Hindawi
Journal of Sensors
Volume 2020, Article ID 8844348, 13 pages
https://doi.org/10.1155/2020/8844348

Research Article

Fast Traffic Sign Detection Approach Based on Lightweight Network and Multilayer Proposal Network

Hoanh Nguyen

Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

Correspondence should be addressed to Hoanh Nguyen; nguyenhoanh@iuh.edu.vn

Received 3 March 2020; Revised 15 May 2020; Accepted 3 June 2020; Published 19 June 2020

Academic Editor: Bin Gao

Vision-based traffic sign detection plays a crucial role in intelligent transportation systems. Recently, many approaches based on deep learning for traffic sign detection have been proposed and showed better performance compared with traditional approaches. However, due to difficult conditions in driving environment and the size of traffic signs in traffic scene images, the performance of deep learning-based methods on small traffic sign detection is still limited. In addition, the inference speed of current state-of-the-art approaches on traffic sign detection is still slow. This paper proposes a deep learning-based approach to improve the performance of small traffic sign detection in driving environments. First, a lightweight and efficient architecture is adopted as the base network to address the issue of the inference speed. To enhance the performance on small traffic sign detection, a deconvolution module is adopted to generate an enhanced feature map by aggregating a lower-level feature map with a higher-level feature map. Then, two improved region proposal networks are used to generate proposals from the highest-level feature map and the enhanced feature map. The proposed improved region proposal network is designed for fast and accuracy proposal generation. In the experiments, the German Traffic Sign Detection Benchmark dataset is used to evaluate the effectiveness of each enhanced module, and the Tsinghua-Tencent 100K dataset is used to compare the effectiveness of the proposed approach with other state-of-the-art approaches on traffic sign detection. Experimental results on Tsinghua-Tencent 100K dataset show that the proposed approach achieves competitive performance compared with current state-of-the-art approaches on traffic sign detection while being faster and simpler.

1. Introduction

Vision-based traffic sign recognition plays an essential role in intelligent transport systems such as an automated driving system and an advanced driver assistance system. A traffic sign recognition system normally includes two stages: traffic sign detection and traffic sign recognition. Traffic sign detection takes images captured from a camera to locate exactly traffic sign regions, while traffic sign recognition classifies each traffic sign into a corresponding class. Detected traffic signs from a detection stage are utilized as inputs for a recognition stage. Thus, the accuracy of traffic sign detection has a dramatic effect on the accuracy of the whole system. Many approaches have been proposed to detect traffic sign [1]. Traditional methods [2–7] are usually based on hand-crafted features such as color, texture, edge, and other low-level features to detect traffic sign in an image. In driving environment, due to the diversity of the traffic sign appearance, the occlusion of traffic sign by other objects, and the effect of lighting conditions, traditional methods for traffic sign detection showed poor performance.

With the fast development of deep learning recently, many deep learning-based approaches for traffic sign detection [8–15] have been proposed and showed outstanding performance compared with traditional approaches. Deep learning-based methods for traffic sign detection first create traffic sign candidates, and classifiers are then used to identify traffic sign and background class. Although deep learning-based methods for traffic sign detection performed well in difficult driving environments, the performance of these methods is still low in challenging conditions as discussed below:

(i) The size of traffic signs is quite small in traffic scene images, and small traffic sign detection is much more

challenging than large traffic sign detection. Most recent methods for traffic sign detection are focused on a large traffic sign. Thus, the performance of these methods is limited on small traffic sign detection

(ii) The inference speed of the traffic sign detection method is a large concern in driving environments where vehicles are unlikely to be equipped with high-end hardware components. Most recent methods for traffic sign detection are implemented on high-end systems. Thus, it is necessary to build a faster framework for traffic sign detection in driving environments

To tackle these issues, this paper designs a deep learning-based framework for fast and efficient traffic sign detection. The proposed framework uses ESPNetv2 network [16] as the based convolution layers for increasing the processing speed. To enhance the performance of the proposed framework on small traffic sign detection, a deconvolution module is used to create an enhanced feature map from convolution feature maps generated by the base network. In the proposal generation process, two improved region proposal networks (RPNs) are adopted to generate both large and small traffic sign proposals from input convolution layers. The improved region proposal network is designed based on the original region proposal network [17] for fast and efficient proposal generation. The proposed approach is evaluated on the German Traffic Sign Detection Benchmark dataset and Tsinghua-Tencent 100K dataset. The results show that the proposed approach achieves competitive performance compared with current state-of-the-art approaches while being faster and simpler. The main contributions of this paper can be summarized as follows:

(i) For the purpose of fast and efficient traffic sign detection, this paper adopts a lightweight deep network as the base network. The proposed base network substantially reduces the computational costs of the whole framework

(ii) To generate proposals, this paper designs a novel RPN based on the original RPN to increase the inference speed and detection accuracy of the proposed framework. In the improved RPN, 1×1 convolution layer is first used after the input feature maps to reduce the number of parameters. Then, dilated convolution is used to enlarge the receptive field, thus including more information from other areas to help recognize the boundaries of objects and maintaining the number of parameters. The proposed RPN can be adopted in other object detection tasks such as pedestrian detection and vehicle detection

(iii) To improve the performance of the proposed network on small traffic sign detection, a deconvolution module is used to generate enhanced feature map, which contains more discriminative representation for small traffic signs. The deconvolution module enhances the shallow feature map from a lower feature layer with a deeper feature map from a higher feature layer

(iv) To further improve the detection performance of the proposed network, a region proposal generation module including two improved RPNs is designed to produce a set of good proposals from the enhanced feature map and the highest-level feature map

(v) Experimental results on public datasets show that the proposed approach achieves comparable accuracy compared with other state-of-the-art approaches on traffic sign detection while being faster and simpler. The proposed framework can be applied in real-time intelligent transport systems

The remaining of this paper is organized as follows. Section 2 reviews the related work. Section 3 details the proposed framework. Section 4 provides the experimental results and comparison between the proposed method and other methods on public datasets. Finally, the conclusions are drawn in Section 5.

2. Related Work

2.1. Traffic Sign Detection. Vision-based traffic sign detection can be divided into two groups: traditional method and deep learning-based method. Traditional methods are usually based on hand-crafted features such as color and shape to detect traffic signs. Bahlmann et al. [2] proposed using a set of Haar wavelet features obtained from AdaBoost training to detect traffic signs and Bayesian generative modeling for classification. Salti et al. [3] proposed an approach based on interest region extraction rather than sliding window detection. In addition, the SVM classifier which takes histogram of oriented gradients in the regions of interest as the input feature is used for classification. In [4], the authors utilized circle detection algorithm and an RGB-based color thresholding technique to detect traffic sign. For traffic sign recognition, an ensemble of features including histogram of oriented gradients, local binary patterns, and Gabor features is employed within a support vector machine classification framework. Timofte et al. [5] proposed to combine 2D and 3D techniques to improve results of traffic sign detection and recognition. In [6], a localization refinement approach for traffic sign candidates was proposed. Color and shape priors are utilized in an iterative optimization approach to accurately segment the traffic signs as foreground objects. In [7], the authors proposed a system with three working stages: image preprocessing, detection, and recognition. The proposed system demonstrated that using RGB color segmentation and shape matching followed by a support vector machine classifier leads to promising results. Manual features, such as histogram of oriented gradients and enhancement information of typical color or geometric shape, tend to fail in many difficult driving environments. Thus, traditional methods for traffic sign detection showed poor performance.

Recently, with the fast development of deep convolutional neural networks (CNN), many methods for traffic sign detection based on deep CNN have been proposed and showed better performance compared with traditional methods. Wu et al. [10] proposed to use support vector machines to transform the original image into the gray scale image at first stage. A convolutional neural network with fixed and learnable layers was then used for detection and recognition traffic signs. In [11], the authors proposed a novel framework with two deep learning components, including fully convolutional network-guided traffic sign proposals and deep CNN for object classification. Zhu et al. [9] proposed a fully convolutional network which performs both detection and classification simultaneously. Liu et al. [8] proposed the multiscale region-based convolutional neural network which uses a multiscale deconvolution operation to upsample the features of the deeper convolution layers and concatenates them to those of the shallow layer to construct the fused feature map. In [12], a detector based on faster R-convolutional neural networks and the structure of MobileNet was designed and implemented for detecting traffic sign in driving environments. Moreover, color and shape information has been used to refine the localizations of small traffic signs. Yang et al. [13] designed a novel detection module based on traffic sign proposal extraction and classification built upon a color probability model and a color histogram of oriented gradient. Then, a convolutional neural network is used to further classify the detected signs into their subclasses within each superclass. Li et al. [14] proposed a model with three stages: channel-wise coarse feature extraction is first adopted to produce coarse feature maps with much information loss, channel-wise hierarchical feature refinement is then used to refine hierarchical features, and hierarchical feature map fusion is used at the final stage to fuse hierarchical feature maps to generate the final traffic sign saliency map. In [15], the authors introduced an attention network to Faster R-CNN for finding potential region of interest and roughly classifying them into three categories according to color feature of the traffic signs. Then, the fine region proposal network is designed to produce the final region proposals from a set of anchors per feature map location. Zhang et al. [18] proposed a cascaded R-CNN to obtain the multiscale features in pyramids and a multiscale attention method to obtain the weighted multiscale features by dot product and softmax to highlight the traffic sign features and improve the accuracy of the traffic sign detection. In [19], the authors proposed novel lightweight networks that can obtain higher recognition precision while preserving less trainable parameters in the models. In addition, a new module combines two streams of feature channels with dense connectivity was introduced to improve the accuracy of traffic sign recognition.

2.2. Small Object Detection. In a traffic scene image, traffic signs usually occupy only a small portion of the entire image. In addition, small object detection is much more challenging than large object detection. Thus, many deep learning frameworks have been proposed to enhance the information representation of small objects in convolution feature maps. Zhang et al. [20] proposed a model that consists of a region proposal network that generates candidate object regions and an object detection network that incorporates multiscale features and global information. In [21], the authors proposed a unified deep CNN for fast multiscale object detection. In this framework, the detection is performed at various intermediate network layers to enable the detection of all object scales. Cao et al. [22] proposed an improved algorithm based on faster region-based CNN for small object detection. An improved loss function based on intersection over union, the multiscale convolution feature fusion, and the improved nonmaximum suppression algorithm is introduced to enhance the performance for small object detection. Li et al. [23] proposed a new perceptual generative adversarial network model that improves small object detection through narrowing representation difference of small objects from the large ones. Wei et al. [24] proposed three enhancements for CNN-based visual object detection for advanced driving assistance systems, including using deconvolution and fusion of CNN feature maps to create enhanced feature maps, adopting soft nonmaximal suppression to address the object occlusion challenge, and setting anchor boxes properly for better object matching and localization. In [25], Lin et al. exploited the inherent multiscale, pyramidal hierarchy of deep convolutional networks to construct feature pyramids with marginal extra cost.

3. Proposed Framework

Figure 1 illustrates the overall framework of the proposed approach. The base network based on ESPNetv2 network [16] first takes input image to generate convolution feature maps. To improve the ability of the proposed framework on detecting of small traffic signs, a deconvolution module is used to aggregate an output feature map at layer 2 with an output feature map at layer 3 to create an enhanced feature map. Then, two improved region proposal networks are adopted to generate proposals from the highest-level convolution feature map and the enhanced convolution feature map. The improved region proposal network includes a 1×1 convolution layer to reduce the number of parameters in the subsequent convolutional layers and a 3×3 dilated convolution to enlarge the receptive field, thus improving the detection accuracy and the inference speed of the proposal generation stage. In the detection network, a region of interest pooling layer is adopted to adjust the size of proposals to fixed size feature maps, and fully connected layers are used for classifying proposals and regressing the bounding box of proposals. In the following sections, the proposed approach is explained in detail.

3.1. The Base Network. Most of deep learning-based approaches for traffic sign detection are focused on detection accuracy. In driving environment, apart from the detection accuracy, the inference speed is also a large concern. Moreover, vehicles in driving environments are unlikely to be equipped with high-end graphic cards as powerful as used in research environments. Thus, it is necessary to build a

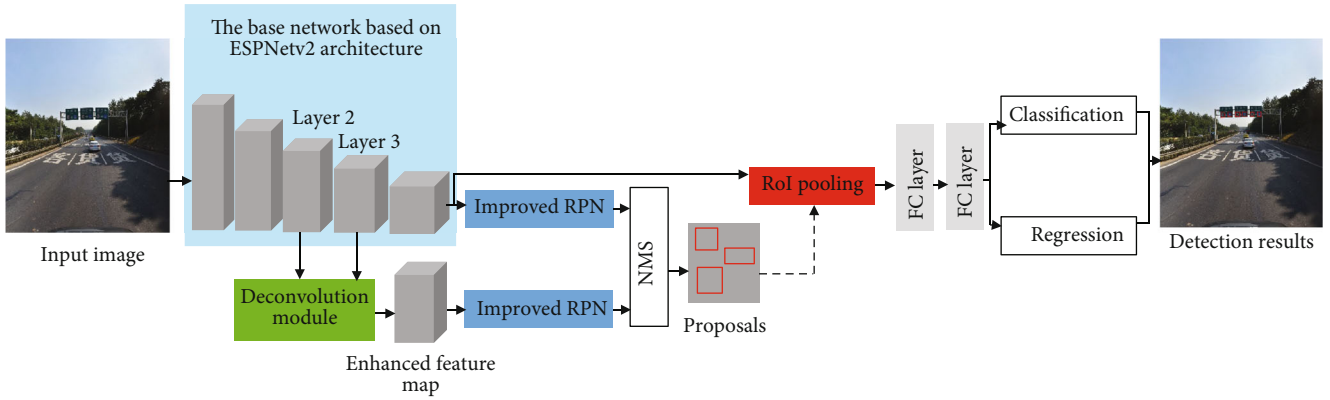

Figure 1: The overall framework of the proposed approach.

Table 1: Performance comparison of different efficient networks on the ImageNet validation set.

Network	# parameters	FLOPs	Top-1 accuracy
MobileNetv1 [27]	2.59 M	325 M	68.4
MobileNetv2 [29]	3.47 M	300 M	71.8
ShuffleNetv1 [30]	3.46 M	292 M	71.5
ESPNetv2	3.49 M	284 M	72.1

faster network for traffic sign detection in driving environments. In [26], Liu et al. showed that about 80% of the forward time is spent on the base convolution layers. Thus, using a faster base network can greatly improve the inference speed of the framework. ESPNetv2 network [16] is a fast and efficient network which uses depth-wise dilated separable convolutions instead of depth-wise separable convolutions [27] to learn representations from a large effective receptive field. Table 1 provides a performance comparison between ESPNetv2 and state-of-the-art efficient networks on the ImageNet 1000-way classification dataset [28]. As shown in Table 1, ESPNetv2 achieves the best performance with the smallest computational budgets (computational budget = 284 million FLOPs). In addition, ESPNetv2 has similarity in the number of parameters compared with ShuffleNetv1 and MobileNetv2 while being more accuracy. Thus, ESPNetv2 architecture is adopted as the base network in this paper for fast and efficient traffic sign detection.

Supposing the resolution of input images is 224×224, the architecture of the ESPNetv2 used in this paper is illustrated in Figure 2 with a kernel size, filter number, and output size. The ESPNetv2 network replaces standard convolutions by extremely efficient spatial pyramid of depth-wise dilated separable convolutions (EESP units) and strided extremely efficient spatial pyramid of depth-wise dilated separable convolutions (strided EESP units). At each layer, ESPNetv2 repeats the EESP units several times to increase the depth of the network. Figure 3 illustrates the structure of EESP unit (a) and strided EESP unit (b). The EESP unit first projects the high-dimensional input feature map into a low-dimensional

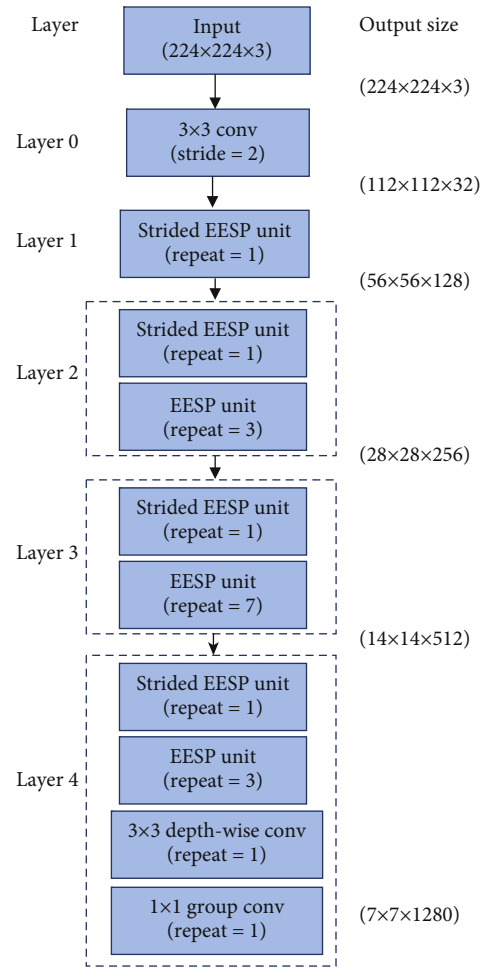

Figure 2: The structure of the ESPNetv2 network used in this paper.

space using group point-wise convolutions and then learns the representations in parallel using depth-wise dilated separable convolutions with different dilation rates. Different dilation rates in each branch allow the EESP unit to learn the representations from a large effective receptive field. With group point-wise and depth-wise dilated

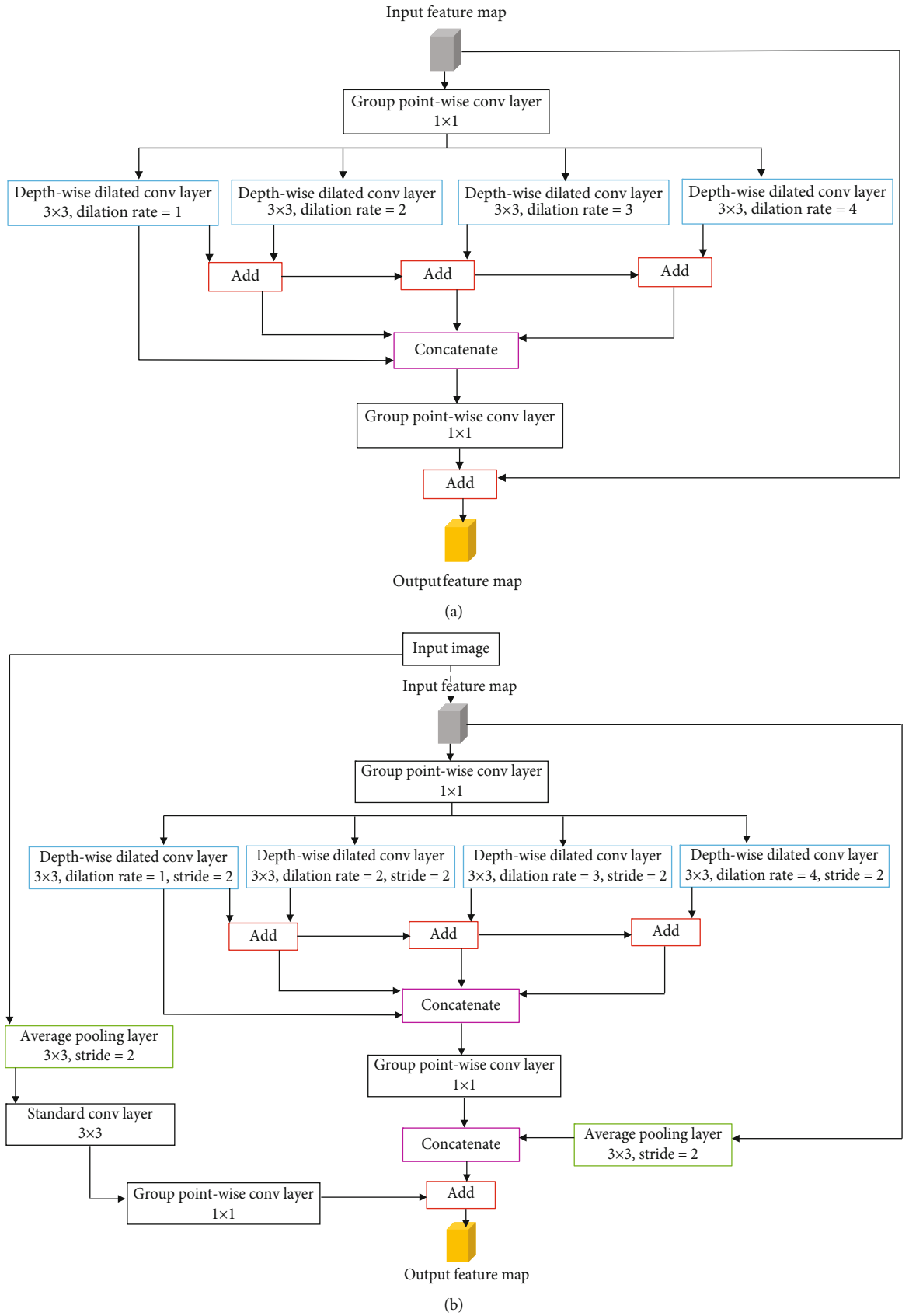

Figure 3: The structure of EESP unit (a) and strided EESP unit (b). ESPNetv2 repeats the EESP units several times to increase the depth of the network.

separable convolutions, the total complexity of the EESP unit is reduced [16]. On the other hand, the strided EESP unit is used to learn representations efficiently at multiple scales. In the strided EESP unit, depth-wise dilated convolutions are replaced with their strided counterpart. In addition, an average pooling layer with stride 2 is added in the strided EESP unit as shown in Figure 3(b). To better encode spatial relationships and learn representations efficiently, an efficient long-range shortcut connection between the input image and the current downsampling unit is added in the strided EESP unit. This connection first downsamples the image to the same size as that of the feature map by repeating a 3×3 average pooling layer and then learns the representations using a stack of two convolutions. The first convolution is a standard 3×3 convolution that learns the spatial representations while the second convolution is a 1×1 point-wise convolution that learns linear combinations between the input and projects it to a high-dimensional space.

Features from the highest-level convolutional layer of the ESPNetv2 architecture could better describe the characteristics of large-scale traffic signs in an image, so the highest-level convolutional layer of the ESPNetv2 network is adopted to generate proposals to ensure the effectiveness of the method for large-scale traffic sign detection. For small-scale traffic signs, features from the larger scales (lower CNN layers) could better describe the characteristics of small-scale traffic signs. However, shallow feature maps from the low layers of feature pyramid inherently lack fine semantic information for object recognition. Thus, this paper adds a deconvolutional module, which fuses the feature map at layer 2 and layer 3 of the ESPNetv2 network, to generate enhanced feature map to better describe the characteristics of small-scale traffic signs. With this deconvolutional module, the semantics from higher layers can be conveyed into lower layers to increase the representation capacity. Details of the deconvolutional module will be explained in the next section.

3.2. Deconvolution Module. Deconvolution module has shown to be helpful for small object detection in DSSD [31]. To enhance a shallow feature map from a lower feature layer with a deeper feature map from a higher feature layer and improve detection performance on small traffic signs, this paper adds a deconvolution module to aggregate the output feature at layer 2 with output feature at layer 3. With a deconvolution module, the semantics from higher feature layer can be conveyed into a lower feature layer to increase the representation capacity. Figure 4 illustrates the architecture of the deconvolution module used in this paper. As shown in Figure 4, a 3×3 convolution layer is first connected with a lower-level feature map extracted after layer 2. For the deconvolution branch, a 2×2 deconvolution layer is used followed by a 3×3 convolution layer to upsample the corresponding higher-level feature map. A batch normalization layer is added after each convolution layer. The higher-level feature map is extracted after layer 3. The deconvolution layer is applied to enlarge the feature map size in order to match the size of the lower-level feature map. Notably, the

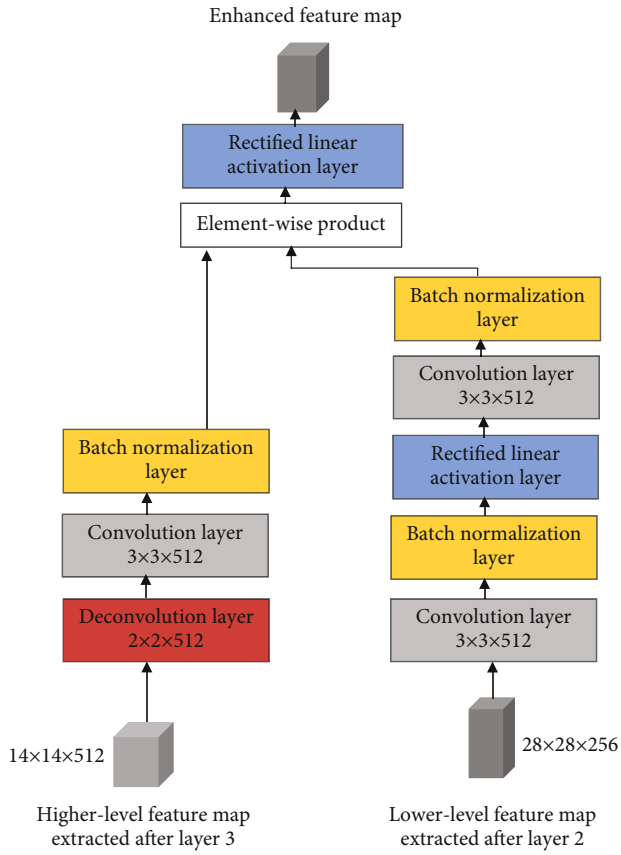

Figure 4: The structure of deconvolution module used in this paper.

deconvolution operation is different from the upsampling operation. Deconvolution operation provides a set of parameters by which to learn a nonlinear upsampling of the features in the deep layers while upsampling operation rescales an image to the desired size by using an interpolation method. Finally, the outputs of these two feature layers, which have the same spatial size and depth, are combined by an element-wise product and processed by a ReLU layer to produce an enhanced feature map.

3.3. Proposal Generation with Improved Region Proposal Network. The region proposal network (RPN) [17] is an efficient and accurate region proposal generation network which showed encouraging performance in general object detection. To increase the inference speed and detection accuracy of the proposed framework, this paper designs an improved RPN based on the RPN [17]. Figure 5 illustrates the structure of the original RPN (a) and the improved RPN (b) proposed in this paper. First, to compress the RPN and increase the inference speed, this paper reduces the number of channels of input feature maps to decrease the number of parameters in the RPN. Decreasing the input channels of input features can further reduce the number of parameters in the subsequent convolutional layers. ResNet [32] used a 1×1 convolution layer to reduce the number of input channels without losing accuracy while also gaining efficiency. Thus, this paper adopts a 1×1 convolution layer after input feature maps to reduce the number of parameters in the RPN. From the

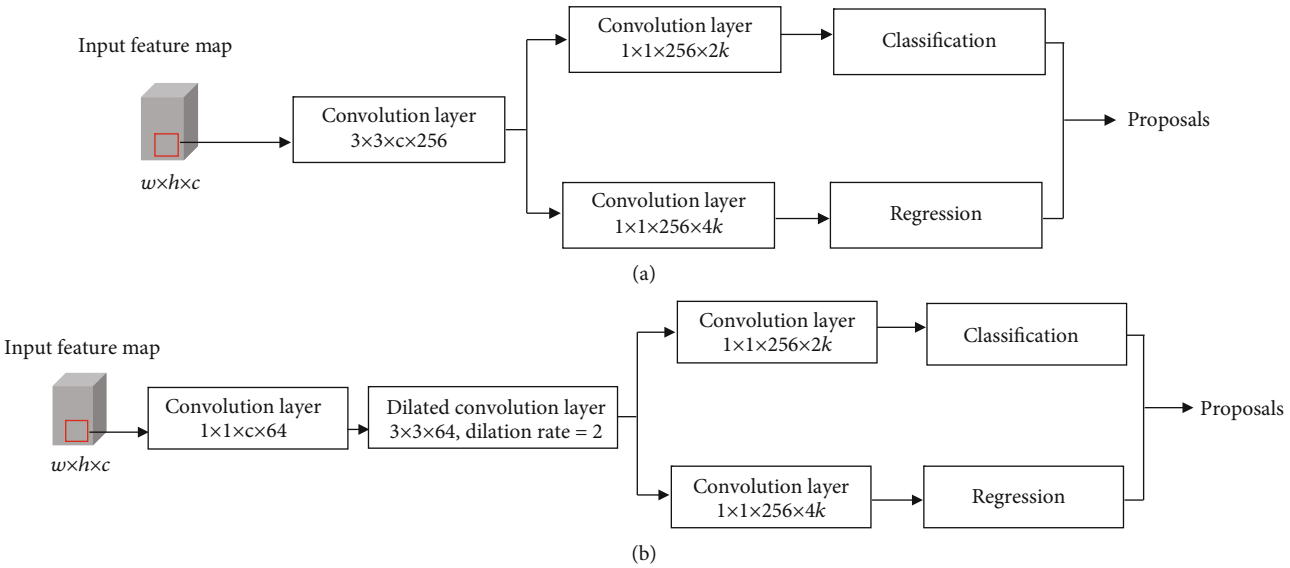

Figure 5: The structure of the original RPN (a) and the improved RPN proposed in this paper (b).

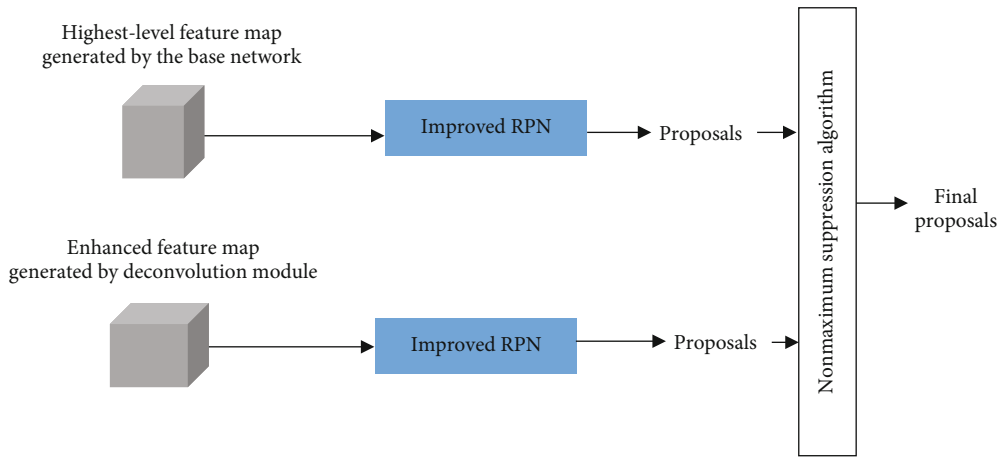

Figure 6: The architecture of the region proposal generation module.

experimental results, this paper reduces the number of input channels to 64 for obtaining good proposals and fast processing speed. Second, to improve the detection accuracy, dilated convolution [33] is adopted to replace standard convolution in the original RPN. Dilated convolution is a powerful module in the context of semantic segmentation. Dilated convolution is used to enlarge the receptive field, thus including more information from other areas to help recognize the boundaries of objects. Another advantage is that using dilated convolution does not increase the number of parameters or the amount of computation because the zero operations are skipped. The improved RPN proposed in this paper can be adopted in other object detection tasks, such as pedestrian detection and vehicle detection.

The enhanced feature map generated by the deconvolutional module could improve the resolution and semantic information for small-scale traffic signs, and the feature map from the highest convolutional layer of the ESPNetv2 network could better describe the characteristics of the large-scale traffic signs in an image. The region proposal generation module is proposed to receive the enhanced feature map and the highest-level convolution feature map and produce a set of proposals to be further processed by the nonmaximum suppression (NMS) algorithm at the end of the region proposal generation module. The region proposal generation module includes two improved RPNs working on different convolution layers for generating proposals as shown in Figure 6. Each improved RPN first generates a set of anchor boxes from the input convolution feature map and then produces two different outputs for each of the anchor box. The first one is an objectness score, which means the probability that an anchor is an object. The second output is the bounding box regression for adjusting the anchors to better fit the object. Since anchors are usually highly overlapped with each other, nonmaximum suppression algorithm is adopted at the end of the region proposal generation module to solve the issue of duplicate proposals.

Table 2: Dataset summary.

Dataset	Number of images	Resolution (width × height)	Ratio between training and testing images	Number of traffic signs	The size of traffic sign
GTSDB	900	1360 × 800	2 : 1	1206	16 × 16 to 128 × 128
TT-100K	100000	2048 × 2048	2 : 1	30000	2 × 7 to 397 × 394

3.4. Detection Network. The detection network includes a region of interest (RoI) pooling layer for adjusting the size of proposals to fixed size feature maps and fully connected (FC) layers for classifying proposals and regressing the bounding box of proposals. The RoI pooling layer uses max pooling to convert the features inside any valid region of interest into a small feature map with a fixed spatial extent of $H \times W$. RoI pooling works by dividing the $h \times w$ RoI proposal into a $H \times W$ grid of subwindows of approximate size $h/H \times w/W$ and then max-pooling the values in each subwindow into the corresponding output grid cell. RoI pooling avoids repeatedly computing the convolutional layers, so it can significantly speed up both train and test time. After extracting fixed size feature maps for each of proposals via RoI pooling, these feature maps are used for classification. The classification stage has two different goals: classify proposals into traffic sign and background class and adjust the bounding box for each of detected traffic sign according to the predicted class. Two FC layers with 2048 neurons are adopted in this paper to extract discriminative features for traffic sign. The features from each RoI are flattened into a vector and fed into two separate FC layers: a box classification layer and a box regression layer. The first FC layer is then fed into the softmax layer to compute the confidence probabilities of being traffic sign and background. The second FC layer with linear activation functions regresses the bounding box of detected traffic sign.

3.5. Training and Loss Function. Training process of the proposed framework includes two phases. First, the region proposal generation network is trained with training samples, and then, both the region proposal generation network and the detection network are trained. For convolution feature with resolution $M \times N$, there are total $M \times N \times k$ anchor boxes. These anchor boxes are highly overlapped with each other. Moreover, there are much more negative anchor boxes than positive anchor boxes, which will lead to bias during the training process if all anchor boxes are used for training. For an anchor box a_k at position (x_m, y_n) in an image, this paper first finds the best matching ground truth box for this anchor box based on the intersection of union (IoU) to create a training sample for this anchor box. The IoU between two anchor boxes a_k and a_i is defined as the following equation:

$$\mathrm{IoU} = \frac{\mathrm{area}\,(a_k \cap a_i)}{\mathrm{area}(a_k \cup a_i)}. \tag{1}$$

Let a_{gtk} represents the best matching ground truth box for anchor box a_k, and $\mathrm{IoU}_{a_k,a_{gtk}}$ represents the IoU overlap between anchor box a_k and ground truth box a_{gtk}. If $\mathrm{IoU}_{a_k,a_{gtk}}$ is higher than 0.7, the anchor box a_k is selected as positive anchor. If $\mathrm{IoU}_{a_k,a_{gtk}}$ is lower than 0.3, the anchor is selected as negative anchor. Otherwise, this anchor will be discarded and is not used as a training sample. Then, 256 anchor boxes from one image are randomly sampled as a minibatch, where the ratio between positive and negative anchors is up to 1 : 1. The regression of the bounding box can be obtained from the anchor coordinates and the ground truth bounding box in a similar way presented in [17].

In this paper, the objective loss function is to minimize the weighted sum of classification loss L_{cls} and localization loss L_{reg} for the region proposal generation network and the detection network. The objective loss function is defined as follows:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_{\mathrm{cls}}(a_i, a_i^*) + \frac{1}{N_p}\sum_{i=1}^{N_p} L_{\mathrm{reg}}(l_i, l_i^*), \tag{2}$$

where N is the size of training samples, N_p is the number of positive samples in training samples, a_i is the predicted probability of anchor i being a traffic sign, a_i^* is the corresponding ground truth label (1 for positive anchor and 0 for negative anchor), l_i is the predicted coordinate offsets for anchor i, and l_i^* is the associated offsets for anchor i relative to the ground truth. Since there is no ground truth bounding box matched with negative anchor boxes, bounding box regression is applied only for positive anchor boxes. In (2), the binary logistic loss is used for box classification, and smooth L1 loss [17] is adopted for box regression. With the above objective function, the network can be trained by backpropagation and stochastic gradient descent strategies.

At test time, test images are fed into feed-forward pass of the network. The proposal generation network generates proposal candidates with bounding boxes and classification confidences. Nonmaximum suppression (NMS) is adopted to select 256 proposals with higher confidences based on the predicted scores, and the detection network further refines the location and class scores for each proposal.

4. Experimental Results and Discussions

In order to compare the effectiveness of the proposed approach with other state-of-the-art approaches on traffic sign detection, this paper conducts experiments on two public datasets: German Traffic Sign Detection Benchmark and Tsinghua-Tencent 100K. The proposed approach is implemented on a Window system machine with Intel Core i7 8700 CPU, NVIDIA GeForce GTX 1080 GPU and 16Gb of RAM. TensorFlow is adopted for implementing deep CNN frameworks.

4.1. Dataset. In this paper, two public datasets are used to evaluate the effectiveness of the proposed approach,

Table 3: The numbers of traffic sign instances in each group size of TT-100K dataset.

Size	Small	Medium	Large
Numbers of instances	8953	10702	1512

including German Traffic Sign Detection Benchmark (GTSDB) [34] and Tsinghua-Tencent 100K (TT-100K) [9]. Table 2 summarizes related information of these datasets.

GTSDB is the most widely used dataset to evaluate traffic sign detection approaches. GTSDB contains 900 images and is divided into 600 training images and 300 testing images. The resolution of each image in this dataset is 1360×800. Each image contains zero to six traffic signs which may appear in every perspective and under every lighting condition. All traffic signs in this dataset can be divided into four categories: prohibitory signs with red color and circular shape, danger signs with red color and triangular shape, mandatory signs with blue color and circular shape, and the rest of traffic signs with different shapes and colors which cannot be classified into these three categories. The sizes of the traffic signs in the images vary from 16×16 to 128×128.

TT-100K contains 100,000 images with resolution 2048×2048 and covers large variations in illuminance and weather conditions. Of these, 10,000 images contain 30,000 traffic signs in total. The ratio between the training set and testing set is 2 : 1. Traffic signs in this dataset can be classified into three categories: prohibitory signs with mostly red circle and black information, mandatory signs with mostly blue circle and white information, and warning signs with mostly yellow triangle and a black boundary and information. Furthermore, based on the size of traffic signs in image, traffic signs in this dataset can be divided into three categories according to their size: small traffic signs (area < 322 pixels), medium traffic signs (322 pixels < area < 962 pixels), and large traffic signs (area > 962 pixel). Table 3 shows the numbers of traffic sign instances in each group size. Compared with the GTSDB dataset, images in TT-100K dataset are higher resolution, and traffic sign instances are smaller and more diverse. To test the ability of the proposed framework on detection of different object sizes, this paper conducts experiments on all group sizes.

4.2. Evaluation Metrics. In order to compare the results of the proposed method with other methods on traffic sign detection, this paper adopts three widely used criteria for evaluating the proposed method, including precision (P), recall (R), and F-measure (F). These criteria are defined as follows:

$$P = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})},$$
$$R = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}, \quad (3)$$
$$F = 2 \times \frac{(P \times R)}{(P + R)},$$

where TP (True Positive) represents the correct detections of traffic signs, FP (False Positive) represents the wrong detections of traffic signs, and FN (False Negative) represents the number of missed traffic signs. A traffic sign is considered as correct detection if the IoU between this traffic sign and ground truth traffic sign is at least 0.5.

4.3. Performance Results. In this section, several experiments are first conducted on the GTSDB dataset to evaluate the effectiveness of each proposed module. Then, this paper compares the performance of the proposed approach with other state-of-the-art approaches on the TT-100K dataset, including Faster R-CNN [17], SSD [26], Multiclass Network [9], and MR-CNN [8].

4.3.1. Experimental Results on GTSDB Dataset. To evaluate the effectiveness of each module of the proposed framework, this paper conducts several experiments on the GTSDB dataset and compares the detection results with original Faster R-CNN framework. In the first experiment, the base network in original Faster R-CNN (VGG-16 [35]) is replaced by the ESPNetv2 network. The RPN and the detection network are kept unchanged as in Faster R-CNN. In the second experiment, the improved RPN is used to replace the original RPN in Faster R-CNN. VGG-16 is kept as the base network in this experiment. In the third experiment, VGG-16 is replaced by ESPNetv2, and two improved RPNs are adopted to generate proposals from a convolution feature map after layer 2 and layer 4 of the ESPNetv2 network. In the final experiment, a deconvolution module is added as described in the previous section. At the same time, two improved RPNs are used to generate proposals from a convolution feature map generated by a deconvolution module and the highest-level convolution feature map. To find the best design of the deconvolution module, this paper uses both element-wise sum and element-wise product in the merging process of a deconvolution module. Table 4 shows the detection results and the inference speed of these experiments on GTSDB dataset. As shown in Table 4, the mAP of Faster R-CNN with ESPNetv2 network is 89.40%. By adding additional improved RPNs after layer 2 of the base network, the mAP increases by 5.14%. This result shows that the additional convolution layer with higher resolution can learn discriminative features of small traffic signs and lead to better performance of the framework. Moreover, by adding a deconvolution module to enhance a shallower convolution feature map with a deeper convolution feature map, the mAP increases by 7.5% and 7.62%. These results demonstrate that the convolution feature map constructed by fusing an upsampled deeper convolution layer with a deconvolution operation and a shallower convolution layer is more suitable for small traffic sign detection. For the inference speed, Faster R-CNN with an improved RPN module improves the speed by 0.13 second. In addition, the inference time of Faster R-CNN with ESPNetv2 network is improved by 0.71 second compared with original Faster R-CNN. These results show the effectiveness of the proposed RPN and ESPNetv2 architecture. Figure 7 shows the comparison of detection results of the proposed method (red boxes) and Faster R-CNN (blue boxes) on GTSDB dataset. As shown in Figure 7, the proposed method can locate exactly small traffic signs while Faster R-CNN cannot detect small traffic signs.

Table 4: Detection results of ablation experiments on GTSDB dataset.

Method	Detection mAP (%)	Inference time (s)
Faster R-CNN	91.17	0.81
Faster R-CNN+ESPNetv2 network	89.40	0.10
Faster R-CNN+improved RPN+VGG-16	92.20	0.68
Faster R-CNN+ESPNetv2 network+two improve RPNs	94.54	0.11
Faster R-CNN+ESPNetv2 network+two improved RPNs+deconvolution module with element-wise sum	96.90	0.16
Faster R-CNN+ESPNetv2 network+two improved RPNs+deconvolution module with element-wise product	97.02	0.18

Figure 7: Detection results of the proposed method (red boxes) and Faster R-CNN (blue boxes) on GTSDB dataset.

4.3.2. Experimental Results on TT-100K Dataset. To evaluate the effectiveness of the proposed method, this paper compares the detection results of the proposed method with the results of other state-of-the-art methods on TT-100K dataset, including Faster R-CNN [17], SSD [26], Multiclass Network [9], and MR-CNN [8]. Table 5 shows the comparison of detection results on all three categories of the TT-100K dataset. As shown in Table 5, the performance of the proposed method outperforms both Faster R-CNN and SSD. More specific, in terms of F-measure, the performance of the proposed method outperforms Faster R-CNN by 51.8%, 19.3%, and 6.4% in a small, medium, and large group, respectively.

Table 5: Detection results of the proposed method and other methods on the Tsinghua-Tencent 100K dataset.

Methods	Small			Medium			Large			Inference time (s)
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	
Faster R-CNN [17]	24.1	49.8	32.5	65.6	83.7	73.6	80.8	91.2	85.7	2.15
SSD [26]	25.3	43.4	32.0	67.8	77.5	72.3	81.5	86.9	84.1	0.12
Multiclass Network [9]	81.7	87.4	84.5	90.8	93.6	92.2	90.6	87.7	89.1	5.62
MR-CNN [8]	82.9	89.3	86.0	92.6	94.4	93.5	92.0	88.2	90.1	—
Proposed method	80.1	89.0	84.3	91.3	94.6	92.9	92.5	91.8	92.1	0.28

Figure 8: Detection results of the proposed method on Tsinghua-Tencent 100K dataset. The regions of detection results are cropped and enlarged at the bottom of each image for better view.

Comparing with SSD, the performance of the proposed method is improved by 51.3%, 20.6%, and 8.0% in a small, medium, and large group, respectively. Since Faster R-CNN and SSD framework are constrained by the size of its output convolution feature maps, these frameworks are unable to clearly detect small traffic signs. Comparing with Multiclass

Network, the proposed approach achieves a competitive result with a small group and outperforms with a medium and large group. For the inference time, the proposed method takes 0.28 second for processing an image, while Faster R-CNN framework takes up to 2.15 seconds. The ESPNetv2 network and improved RPN used in this paper dramatically improve the inference time of the proposed approach. SSD is the fastest framework which takes only 0.12 second, but SSD shows worse performance than the proposed approach. From Table 5, MR-CNN achieves the best performance in a small and medium group. Because the inference speed is not mentioned in the original paper of MR-CNN and the code of the paper is not public, the inference time cannot compare directly. However, MR-CNN framework uses multiple deconvolution and concatenation operations at different layers of the base network, so this framework is high computational cost. The proposed approach achieves the best performance in a large group. Moreover, the proposed approach meets the real-time detection standard and can be directly applied to hardware used in driving environments. Figure 8 shows some detection results of the proposed method on Tsinghua-Tencent 100K dataset. As shown in Figure 8, the proposed method can detect small traffic sign instances which occupy less than 2% of image.

5. Conclusions

This paper proposes a deep learning-based framework for fast and efficient traffic sign detection. To improve the inference speed of the proposed framework, ESPNetv2 network is adopted as the based network. A deconvolution module is used to create an enhanced feature map which contains more representation capacity. Furthermore, an improved region proposal network, which includes a 1×1 convolution layer to reduce the number of parameters in the subsequent convolutional layers and a 3×3 dilated convolution to enlarge the receptive field, is designed to increase the performance of a proposal generation stage. In the experiments, two widely used datasets are adopted to evaluate the effectiveness of each enhanced module and the whole framework, including GTSDB dataset and TT-100K dataset. Experimental results on these datasets show that the proposed framework improves the performance of traffic sign detection under challenging driving conditions and meets the real-time requirement of an advanced driver assistant system.

Data Availability

The codes used in this paper are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] C. Liu, S. Li, F. Chang, and Y. Wang, "Machine vision based traffic sign detection methods: review, analyses and perspectives," *IEEE Access*, vol. 7, pp. 86578–86596, 2019.

[2] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler, "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005*, pp. 255–260, Las Vegas, NV, USA, 2005.

[3] S. Salti, A. Petrelli, F. Tombari, N. Fioraio, and L. Di Stefano, "Traffic sign detection via interest region extraction," *Pattern Recognition*, vol. 48, no. 4, pp. 1039–1049, 2015.

[4] S. K. Berkaya, H. Gunduz, O. Ozsen, C. Akinlar, and S. Gunal, "On circular traffic sign detection and recognition," *Expert Systems with Applications*, vol. 48, pp. 67–75, 2016.

[5] R. Timofte, K. Zimmermann, and L. V. Gool, "Multi-view traffic sign detection, recognition, and 3D localisation," in *2009 Workshop on Applications of Computer Vision (WACV)*, pp. 1–8, Snowbird, UT, USA, December 2009.

[6] Z. Zhu, J. Lu, R. R. Martin, and S. Hu, "An optimization approach for localization refinement of candidate traffic signs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3006–3016, 2017.

[7] S. B. Wali, M. A. Hannan, A. Hussain, and S. A. Samad, "An automatic traffic sign detection and recognition system based on colour segmentation, shape matching, and SVM," *Mathematical Problems in Engineering*, vol. 2015, Article ID 250461, 11 pages, 2015.

[8] Z. Liu, J. Du, F. Tian, and J. Wen, "MR-CNN: a multi-scale region-based convolutional neural network for small traffic sign recognition," *IEEE Access*, vol. 7, pp. 57120–57128, 2019.

[9] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2110–2118, Las Vegas, NV, June 2016.

[10] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu, "Traffic sign detection based on convolutional neural networks," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Dallas, TX, 2013.

[11] 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, 2016.

[12] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient CNNs in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 975–984, 2019.

[13] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2022–2031, 2016.

[14] C. Li, Z. Chen, Q. M. J. Wu, and C. Liu, "Deep saliency with channel-wise hierarchical feature responses for traffic sign detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 7, pp. 2497–2509, 2019.

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

[16] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: a light-weight, power efficient, and general purpose convolutional neural network," in *2019 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 9190–9200, Long Beach, CA, USA, June 2019.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[18] J. Zhang, Z. Xie, J. Sun, X. Zou, and J. Wang, "A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection," *IEEE Access*, vol. 8, pp. 29742–29754, 2020.

[19] J. Zhang, W. Wang, C. Lu, J. Wang, and A. K. Sangaiah, "Lightweight deep network for traffic sign classification," *Annals of Telecommunications*, 2019.

[20] 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 Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8019–8030, 2018.

[21] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*, pp. 354–370, Springer, 2016.

[22] C. Cao, B. Wang, W. Zhang et al., "An improved faster R-CNN for small object detection," *IEEE Access*, vol. 7, pp. 106838–106846, 2019.

[23] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1951–1959, Honolulu, HI, 2017.

[24] J. Wei, J. He, Y. Zhou, K. Chen, Z. Tang, and Z. Xiong, "Enhanced object detection with deep convolutional neural networks for advanced driving assistance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1572–1583, 2020.

[25] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, Honolulu, HI, 2017.

[26] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Computer Vision – ECCV 2016*, pp. 21–37, Springer, 2016.

[27] A. G. Howard, M. Zhu, B. Chen et al., *MobileNets: efficient convolutional neural networks for mobile vision applications*, CoRR, 2017.

[28] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June2018.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, Salt Lake City, UT, USA, June 2018.

[31] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : deconvolutional single shot detector," 2017, https://arxiv.org/abs/1701.06659.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[33] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, https://arxiv.org/abs/1511.07122.

[34] 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 *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Dallas, TX, USA, August 2013.

[35] K. Simonyan and A. Zisserman, "Very Deep deep Convolutional convolutional Networks networks for Largelarge-Scale scale Image image Recognitionrecognition," 2014, https://arxiv.org/abs/1409.1556.