IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 15, 2020, accepted February 3, 2020, date of publication February 7, 2020, date of current version February 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972338

A Cascaded R-CNN With Multiscale Attention and Imbalanced Samples for Traffic Sign Detection

JIANMING ZHANG, (Member, IEEE), ZHIPENG XIE, JUAN SUN, XIN ZOU, AND JIN WANG, (Senior Member, IEEE)
School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China

Corresponding author: Jin Wang (jinwang@csust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972056 and Grant 61772454, in part by the Natural Science Foundation of Hunan Province of China under Grant 2019JJ50666, in part by the Double First-Class International Cooperation and Development Scientific Research Project of Changsha University of Science and Technology under Grant 2019IC34, in part by the Postgraduate Training Innovation Base Construction Project of Hunan Province under Grant 2019-248-51, and in part by the Practical Innovation and Entrepreneurship Ability Improvement Plan for Professional Degree Postgraduate of Changsha University of Science and Technology under Grant SJCX201972.

ABSTRACT In recent years, the deep learning is applied to the field of traffic sign detection methods which achieves excellent performance. However, there are two main challenges in traffic sign detection to be solve urgently. For one thing, some traffic signs of small size are more difficult to detect than those of large size so that the small traffic signs are undetected. For another, some false signs are always detected because of interferences caused by the illumination variation, bad weather and some signs similar to the true traffic signs. Therefore, to solve the undetection and false detection, we first propose a cascaded R-CNN to obtain the multiscale features in pyramids. Each layer of the cascaded network except the first layer fuses the output bounding box of the previous one layer for joint training. This method contributes to the traffic sign detection. Then, we propose a multiscale attention method to obtain the weighted multiscale features by dot-product and softmax, which is summed to fine the features to highlight the traffic sign features and improve the accuracy of the traffic sign detection. Finally, we increase the number of difficult negative samples for dataset balance and data augmentation in the training to relieve the interference by complex environment and similar false traffic signs. The data augment method expands the German traffic sign training dataset by simulation of complex environment changes. We conduct numerous experiments to verify the effectiveness of our proposed algorithm. The accuracy and recall rate of our method are 98.7% and 90.5% in GTSDB, 99.7% and 83.62% in CCTSDB and 98.9% and 85.6% in Lisa dataset respectively.

INDEX TERMS Traffic sign detection, convolutional neural network, attention, object detection, Multiscale.

I. INTRODUCTION

A. BACKGROUND

Traffic signs detection is not only an important part of automatic driving and assisted driving but also a crucial function of Cooperative Intelligent Transport Systems (CITS). The drivers can receive the information obtained by automatic traffic sign detection in time to regulate the behavior of drivers and enhance the security and comfort of motor vehicle driving. Meanwhile, automatic traffic sign detection is the basis of theory and practical application of automatic driving, traffic sign management and robot. It has a prospective application in the future. Whereas, the real-world traffic scenes are complicated as a result of illumination variation, bad weather and similar false signs etc. Therefore, the research on the traffic sign detection still faces great challenges. The key to achieving the robustness and accuracy of traffic sign detection is to detect the traffic signs of small size in the complex environment. Wide research has been done in the traffic sign detection. The color and shape based traffic sign detection algorithms are proposed [1]–[3] because the shape of traffic signs are triangles, circles and rectangles with bright color, which extract the color and shape information to output the features extracted from the region of

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

interest (ROI) containing the traffic signs. Deep learning is prevailing in the field of traffic sign detection recently. The algorithms [3]–[8] based on the improved network structure increase the detection accuracy aimed at traffic sign with small size. The methods [9], [10]based on the image segmentation to detect the traffic signs under complex environment. Some attention-based detection methods [11], [12] gain the ROI from the input image by attention module to fine the features in a large sophisticated background. These two ways enhance the performance of small traffic sign detection and reduce the cases of false detection. However, because the deployment of deep neural network in mobile platforms is time-consuming and has large computation, it is challenging to detect the traffic signs in mobile devices. Therefore, methods [13], [14] based on light-weight network designed by model compression to reduce the calculation of network parameters, achieving the real-time traffic sign detection.

B. MOTIVATION

Features of images can be gradually learned by deep learning. The performance of detection is greatly improved to extract deep features to recognize the traffic signs. References [15]–[17] exploit deep learning methods to detect the traffic signs with improved performance. Traffic signs are of small size which is more difficult to detect than traffic signs with large size. The general traffic sign detection network, containing top-down convolution, uses the feature map from the last layer to predict the signs which result in declining performance of small traffic sign detection. Hence, we propose a multiscale cascaded R-CNN to obtain the multiscale features in pyramids that improve the performance of traffic sign detection.

The real-world traffic environment is complex generally. The performance of traffic detection is interfered by the complex environment and similar traffic signs. References [9], [10] use image segmentation to solve the problem. It is proved that such kind of methods is helpful to traffic sign detection in the complex environment. However, it is hard to use image segmentation in practical application on account of high computational cost. Attention-based detection methods [11], [12] are proposed to reduce the false detection cases arisen from similar false traffic signs. The network structures named Transformer with Multi-Head Attention [19] are designed for the machine translation in Natural Language Processing Transformer network structure discards the recursive unit and Multi-Head Attention structure explores the relationship between input and output in parallel to increase the computational speed. Motivated by the Multi-Head Attention[19], we proposed multiscale attention used after each output of multiscale and cascaded object detection network in order to reduce the false sign detection cases in sophisticated environment.

Nowadays, traffic sign detection methods based on the deep learning usually exploit random sampling to gain the positive and negative samples. A research [20] shows that the overlap of more than 60% hard negative samples is greater than 0.05, however, the random sampling only provides 30% hard negative samples whose overlap is greater than 0.05 as training samples. Many hard samples are lost so that the accuracy is decreased. Figure 1 shows IoU distribution of random selected samples and hard negatives. Then, we increase the number of hard samples to balance the distribution of positive samples and negative samples. Meanwhile, to reduce the undetection and false detection in complex environment, we exploit 15 data augmentation methods such as add noise, blur image to simulate the real-world environment.

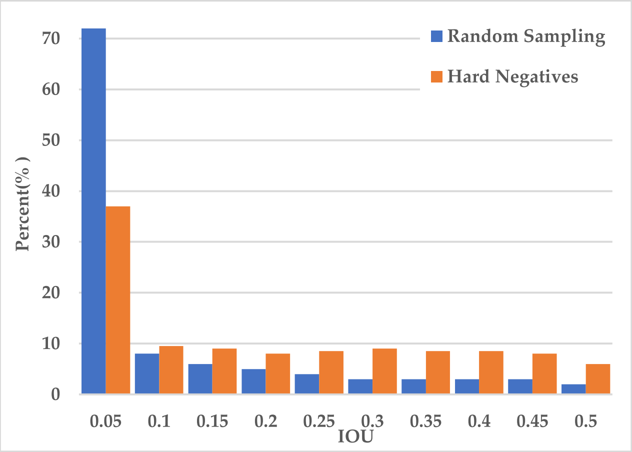


FIGURE 1. IoU distribution of random selected samples, and hard negatives.

C. MAIN CONTRIBUTIONS

In this paper, we propose a multiscale cascaded R-CNN and multiscale features in pyramids which fuses high level semantic information and low level spatial information. The cascaded network is trained by positive samples which adjusts the bounding boxes with higher overlap. Moreover, we put forward a multiscale attention to enhance the ability to detect the true traffic signs other than false signs. The features extracted from the ROI for each scale are fined and then fused to improve the accuracy of traffic sign detection. The contributions of our paper are as follows:

(1) We propose a multiscale cascaded object detection network and introduce multiscale features in pyramids to obtain feature of each scale. Each estimated position by skip connection of every feature is trained to obtain features of other scales, which relieves the overfitting. Finally, the high level semantic information and low level spatial information extracted from the multiscale object detection network are fused.

(2) We propose a multiscale attention method to do dot-product and softmax by multiscale features itself to gain the weights, which is summed to fine the features to highlight the traffic sign features and better detect the traffic signs in complex background.

(3) The similar objects are generally confounded with true objects in complex environment. Aimed at the balancing the

distribution of traffic sign categories, we increase the number of hard negative samples in the training stage. The German Traffic Sign Dataset Benchmark (GTSDB) is expanded by 15 data augmentation methods such as add noise, blur image. We test our proposed method in GTSDB, Chinese Traffic Sign Detection Benchmark (CCTSDB) and US Traffic Sign Dataset (Lisa dataset). The experimental results prove the effectiveness of our contributions. The accuracy and recall rate of our method are 98.7% and 90.5% in GTSDB, 99.7% and 83.62% in CCTSDB and 98.9% and 85.6% in Lisa dataset respectively.

II. RELATED WORK

In this section, we review the traffic sign detection algorithms related to our methods. The traditional traffic sign detection algorithms, the CNN-based traffic sign detection algorithms, generic object detection and attention in image processing are included in this section.

A. TRADITIONAL TRAFFIC SIGN DETECTION ALGORITHMS

Traditional traffic sign detection algorithms [21]–[23] use physical characteristics of the detected traffic signs including color and shape based methods. Color based traffic sign detection methods are usually following three steps: 1) Extract points with specific color from images; 2) the points are connected to surround the biggest region; 3) obtain the features in the ROI containing traffic signs. Contrast-limited adaptive histogram equalization (CLAHE) in Reference [24] is employed to normalize colors when inputting the images. Reference [25] raises a new approach based SVM, which can use color and shape to extract candidate regions of traffic signs, achieving high performance.

B. CNN-BASED TRAFFIC SIGN DETECTION ALGORITHMS

The deep learning based traffic sign detection develops faster and faster at the same time because of the excellent performance. Traffic sign is a specific object from general objects. The deep learning based traffic sign detection methods often improve existing generic object detection network. These methods [26]–[29] use deep learning to process image. Improved Faster R-CNN [30] uses Attention Network (AN) rather than Region Proposal Network (RPN). Its model achieves accuracy of 96.05% and 98.31% in TT100K and BTSD respectively. The improved YOLO V2 [30] achieves accuracy of 99.61% and 96.69% in CCTSDB and GTSDB. Yu *et al.* [4] designs a visual co-saliency to detect the traffic sign, which includes two stage: in first stage, co-saliency model based on cluster is adopt to procedure the final co-saliency map, and then construct geometric structure constraint model to recognize objects which are prominent. Luo *et al.* [5] puts forward a new data drive system and a multitask CNN to cope with the ROI refinement and classification aimed at challenging text-based traffic sign detection. Meanwhile, it exploits the compositive traffic signs and labelled images from street views aimed at the small number of labelled traffic sign detections. A novel traffic sign detection system [10] is proposed to predict the position and precise boundary by CNN. The end-to-end boundary estimation is transformed into pose estimation problem which is more robust to occlusion and small objects than contour estimation and image segmentation. It achieves speed at 7 fps in mobile devices. Potential ROI is determined by attention module [11]. The deconvolution added in the convolutional layers to adapt objects of small size, which deal with the small object recognition. As a result of the diversity of the text in traffic signs, the obvious size changes and illumination variation, the real-time requirement and high accuracy are hard to meet. A traffic sign detection framework based text is proposed in Reference [32], which apply a fully convolutional network to segment candidate areas of traffic sign. The text ROI of candidate areas is detected by CNN. The problem of multiscale for the text detection part by narrowing the text detection area.

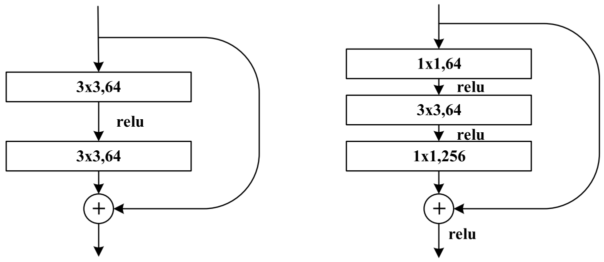


FIGURE 2. The unit of residual network. We use the unit of residual network as our backbone network.

C. GENERIC OBJECT DETECTION

Ross *et al.* [33] propose RCNN network using selective search to obtain the region proposals. It combines region proposals and CNN to localize the object. Residual network [34], whose units are shown in Figure.2, introduces shortcut to increase identity mapping to relieve the model degradation. It has been applied to many mainstream fields and often used as a general object detection backbone network. Ren *et al.* [35] put forward a Faster R-CNN based on Region Proposal Network (RPN) to generate region proposals, extract features, classification and precise object localization. It improves the speed and accuracy greatly. Lin *et al.* [36] propose the feature pyramid network to deal with the multiscale object detection. Feature pyramids with marginal extra cost is constructed by inherent multiscale, pyramidal hierarchy of CNN and a top-down network structure having lateral connections. All scales of high-level feature maps containing semantic information are constructed as a generic feature extractor, which presents great improvements. Mask R-CNN [37] extends based on Faster RCNN with a branch to predict an object mask with existing branch in parallel. Mask RCNN achieves effective object detection and high-quality semantic segmentation at the same time. Cascade R-CNN [38] uses different IoU thresholds to train several cascaded detector to address the problem of noisy interference and imprecise object detection.

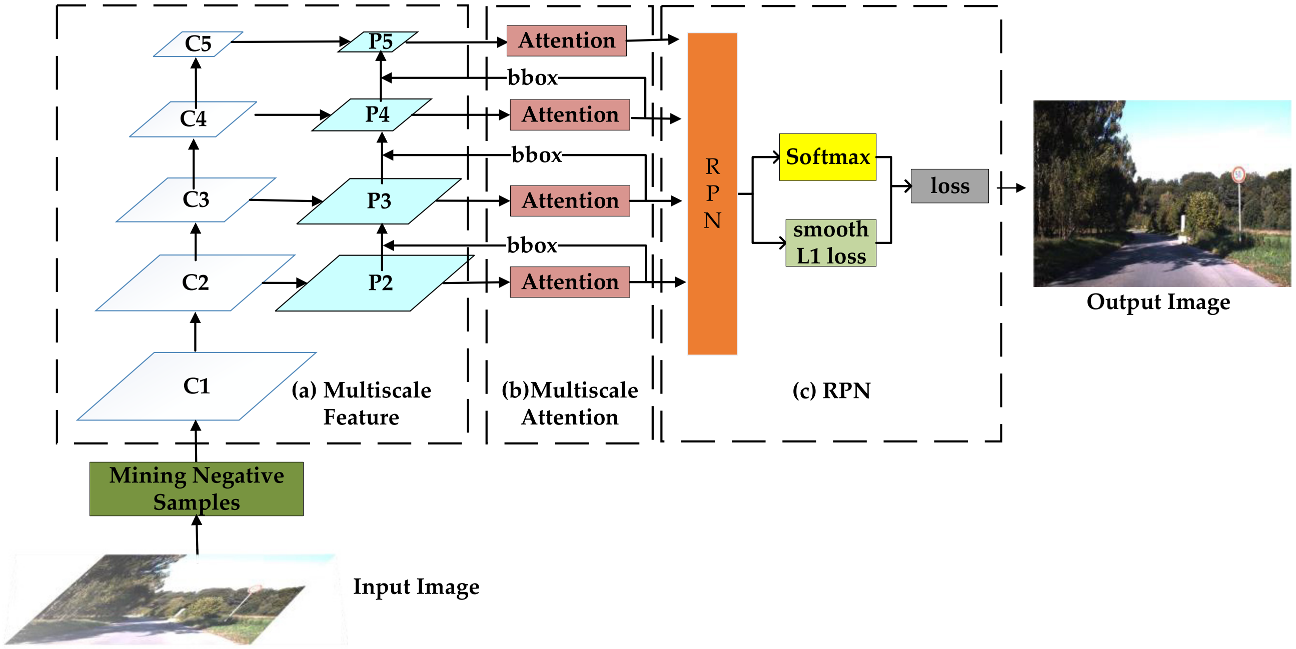


FIGURE 3. Overview of our proposed approach: cascaded R-CNN with multiscale attention and imbalanced samples.

D. ATTENTION IN IMAGE PROCESSING

Attention-based methods can emphasize the important information and neglect inconsequential information. Mnih et al. [39] propose a recurrent attention model based on RNN for the first time. It selects a sequence of regions or locations adaptively and only process the regions and extract features at a high resolution. Spatial transformer networks [40] is proposed. It considers pooling in convolutional network as a violent way so that the direct combination of information will cause the ignorance of key information. The spatial transformation of spatial information in images is helpful to extract key information. Residual attention network [41] uses not only spatial attention but also channel attention, which are formed meanwhile by weight of each element considered as a mask. Reference [42] improves the network representation ability by recalibrating channel-wise feature, and the weights of features are learned by loss function. Inspired by the traditional non-local means, Ref. [43] captures long-range dependencies by non-local operations. This non-local operation is used to compute the feature weight of each position, then feature weights of all position are summed. Woo *et al.* [44] combined spatial attention [40] and channel self-attention [43]. It directly uses non-local Matmul to avoid handcraft pooling and multilayer perceptron. A simplified query-independent method is proposed to gain global context block [45]. It reduces the computation caused by the same context information at every position.

III. OUR APPROACH

A. OVERVIEW

The overview of our proposed approach is shown Figure.3. Traffic signs has obvious features like color and shape. Traditional machine learning methods extract above features fast. Traditional methods with deep neural network raise the accuracy of localization and recognition rate. But feature map extracted from the last layer of traditional object detection networks is utilized to predict so that the performance of small object detection drops sharply. At present, two stages detection network extracts region proposals containing objects. The region proposals are then input to CNN to judge whether the region proposals contain the object or not. However, in the real-world traffic scenes, traffic signs are usually of small size and there is no context information, therefore, it is difficult to distinguish the false traffic signs similar to the real traffic sign objects. The multiscale cascaded object detection network is shown in part (a) in Figure.3. Part (a) uses the structure of ResNet50 as its backbone network. The feature maps named {C1, C2, C3, C4, C5} in five scales, are extracted by top-down convolution to relieve the degradation with the increasing depth of convolutional layers. In ResNet50, C1 represents the feature output by the conv1x layer, C2 represents the feature output by the res2bx layer, C3 represents the feature output by the res3ax layer, C4 represents the feature output by the res4b22x layer, and C5 represents the feature output by the res5cx layer. Each feature map in five scales is then processed by the attention model in Part (b). The outputs of attention models are all input to Part (a) back layer by layer. More specifically, the output of C_{i-1}, P_{i-1} and bbox generated containing position information processed by P_{i-1} after attention, are all used to train the parameters of P_i. The feature maps named {P2, P3, P4, P5}, generated by feature pyramid network, have spatial information in high resolution and semantic information in low resolution. The final feature map is fused by the previous feature maps, then it is passed through RPN and the detection result is obtained in the end.

B. MINING HARD NEGATIVE SAMPLES

General object detection methods [35], [37], [38] use random sampling to mine negative and positive samples. The hard-to-detect negative samples is defined as negative samples that has false objects similar to true objects and is prone to be determined as positive samples. We conduct a research on distribution of hard negative samples. It is found that IoU of more than 60% hard negative samples is greater than 0.05. However, only 30% training samples is hard samples using random sampling. The low proportion of hard samples causes false detection easily and decreases the model accuracy. Therefore, we mine the hard negative samples and increase the number of these samples according to the distribution of training samples. According to the distribution of IoU, the interval is divided equally. The hard negative samples are divided equally in each bin. Correspondingly, the quantity of hard negative samples whose IoU is greater than 0.05 is increased. In the stage of training, the quantity of hard negative samples is known to us. The method of assuming an average K divided intervals, hard negative samples is mined as follows:

$$P = \frac{NS}{\sum_{k=1}^{K} N_k} \tag{1}$$

where P denotes the probability of selected hard negative samples, NS represents the total quantity of negative samples. $\sum_{k=1}^{K} N_k$ is the total number of samples. Hard negative samples is mined for each interval as follows:

$$P_k = \frac{NS}{K} * \frac{1}{N_k} \tag{2}$$

where P_k denotes the probability of selected hard negative samples for each bin, K denotes the number of bins. NS represents the quantity of negative samples. N_k is the number of samples in the k-th bin. Hence, the amount of hard negative samples in each bin is same. In our practical experiment, K is set to 4 and the interval is divided as four bins: [0, 0.125], [0.125, 0.25], [0.25, 0.375], [0.375, 0.5].

C. MULTISCALE CASCADED R-CNN

Owing to the small size of traffic signs, it is harder to detect than generic object with large size. Therefore, we propose the multiscale cascaded R-CNN. The high level semantic information and low level spatial information extracted from the multiscale cascaded R-CNN are fused. Each layer of the cascade network except the first layer fuses the out- put bounding box of the previous one layer for joint training. The optimization object of our multiscale cascaded R-CNN is described as:

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{loc}(t^u, v) \tag{3}$$

where L_{cls} is the classification loss using the softmax function. p and u represent the predictions and targets respectively. L_{loc} is the localization loss. t^u represents the corresponding regression results of the u-th category. v is regression target. λ is the learning rate. The classification loss L_{loc} is described as:

$$L_{loc} = \sum\nolimits_{i\in\{x,y,w,h\}} (t_i^u - v_i) \tag{4}$$

where t_i^u is the regression result. v is regression target. i is the input.

We adopt the bounding box regression method proposed in [38], using a regressor $f(x, b)$ to regress the candidate bounding b into a target bounding box g, and L_{loc} operation the defined distance vector $\Delta = (\delta x, \partial y, \partial w, \partial h)$ as follows:

$$\begin{aligned} \partial_x &= (g_x - b_x)/b_w, \quad \partial_y = (g_y - b_y)/b_h, \\ \partial_w &= log(g_w/b_w), \quad \partial_h = log(g_h/b_h). \end{aligned} \tag{5}$$

where x, y, w and h present the center coordinates of the box as well as width and height, b is the candidate bounding box, and g represents the target bounding box. The L_{loc} uses $smooth_{L_1}$ loss [30] as follows:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if\,|x| < 1 \\ |x| - 0.5, & othervise, \end{cases} \tag{6}$$

$smooth_{L_1}$ loss converges faster than L_1 loss function and is more insensitive to outliers than L_2 loss function. Moreover, gradient changes more slightly and it also does not vanish in the stage of training.

D. MULTI SCALE ATTENTION

To improve the detection accuracy in the real-world complex scenes and enhance the ability to find the hard negative samples similar to true traffic signs, we propose a multiscale attention method shown in Figure 3. This method conducts dot-product and softmax by multiscale features itself to gain the weights, which is summed to fine the features to highlight the traffic sign features and improve the detection accuracy. The final feature map is fused by the each layer after attention processing. Multiscale attention is define as:

$$Attention(Q, K, V) = softmax(\frac{f(Q^T, K)}{\sum_{j=1}^{m} e^{f(Q,K_j)}})V \tag{7}$$

where Q, K, V are same multiscale feature maps by feature pyramids, $Q = K = V$. m represents the number of multiscale feature maps. First, the dot-product is conducted between Q^T and K to obtain their similarity measurement as follows:

$$f(Q^T, K_j), \quad j = 1, 2, \ldots, m \tag{8}$$

where $f(.)$ is the dot-product operation. j is the j-th feature map. The normalization is achieved as:

$$s_j = softmax(\frac{f(Q^T, K)}{\sum_{j=1}^{m} e^{f(Q,K_j)}}), \quad j = 1, 2, \ldots, m \tag{9}$$

Finally, the weights are summed to gain the vector of multiscale attention:

$$\sum\nolimits_{i=1}^{m} s_i V_i, \quad m = 4 \tag{10}$$

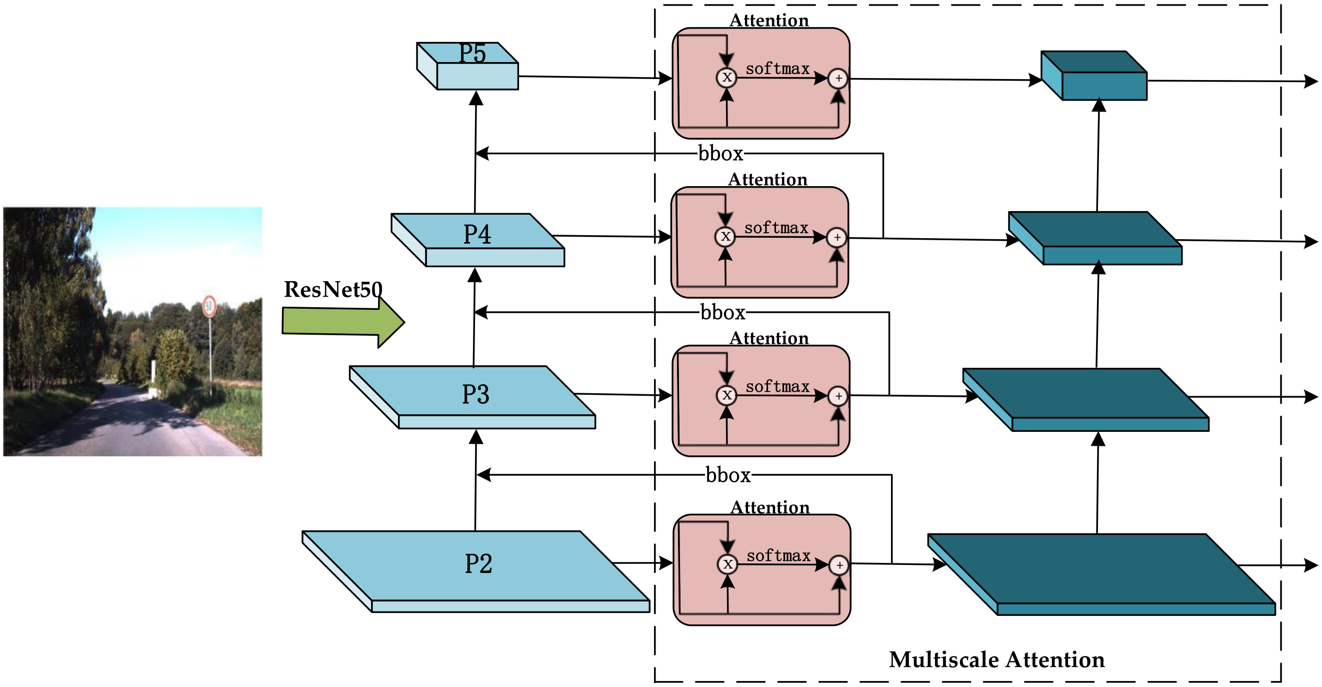


FIGURE 4. Multiscale attention.

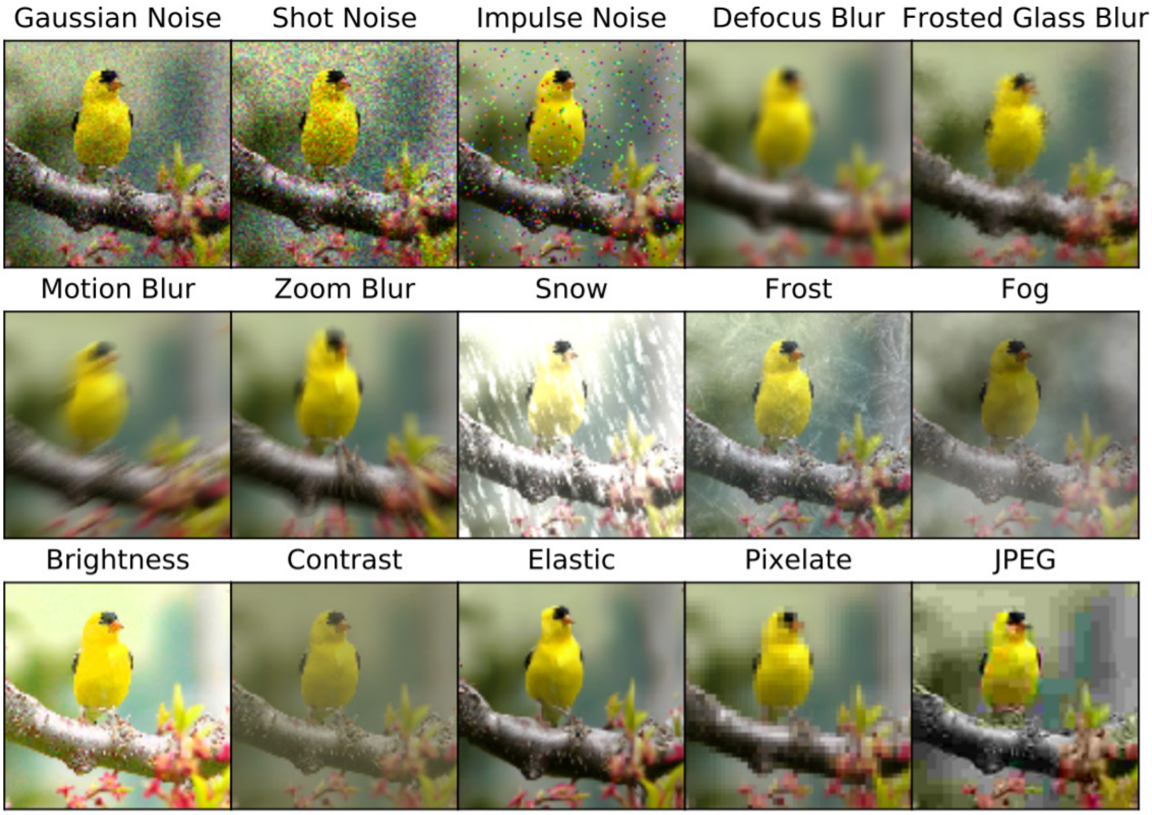


FIGURE 5. The IMAGENET-C dataset [35] contains noise, blur, weather, and digital category damage generated by 15 algorithms.

E. DATA AUGMENTATION

In the real-world traffic environment, there are often many complex environmental factors. Dan and Dietterich *et al.* [46] enhance the robustness of corruption and disturbance of image data in view of the fact that deep learning network is disturbed by many forms of image corruption, such as snow,

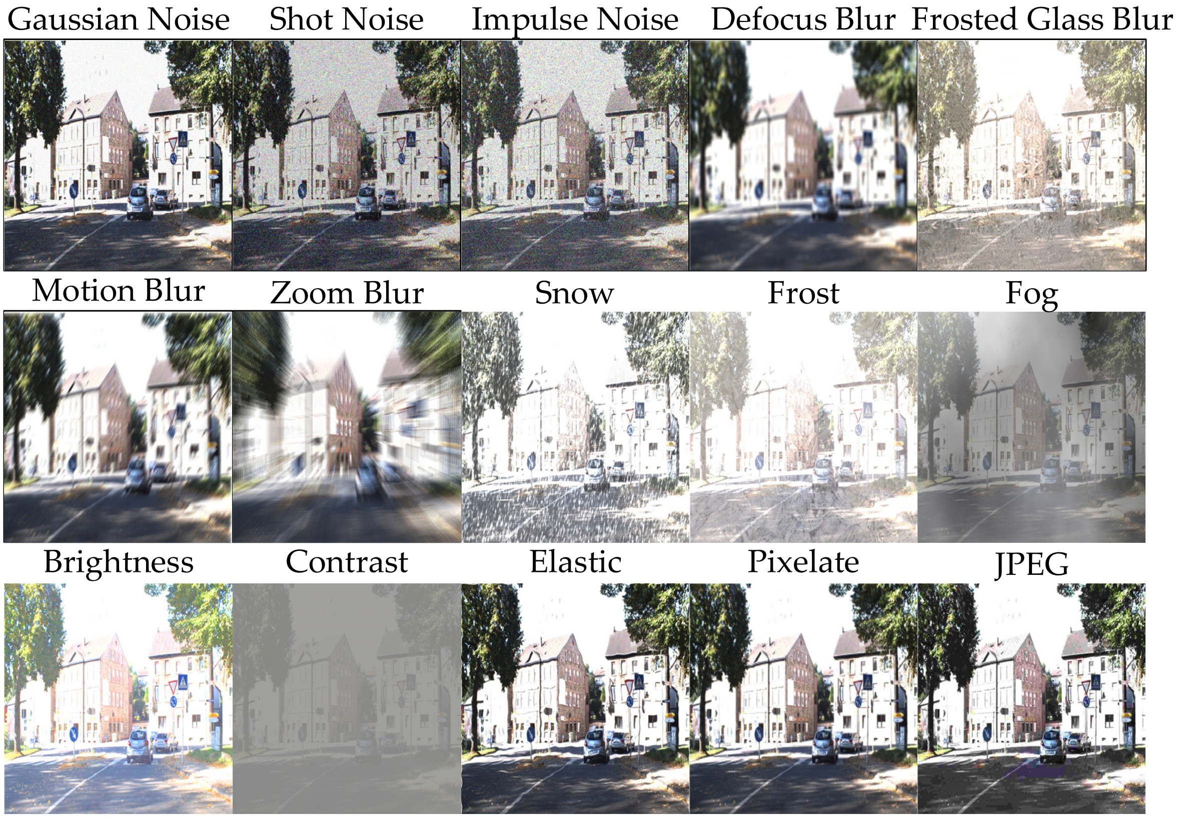


FIGURE 6. Randomly select an image from the GTSDB dataset, and use 15 data enhancement methods to generate the image.

blur, pixelation, etc. Corruption types include noise, blur, weather and data category damage generated by 15 algorithms. Each type of corruption can be categorized into five severity degrees, as shown in Figure 5. There are three kinds of noise corruption: Gaussian noise, shot noise and impulse noise. Four kinds of blurring corruption are defocus blur, frosted glass blur, motion blur and zoom blur. Five kinds of weather corruption are snow, frost, fog, brightness and contrast. Two kinds of data corruption are pixelate and JPEG corruption. In the face of the complex environment in traffic sign detection, we follow the same method proposed in [32] to enhance the robustness of corruption and disturbance, and enhance the GTSDB training dataset for real-world traffic sign environment, including 15 kinds of data enhancement strategies such as adding noise, blur, digital and weather to the image. Because each type of corruption has five severity degrees, we select two kinds of severity to expand the training dataset considering the environmental conditions and the size of traffic signs dataset. Eventually, we expand the GTSDB training dataset to be 18600 images. After data augmentation, the enhancement results of one picture in the GTSDB is shown in Figure 6.

IV. EXPERIMENTS

Experimental evaluations of proposed algorithm are summarized in this section. First, we use Python deep learning open-source framework to implement our traffic sign detection method. Then, our algorithm are evaluated on GTSDB [30], CCTSDB [31] and Lisa [3] datasets.

A. EXPERIMENTAL SETUP

Implementation Details. Our experiment is carried out on a computer-based on the python environment, which uses Intel (R) Xeon (R) CPU e5-2640 2.40 GHz CPU and NVIDIA GeForce GTX 1080 Ti GPU (11 g memory). Using the mmdetection toolbox to extract deep features from Resnet-50. The parameters of our algorithm are set as follows: the learning rate is set to 0.02, and the weight attenuation is set to 0.0001. Momentum is set to 0.9. In all experiments, the same super parameters are used to train 12 batches. The 8th and 11th batches of our method reduced the learning rate to 1/3 of the original. We training on GTSDB, CCTSDB and US traffic sign datasets, two GPUs are used for training.

B. EFFECTIVENESS ANALYSIS OF OUR APPROACH

1) DATASETS

We conduct comparative experiments on the German, Chinese and US traffic sign datasets to fully verify the effectiveness of proposed algorithm. The GTSDB dataset, as shown in Figure 7, contains 600 training images and 300 test images for, and resolution of each image is 1360 × 800. The sizes of traffic signs range from 16 to 128 pixels. We expand the training data to 18600 images by data augmentation, such as noise, blur, digital, and weather.

FIGURE 7. Some examples of GTSDB dataset.

FIGURE 8. A few examples from the CSUST dataset in China (CCTSDB).

CCTSDB have nearly 20,000 images and nearly 40,000 traffic signs. Several examples selected from CCTSDB is shown in Figure 8. Amplifying the dataset with limited image transformation. Image transformation is based on existing images, such as affine transformation, noise processing, translation transformation, rotation transformation, scaling processing, scaling transformation, brightness processing, etc. To simulate the real situation. Low pixels, distorted logos, different lighting, rain, and snow, etc. Lisa Dataset contains American traffic signs videos and video frames are annotated. These include 47 types of US traffic signs, with a total of 7,855 images and 6610 signs. Traffic sign sizes range from 6x6 to 167x168 pixels. As shown in Figure 9. We divide traffic signs into three categories according to [24], including warning, speed limit, and noTurn.

2) EVALUATION METRICS

We use accuracy [47], recall [47], loss rate and F-measure [47] to evaluate the proposed algorithm. More specifically, t denotes the total number of traffic signs, t_p is true positive (TP) which indicates the detections of traffic signs are correct. f_p is false positive (FP) which indicates the detections of traffic sign are wrong. f_n is false negative (FN) and it represents the number of lost traffic signs. β is set to 1. Therefore, the metrics are evaluated as follows:

$$Precision = \frac{t_p}{t_p + f_p} \tag{11}$$

FIGURE 9. Some examples of Lisa dataset.

$$Recall\_rate = \frac{t_p}{t_p + f_n} \tag{12}$$

$$Missing\_rate = 1 - \frac{t_p}{t_p + f_n} \tag{13}$$

$$F-meature = (1+\beta^2)\frac{Precision * Recall\_rate}{\beta^2 * Precision + Recall\_rate} \tag{14}$$

3) PERFORMANCES ON GTSDB

To test the effectiveness of our method, we implement our model on GTSDB dataset. The GTSDB dataset contains 600 training images and 300 test images, and the resolution of each image is 800 x 1360. There are 42 types of traffic signs, which can be divided into three categories: prohibition, mandatory, and danger. These categories on training dataset contain 59.5%, 17.1%, and 23.4% traffic signs, respectively. There are often many complicated environmental factors to disturb in the real road traffic environment. Meanwhile, we use the data augmentation method to augment the training dataset because the amount of image in GTSDB is small. We use 15 data enhancement strategies including noise, blur, digital, and weather, etc. to augment training dataset from 600 images to 18600, and train our model on this training dataset. Faster R-CNN [35], Mask R-CNN [37] Cascade R-CNN [38] have provided code tests to obtain detection results. Also, the overall average accuracies and recalls of true positive, false positive, and false-negative samples of three categories are computed. The results of our method and other seven algorithms are presented in Table 1. Our model obtains good results with accuracy and recall of 98.7% and 90.5%, respectively. Except for the indicators given in Table 1, the average execution time of our model is 7.6 fps,while the average execution time in Faster R-CNN, Mask R-CNN, and Cascade R-CNN is 2.26 fps, 5.7 fps, 7 fps. The time consumption of our proposed method is near to the original method. In particular, our proposed models are performed in different hardware and software platform, so these models cannot be compared directly. According to the calculation results of the three categories, our model performs better than the Cascade R-CNN network. Therefore, the results on the GTSDB dataset prove the scalability of our method.

TABLE 1. Detection result on GTSDB.

Method	Precision(%)	Recall_rate(%)	Missing_rate(%)	F-measure (%)
Faster R-CNN [35]	96.10%	86.30%	13.70%	90.94%
Mask R-CNN [37]	97.10%	86.90%	13.10%	91.72%
Cascaded R-CNN [38]	96.80%	88.60%	11.40%	92.52%
Xu X[1]	93.96%	95.27%	4.73%	94.61%
Kamal U [9]	95.29%	89.01%	10.99%	92.04%
Li J [16]	84.5%	97.81%	2.19%	90.67%
Li C [17]	90.7%	81.7%	18.3%	86%
Ours w/o data augmentation	98.30%	85.00%	15.00%	91.17%
Ours	**98.70%**	90.50%	9.50%	94.42%

TABLE 2. Effects of each component in our Method. Results are reported on GTSDB.

Difficult sample mining	Multiscale Attention	Data augmentation	Precision (%)	Recall_rate (%)	Missing_rate (%)	F-measure (%)
			96.80%	82.10%	17.90%	92.52%
✓			97.10%	83.80%	16.20%	90.89%
✓	✓		98.3%	85.0%	15.0%	91.17%
✓	✓	✓	**98.70%**	**90.50%**	**9.50%**	**94.42%**

TABLE 3. Detection result on CCTSDB.

Method	Precision	Recall_rate (%)	Missing_rate (%)	F-measure (%)
Li C [17]	86.7%	75.6%	24.4%	80.08%
HOG+SVM[30]	82.2%	62.9%	37.1%	71.3%
RBD[48]	76.9%	54.1%	45.9%	63.5%
SRM[50]	81.2%	60.0%	40.0%	69.0%
SMD[49]	80%	53.3%	46.7%	64%
CCNN[32]	81.7%	63.8%	36.2%	71.6%
Ours	**99.7%**	**83.62%**	**16.38%**	**90.82%**

TABLE 4. Effects of each component in our Method. Results are reported on Lisa.

Difficult sample mining	Multiscale Attention	Precision (%)	Recall_rate (%)	Missing_rate (%)	F-measure (%)
		97.10%	84.70%	15.30%	90.48%
✓		97.50%	84.80%	15.20%	90.59%
✓	✓	**98.90%**	**85.60%**	**14.40%**	**91.77%**

4) ABLATION EXPERIMENTS ON GTSDB

We report the overall ablation study in Table 2. On the Cascade R-CNN baseline, the mining negative samples, multiscale attention, and data augmentation are gradually added. The experiments of the Ablation study are implemented using the same pre-calculation scheme for a fair comparison. The bounding box results of traffic signs on GTSDB are shown in Figure 12.

a: MINING HARD NEGATIVE SAMPLES

Mining negative samples is 0.3% and 0.9% higher than the Cascade R-CNN baseline in accuracy and recall rate, respectively. The results prove the effectiveness of our mining hard negative samples.

b: MULTISCALE ATTENTION

Multiscale attention based on the mining hard negative samples improves the average accuracy from 97.1% to 98.3% and recall rate from 83.8% to 85%, and the average accuracy and the recall rate are increased by 1.2% and 1.2% respectively. The results prove the effectiveness of our multiscale attention.

c: DATA AUGMENTATION

Data augmentation based on the multiscale attention and mining hard negative samples, the average accuracy is increased from 98.3% to 98.7%, and recall rates is increased from 85% to 90.5%. The results prove the effectiveness of data augmentation.

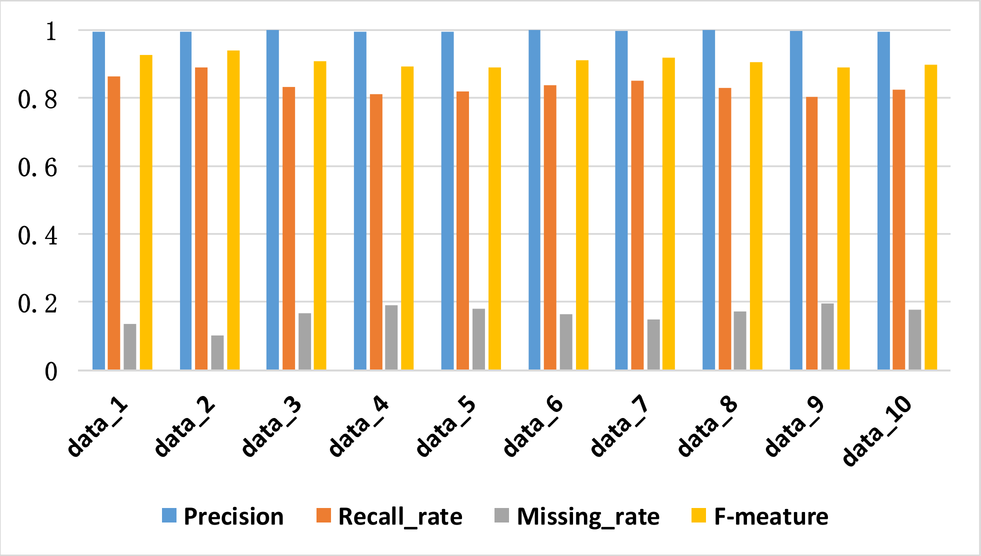


FIGURE 10. Use 10% cross-validation test results on CCTSDB dataset.

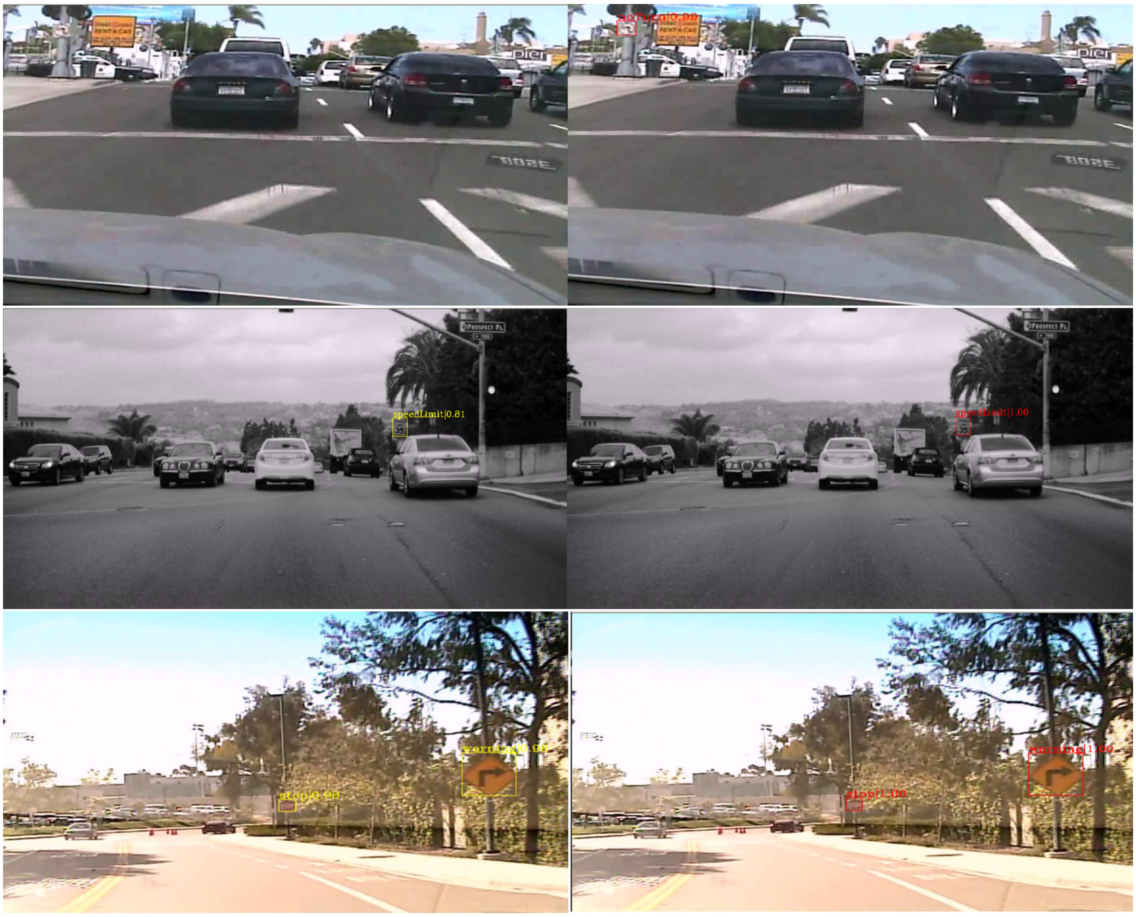


FIGURE 11. Traffic sign detection results in Lisa dataset. The detection results of cascaded R-CNN network are shown on the left, and the results of network with our proposed method are shown on the right. Especially, the target of images in the first row is undetected using the cascaded R-CNN network on the left while it is detected using our proposed method on the right.

5) PERFORMANCE OF CCTSDB

We also perform experiments on the CCTSDB. The experiments follow the settings of Ref. [17]. To better evaluate our model, ten-folder cross-validations are used. The method randomly divide 3000 image from dataset into ten parts, in which nine parts are used for training and one for test.

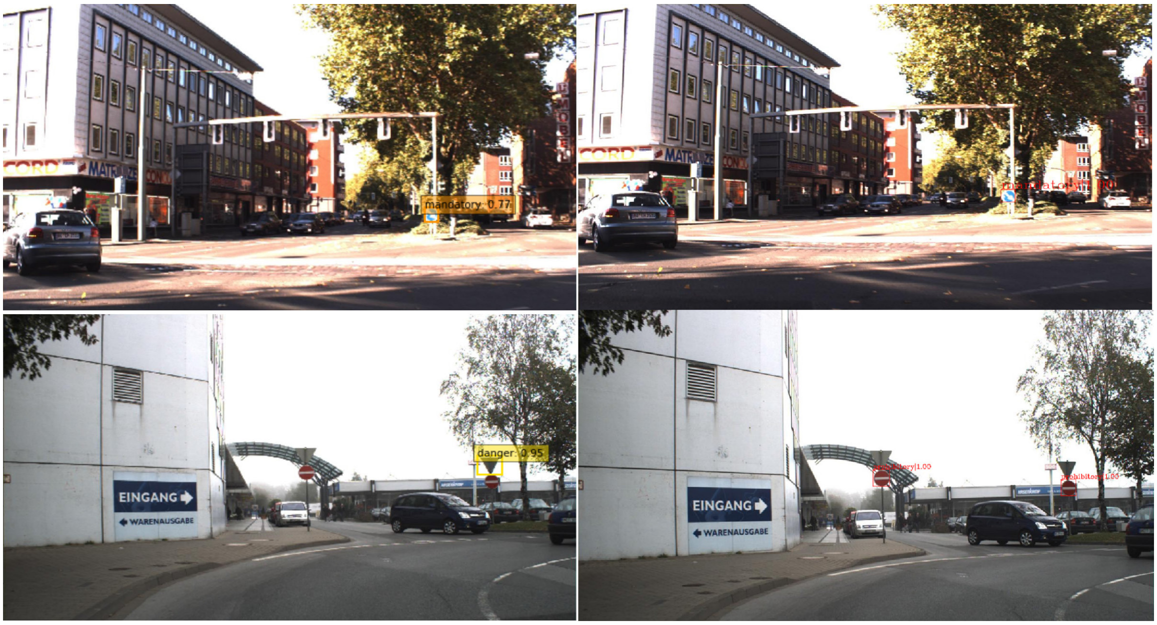


FIGURE 12. Traffic sign detection results in GTSDB. The detection results of cascaded R-CNN network are shown on the left, and the results of network with our proposed method are shown on the right. Especially, one target of images in the second row is undetected and another is false detected using the cascaded R-CNN network on the left while they are both detected on the right.

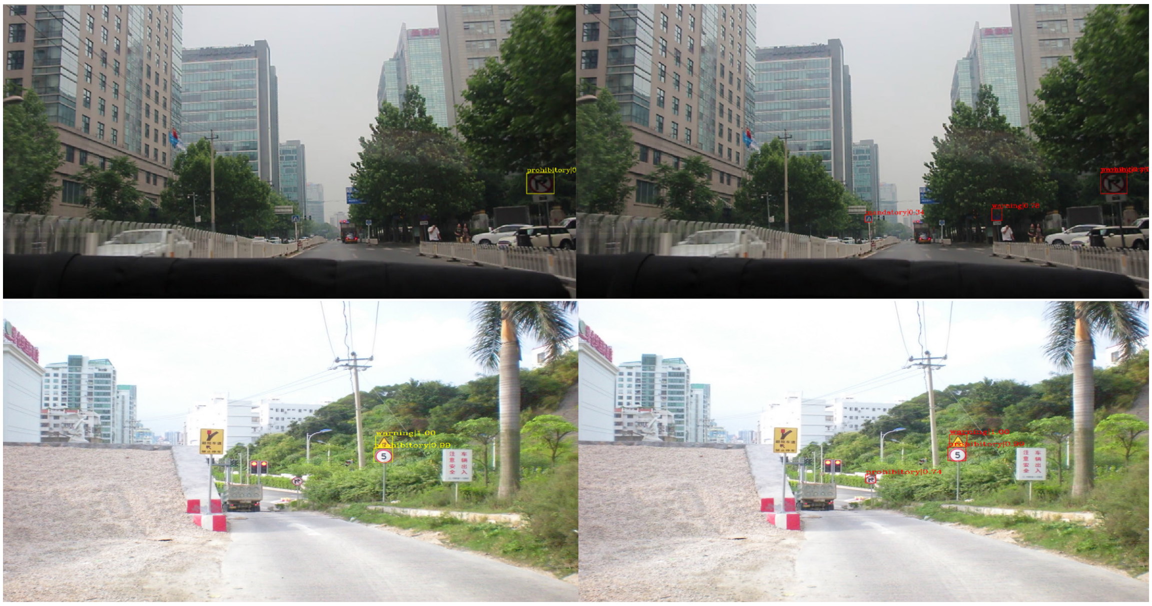


FIGURE 13. Traffic sign detection results in CCTSDB. The detection results of cascaded R-CNN network are shown on the left, and the results of network with our proposed method are shown on the right. The targets of images using the cascaded R-CNN network on the left are undetected and false detected while they are detected on the right.

The average result of 10 times is regarded as the accuracy of the proposed algorithm. The average accuracy and recall of our model are 99.7% and 83.62%. The results of our method exceed Ref. [17], HOG+SVM [30], CCNN [32], RBD [48], SMD [49], and SRM [50], the results are shown in Table 3, and the specific experimental details are shown in Figure 10. The bounding box results of traffic signs on CCTSDB are shown in Figure 13. It is worth pointing out that Ref [31] has also been tested on the CCTSDB dataset, but its experimental setting is different from [17], so we not compare with the experimental results of Ref. [31].

6) PERFORMANCE OF LISA DATASET

Our method is tested on traffic datasets in different countries, we also perform experiments on the Lisa dataset. The experiment follows the setting of Ref. [24]. We divide 47 types of traffic signs into Superclasses including warning, speed limit, stop and noTurn. We use 20% of the Lisa dataset as the test dataset, totally 1571 images. To prove the effectiveness of our module, and the ablation experiments are also implemented on Lisa dataset, we gradually add the mining difficult sample and multiscale attention to Cascade R-CNN baseline. The data volume of the US traffic sign dataset is relatively large, and using the data augmentation method will greatly increase the training time, so we do not compare the data enhancement method in this experiment. Compared with the Cascade R-CNN bsaseline, our accuracy increased by 0.4%, the recall rate increased by 0.1% when adding the mining difficult sample and multiscale attention, respectively. The accuracy increased by 1.8%, and the recall rate increased by 0.9%. The bounding box results of traffic signs on Lisa dataset are shown in Figure 11.

V. CONCLUSION

In this work, we propose an outstanding algorithm for traffic sign detection. Aiming at the problem that it difficult to detect the traffic signs of small size, a multiscale cascaded R-CNN is proposed. Then, we propose a multiscale attention method to deal with the problem that similar traffic signs in real-world traffic scenes often lead to false detection. By performing dot-product and softmax with the input multiscale features, a similarity measurement is obtained, and then the features are weighted and summed to refine the features to improve the accuracy of object detection. Finally, to alleviate the interference of environmental factors and improve the detection accuracy, we increase the number of hard negative samples during the training stage and expand the GTSDB training dataset by generating real-world pictures containing traffic signs in situations such as lighting, weather changes. A cascaded R-CNN with multiscale attention and imbalanced samples we propose has excellent performance. The effectiveness of our proposed method is proved by the extensive experimental results on the GTSDB dataset, the CCTSDB dataset, and the US traffic sign dataset.

REFERENCES

[1] X. Xu, J. Jin, S. Zhang, L. Zhang, S. Pu, and Z. Chen, "Smart data driven traffic sign detection method based on adaptive color threshold and shape symmetry," *Future Gener. Comput. Syst.*, vol. 94, pp. 381–391, May 2019.

[2] Y. Yang and F. Wu, "Real-time traffic sign detection via color probability model and integral channel features," in *Proc. Chin. Conf. Pattern Recognit. (CCPR), CCIS*, vol. 484, 2014, pp. 545–554.

[3] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1484–1497, Dec. 2012.

[4] L. Yu, X. Xia, and K. Zhou, "Traffic sign detection based on visual co-saliency in complex scenes," *Appl. Intell.*, vol. 49, no. 2, pp. 764–790, Feb. 2019.

[5] H. Luo, Y. Yang, B. Tong, F. Wu, and B. Fan, "Traffic sign recognition using a multi-task convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1100–1111, Apr. 2018.

[6] S. Hussain, M. Abualkibash, and S. Tout, "A survey of traffic sign recognition systems based on convolutional neural networks," in *Proc. IEEE Int. Conf. Electro/Inf. Technol.(EIT)*, May 2018, pp. 0570–0573.

[7] á. Arcos-García, J. A. álvarez-García, and L. M. Soria-Morillo, "Evaluation of deep neural networks for traffic sign detection systems," *Neurocomputing*, vol. 316, pp. 332–344, Nov. 2018.

[8] S. Zhou, W. Liang, J. Li, and J.-U. Kim, "Improved VGG model for road traffic sign recognition," *Comput., Mater. Continua*, vol. 57, no. 1, pp. 11–24, 2018.

[9] U. Kamal, T. I. Tonmoy, S. Das, and M. K. Hasan, "Automatic traffic sign detection and recognition using SegU-Net and a modified Tversky loss function with L1-constraint," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/tits.2019.2911727.

[10] H. S. Lee and K. Kim, "Simultaneous traffic sign detection and boundary estimation using convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1652–1663, May 2018.

[11] T. Yang, X. Long, A. K. Sangaiah, Z. Zheng, and C. Tong, "Deep detection network for real-life traffic sign in vehicular networks," *Comput. Netw.*, vol. 136, pp. 95–104, May 2018.

[12] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, Jul. 2019.

[13] J. Zhang, W. Wang, C. Lu, J. Wang, and A. K. Sangaiah, "Lightweight deep network for traffic sign classification," *Annals of Telecommunications*, to be published, doi: 10.1007/s12243-019-00731-9.

[14] S. Song, Z. Que, H. Hou, S. Du, and Y. Song, "An efficient convolutional neural network for small traffic sign detection," *J. Syst. Architect.*, vol. 97, pp. 269–277, Aug. 2019.

[15] D. Tabernik and D. Skocaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/tits.2019.2913588.

[16] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient CNNs in the wild," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 975–984, Mar. 2019.

[17] C. Li, Z. Chen, Q. J. Wu, and C. Liu, "Deep saliency detection via channel-wise hierarchical feature responses," *Neurocomputing*, vol. 322, pp. 80–92, Dec. 2018.

[18] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2110–2118.

[19] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5999–6009.

[20] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2019, pp. 821–830.

[21] J. Wang, Y. Gao, W. Liu, W. Wu, and S.-J. Lim, "An asynchronous clustering and mobile data gathering schema based on timer mechanism in wireless sensor networks," *Comput., Mater. Continua*, vol. 58, no. 3, pp. 711–725, 2019.

[22] J. Wang, C. Ju, H.-J. Kim, R. S. Sherratt, and S. Lee, "A mobile assisted coverage hole patching scheme based on particle swarm optimization for WSNs," *Cluster Comput*, vol. 22, no. S1, pp. 1787–1795, Jan. 2019.

[23] Y. Chen, J. Xiong, W. Xu, and J. Zuo, "A novel online incremental and decremental learning algorithm based on variable support vector machine," *Cluster Comput.*, vol. 22, no. S3, pp. 7435–7445, May 2019.

[24] A. Mogelmose, D. Liu, and M. M. Trivedi, "Detection of U.S. Traffic signs," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3116–3125, Sep. 2015.

[25] T. Le, S. Tran, S. Mita, and T. Nguyen, "Real time traffic sign detection using color and shape-based features," in *Proc. Asian Conf. Intell. Inf. Database Syst.*, in Lecture Notes in Computer Science, vol. 5991, 2010, pp. 268–278.

[26] J. Zhang, Y. Wu, W. Feng, and J. Wang, "Spatially attentive visual tracking using multi-model adaptive response fusion," *IEEE Access*, vol. 7, pp. 83873–83887, 2019.

[27] J. Zhang, C. Lu, X. Li, H.-J. Kim, and J. Wang, "A full convolutional network based on DenseNet for remote sensing scene classification," *Math. Biosci. Eng.*, vol. 16, no. 5, pp. 3345–3367, 2019.

[28] J. Zhang, X. Jin, J. Sun, J. Wang, and K. Li, "Dual model learning combined with multiple feature selection for accurate visual tracking," *IEEE Access*, vol. 7, pp. 43956–43969, 2019.

[29] J. Zhang, X. Jin, J. Sun, J. Wang, and A. Sangaiah, "Spatial and semantic convolutional features for robust visual object tracking," *Multimedia Tools Appl.*, to be published, doi: 10.1007/s11042-018-6562-8.

[30] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Dallas, Tx, USA, Aug. 2013, pp. 1–8.

[31] J. Zhang, M. Huang, X. Jin, and X. Li, "A real-time chinese traffic sign detection algorithm based on modified YOLOv2," *Algorithms*, vol. 10, no. 4, p. 127, Nov. 2017.

[32] Y. Zhu, M. Liao, M. Yang, and W. Liu, "Cascaded segmentation-detection networks for text-based traffic sign detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 209–219, Jan. 2018.

[33] 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.*, Portland, OR, USA, Jun. 2014, pp. 580–587.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[35] 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.

[36] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jul. 2017, pp. 936–944.

[37] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Honolulu, HI, USA, Oct. 2017, pp. 2980–2988.

[38] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6154–6162.

[39] V. Mnih, N. Hees, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2204–2212.

[40] M. Jaderberg, "Spatial transformer networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 2017–2025.

[41] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jun. 2017, pp. 7132–7141.

[43] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jun. 2017, pp. 7794–7803.

[44] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 3–19.

[45] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," Apr. 2019, *arXiv:1904.11492*. [Online]. Available: https://arxiv.org/abs/1904.11492

[46] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and surface variations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, 2018, pp. 1–13.

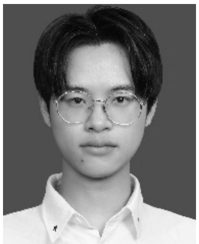
[47] X. Yuan, J. Guo, X. Hao, and H. Chen, "Traffic sign detection via graph-based ranking and segmentation algorithms," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 12, pp. 1509–1521, Dec. 2015.

[48] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.

[49] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[50] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4039–4048.

JIANMING ZHANG (Member, IEEE) received the B.S. degree from Zhejiang University, in 1996, the M.S. degree from the National University of Defense Technology, China, in 2001, and the Ph.D. degree from Hunan University, China, in 2010. He is currently a Professor with the School of Computer and Communication Engineering, Changsha University of Science and Technology, China. He has published more than 80 research articles. His current research interests include computer vision, data mining, and wireless ad hoc and sensor networks. He is a Senior Member of CCF.



ZHIPENG XIE received the B.S. degree from the Hunan University of Science and Engineering, China, in 2018. He is currently pursuing the M.S. degree in computer science and technology with the Changsha University of Science and Technology. His research interests include deep learning, computer vision, and object detection.

JUAN SUN received the B.S. degree from the Changsha University of Science and Technology, China, in 2018. She is currently pursuing the M.S. degree in computer science and technology with the Changsha University of Science and Technology. Her research is mainly about computer vision and object tracking.

XIN ZOU received the B.S. degree from the Changsha University of Science and Technology, in 2019, China. He is currently pursuing the M.S. degree in computer science and technology with the Changsha University of Science and Technology. His research interests include computer vision, deep learning, and object detection.

JIN WANG (Senior Member, IEEE) received the B.S. and M.S. degrees from the Nanjing University of Posts and Telecommunications, China, in 2002 and 2005, respectively, and the Ph.D. degree from Kyung Hee University, South Korea, in 2010. He is currently a Professor with the School of Computer and Communication Engineering, Changsha University of Science and Technology. His research interests mainly include wireless communications and networking, performance evaluation, and optimization. He is a member ACM.