This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2975828, IEEE Access

Multi-Feature Fusion and Enhancement Single Shot Detector for Traffic Sign Recognition

Yanmei Jin[1], Yusheng Fu[1]*, Wenqin Wang[1], Jinhong Guo[1], Chunhui Ren[1], Xin Xiang[2]

1 School of Information and Communication Engineering, University of Electronic Science and Technology of China, China
2 Air Force Engineering University, Xian, China

Corresponding author: Yusheng Fu (e-mail: yushengf@uestc.edu.cn)

ABSTRACT Road traffic sign detection and recognition play an important role in advanced driver assistance systems (ADAS) by providing real-time road sign perception information. In this paper, we propose an improved (Single Shot Detector) SSD algorithm via multi-feature fusion and enhancement, named MF-SSD, for traffic sign recognition. First, low-level features are fused into high-level features to improve the detection performance of small targets in the SSD. We then enhance the features in different channels to detect the target by enhancing effective channel features and suppressing invalid channel features. Our algorithm gets good results in domestic real-time traffic signs. The proposed MF-SSD algorithm is evaluated with the German Traffic Sign Recognition Benchmark (GTSRB) dataset. The experimental results show that the MF-SSD algorithm has advantages in detecting small traffic signs. Compared with existing methods, it achieves higher detection accuracy, better efficiency, and better robustness in complex traffic environment.

INDEX TERMS traffic sign detection; small target detection; single shot detector; feature fusion; feature enhancement

I. INTRODUCTION

The detection and recognition of road traffic signs are meaningful in advanced driver assistance systems [1] (ADAS) for enhanced driving safety. As traffic signs usually consist of specific shapes (circles, squares, and triangles) and colors (red, blue, and yellow), which have significant visual effects in road environments, traffic sign detection methods can be divided into color-based, shape-based, and color-based methods [2]-[3][4][5]. In color-based methods, RGB images are usually converted into other color spaces, such as HSI [6], CIELab [7], and HSL [8]. Then, the traffic signs are extracted via color threshold segmentation through intelligent data processing [9]. Color-based detection methods are usually vulnerable to complex lighting conditions in the traffic scene. In shape-based traffic sign detection, the geometric contour shape of traffic sign is detected by geometric symmetry [7]-[8][10][12]. Compared with the template matching in complex lighting environment, geometric moment invariant detection has better adaptability, but requires higher computational complexity. Nevertheless, the recognition rate of these methods should be further improved.

In recent years, the deep convolution neural network (CNN) for feature extraction has received much attention [13]-[14][15]. Benchmark works include GTSRB [16] and GTSDB [17]. The faster region-based CNN (Faster R-CNN) [18] is a representative two-stage target detection framework that has become a popular object detection framework, but it still has difficulty detecting small objects. In recent years, some new methods have been proposed to identify traffic signs [19]-[20][21].

Due to its small proportion in the image, the recognition of small traffic signs plays an important role for ITS security, but is difficult due to low resolution and noise effects. For instance, although PASCAL VOC and MSCOCO can achieve satisfactory performance for large objects, small object detection is still a challenge [22]. In this paper, we propose a small traffic sign recognition method that is different from previous ones based on GTSRB and GTSDB datasets.

The reasons for the difficulty of small target detection are summarized below.

A small target occupies fewer pixels with fewer features and is difficult to detect.

In CNN methods, low-level features may contain smaller target information but less semantic information, whereas high-level features contain abundant semantic information but less small target information. Consequently, small targets are not easy to detect.

Large computation complexity is required to detect small objects.

To improve the detection accuracy and speed for small objects, we propose an improved SSD algorithm by jointly exploiting feature fusion and enhanced SSD algorithm MF-SSD. The proposed traffic sign detection process is listed in Fig. 2. To resolve the problem that the SSD algorithm is not effective in detecting small objects, this paper proposes an improved SSD algorithm through feature fusion and enhancement, named MF-SSD. We fuse low-level features into high-level features to enhance the detection performance and detection efficiency of small targets in SSD.

II. Related Work

In recent years, the deep convolution neural network has been successfully applied to object recognition and target detection, with AlexNet being the representative case [23]. In 2012, Krizhevsky et al. demonstrated the CNN' s ability to

significantly improve image classification accuracy in the ImageNet Large-scale Visual Recognition Challenge Competition. Inspired by AlexNet′ s work, Ross Girshick et al. [24] proposed a deep learning model named R-CNN, which also has been applied to target detection. The model first uses a selective search algorithm to calculate the candidate regions of images and then inputs all candidate regions into R-CNN model. The feature is extracted from type A and the classification is completed in SVM [25]. Moreover, the model designs a bounding box regression algorithm to calculate the coordinates of candidate regions and tests it on the target detection set of PASCAL VOC. The average accuracy is about 20% higher than the non-neural network algorithm.

Model preprocessing also is applied in the above method. First, the weight of the network is initialized on the small dataset of ImageNet and then the network is fine-tuned on the PASCAL VOC dataset. In doing so, although the R-CNN accuracy is greatly improved, a large computation complexity is needed because there are about 2,000 candidate regions in each image. In the application of SPPnet [26] to target detection, Microsoft Asia Research Institute first makes a mapping and calculates the position of candidate regions mapped to the feature map of the highest convolution layer, then the pooling layer based on SPP algorithm is used to reduce the dimension and, finally, a feature layer of a specific size is obtained. Although its accuracy is similar to R-CNN, the running time is greatly reduced. In 2015, Ross Girshick further combined the idea of SPPnet with R-CNN to propose a convolutional neural network model, Faster R-CNN [18], and then replaced the SVM classifier with soft Max [27] regression to reduce the space and time overhead. The whole training process does not need to be graded and the detection process is more efficient and accurate. After training and testing on the GPU, the experimental results show that the extraction time of candidate regions is significantly shortened, the detection time is shortened to one-tenth, and the classification accuracy is increased.

In 2016, Liu Wei et al. combined the structure of the YOLO network with Girshick′ s Faster RCNN and proposed an SSD (Single Shot multibox Detector) target detection algorithm [28]-[29]. The SSD network is much faster than Faster R-CNN, but its working mode is significantly different. Faster R-CNN

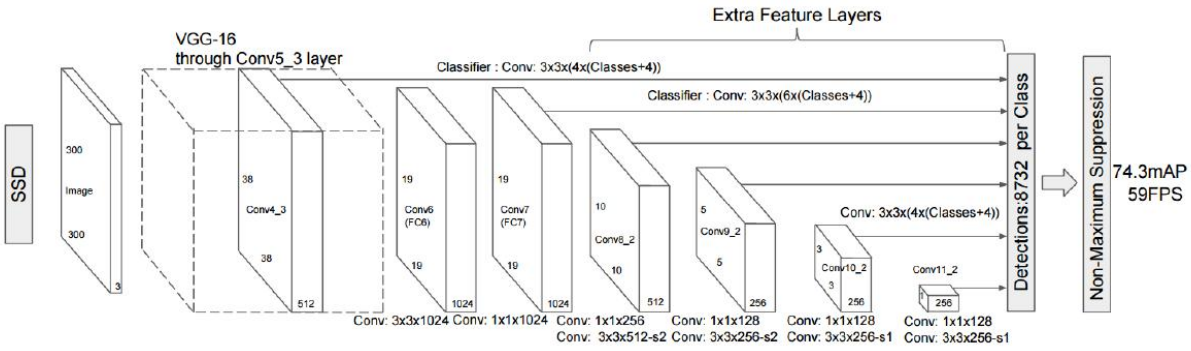

FIGURE 1. SSD scheme

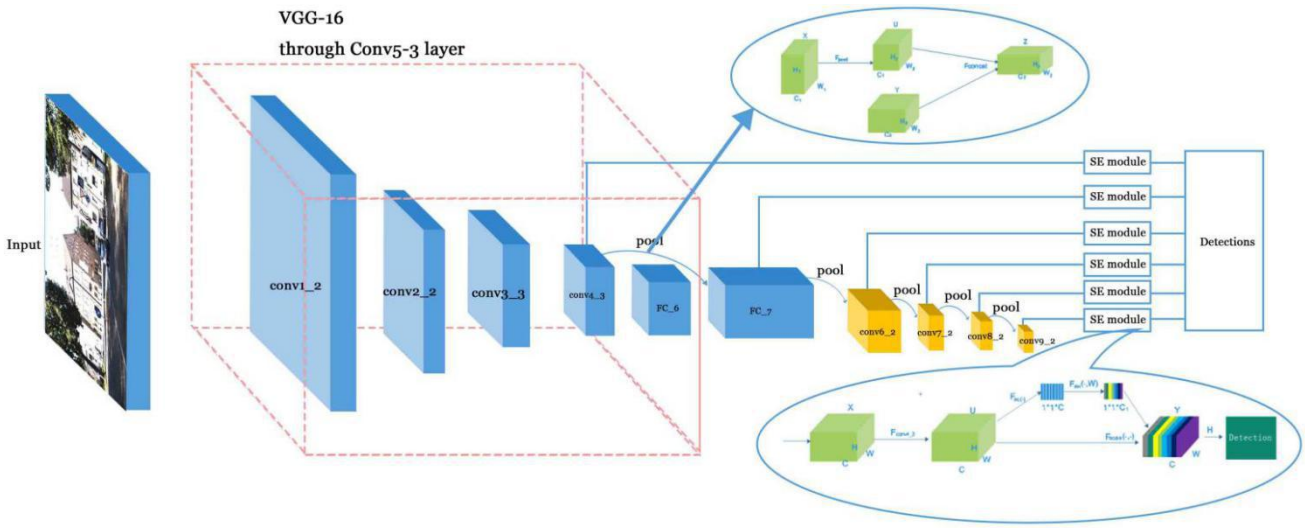

FIGURE 2. The architecture of multi-feature fusion and enhancement single shot fetector (MF-SSD)

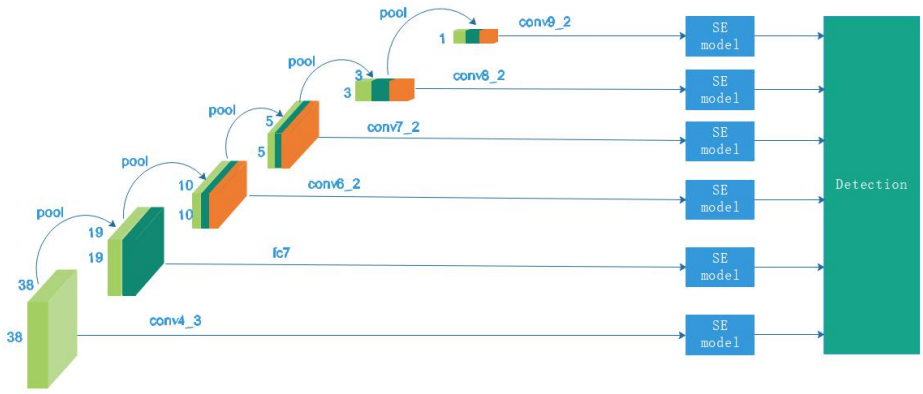

FIGURE 3. Structural block diagram of the proposed MF-SSD algorithm

[18] uses region inference to generate candidate regions and uses a classification algorithm to generate target frames in each candidate region. In contrast, the SSD algorithm generates target boundary frames of various sizes directly on the whole image and uses non-maximum suppression technology to integrate highly overlapping boundary frames into one. The candidate regions are transformed into a linear regression problem to find the prediction frame closest to the target so as to improve the calculation speed and accuracy. In 2017, SENet and SE modules were proposed [30]. SENet enables the network to enhance effective channel features and suppress invalid channel features according to global information. The SE module is not a complete network structure but a sub-structure, which can be embedded in other classification or detection networks. The method of embedding the SE module into the ResNet network in literature is the first in the ILSVRC2017 classification project. The working mode of SE module is to learn feature weights according to global letters, which makes the weight of effective channel features increase and the weight of ineffective or ineffective channel features decrease. Although embedding the SE module in the original classification or detection model will increase some parameters and computational complexity, the additional parameters and computational complexity are very small.

Recently, some small object detection methods based on original Faster R-CNN have been proposed, e.g., multi-scale input [31], multi-scale detector [32][33], multi-task learning [34][35], and multi-scale features [36] − [37][38]. However, these methods easily lead to heavy computation time in the training stage. To enhance the information representation ability of small objects in the feature map, the multi-input method [31] produces a high-resolution feature map. In references [32] and [33], the multi-scale detector is used to extract features from multiple consecutive layers to increase context information. However, the multi-detector also increases the computation cost in the training and testing stages. In literature [34][35], the multi-task learning method is used to improve detection performance. However, the feature map is only the output of the last layer and the information contained is not enough for small object detection. By combining the features of different layers, the representation of small objects in the feature map can be effectively enhanced. The multi- scale feature method [36]−[38] has attracted more attention than other methods in the field of small object detection.

Most of the existing SSD improvement algorithms are based on feature fusion. In reference [39], the RSSD network structure is proposed, where low-level features are fused to high-level features while high-level features are fused to low-level features. Literature [40] studies the FPN network. By extracting features from the image moving from the bottom to the top, a set of pyramid features are constructed and then the feature fusion is realized by using the up−sampling; finally, the target detection accuracy is improved. In reference [41], multiple low-level features are fused to enhance SSD detection of small targets.

III. Our Proposed Approach

Figure 1 shows the scheme of SSD, while the proposed MF-SSD is illustrated in Fig. 2, which is significantly modified to improve the performance of small traffic sign recognition. The overall framework of the algorithm is shown in Fig. 2. On the basis of the SSD framework, a feature fusion layer is added and the SE module is added to the feature extraction layer after fusion. The detailed process will be described next. Note that the feature fusion method is to fuse low-level features into high-level features. In the MF-SSD, the features at conv4_3 and fc7 are taken as low-level features, and the features at fc7, conv6_2, conv7_2, conv8_2, and conv9_2 are fused by pooling operation.

Figure 3 shows the process of feature fusion from conv4_3 to the feature group to be detected. A set of feature fusion processes are described in detail, with examples of feature fusion from conv4_3 to fc7, conv6_2, conv7_2, conv8_2, and conv9_2. The features at conv4_3 with a width of 38×38 are converted into the features of conv4_3_pool_fc7 with a width of 19×19 by pooling operation and the features of conv4_3_pool_fc7 are the leftmost features in the penultimate line. Then the features of fc7 and conv4_3_pool_fc7 with width and height of 19×19 form new features through series operation. After dyeing, these features are enhanced and finally

enter the detector. The following feature fusion process can be performed in a similar manner. The advantage lies in the shared pooling features, which will increase the relationship between layers. In the following, we provide an example of feature fusion from conv4_3 to fc7 to illustrate the details of feature fusion.

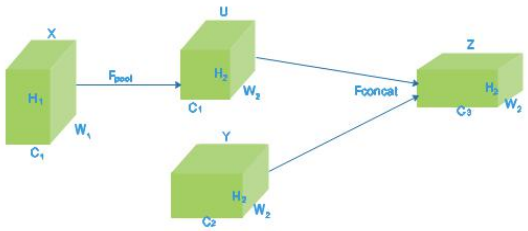

FIGURE 4. Process of feature fusion from conv4_3 to fc7

As shown in Fig. 4, $X \in R^{W_1 \times H_1 \times C_1}$ represents the feature map at conv4_3, $Y \in R^{W_2 \times H_2 \times C_2}$ represents the feature map at fc7, X is transformed into U through pooling operation, $U \in R^{W_2 \times H_2 \times C_2}$, and then U and Y are converted into Z through series operation, $Z \in R^{W_2 \times H_2 \times C_2}$.

The pooling operation can then be formulated as follows:

$$U_C(x,y) = F_{pool}(X_C) = \max(X_C(i,j) \mid i = x+l, \cdots, x+k; j = y+1, \cdots y+k) \quad (1)$$

where U_c denotes the data of the C channel of U characteristic graph, $U_c(x,y)$ denotes the data at the height equal to y width equal to X. Similarly, it can be inferred that X_c and $X_c(i,j)$. $k \times k$ is the pooling core with K being the step size and the patch is 0.

$$x = 0, \cdots, \frac{W_1}{k} - 1, y = 0, \cdots, \frac{H_1}{k} - 1 \quad (2)$$

The concatenation operations can then be expressed as

$$F_{conv4\_3\_fc7} = F_{conv4\_3} \circ F_{fc7} \quad (3)$$

where $\circ$ denotes the series operation, F_{conv4_3} and F_{fc7} are connected in series on the channel dimension, $Z \in R^{W_2 \times H_2 \times C_3}$, $C_3 = C_1 + C_2$.

After feature fusion, the features at fc7, conv6_2, conv7_2, conv8_2, and conv9_2 increase more at the lower level. These features contain more semantic information. Adding low-level features that contain additional small target information is helpful for subsequent small target detection. In the feature fusion, it should be noted that U and Y need Batch Normalization (BN) before concatenation in order to construct consistent scales for different feature maps. The batch standardization algorithm is listed as Table 1:

The principle of the SE feature enhancement module is as follows. Figure 5 illustrates the process of using the SE module to enhance the features at conv4_3, which enhances the feature U at conv4_3 to Y in the graph. The operation of converting feature X into feature U is a convolution operation, which belongs to SSD itself.

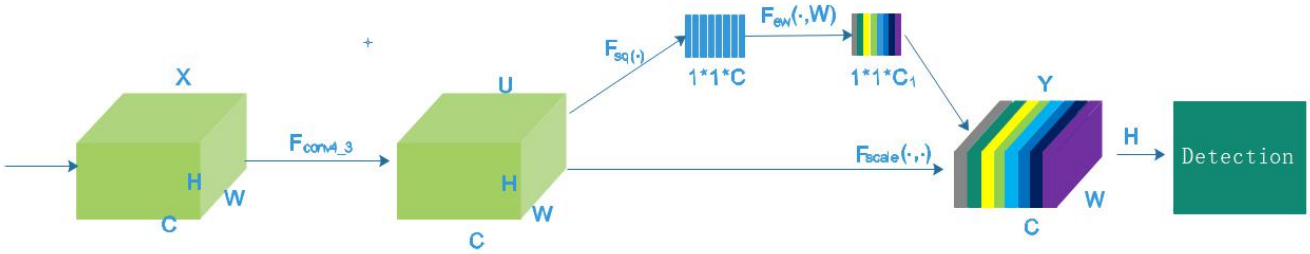

FIGURE 5. Enhances features at conv4_3 by using the SE module

TABLE I
STANDARD ALGORITHM FOR BATCH QUANTIZATION

Batch standardization (BN) algorithm

Input: small batch sample set: $B = \{x_{1 \cdots m}\}$

Parameters to learn: γ, β

Output: $\{y_i = BN_{\gamma,\beta}(x_i)\}$

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\delta_B^2 = \frac{1}{m}(x_i - \mu_B)^2$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\delta_B^2 + \varepsilon}}$$

$$y_i = \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$$

The operation of converting feature U to feature Y is the Squeeze-and-Excitation (SE) model enhancement operation. It is seen that after enhancement of the SE module, the size of feature U is not changed while the feature weights of different channels of feature U are. Finally, the enhanced feature Y of the SE model is sent to the detector for subsequent classification and detection.

F_{conv_3} is the convolution operation of conv4_3 in which the convolution core is 3×3, the patch is 1, and the step is 1. X is a three-dimensional matrix of size $W\times H\times C$ and U is a three-dimensional matrix of size $W\times H\times C$. After conv4_3 convolution operation, the size of X remained unchanged.

The formula of F_{conv_3} is as follows:

$$u_c = v_c * X = \sum_{s=1}^{C} v_c^s * X^s \tag{4}$$

where v_c denotes the convolution core of c, X^s denotes the input of s, U denotes the three-dimensional matrix of C with the size of $W\times H$, u_c denotes the two-dimensional matrix of C in U, and C denotes the channel.

The operation after F_{conv_3} is extrusion, where global average pooling is adopted. The formula is as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} u_c(i,j) \tag{5}$$

The number of channels of z is C and the number of channels of input feature graph is C. z has global information to some extent, representing the response intensity of each channel in the feature graph.

The extrusion operation is followed by an excitation operation, which converts the second matrix in Fig. 5 to the third matrix in Fig. 5. The activation formula is as follows:

$$s = F_{ex}(z,W) = \delta(g(z,W)) = \sigma(W_2\delta(W_1 z)) \tag{6}$$

where z is the output of the front extrusion operation. $W_1 z$ denotes the full connection layer process and $W_2\delta(W_1 z)$ denotes a full connection operation after the previous full connection. Finally, the sigmoid function is applied to the previous results and the s of dimension $1\times1\times C$ is obtained.

The final operation is to multiply s as a weight and U. The formula is as follows:

$$\tilde{X}_c = F_{scal}(u_c, s_c) = s_c \bullet u_c \tag{7}$$

where s_c denotes the number C in S.

Model evaluation index: the formulas of precision, recall and mean Average Precision (mAP) are as follows:

$$\Pr ecision = \frac{TP}{TP+FP} \tag{8}$$

$$\mathrm{Re}call = \frac{TP}{TP+FN} \tag{9}$$

$$mAP = \frac{1}{|Q_R|}\sum_{q\in Q_R} AP(q) \tag{10}$$

The formulas of $F1-Measure$ are as follows:

$$F1-Measure = \frac{2PR}{P+R} \tag{11}$$

where

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN} \tag{12}$$

as far as the location of the target is concerned, it is necessary to introduce an Intersection Over Union (IoU) to determine the positive case as a normal or a negative case. The formula is as follows:

$$IoU = \frac{Gt \cap Dr}{Gt \cup Dr} \tag{13}$$

where $Gt \cap Dr$ is the intersection of Gt and Dr, $Gt \cup Dr$ is the union of Gt and Dr.

The range of IoU is 0–1. Note that, in this paper IoU is set to 0.5. Once the detection position and label location are achieved, the target position is determined accordingly:

$$IoU = \frac{Gt \cap Dr}{Gt \cup Dr} \geq 0.5 \tag{14}$$

IV. Experiment Results

A. Datasets

There are two kinds of datasets used in this paper. One is the domestic (Chinese) traffic sign dataset, which contains 1,465 pictures. At present, seven kinds of image samples are marked as shown in Fig. 6. The image comes from a real picture of the city.

FIGURE 6. China traffic sign detection data

FIGURE 7. Example of pictures in GTSDB

Seven of them are shown in Fig. 6: right, straight, stop, nohonk, crosswalk, left, and background. The other data comprise German Traffic Sign Detection Benchmark (GTSDB) traffic signs (Fig. 7). There are 1,000 pictures and 43 kinds of marks as shown in Fig. 8. To test the detection effect of MF-SSD on each kind of traffic signs and evaluate our method for both small and large traffic sign detection, we divided the traffic signs into three size groups: small (0–32 pixels), medium (32–96 pixels), and large (96–200 pixels). In addition, it is worth noting that all the traffic signs used occupy less than 1% of the original image.

FIGURE 8. GTSDB signs (43 categories)

B. Detection performance

FIGURE 9. Examples of Chinese road signs detected by our model

As is seen in Fig. 9, our model obtains better detection results for Chinese traffic signs.

We use GTSDB datasets to compare experiments.

Table 2 provides detailed test indicators for each of the five methods, demonstrating that MF-SSD achieves the best performance in most categories. The experiments were run on a Linux PC with an Intel Core i5-8400K, 8 GB of memory, and one GeForce GTX 1060 GPUs.

We evaluated the performance of traffic signs through recall and accuracy.

First, we divided the traffic signs into 43 categories: Maximum speed limit (20), Maximum speed limit (30), Maximum speed limit (50), Maximum speed limit (60), Maximum speed limit (70), Maximum speed limit (80), End of speed limit (80), Maximum speed limit (100), Maximum speed limit (120), No passing, No passing for vehicles over 3.5 t, Priority, Priority road, Yield, Stop, Road closed, Vehicles over 3.5 t prohibited, Do not enter, General danger, Curve (left), Curve (right), Double curve, Rough road, Slippery when wet or dirty, Road narrows (right or side), Road work, Traffic signals ahead, Pedestrians, Watch for children, Bicycle crossing, Beware of ice/snow, Wild animal crossing, Lane added (left), Mandatory direction of travel (right), Mandatory direction of travel (left), Mandatory direction of travel (straight), Mandatory direction of travel (straight or right), Mandatory direction of travel (straight or left), Pass by on right, Pass by on left, Yield to roundabout, End of no passing zone, End of no passing zone for vehicles over 3.5 t. The worst precision results were obtained by the category of compulsory traffic signs in almost all models tested. The comparison of results in Table 2 shows that our algorithm has higher precision.

TABLE II
COMPARISON OF THE RECOGNITION PERFORMANCE OF FIVE MODELS ON DIFFERENT CATEGORIES

mAP	ssd_mobilenet_v2_coco	faster_rcnn_inception_v2_coco	ssd_mobilenet_v1_coco	ssd_inception_v2_coco	FSSD	*Our model*
Classification_loss	9.27	0.13	10.67	10.54	9.25	9.18
Localization_loss:	1.12	0.56	1.13	1.32	1.10	1.00
PascalBoxes_Precision	0.275	0.31	0.29	0.26	0.25	0.28
Bicycle crossing	0	0.07	0	0	0	0
Curve(left)	0	0	0	0	0	0
Curve(right)	0	0	0.25	0.01	0.25	0.04
Do not enter	0.33	0.34	0	0.33	0.33	0.33
End of no passing zone for	0	1	0.5	0	0	0

vehicles over 3.5t						
End of speed limit(80)	1	1	1	1	1	1
General danger	0.5	0.56	0.5	0.1	0.2	0.33
Lane added(left)	0	0	0	0.04	0	0
Mandatory direction of travel(left)	0	0	0	0.13	0.13	0.13
Mandatory direction of travel(straight)	0.33	0.33	0.12	0.17	0.33	0.37
Maximum speed limit(100)	0	0.05	0.5	0	0.05	0.08
Maximum speed limit(30)	0.46	0.22	0.36	0.32	0.36	0.47
Maximum speed limit(50)	0.23	0.48	0.06	0.20	0.18	0.14
Maximum speed limit(60)	0.07	0.16	0.08	0.20	0.12	0.13
Maximum speed limit(70)	0.26	0.52	0.015	0.10	0.24	0.10
Maximum speed limit(80)	0.71	0.50	0.54	0.60	0.64	0.27
No passing for vehicles over 3.5t	0.02	0.68	0.25	0.14	0.16	0.33
No passing	0.82	0.56	0.88	0.57	0.72	0.60
Pass by on right	0.40	0.54	0.38	0.44	0.41	0.43
Priority road	0.60	0.18	0.59	0.55	0.65	0.7
Priority	0.52	0.29	0.78	0.60	0.63	0.87
Road closed	0.11	0.03	0	0	0.18	0.25
Road narrows(right or side)	1.00	1.00	1.00	1.00	1.00	1.00
Road work	0.06	0.07	0	0	0.1	0.17
Rough road	0	0.08	0.50	0.50	0.33	0.33
Slippery when wet or dirty	0.22	0.18	0.17	0.40	0.20	0.087
Stop	0.50	0.30	0.33	0.50	0.50	0.53
Traffic signals ahead	0	0.03	0	0	0	0
Vehicles over 3.5t prohibited	0.10	0.06	0	0	0	0
Watch for children	0.02	0.04	0.25	0.08	0	0
Yield to roundabout	0.17	0.50	0	0	0.20	0.10
Yield	0.33	0.23	0.38	0.25	0.31	0.31

Examples of detection using five different models in a road scene are illustrated in Fig. 10. All detections are correct in the examples. As can be seen from the figure, our method has the highest detection rate.

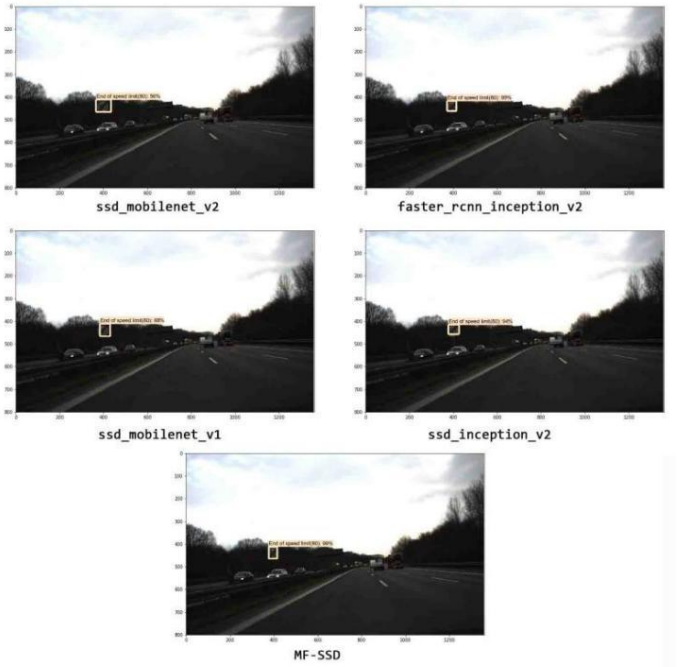

FIGURE 10. Examples of road sign detection using five different models

Second, we divided traffic signs into three size categories: small (0–32 pixels), medium (32–96 pixels), and large (96–200 pixels). For more intuitive comparisons,

we also use F1_measure as an additional metric. To verify the effectiveness of the method, we compared MF-SSD with SSD, faster_rcnn, and FSSD [40]. Faster_rcnn is a detection method for multi-scale objects, which achieves better performance on MS COCO and PASCAL VOC datasets. FSSD, as proposed by Zuoxin Li and Fuqiang Zhou, has higher accuracy and speed than the conventional SSD by a large margin.

TABLE III
COMPARISON OF RECOGNITION PERFORMANCE FOR DIFFERENT SIZE GROUPS

Method	Metrics	Small	Medium	Large
ssd_mobilenet_v2	Recall	43.4	77.5	86.9
	Precision	25.3	67.8	81.5
	F1_measure	32.0	72.3	84.1
faster_rcnn_inception_v2	Recall	49.5	84.4	90.1
	Precision	24.8	65.7	80.0
	F1_measure	35.6	77.5	84.3
ssd_mobilenet_v1	Recall	41.5	76.5	85.7
	Precision	24.6	64.8	86.4
	F1_measure	35.7	73.6	82.5
ssd_inception_v2	Recall	43.6	76.7	86.3
	Precision	26.5	65.4	82.9
	F1_measure	31.4	71.5	83.8
FSSD	Recall	44.1	78.4	87.4
	Precision	28.5	67.6	81.6
	F1_measure	34.8	70.1	83.9
Our model	**Recall**	**45.6**	**79.2**	**88.1**
	Precision	**28.8**	**67.5**	**82.6**
	F1_measure	**35.9**	**73.7**	**85.8**

Table 3 provides a comparison of the performance of these three methods on different traffic sign size groups. The precision measurements of small and medium sizes obtained by the proposed MF-SSD are 28.8 % and 67.5 % respectively. The precision value of the large size is 82.6 %, which is superior to the precision of other methods. This shows that MF-SSD can accurately identify small traffic signs and medium or large traffic signs.

Figure 11 shows the partial visualization results of the test dataset under different weather conditions. It can be seen that each traffic sign instance is very small, accounting for less than 1% of the whole scene; nevertheless, our method can recognize them accurately.

The German Traffic Sign Detection Benchmark (GTSDB) used in this paper is highly accepted and widely used in traffic sign detection methods in comparative literature. GTSDB includes natural traffic scenarios recorded in various types of roads (roads, villages, cities) during daytime and dusk, and numerous weather conditions. The dataset consists of 900 complete images containing 1,206 traffic signs, which are divided into 600 training sets (846 traffic signs) and 300 test sets (360 traffic signs). Each image contains zero or one or more traffic signs, which are usually affected by differences in direction, light conditions, or occlusion. The signs are classified into four categories: mandatory, prohibitive, dangerous, and other. The training set contains 396 prohibitions (59.5%), 114 (17.1%) mandatory, and 156 (23.4%) dangerous traffic sign samples while the test set includes 161 prohibitions, 49 mandatory, and 63 dangerous traffic sign images. Figure 6 shows some images of the dataset. We divided the traffic signs into three size categories—small (0–32 pixels), medium (32–96 pixels), and large (96–200 pixels)—and tested the detection effect of each traffic sign by MF-SSD.

(a) Sunny day (b) Weak light

(c) Rainy day (d) Motion blur

(e) Small size sign (f) Sign occlusion

FIGURE 11. Example images from GTSDB data set

V. Conclusion

To improve small target detection performance, this paper proposed an improved algorithm named MF-SSD, which combines low-level features with high-level features and adds the SE module to improve the detection accuracy. The experimental results verified that the proposed method outperforms conventional methods detecting small objects with respect to detection accuracy and efficiency. Although our method has a great improvement in detecting small target image, there is still large room for improving the accuracy for a real-time application. In future work, we will continue to improve the algorithm and strive to apply the framework to the domestic traffic sign dataset to achieve real-time application.

REFERENCES

[1] G. Anjan, C. Shreesha, U. Raghavendra, "A review on automatic detection and recognition of traffic sign," Multimed. Tools Appl., vol. 75, no. 1, pp. 333-364, 2016.

[2] G. Piccioli, E. DeMicheli, P. Parodi, M. Campan, "Robust method for road sign detection and recognition," Image Vis. Comput., vol. 14, no. 3, pp. 209-223, 1996.

[3] Y. Tsai, P. Kim, Z. H. Wang, "Generalized Traffic Sign Detection Model for Developing a Sign Inventory," J. Comput. Civil. Eng., vol. 23, no. 5, pp. 266-276, 2009.

[4] X. Yuan, X. L. Hao, H. J. Chen, X. Y. Wei, "Robust Traffic Sign Recognition Based on Color Global and Local Oriented Edge Magnitude Patterns," IEEE Trans. Intell. Transp. Syst., vol. 15, no. 4, pp. 1466-1477, 2014.

[5] H. J. Li, F. M. Sun, L. J. Liu, "A novel traffic sign detection method via color segmentation and robust shape matching," Neurocomputing., vol. 169, pp. 77-88, 2015.

[6] A. D. Escalera, L. E. Moreno, M. A. Salichs, J. M. Armingol, "Road traffic sign detection and classification," IEEE Trans. Ind. Electron., vol. 44, no. 6, pp. 848-859, 1997.

[7] J. F. Khan, M. A. Bhuiyan, R. Adhami, "Image Segmentation and Shape Analysis for Road-Sign Detection," EEE Trans. Intell. Transp. Syst., vol. 12, no. 1, pp. 83-96, 2011.

[8] F. Perez, C. Koach, "Toward color image segmentation in analog vlsi: Algorithm and hardware," Int. J. Comput. Vis., vol. 12, no. 1, pp. 17-42, 1994.

[9] G. Anjan, C. Shreesha, U. Raghavendra, U. Acharya, "Multiple Thresholding and Subspace Based Approach for Detection and Recognition of Traffic Sign," Multimedia Tools and Applications, vol. 76, no. 5, pp. 6973-6991, 2017.

[10] Z. X. Cai, M. Q. Gu, "Traffic sign recognition algorithm based on shape signature and dual-tree complex wavelet transform," J. Cent. South Univ., vol. 20, no. 2, pp. 433-439, 2013.

[11] X. Yuan, X. L. Hao, H. J. Chen, X. Y Wei, "Traffic sign recognition based on a context-aware scale-invariant feature transform approach," J. Electron. Imaging., vol. 22, no. 4, pp. 1-17, 2013.

[12] S. K. Berkaya, H. Gunduz, O. Ozsen, C. Akinlar, S. Gunal, "On circular traffic sign detection and recognition," Expert Syst. Appl., vol. 48, pp. 67-75, 2016.

[13] G. Jack, M. Majid, "Real-Time Detection and Recognition of Road Traffic Signs," IEEE Trans. Intell. Transp. Syst., vol. 13, no. 4, pp. 1498-1506, 2012.

[14] A. Shustanov, P. Yakimov, "CNN design for real-time traffic sign recognition," Procedia Eng., vol. 201, pp. 718-725, 2017.

[15] J. Jin, K. Fu, C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," IEEE Trans. Intell. Transp. Syst., vol. 15, no. 5, pp. 1991-2000, 2014.

[16] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark. Proc," IEEE Int. Joint Conf. Neural Netw., pp. 1-8, 2013.

[17] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, "The German traffic sign recognition benchmark: A multi-class classification competition," Proc. IEEE Int. Joint Conf. Neural Netw., pp. 1453-1460, 2011.

[18] S. Ren, K. He, R. Girshick, 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, 2017.

[19] M. A. B. Mahmoud, P. Guo, "A Novel Method for Traffic Sign Recognition Based on DCGAN and MLP With PILAE Algorithm," IEEE Access, vol. 7, pp. 74602-74611, 2019.

[20] G. Anjan, C. Shreesha, U. Raghavendra, U. Acharya, "An efficient traffic sign recognition based on graph embedding features," Neural Computing and Applications, vol. 31, pp. 395–407, 2019.

[21] G. Anjan, C. Shreesha ; U. Raghavendra, U. R. Acharya, "Local texture patterns for traffic sign recognition using higher order spectra," Pattern Recognition Letters, vol. 94, pp. 202–210, 2017.

[22] S. B. Wali, M. A. Abdullah, M. A. Hannan, A. Hussain, S. A. Samad, P. J. Ker, M. B. Mansor, "Vision-Based Traffic Sign Detection and Recognition Systems: Current Trends and Challenges," Sensors , Basel, Switzerland, vol. 19, no. 9, pp. 2093, 2019.

[23] K. Alex, S. Llya, H. Geoffrey, "ImageNet Classification with Deep Convolutional Neural Networks," Commun. ACM., vol. 60, no. 6, pp. 84-90, 2017.

[24] R. Girshick, J. Donahue, T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," In: IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[25] M. B. Saturnino, L. A. Sergio, G. J. Pedro, "Road-sign detection and recognition based on support vector machines," IEEE Trans. Intell. Transp. Syst., vol. 8, no. 2, pp. 264-278, 2007.

[26] K. M. He, X. Y. Zhang, S. Q. Ren, 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, 2015.

[27] G. Bouchard, "Clustering and classification employing softmax function including efficient bounds," U.S. Patent 8,065, vol. 246, 2011.

[28] J. Redmon, S. Divvala, R. Girshick, "You Only Look Once: Unified, Real-Time Object Detection," In: IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016.

[29] W. Liu, D. Anguelov, D. Erhan, "SSD: Single shot multibox detector," Proc. Eur. Conf. Comput. Vis., pp. 21-37, 2016.

[30] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, "Squeeze-and-excitation networks," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 4, 2018.

[31] X. Z. Chen, K. Kundu, Y. Zhu, S. Fidle, R. Urtasun, H. Ma, "3D object proposals using stereo imagery for accurate object class detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1259-1272, 2018.

[32] P. Hu, D. Ramanan, "Finding tiny faces," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 951-959, 2017.

[33] Q. Chen, X. Meng, W. Li, X. Fu, X. Deng, J. Wang, "A multi-scale fusion convolutional neural network for face detection," Proc. IEEE Conf. Syst. Man Cybern., pp. 1013-1018, 2017.

[34] J. Dai, Y. Li, K. He, J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," Proc. Adv. Neural Inf. Process. Syst., pp. 379-387, 2016.

[35] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN," Proc. IEEE Int. Conf. Comput. Vis., pp. 2961-2969, 2017.

[36] S. Bell, C. L. Zitnick, K. Bala, R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2874-2883, 2016.

[37] T. Kong, A. Yao, Y. Chen, F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 845-853, 2016.

[38] H. Zhang, K. Wang, Y. Tian, C. Gou, F. Y. Wang, "MFR-CNN: Incorporating multi-scale features and global information for traffic object

detection," IEEE Trans. Veh. Technol., vol. 67, no. 9, pp. 8019-8030, 2018.
[39] J. Jeong , H. Park , N. Kwak, "Enhancement of SSD by concatenating feature maps for object Detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017.
[40] Z. Li, F. Zhou, "FSSD: Feature Fusion Single Shot Multibox Detector," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 3, 2017.
[41] G. Cao, X. Xie, W. Yang, "Feature-Fused SSD: Fast Detection for Small Objects," in Proceedings of the Ninth International Conference on Graphic and Image Processing, vol. 10615, pp. 106151E-106151E-8, 2018.

Yusheng Fu received his Bachelor's degree in avionics engineering from the Air Force Engineering University, Xi'an, China in 1995, Master's degree from the University of Electronic Science and Technology of China, Chengdu, China in 2000, and PhD from the University of Electronic Science and Technology of China, Chengdu, China in 2004.

His research over the past five years has mainly focused on signal processing, aeroelectronics, biomedical electronics engineering and, more recently, network science and technology. Over the past five years, a total of 10 million yuan has been spent on scientific research, with an average annual expenditure of more than 1 million yuan. More than 10 academic papers have been published, including three SCI papers and more than 10 EI papers. He is a reviewer of Circuit System and Signal Processing and other academic journals.

Yanmei Jin received her Bachelor's degree in communications engineering from Zhengzhou University in 2011.

Currently, she is a Master's candidate in the University of Electronic Science and Technology of China, Chengdu, China.

Jinhong Guo received his Bachelor's degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China in 2010 and PhD degree in biomedical engineering from Nanyang Technological University in 2014.

Currently, he is a full professor in the School of Communication and Information Engineering, University of Electronic Science and Technology of China and Chengdu University of Traditional Medicine, Chengdu, China. After his doctoral studies, he was a postdoctoral fellow in the Pillar of Engineering Design at MIT-SUTD Singapore from 2014 to 2015. He then worked as a Visiting Professor in the School of Mechanical Engineering at the University of Michigan, Ann Arbor, from January to July 2016. His current research focuses on electrochemical sensors and lab-on-a-chip devices for clinical Point of Care Testing (POCT). He is a recipient of the China Sichuan Thousand Talents Plan for Scholars Award (2015) and Chengdu Expert in Science and Technology Award (2015). He is also appointed as Chief Scientist at Longmaster Information Co., Ltd. (one listed corporation in China, stock ID: 300288) and is in charge of the research and development center for POCT. He has published over 70 research papers in high-impact journals such as IEEE TII, TBME, TBioCAS, Analytical Chemistry, Biosensor and Bioelectronics, etc.

Chunhui Ren received her Bachelor's, Master's, and PhD degrees from the University of Electronic Science and Technology of China, Chengdu, China in 1992, 1998 and 2006, respectively.

In recent years, her research has mainly focused on electronic countermeasures, statistical signal processing, non-cooperative signal processing, and other fields. She has published more than 10 academic papers, including several papers published in SCI (including SCIE) source journals and several EI-retrieved journals and conference papers.

Xin Xiang received his PhD degree from Xidian University, Xi'an, China. His research interest is in communication signal processing. He is now a professor at the Air Force Engineering University, Xi'an, China.