Computer Networks 136 (2018) 95–104

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

# Deep detection network for real-life traffic sign in vehicular networks

Tingting Yang[a], Xiang Long[a], Arun Kumar Sangaiah[b], Zhigao Zheng[c], Chao Tong[a,*]

[a] State Key Laboratory of Virtual Reality Technology and Systems, School of Computing Science and Engineering, Beihang University, Beijing, China
[b] School of Computing Science and Engineering, Vellore Institute of Technology (VIT), Vellore-632014, India
[c] Big Data Technology and System Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

*Article history:*
Received 13 November 2017
Revised 4 February 2018
Accepted 27 February 2018
Available online 2 March 2018

*Keywords:*
Traffic sign detection
Attention network
Fine region proposal network
Convolutional neural network

ABSTRACT

The challenge for real-life traffic sign detection lies in recognizing small targets in a large and complex background, making state-of-the-art general object detection methods not work well in both detection speed and precision. The existing deep learning models for traffic signs detection fail to use the fixed feature of the targets. This paper proposes a novel end-to-end deep network that extracts region proposals by a two-stages adjusting strategy. Firstly, we introduce an AN (Attention Network) to Faster-RCNN for finding all potential RoIs (Regions of Interest) and roughly classifying them into three categories according to colour feature of the traffic signs. Then the FRPN (Fine Region Proposal Network) produces the final region proposals from a set of anchors per feature map location extracted by the AN. We also modify the model by (1) adding a deconvolutional structure to convolutional layers to fit the small size of targets, and (2) replacing the classifier with three softmax corresponding to three coarse categories obtained by the AN. Our method is evaluated on two publicly available traffic sign benchmarks which are collected in real road condition. The experiments show our method generates only 1/14 of the anchors generated by Faster-R-CNN so the detection speed is increased by about 2 *fps* with ZF-Net and it reaches an average mAP of 80.31% and 94.95% in two benchmarks, 9.69% and 7.88% higher than Faster-R-CNN with VGG16, respectively.



## 1. Introduction

In modern cities, the quick increase of vehicles is a great puzzling problem, along with traffic accidents. The vehicular networks and the intelligent systems [1–6] are considered as new solutions to improve transportation efficiency and safety. Traffic sign detection and recognition are indispensable for intelligent terminals of vehicular networks or self-driving cars. Deep learning can be one efficient solution for the complicated detection task which requires high accuracy in real-time for multi-object detection. An example is shown in Fig. 1.

In comparison with normal detection tasks, the traffic signs occupy small proportion of each image in the real-driving scenario [7]. Instead, the majority of the picture is filled by complex background such as skies, roads, pedestrians, vehicles and streetscape. The environmental factors, like billboards, usually have colour saturation and contrast in common with traffic signs, which badly disturb detection accuracy. Other difficulties are particularly from variable illumination, traffic sign colour fading, different angles deformation or rotation, target occlusion and shadow interference. Fig. 2 shows some difficult examples, making the traffic signs detection task still an open problem. The four in the top row show difficult examples because of complex background, occlusion, angles deformation and shadow interference. The ones in the second row show difficult examples because of variable illumination.

Recent years witnessed extremely rapid development of deep learning technology. The CNNs (Convolutional Neural Networks) have shown great strength for image classification [8], especially in ImageNet LSVRC [9]. Object detection, including classification and location, has achieved good practical results through ingenious combination of region proposal method and CNNs. RCNN series works [10–12], are typical representatives in the methods. However, these works are not applicable for real-life traffic sign detection, for the reason that they are designed for the PASCAL VOC [13] task whose target objects typically occupy a large part of each image. Normally, most visual fields of drivers are background like streetscape thereby overwhelming traffic sign targets. The existing methods based on region proposal fail to detect small-size objects. Researchers have proposed some deep learning models [14,15] improved from the general object detection methods. However, the existing deep learning models fail to take advantage of features of the traffic signs. There are fixed colour and shape features in traf-

* Corresponding author.
*E-mail address:* tongchao@buaa.edu.cn (C. Tong).

https://doi.org/10.1016/j.comnet.2018.02.026
1389-1286/

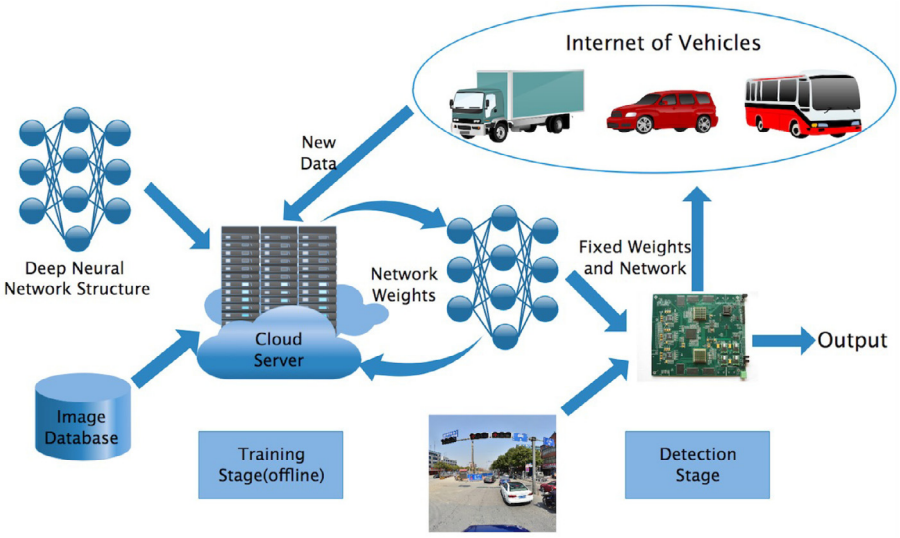


**Fig. 1.** An efficient solution is to train a deep neural network in cloud server offline. The pre-trained network with fixed network weights is ported into the vehicle-mounted intelligent terminals for detection. The new data collected by the terminals is transmitted to cloud server through vehicular networks to promote the performance detection model.

**Fig. 2.** Some examples for the small traffic signs in large and complex background.

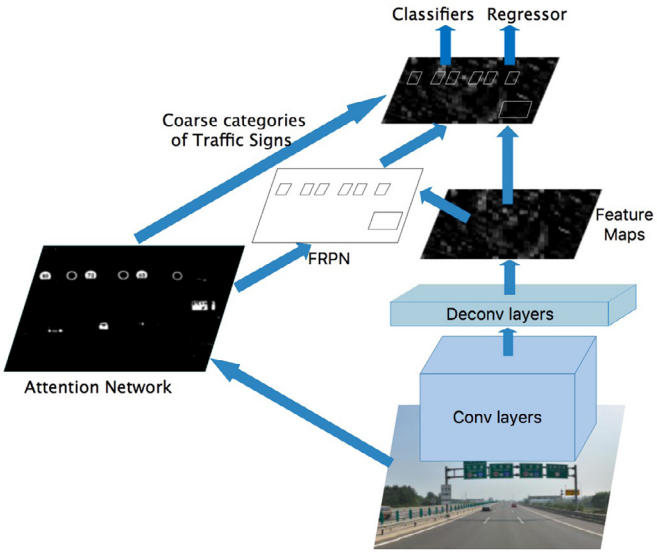


**Fig. 3.** The overview of deep real-life traffic-signs detection system. The input images are processed by the convolutional layers and the attention network in parallel. Their outputs are used for region proposals extracting in the FRPN and detecting targets in the classifiers and regressor.

fic signs for the reason that most countries follow the international patterns.

In order to deal with these problems, we propose a novel and efficient method to detect real-life traffic-signs. The distinctive colour property of the traffic signs is fully used to locate RoIs (Regions of Interest) and narrow the search range effectively. The entire system, as depicted in Fig. 3, is an end-to-end network combining computer vision technology with Faster-R-CNN.

We modify state-of-the-art detection network Faster-RCNN to improve precision and detection speed. The proposed deep detection network makes full use of small targets' properties so as to segment RoIs. Motivated by the biological visual attention mechanism, we introduce a much faster region proposal method by adding a novel AT (Attention Network) which locate all potential RoIs by the traditional computer vision technology according to colour feature of traffic signs. The FRPN (Fine Region Proposal Network) filters the RoIs and fine-tunes the position of the rest ones to produce region proposals. The attention networks can be implemented on CPU, while CNNs take advantage of GPUs. They can work together in parallel to accelerate proposal computation. The attention network can also improve the final classification performance by rough classification in advance. In fact, it offers two kinds of information: RoIs and coarse categories of traffic signs. We also modify the model by adding a deconvolutional structure to convolutional neural network to fit the small size of targets. The framework offers a new thought for the detection task in which small targets are in large and complex background. In summary, the main contributions of this paper can be concluded in three aspects:

- The novel module called attention network, whose main role is to constrains the searching area, tells the entire system where to look according to the colour feature of traffic signs. It offers two kind of information: all possible RoIs and the coarse categories of traffic signs.
- A new framework is proposed which extracts region proposals in a coarse-to-fine manner avoiding to miss the small targets. It integrates the deep CNNs with traditional computer vision algorithms ingeniously, making full use of the characteristics of the target and the strong ability of CNNs to extract features.
- A deconvolutional structure is added to convolutional layers to fit the small size of targets, and the final classifier is replaced

by three softmax classifiers corresponding to three coarse categories obtained by the AN to improve the final precision.

Our method is aimed at small targets detection in large and complex background. Using VGG-16 network, our method yields 80.31% mAP on the Tsinghua–Tencent 100k benchmark and 94.95% on the Belgium Traffic Sign Dateset. The experiments show that our method performs better on small size targets. The remainder of the paper is organized as follows: Section 2 presents related work about the object detection methods, the traffic sign detection methods and the traffic sign benchmarks. Section 3 details the proposed method of this paper, mainly including the architecture of our deep detection network and two-stages adjusting strategy for region proposals extraction. Section 4 introduces the experimental results and discussions. The last section presents the conclusions and future works.

## 2. Related work

### 2.1. Object detection by CNNs

The object detection task is usually divided into two key subtasks: the object classification and the object location. So far, there are two types of strategies. One is object detection based on regression represented by OverFeat [16], YOLO [17] and SSD [18]. The OverFeat combines deep CNN, classification network and regression network. The CNN, elaborated as a feature extractor, is followed by the classification network and the bounding box regression network. YOLO is a real-time end-to-end network, regarding object detection as a regression problem of spatially separated bounding boxes and associated class probabilities. SSD uses small convolutional filters to discretize the output space of bounding boxes into a set of boxes over different aspect ratios and scales. It generates scores for the presence of each object category in each default box and adjusts the boxes to better match the object shapes.

The other one is object detection based on region proposal on behalf of R-CNN series works. These methods hold a lead for different benchmarks considering the accuracy. Faster-R-CNN and R-FCN are the good ones. They both accelerate computing efficiency by sharing computation to get nearly cost-free region proposals. A RPN (Region Proposal Network) is introduced to Faster-R-CNN whose full-image convolutional features are shared with the detection network. It is a fully convolutional network that simultaneously predicts object bounding boxes and objectness scores. The R-FCN is a region-based, fully convolutional network which introduces position-sensitive score maps to address a dilemma of translation-invariance.

Both strategies perform well in PASCAL VOC and other tasks whose targets occupy a large fraction of an image. However, in real-life traffic sign detection tasks, the size of the targets is very small. They ignore many small targets because of insufficient features. Moreover, they generate a great number of candidates so that they have high complexity.

### 2.2. Traffic sign detection and recognition

Although the research of traffic sign detection and recognition has last for decades and has achieved a lot of theoretical achievements, it is difficult to apply these research achievements to the actual industry. The traditional computer vision methods mainly avail oneself of the characteristics of the traffic signs, divided into three classes: the colour-based detection approaches [19–22], the shape-based detection approaches [23–25] and the approaches combining colour and shape detection [26–29]. The colour segmentation is popular to separate the traffic sign targets from background with a threshold in different colour spaces for the reason that traffic signs usually have fixed colours. Commonly used colour spaces include RGB, HSV, HIS, LAB, LUV, YCbCr and so on. The colour-based detection approaches have the advantage of fast computing speed and almost meet the real-time requirement. The shortcomings of low accuracy is also obvious, especially when the background is complex. Because of the regular shapes of the traffic signs, shape-based detection approaches are a common way to use geometric information to locate traffic sign targets. Hough transform [30] is one of the popular approaches to detect circles, triangles and rectangles. In [31], Haar-like and HOG features are combined to detect traffic sign. Moreover, in order to improve the performance, 3D technique is used with SVM and adaboost classifiers. In [32], segmented contours described by Fourier descriptors, together with an implicit star-shaped model are used as prototypes of traffic sign classes. These methods are robust but time-consuming. The colour-based approaches are sensitive to illumination changing and colour fading, while the shape-based approaches are sensitive to targets occlusion and noise interference in clutter environment. The detection approaches combining the colour and shape features have advantages of both approaches above. Unfortunately, all the approaches above are likely to miss small targets for different reasons. Approaches based sliding window have been proposed to improve the recall. However, a great deal of candidates are produced by sliding windows, hence the high time complexity makes these methods far away from mature.

In [14], multi-scale architecture based on CNNs is proposed to detect traffic signs as early as 2011. In [15], two deep CNNs are included to locate and classify traffic signs together. The FCNs (Fully Convolutional Networks) and the EdgeBox [33] algorithm are utilized for coarse proposals extraction and fine proposals extraction, respectively. Finally, deep CNNs classify the proposals to different classes. In [34], the local binary pattern detector and the AdaBoost classifier are used to locate region proposals and final detection results are produced by cascaded CNNs. However, the above methods are not end-to-end networks, hence they make training task difficult and impractical in real applications. In [35], an end-to-end CNN is proposed to simultaneously classify and locate traffic signs. A six layers CNN is used for extracting features. After convolutional layers, the network branches into three parallel structures, a bounding box layer, a pixel layer and a label layer.

### 2.3. Traffic sign benchmark

There are three famous traffic sign benchmarks: GTSRB (German Traffic Sign Recognition Benchmark), STSD (Swedish Traffic Sign Dataset) and BTSD (Belgium Traffic Sign Dateset) [36]. Last year, Tsinghua University published Chinese traffic sign detection benchmark Tsinghua–Tencent 100k [7], *TT*100*k* for short. There are about 10,000 images containing 30,000 Chinese traffic sign instances. These instances cover a great many variations illuminance. The resolution of the images is up to $2048 \times 2048$, while the traffic sign targets always occupy less than 0.1% of an image. The images in this benchmark are similar to real visual field of drivers. However, detection for these small targets in large and complex background is difficult.

## 3. Proposed method

Urban cities deploy facilities for traffic, e.g. traffic lights and traffic signs. However, urban road condition is complex and interference information is numerous. The original intention of traffic sign design is attracting drivers' attention by vivid colours. In order to take full advantage of the feature, our system, called deep detection network for real-life traffic sign, combining traditional computer vision technology with the state-of-the-art object detection

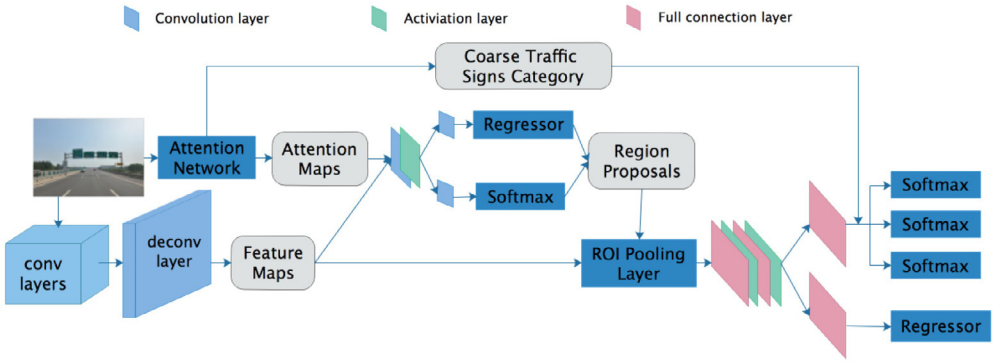


**Fig. 4.** The architecture of deep detection network for real-life traffic-signs. It is a unified network combining computer vision technology and deep learning. The attention network tells the network where to look by generating attention map marking the noticed position. Two FCNs are used to produce region proposals fed into a detector.

network. The entire system using attention mechanisms is a unified end-to-end network for small traffic sign targets detection in a large and complex background.

### 3.1. Network architecture

The proposed deep detection network is composed of four modules. The first module is CNN layers, whose effect is computing features. The output feature maps can not only be effective for feature representation of the full-image but also share computation for the follow-up detection operations. The second module called attention network takes advantage of the intrinsic properties of targets by computer vision technology. The purpose of this module is to call attention of the whole system. The attention maps, as output of the attention network, can offer guidance on where to generate region proposals with the colour model. Furthermore, attention network can acquire information which is useful for the final classification. The attention network coarsely locates the candidate region proposals and performs rough classification. The module does not cost extra computing time because it works in parallel with the CNN layers. The third module FRPN is a fully convolutional network which produce the final region proposals as candidates for the small targets. It works closely with the attention network to generate region proposals in coarse-to-fine manner. For small targets detection tasks, this kind of cooperation can largely narrow the search range and improve the computational efficiency. The last module, functioning as a detector, is improved from Fast-R-CNN. The design of the module, acting as the classifier and regressor, synthesizes various information of above modules. Fig. 4 illustrates the detailed architecture of our deep detection network.

### 3.2. Two-stages adjusting strategy for proposals extraction

There are common characteristics in traffic signs, i.e. striking colours, regular shapes and simple patterns, which are designed for informing and warning drivers. The colours of traffic signs in most countries are blue, yellow, and red. Traffic signs often fall into three categories: Mandatory signs, warning signs and prohibition signs. In Chinese, mandatory signs mostly have blue circles with white information. Warning signs mostly have the yellow background surrounded by black triangles. Prohibition signs always have red circle surrounding white information, and possibly have a diagonal bar. In this way, these distinctive colour property should be made full use of to achieve the purpose of rapid location and detection. It can reduce the computing complexity by coarse location, particularly when the major of the image is useless background. As stated earlier, the property of traffic sign targets can not only provide guidance for generating region proposals, but also coarsely classifies to three categories to make the finally classification and recognition more accurate.

Locating RoIs is not a trivial task, especially when the targets occupy a small fraction of an image. We adopt a two-stages adjusting strategy in a coarse-to-fine manner to speed up object detection process. The strategy is to obtain RoIs firstly according to the colour characteristics of the traffic signs. Finally, the fine region proposals are located by the FRPN. The pipeline of two-stages adjusting strategy for proposals extraction is illustrated in Fig. 5, which is chiefly composed of two stages. The first one is the coarse proposals extraction stage implemented by Attention Network, and the second one is for fine proposals extraction implemented by FRPN. Compared with the previous object proposal methods for small targets, include EdgeBox, selective search, ours provide a smaller number of candidates bounding boxes with higher recall. Given a real-life driving image, the Attention Network generates coarse RoIs by extracting the inherent characteristics of the traffic signs firstly. Then the FCNs with box classifier and box regressor implements process of the filtering negative samples in RoIs and adjusting the bounding boxes of the positive samples. The results show that using the fusion of two-stages has less computational overhead, which is conducive to traffic sign detection of intelligent terminal on vehicular networks.

#### 3.2.1. Attention network for coarse stage

An attention network performs coarse target location and coarse target classification. It takes a real-life driving image of any size as input and outputs image segmentation called attention map and rough traffic signs classification information, for example mandatory signs, prohibition signs and warning signs. We model this process with traditional computer vision methods described in this section. The attention network draws attention to the position where particular colours occur. We adopt a simple but effective method to extract the pixel masks of all the possible RoIs inspired by the human vision mechanism whose attention can be called by bright colours. Its marginal cost for computing coarse region proposals is small.

We use a colour enhancement technique to segment blue, yellow and red regions. The components of RGB colour space are all closely related to lighting levels. As long as the lighting changes, the three components will change accordingly. This phenomenon hampers segmentation process. Therefore, the RGB colour space is suitable for the display system rather than image processing. The transform between the RGB space and the HSV apace can eliminate the correlation between colour components. The little time overhead of the transform can meet the real-time needs of traffic sign detection of vehicular networks. We select the HSV colour space to perform the colour-based segmentation. The transforma-

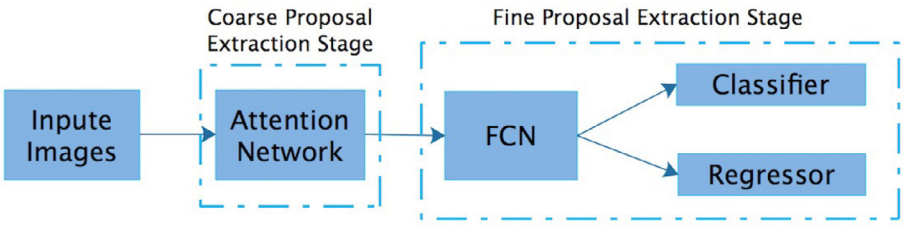


**Fig. 5.** The pipeline of two-stages adjusting strategy for proposals extraction.

**Table 1**
The empirical thresholds for HSV colour space.

| | Blue | Yellow | Red1 | Red2 |
|---|---|---|---|---|
| min($H$) | 90 | 6 | 0 | 140 |
| max($H$) | 130 | 28 | 15 | 185 |
| min($S$) | 120 | 150 | 40 | 10 |
| min($S$) | 255 | 255 | 255 | 255 |
| min($V$) | 35 | 60 | 0 | 0 |
| max($V$) | 255 | 255 | 255 | 2 55 |

tion formula from RGB colour space to HSV colour space is defined as:

$$H = \begin{cases} \dfrac{G-B}{6(MAX-MIN)} & if \quad R = MAX \\ \dfrac{B-R+12}{6(MAX-MIN)} & if \quad G = MAX \\ \dfrac{R-G+24}{6(MAX-MIN)} & if \quad G = MAX \end{cases} \tag{1}$$

$$S = \frac{MAX-MIN}{MAX} \tag{2}$$

$$V = MAX \tag{3}$$

where $MAX$ and $MIN$ denote $max(R, G, B)$ and $min(R, G, B)$ respectively.

After a great quantity of experiments on sample images, the acceptable empirical thresholds (as indicated in Table. 1) are selected for the traffic sign targets segmentation among.

*3.2.2. FRPN for fine stage*

At this stage, a FRPN takes the convolutional feature maps generated by the CNN layers and attention maps produced by the AN and outputs a set of rectangular proposals pre with classification scores of being an object or not. The FRPN includes two FCNs, a box classifier whose results are $2n$ objectness scores and a box regressor whose results are $4n$ coordinates of bounding box, as shown in Fig. 6. The FCN which is in the lead of the object segmentation recently can perform pixel-level prediction on features maps from the last CNNs layer to generate hierarchical representation. The running speed of FRPN is very fast for two reasons. One is that the attention network helps to find coarse RoIs to constrain search ranges and reduce the candidates of the region proposals. The other one is that the FRPN shares convolutional features computation with the CNN layers.

In order to accelerate proposal computation, the attention map is firstly mapped to the feature map through the feat stride. A $3 \times 3$ convolution kernel followed by a ReLU activation layer is selected to slide on the convolutional feature map so as to fuse spatial information. Then we eliminate the positions which have not been noticed by the mapped attention map. The feature vectors (e.g. $256-d$ for ZF-Net) at each noticed position is fed into two $1 \times 1$ FCNs for objectness classification and bounding boxes regression respectively. Each noticed position is considered as the center to

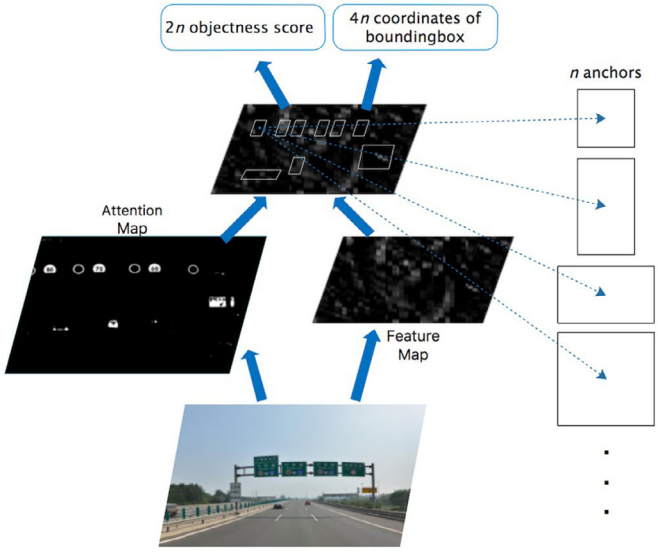


**Fig. 6.** Fine region proposal network.

simultaneously generate $n$ anchors over multiple scales and different aspect ratios. The number of convolutional filters in the FCN for box regression is $4n$ to encode the coordinates of n bounding boxes, while it is $2n$ corresponding to classification probability scores for existing a target or not. Therefore, the fine region proposals extraction can be realized for synthesizing the comprehensive information from the two FCNs at per noticed position in feature map. The objectness classification helps to get rid of the rectangular anchor in which there is no targets. The boxes regression helps to adjust position of fine detection bounding boxes from original anchors. Fig. 6 illustrates fine region proposal network.

For training the FRPN, we select the multi-task loss function defined as:

$$L(\{p_i\},\{b_i\}) = -\frac{1}{N_c}\sum_{i=1}^{n}\log[p_i^* p_i + (1-p_i^*)(1-p_i)] + \lambda\frac{1}{N_b}Smooth_{L1}(b_i - b_i^*) \tag{4}$$

where $i$ denotes the index of anchor. $p_i^*$ denotes a binary objectness label and $p_i$ is a probability predicting anchor $i$ to be a target or not. $b_i^*$ is a 4-dimensional vector representing coordinates of ground truth box while $b_i$ is coordinates vector of predicted bounding-box. A $smooth_{L1}$ is chosen as bounding-box regression loss-function, which is defined as:

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2 & if \quad |x| < 1 \\ |x| - 0.5 & if \quad |x| > \end{cases} \tag{5}$$

### 3.3. Traffic sign classification and regression

After extracting region proposals, we utilize a detector improved from Fast-R-CNN to perform the task of final classification and regression. The inputs of this module include feature maps, region proposals and the coarse traffic signs classification information generated by the attention network. The RoI max pooling layer is adopted to convert region proposals with any size into a fixed feature vectors as the inputs of full connection layer. The coarse traffic signs classification information helps to classify the region proposals into three major categories. There are three softmax classifiers further differentiating the three major categories into specific categories and background. Training three softmax classifiers can learn more details than only one classifiers. Our experimental results show that pre-categorized data and increased number of classifiers improve the accuracy. The bounding boxes are adjusted for twice throughout the entire process. The first time is in the proposals extraction module, and the final results are generated for the second time in this module.

#### 3.3.1. Spacial translation invariance

Spacial translation invariance refers to when the targets are transformed with scaling or translation to any position of the image, accurate detection is immune to this change. Our two-stages adjusting strategy for proposals extraction provides a form of spacial translation invariance. For the first stage, the attention map comes out from the original image with pixel-level colour segmentation method which guarantees spacial translation invariance. For the second stage, the FCN architecture is responsible for the translation invariant property. When it comes to combination of two stages, mapping the attention map into the convolution feature map is mainly related to pooling operation. The pooling regimes make convolution process invariant to translation, rotation and shifting. Most widely used pooling operation is max-pooling. Even an image is translated, highest activation in the sliding window of max-pooling is kept to capture commonality between original image and feature map.

The translation-invariant property is beneficial to detect small targets. The attention network extracts possible RoIs with pixel-level operation, therefore there is little possible to miss the targets even they are very small. The translation-invariant property can also make less risk of overfitting by reducing the number of model parameters.

### 3.4. Implementation details

Just 50 classes traffic sign in TT100k are choosen because of data imbalance, while 28 classes in BTSD. We implement data augmentation to solve the problem of imbalance between different traffic sign classes. The standard traffic signs templates are used to be added to images manually after being scaled and rotated randomly.

For the TT100k, to fit the size of the targets, we select 5 scales with bounding box areas of $16^2$, $32^2$, $64^2$, $128^2$ and $256^2$ pixels, and 3 aspect ratios of 1: 1, 1: 2, and 2: 1 by statistical analysis for the areas of targets. For the BTSD, we select 5 scales with bounding box areas of $48^2$, $80^2$, $112^2$, $144^2$ and $176^2$ pixels, and 3 aspect ratios of 1: 1, 1: 2, and 2: 1 by statistical analysis for the areas of targets.The anchors generated by this range of scales and aspect rations can cover different size-targets.

The convolutional layers are initialized with the ImageNet-pretrained parameters and the rest layers are initialized with Gauss distribution random parameters. We fine-tune the model using a learning rate of 0.001, momentum of 0.9, and learning rate decay of 0.1 for every epoch.

**Table 2**
The statistics about the number of different targets sizes.

| Area | $(0, 32^2)$ | $[32^2, 96^2)$ | $[96^2, \infty)$ |
|---|---|---|---|
| *TT100k* | 8953 | 10,702 | 1512 |
| *BTSD* | 800 | 7892 | 4788 |

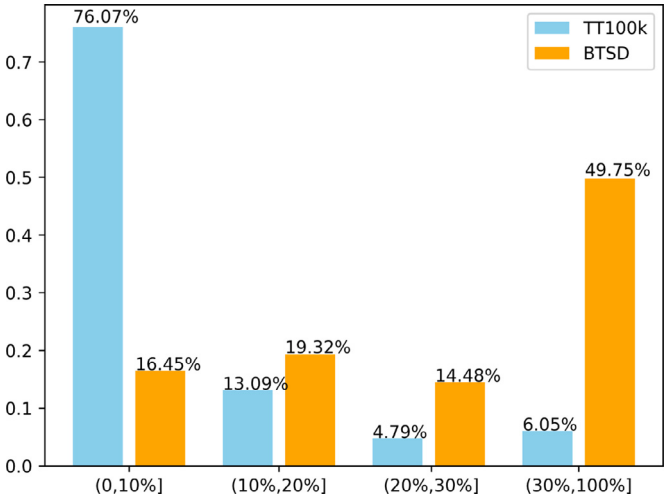


**Fig. 7.** The statistics about the area ratios of traffic sign target to image.

**Table 3**
The comparison of detection results.

| Method | Benchmark | Model | mAP(%) |
|---|---|---|---|
| Faster-RCNN | VGG | TT100k | 70.63% |
| | | BTSD | 87.07% |
| Baseline (w/o deconv) | VGG | TT100k | 73.02% |
| | | BTSD | 91.98% |
| Baseline (w/o AT) | VGG | TT100k | 75.21% |
| | | BTSD | 89.82% |
| Baseline (only 1 softmax) | VGG | TT100k | 79.06% |
| | | BTSD | 93.88% |
| Proposed method | VGG | TT100k | 80.31% |
| | | BTSD | 94.95% |

## 4. Experiments

We chose two benchmarks, *TT100K* for Chinese traffic signs and *BTSD* for Belgian traffic signs, to evaluate the proposed method. Their sizes of images are $2048 \times 2048$ and $1628 \times 1236$ respectively. Their numbers of images are 9180 and 9007 respectively.

We divided the traffic sign targets into three categories according to size (in pixels): Small targets whose areas are less than $32^2$, medium targets whose areas are between $32^2$ and $96^2$ and large targets whose areas are more than $96^2$. The statistics about the number of different targets sizes is give in Table 2, which notes that there are more small-size targets in *TT100k* than in *BTSD*.

The images in *TT100k* has been cropped to $2048 \times 1648$, while the ones in *BTSD* have not. Fig. 7 shows the statistics about the area ratios of traffic sign target to image, which notes that the detection task for *TT100k* is more difficulty than *BTSD*.

### 4.1. Detection performance

In this section, we evaluate the performance of our method with mAP (mean Average Precision), for the reason that it is the actual metric for object detection. We chose ZF-net which consists of 5 convolutional layers and 3 fully-connected layers, and VGG-16 which consists of 13 convolutional layers and 3 fully-connected layers. Table 3 shows the detection results of our method, together

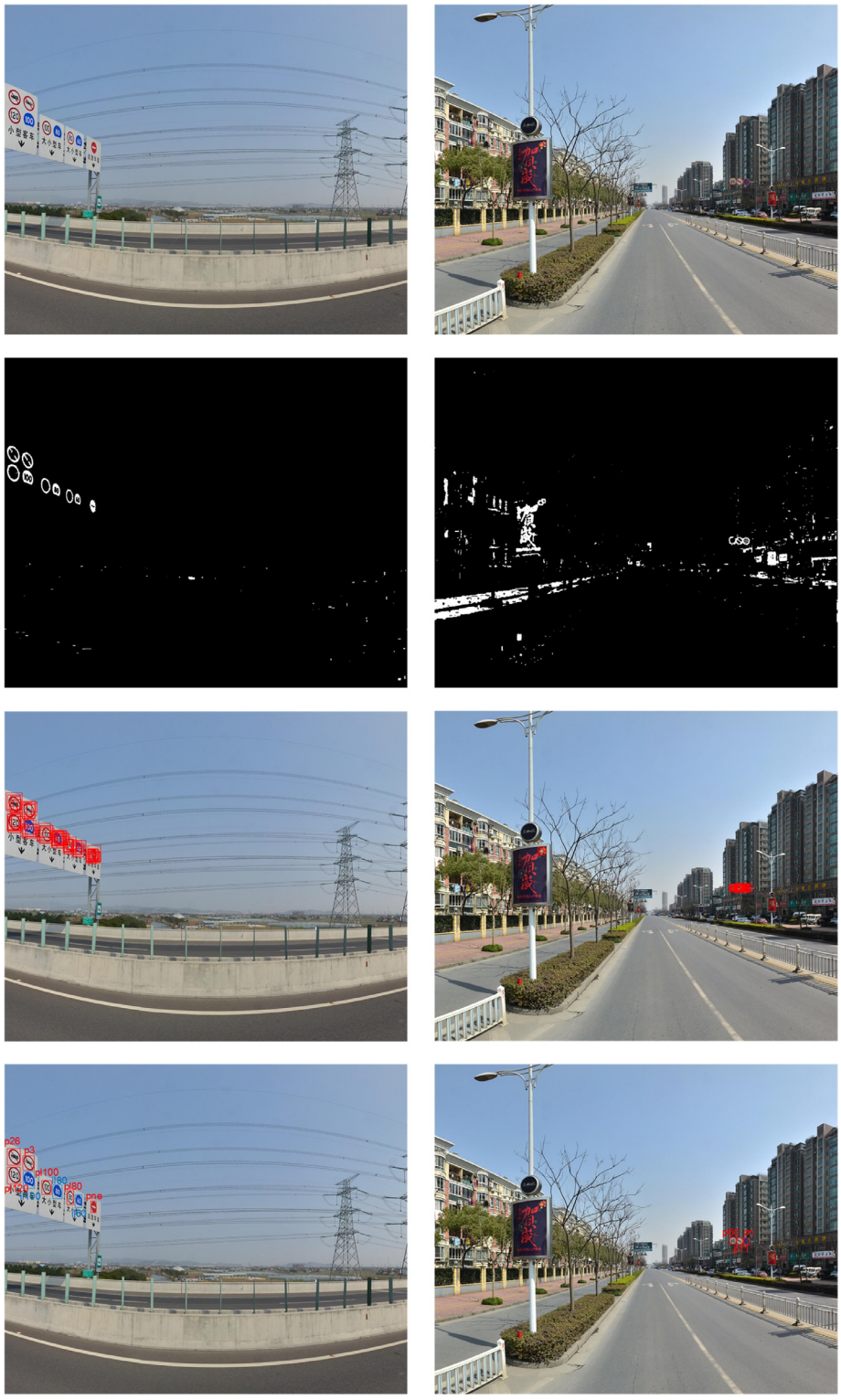

**Fig. 8.** Examples of the proposals extraction stages results and final detection results. The top row shows the original images. The second row shows the attention maps produced by the attention network. The third row shows the region proposals generated by the FRPN and the fourth row shows the final detection results.

with Faster-RCNN in two benchmarks. This table illustrates that our method outperforms Faster-RCNN. We also make comparison to three baseline methods of the improved Faster-RCNN. "Baseline(w/o deconv)" is a baseline method where the deconvolutional layer is omitted. The comparison between 'Ours" and it verifies the effectiveness of the deconvolutional layer. It performs more effectively in TT100k whose targets are more small than BTSD. "Baseline(w/o attention)" is a baseline method who remove the attention network, and "Baseline(1 softmax)" is a baseline method where the final classifier has only one softmax. The comparison shows that the attention network and the increased classifiers both improve the mAP in different degree. We can know that deconvolutional layer help to improve higher mAP for TT100k than for BTSD, because the targets in TT100K is smaller.

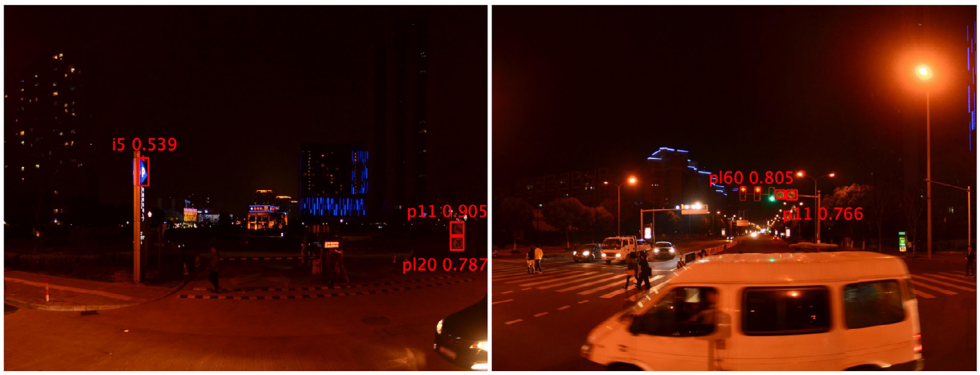


**Fig. 9.** Examples of detecting traffic signs at night.

**Table 4**
The comparison of detection results of our approach and Faster-RCNN for small, medium and large targets.

| Method | Benchmark | Size | AP(%) |
|---|---|---|---|
| Faster-RCNN | TT100k | Small | 31.22% |
| | | medium | 77.17% |
| | | Large | 94.05 |
| | BTSD | Small | 43.93% |
| | | medium | 88.05% |
| | | Large | 96.82 |
| Proposed method | TT100k | Small | 49.81% |
| | | medium | 86.9% |
| | | Large | 96.05% |
| | BTSD | Small | 50.82% |
| | | medium | 97.8% |
| | | Large | 98.31% |

Table 4 shows the detection performance of our approach and Faster-RCNN for small, medium and large traffic sign targets. It is noteworthy that the proposed method performs betters on object targets of small sizes, which reaches the an AP of 49.81%, 18.59% higher than Faster-RCNN for small targets in TT100K, and 50.82%, 6.89% higher than Faster-RCNN in BTSD.

Two examples of the proposal extraction stages results and final detection results are shown in Fig. 8.

For some difficulty condition like, we give two examples when tested at night in Fig. 9. There are a mistake in the image in left, which is the "i5" with low confidence of 0.539.

### 4.2. Time discussion

In this section, we discuss the cost time of detection method in TT100k benchmark. We resize the cropped images to $1024 \times 800$ limited by the memory of GPU. For a $1024 \times 800$ image in the benchmark, Fater-R-CNN will produce $48000(64 \times 50 \times 15)$ original anchors. There are about 20,000 anchors with cross-boundary ones ignored. With our attention network extracting RoIs, only $1000 - 5000$ original anchors are produced and only 600–2000 anchors are left after cross-boundary ones ignored. The less RoIs can also help to reduce the running time of "Region-wise" including NMS, pooling, full connected, and softmax layers. The running time of the attention network is less than 90*ms*, which runs in parallel with CNNs thereby its cost time is contained within the cost time of the CNNs. The cost time of "Proposal" Table 5 lists the cost time (ms) on Tesla K20 GPU with Mxnet. The CNNs in our model cost more time than the CNNs in Faster-RCNN because of the deconvolutional layer. However, there are fewer anchors and proposals thanks to attention networks. Our approach with ZF net takes in total 128 ms, 37 ms faster than the Faster-RCNN. While with VGG-16, it takes 458 ms, 77 ms faster than the Faster-RCNN.

**Table 5**
The cost time for different stages.

| System | model | Conv | Proposal | Region-wise | Total | Rate |
|---|---|---|---|---|---|---|
| Faster-R-CNN | ZF-Net | 90 | 9 | 66 | 165 | 6 fps |
| Faster-R-CNN | VGG | 396 | 29 | 125 | 550 | 1.8 fps |
| Ours | ZF-Net | 98 | 3 | 27 | 128 | 7.8 fps |
| Ours | VGG | 411 | 11 | 51 | 473 | 2.1 fps |

## 5. Conclusions

In this paper, we propose a novel network called deep detection network for real-life traffic sign, updating Faster-R-CNN to adapt to detect small targets in large and complex background. The two-stages adjusting strategy is adopted to extract region proposals in a coarse-to-fine manner. A attention network using tradition computer vision technology is introduced into the framework to extract RoIs coarsely and classify the proposals into three categories coarsely. It largely reduces the search range to reduce the time overhead and improve the accuracy of detection results. We test our method with *TT*100*k* benchmark whose traffic sign targets is extremely small. Our experimental results show that our network outperforms the previous work in both speed and mAP.

In future, we will seek a small network model like MobileNet to run the detection network on mobile devices or intelligent terminals. The shape feature of traffic signs is abandoned because the shape-based proposal extracting methods we have tried are all time-consuming. We will seek a fast and effective shape-based proposal extracting method to further reduce the search domain.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (61472024, U1433203).

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.comnet.2018.02.026.

## References

[1] Z. Zhang, T. Tan, Y.W.K. Huang, Practical camera calibration from moving objects for traffic scene surveillance, Circuits Syst. Video Technol. 23 (3) (2013) 518–533.
[2] Z. Zhang, K. Huang, M.L. Y. Wang, View independent object classification by exploring scene consistency information for traffic scene surveillance, Neurocomputing 99 (2013) 250–260.
[3] C. Yao, W.L. X. Bai, L. Latecki, Human detection using learned part alphabet and pose dictionary, Proc. ECCV (2014).
[4] C. Hu, X. Bai, L. Qi, X. Wang, G. Xue, L. Mei, Learning discriminative pattern for real-time car brand recognition, Intell. Transp. Syst 16 (6) (2015) 3170–3181.

[5] Y. Xia, W. Xu, L. Zhang, X. Shi, K. Mao, Integrating 3d structure into traffic scene understanding with rgb-d data, Neurocomputing 151 (2015) 700–709.
[6] Q. Ling, J. Yan, F. Li, Y. Zhang, A background modeling and foreground segmentation approach based on the feedback of moving objects in traffic surveillance systems, Neurocomputing 133 (2014) 32–45.
[7] Z. Zhu, D. Liang, S. Zhang, X. Huang, Traffic-sign detection and classification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2110–2118.
[8] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv Neural Inf. Process. Syst. 25 (2012) 1097–1105.
[9] O. Russakovsky, J. Deng, J.K.H. Su, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. (IJCV) (2015) 1–42.
[10] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[11] R.B. Girshick, Fast r-cnn. corr, abs 04 (2015) 8–83.
[12] S. Ren, K. He, R.B. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal net- works, abs 06 (2015) 14–97.
[13] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[14] P. Sermanet, Y. LeCun, Traffic sign recognition with multi-scale convolutional networks, in: The 2011 International Joint Conference on Neural Networks (IJCNN), IEEE, San Jose, California, USA, 2011, pp. 2809–2813.
[15] Y. Zhu, C. Zhang, D. Zhou, X. Wang, X. Bai, W. Liu, Traffic sign detection and recognition using fully convolutional network guided proposals, Neurocomputing (2016).
[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition, localiztion and detection using convolutional networks, abs 12 (2013).
[17] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once:unified, real–time object detection, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
[18] W. Liu, D. Anguelov, D. Erhan, Ssd: single shot multibox detector, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
[19] A. la Escalera, J.M. Armingol, M. Mata, Traffic sign recognition and analysis for intelligent vehicles, Image Vis. Comput 21 (3) (2003) 247–258.
[20] S. Maldonado-Bascón, S. S.Lafuente-Arroyo, P. P.Gil-Jimenez, H. Gómez-Moreno, F. López-Ferreras, Road-sign detection and recognition based on support vector machines, IEEE Trans. Intell. Transp. Syst 8 (2) (2007) 264–278.
[21] A. Ruta, Y. Li, X. Liu, Real-time traffic sign recognition from video by class-specific discriminative features, Pattern Recognit. 43 (1) (2010) 416–430.
[22] J. Lillo-Castellano, I. Mora-Jimenez, C. Figuera-Pozuelo, J. Rojo-Alvarez, Traffic sign segmentation and classification using statistical learning methods, Neurocomputing 153 (2015) (2015) 286–299.
[23] M. Garca-Garrido, M. Sotelo, E. Martín-Gorostiza, Fast road sign detection using hough transform for assisted driving of road vehicles, in: Computer Aided Systems Theory?EUROCAST 2005, 153, Springer, Las Palmas de Gran Canaria, Spain, 2005, pp. 543–548.
[24] G. Loy, N. Barnes, Fast shape-based road sign detection for a driver assistance system, in: 2004 Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), vol. 1, IEEE, Sendai, Japan, 2004, pp. 70–75.
[25] X. Bai, S. Bai, Z. Zhu, L. Latecki, 3D shape matching via two layer coding, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2361–2373.
[26] H. Li, F. Sun, L. Liu, L. Wang, A novel traffic sign detection method via color segmentation and robust shape matching, Neurocomputing 169 (2015) 77–88.
[27] G. Piccioli, E. Micheli, P. Parodi, M. Campani, Robust method for road sign detection and recognition, Image Vis. Comput. 14 (3) (1996) 209–223.
[28] X.W. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, N. Shevtsova, Re- cognition of traffic signs based on their colour and shape features extracted using human vision models, J. Vis. Commun. Image Represent 17 (4) (2006) 675–685.
[29] X.B.Y. Zhou, W. Liu, L. Latecki, Similarity fusion for visual tracking, J. Vis. (2016) 1–27. https://doi.org/10.1007/s11263-015-0879-9.
[30] I.M. Creusen, R.G. Wijnhoven, E. Herbschleb, P.D. With, Color exploitation in hog-based traffic sign detection, in: Proceedings of ICIP, IEEE, Hong Kong, China, 2010. 2669–2669.
[31] R. Timofte, K. Zimmermann, L. Gool, Multi-view traffic sign detection, Recognition, and 3d Localisation 25 (3) (2014) 633–647.
[32] F. Larsson, M. Felsberg, Using fourier descriptors and spatial models for traffic sign recognition, in: IImage Analysis, Springer, 2011, pp. 238–249.
[33] C. Zitnick, P. Dollár, Edge boxes: locating object proposals from edges, in: Proceedings of ECCV, Springer, Zurich, Switzerland, 2014, pp. 391–405.
[34] D. Zang, J. Zhang, D. Zhang, M. Bao, J. Cheng, K. Tang, Traffic sign detection based on cascaded convolutional neural networks, in: 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). IEEE, 2016, pp. 201–206.
[35] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, S. Hu, Traffic-sign detection and classification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2110–2118.
[36] M. Mathias, R. Timofte, R. Benenson, L.J.V. Gool, Traffic sign recognition - how far are we from the solution? in: International Joint Conference on Neural Networks (IJCNN), 2013, pp. 1–8.

**Tingting Yang** received her BS degree in Communication Engineering from Tianjin Polytechnic University in 2011, received her MS degree in Computer Science from Civil Aviation University of China in 2015. Now she is pursuing her Ph.D. degree in Beihang University. Her research interests include data mining, deep learning and computer vision.

**Xiang Long** received his BS degree in Mathematics from Peking University in 1985, received the MS and PhD degrees in Computer Science from Beihang University in 1988 and 1994. He has been a professor at Beihang University since 1999. His research interests include parallel and distributed system, computer architecture, real-time system, embedded system and multi-/many-core oriented operating system.

**Arun Kumar Sangaiah** has received his Master of Engineering (ME) degree in Computer Science and Engineering from the Government College of Engineering, Tirunelveli, Anna University, India. He had received his Doctor of Philosophy (PhD) degree in Computer Science and Engineering from the VIT University, Vellore, India. He is presently working as an Associate Professor in School of Computer Science and Engineering, VIT University, India. His area of interest includes software engineering, computational intelligence, wireless networks, bio-informatics, and embedded systems. He has authored more than 100 publications in different journals and conference of national and international repute. His current research work includes global software development, wireless ad hoc and sensor networks, machine learning, cognitive networks and advances in mobile computing and communications. Also, he was registered a one Indian patent in the area of Computational Intelligence. Besides, Prof. Sangaiah is responsible for Editorial Board Member/Associate Editor of various international journals.

**Zhigao Zheng** is with Services Computing Technology and System Lab/Cluster and Grid Computing Lab/Big Data Technology and System Lab, School of Computer Science and Technology, Huazhong University of Science and Technology. He is the guest editor of ACM/Springer Mobile Networks and Applications, Multimedia Tools and Applications, Journal of Intelligent & Fuzzy Systems, Computers & Electrical Engineering, International Journal of Networking and Virtual Organisations (IJNVO) and so on, he is also the reviewer of many journals such as IEEE Transactions on Big Data, IEEE Transactions on Industrial Informatics, Journal of Network and Computer Applications, The Journal of Supercomputing, Multimedia Tools and Applications and some top conference such as SC?16, CCGrid?16, NPC?15 and NPC?16. His main research interest is focused on parallel and distributed computing. His is a member of CCF, IEEE and ACM.

**Chao Tong** received his Ph.D. degrees in 2009 in computer science from Beihang University (BUAA). He is an associate professor in the School of Computer Science and Engineering, BUAA. He is also a visiting professor in the School of Computer Science, McGill University. He has published more than 40 referred papers and filed more than 20 patents. His current research interests include machine learning, mobile computing and social networks analysis. He is a reviewer of many journals and conferences, such as IEEE Transactions on Services Computing, IEEE Transactions on Intelligent Transportation Systems, IEEE Communications Magzine, IEEE Network Magazine, INFOCOM, ICC, etc.