Traffic Sign Detection Using a Multi-Scale Recurrent Attention Network

Yan Tian, Judith Gelernter, Xun Wang, *Member, IEEE*, Jianyuan Li, and Yizhou Yu, *Fellow, IEEE*

***Abstract*— Traffic sign detection plays an important role in intelligent transportation systems. But traffic signs are still not well-detected by deep convolution neural network-based methods because the sizes of their feature maps are constrained, and the environmental context information has not been fully exploited by other researchers. What we need is a way to incorporate relevant context detail from the neighboring layers into the detection architecture. We have developed a novel traffic sign detection approach based on recurrent attention for multi-scale analysis and use of local context in the image. Experiments on the German traffic sign detection benchmark and the Tsinghua-Tencent 100K data set demonstrated that our approach obtained an accuracy comparable to the state-of-the-art approaches in traffic sign detection.**

***Index Terms*— Traffic sign detection, intelligent transportation system, deep learning.**

I. Introduction

TRAFFIC sign detection is a vital topic for both academia and industry, and it has been an active area of research over the past decade. A fast and robust traffic sign detection mechanism can assist and release the driver, and thus, can significantly increase driving comfort and safety. For example, it can remind the driver of traffic constraints, stopping him from performing inappropriate actions such as passing a car in a no passing zone, entering a one-way street, unwanted speeding, to name a few. Further, it could be integrated into an Automated Driving System (ADS) and Advanced Driver Assistance System (ADAS).

Traditional approaches for traffic sign detection include a wide variety of algorithms and various representations, including adaboost [1], support vector machine (SVM) [2], Hough transform [3] and so on, which use color, texture, edge and other low-level features to detect the area or the edge of a traffic sign in an image. These approaches based on low-level features do not detect or recognize traffic signs well because of: 1) variations in the traffic sign appearance due to different sign shapes or colors (e.g., triangle, square, circular symbol, and their color, yellow, red or blue); 2) objects such as trees and vehicles which may occlude the traffic signs; and 3) variations in lighting at different times of day make the same traffic sign appear to be different.

Recently, deep learning, especially Convolution Neural Networks (CNN) [4], has been applied with success in this detection and recognition task. The philosophy underlying the CNN-based approaches is that if we collect and annotate diverse data for training, we can increase the detection performance. However, despite the data set large size, the data are not heterogeneous enough. The small size of traffic signs makes them hard to detect, and useful context information is not exploited fully by CNN-based approaches.

A popular solution to this problem in CNNs is to combine information from the background [5] or relationships among the objects [6], which combines finer details from multiple convolution layers with different local receptive fields. But it has been found that simply concatenating these feature maps does not significantly improve the accuracy due to over-fitting caused by curse of dimensionality. Therefore, we need what has been called an attention mechanism [7] to select relevant features from the convolution layers.

Our theory is that, if we use more information around all objects, and we use more information from non-cluttered regions of the complete image, we can detect more traffic sign objects successfully. In this paper, we develop a novel traffic sign detection approach based on the Deconvolutional Single Shot Detector (DSSD) [8] for multi-scale analysis. Our approach gradually revises the potential object region by making use of the information from the neighboring receptive fields in a recurrent manner, which is illustrated in Fig. 1. The novelty of this approach is that:

- We introduce an attention mechanism [7] into the traffic sign detection task, which focuses on local context information to improve the detection results.
- We propose recurrent attention, assuming that attention maps in neighboring receptive fields are relevant, and use late fusion to combine forms of local information.
- We revise the base network according to the traffic sign size in order to further enhance recall.
- Experiments on the German Traffic Sign Detection Benchmark (GTSDB) [9] and the Tsinghua-Tencent 100K (TT-100K) data set [10] show that the proposed approach is competitive when compared with state-of-the-art approaches for traffic sign detection.

The rest of this paper is organized as follows. Section II reviews studies on traffic sign detection and small

Manuscript received March 27, 2018; revised October 26, 2018; accepted December 7, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant U1609215, Grant 61602407, Grant 61672460, and Grant 61702453, in part by the Natural Science Foundation of Zhejiang Province under Grant LY19F030005 and Grant LY18F020008, and in part by the Opening Foundation of Engineering Research Center of Intelligent Transport of Zhejiang Province under Grant 2017ERCITZJ-KF04. The Associate Editor for this paper was H. G. Jung. *(Corresponding author: Yan Tian.)*

Y. Tian and X. Wang are with the School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310014, China (e-mail: xunwangzjgsu@163.com).

J. Gelernter is with the Information Science Department, Rutgers University, New Brunswick, NJ 08901 USA.

J. Li is with Enjoyor Co., Ltd., Hangzhou 310030, China.

Y. Yu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China.

Digital Object Identifier 10.1109/TITS.2018.2886283

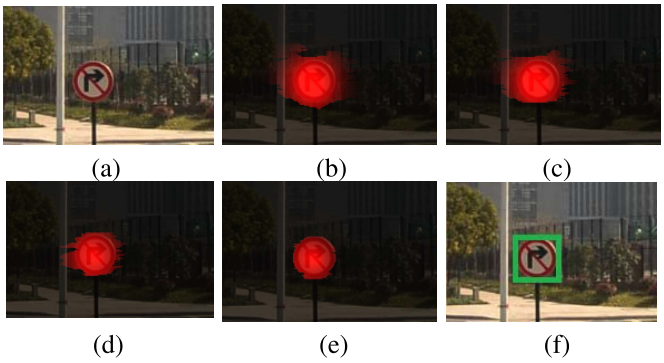


Fig. 1. Illustrations of the recurrent attention in traffic sign detection. Our approach gradually revises the potential object region by making use of the information from the neighboring receptive fields in a recurrent manner. (a) is the input image, (b-e) are recurrent attention maps in iteration 2, 4, 6 and 8. The detected bounding box is illustrated in (f).

object detection. Section III introduces the DSSD network that is the basis for our method. Section IV presents our multi-scale and recurrent attention mechanism for traffic sign detection. Section V presents experimental results. Concluding remarks appear in Section VI.

II. Related Work

A. Traffic Sign Detection

Traditional approaches to traffic sign detection include a wide variety of algorithms and various representations. Escalera *et al.* [11] took advantage of color and shape features to detect road traffic signs, while Shadeed *et al.* [12] used histogram equalization, light control and color segmentation to locate road signs. Later, Garcia-Garrido *et al.* [3] employed the Hough transform to get the information from the edges in the image, but the computational complexity was high so that it hindered the real-time application. To deal with the efficiency problem, Bahlmann *et al.* [1] detected traffic signs using a set of Haar wavelet features obtained from AdaBoost training [13]. To balance effectiveness and efficiency, Salti *et al.* [2] proposed an approach in which the regions of interest rather than the sliding window were extracted at first, and then a histogram of oriented gradients (HOG) in the regions of interest was extracted, to be the input feature of the SVM classifier [14]. Recently, Berkaya *et al.* [15] extended this approach by using an ensemble of features including HOG, local binary patterns (LBP) and Gabor features within an SVM classification framework. To improve results obtained by single view analysis, Timofte *et al.* [16] combined 2D and 3D techniques to generate and evaluate 3D proposals. For more information about traditional approaches, Mogelmose *et al.* [17] provided a survey of traditional approaches to traffic sign detection.

Deep convolution neural networks have made huge progress in object detection and other computer vision tasks, and experts have begun to think about using this approach for traffic sign detection. John *et al.* [18] used CNN to extract features and detect road traffic signs, making a saliency map containing the traffic light location. For network optimization, Jin *et al.* [19] suggested a hinge loss stochastic gradient descent (HLSGD) method to train a detection network. To perform fast and accurate traffic sign detection and recognition, Zhu *et al.* [20] employed a holistically-nested edge detection network [21]. All these approaches are able to detect some traffic signs. However, constrained by feature map size, none show significant advantages.

B. Small Object Detection

The Region Proposal Network (RPN) [22] showed encouraging performance in general object detection by using a deep convolution neural network, for example, the VGG [23] network. Nevertheless, it still could not detect small objects such as traffic lights or traffic signs because of the coarseness of its feature maps.

It is known from feature of the pooling layer and the convolution layer that the first layer of the network contains most of the location information, while the last layer contains the least. Hariharan and Arbelez [24] presented a hyper-column feature that combined feature maps from all the layers to localize the object when up-sampled. Hu and Ramanan [25] later extended this approach to detect small faces in photographs. Kong *et al.* [26] then developed a new hyper-feature in which a deconvolution is used on the 5th layer, and a max-pooling is employed on the 1st layer, so that feature maps from multiple layers were adjusted to the same size and concatenated together. The drawback of this method was that it was very slow.

Then, Liu *et al.* [27] proposed the Single Shot Multi-Box Detector (SSD) that discretized the output space of the bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. It encapsulated all computation into a single network. However, it is still weak in small object detection. To combine low-resolution and semantically strong features with high-resolution and semantically weak features, an hourglass structure network was employed in the Feature Pyramid Network (FPN) [28], the DSSD [8], and the Zoom-Out-And-In network [29]. However, none of these approaches have proved effective in traffic sign detection.

III. Deconvolutional Single Shot Detector

Our method is based on the DSSD structure but it solves what the DSSD cannot do well. The SSD approach does not obtain satisfactory results in small object detection because grids in the feature map in the relatively shallow layer regress to the proposals. Therefore, the DSSD approach uses an "encoder-decoder" structure to project the feature maps in the deeper layers (with more semantic information) to the classification and localization results. The DSSD structure is illustrated in Fig. 2, and the resolution and channel number of the output feature maps in each layer can be seen in Table I.

The DSSD network is set up as follows: The base network is the Residual-101 [30] network instead of the VGG network to achieve better accuracy. It is pre-trained on the ILSVRC CLS-LOC dataset [31]. The convolution layer with stride 2 in the *conv*5_*x* stage is modified to 1 to increase the

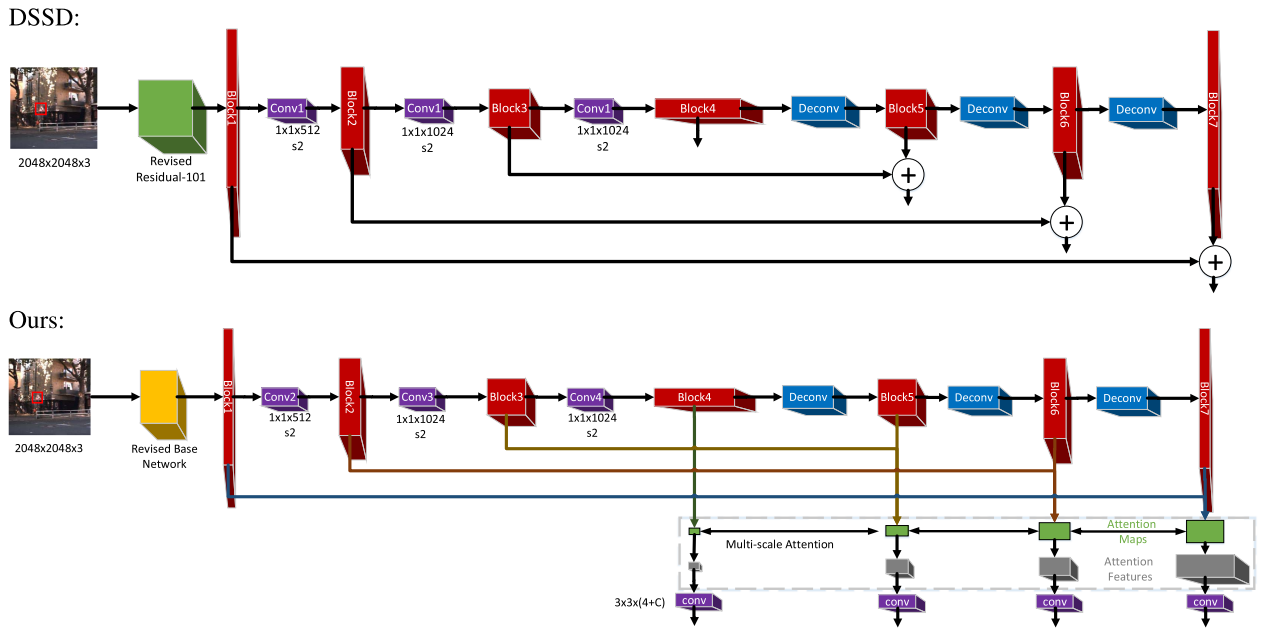


Fig. 2. Architecture of the DSSD and the corresponding network with multi-scale attention structure.

TABLE I
FEATURE MAP PIXEL INFORMATION IN DSSD & FPN VS OUR APPROACH

DSSD [8]	Block1	Block2	Block3	Block4	Block5	Block6	Block7
Resolution	128×128	64 × 64	32 × 32	16 × 16	32 × 32	64 × 64	128×128
Channel	256	512	1024	1024	1024	512	256
FPN [28]	Block1	Block2	Block3	Block4	Block5	Block6	Block7
Resolution	256×256	128×128	64 × 64	32 × 32	64 × 64	128×128	256×256
Channel	256	256	256	256	256	256	256
Ours	Block1	Block2	Block3	Block4	Block5	Block6	Block7
Resolution	256×256	128×128	64 × 64	32 × 32	64 × 64	128×128	256×256
Channel	256	512	1024	1024	1024	512	256

feature maps' resolution. Beyond the base network, pooling layers or convolution layers with stride 2 are used to compress the feature maps to a lower resolution. At each down-sampling step, a residual block [30] is employed for information processing in that receptive field. This procedure continues until the lowest resolution is obtained. Then the network begins the expansion process, that is, up-sampling of the lower resolution feature maps by using the deconvolution layer. The structure of the DSSD network is symmetrical: for each layer presented at a down-sampling layer there is a corresponding up-sampling layer.

To fuse information from the corresponding layers, feature maps from two layers with the same resolution are merged. Then, a 1×1 convolution layer is employed to obtain the position and prediction confidence of specific objects.

The "encoder-decoder" structure increases the recall on the small object detection. However, context information is neglected in this approach, and clutter in the background may confuse the object detection task and make it even more difficult.

IV. OUR NETWORK

To deal with the problems that DSSD cannot solve well, we introduced an attention mechanism into the "encoder-decoder" structure. The framework is shown in the bottom diagram of Fig. 2.

Attention is a mechanism to dynamically extract salient features that the current stage needs, rather than compress the entire frame information into a static representation. This method is especially useful when there is a clutter background in an image, where high activation values in the background region confuse the classifier, and we need to distill the salient foreground region. Unfortunately, there is a potential disadvantage in the distill phase that the valuable foreground information may also be removed with the irrelevant background.

Detection and segmentation are relevant tasks because an accurate bounding box gives the borders of the object of interest and an accurate segmentation provides the region cues for detection. Compared with approaches that use a complex model for the segmentation, we propose an iterative method to revise the attention map in the detection task by using the information from receptive fields near the object and dynamically extracting salient features.

A. Multi-Scale Attention

Small objects in the traffic environment such as traffic signs, traffic lights, and lane marking patches, not to mention the exact locations of the correct bounding boxes, can hardly

Fig. 3. Examples of the attention maps at receptive fields 1, 2, 3 and 4 of the decoder (with 1 the lowest and 4 the highest resolution).

be detected in most approaches. This is because their tiny size makes them get lost in the feature maps. Moreover, after several pooling layers or convolution layers with a stride greater than two, the feature maps corresponding to small objects are totally omitted. Therefore, information from the relevant context (e.g. multi-scale information or features around an occluded region) should be paid more attention, while information from the irrelevant regions should be paid less attention. That is why it is effective to use some form of context aware refinement procedure to remove what is irrelevant.

Instead of multi-context attention [32] using a single 1×1 convolution layer to obtain attention maps independently at different scales, we found that the attention maps in the multi-scale analysis were near each other (examples can be seen in Fig. 3). We therefore we added to the architecture a multi-scale attention module (MSAM).

Fig. 3 shows a traffic sign (in the bounding box) image and the corresponding attention maps from different receptive fields. Notice that none of the attention maps at different receptive fields obtains an accurate location for the traffic sign. Attention maps in the shallow layer (at low resolution–first box to the left of the traffic sign in Fig. 3) of the decoder have high activation outputs and inaccurate localization information, whereas attention maps in the deep layer (at high resolution–fourth box to the left of the traffic sign in Fig. 3) of the decoder have rich localization information and low activation outputs. We found by experiment that the multi-scale attention module can compensate for the accuracy in the attention map by using information from neighboring receptive fields.

We denote k-th scale feature maps with $W_k \times H_k \times D_k$ dimension as $\mathbf{x}_k$, where W_k, H_k, D_k are width, height and channel number in the feature maps. An encoder is applied at k-th scale to learn an attention map α_k measuring the pixel importance in the feature map at the k-th scale.

The basic structure of the attention part is a 3×3 convolution layer to get the activation response at each scale. The order of convolution layer is BN-ReLU-Conv for information propagation. Another 1×1 convolution layer is added at each scale to obtain the activation map $\mathbf{h}_k \in R^{W_k \times H_k}$.

Then, non-normalized attention map $\mathbf{z}_k$ is obtained by using activation maps at the neighboring scales

$$\mathbf{z}_k = f(\mathbf{h}_{k-1}, \mathbf{h}_k, \mathbf{h}_{k+1}), \tag{1}$$

where f is the recurrent combination function (which will be introduced in the subsection following).

A softmax function is applied to $\mathbf{z}_k$ to ensure all the attention weights sum to one.

$$\alpha_k = \frac{exp(\mathbf{z}_k)}{\sum_k exp(\mathbf{z}_k)} \tag{2}$$

Next, the output feature maps α_k in each branch are expanded to 3D tensor $\alpha_k^3 \in R^{W \times H \times D}$ and combined with the corresponding feature maps $\mathbf{x}_k$ to get the context vector $\mathbf{c} \in R^{W \times H \times D}$ by a Hadamard production:

$$\mathbf{c} = \sum_k \alpha_k^3 \odot \mathbf{x}_k \tag{3}$$

Given the proposed attention mechanism, the module can focus on certain local regions at each scale rather than treating all the regions equally. The input attention mechanism is a feed forward network that can be trained jointly with other components.

B. Recurrent Attention Map

We want to find attention maps containing contextual information for different objects of interest, where all the aforementioned contextual information can be obtained either from that objects higher level or lower level counterparts. This approach has less computationally complexity and is easier for training the network.

It is possible to seamlessly add the context aware refinement procedure to the simple network. The insight is that such a procedure can be implemented by using a novel recurrent convolution network. That is, contextual information at the neighboring scale can be selectively introduced into the current attention map.

In step $t = 1, 2, \ldots T$, we use deep residual learning and late fusion to combine context attention maps at scales $k-1$, $k, k+1$ to obtain the attention map at scale k in step $t+1$, and Eq (1) can be expressed as sum operation

$$\mathbf{h}_k^{t+1} = \mathbf{W}_{up}^k \mathbf{h}_{k-1}^t + \mathbf{h}_k^t + \mathbf{W}_{down}^k \mathbf{h}_{k+1}^t, \tag{4}$$

or concatenation operation

$$\mathbf{h}_k^{t+1} = \mathbf{W}_k \; Concate(\mathbf{W}_{up}^k \mathbf{h}_{k-1}^t, \mathbf{h}_k^t, \mathbf{W}_{down}^k \mathbf{h}_{k+1}^t) + \mathbf{h}_k^t, \tag{5}$$

where $\mathbf{W}_{up}^k$ and $\mathbf{W}_{down}^k$ are network parameters to fulfill attention map up-sampling and down-sampling. Sampling is implemented by sub-pixel convolution [33] and convolution (with stride 2). $\mathbf{W}_k$ is a 1×1 convolution layer which projects the concatenation result back to an image.

In the experiments, we found that the concatenation approach obtains a slightly better performance. The concatenation details can be seen in Fig. (4), where the left sub-figure is the architecture of the recurrent attention structure, and the right sub-figure is the detail in a specific step.

Compared to feature combination approaches such as hyper-feature [26] and recurrent rolling convolution [34], our approach works on the 2D attention map rather than the 3D feature maps, so that local context information from multiple layers can be combined effectively. The whole process is fully data driven and can be trained end-to-end.

C. Revised Base Network

DSSD uses the revised ResNet-101 [8] as the base network, which decreases the image size 16-fold. We argue that this is too constricting to detect traffic signs. Moreover, the semantic feature will be lost if the whole base network is deleted.

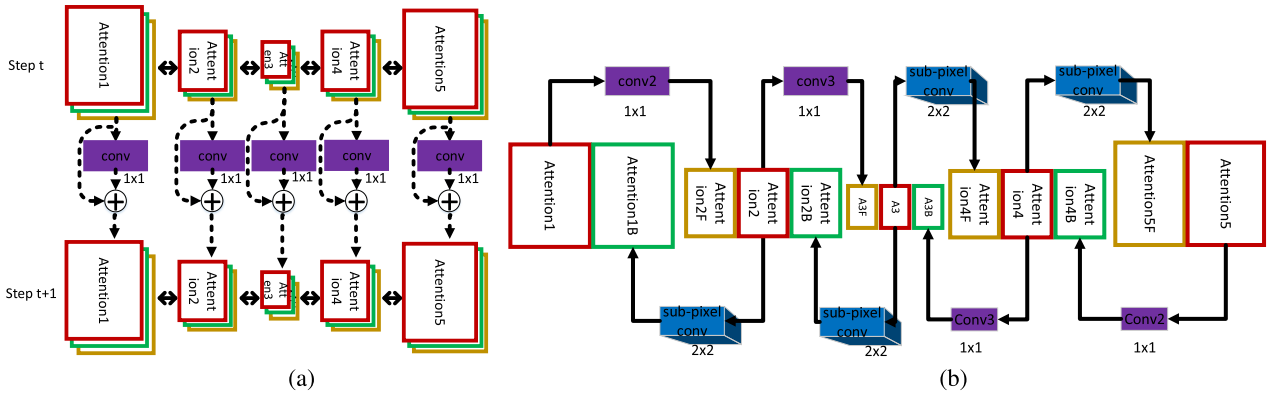


Fig. 4. The procedure to obtain the multi-scale recurrent attention. (a) The architecture of the recurrent attention structure, where the attention maps in step $k+1$ are obtained by residual learning of the attention maps in step k. (b) The detail for each step, where the attention map in each receptive field is obtained by the concatenation of the neighboring attention maps.

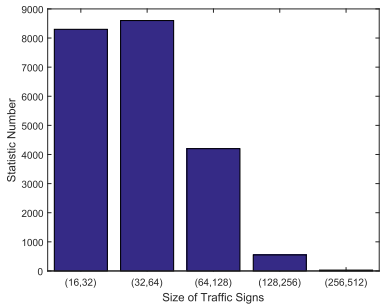


Fig. 5. Number of traffic sign instances of each size in the TT-100K data set.

We need a method to balance feature map resolution and semantic information extraction.

We gathered statistics of the sizes in pixels of traffic signs in the large scale data set, such as TT-100K data set, with results presented in Fig. 5. See in Fig. 5 that height of the smallest traffic signs appear as 16 pixels, and most of the pictured traffic signs are under 128 pixels.

We modified the base network in DSSD according to the sizes of traffic signs found in our data set. The first convolution layer with stride 2 in the *conv*4_*x* stage is modified to 1 to increase the resolution of the feature maps. Then the dilation of all convolution layers in *conv*4_*x* stage with a kernel size larger than 1 is increased from 1 to 2 to fix the gaps caused by the reduced stride. Though *conv*5_*x* is deleted in our revised base network, additional convolution layers in the encoder-decoder part can be trained to extract the high level semantic feature.

Now, given an image with the resolution of 2048 $\times$ 2048 pixels, which is the resolution of most images in the TT-100K data set, the revised base network can output feature maps with the resolution of 256 $\times$ 256 pixels, and each pixel in the feature maps corresponds to a proposal with the resolution of 16 $\times$ 16 pixels. As a result, small traffic signs far from the observation spot have the potential to be recognized.

V. Results

In this section, we compare the performance and efficiency of our multi-scale and recurrent attention approach to other approaches in traffic sign detection. We use the TT-100K and the GTSDB data sets for comparison of performance and efficiency (and for efficiency comparison, we also make hardware setup uniform).

A. Hardware and Software Environment

We use a workstation with Intel i7-4790 3.6GHz CPU, 32GB memory, and a single NVIDIA GTX Titan X graphics card. Our algorithm uses TensorFlow [35] to verify performance and computational efficiency.

B. Data Sets

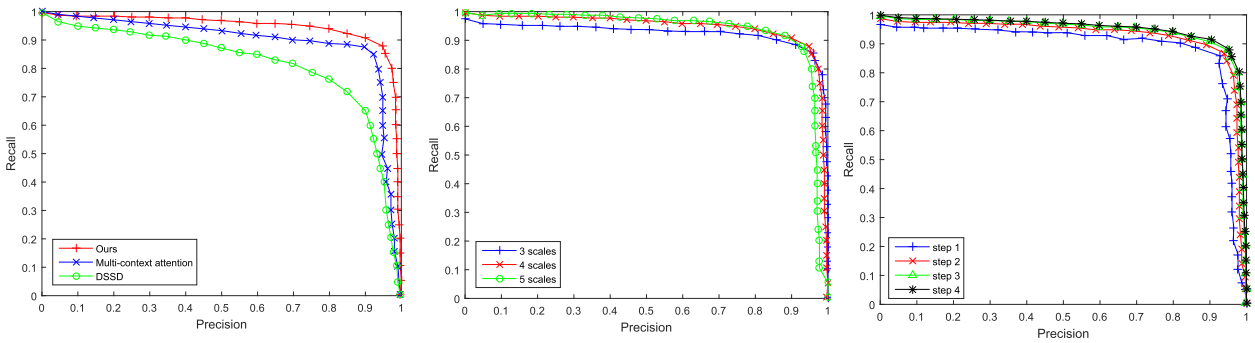
We verify our proposed approach on the GTSDB [9] and the TT-100K data set [10].

1) German Traffic Sign Detection Benchmark: GTSDB is the most widely-used data set in traffic sign detection. It contains 900 images, divided into 600 training images and 300 testing images, each with the size of 1360×800 pixels. Its traffic signs can be divided into four classes: 161 Prohibitory signs (usually of red color and circular shape), 49 Mandatory signs (usually of blue color and circular shape), 63 Danger signs (usually of red color and triangular shape), and other signs with different shapes and colors which cannot be classified into these three categories.

2) Tsinghua-Tencent 100K Data Set: It provides 100,000 images and 30,000 traffic sign samples which can be classified into three categories: Prohibitory (red circle with black information), Mandatory (blue circle with white information), and Warnings (yellow triangle with a black boundary and information). A typical traffic sign is about 80 $\times$ 80 pixels in a 2048 $\times$ 2048 pixel image, or just 0.2% area of the image.

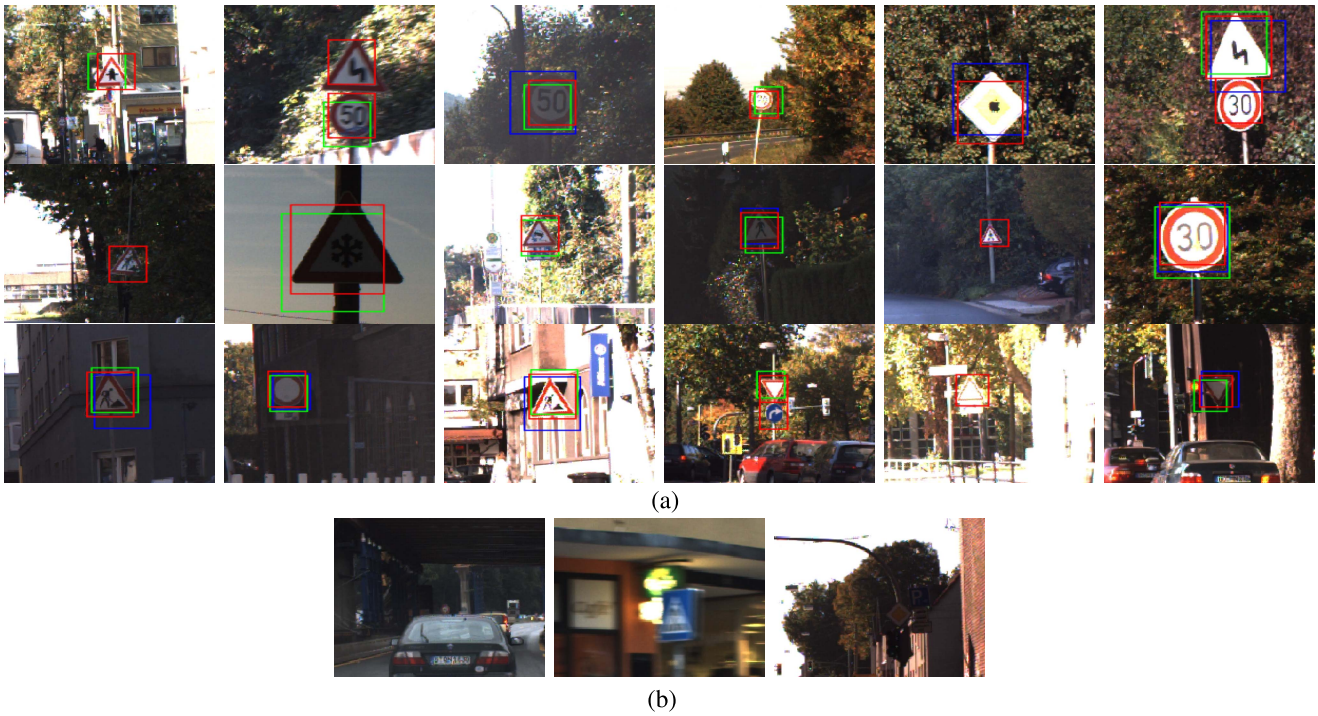
C. Evaluation Criterion

We use evaluation criteria that others have used in published research in order to compare our work on the same datasets to state-of-the-art approaches. For this reason, we use area under curve (AUC) values as the evaluation measure in the GTSDB data set, and precision-recall for the evaluation criteria for the TT-100K data set. Recall is the proportion of positive ground truth images being detected in the sample, and precision is



(a) (b) (c)

Fig. 6. Evaluation on attention mechanism, multi-scale analysis, and recurrent steps on the TT-100K data set. (a) Evaluation on attention. (b) Evaluation on scale number. (c) Evaluation on step number.



(a)

(b)

Fig. 7. Experiment Results on the GTSDB data set. (a) Successful detection results, bounding boxes are obtained by using Faster R-CNN (blue), DSSD (green), and our approach (red). (b) Unsuccessful detection results.

the proportion of all the detected examples that are positive ground truth. The precision/recall curve is computed from the ranked output.

D. Implementation Details

Down-sampling is performed by convolution layers with a stride of 2, and up-sampling is performed by sub-pixel convolution layers.

The parameters are chosen following the work of [20]. The whole network is trained with a Stochastic Gradient Descent (SGD) that has a momentum of 0.9, and weight decay of 0.0005. Learning rate is set to 0.01 for first 60K iterations, and 0.001 for the next 20K iterations. Each mini-batch contains 16 positive and 48 negative samples that are selected from four different training images. Positive samples are defined as having a minimum intersection over union (IoU) of 0.6 between the proposal bounding box and ground truth. Negative samples are selected from the cluttered background.

When evaluating the results, a threshold of 0.7 is used for the confidence score and an IoU of 0.6 is used for object localization.

E. Ablation Study

We considered four questions to determine general effectiveness and parameters of our method in preparation for comparison experiments. We used the TT-100K data set, with results in Fig. 6.

Questions 1 and 2: Is the attention mechanism effective at object detection? If yes, does the attention interaction get more discriminating information than the normal attention mechanism? To answer these questions, we compared our approach

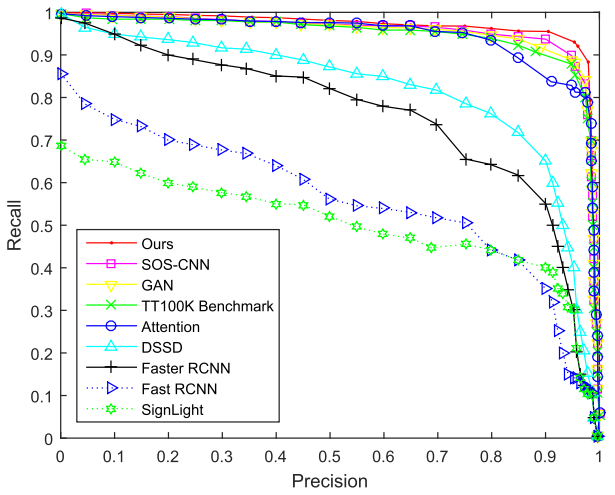


Fig. 8. Experimental Results (precision-recall curves) on the TT-100K data set.

TABLE II
EXPERIMENTAL RESULTS ON THE GTSDB. '-' INDICATES THAT THE RESULTS ARE NOT REPORTED. THE RESULTS OF ALL APPROACHES ARE OBTAINED FROM THE ORIGINAL PAPERS

Method	GTSDB dataset		
	Prohibitory (AUC)	Mandatory (AUC)	Danger (AUC)
Integral Channel Feature [37]	97.46	93.45	91.12
HOG+SVM [2]	99.43	95.01	97.22
HOG+Color [38]	99.29	96.74	97.13
Template Matching [40]	100.00	92.00	98.85
Coarse-to-fine [36]	100.00	100.00	99.91
Modified YOLOv2 [41]	96.81	94.02	96.12
SVM+CNN [39]	-	97.62	99.72
Faster R-CNN [42]	99.70	94.50	-
DSSD [8]	99.70	94.59	97.31
Boundary Estimation [43]	99.89	99.16	99.93
Ours	99.70	99.72	99.72

with the DSSD [8] and the multi-context attention [32]. The resulting precision/recall curve can be seen in Fig. 6(a). Multi-context attention out-performs DSSD by a large margin in average recall and precision, which demonstrates the effectiveness of the attention mechanism in object detection. Our multi-scale recurrent attention increases the average recall and precision by another 2.5% and 2.9% respectively, which demonstrates that our approach can effectively incorporate fine-grained details from local context to improve the traffic sign detection.

Question 3: How many scales should be processed in traffic sign detection? To answer this, schemes with various scale numbers are tested in the experiments, and the precision/recall curve can be seen in Fig. 6(b). If more scales are used in the scheme, the recall will improved. However, large size bounding boxes will lessen overall effectiveness by reducing precision. In our experiment, 4 scales show the best performance in the precision/recall curve. Hence, in the next experiments we use a scheme with 4 scales to compare the effectiveness and the efficiency.

Question 4: How many steps should be used to find the optimal result? We know that step number should balance the effectiveness and the efficiency. The attention map can be improved as the step number increases, which can be seen in Fig. 6(c). However, the performance improvement is limited when the step number t is greater than 3, while the computational complexity continues to increase according to the recurrent steps. Therefore, in the next experiments, the recurrent steps t is fixed to 3 to obtain the satisfactory performance without unnecessary computational load.

To recapitulate, our findings were that our approach is effective, even more effective than DSSD and multi-context attention. We learned that we should use 4 scales to compare effectiveness and efficiency, with recurrent steps set at 3 for optimum performance.

F. Experimental Results on the German Traffic Sign Detection Benchmark

We compare our approach to deep learning and other machine learning approaches. Experimental results are shown in Table II. In most of the traditional approaches [2], [36]–[38], proposals that exhibit a uniform value for the main colors of a sign are extracted first, then manually designed HOG features are extracted for an SVM classifier to detect the traffic signs. In 'SVM+CNN' [39], the image is first pre-processed to grey scale by an SVM and then fed into layers of CNN.

Manually designed features, such as HOG, and classical machine learning approaches, such as SVM, obtain satisfactory results owing to the discriminant shape and color of traffic signs. For example, the 'HOG+Color' approach obtains AUC value 99.29 in 'Prohibitory', 96.74 in 'Mandatory', and 97.13 in 'Danger'. The coarse-to-fine approach [36] is most effective, but its computational complexity is high – which makes it take more than 3500 milli-seconds to process a single frame. By contrast, our approach is dozens of times faster, at only 170 milli-seconds.

Some CNN-based approaches do not show notable improvement over manually-designed approaches in the traffic sign detection task. Faster R-CNN only gets AUC value 99.70 in 'Prohibitory' and 94.50 in 'Mandatory'. Constrained by the size of its convolution feature map output, Faster R-CNN is unable to clearly detect small objects. DSSD improves the performance by introducing the multi-scale analysis, but its effectiveness is limited because the ResNet-101 base network still decreases the feature map size dramatically. Boundary Estimation [43] tries to use edge cue as context information to assist the detector, and this is somewhat effective.

Our approach modifies the base network (see Fig. 2) in DSSD to increase the resolution of the feature maps, while semantic information is retained and local context information is obtained by using the multi-scale recurrent attention (see Fig. 4). Hence, our approach detects almost all the traffic signs.

Detection results of Faster R-CNN (in blue), DSSD (in green), and our approach (in red) are shown in Fig. 7.

The Fig. 7(a) compares different approaches. DSSD obtained better results than Faster R-CNN because DSSD can get semantic proposals directly from the high resolution feature maps. For example, DSSD can detect traffic signs with a height less than 60 pixels in the image, while the Faster R-CNN cannot detect such small size objects because

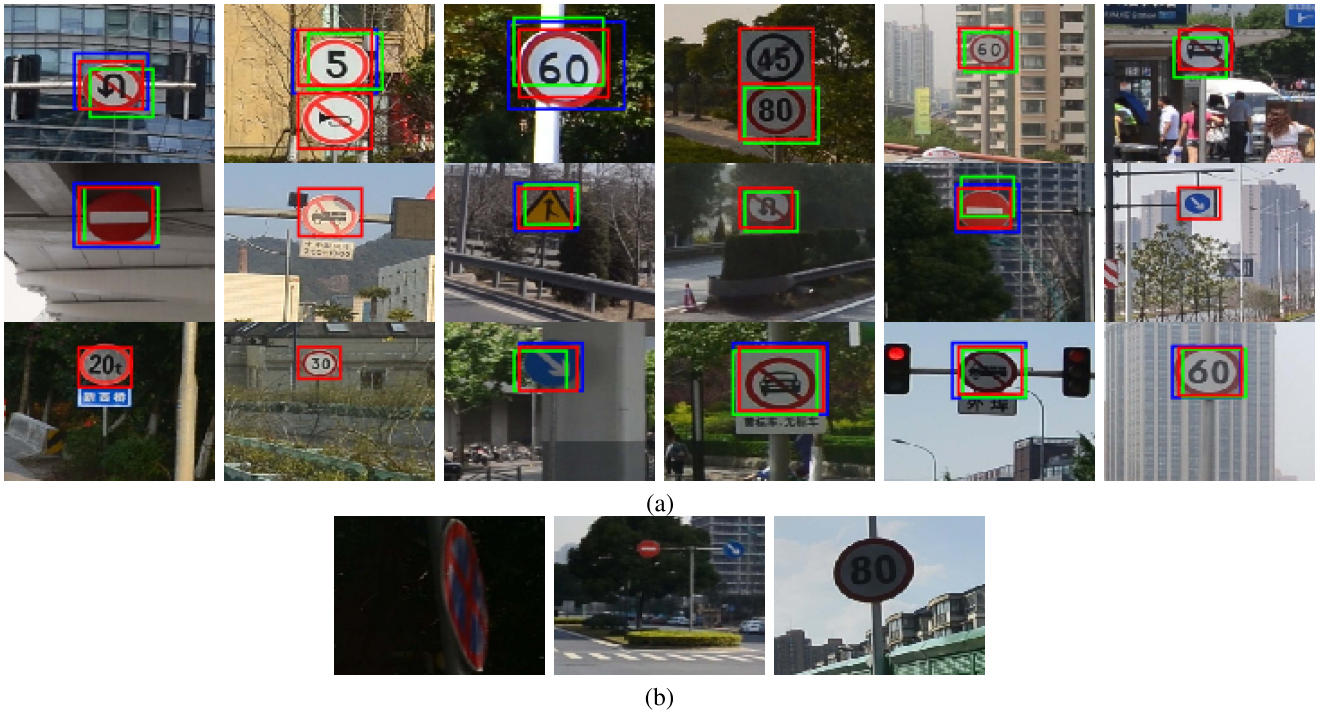


Fig. 9. Experiment result on the TT-100K data set. (a) Successful detection results, bounding boxes are obtained by using Faster R-CNN (blue), DSSD (green), and our approach (red). (b) Unsuccessful detection results.

its 5 pooling layers filter out useful local information. Our approach gets more stable results than the DSSD in part because the cluttered background near the object is removed so more salient context feature maps are obtained.

The Fig. 7(b) shows the unsuccessful detection results. We see that traffic sign in small size, blur of the image, and backlight, all deteriorate detection effectiveness. Our detection results could be improved by enlarging the amount of the training data and enriching the diversity of the samples (the GTSDB has only 600 training images).

G. Experimental Results on the Tsinghua-Tencent 100K Data Set

We discuss our approach vis a vis other approaches in Fig. 8 (accuracy) and Table III (timing), with some image output results pictured in Fig. 9.

Fig. 8 compares effectiveness of our approach to leading CNN-based-approaches. Fast R-CNN [44], Faster R-CNN and DSSD are general object detection approaches. 'TrafficSign&Light Detection' [45] uses multi-task learning to learn two classifier for traffic sign and light respectively. In 'TT-100K benchmark' [20], a end-to-end network simultaneously detects and classifies traffic signs. In 'Perceptual GAN' [46], perceived poor representations of small objects are transferred to super-resolved ones, and the discriminator competes with the generator to identify the output. In 'SOS-CNN' [47], the original image is down-sampled to form an image pyramid, and truncated SSD is employed in each level of its image pyramid. 'Attention' [48] also introduces the attention mechanism into the detection framework, and it uses attention to improve proposal generation in two-stage detection methods.

TABLE III

EFFICIENCY COMPARISON ON THE TT-100K DATA SET. THE RESULTS OF ALL APPROACHES ARE OBTAINED FROM THE ORIGINAL PAPERS

Method	time (ms)
TrafficSign&Light Detection [45]	**15**
Faster R-CNN [22]	231
DSSD [8]	620
Attention [48]	182
TT-100K Benchmark [20]	4081
Perceptual GAN [46]	600
Ours	658

It can be seen in Fig. 8 that Fast R-CNN, Faster R-CNN and DSSD cannot obtain satisfactory results due to the reduced resolution in the base network and overlooking context information.

By using information from different tasks (refer again to Fig. 8), the 'TT-100K Benchmark' approach improves detection effectiveness, that is, the precision is 84.8% when the recall is 92.3%. 'Perceptual GAN' raises precision by 1.0% owing to its effective generator which acquires super-resolved feature maps for things like traffic signs. Based on the SSD, 'SOS-CNN' employs multi-scale analysis in both image and feature maps, and as a result, it effectively improves recall. Our approach also employs multi-scale analysis, and the recall and precision can be improved about 1.0% with the use of local context.

We also compare the efficiency of different approaches in Table III. Here, we checked the hardware configuration of each approach and converted the process time to a general platform with a single NVIDIA GeForce GTX TITAN X GPU with 12GB to make the performance comparison consistent. 'Perceptual GAN' takes about 600 milli-seconds to process an

image of 2048 × 2048 pixels. The authors of 'SOS-CNN' do not give the processing speed, but only mention that it is time consuming. Our method is not as fast as 'Perceptual GAN', but GPU computing is developing fast, and our approach would speed up if more CUDA units were deployed. Even so, our method can be adapted to pixel-size of traffic signs in different applications to achieve excellent detection performance.

Actual results are shown in Fig. 9. The Fig. 9(a) compares different approaches. Our method is robust to traffic sign shape, color, and cluttered background. If lighting conditions are not too bad to distinguish traffic sign, feature maps in the foreground can be extracted pretty well.

The Fig. 9(b) shows the unsuccessful detection results. Notice that traffic signs that are foreshortened, tiny or backlit, are not well detected by any method. In the future, we will extend our method to handle situations where the lighting is uneven, perspective irregular or the image is blurry. We will also work to increase efficiency.

VI. Conclusion

We present here a new method that we call multi-scale and recurrent attention to detect the traffic sign in photos or video. Our method is successful because it improves the use of context within the image. We show a 5.2% AUC value improvement over state-of-the-art methods in the German Traffic Sign Detection Benchmark and a 1.0% precision & recall improvement in the Tsinghua-Tencent 100K data set.

References

[1] 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 *Proc. IVS*, Jun. 2005, pp. 255–260.

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

[3] M. A. Garcia-Garrido, M. A. Sotelo, and E. Martin-Gorostiza, "Fast traffic sign detection and recognition under changing lighting conditions," in *Proc. ITSC*, Sep. 2006, pp. 811–816.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[5] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "CRAFT objects from images," in *Proc. CVPR*, Jun. 2016, pp. 805–813.

[6] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," in *Proc. ICCV*, Oct. 2017, pp. 117–125.

[7] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. NIPS*, 2014, pp. 2204–2212.

[8] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. (2017). "DSSD : Deconvolutional single shot detector." [Online]. Available: https://arxiv.org/abs/1701.06659

[9] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in *Proc. IJCNN*, Aug. 2013, pp. 1–8.

[10] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proc. CVPR*, Jun. 2016, pp. 2110–2118.

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

[12] W. G. Shadeed, D. I. Abu-Al-Nadi, and M. J. Mismar, "Road traffic sign detection in color images," in *Proc. ICECS*, vol. 2, Dec. 2003, pp. 890–893.

[13] G. Rätsch, T. Onoda, and K. Müller, "Soft margins for AdaBoost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, Mar. 2001.

[14] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

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

[16] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3D localisation," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 633–647, Apr. 2014.

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

[18] V. John, K. Yoneda, B. Qi, Z. Liu, and S. Mita, "Traffic light recognition in varying illumination using deep learning and saliency map," in *Proc. ICITS*, Oct. 2014, pp. 2286–2291.

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

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

[21] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. ICCV*, Dec. 2015, pp. 1395–1403.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, Apr. 2015, pp. 365–374.

[24] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. CVPR*, Jun. 2015, pp. 447–456.

[25] P. Hu and D. Ramanan. (2016). "Finding tiny faces." [Online]. Available: https://arxiv.org/abs/1612.04402

[26] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. CVPR*, Jun. 2016, pp. 845–853.

[27] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. (ECCV)*, Jun. 2016, pp. 21–37.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belon, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 2117–2125.

[29] H. Li, Y. Liu, W. Ouyang, and, X. Wang. (2017). "Zoom out-and-in network with recursive training for object proposal." [Online]. Available: https://arxiv.org/abs/1702.05711

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016, pp. 630–645.

[31] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[32] X. Chu, W. Yang, W. Ouyang, C. Ma, A. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, Feb. 2017, pp. 1831–1840.

[33] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. CVPR*, Jun. 2016, pp. 1874–1883.

[34] J. Ren *et al.*, "Accurate single stage detector using recurrent rolling convolution," in *Proc. CVPR*, 2017, pp. 67–75.

[35] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467

[36] G. Wang, G. Ren, Z. Wu, Y. Zhao, and L. Jiang, "A robust, coarse-to-fine traffic sign detection method," in *Proc. IJCNN*, Aug. 2013, pp. 101–105.

[37] Y. Yang and F. Wu, "Real-time traffic sign detection via color probability model and integral channel features," in *Proc. CCPR*, 2014, pp. 545–554.

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

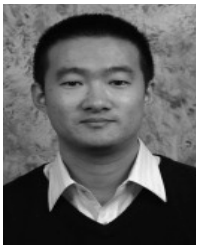
[39] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu, "Traffic sign detection based on convolutional neural networks," in *Proc. IJCNN*, Aug. 2013, pp. 1–7.

[40] M. Liang, M. Yuan, X. Hu, J. Li, and H. Liu, "Traffic sign detection by ROI extraction and histogram features-based recognition," in *Proc. IJCNN*, Aug. 2013, pp. 201–208.

[41] J. Zhang, M. Huang, X. Jin, and X. Li, "A real-time chinese traffic sign detection algorithm based on modified YOLOv2," *Algorithms*, vol. 10, no. 4, pp. 127–133, Nov. 2017.

[42] E. Peng, F. Chen, and X. Song, "Traffic sign detection with convolutional neural networks," in *Proc. ICCSSP*, 2016, pp. 214–224.

[43] H. S. Lee and K. Kim, "Simultaneous traffic sign detection and boundary estimation using convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1652–1663, May 2018.
[44] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.
[45] A. Pon, O. Andrienko, A. Harakeh, and S. Waslander, "A hierarchical deep architecture and mini-batch selection method for joint traffic sign and light detection," in *Proc. CRV*, 2018, pp. 122–130.
[46] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. CVPR*, Jul. 2017, pp. 1222–1230.
[47] Z. Meng, X. Fan, X. Chen, M. Chen, and Y. Tong. (2017). "Detecting small signs from large images." [Online]. Available: https://arxiv.org/abs/1706.08574
[48] Y. Lu, J. Lu, S. Zhang, and P. Hall, "Traffic signal detection and classification in street views using an attention model," *Comput. Vis. Media*, vol. 4, no. 3, pp. 253–266, Sep. 2018.



Yan Tian received the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2011. He held a post-doctoral research fellow position (2012–2015) in the Department of Information and Electronic Engineering, Zhejiang University, Hangzhou, China. He is currently an Associate Professor of computer science and technology with the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. His research interests are in computer vision.

Judith Gelernter received the Ph.D. degree in information science from Rutgers University in 2008. She did research, from 2008 to 2015, at the Language Technologies Institute, School of Computer Science, Carnegie Mellon University. She was a Research Scientist with the Information Technology Laboratory, National Institute of Standards and Technology, from 2015 to 2018. She is currently affiliated with Rutgers University.

Xun Wang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2006. He is currently a Professor with the School of Computer Science and Information Engineering, Zhejiang Gongshang University, China. His current research interests include mobile graphics computing, computer vision, image/video processing, virtual reality, and visual analytics. He is a member of the IEEE and ACM, and a senior member of CCF.

Jianyuan Li received the Ph.D. degree in Computer Science and Technology from Tongji University, Shanghai, China, in 2012. He held a post-doctoral research fellow position with Enjoyor Co. Ltd., Hangzhou, China, from 2014 to 2017, and where he is now the Chief Technical Officer. His research interests include machine learning and data mining.

Yizhou Yu received the Ph.D. degree from the University of California, Berkeley, CA, USA, in 2000. He is currently a Professor with Zhejiang University. He was a faculty member with the University of Illinois, Urbana–Champaign from 2000 to 2012. He received the 2002 National Science Foundation CAREER Award. He is an IEEE Fellow and is on the editorial boards of *IET Computer Vision*, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS and the *International Journal of Software and Informatics*. His current research interests include deep learning methods for computer vision, computational visual media, geometric computing, and video analytics.