and the real box.

The difference between IoU and GioU: As a distance, GioU is similar to IoU. Compared to IoU, GIoU not merely focuses on overlapping areas. When A and B are not well aligned with each other, the blank space between the two symmetrical shapes A and B in the closed shape C increases. Therefore, the value of GIoU can better reflect the overlap between two symmetrical objects. It can then be used as an alternative to IoU.

2.4 The Improved SSD Model based on GioU

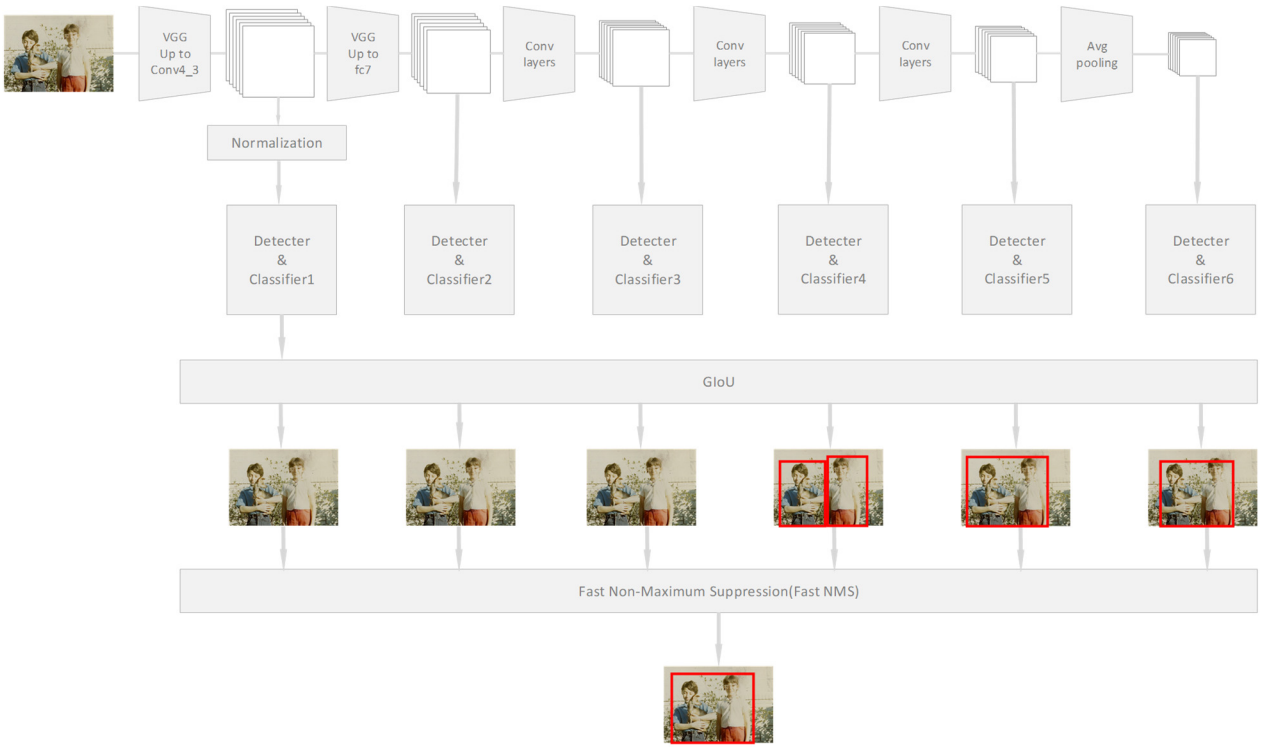

Figure 3. The improved model

3. Third the Improved Ssd Network Experiment based on Giou

3.1 Experimental Platform

8GB memory, NVIDIA GeForce RTX 2080 GPU, Inter(R) Core(TM) i7-9700K CPU are used as the hardware platform. Operating system is Linux Ubuntu 16.04. Deep Neural Network library version is CUDA8.0. GPU-accelerated library is CUDNN V6.0. The deep learning framework based on Python programming language is tensorflow-gpu==1.9.0. This platform trains and verifies the SSD object detection network model improved in this paper.

3.2 Experimental Data

This paper uses the standard test dataset PASCAL VOC as experimental data. The PASCAL VOC dataset provides a standard image annotation dataset and standard evaluation system for detection algorithms and learning performance. The PASCAL VOC database consists of 20 catalogues: humans; animals (birds, cats, cattle, dogs, horses, sheep); vehicles (aircrafts, bicycles, boats, buses, cars, motorcycles, trains); indoor appliance (bottles, chairs, dining tables, potted plants, sofas, televisions). The dataset has good image quality and complete labeling, which is suitable for testing algorithm performance. The training set contains 17125 images; the validation set contains 5011 images; and the test set contains 4952 images. The original image has different pixel sizes: the horizontal image size is about 500*375, and the vertical image size is about 375*500. The training process will resize

these images to 300*300 through the reshape operation, so that the original images will not be too far from the standard. The resizing images in this paper are the image data used for training and test verification, and then complete the experiment.

3.3 Format Pre-preprocessing

In order to read the data faster for processing and avoid blocking on the input layer during the network training, the TFRecord format conversion is performed on the experimental data set VOC. TFRecord is a file format provided by tensorflow for reading data at a high speed. In the process of reading TFRecord data, the input queue is created with the corresponding TF file as the parameter, and the input and output of the data is performed independently of the network computing, so that it can perform multi-threaded acceleration.

3.4 The Instruction of Model Parameter Settings

In order to save unnecessary training time and make the training effect more obvious, we use Transfer learning to train the deep learning network model. We first load the VGG16 pre-training model parameters, and the VGG16 pre-training model does not interfere with the regression training of object detection box and the different experimental results between IoU and GIoU. The remaining parameters are randomly initialized by a Gaussian distribution function whose mean is 0 and standard deviation is 0.01.

The adaptive gradient descent algorithm is used for optimization. The batch-size is set as 32; the weight attenuation coefficient is 0.005, the initial learning rate is 0.001; and the exponential decay parameter is 0.94. In this paper, the maximum number of iterations is set as 20,000 times. The training of IoU original model and the SDK model improved by GIoU are iterated by 20,000 times of training respectively. The network model is saved once every 5,000 times, and the hyperparameter value is printed once every 10 steps. After the training, save the parameters of the final model for testing, record the AP parameters of various classes, and compare the Map gap between the two.

4. Fourth Results and Analysis

4.1 Evaluation Indicators

In the course of the experiment, the evaluation index that often used are Precision and Recall. If a user submits a query to the system (for example, "What is a cat?"), the system returns a series of results, which are determined by what the user submitted. That is precision (which describes the accuracy of finding an object) and recall (the miss rate of finding an object). A new AP (Average Precision) value for each category is first calculated respectively, which is an important index for evaluating the detection effect. When the intersection ratio (IoU) of the object bounding predicted by the model to the bounding box in the labeled data corresponding to the test set is greater than or equal to the set threshold, the detection result is considered to be correct TP, otherwise it is regarded as the detection error FP. The formula is as follows:

$$\begin{aligned} p &= \frac{Tp}{Tp + Fp} \\ R &= \frac{Tp}{Tp + F_N} \\ AP1 &= \frac{R1' + R2' + R3' + ... + RN'}{N'} \\ MAP &= \frac{AP1 + AP2 + ... + APX}{X} \end{aligned} \tag{5}$$

Table 1. Judge the correctness of the detection object

	Detection as a object	Detection as non-object
object	TP (true)	TN (true error)
non-object	FP (false true)	FN (false error)

P ——The accuracy of the evaluation index of the detection accuracy, which means how many test results are given by the model are correct.

R ——The recall rate the evaluation index of the detection accuracy, which means how many the correct objects are detected actually.

AP ——The average accuracy mean

MAP —— The average accuracy mean of the composite indicators, the average of the sum of the APs in each category

The MAP indicator is used to evaluate the performance accuracy of the object detection model, which can avoid the extremes of certain categories that weakens the performance of other categories.

4.2 Experimental Results

After 20,000 times' training iterations, the final saved model parameters are uesed to test the SSD object detection model before and after the improvement by GioU, and the AP values of the 21 objects on the VOC data set and the final Map value are respectively collected. The improved network model increases the range of AP values for each category from 0.05 to 0.4. The improved network model Map is 0.712, which is 0.4 percentage points higher than the original model, indicating that the application of GioU to the classic object detection model SSD is performing well and improves its detection accuracy.

Table 2. The experimental comparison data

Class / AP	AP		**Improvement%**
	IoU	**GIoU**	
Aeroplane	0.7360	0.7729	+0.369
Bicycle	0.7881	0.7720	-0.161
Bird	0.7098	0.6954	-0.144
Boat	0.6189	0.6613	+0.424
Bottle	0.3986	0.3754	+0.680
Bus	0.7980	0.8147	+0.167
Car	0.7852	0.8177	+0.325
Cat	0.8534	0.8203	-0.331
Chair	0.5246	0.5029	-0.217
Cow	0.7648	0.7882	+0.234
Diningtable	0.6779	0.7296	+0.517
Dog	0.8188	0.8173	-0.150
Horse	0.8090	0.7864	-0.229
Motorbike	0.7549	0.7543	-0.600
Person	0.7445	0.7574	+0.129
Potted Plant	0.4423	0.3843	-0.580
Sheep	0.7131	0.7450	+0.319
Sofa	0.7216	0.7613	+0.397
Train	0.8172	0.8050	-0.122
Tvmonitor	0.7049	0.6970	-0.790
MAP	**0.7091**	**0.7129**	**+0.380**

4.3 Comparison of Experimental Results

Figure 4. Each group from left to right respectively is the original image, SSD detection image after using the IOU, SSD detection image after using the GIOU

5. Fifth Conclusion

The GIoU bounding box loss function proposed based on the latest CVPR2019 paper in object detection field is innovatively applied to the current object detection mainstream framework SSD network model. By loading the pre-trained VGG16 model, the convergence speed is improved without interfering the experimental comparison. By testing on the standard test dataset PASCAL VOC, the improved SSD network model Map reached 71.29%, which is 0.4 percentage points higher than the original network model Map value (70.91%). Therefore, the improved SSD network model has a higher accuracy and a better effect. The next step is to add more experimental comparison, test on the COCO dataset and optimize the GioU's inability to distinguish alignments, thus improving the accuracy of the SSD object detection model.

References

[1]. LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multiboxdetector [C]. European Conference on Computer Vision, 2016: 21-37.

[2]. M. A. Rahman and Y. Wang. Optimizing intersection-overunion in deep neural networks for image segmentation. In International Symposium on Visual Computing, pages 234– 244, 2016. 3.

[3]. M. B. A. R. T. Matthew and B. Blaschko. The lov´aszsoftmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 3.

[4]. PENG Hongxing, HUANG Bo, SHAO Yuanyuan, et al. General improved SSD model for picking object recognition of multiple fruits in natural environment[J/OL]. Transactions of the CSAE,2018, 34(16): 155-162.

[5]. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In ECCV, 2016.

[6]. Fu C Y, Liu W, Ranga A, et al. DSSD : Deconvolutional Single Shot Detector[J]. 2017. He K, Zhang X, Ren S, etal. Deep Residual Learning for Image Recognition [J]. 2015:770-778.

[7]. S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. In CVPR, pages 789–798, 2016. 3.

[8]. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. arXiv preprint arXiv:1711.07752(2017).

[9]. Gao Jianwei, Li Lei, Yao Rui, et al. Real-time Detection and Tracking of Weak and Small Objects Based on Kalman Filter[J]. Computer Engineering, 2012, 38(2).

[10]. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second.In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2903–2910. IEEE (2012).

[11]. C. L. Zitnick and P. Doll´ar. Edge boxes: Locating object proposals from edges. In Computer Vision–ECCV 2014, pages 391–405. Springer, 2014. 2, 4.

[12]. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5079–5087 (2015).

[13]. Liu Kun, Liu Weidong. Infrared weak object detection algorithm based on weighted fusion feature and Ostu segmentation [J]. Computer Engineering, 2017, 43(7) : 253-260.

[14]. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1259–1267 (2016).

[15]. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. arXiv preprint arXiv:1711.07752 (2017).