IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 10, 2019, accepted April 21, 2019, date of publication April 29, 2019, date of current version May 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913882

MR-CNN: A Multi-Scale Region-Based Convolutional Neural Network for Small Traffic Sign Recognition

ZHIGANG LIU[1,2], JUAN DU[1], FENG TIAN[1], AND JIAZHENG WEN[3]
[1]School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583
[3]Department of Computer Science for Artificial Intelligence, University of Nottingham Ningbo, Ningbo 315100, China

Corresponding author: Zhigang Liu (zhigangliu313@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502094, Grant 51774090, and Grant 51104030, in part by the Natural Science Foundation of Heilongjiang Province under Grant F2016002, Grant E2016008, and Grant F2015020, and in part by the Youth Science Foundation of Northeast Petroleum University under Grant 2017PYZL-06 and Grant 2018YDL-22.

ABSTRACT Small traffic sign recognition is a challenging problem in computer vision, and its accuracy is important to the safety of intelligent transportation systems (ITS). In this paper, we propose the multi-scale region-based convolutional neural network (MR-CNN). At the detection stage, MR-CNN uses a multi-scale deconvolution operation to up-sample the features of the deeper convolution layers and concatenates them to those of the shallow layer to construct the fused feature map. The fused feature map has the ability to generate fewer region proposals and achieve higher recall values. At the classification stage, we leverage the multi-scale contextual regions to exploit the information surrounding a given object proposal and construct the fused feature for the fully connected layers. The fused feature map inside the region proposal network (RPN) focuses primarily on improving the image resolution and semantic information for small traffic sign detection, while outside the RPN, the fused feature enhances the feature representation by leveraging the contextual information. Finally, we evaluated MR-CNN on the largest dataset, Tsinghua-Tencent 100K, which is suitable for our problem and more challenging than the GTSDB and GTSRB datasets. The final experimental results indicate that the MR-CNN is superior at detecting small traffic signs, and that it achieves the state-of-the-art performance compared with other methods.

INDEX TERMS Context, deconvolution, small traffic sign, Tsinghua-Tencent 100K.

I. INTRODUCTION

Traffic sign recognition is playing an increasingly important role in intelligent transportation systems (ITS). In the real world, traffic signs vary due to different viewpoints, motion blur, illumination, etc., all of which increase the difficulty of accurate recognition. Previous work usually used hand-crafted features and machine learning models to recognize traffic signs. However, designing hand-crafted features for a specific task, including specific colours [1], [2], shapes [3], [4] and other discriminative features [5]–[9] is both labour-intensive and difficult. As computational capacity has increased and large-scale datasets have become available, deep convolutional neural networks (CNNs) have become suitable for extracting features from raw images without the need for hand-crafted features; thus, they have been applied in many applications. Various CNN-based methods [4], [10]–[17] have been proposed and obtained state-of-the-art results on the GTSRB [18] and GTSDB [19] benchmark datasets. Unfortunately, these methods do not work well in real-world applications, primarily because of the inadequacy of these benchmark datasets.

The GTSDB dataset provides only one of the 4 major categories of traffic signs for detection, while each traffic sign in the GTSRB images occupies a large proportion of the image. Due to their use of both these datasets, most of the existing studies separated traffic sign recognition into two independent tasks, classification and detection, causing a gap between these tasks. In 2016, Zhu et al. [16] proposed the Tsinghua-Tencent 100K dataset, the largest and

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

most challenging traffic sign dataset. The Tsinghua-Tencent 100K dataset includes 100,000 images in 100 classes and 30,000 traffic sign instances. Compared with the GTSRB and GTSDB datasets, each image has a higher resolution (2,048×2,048), and each traffic sign instance generally occupies a smaller proportion of the image, e.g., 1%, making the Tsinghua-Tencent 100K dataset more suitable for the task of small traffic sign recognition.

Because the proportions of traffic signs in the image are determined by their distance from the camera and the overall size of the traffic scene, small traffic sign recognition is very important for ITS safety. Thus, this paper mainly aims to small traffic sign recognition, which differs from the previous works on the GTSRB and GTSDB datasets. Note that small traffic sign recognition is a difficult problem due to its low resolution and noisy representation. Although the novel object detection frameworks have been constantly proposed recently and achieved the state-of-the-art performance on PASCAL VOC and MS COCO, small object detection is much more challenging than normal object detection and good solutions are still rare so far.

Recently, faster region-based convolutional neural network (Faster R-CNN) [23] is the representative of two-stage object detection framework and has become a popular object-detection framework; nonetheless, it still struggles to detect small objects for the following reasons. First, when the VGG-16 extracts the features as the backbone structure, the feature map has low resolution and a large receptive field in each pixel due to the small size of the feature map (approximately 1/16) compared with the original image. This coarseness leads to poor localization performance for small traffic signs. For example, the dimensions of a traffic sign might be 32×32 pixels in the original image but are represented by only 2×2 pixels in the feature map, which is insufficient to encode discriminative features. Second, the smaller the dimensions of a region of interest (RoI) are, the more disruptive information is imported from outside the RoI to each pixel of the feature map, which unnecessarily increases the uncertainty of traffic sign recognition.

To enhance the recognition accuracy, we propose the multi-scale region-based convolutional neural network (MR-CNN) detection framework that simultaneously employs fused feature representations in the detection and classification stages. In the detection stage, the multi-scale deconvolution operator up-samples the output of the deeper convolution layer. The features from the shallow convolution layer and the deconvolution layers are normalized to the same scale by L2 normalization and concatenated along the channel axis. This approach constructs the fused feature map by compressing the concatenated feature using pointwise convolution. In the classification stage, we assume that contextual information can provide some important cues for small traffic sign recognition (e.g., cues from the surrounding environment or discriminative parts). Thus, the fused feature is designed to provide more discriminative representation for small traffic signs. It uses the contextual regions to leverage the information surrounding an object proposal region. The final fused feature is constructed by RoI-pooling, concatenation and compression operations. In addition, pointwise convolution is employed to weaken the interference caused by background noise, which is potentially introduced by the contextual information. Finally, we evaluated the performance of the MR-CNN on the challenging new Tsinghua-Tencent 100K dataset, where it obtained state-of-the-art results and achieved a considerable improvement compared with other methods. The experimental results demonstrate the superiority of MR-CNN for small traffic sign recognition. Our contributions can be summarized as follows.

1) In the detection stage, we combine the CNN features and design a multi-scale fused feature map. To ensure different feature maps have the same dimension, we leverage the deconvolution alleviated the loss of information, which is obviously different from the up-sampling operation used in the previous work.
2) For each object proposal, multiple contextual regions with the same centre coordinate as the predefined factors are generated. During the classification stage, the fused feature leverage the multi-scale contextual information and thus enhances the feature representation.
3) Using a two-stage fusion strategy, MR-CNN is evaluated on the Tsinghua-Tencent 100K dataset and obtains state-of-the-art results (an F1-measure of 86.0% for small (sizes$\in$ $(0, 32]$ pixels), 93.5% for medium (sizes$\in$ $(32, 96]$ pixels), and 90.1% for large (sizes$\in$ $(96, 200]$ pixels).

The remaining of this paper is organized as follows. Section II reviews the related work. Section III details the proposed object detection framework. Section IV provides the experimental results and comparison between our method and other frameworks on Tsinghua-Tencent 100K dataset. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

Previous research work usually used intuitive features (e.g., colour and shape) or more discriminative and sophisticated hand-crafted features to detect the traffic signs. However, these features show the limited feature representation power and are not robust enough to accurately detect the traffic signs. In addition, designing hand-crafted features is labour-intensive and difficult because it consumes a lot of time and needs abundant expertise. Recently, due to the increase in computational capacity and the advent of large-scale datasets, deep convolutional neural networks (CNNs) have demonstrated their capabilities to self-learn features from raw images. Thus, some efforts using CNNs have been devoted to addressing traffic sign detection and classification. Regarding traffic sign detection, Wu et al. [14] began employing CNNs to address traffic sign detection as a candidate region classifier. Zhu et al. [15] proposed a novel framework in which a fully convolutional network generated the region proposals and CNNs was used for classification. In the literature [16], a pipeline based on a fully convolutional network was designed to perform both detection and classification simultaneously. Unfortunately, the aforementioned methods do not work well in real-world applications, primarily because of the inadequacy of GTSDB and GTSRB datasets.

Note that small traffic sign recognition is quite important for ITS safety because the proportions of signs in the image are determined by their distance from the camera. However, small object detection is much more challenging than normal object detection and good solutions are still rare so far. The novel object detection frameworks have been constantly presented in recent years and obtained state-of-the-art performance on PASCAL VOC and MS COCO; nevertheless, the objects in VOC and COCO had a large size. Recently, Faster R-CNN [23], You Only Look Once (YOLO) [24] and Single Shot MultiBox Detector (SSD) [25] have become popular object detection frameworks. YOLO and SSD show the fast detection speed while struggling to precisely localize small objects because they both divide images into many large grids that contain perhaps two or more small objects. Faster R-CNN, the two-stage object detection framework, shows a higher accuracy than YOLO and SSD. However, the coarseness of the feature map easily leads to poor localization performance for small traffic sign due to its low resolution.

Most recently, some efforts based on the original Faster R-CNN have been presented for small object detection and fall into three categories: multi-scale input [26], multi-scale detector [27], [28], multi-task learning [29], [30], and multi-scale features [31]–[33]. To enhance the information representation power of small object in the feature map, multi-input method [26] produced the high-resolution feature map. However, simply increasing the scale of images easily leads to heavy time computation in the training stage. In the literature [27], [28], the multi-scale detector was employed to extract features from multiple consecutive layers to increase the contextual information. However, multi-detectors also increase the computation overhead in the training and testing stage. In the literature [29], [30], multi-task learning method was employed to improve the detection performance. However, the feature map is the output only by the last layer, and contains insufficient information for small object detection. The multi-scale feature method [31]–[33] received more attention than other aforementioned methods in the field of small object detection. It can effectively enhance the representation power of the small object in the feature map by combining the features from different layers. To make the lower-level features the same size as the higher-level features, pooling was applied to the lower-level features. Note that pooling is down-sampling operation and further leads to the loss of small object details in the feature map.

Inspired by these methods, our work follows a similar philosophy of gathering expressive features from different convolution layers. But we make the notable modification and use the deconvolution operation to up-sample the higher-level features. In addition, the multi-scale contextual information is leveraged to enhance the feature representation at the classification stage.

III. OUR PROPOSED APPROACH

Our proposed MR-CNN makes notable modifications to improve the small traffic sign recognition performance. As shown in Figure 1, the input traffic scene passes through several convolution layers (conv3, conv4, and conv5), which are operationally concatenated by a deconvolution and a normalization layer and compressed into a fused feature map. Then, the subsequent RPN generates the region proposals from the fused feature map. Finally, a RoI-pooling layer extracts a fixed size feature vector from each object proposal and its proposed contextual regions. We design the fused feature by concatenating and compressing the feature vectors, and fed them into a subsequent fully connected layers.

A. MULTI-SCALE FUSED FEATURE MAP

For notational convenience, $C = \{C_i \,|i = 1, 2, \ldots, 5\}$ denotes the outputs of the different convolution layers in VGG-16. The number of output channels of different convolution layer is 64, 128, 256, 512, and 512 in sequence. The size of kernel, stride, and padding used in each convolution layer is 3×3, 1, and 1, respectively.

To detect small traffic signs, we empirically compare the performance of different convolution layers and find that C_3 is the most suitable for localization because it possesses smaller receptive fields and higher resolution compared with C_4 and C_5. However, when used as a feature map, C_3 leads to poor detection performance because it contains less semantic information. Thus, we design a fused feature map that improves the resolution of a small traffic sign and simultaneously includes more semantic information, which improves the performance of the region proposals.

First, we use the multi-scale deconvolution operation to up-sample the output of the deeper convolution layers. Notably, this deconvolution operation is different from the original up-sampling operation; it provides a set of parameters by which to learn a nonlinear up-sampling of the features in the deep layers. The outputs of the deconvolution layer are denoted as $D = \{D_i \,|i = 1, 2, \ldots, 5\}$, where D_i is defined as

$$D_i = \mathrm{Deconv}(C_i, o_i, k_i, s_i, m_i) \tag{1}$$

where $\mathrm{Deconv}(\cdot)$ specifies the deconvolution operation. The parameters o_i, k_i, s_i and m_i specify the size of the output channel, kernel, stride, and padding, respectively.

Next, the features of different layers, C_3, D_4, and D_5, are assembled to concatenated feature (CF) that can be defined as

$$\begin{aligned} CF_{\{i=3,4,5\}} &= \mathrm{concat}(C_3, D_4, D_5) \\ &= L_2(C_3) \oplus L_2(D_4) \oplus L_2(D_5) \end{aligned} \tag{2}$$

where the concatenation operation is denoted as $\oplus$, L_2 specifies the L2 normalization and $D_4 = \mathrm{Deconv}(C_4, 256, 4, 2, 1)$, $D_5 = \mathrm{Deconv}(C_5, 256, 8, 4, 2)$.

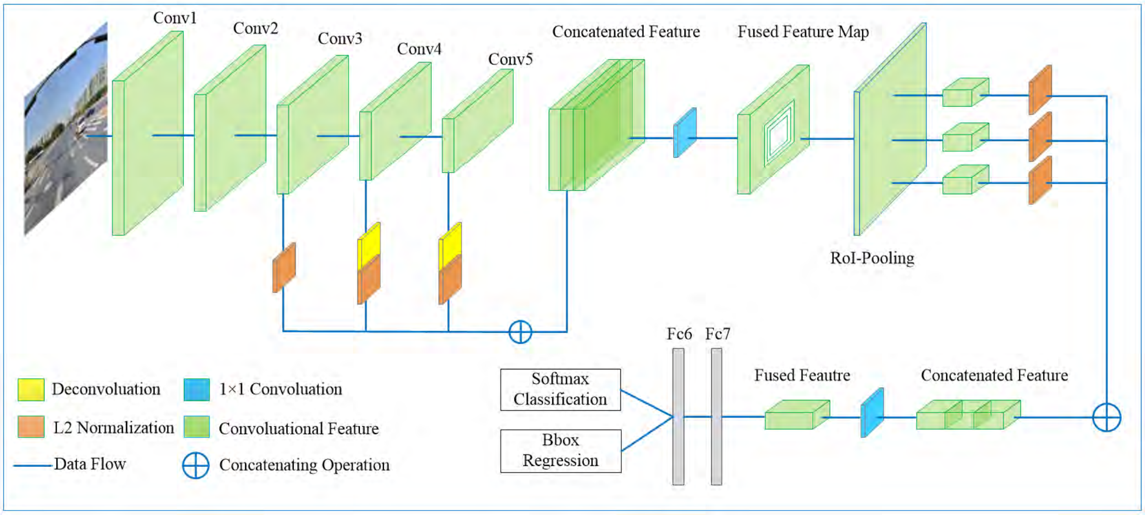

FIGURE 1. The architecture of our proposed multi-scale region-based convolutional neural network (MR-CNN).

The features of different convolution layers have a different scale of value, and feature values from the shallower layer are generally larger than them from the deeper layer. Directly concatenating them easily leads to the smaller values being dominated by the larger values. Thus, L2 normalization is a crucial step before the concatenation operation because it can effectively keep the feature values from the different convolution layer on the same scale. For each pixel vector $x = (x_1, x_2, \ldots, x_d)$ in the concatenated feature, L2 normalization is defined as

$$\hat{x} = x/||x||_2 \;= x/\left(\sum_{i=1}^{d} |x_i|^2\right)^{1/2} \tag{3}$$

where $\hat{x}$ specifies the normalized vector, $||x||_2$ specifies the L2 normalization of x, and the number of channels is denoted as d.

Finally, to compress the number of channels within the concatenated feature, we use the pointwise convolution to operate $CF_{\{i=3,4,5\}}$. The final fused feature map $Fmap_{\{i=3,4,5\}}$ can be defined as

$$Fmap_{\{i=3,4,5\}} = \text{Conv}\left(CF_{\{i=3,4,5\}}, o, k, s, m\right) \tag{4}$$

where $\text{Conv}(\cdot)$ specifies the convolution operation, $o = 256$, $k = s = 1$, and $m = 0$.

B. MULTI-SCALE CONTEXTUAL INFORMATION

The fused feature map in the RPN is intended to improve the resolution and semantic information for small traffic sign detection. The contextual information, which is drawn from the neighbourhood of the object proposal outside of the RPN, can provide important cues for object classification (e.g., the surrounding environment or a discriminative portion of an object). In our problem, we hold that it can also benefit small traffic sign classification. Thus, we design a simple method to leverage contextual information for the object proposal; the structure of this method is illustrated in Fig.1.

We denote an object proposal as $p = (p_x, p_y, p_w, p_h)$, where (p_x, p_y) specifies its centre coordinates and (p_w, p_h) specifies its width and height. The contextual region, $\hat{p}_i$, is cropped from the fused feature map $Fmap_{\{i=3,4,5\}}$ at two scales (i.e., $\varphi_1 = 1.2$ and $\varphi_2 = 1.6$) and can be defined as

$$\hat{p}_i = \left\{(p_x, p_y, \varphi_i p_w, \varphi_i p_h) | i = 1,\ 2\right\} \tag{5}$$

Note that in our proposed method, the centre coordinates of each contextual region, $\hat{p}_i$, are the same as those of the given object proposal, p.

Then, the object proposal p and its contextual regions $\hat{p}_i$ are fed into the RoI-pooling layer, which outputs the feature representation vector, f_i. This operation can be defined as

$$f_i = \begin{cases} \text{Pool}(p), & i = 0 \\ \text{Pool}(\hat{p}_i), & i = 1, 2 \end{cases} \tag{6}$$

where $\text{Pool}(\cdot)$ specifies the pooling operation of the RoI-pooling layer. In our problem, each feature representation has a fixed size of 7×7×256.

In addition, due to the different scales of the three feature representations in f_i, we apply the L2 normalization. Then, we concatenate the three feature representation vector into K, which is defined as

$$K = \text{concat}_{\{i=0,1,2\}}\left(L_2(f_i)\right) \tag{7}$$

where $\text{concat}(\cdot)$ specifies the concatenation operation along the channel axis.

Next, a 1×1 convolution is employed to compress the number of channels within the concatenated feature, K, from 7×7×3×256 to 7×7×256. The final fused feature can be defined as

$$\begin{aligned} F &= \text{Conv}(K, o, k, s, m) \\ &= \text{Conv}\left(\text{concat}_{\{i=0,1,2\}}(L_2(f_i)), o, k, s, m\right) \end{aligned} \tag{8}$$

where $o = 7 \times 7 \times 256$, $k = 1$, $s = 1$ and $m = 0$.

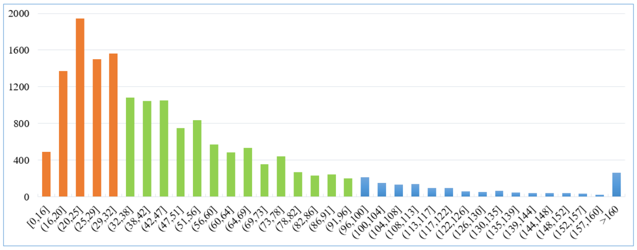

FIGURE 2. Number of instances for each traffic sign size in the Tsinghua-Tencent 100K dataset.

In addition, note that the contextual information is not always helpful because it can potentially introduce invalid background noise to small traffic sign recognition. Thus, the pointwise convolution used in MR-CNN provides a set of parameters for learning the weights between the object features and the contextual features, which weaken the interference caused by invalid background noise.

Last, the final fused feature vector is fed into the fully connected layers (i.e., fc6 and fc7) that branch into two sibling output layers. One layer is used for the object classification according to the estimated probability calculated by softmax. Another layer is implemented to the bounding-box (bbox) regression and outputs the localization of each object.

IV. EXPERIMENTS

A. DATASETS

We evaluate MR-CNN on the Tsinghua-Tencent 100K dataset. Compared with the GTSRB and GTSDB datasets, the images are higher resolution and the traffic sign instances are smaller, more variable, and more numerous. The number of instances in the size range [0, 32] pixels and (32, 96] pixels are approximately 41.6% and 49.1% of the total, respectively. This shows that the Tsinghua-Tencent 100K dataset is more suitable for small traffic sign recognition. Fig.2 provides the number of instances of each traffic sign size.

We select 45 categories of traffic signs, and each category includes more than 100 images. Each instance has a unique label in the dataset. As shown in Fig. 3, some labels (e.g., ‘il*’, ‘pl*’, ‘pm*’, and ‘ph*’) are representatives of the same family. For example, they denote the different height restrictions by replacing ‘*’ in ‘ph*’ with a specific value. The size ratio between the training and testing dataset is 1:2. To ensure that each category has at least 1,000 instances for sample balancing, we use the re-sampling method for the categories with less than 1,000 images in each epoch. The experiments were run on a Linux PC with an Intel Core i7-7700K, 32 GB of memory, and two GeForce GTX 1080 GPUs.

B. DETECTION PERFORMANCE

The traffic sign detection performance of the proposed method is evaluated by the standard detection metrics of recall and precision, which are the same as those used in the previous work [16]. To enable a more intuitive comparison, we also used the F1-measure as an additional metric.

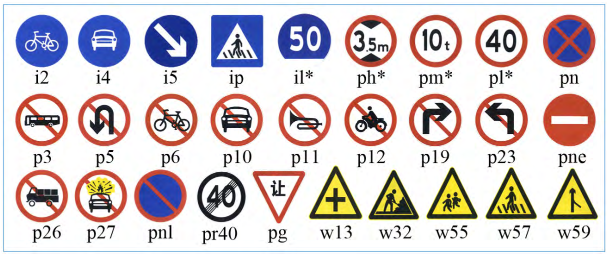

FIGURE 3. The selected 45 categories and each category has more than 100 images.

TABLE 1. Comparison of recognition performance for different size groups. (in %).

Methods	Metrics	Small	Medium	Large
Faster R-CNN	recall	49.8	83.7	91.2
	precision	24.1	65.6	80.8
	F1_measure	32.5	73.6	85.7
SSD	recall	43.4	77.5	86.9
	precision	25.3	67.8	81.5
	F1_measure	32.0	72.3	84.1
FPN	recall	78.6	88.4	90.1
	precision	77.3	86.7	88.0
	F1_measure	77.9	87.5	89.0
Zhu et al.	recall	87.4	93.6	87.7
	precision	81.7	90.8	90.6
	F1_measure	84.5	92.2	89.1
MR-CNN	recall	**89.3**	**94.4**	**88.2**
	precision	**82.9**	**92.6**	**92.0**
	F1_measure	**86.0**	**93.5**	**90.1**

In addition, to evaluate our method for both small and large traffic sign detection, we divided the traffic signs into three size groups: small (sizes∈ (0, 32] pixels), medium (sizes∈ (32, 96] pixels) and large (sizes∈ (96, 200] pixels). In addition, it is worth noting that all the traffic signs used occupy less than 1% of the original image.

To validate the effectiveness of the proposed method, we compare MR-CNN with the original Faster R-CNN [23], SSD [25], the method proposed by Zhu et al. [16], and the feature pyramid network (FPN) [34]. In this comparison, Faster R-CNN and SSD are the representative of two-stage object detection and one-stage object detection, respectively. Zhu et al. achieved state-of-the-art recognition results on the Tsinghua-Tencent 100K dataset. In addition, FPN is well-known multi-scale object detection method that has achieved state-of-the-art performances on the MS COCO and PASCAL VOC datasets. Here, FPN uses the ResNet-50 as its backbone architecture and employs Faster R-CNN as backbone detector.

Table 1 provides a comparison of the detection performance of these five methods on the different traffic sign size groups. The F1-measure obtained by our proposed MR-CNN was 86.0% for small (sizes∈ (0, 32] pixels), 93.5% for medium (sizes∈ (32, 96] pixels) and 90.1% for large (sizes∈ (96, 200] pixels) size group. It outperforms Gudigar et al. [17] by 1.5 points, 1.3 points and 1.0 points, respectively, the

TABLE 2. Comparison of the recognition performance of five methods on 45 categories. (in %).

Methods	Metrics	i2	i4	i5	il100	il60	il80	io	ip	p10	p11	p12	p19	p23	p26	p27
Faster R-CNN	recall	63.4	77.8	81.3	72.6	89.1	78.5	73.7	66.7	63.7	63.3	54.6	74.6	76.5	79.2	83.6
	precision	46.1	45.2	47.8	44.3	58.9	66.4	43.2	41.6	47.2	39.6	61.3	60.3	66.3	51.6	81.2
	F1-measure	53.4	57.2	60.2	55.0	70.9	71.9	54.5	51.2	54.2	48.7	57.8	66.7	71.0	62.5	82.4
SSD	recall	56.7	64.6	74.3	63.5	80.7	65.7	68.0	51.7	60.8	56.2	51.7	64.1	71.0	70.7	81.2
	precision	41.0	50.1	45.8	43.1	61.3	71.3	48.5	48.0	51.3	41.5	57.3	58.9	59.1	53.6	73.5
	F1-measure	47.6	56.4	56.7	51.3	69.7	68.4	56.6	49.8	55.6	47.7	54.4	61.4	64.5	61.0	77.2
FPN	recall	78.2	83.2	87.5	90.1	81.4	91.3	83.5	77.2	88.4	81.5	82.1	88.4	85.9	87.6	84.2
	precision	71.6	86.7	90.4	88.3	90.5	83.1	71.6	82.7	71.4	82.3	85.6	87.8	87.3	82.1	81.7
	F1-measure	74.8	84.9	88.9	89.2	85.7	87.0	77.1	79.9	79.0	81.9	83.8	88.1	86.6	84.8	82.9
Zhu et al.	recall	82.4	91.3	91.3	94.9	86.4	96.9	88.0	78.9	92.0	86.3	89.4	93.9	91.2	92.5	93.6
	precision	75.9	86.8	95.0	94.9	97.6	84.5	78.5	88.6	76.2	88.8	92.2	93.9	91.2	79.6	88.0
	F1-measure	79.0	89.0	93.1	94.9	91.7	90.3	83.0	83.5	83.4	87.5	90.8	93.9	91.2	85.6	90.7
MR-CNN	recall	**80.7**	**88.6**	**92.4**	93.1	**90.3**	**95.5**	**89.6**	**84.1**	90.4	**84.3**	88.6	**94.5**	**89.7**	**88.2**	92.3
	precision	**82.2**	**91.8**	**94.5**	92.2	**94.8**	**87.7**	**80.3**	**87.4**	74.7	**91.2**	90.3	**95.1**	**93.4**	**83.5**	87.0
	F1-measure	**81.4**	**90.2**	**93.4**	92.6	**92.5**	**91.4**	**84.7**	**85.7**	81.8	**87.6**	89.4	**94.8**	**91.5**	**85.8**	89.6

Methods	Metrics	p3	p5	p6	pg	ph4	ph4.5	ph5	pl100	pl120	pl20	pl30	p140	pl5	pl50	pl60
Faster R-CNN	recall	54.6	86.3	57.2	86.2	59.3	81.4	47.8	88.2	78.7	44.4	62.5	69.3	68.5	63.6	66.4
	precision	49.9	59.6	78.7	81.7	69.2	59.6	52.5	69.1	68.1	53.6	46.1	54.2	54.6	41.3	54.9
	F1-measure	52.1	70.5	66.2	83.9	63.9	68.8	50.0	77.5	73.0	48.6	53.1	60.8	60.8	50.1	60.1
SSD	recall	58.3	78.7	58.3	78.5	62.8	76.3	52.6	82.6	63.5	51.2	52.8	58.3	61.1	57.5	61.0
	precision	51.7	61.0	71.5	73.2	57.1	62.5	47.2	71.9	74.7	55.6	57.2	51.6	53.5	51.3	52.7
	F1-measure	54.8	68.7	64.2	75.8	59.8	68.7	49.8	76.9	68.6	53.3	54.9	54.7	57.0	54.2	56.5
FPN	recall	83.8	83.4	80.3	83.2	80.7	84.3	83.1	88.6	92.5	78.7	83.5	91.3	88.4	83.6	90.5
	precision	73.5	90.1	84.6	84.5	77.8	82.6	71.0	90.2	90.3	81.6	78.2	84.7	81.5	86.3	78.1
	F1-measure	78.3	86.6	82.4	83.8	79.2	83.4	76.6	89.4	91.4	80.1	80.8	87.9	84.8	84.9	83.8
Zhu et al.	recall	89.7	95.8	89.7	90.9	75.7	85.0	71.8	98.1	93.1	80.3	89.2	89.1	87.1	91.5	80.3
	precision	74.3	87.9	74.5	90.9	82.4	85.0	87.5	90.6	98.8	90.0	89.2	92.1	91.1	85.4	95.6
	F1-measure	81.3	91.7	81.4	90.9	78.9	85.0	78.9	94.2	95.9	84.9	89.2	90.6	89.1	88.3	87.3
MR-CNN	recall	**88.4**	**92.1**	**88.9**	**91.5**	**78.7**	**88.0**	**75.9**	**93.9**	94.2	**85.3**	**91.7**	**91.4**	85.3	**92.2**	**83.7**
	precision	**76.6**	**93.6**	**76.7**	**93.2**	**80.5**	**84.2**	**82.8**	**94.7**	91.4	**90.6**	**90.8**	**90.5**	87.6	**86.5**	**91.8**
	F1-measure	**82.1**	**92.8**	**82.4**	**92.3**	**79.6**	**86.1**	**79.2**	**94.3**	92.8	**87.9**	**91.2**	**90.9**	86.4	**89.3**	**87.6**

Methods	Metrics	pl70	pl80	pm20	pm30	pm55	pn	pne	po	pr40	w13	w32	w55	w57	w59	wo
Faster R-CNN	recall	70.3	69.6	65.2	63.7	79.8	78.1	85.5	63.6	98.0	72.5	59.1	64.6	78.1	79.7	51.5
	precision	63.7	54.9	60.8	67.1	61.2	52.3	51.9	35.1	74.8	35.2	56.7	41.5	49.2	40.1	39.2
	F1-measure	66.8	61.4	62.9	65.4	69.3	62.6	64.6	45.2	84.8	47.4	57.9	50.5	60.4	53.4	44.5
SSD	recall	68.4	67.3	63.5	65.2	63.6	64.2	81.7	51.6	87.2	68.6	58.4	63.1	71.5	75.2	44.6
	precision	55.2	51.0	61.7	61.4	68.3	50.1	53.1	41.0	71.3	41.0	51.5	51.3	51.9	46.5	36.9
	F1-measure	61.1	58.0	62.6	63.2	65.9	56.3	64.4	45.7	78.5	51.3	54.7	56.6	60.1	57.5	40.4
FPN	recall	81.6	88.9	90.3	88.6	77.3	90.8	88.6	74.3	88.5	81.6	78.6	88.2	86.8	77.4	47.5
	precision	83.2	82.5	88.4	86.2	87.6	88.4	87.1	71.7	85.9	77.3	72.1	70.4	78.3	67.5	47.0
	F1-measure	82.4	85.6	89.3	87.4	82.1	89.6	87.8	73.0	87.2	79.4	75.2	78.3	82.3	72.1	47.2
Zhu et al.	recall	90.1	93.4	85.7	90.6	94.7	89.1	91.0	67.6	93.7	80.6	55.6	55.0	80.3	66.7	28.9
	precision	88.9	87.7	93.3	87.9	72.0	93.1	93.2	78.5	92.2	80.6	95.0	97.1	90.7	81.6	45.8
	F1-measure	89.5	90.5	89.3	89.2	81.8	91.1	92.1	72.6	92.9	80.6	70.1	70.2	85.2	73.4	35.4
MR-CNN	recall	88.6	92.3	**88.4**	**91.8**	**93.5**	88.2	92.5	**70.6**	**92.8**	**83.2**	68.6	63.1	**84.6**	**74.5**	42.8
	precision	84.5	86.3	**92.7**	**88.9**	**78.1**	90.4	89.2	**77.5**	**93.3**	**82.5**	82.3	82.5	**89.3**	**75.1**	41.6
	F1-measure	86.5	89.2	**90.5**	**90.3**	**85.1**	89.3	90.8	**73.9**	**93.0**	**82.8**	74.8	71.5	**86.9**	**74.8**	42.2

original Faster R-CNN by 53.5 points, 19.9 points and 4.4 points, respectively. It also outperforms other object detection frameworks (e.g., SSD and FPN). It demonstrates that the MR-CNN can accurately recognize small traffic signs as well as medium or large ones.

Table 2 provides the detailed detection metrics for each traffic sign category for five methods which show that MR-CNN achieves the best performance in most categories. The experimental results demonstrate that our proposed fused feature map and multi-scale contextual information can effectively enhance feature representation power and boost the performance of small traffic sign detection.

Fig.4 shows the comparisons of detection performance about different methods for small, medium and large signs by using the precision-recall curves. In the object detection, the precision-recall curve is widely used to evaluate the detection performance of the method. According to the area enclosed by the curve, X axis and Y axis, the larger the area, the better the detection performance. Thus, Fig.4 demonstrates that our proposed method consistently outperforms to other methods on different traffic sign sizes, especially when the sign size is small.

Fig.5 shows the partially visualized detection results on the testing dataset. As can be observed, each traffic sign instance is quite small and occupies less than 1% of the whole scene; nonetheless, our approach recognized them accurately. To facilitate observe, the recognition results are enlarged and placed at the bottom left or bottom right of the traffic scene.

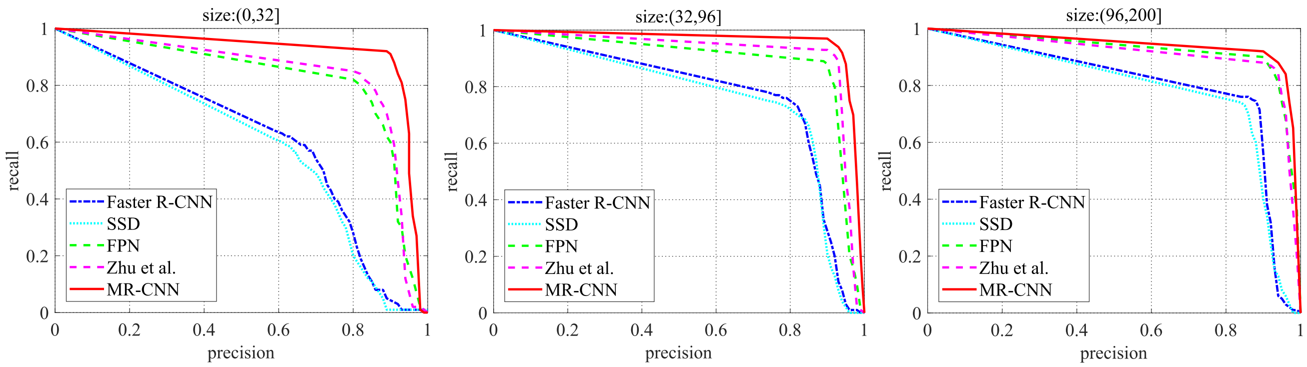

FIGURE 4. Comparisons of the detection performance about five methods for three size groups.

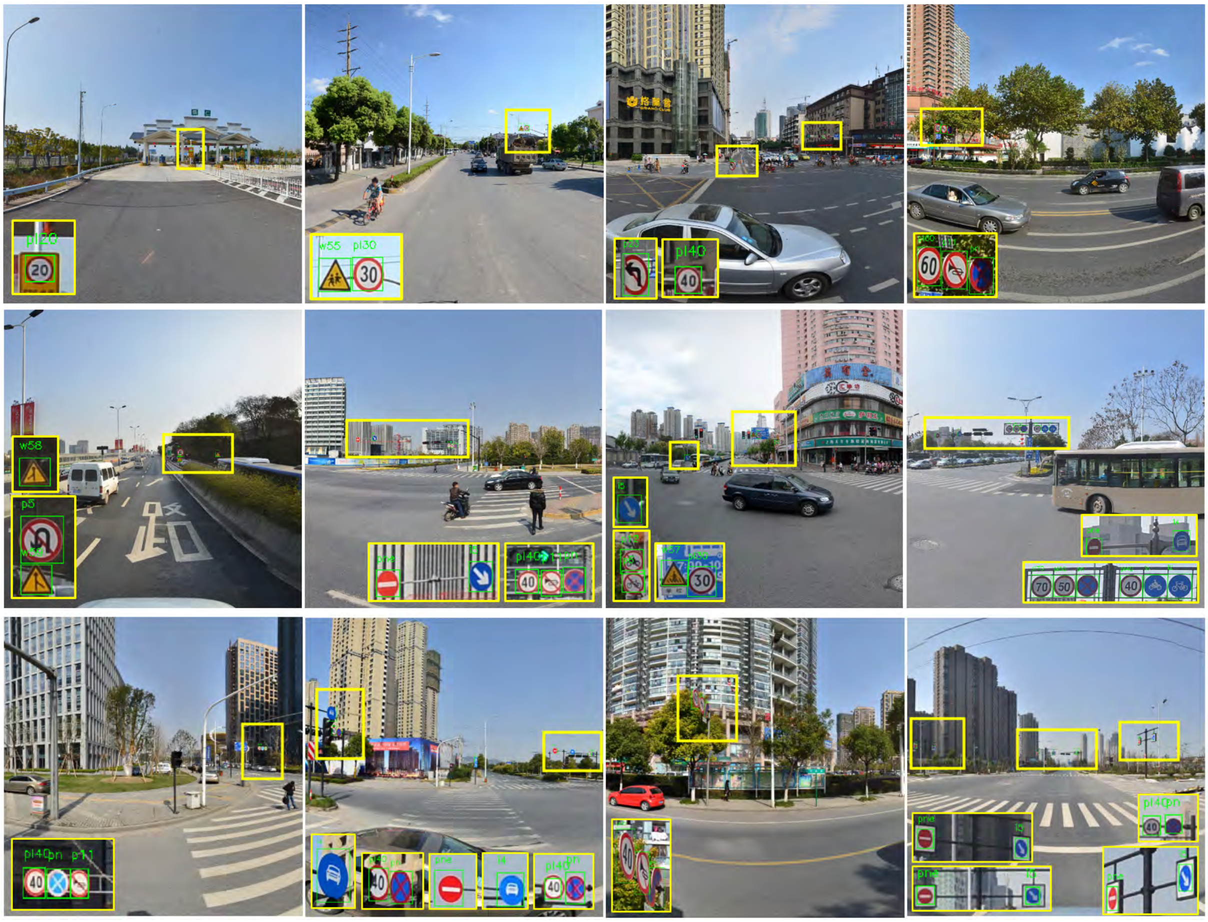

FIGURE 5. The detection results of the proposed methods. 'pl40', 'pne', and the other strings are the traffic sign labels.

In addition, the traffic sign in the scene shown at the third row of the third column in Fig.5 is seriously deformed due to the viewpoint; nonetheless, our model can still recognize it correctly.

C. ABLATION ANALYSIS

Small traffic sign detection is a challenging problem in computer vision because it is more difficult to localize due to their low resolution. For an object detection framework, when the targets do not appear in the region proposal set, the subsequent classification will be invalid. To highlight the impact of the proposed fused feature map and the advantage gained by the architecture, we compare the detection performance of region proposals for different feature maps in Table 3.

We can observe that D_4 and D_5 outperform C_4 and C_5, respectively. This demonstrates that the feature maps

TABLE 3. Detection performance regarding different feature maps for 500 region proposals (IoU=0.5).

Feature map	Proposal recall	Detection mAP
C_5	0.648	0.532
C_4	0.722	0.591
C_3	0.746	0.617
D_5	0.693	0.584
D_4	0.738	0.605
$F_{\{i=3,4\}}$	0.833	0.648
$F_{\{i=3,4,5\}}$	**0.894**	**0.715**
$F_{\{i=2,3,4\}}$	0.860	0.671
$F_{\{i=1,2,3\}}$	0.827	0.629

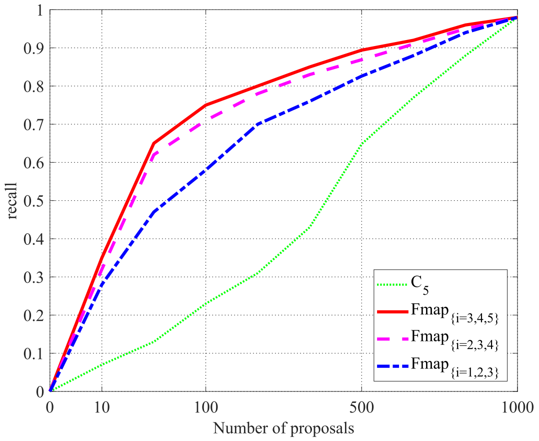

FIGURE 6. Region proposal performance of feature maps from different convolution layers.

constructed by up-sampling the deeper convolution layer with a deconvolution operation are more suitable for small traffic sign detection than are the original maps. However, their performances are still worse than that of C_3. We consider that D_4 and D_5 lose some detail information due to the continuous down-sampling and up-sampling operations.

Furthermore, combining the deeper layers with the shallower layer, e.g., $F_{\{i=3,4\}}$, $F_{\{i=3,4,5\}}$, $F_{\{i=2,3,4\}}$, and $F_{\{i=1,2,3\}}$, outperforms a single convolution layer because there more semantic information exists in the deeper layers. In all combining strategies, our proposed fused feature map, $F_{\{i=3,4,5\}}$, achieves the best performance. The recall and detection accuracy values improved from 64.8% to 89.4% and from 53.2% to 71.3%, respectively, compared with the original feature map, C_5, in the Faster R-CNN.

Fig.6 shows a comparison of the performance regarding the region proposals between the original feature map C_5 in the Faster R-CNN and the different combining strategies. These results demonstrate that our proposed fused feature map not only generates fewer region proposals but also achieves higher recall values.

V. CONCLUSION AND FUTURE WORK

This paper proposes a small traffic sign detection framework named MR-CNN that employs the two-stage fusion strategy. The MR-CNN framework integrates multiple levels of convolution feature and multiple levels of contextual information. At the detection stage, the region proposals are generated from the fused feature map with sufficient information. We design the fusion strategy of different convolution layers by using deconvolution, normalization, and compression. At the classification stage, we construct a fused feature for the fully connected layer and leverage the multi-scale contextual regions to exploit the surrounding information for a given object proposal. The fused feature map is focused on improving the resolution and semantic information of small traffic signs, while the fused feature provides more discriminative representations of the contextual regions. The final experimental results show our method's superiority for detecting small signs, and it achieved state-of-the-art performance compared with other methods. In addition, note that the hard negative samples are important for efficient training and detection accuracy. In future work, we plan to focus on designing an efficient training algorithm to differentiate hard negative samples from easy positive samples.

REFERENCES

[1] J. F. Khan, S. M. A. Bhuiyan, and R. R. Adhami, "Image segmentation and shape analysis for road-sign detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 83–96, Mar. 2011. doi: 10.1109/TITS.2010.2073466.

[2] S. K. Berkaya, H. Gunduz, O. Ozsen, C. Akinlar, and S. Gunal, "On circular traffic sign detection and recognition," *Expert Syst. Appl.*, vol. 48, pp. 67–75, Apr. 2016. doi: 10.1016/j.eswa.2015.11.018.

[3] M. Mathias, R. Timofte, R. Benenson, and L. Van Gool, "Traffic sign recognition—How far are we from the solution," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Aug. 2013, pp. 1–8. doi: 10.1109/IJCNN.2013.6707049.

[4] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2022–2031, Jul. 2016. doi: 10.1109/TITS.2015.2482461.

[5] S. Khalid, N. Muhammad, and M. Sharif, "Automatic measurement of the traffic sign with digital segmentation and recognition," *IET Intell. Transp. Syst.*, vol. 13, no. 2, pp. 269–279, 2019. doi: 10.1049/iet-its.2018.5223.

[6] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 920–933, Apr. 2017. doi: 10.1109/TCYB.2016.2533424.

[7] X. Lu, Y. Wang, X. Zhou, Z. Ling, and Z. Zhang, "Traffic sign recognition via multi-modal tree-structure embedded multi-task learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 960–972, Apr. 2018. doi: 10.1109/TITS.2016.2598356.

[8] F. Zaklouta and B. Stanciulescu, "Real-time traffic-sign recognition using tree classifiers," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1507–1514, Apr. 2012. doi: 10.1109/TITS.2012.2225618.

[9] C. Liu, F. Chang, and C. Liu, "Occlusion-robust traffic sign detection via cascaded colour cubic feature," *IET Intell. Transp. Syst.*, vol. 10, no. 5, pp. 354–360, 2015. doi: 10.1049/iet-its.2015.0099.

[10] 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. doi: 10.1109/TITS.2018.2843815.

[11] 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. doi: 10.1109/TITS.2018.2801560.

[12] Z. Zhu, J. Lu, R. R. Martin, and S. Hu, "An optimization approach for localization refinement of candidate traffic signs," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3006–3016, Nov. 2017. doi: 10.1109/TITS.2017.2665647.

[13] 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. doi: 10.1109/TITS.2017.2714691.

[14] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu, "Traffic sign detection based on convolutional neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Aug. 2013, pp. 1–7. doi: 10.1109/IJCNN.2013.6706811.

[15] Y. Zhu, C. Zhang, D. Zhou, X. Wang, X. Bai, and W. Liu, "Traffic sign detection and recognition using fully convolutional network guided proposals," *Neurocomputing*, vol. 214, pp. 758–766, Nov. 2016. doi: 10.1016/j.neucom.2016.07.009.

[16] 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.*, Jun. 2016, pp. 2110–2118. doi: 10.1109/CVPR.2016.232.

[17] A. Gudigar, S. Chokkadi, U. R. Acharya, and U. Raghavendra, "An efficient traffic sign recognition based on graph embedding features," *Neural Comput. Appl.*, vol. 31, no. 2, pp. 395–407, 2019. doi: 10.1007/s00521-017-3063-z.

[18] 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. IEEE Int. Joint Conf. Neural Netw.*, Aug. 2013, pp. 1–8. doi: 10.1109/IJCNN.2013.6706807.

[19] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: A multi-class classification competition," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul./Aug. 2011, pp. 1453–1460. doi: 10.1109/IJCNN.2011.6033395.

[20] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 385–400.

[21] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1925–1934. doi: 10.1109/CVPR.2017.549.

[22] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 528–537. doi: 10.1109/CVPR.2018.00062.

[23] 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. doi: 10.1109/TPAMI.2016.2577031.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[25] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[26] X. Z. Chen, K. Kundu, Y. Zhu, S. Fidle, R. Urtasun, and 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, May 2018. doi: 10.1109/TPAMI.2017.2706685.

[27] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 951–959. doi: 10.1109/CVPR.2017.166.

[28] Q. Chen, X. Meng, W. Li, X. Fu, X. Deng, and J. Wang, "A multi-scale fusion convolutional neural network for face detection," in *Proc. IEEE Conf. Syst., Man, Cybern.*, Oct. 2017, pp. 1013–1018. doi: 10.1109/SMC.2017.8122743.

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

[30] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969. doi: 10.1109/ICCV.2017.322.

[31] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2874–2883. doi: 10.1109/CVPR.2016.314.

[32] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 845–853. doi: 10.1109/CVPR.2016.98.

[33] H. Zhang, K. Wang, Y. Tian, C. Gou, and 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, Sep. 2018. doi: 10.1109/TVT.2018.2843394.

[34] T.-Y. Lin, P. Dollar, R. Girshick, B. Hariharan, S. Belongie, and K. He, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125. doi: 10.1109/CVPR.2017.106.

ZHIGANG LIU was born in Jiaohe, China, in 1979. He received the Ph.D. degree from Northeast Petroleum University, Daqing, China, in 2016, where he has been an Associate Professor with the School of Computer and Information Technology, since 2012. From 2018 to 2019, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, National University of Singapore. His main research interests include machine learning, computer vision, and their applications in the intelligent transportation systems. He is currently a member of the Association for Computing Machinery and China Computer Federation. He has been granted several scholarships and funding projects in his academic research. He is the Reviewer of several international journals and conferences.

JUAN DU received the B.S. and M.S. degrees in computer science and technology from Northeast Petroleum University, Daqing, China, in 2003 and 2009, respectively, where she is currently an Associate Professor with the School of Computer and Information Technology. Her research interests mainly include computer vision and image processing.

FENG TIAN was born in Anda, China, in 1980. He received the Ph.D. degree from Beihang University, China, in 2012. From 2015 to 2016, he was a Visiting Scholar with the Department of Computer Science, National University of Singapore. He is currently a Professor with the School of Computer and Information Technology, Northeast Petroleum University. His research interests include computer vision and pattern recognition.

JIAZHENG WEN is an undergraduate and currently pursuing the B.E. degree with the University of Nottingham Ningbo, China, and will graduate, in 2021. His major is computer science for artificial intelligence. His perspective for the future is artificial intelligence territory and software construction for mobile electronic equipment. His research interest is the object detection-based deep learning.