A DEEP LEARNING FRAMEWORK FOR SMALL TRAFFIC LIGHT RECOGNITION IN TRAFFIC SCENE IMAGE

HOANH NGUYEN

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

E-mail: nguyenhoanh@iuh.edu.vn

ABSTRACT

Traffic light recognition plays a crucial role in intelligent transport systems. In traffic scene images, traffic light instances usually occupy a small region. Thus, recent state-of-the-art object detectors such as Faster R-CNN and SSD obtain low accuracy on traffic light recognition in traffic scene images. This paper presents a deep learning framework for traffic light recognition in traffic scene image. Considering that feature maps at shallow layers have higher resolution which will improve small traffic light detection, and feature maps at deep layers contain more discriminative representation which will improve traffic light classification task, this paper designs a feature fusion subnet for feature extraction to solve the problem of small traffic light detection. The feature fusion subnet fuses feature maps at different layers. Thus, the feature fusion subnet not only can preserve the information of small traffic lights but also enhance the semantic information. Furthermore, a detection subnet is designed at the detection prediction stage. The detection subnet includes multiple detection layers, and each layer performs detection predictions with a coarse-to-fine detection strategy. The coarse-to-fine detection strategy is applied to improve the classification performance of the detection network. The proposed approach is evaluated on Bosch Small Traffic Lights dataset. Experimental results show that the proposed approach obtains higher accuracy compared with recent state-of-the-art detectors such as Faster R-CNN and SSD.

Keywords: *Traffic Light Recognition, Deep Learning, Intelligent Transportation Systems, Feature Fusion, Advanced Driving Assistance System*

1. INTRODUCTION

Traffic light recognition is a key problem in intelligent transport systems such as advanced driving assistance system and an autonomous vehicle system. Many methods for traffic light recognition have been proposed in recent years. Normally, a traffic light recognition approach includes two stages: proposal generation and proposal classification. Traditional methods for traffic light recognition usually use hand-crafted features such as colour and shape for generating traffic light proposals and classifying these proposals. The main drawbacks of traditional methods are runtime and accuracy. Traditional methods are not real-time capable, often requiring several seconds for processing an image. On the other hand, hand-crafted features for traffic light detection are not able to obtain sufficient accuracy. In addition, traditional methods usually require the traffic lights to be at least a certain size for the algorithm to work, or on a distinctive background such as suspended traffic lights in front of the sky or assume the existence of maps that contain prior knowledge about the locations of all traffic lights in the environment.

With the success of deep convolutional neural networks (CNNs) in recent years, deep CNN-based methods for traffic light recognition have been proposed and achieved better performance compared with traditional methods. Deep CNN-based methods first apply a series convolution layers for generating base feature maps. Then, traffic light candidates are generated based on the base feature maps. Finally, a CNN classifier is applied to classify traffic light candidates. Popular deep CNN-based methods include Fast R-CNN [23], Faster R-CNN [24], YOLO [16], and SSD [25]. Fast R-CNN uses selective search to generate object candidates and applies fully connected neural network to classify objects. YOLO uses a simple CNN approach to achieve real-time processing by enhancing detection

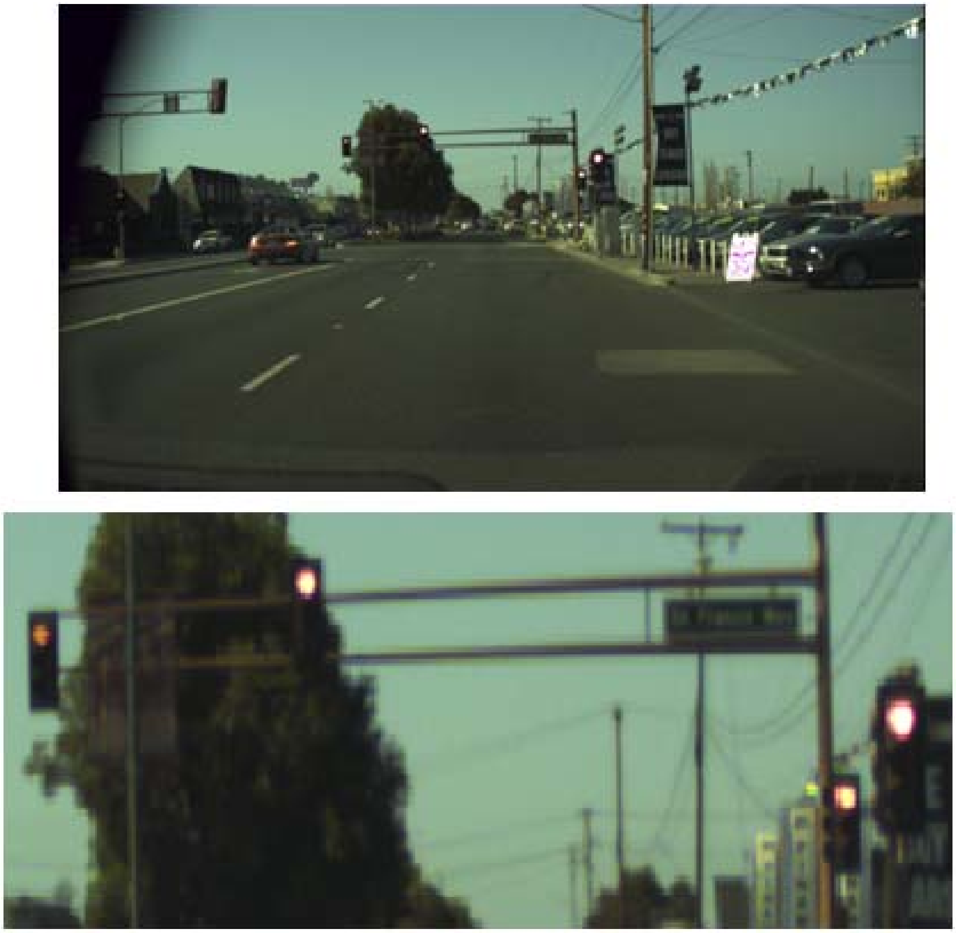


Figure 1: Traffic Light Instances (Bottom) in Traffic Scene Image (Top). Traffic Light Instances Have Resolution of about 8×14 Pixels, and Original Image Has Resolution of 1280×720 Pixels.

accuracy and reducing computational complexity attaining. Faster R-CNN replaces the selective search in Fast R-CNN by region proposal network. The region proposal network is a fully convolutional network that simultaneously predicts the object boundaries and object scores at each position. SSD uses multiple sized convolutional feature maps to achieve a better accuracy.

The drawback of deep CNN-based methods for traffic light recognition currently is the detection of small traffic lights. It is mainly caused by pooling operations, which increase the receptive field and reduce the computational effort. However, pooling also decreases the image resolution leading to difficulties for accurate localization of small objects. On the other hand, traffic light instances usually occupy small regions in traffic scene images. Figure 1 shows an example of traffic lights in traffic scene image. As shown, the image is taken at the resolution of 1280×720, and the traffic light instances have the resolution of about 8×14 pixels. Thus, deep CNN-based methods for traffic light recognition achieve low detection accuracy. To tackle this problem, this paper presents a deep learning-based framework for small traffic light recognition in traffic scene images. In the proposed framework, a feature fusion subnet which fuses feature maps at different layers of the base network is designed to enhance the accuracy on small traffic lights. In addition, a detection subnet is designed at the detection prediction stage. The detection subnet includes multiple detection layers, and each layer performs detection predictions with a coarse-to-fine detection strategy. The coarse-to-fine detection strategy is applied to improve the classification ability of the detection network.

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

2. RELATED WORK

2.1 Traditional Traffic Light Detection Method

Traditional traffic light detection methods usually use a simple thresholding in different color spaces such as RGB and HSV and other hand-crafted features such as shape and texture to locate traffic lights in an image. Chiang et al. [1] proposed a novel genetic approach to handle appearance changes caused by perspective shape deformations and partial occlusions. In addition, a spatial texture layout feature was designed to handle illumination variations under different weather conditions and eliminate false alarms from irrelevant scene backgrounds. Diaz-Cabrera et al. [2] proposed a novel technique to detect suspended traffic lights based on colors and features such as black area of traffic lights or area of lighting lamps. Gong et al. [3] first used the threshold segmentation method and the morphological operation to extract the candidate region of the traffic light. Then, the recognition algorithm of the traffic light based on machine learning is employed. In [4], the authors first extracted candidate blobs of traffic lights from RGB color image. Then shape filter was used for noise reduction by using information of blobs such as length, area, area of boundary box. Lindner et al. [5] proposed a method for real-time detection and recognition of traffic signals including three main modules: detection, tracking, and sample-based classification. Furthermore, additional sensor information such as vehicle data, GPS, and enhanced digital maps were used to enhance the performance and robustness of the system. De Charette et al. [6] introduced a new real-time traffic light recognition system for on-vehicle camera applications based on a spot detection algorithm. The system was able to detect lights from a high distance with the main advantage of being not so sensitive to motion blur and illumination variations. In [7], the authors presented a probabilistic framework for increasing online object detection performance when given a semantic 3D scene prior. The framework was applied to the task of traffic light detection for autonomous vehicles. In [8], presented a method for automatically mapping the three-dimensional positions of traffic lights and robustly detecting traffic light state onboard cars with cameras. Müller et al. [9] used a two-camera setup for traffic light recognition. A stereo camera is used for far and moderate range, while for close range a wide-angle camera is used.

2.2 Traffic Light Detection Based on Deep Learning

In recent years, deep learning has become very popular in vision-based object recognition due to its high performances of classification and detection. Thus, some traffic light recognition approaches based on deep learning have been proposed and showed better performance compared with traditional approaches. Bach et al. [10] proposed a traffic light detection system consisting of multiple on-board cameras. The system was based on tracking techniques using a Labeled Multi-Bernoulli filter in combination with the fusion of classifications based on the Dempster-Shafer theory of evidence. Behrendt et al. [11] proposed a complete system consisting of a traffic light detector, tracker, and classifier based on deep learning, stereo vision, and vehicle odometry which perceives traffic lights in real-time. In [12], the authors proposed a camera-based system for real-time detection and classification of traffic lights. The system was able to detect traffic lights on the whole camera image without any pre-segmentation. This was achieved by classifying each fine-grained pixel region of the input image and performing a bounding box regression on regions of each class. Saini et al. [13] presented a vision-based traffic light structure detection and convolutional neural network-based state recognition method. In this approach, traffic light candidate regions are generated by performing HSV based color segmentation, and traffic lights are recognized by using CNN. In addition, Histogram of Oriented Gradients features were extracted for each region and traffic light structures were validated using Support Vector Machine to further validate the traffic light candidate regions. Lee et al. [14] proposed a method that combines a conventional approach and a DNN approach. The conventional approach is fast but weak to false positives, while the DNN approach is not suitable for detecting small objects but a very powerful classifier. In [15], the authors applied the state-of-the-art, real-time object detection system You Only Look Once, (YOLO) [16] on the public LISA Traffic Light dataset [17]. This dataset contains a high number of annotated traffic lights, captured in varying light and weather conditions. Kim et al. [26] proposed to apply six color spaces and three types of network models to investigate how to design a deep learning-based high-performance traffic light detection system. The simulations showed that the best performance is achieved with the combination of RGB color space and Faster R-CNN model.

3. PROPOSED APPROACH

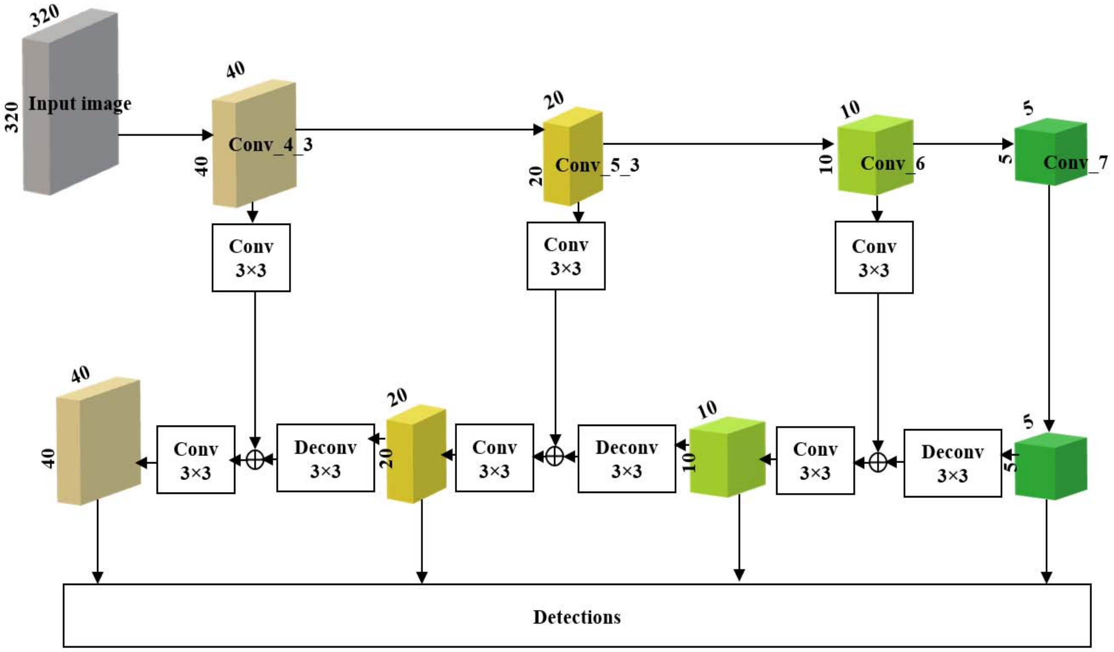


Figure 2: The Structure of The Feature Fusion Module Used in This Paper.

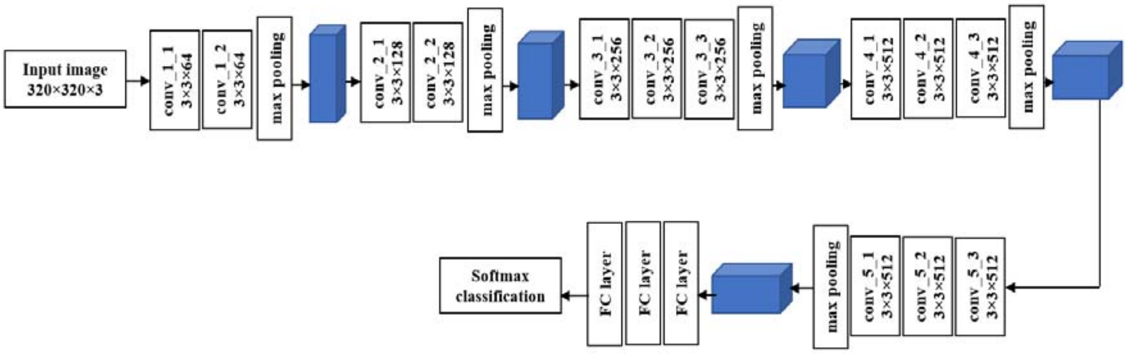


Figure 3: The Architecture of VGG-16 Network.

3.1 Feature Extraction with Feature Fusion

In deep CNN, feature map generated by convolutional layer becomes smaller as the network deepens, and the information of small objects is also lost. On the other hand, shallow convolutional layers generate high resolution feature maps with less discriminative representations. Using only feature map at shallow or deep convolutional layer will cause a low detection accuracy for small objects. Considering that feature maps at shallow layers have higher resolution which will improve small object detection, and feature maps at deep layers contain more discriminative representation which will improve classification task, this paper designs a feature fusion network based on FPN [18] to solve the problem of small traffic light detection. The feature fusion network fuses feature maps at different layers and independently predicts object position at each feature layer. Thus, the feature fusion network not only can preserve the information of small traffic lights but also enhance the semantic information. Figure 2 illustrates the structure of the feature fusion network used in this paper. As shown, VGG-16 architecture [19], which contains 13 convolution layers and 4 max pooling layers as shown in Figure 3, is adopted to generate base feature maps. To improve the detection performance, this paper discards the last max pooling layer, and adds two convolutional layers (conv_6 and conv_7) at the end of the network. The

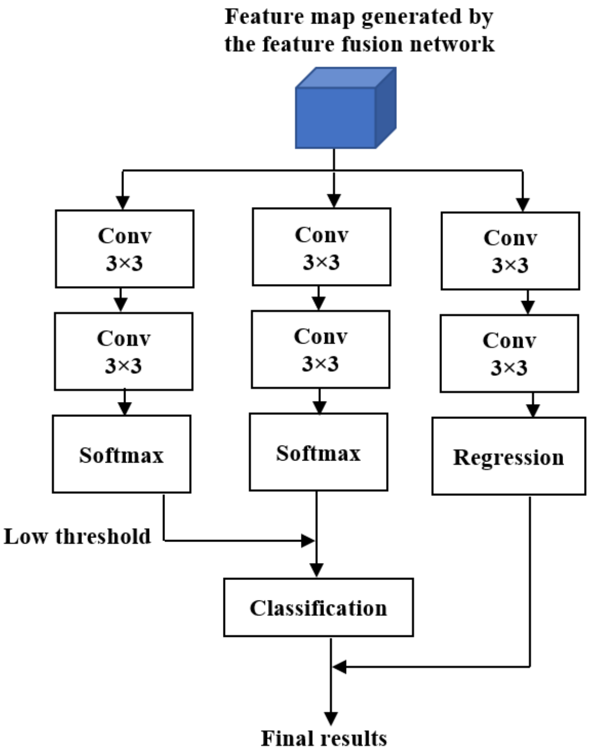


Figure 4: The Structure of a Detection Layer.

two additional convolutional layers reduce the resolution of the feature map to half in sequence. With the deepening of the network, the resolution of the output feature map gets smaller, but the semantic information is more abundant. Four feature maps at layers conv_4_3, conv_5_3, conv_6 and conv_7 are used to fuse to generate fused feature maps. The strides of these feature maps are 8, 16, 32 and 64 respectively. Supposing that the input size of the network is 320×320 pixels, the resolutions of the feature maps at these layers are 40×40, 20×20, 10×10, and 5×5 respectively. At each fusing process, the high-level feature map is upsampled by a factor of 2 using deconvolution layer, and then it is combined with the corresponding previous feature map in the base network by using element addition. The previous feature map in the base network would be subjected to a 3×3 convolution layer to change the dimensions, which should be the same as the dimensions in the next layer. This process is repeated iteratively until the finest feature map is generated. The deepest feature map generated by conv_7 layer is directly considered as the feature map of the last detection layer input. After each fusing process, a 3×3 convolution layer is used on each merged feature map to generate the fused feature map in order to eliminate the aliasing effect of upsampling. The fused feature map serves as the input of the detection layer.

3.2 Detection Network with Coarse-to-Fine Detection Strategy

The detection network includes multiple detection layers, and each layer performs detection predictions with a coarse-to-fine detection strategy. The coarse-to-fine detection strategy is applied to improve the classification ability of the detection network. Figure 4 illustrates the structure of a detection layer. As shown, feature map generated by the feature fusion network is fed into three parallel branches. The first branch includes 3×3 convolution layers and a softmax layer for coarse classification. The second branch includes 3×3 convolution layers and a softmax layer for fine classification. The final branch includes 3×3 convolution layers for bounding box regression. At each position on the input feature map, the bounding box regression branch predicts the offsets relative to the anchor shapes, and the coarse classification branch predicts the confidence which indicates the presence of traffic light in each of bounding boxes. Then, the image block contained in the bounding box which has a confidence higher than the threshold is further classified into traffic light classes and background class by the fine

Table 1: The Size of Traffic Light Instances in Test Set.

	Minimum (pixel)	Average (pixel)	Median (pixel)	Maximum (pixel)
Width× Height	1.875×3.250	9.430×26.745	8.500×24.500	48.375×104.500
Area	11.718	313.349	212.109	4734.000

classification branch to obtain the final detection result.

3.3 Loss Function

The proposed framework is trained in an end-to-end fashion using a multi-task loss function. Beside the conventional classification loss for fine classification and regression loss for bounding box regression, this paper adds additional loss function for coarse classification. Thus, the multi-task loss function is defined as follow:

$$L = \frac{1}{N_{cls}}\sum L_{cls}(b_i, b_i') + \frac{1}{N_{reg}}\sum L_{reg}(t_i, t_i') + \frac{1}{M_{cls}}\sum L_{cls}(b_j, b_j') \qquad (1)$$

where i represents the index of an anchor from the coarse classification subnet and the bounding box regression subnet in a minibatch; j represents the index of an anchor from the coarse classification subnet; b_i represents the predicted probability that the anchor i is a traffic light. If the anchor is positive, the ground truth label b_i' is 1 and vice versa; t_i is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t_i' is that of the ground-truth box associated with a positive anchor. N_{cls} represents the number of positive and negative anchors from the coarse classification subnet in the minibatch. N_{reg} represents the number of positive anchors from the bounding box regression subnet in the minibatch, and M_{cls} represents the number of positive and negative anchors from the fine classification subnet in the minibatch.
In (1), the classification loss L_{cls} is the log loss, and the regression loss is the smooth L1 loss as follows:

$$L_{cls}(b_i, b_i') = -\log(b_i b_i' + (1 - b_i)(1 - b_i')) \qquad (2)$$

$$L_{reg}(t_i, t_i') = smooth_{L1}(t_i, t_i') \qquad (3)$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & if\ \ |x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \qquad (4)$$

For generating training samples in bounding box regression and coarse classification subnet, this paper randomly samples 256 anchors from one image as a minibatch, where the ratio between positive and negative anchors is 1:1. The anchors that have Intersection over Union (IoU) scores larger than 0.5 with any ground-truth bounding box are selected as positive anchors, while anchors with IoU lower than 0.3 are selected as negative anchors. The anchors with the highest IoU scores are also regarded as positives to make sure that every ground-truth box has at least one positive anchor. For generating training samples in fine classification subnet, the anchors selected from coarse classification subnet are further given positive and negative label. However, the IoU scores for selecting the positive anchor is raised from 0.5 to 0.7. The larger IoU scores means that the positive anchor selected is closer to the ground-truth bounding box, which makes the classification more precise.

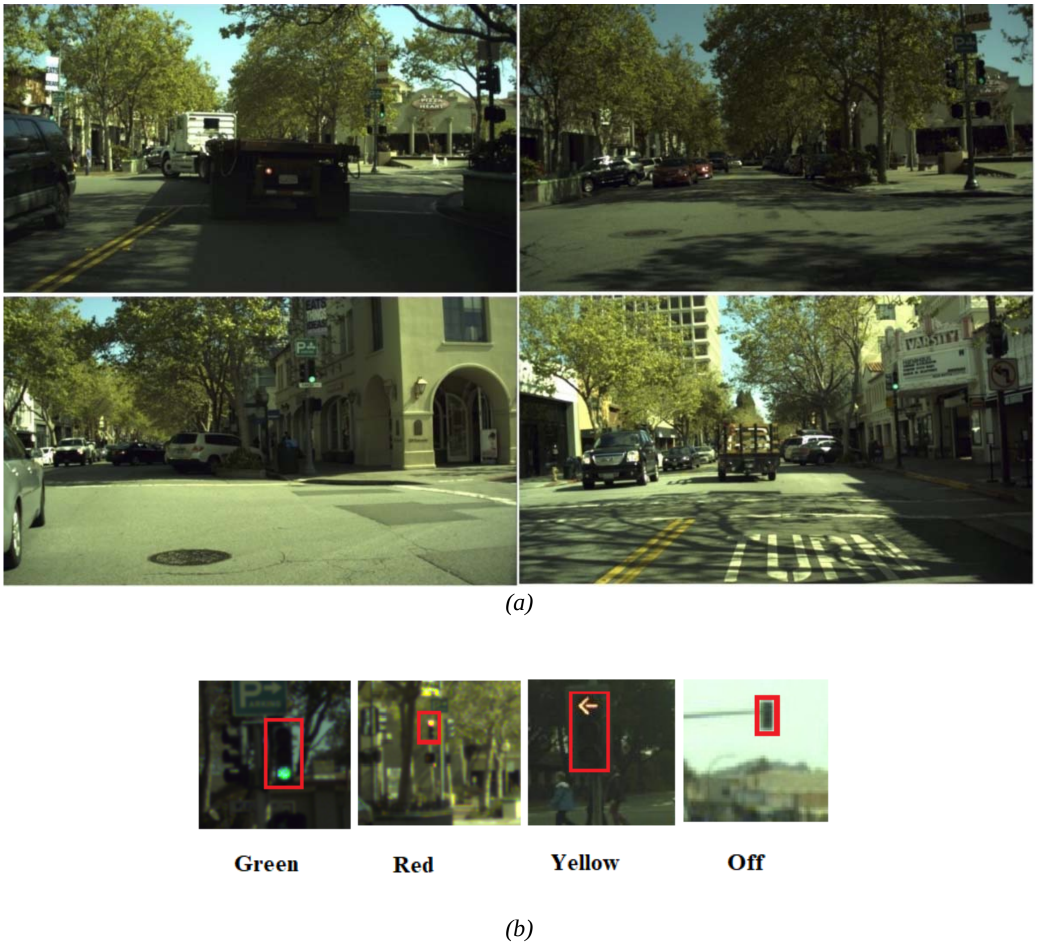
4. EXPERIMENTAL RESULTS

4.1 Dataset

In this paper, Bosch Small Traffic Lights dataset [11] is used to evaluate the performance of the proposed method on traffic light recognition. This dataset aims for evaluating small traffic light detection in large images. The images included in this dataset contain small traffic lights under diverse driving environments such as busy street scenes inner-city, suburban multilane roads with varying traffic density, and dense stop-and-go traffic as shown in Figure 5 (a). There are total 13,427 images at a resolution of 1280×720 pixels with about 24,000 annotated traffic lights in this dataset. All images are divided into training and testing set. The training set includes 5,093 images with 10,756 annotated traffic lights. There are 15 different labels in the training set. To increase the number of images in the training set, this paper adopts data augmentation strategies used in [20]. The testing set includes 8,334 consecutive images with 13,486 annotated traffic lights. There are 4 different labels in the testing set, including “red”, “yellow”, “green”, and “off”. Figure 5 (b) shows example images of these labels. The size of traffic light instances in each image of the testing set is varied as shows in Table 1. This dataset provides a challenging benchmark for detecting traffic lights or more generally, small objects in images.

4.2 Metrics

This paper adopts average precision (AP) and mean average precision (mAP) to evaluate the performance of the proposed method on traffic light



(a)

(b)

Figure 5: Example Images in The Bosch Small Traffic Lights Dataset (a) and Four Different Labels in The Test Set (b).

detection. These metrics have been widely used in VOC challenge [21] and the COCO 2015 detection challenge [22]. For better expression, true positives, false positives, and false negatives are denoted as TP, FP and FN. A predicted bounding box is considered to be TP if the IoU between predicted bounding box and ground truth is larger than 0.5. Otherwise, it would be considered as FP. FN denotes that the actual annotated object has no predicted bounding box. Specifically, TP means the correct retrieval of an object. The precision indicator measures the proportion of detections that are TP and the recall indicator measures the fraction of practical annotations that are classified correctly. The calculation formulas of precision and recall are as follows:

$$Precision = \frac{TP}{(TP+FP)} \tag{5}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{6}$$

AP is the precision averaged across all values of recall between 0 and 1. Here, AP is calculated by averaging the interpolated precision over eleven equally spaced interval of recall value [0, 0.1, 0.2, . . .0.9, 1.0]. To evaluate the performance for two or more classes, the average of AP, mAP is calculated by averaging APs over every class. This paper conducts all of the experiments on a desktop computer which is equipped with an Intel Core I7-8700 CPU (6 Core, 3.2 GHz), 32 GB of memory, a Nvidia GeForce GTX 1080 GB GPU and Window 10 OS.

4.3 Performance Results

To evaluate the effectiveness of the proposed method, this paper conducts experiments on Bosch Small Traffic Lights dataset and compares the recognition results with other object detection algorithms, including Fast R-CNN [23], Faster R-CNN [24], and SSD [25]. Fast R-CNN is a two-stage object detection framework based on deep CNN. A proposal generation stage based on sliding-window

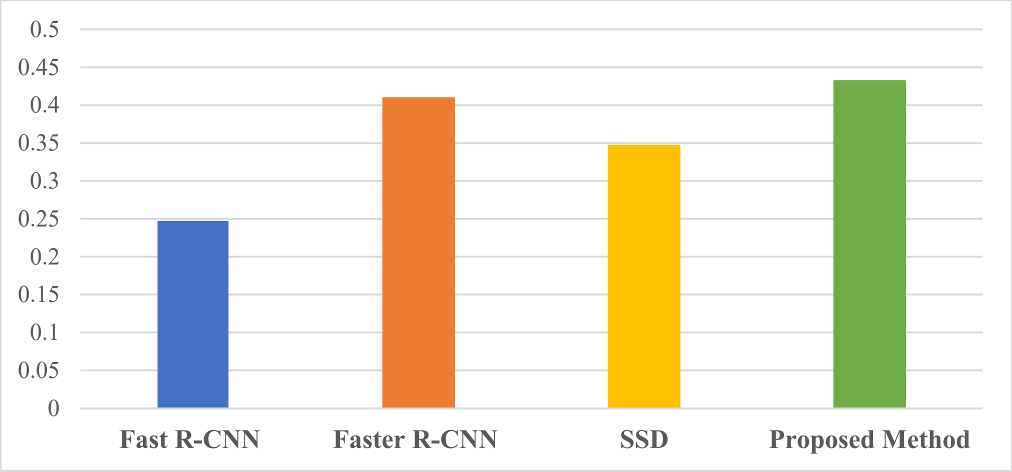


Figure 6: The mAP Results of The Proposed Method and Other Methods on Bosch Small Traffic Lights Dataset.

Table 2: Comparisons of AP for Several Methods.

Method	AP (%)			
	Green	Red	Yellow	Off
Fast R-CNN	51.7	30.6	6.22	10.4
Faster R-CNN	75.2	58.4	14.6	15.5
SSD	68.3	50.1	8.0	12.1
Proposed method	80.4	60.2	16.5	15.6

method is first used to generate proposals, and a deep CNN-based classifier is then used to classify each proposal. Faster R-CNN proposed to use region proposal network to replace sliding-window method in Fast R-CNN. Region proposal network provides a fast and efficient approach for generating proposals. SSD is a one-stage object detection framework, which performs detection predictions at multi-scale of the base feature maps. Comparing with Faster R-CNN, SSD achieves comparable detection performance while being faster.

Figure 6 shows the mAP results of the proposed method and other methods on Bosch Small Traffic Lights dataset. It is evident that Faster R-CNN obtains better performance than Fast R-CNN and SSD. This is mainly because the region proposal network employed by Faster R-CNN is more suitable to detect small traffic lights. SSD detects traffic lights at deeper layers of the base network where the resolution of feature maps is small, thus destroying the semantic information of small traffic lights. Fast R-CNN uses the last convolution layer of the base network for locating proposals, thus the ability of detecting multi-scale objects is limited. Meanwhile, the proposed method is compared with other methods. As can be seen from Figure 6, the mAP of the proposed method is higher than that of any other methods. More specific, the mAP of the proposed method is improved by 18.45%, 2.25%, 8.55% compared with Fast R-CNN, Faster R-CNN, and SSD respectively. Because the proposed framework utilizes feature fusion strategy to extract features and coarse-to-fine detection strategy to improve the classification ability of the detection network, the performance of the proposed method on traffic light detection is significantly improved, especially with small traffic lights. Table 2 shows the comparisons of AP for several methods. The results in Table 2 shows that the proposed method outperforms other methods on all group labels. More specific, compared with the highest accuracy of reference methods, the proposed method can improve the AP by 5.2%, 1.8%, 1.9%, 0.1% in “green”, “red”, “yellow”, and “off” group respectively. Figure 7 shows example of detection results of the proposed method on Bosch Small Traffic Lights dataset. As shown, the proposed network can recognize exactly small traffic lights in large traffic scene images.

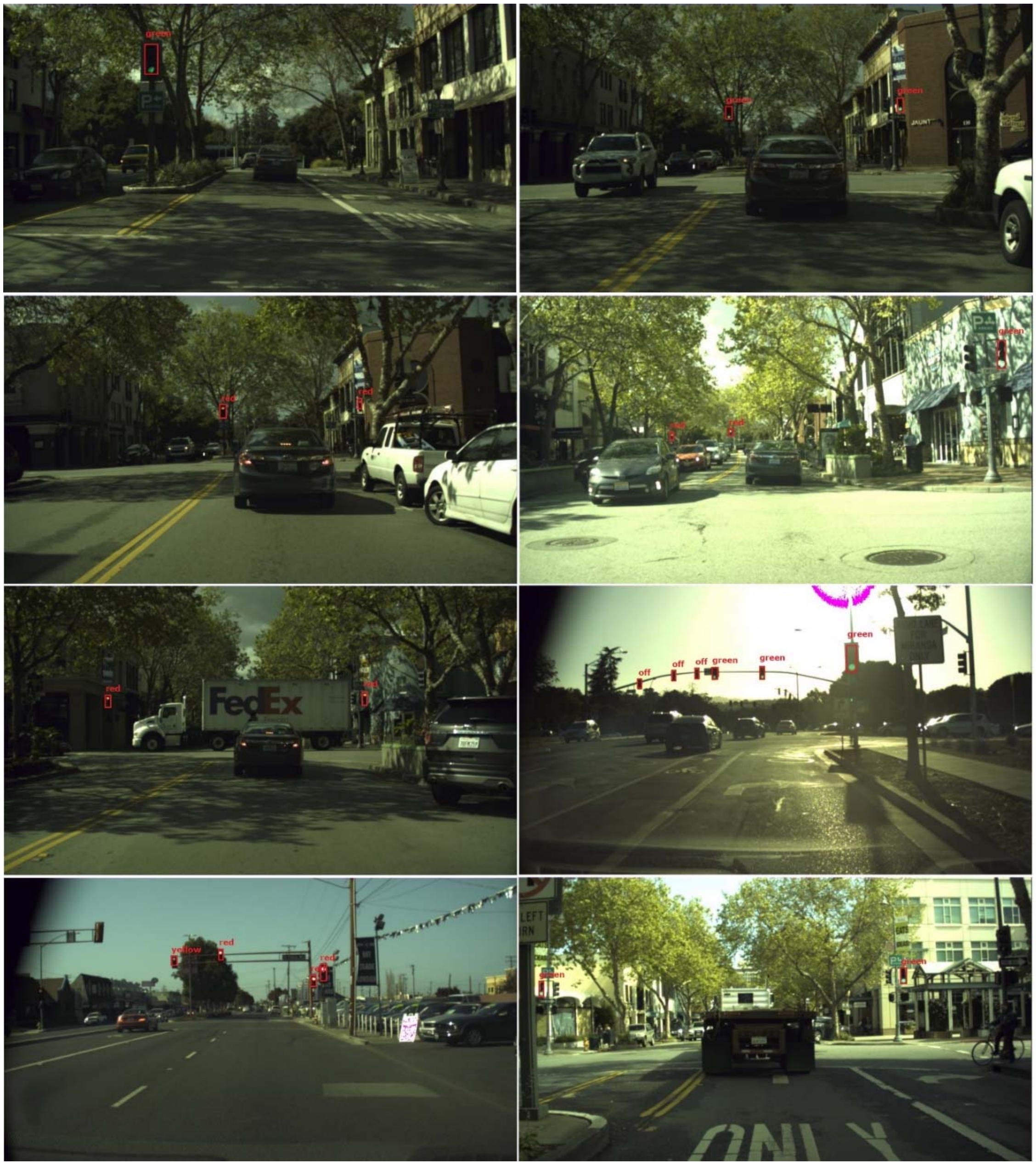


Figure 7: Example of Detection Results of The Proposed Method on Bosch Small Traffic Lights Dataset.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a deep learning framework for small traffic light recognition in traffic scene images. In the proposed framework, a feature fusion subnet, which fuses feature maps at different layers of the base network, is designed to enhance the accuracy on small traffic lights. In addition, a detection subnet is designed at the detection prediction stage. The detection subnet includes multiple detection layers, and each layer performs detection predictions with a coarse-to-fine detection strategy. The coarse-to-fine detection strategy is applied to improve the classification ability of the detection network. The proposed approach is evaluated on Bosch Small Traffic Lights dataset. Experimental results show that the proposed approach achieves better accuracy compared with recent state-of-the-art deep learning object detectors. This paper will investigate more enhancements to improve the accuracy and the detection speed of the proposed network in future work.

REFERENCES:

[1] Chiang, Cheng-Chin, Ming-Che Ho, Hong-Sheng Liao, Andi Pratama, and Wei-Cheng Syu. "Detecting and recognizing traffic lights by genetic approximate ellipse detection and spatial texture layouts." *International Journal of Innovative Computing, Information and Control* 7, no. 12 (2011): 6919-6934.

[2] Diaz-Cabrera, Moises, Pietro Cerri, and Javier Sanchez-Medina. "Suspended traffic lights detection and distance estimation using color features." In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pp. 1315-1320. IEEE, 2012.

[3] Gong, Jianwei, Yanhua Jiang, Guangming Xiong, Chaohua Guan, Gang Tao, and Huiyan Chen. "The recognition and tracking of traffic lights based on color segmentation and camshift for intelligent vehicles." In *2010 IEEE Intelligent Vehicles Symposium*, pp. 431-435. Ieee, 2010.

[4] Kim, Hyun-Koo, Young-Nam Shin, Sa-gong Kuk, Ju H. Park, and Ho-Youl Jung. "Night-time traffic light detection based on svm with geometric moment features." *International Journal of Computer and Information Engineering* 7, no. 4 (2013): 472-475.

[5] Lindner, Frank, Ulrich Kressel, and Stephan Kaelberer. "Robust recognition of traffic signals." In *IEEE Intelligent Vehicles Symposium, 2004*, pp. 49-53. IEEE, 2004.

[6] De Charette, Raoul, and Fawzi Nashashibi. "Real time visual traffic lights recognition based on spot light detection and adaptive traffic lights templates." In *2009 IEEE Intelligent Vehicles Symposium*, pp. 358-363. IEEE, 2009.

[7] Barnes, Dan, Will Maddern, and Ingmar Posner. "Exploiting 3D semantic scene priors for online traffic light interpretation." In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pp. 573-578. IEEE, 2015.

[8] Fairfield, Nathaniel, and Chris Urmson. "Traffic light mapping and detection." In *2011 IEEE International Conference on Robotics and Automation*, pp. 5421-5426. IEEE, 2011.

[9] Müller, Julian, Andreas Fregin, and Klaus Dietmayer. "Multi-camera system for traffic light detection: About camera setup and mapping of detections." In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 165-172. IEEE, 2017.

[10] Bach, Martin, Stephan Reuter, and Klaus Dietmayer. "Multi-camera traffic light recognition using a classifying labeled multi-bernoulli filter." In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1045-1051. IEEE, 2017.

[11] Behrendt, Karsten, Libor Novak, and Rami Botros. "A deep learning approach to traffic lights: Detection, tracking, and classification." In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1370-1377. IEEE, 2017.

[12] Weber, Michael, Peter Wolf, and J. Marius Zöllner. "DeepTLR: A single deep convolutional network for detection and classification of traffic lights." In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pp. 342-348. IEEE, 2016.

[13] Saini, Sanjay, S. Nikhil, Krishna Reddy Konda, Harish S. Bharadwaj, and N. Ganeshan. "An efficient vision-based traffic light detection and state recognition for autonomous vehicles." In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 606-611. IEEE, 2017.

[14] Lee, Gwang-Gook, and Byung Kwan Park. "Traffic light recognition using deep neural networks." In *2017 IEEE international conference on consumer electronics (ICCE)*, pp. 277-278. IEEE, 2017.

[15] Jensen, Morten B., Kamal Nasrollahi, and Thomas B. Moeslund. "Evaluating state-of-the-art object detector on challenging traffic light data." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 9-15. 2017.

[16] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.

[17] Jensen, Morten Bornø, Mark Philip Philipsen, Andreas Møgelmose, Thomas Baltzer Moeslund, and Mohan Manubhai Trivedi. "Vision for looking at traffic lights: Issues, survey, and perspectives." *IEEE Transactions on Intelligent Transportation Systems* 17, no. 7 (2016): 1800-1815.

[18] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017.

[19] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[20] KHALIFA, NOUR ELDEEN M., MOHAMED LOEY, and MOHAMED HAMED N. TAHA. "Insect pests recognition based on deep transfer learning models." *Journal of Theoretical and Applied Information Technology* 98, no. 01 (2020).

[21] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." *International journal of computer vision* 88, no. 2 (2010): 303-338.

[22] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115, no. 3 (2015): 211-252.

[23] Girshick, Ross. "Fast r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 1440-1448. 2015.

[24] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In *Advances in neural information processing systems*, pp. 91-99. 2015.

[25] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In *European conference on computer vision*, pp. 21-37. Springer, Cham, 2016.

[26] Kim, Hyun-Koo, Ju H. Park, and Ho-Youl Jung. "An efficient color space for deep-learning based traffic light recognition." *Journal of Advanced Transportation* 2018 (2018).