IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 31, 2019, accepted November 20, 2019, date of publication November 25, 2019,
date of current version December 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955685

Occlusion Problem-Oriented Adversarial Faster-RCNN Scheme

QINGYANG XU[1], XIAOFENG ZHANG[1], RUOSHI CHENG[1], YONG SONG[1], AND NING WANG[2], (Senior Member, IEEE)

[1]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China
[2]School of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China

Corresponding author: Qingyang Xu (qingyangxu@sdu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603214, Grant 61973184, Grant 61673245, and Grant 61803227, in part by the National Key Research and Development Plan of China under Grant 2017YFB1300205, in part by the Shandong Province Key Research and Development Plan under Grant 2018GGX101039 and Grant 2016ZDJS02A07, in part by the China Postdoctoral Science Foundation under Grant 2018M630778, and in part by the Independent Innovation Foundation of Shandong University under Grant 2018ZQXM005.

ABSTRACT In the practical scene, object detection faces a very complicated situation. The occlusion problem always occurs in actual scene, which may affect the accuracy of object detection, especially for the occluded objects. For the deep models, a larger dataset with sufficient occlusion samples will improve the performance of the object detection models. However, the sample with occlusion problem is too hard to obtain. Therefore, a global average pooling(GAP) based adversarial Faster-RCNN is proposed to generate the hard samples and enhance the performance of object detection algorithm. Sufficient hard samples can be generated with the help of this model. Therefore, the object detection model can be trained adequately for the occluded objects. The hard sample generation is carried out in the space of image feature instead of image generation directly. The class-dependent part is obtained by the GAP network, and it is obscured to generate the feature map of hard sample for model reinforcement training. Therefore, the better object detection model can be trained using a conventional dataset. The Faster-RCNN is adopted as the baseline. The Faster-RCNN and GAP have a joint training to improve the performance of the proposed model. The simulation results exhibit the validation of the proposed algorithm.

INDEX TERMS Faster-RCNN, object detection, occlusion problem-oriented, global average pooling.

I. INTRODUCTION

Object detection is a hot topic in visual perception, which provides the information for the image and video understanding [1], [2]. It has a wide range applications, such as automatic vehicles [3], video surveillance, and robotics. The object detection algorithm has developed from the classical methods to deep learning based methods [4]. The classical methods include the process of sliding window generation, image feature extraction and classification. However, the accuracy and speed are limited. The deep learning based methods have enhanced the performance of object detection algorithm to a new level, which can be categorized into two types, namely region and region-free based algorithms. The deep learning based algorithms make use of improved architectures, larger training sets, and end-to-end training mode to strengthen the performance of the object detect algorithm. The representative region-based models include Region-based Convolutional Neural Network (R-CNN), SPP-Net [5], Fast R-CNN [6] and Faster R-CNN [7]. The representative region-free approaches include Single Shot MultiBox Detector (SSD) [8] and You Only Look Once (YOLO) [9]. Unfortunately, the object detection algorithms suffer from several challenges, such as scene complexity, deformation and occlusion [10]–[12]. Deformation and partial occlusion problem are still a major challenge to the state-of-the-art object detectors [13], [14]. In general, the occlusion problem can be divided into inter-class occlusion and intra-class occlusion type problem. The object is occluded by the fixed stuff or some other typed objects can be considered as an inter-class occlusion problem. If the object is occluded by the same type of object, it can be seen as an intra-class occlusion problem which also can be seen as a crowd occlusion problem [15]. For the intra-class occlusion problem, the pedestrian detection is a popular problem due to the application requirements, such as auto-driving, video surveillance etc. Also, the inter-class occlusion problem occurs frequently, such as the pedestrian occluded by

The associate editor coordinating the review of this manuscript and approving it for publication was Xian Sun.

car. In order to handle these problems, various star models [14], [16], tree models [17] and graph models [18] etc. were employed, and achieved a better performance. For the occlusion problem handling, the classical approach is the partial detection method [19]. Since visibility estimation plays a key role for occlusions handling, various approaches were proposed to estimate the visibilities of parts [20]–[23] which had improved the performance of object detection algorithm to some extent. The artificial features of images, such as Histogram of oriented gradient (HOG) and local binary pattern (LBP) features etc., which will be fed into the classifiers for classification. Sometimes, the boosting technology is adopted to train kinds of detectors for specific occlusion problem, and then the result is the output of ensemble model. The performance of the model degrades when the object becomes partially occluded. In order to deal with the inter-class problem better, the joint part combination approach was adopted which obtained the occlusion map by occlusion reasoning or segment results [20]. Tang et al. proposed a different model to identify the occlusion patterns [24], Ouyang et al. proposed an integrated framework to improve the capability of object detection algorithm for occluded object [25], and Pepik et al. incorporated the appearance of occluded object directly [13]. However, these algorithms rely on the detection scores of parts and always fail to capture the correlations of random parts, especially for the complex occlusion patterns. On the other hand, the time consuming is huge. For deep learning based object detection algorithm, the typical solution to overcome these problem is to collect large scale dataset which contains kinds of object instances [26], such as COCO dataset [27] which contains 10K samples under different occlusions and deformations. However, it is believed that some occlusions follow a long-tail distribution, namely some occlusions are too rare to appear in large scale dataset [26], and some occlusions occur frequently than others especially in crowded scenes [14]. Data is an important factor affecting the performance of deep learning model. The richness of data determines the convergence and accuracy of deep neural networks. Therefore, the rich type can be detected well, the scare one may be detected bad.

However, in the conventional dataset, the frequency of occlusion appears to be rare, especially for the scare one, and it is always considered as the hard samples. Hard samples are relatively indistinguishable for the deep model, which loses partial information [26]. If kinds of hard samples are gathered and used to reinforce the object detection model training, the performance of object detection algorithm will be improved. Therefore, hard samples mining is an important research topic in deep learning. In the context of class imbalance in object detectors training, on-line hard sample mining (OHEM) was designed to emphasize hard samples [28]. The OHEM algorithm made use of Fast-RCNN [29] as the baseline, which combines two fully connected layers. One full connected layer is responsible for the scoring of the proposed region, and the hard sample has the higher score with larger loss value. Another fully connected layer is adjusted by the BP algorithm according to the score to enhance the performance of Fast-RCNN. Repulsion Loss [30] and Occlusion-aware R-CNN [31] attempted to add a penalty term in the loss function to reduce the gap between the proposed region and the Ground Truth. Zhang et al. integrated the attention mechanism to the faster-RCNN to improve the occluded objects detection accuracy [32]. These researches are relied on the model improvement or better loss function participation. These approaches make use of additional parts to improve the occluded objection detection accuracy which increase the complexity of the original models sometimes.

A different way to improve the performance of detectors is to better exploit the dataset which is also the source power of deep learning. Shrivastava et al. tried to incorporate the hard sample for training region based CovNets by hard sample mining [28]. Recently, some studies made use of generative models to generate hard samples, such as GAN [33] to generate as many images as possible with occlusion objects for data enhancement. The GAN network consists of two parts: the discrimination model and the generated model. The two models continually reinforce each other and ultimately improve each other's capabilities by generation adversarial samples through the zero-sum game theory. The adversarial sample is the sample that little changes will make the machine learning algorithm output an erroneous value [34]. However, it is not a feasible solution due to the requirement of large scale dataset containing occluded objects. In addition, the space of occlusion problem is huge. A-Fast-RCNN [26] aimed to generate adversarial samples in the feature space of deep neural network, and the random mask is compounded with the feature to generate a new feature. These features are then sent to the fully connected layer for scoring, and the feature with high scores drops will be used to the hard sample. The performance of Fast-RCNN is enhanced for better detecting the occluded object by this way. It is an easier operation than the image generation. However, it is a random shelter of features without purpose in A-Fast-RCNN. Inspired by this idea, an improved scheme is proposed to deal with a wide range of occlusion problem in object detection. The Faster-RCNN [7] is selected as the baseline of the object detector, and has a fusion with the global average pooling (GAP) [35] to gather the Class Activation Mapping (CAM) of Faster-RCNN. Finally, the class-dependent (Class-specific) is located in the feature of Faster-RCNN and shielded off to generate the feature of hard sample. This feature of hard samples is used for the training of Faster-RCNN, which endows Faster-RCNN with the ability of hard sample detection and occlusion object detection.

II. GLOBAL AVERAGE POOLING BASED CLASS DEPENDENT FETURE GATHERING

A. CLASS DEPENDENT FEATURE

The selection of features in pattern recognition tasks is critical. Some important features of object are extracted to distinguish it from other classes, namely class-dependent

features [35]. Therefore, the purpose of class-dependent feature selection is to extract the decisive feature for the class discrimination. For the deep learning based classifier and detector, the hard sample can be obtained by the masking of class dependent feature of the convolution neural network as ref [26]. Therefore, this paper adopts the method of processing the feature of convolutional neural network to obtain hard samples and strengthen the training of the object detection model instead of relying on hard sample image generation. Thereby, it is especially important to locate the class dependent feature. The feature of hard sample is generated by the processing of class dependent feature.

B. GLOBAL AVERAGE POOLING

Deep learning is less interpretable than other models. Recently there are many ways to better understand CNN by visualizing CNN [36]. Deconvolution based feature visualization is the first work in the field visual understanding of CNN [37]. Deconvolution is used to visualize features and explain CNN's feature learning process and the feature characteristics of each layer. Through feature visualization, the feature extraction process of convolutional neural networks can be understood. However, which features play a leading role for the final classification. GAP is used to reduce the complexity of the convolutional neural network firstly [38]. The CNN is always followed by fully connected layer. The fully connected layer is replaced by the GAP with lower dimension and good performance. Additionally, the GAP can maintain a better spatial information. Therefore, the Class Activation Mapping (CAM) technique made use of GAP to find the class dependent feature [35], and it makes us to visually track the crucial part of the image corresponding to the class dependent feature. Therefore, the GAP is adopted to analyze the extracted features by CNN, and explain which one is the class-dependent features for the hard sample feature generation.

For a certain CAM, it represents which partial feature of CNN is the discriminant basis of this class, namely the class-dependent part, and explains the basis for the model to classify the object into a certain class. The input image is convoluted by the filter layer by layer, and the last feature map is followed by GAP instead of full connected layer. The mean value of each feature map is obtained by GAP, and then the mean values have a weighted sum as the input of softmax function as equation (1).

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y) \quad (1)$$

where w_k^c is weight of the feature map of category c. $f_k(x,y)$ is the feature map.

After the computation of $w_k^c f_k(x,y)$, the one with biggest value is the class dependent section, and the heat map corresponding to each feature can be created. Since the size of heat map is inconsistent with the input image, an upsampled operation is carried out and it is superimposed on the original image. Figure 1 shows the class-dependent portion of the convolutional neural network obtained by GAP method and the heat map is mapped back to the corresponding position of the image. The highlighted portion of the CAM is the class-dependent portion. After the obtaining of class-dependent features, the class-dependent part of image also can be gathered by CAM, which can exhibit the crucial part of the image for classification.

III. GLOBAL AVERAGE POOLING BASED ADVERSARIAL FASTER-RCNN

A. FRAMEKWORK OF ADVERSARIAL FASTER-RCNN

Fast-RCNN makes use of the convolutional neural network to extract the feature map of input image. The proposal regions are generated according to the Selective Search algorithm [6] and the coordinate of the proposal regions are mapped to the feature map to get the corresponding block of feature, which only have one convolution operation unlike RCNN [39] convoluting for every proposal region.

Faster-RCNN includes Fast-RCNN and RPN two main sections. RPN is the core network which is used to select the proposal region for ROI region generation, and then a coordinate regression and classification are carried out for ROI region gathering instead of all regions.

Supposing the definition of object detection network is $F(X)$, X is the proposal region. The two outputs of the network are categories F_C and coordinates of region F_l. C is the label of X, L is the coordinate of Ground Truth. The loss function of object detection model can be described as eq. (2).

$$L_F = L_{softmax}(F_c(X), C) + [C \notin bg] L_{bbox}(F_l(X), L) \quad (2)$$

where $L_{softmax}$ is the *softmax* function of category, L_{bbox} is the loss function of region coordinate.

Supposing the hard sample is defined as $A(X)$, X is the foreground object feature of proposal region. The loss function of the hard sample is defined as equation (3).

$$L_A = L_{softmax}(F_c(A(X), C)) \quad (3)$$

The label information of hard sample is unchanged. Therefore, the loss function used for the hard sample is shown as equation (4), so as to improve the robustness of the object detection algorithm.

$$\begin{aligned} L_{total} &= L_F + L_A = L_{softmax}(F_c(X), C) \\ &\quad + L_{softmax}(F_c(A(X), C)) \\ &\quad + [C \notin bg] L_{bbox}(F_l(X), L) \end{aligned} \quad (4)$$

where,

$$L_{bbox} = \sum_{i \in (x,y,w,h)} smooth_{L1}(t_i^u - v_i) \quad (5)$$

The loss function of RPN is defined as equation (6).

$$\begin{aligned} L(\{p_i\}, \{t_i\}) &= \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) \\ &\quad + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \end{aligned} \quad (6)$$

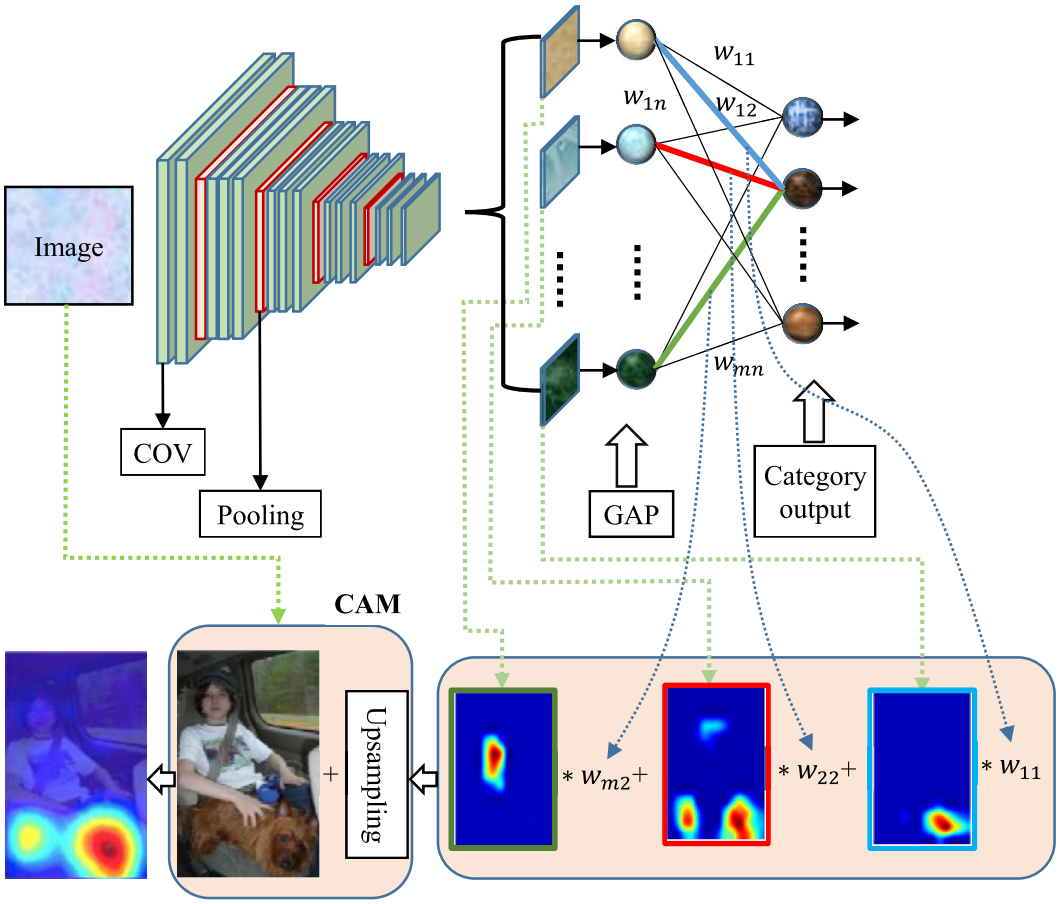


FIGURE 1. Global average pooling.

where p_i is the predicted probability of each anchor, p_i^* represents whether the anchor is foreground or background, t_i and t_i^* are the coordinates of foreground anchor and Ground Truth, N_{cls} is set as 256, N_{reg} is about 2000 and λ is set to 10. The parameters is set according to ref. [6].

B. FUSION OF FASTER-RCNN AND GAP

Faster-RCNN is adopted as the baseline of object detector. The input image is convoluted by filters, and the RPN network is used for the proposal region selection. The ROI Pooling layer has a fixed size feature map for classification and regression, and gets the detection result. It is a standard training process of Faster-RCNN. The GAP part is added to the model as figure 2. The ROI Pooling layer is followed by GAP part to gather the class-dependent feature. The class dependent feature coming from the GAP will be removed from feature map channel to generate the feature of hard sample which is used for the model training. The GAP and Faster-RCNN are trained jointly.

In the training of GAP, partial parameters of Faster-RCNN are fixed and GAP is trained as a classifier. Since, there is only one full connected layer in GAP, a 3*3 and 1*1 filters are added before the GAP to overcome the unfitting phenomenon as figure 3.

The loss function of GAP training is as equation (7).

$$Loss = L_{softmax}(G(X), C) \tag{7}$$

where $G(X)$ is the output of GAP, C is the label.

In order to exhibit the effect of the fusion of GAP and Faster-RCNN intuitively, all the ω_k^c corresponding to the detected categories are gathered, and the weighted sum of them and the feature maps of corresponding proposal region are obtained. After the training, the weights of each category are extracted and sum them with the corresponding feature maps to obtain CAM.

Since the size of the result is inconsistent with the feature map, it needs to be upsampled and superimposed on the original image. The final results are shown as figure 4. In figure 4, the first column is the input image with bounding box, the second column is the proposal region with CAM indication, the third column is the feature map of proposal region in Faster-RCNN fused with GAP and the highlight part is the class-dependent part.

The training of the proposed model contains the process of Faster-RCNN and GAP training as shown in Algorithm 1. The Faster-RCNN is trained firstly, and then the GAP is trained based on the pre-trained Faster-RCNN. Then, the Faster-RCNN and GAP are pre-trained completely.

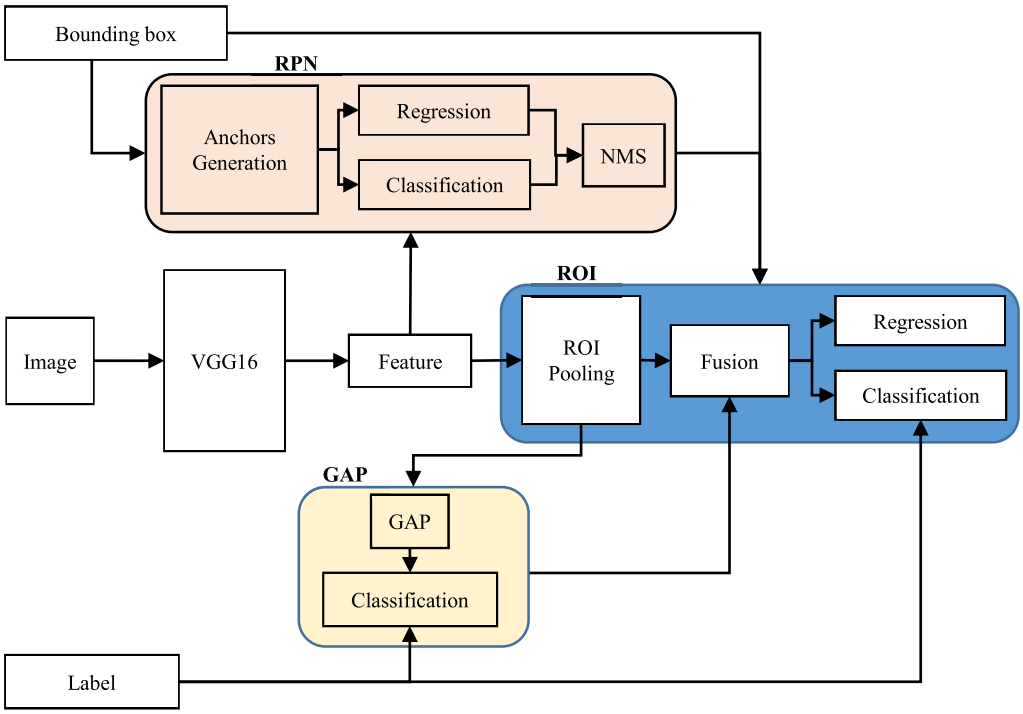


FIGURE 2. Framework of the proposed scheme.

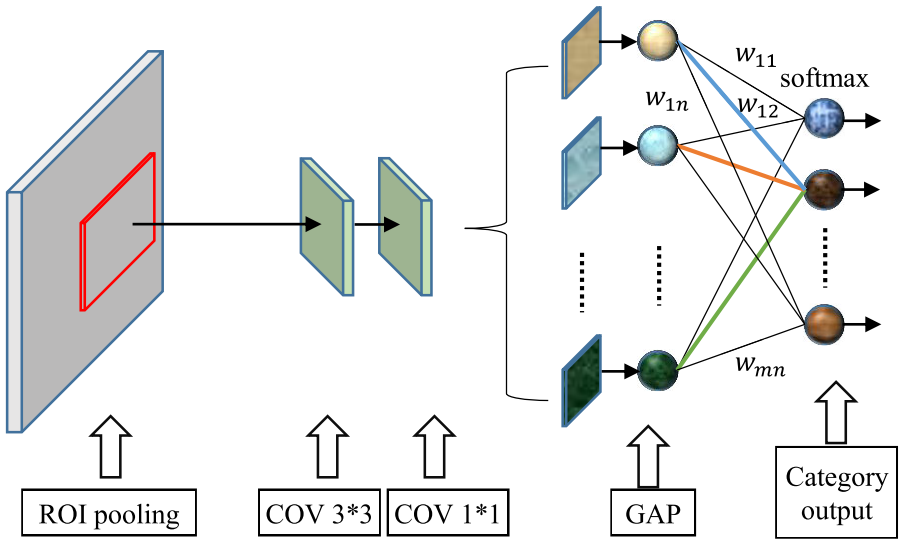


FIGURE 3. GAP fused with faster-RCNN.

Then, the GAP and Faster-RCNN are trained jointly. Namely, the GAP is with fixed parameters when the Faster-RCNN is in training, and then the Faster-RCNN is with the fixed parameters when the GAP is in training. The pseudocode of the proposed algorithm is shown in Algorithm1.

C. MULTI-SCALE FEATURE FUSION

The feature size of each foreground object X is d^*d^*c, d is the width and length of feature map, and c is the number of channels. For traditional Faster-RCNN, the feature map size is 7*7. In this paper, the 7*7 feature map is also used for the hard sample feature generation with 512 channels, which is sent to the GAP to obtain the class-dependent part. Then, the class-dependent part is set to 0 in order to generate the hard sample feature. The 7*7 class-dependent partial feature maps are mapped to a fixed length vector, and the 5*5 and 3*3 feature maps are also generated and mapped to the corresponding length vector. The three sizes vectors are combined to achieve feature fusion [5], and then the fused feature inputs to the full connected layer for classification and regression which increases the richness of features [40].

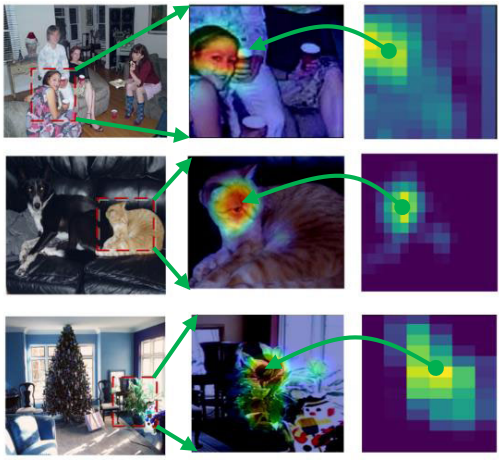


FIGURE 4. CAM of region proposal based on GAP.

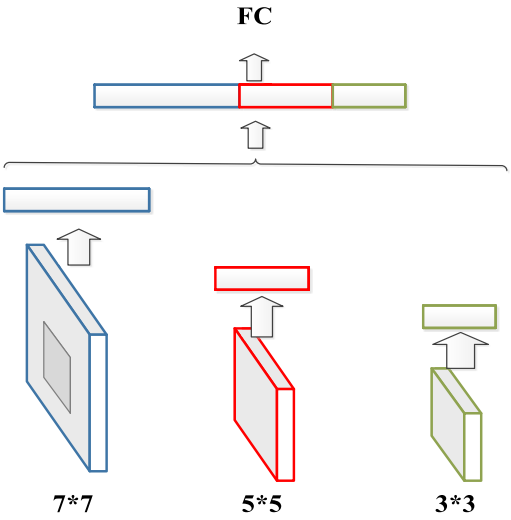


FIGURE 5. Diagram of multi-scale feature fusion.

Algorithm 1 Pseudocode of the Proposed Algorithm

1. Pre-trained Faster-RCNN
   1) RPN training based on the pre-trained model by ImageNet
   2) Proposals collection based on RPN
   3) Fast-RCNN training firstly
   4) RPN training based on the training of Fast-RCNN
   5) Fast-RCNN training again
2. Pre-trained GAP
   GAP training based on the Pre-trained Faster-RCNN
3. GAP and Faster-RCNN training jointly
   For $i = 1$ to n
      1) GAP training
      2) Faster-RCNN training
   End For

IV. SIMULATION STUDIES

A. PRAMETERS SETTING

In order to testify the performance of the proposed scheme, the PASCAL VOC and COCO datasets are adopted. The PASCAL VOC is a benchmark for object classification and detection of visual models [41], includes 20 categories, such as human, animals (birds, cats, cows, dogs, horses, sheep), vehicles (aircraft, bicycles, boats, buses, cars, motorcycles, trains), and indoors (bottles, chairs, dining tables, Potted plants, sofas, TV). The VOC2007 includes 5K images for training and 5K images for testing. The COCO dataset is proposed by Microsoft team for the evaluation of image detection and segmentation task. The COCO dataset has more than 200,000 images and 80 object categories. All object instances are labeled with a detailed split mask, and there are more than 500,000 object entities with label. The COCO data set is much larger than the PASCAL VOC with more object types and it is more difficult for the visual task.

The Tensorflow framework is adopted to implement the entire algorithm with Cuda version is 9.2, and the cudnn version is 6.1. The image is processed by OPENCV3.4.1. The hardware system is configured with a 11G NVIDIA GTX 1080ti graphics card, Inter Xeon (R) E5-1620 (3.50GHz, 8 core) CPU, 16G RAM.

A pre-trained model by ImageNet is adopted to initialize the VGG network. The GAP shares the VGG and RPN networks with the Faster-RCNN. In the experiment, it was a crucial step to pre-train the class-dependent feature extraction part. According to Faster-RCNN's two-step training method, Fast-RCNN and RPN network have a joint training firstly. Then, the pre-trained Faster-RCNN is used to pre-train the class dependent feature extraction part. After 7K iterations, the GAP network can extract the feature of class-dependent part. The loss value of GAP part training is shown in figure 6.

Subsequently, the pre-trained GAP is jointly trained with the Faster-RCNN. The specific details are as follows: the weights of the Faster-RCNN are fixed when the GAP is in training, and the parameter of the GAP part are fixed when the Faster-RCNN is in training, and the training process is alternately performed. The model is trained by an end-to-end mode. The 7*7, 5*5 and 3*3 feature maps are obtained after the RPN network, and the 7*7 feature maps are sent to the GAP to obtain the class-dependent part. And then, the class-dependent part is set to 0 in order to generate the hard samples. The feature of hard sample is fused with the other two size maps to train the Faster-RCNN.

The Momentum optimizer is used in Faster-RCNN, and the Adam is adopted for the training of GAP. During the joint training process, the learning rate of Faster-RCNN is initially set to 0.0001 when training on the VOC2007 data set, and is attenuated to 0.00001 after 30,000 training iterations. The momentum value is set to 0.9. The learning rate of GAP is set to 0.00001. In joint training, the initial learning rate is set to 0.0001 and decays to 0.00001 after 30,000 training iterations. During the training, the minibatch size is set to 2, resulting in 256 recommended areas. The three parts are each trained for 70,000 iterations, and consumes about 20 hours. The parameter settings trained on the COCO data set are

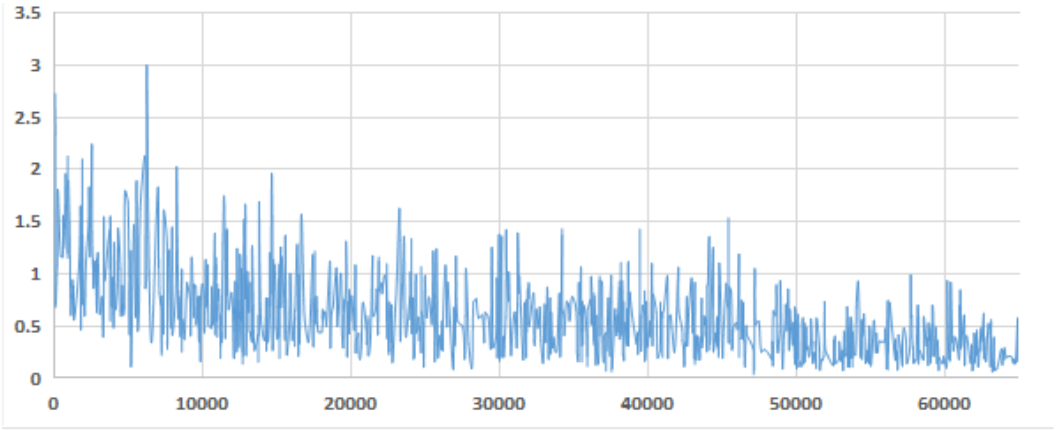


FIGURE 6. The loss value of GAP.

the same as the training parameter settings on the VOC data set. Due to the large amount of data in the COCO dataset, the three parts are trained 320,000 iterations respectively, and the learning rate began to decay after 280,000 iterations. The entire training process consumes about 90 hours.

B. EVALUATION METRICS

The universal evaluation metrics are adopted in his paper. True Positive (TP) means that the positive samples are predicted as positive one. False Positive (FP) means that the negative samples are predicted as the positive one falsely. False Negative (FN) means that the positive samples are predicted as the negative one correctly. According to detection results, the precision can be computed by the following equation:

$$Precision_{iC} = \frac{TP_{iC}}{TP_{iC} + FP_{iC}} = \frac{TP_{iC}}{N_{iC}} \tag{8}$$

$$AvePrecesion_C = \frac{\sum_{image} Precision_{iC}}{NC_{image}} \tag{9}$$

$$mAP = \frac{\sum_C AvePrecision_C}{N_{classes}} \tag{10}$$

where iC is the object label in an image, N_{iC} is the object number in an image, $Precision_{iC}$ is the detection precision for class C in single image, NC_{image} is the object number of class C in the dataset, $AvePrecesion_C$ is the average precision of class C in the dataset, $N_{classes}$ is the class numbers in the dataset, mAP is the mean average precision for all types object.

C. ABLATION STUDY

In order to investigate the effectiveness of the GAP part and fusion module, the ablation studies are carried out based on the VOC2007 dataset.

According to figure 2, the Faster-RCNN is adopted as the baseline of this model. Therefore, this model will degrade into Faster-RCNN when the GAP and fusion model are removed from the proposed model. The results are shown in Table 1 (FO is the failure detection sample with occlusion).

TABLE 1. The ablation studies based on VOC2007 data set.

	mAP	FO
Faster-RCNN	68.1	99%
Faster-RCNN+GAP	70.0	5%
Faster-RCNN+feature fusion	68.9	97%
Faster-RCNN+GAP+feature fusion	**70.2**	**2%**

1) EFFECTIVENESS OF THE GAP PART

The GAP part is the key module for hard sample generation. With the help of GAP, the dependent feature of specific categories will be gathered, and the feature of hard samples can be obtained by feature fusion without the dependent feature. Therefore, the Faster-RCNN can be trained by more samples, especially with more hard samples. According to Table 1, the effectiveness of the GAP can be observed. Compared with traditional Faster-RCNN, the model with GAP can improve the object detection accuracy, especially for some occluded objects. For Faster-RCNN, the detection failure samples consist of 99% occluded objects. With the help of GAP, the occluded objects can be detected properly.

2) EFFECTIVENESS OF THE FEATURE FUSION

According to figure 5, the feature fusion module integrates features with different scales. By the integration, there will be many rich features used for classification to improve the detection accuracy as Table 1. However, the improvement is limited, some occluded objects are also hard for detection due to the absence of GAP part.

3) EFFECTIVENESS OF THE PROPOSED MODEL

By the integration of GAP and feature module into traditional Faster-RCNN, the sample number is increased especially for the hard samples, and the features used for classification become richer. Therefore, the objection detection accuracy is improved especially for the occluded objects as Table 1.

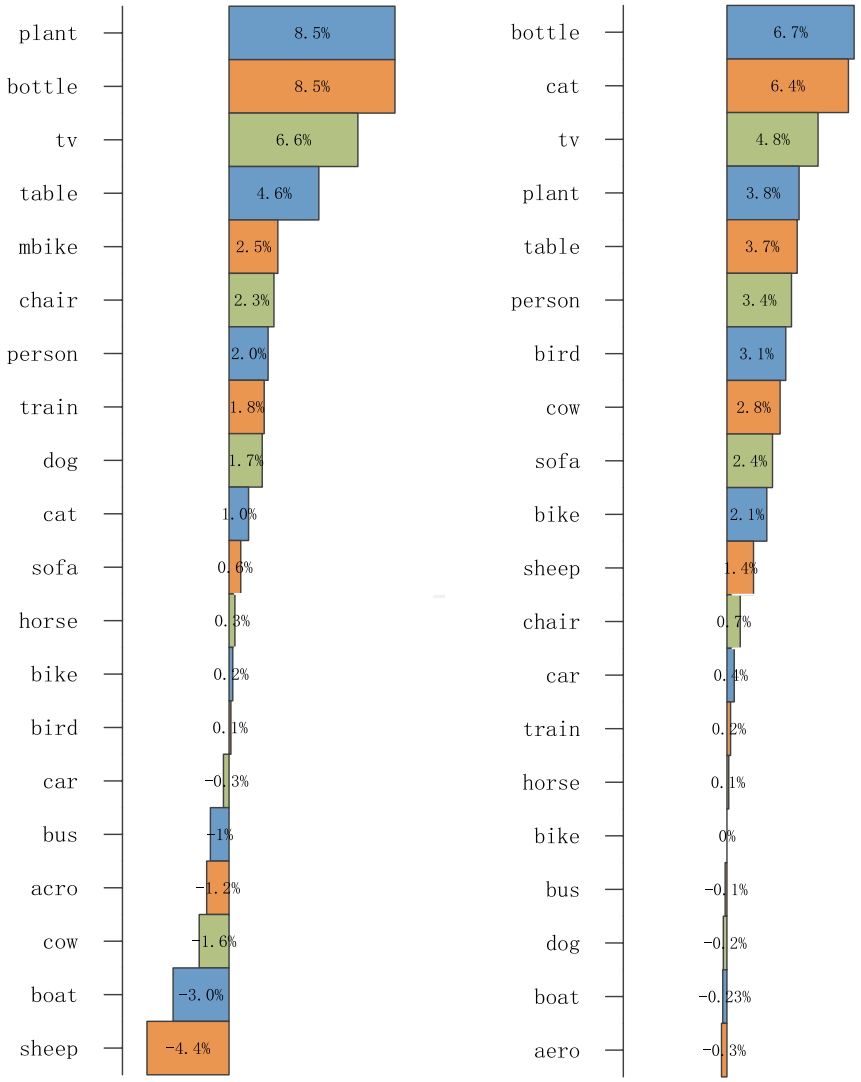


FIGURE 7. The comparison of the accuracy improvement of the object detection for each category.

TABLE 2. The detection results comparison on VOC2007.

Methods	mAP
RCNN(Alex) [39]	68.1
Faster-RCNN(VGG)	68.1
SPP-net(ZF) [5]	68.5
GCNN [43]	68.3
SubCNN [42]	**70.2**
A-Fast-RCNN [26]	69.9
Our algorithm	**70.2**
Faster-RCNN(ReIm) [30]	79.5
Faster-RCNN(ReIm)+Rep [30]	***79.8***

D. EXPERIMENTAL RESULTS

For the conventional object detection algorithm, the purpose of the improvement is to enhance the mAP with some improved architectures, and got an impressive promotion. For example, SubCNN [42] made use of a complex subcategory classifying layer to detect the object. Although the performance of the algorithm is improved, it doesn't

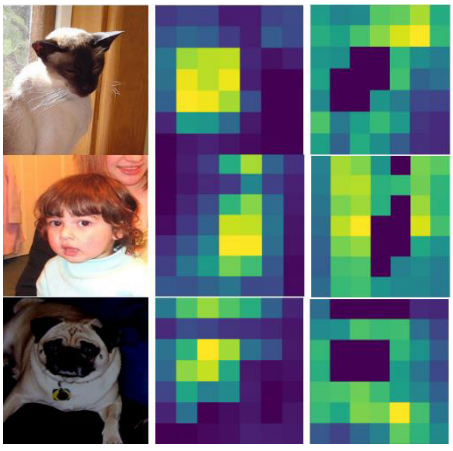


FIGURE 8. The visualization of class dependent feature masking.

(a) Object detection by conventional Faster-RCNN (b) Object detection by the proposed method

FIGURE 9. The detection results of VOOC2007 DATASET.

face the occlusion problem. For the occluded object detection problem based on deep learning technique, detection model improvement is also the first choice for the researchers. A compound loss function is defined to evaluate the occluded object which is called repulsion loss [30]. The new loss contains the attraction of ground truth and the repulsion of surrounding objects. Although it has got a promising results, the complexity of the model is increased. A direct way is to increase the scale of dataset, especially for the proportion of occluded objects. The performance of Faster-RCNN for occluded objects detection may be improved by the augment of dataset. Therefore, a comparison is carried out based on the coco dataset for the common occluded objects detection. Table 2 shows the performance of different object detection models and the proposed method on the VOC2007 data set. It can be seen from Table 2 that the accuracy based on the global average pooling against other models is improved, except for the SubCNN. It has the same performance

TABLE 3. The detection results comparison on COCO.

Methods	mAP
Fast RCNN [6]	39.9
ION [44]	43.2
NOC+FRCN(VGG) [45]	41.5
OHEM+FRCN [28]	42.5
SSD+Rep [30]	40.9
A-Faster-RCNN [26]	42.7
Our algorithm	**44.5**
NOC+FRCN(Google) [45]	44.4
NOC+FRCN(ResNet101) [45]	48.4
R-FCN(ResNet101) [46]	***51.5***

as SubCNN. Compared with A-Fast-RCNN, the accuracy is also improved. The backbone network also affects the performance of the object detection algorithm which decides the capability of feature extraction. Faster-RCNN(ReIm) is a re-implement of Faster-RCNN by ref [30] which adopts the ResNet50 as the backbone network. The re-implement Faster-RCNN with ResNet50 backbone network got a performance of 79.5%mAP. With the help of Repulsion loss function, the Faster-RCNN(ReIm) got a 79.8%mAP. Therefore, compared with Faster-RCNN(ReIm), the performance of our algorithm is worse due to capability of backbone network.

The proposed method is evaluated by anther data set COCO. Table 3 shows the performance of some object detection models and the proposed method based on the COCO dataset. For conventional object detection algorithm with VGG backbone network, the performance of our model is promising. With the help of backbone network, the object detection algorithms get a better feature extraction capability. Therefore, a deeper backbone network (Google, ResNet101) can improve the performance of conventional object detection algorithm.

Figure 7 shows the comparison of the accuracy improvement of the object detection for each category by A-Fast-RCNN and our proposed method. Figure 7 (a) is the accuracy improvement of each category detection by A-Fast-RCNN, and Figure 7 (b) is the accuracy improvement of each category detection by our proposed method. It can be seen that both methods have a greater accuracy improvement in the detection of bottle, cat, etc., because these objects are occluded at a higher frequency in the data set. However, for some object such as boat, bus etc., the improvement is limited. Because they are always without blocking the view. Compared with A-Fast-RCNN, the proposed method does not have a higher improvement of detection accuracy than A-Fast-RCNN, but it is more stable than A-Fast-RCNN. For example, A-Fast-RCNN's detection accuracy for sheep is reduced by 4.4%, and the detection accuracy of the boat is reduced by 3.0%, which is a large performance degradation. The maximum amplitude of performance degradation for our method is about 0.3%. therefore, the proposed method is more stable.

Figure 8 is the visualization of the class dependent feature masking process for the proposal region base on VOC 2007. In figure 8, the first column image is the proposal region image, the second column is the positioning of the class dependent part in the feature by the GAP, and the third column is the feature map of masking the class dependent part.

Figure 9 shows the detection of occluded object selected in the VOC2007 data test set. The first column (a) is the detection results by the conventional Faster-RCNN, and the second column (b) is the detection results by the proposed algorithm. For the picture (a1), the conventional Faster-RCNN takes the cat for a dog due to half of the cat's face was occluded, and the proposed model successfully detected the cat as (b1). For the picture (a2), conventional the Faster-RCNN cannot detect the cat caused by the occlusion of window's fence. The improved Faster-RCNN can successfully detect the cat as (b2). For the picture (a3), there is severe occlusion problem of people. Therefore, only three persons are detected for the traditional Faster-RCNN. For the proposed algorithm, the five persons are detected completely. The picture (a4) also misses a plottedplant by Faster-RCNN, and it is detected the proposed model. For the conventional Faster-RCNN detection results, the missed one is marked by the green boxes.

V. CONCLUSION

There are many difficulties in the object detection task, such as occlusion, illumination, etc., leading to missed detection and false detection. Some methods are often devoted to data enhancement by augmenting data sets to enhance the performance of object detection model. In this paper, a global average pooling based adversarial Faster-RCNN is proposed, which makes use of the class dependent feature selection and masking to generate hard sample feature. The reinforcement training of the hard sample, the performance of occluded object detection of Faster-RCNN is improved. Therefore, we can make use of a conventional dataset to train a better object detection model.

Although the experiments show a promising results for the proposed scheme, there are also some shortcomings. The speed is still a common problem related to the region-based object detection algorithm. Besides, in the analysis of class dependent part, the partial occlusion of feature is not effective for some objects due to the correlation of foreground and background. Therefore, the auxiliary semantic analysis will be our future works.

ACKNOWLEDGMENT

The authors would like to thank for the reviewers' help and suggestions, and tanks for the source code provider of https://github.com/bailvwangzi/ repulsion_loss_ ssd.

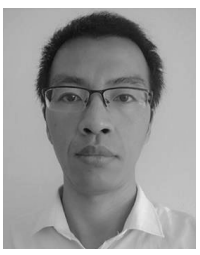
REFERENCES

[1] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion and online hard example mining," *IEEE Access*, vol. 6, pp. 19959–19967, 2018.

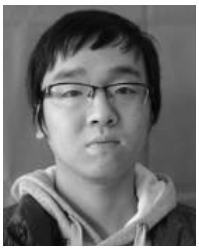
[2] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[3] Z. Liang, J. Shao, and D. Zhang *et al.*, "Traffic sign detection and recognition based on pyramidal convolutional networks," *Neural Comput. Appl.*, pp. 1–11, Mar. 2019, doi: 10.1007/s00521-019-04086-z.

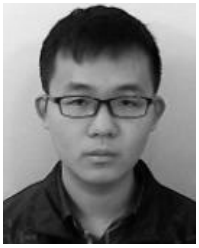
[4] F. Yang, H. Chen, J. Li, F. Li, L. Wang, and X. Yan, "Single shot multibox detector with Kalman filter for online pedestrian detection in video," *IEEE Access*, vol. 7, pp. 15478–15488, 2019.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[9] 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.
[10] J. Chu, T. Zhu, and J. Miao, "Target tracking based on occlusion detection and spatio-temporal context information," *Pattern Recognit. Artif. Intell.*, vol. 30, no. 8, pp. 718–727, 2017.
[11] S. K. Choudhury, P. K. Sa, S. Bakshi, and B. Majhi, "An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios," *IEEE Access*, vol. 4, pp. 6133–6150, 2017.
[12] G. Yu, Z. Hu, H. Lu, and W. Li, "Robust object tracking with occlusion handle," *Neural Comput. Appl.*, vol. 20, pp. 1027–1034, Oct. 2011.
[13] B. Pepikj, M. Stark, P. Gehler, and B. Schiele, "Occlusion patterns for object class detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3286–3293.
[14] W. Ouyang, X. Zeng, and X. Wang, "Partial occlusion handling in pedestrian detection with a deep model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2123–2137, Nov. 2016.
[15] S. Tang, M. Andriluka, and B. Schiele, "Detection and tracking of occluded people," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 58–69, 2014.
[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[17] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.
[18] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr, "A study of parts-based object class detection using complete graphs," *Int. J. Comput. Vis.*, vol. 87, p. 93, Mar. 2010.
[19] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 990–997.
[20] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 185–204, Apr. 2009.
[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[22] J. Marin, D. Vazquez, A. M. Lopez, J. Amores, and L. I. Kuncheva, "Occlusion handling via random subspace classifiers for human detection," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 342–354, Mar. 2014.
[23] C. Zhang, J. Zhang, H. Zhao, and J. Liang, "A part-based probabilistic model for object detection with occlusion," *PLoS ONE*, vol. 9, no. 1, 2014, Art. no. e84624.
[24] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, "Learning people detectors for tracking in crowded scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1049–1056.
[25] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063.
[26] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2606–2615.
[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, vol. 8693, 2014, pp. 740–755.
[28] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 761–769.
[29] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[30] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7774–7783.
[31] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 637–653.
[32] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6995–7003.
[33] I. Goodfellow, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, and J. Pouget-Abadie, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[34] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: https://arxiv.org/abs/1412.6572
[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[36] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*, vol. 1681. Berlin, Germany: Springer-Verlag, 1999, pp. 319–345.
[37] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision* (Lecture Notes in Computer Science), vol. 8689. Cham, Switzerland: Springer, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. 2014, pp. 818–833.
[38] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Banff, AB, Canada, 2014, pp. 1–10.
[39] 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.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.
[40] Z. Yan, M. Yan, H. Sun, K. Fu, J. Hong, J. Sun, Y. Zhang, and X. Sun, "Cloud and cloud shadow detection using multilevel feature fused segmentation network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1600–1604, Oct. 2018.
[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[42] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, Mar. 2017, pp. 924–933.
[43] M. Najibi, M. Rastegari, and L. S. Davis, "G-CNN: An iterative grid based object detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2369–2377.
[44] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2874–2883.
[45] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1476–1481, Jul. 2017.
[46] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Barcelona, Spain: Curran Associates, 2016, pp. 379–387.



QINGYANG XU received the M.S. and Ph.D. degrees in control theory and engineering from Dalian Maritime University, Dalian, China, in 2007 and 2010, respectively. From August 2010 to July 2012, he was a Postdoctoral Researcher with the Dalian Institute of Chemical Physics, Chinese Academy of Sciences. He is currently an Associate Professor with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China. His research interests include artificial intelligence, deep learning, and their applications on robot.



XIAOFENG ZHANG was born in Yushu, Changchun, Jilin, China, in 1993. He received the B.S. and M.S. degrees from Shandong University, Weihai, in 2016 and 2019, respectively. He has done some works about the object detection and tracking based on the deep learning techniques. His interests include deep learning, image understanding, and intelligent robot.



RUOSHI CHENG was born in Huozhou, Shanxi, China, in 1996. He received the B.S. degree from Shandong University, Weihai, in 2018. He is currently pursuing the M.Sc. degree under the supervision of Asso. Prof. Q. Xu. He has done some works about the object detection. His interests include deep learning, image recognition, and computer stereo vision.

YONG SONG received the B.S. degree in control science from Shandong University, Weihai, in 2001, and the M.S. and Ph.D. degrees in pattern recognition and intelligent system from Shandong University, in 2008 and 2012, respectively. He is currently an Associate Professor with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai. His current research interests include intelligent robot control, machine learning, and swarm intelligence robotics.

NING WANG received the B.Eng. degree in marine engineering and the Ph.D. degree in control theory and engineering from Dalian Maritime University (DMU), Dalian, China, in 2004 and 2009, respectively. From August 2014 to August 2015, he was a Visiting Scholar with the University of Texas at San Antonio. He is currently a Full Professor with the Marine Engineering College, DMU. His research interests include fuzzy neural systems, deep learning, nonlinear control, self-organizing fuzzy neural modeling and control, unmanned vehicles, and autonomous control. He has authored two books, two book chapters, and more than 150 refereed journal and conference papers in his research areas of interest. Dr. Wang was financially supported by China Scholarship Council (CSC) to work as a Joint Training Ph.D. Student with the Nanyang Technological University (NTU), Singapore, from September 2008 to September 2009. In light of his significant research at NTU, he received the Excellent Government-funded Scholars and Students Award, in 2009. He received the Nomination Award of Liaoning Province Excellent Doctoral Dissertation; the DMU Excellent Doctoral Dissertation Award and the DMU Outstanding Ph.D. Student Award, in 2010, respectively. He also received the Liaoning Province Award for Technological Invention (First Class) and the honor of Youth Leading Talents for Transportation Science and Technology Innovation Talents Promotion Plan, Liaoning BaiQianWan Talents (First Level), Liaoning Excellent Talents, Science and Technology Talents the Ministry of Transport of China, Youth Science and Technology Award of China Institute of Navigation, Dalian Distinguished Young Scholars, and Dalian Leading Talents. He serves as associate editors of the *Neurocomputing*, the *International Journal of Fuzzy Systems* and the *International Journal of Intelligent Autonomous Systems*, and a Guest Editor of the *Advances in Mechanical Engineering*.