Journal of Theoretical and Applied Information Technology
31st December 2019. Vol.97. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MULTI-SCALE DEEP LEARNING NETWORK FOR VEHICLE DETECTION

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

Vehicle detection is an important subject in intelligent transportation systems. With the development of deep learning, many methods for vehicle detection based on deep learning have been proposed and showed better performance compared with traditional methods. However, due to difficult conditions such as occlusion or truncation of vehicle in image, small size of vehicle, and environment conditions, the performance of deep learning-based vehicle detection is still limited. This paper proposes a multi-scale deep learning framework for vehicle detection in traffic scene. To improve the performance of small vehicle detection, an enhanced multi-scale feature map generation module is designed to fuse different convolution layers at different scales of the base network and create the enhanced feature map which improves the resolution of small vehicle and simultaneously includes more semantic information. In the detection stage, the information surrounding a given object proposal is exploited to enhance the feature representation of proposals. At final stage, a classifier including a region of interest pooling layer and fully connected layers is used for classification and bounding box regression. For evaluating the proposed framework, the KITTI dataset is adopted. Experimental results on the KITTI dataset show that the proposed method achieves nearly as performance as other state-of-the-art methods on vehicle detection.

Keywords: *Vehicle Detection, Convolutional Neural Network, Intelligent Transportation System, Object Detection, Deep Learning*

1. INTRODUCTION

Vision-based vehicle detection from images is an essential prerequisite for many intelligent transportation systems, with a wide range of real-world applications, such as ADAS and autonomous driving, intelligent traffic management systems and so on. Many methods for vision-based vehicle detection have been proposed. Traditional methods are usually based on hand craft features such as colour, shape, energy, and so on to locate vehicle in image [8]-[15]. Because that vehicle detection systems are used in real-world applications, they should be robust to illumination variations, partial detections, occlusions, camera viewpoint changes, scale changes, and so on. In urban driving, frequent occlusions, along with a variety of vehicle orientations and scales, make vehicle detection difficult. Therefore, improvements in the detection accuracy of real-world vehicle-detection systems have become major issues.

In recent years, deep convolutional neural networks (CNNs) have achieved incredible success on vehicle detections as well as various other object detection tasks. However, when applying CNNs to vehicle detection, real-time vehicle detection in driving environment is still very challenging. It is observed that the object detection performance of the popular CNN detectors including Faster-RCNN [2] and SSD [5] without modification is not very good over the KITTI benchmark datasets [3]. KITTI is the largest public dataset dedicated to ADAS and autonomous driving benchmarking. One of the main challenges is that traditional CNNs are sensitive to scales while it is quite common that in-car videos or transportation surveillance videos contain vehicles with a large variance of scales. Current methods are based on modifying the popular CNN detectors to enhance the performance of detection results. These methods focus on making the network fit different scales by utilizing input images with multiple resolutions. However, these methods introduce expensive computational overhead and thus are still incapable of fast vehicle detection, which is essential for autonomous driving systems, real time surveillance and prediction systems.

In view of the above research challenges, this paper proposes a multi-scale deep learning

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

framework to increase the visual vehicle detection accuracy. To improve the performance of small vehicle detection, an enhanced multi-scale feature map generation module is designed to fuse different convolution layers at different scales of the base network and create the enhanced feature map which improves the resolution of small vehicle and simultaneously includes more semantic information. In the detection stage, the information surrounding a given object proposal is exploited to enhance the feature representation of proposals. At final stage, a classifier including a region of interest pooling layer and fully connected layers is used for classification and bounding box regression. The proposed approach is evaluated over KITTI benchmark dataset. With this dataset, the proposed method achieves comparable detection results with other state-of-the-art methods.

This paper is organized as follows: an overview of previous methods is presented in Section 2. Section 3 describes detail the proposed method. Section 4 demonstrates experimental results. Finally, the conclusion is made in Section 5.

2. RELATED WORK

In this section, this paper introduces previous vehicle detection methods, including traditional methods and deep learning-based methods.

Vision-based vehicle detection methods can be divided into two groups [23]: Traditional method and deep learning-based method. Traditional methods are usually based on hand craft feature such as colour, shape, energy, and so on to locate vehicle in image. Li et al. [8] proposed a method of learning reconfigurable hierarchical And-Or models which represents the regularities of car-to-car context and occlusion patterns at three levels to integrate context and occlusion for car detection. In [9], the structure of the And-Or model is learned with three components, and the model parameters are jointly trained using Weak-Label Structural SVM. Chen et al. [10] proposed a method based on background Gaussian Mixture Model and shadow removal method to deal with sudden illumination changes and camera vibration. Furthermore, A Kalman filter tracks a vehicle to enable classification by majority voting over several consecutive frames, and a level set method has been used to refine the foreground blob. In [11], the authors used Haar and Adaboost algorithm to detect the vehicle. In addition, simplified Lucas-Kanade algorithm and virtual edge were used to remove false positive detection and use automatic image matting to do detection refinement. In [12], two-dimensional discrete wavelet transform is used first for extracting features from the images which has a good location property in time and frequency domains. Moreover, road detection is proposed to determine the zone of interest, this technique is used one time at the beginning of the processing to solve the problem of unimportant movement of the background and also to reduce the processing time. To detect vehicles, the Background subtraction method is used, followed by the connected components method to improve the results of the detection. In [13], the features of vehicles are learned as a deformable object model through the combination of a latent support vector machine and histograms of oriented gradients. The detection algorithm combines both global and local features of the vehicle as a deformable object model. Hsieh et al. [14] proposed a new symmetrical SURF descriptor to enrich the power of SURF to detect all possible symmetrical matching pairs through a mirroring transformation. A vehicle-make and model recognition application are then adopted to prove the practicability and feasibility of the method. Yebes et al. [15] carried out a discussion on the supervised learning of a car detector built as a Discriminative Part-based Model from images in the KITTI benchmark suite as part of the object detection and orientation estimation challenge. In [22], the authors proposed a general active-learning framework for robust on-road vehicle recognition and tracking. This framework takes a novel active-learning approach to building vehicle-recognition and tracking systems. Makris et al. [24] proposed a method that uses local image features and follows the part-based detection approach. The method fuses intensity and depth information in a probabilistic framework. The depth of each local feature is used to weigh the probability of finding the object at a given distance. In [25], the authors used active learning to train independent-part detectors. A semisupervised approach is used for training part-matching classification, which forms sideview vehicles from independently detected parts. In [26], strategies for occlusion and orientation handling are explored by learning an ensemble of detection models from visual and geometrical clusters of object instances. An AdaBoost detection scheme is employed with pixel lookup features for fast detection.

Recently, deep learning-based methods have become the leading method for high quality general object detection. Faster region-based convolutional neural network (Faster R-CNN) [2] defined a region proposal network (RPN) for generating region proposals and a network using these proposals to detect objects. RPN shares full-image convolutional

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features with the detection network, thus enabling nearly cost-free region proposals. This method has achieved state-of-the-art detection performance and become a commonly employed paradigm for general object detection. SSD framework [5] predicted category scores and box offsets for a fixed set of default bounding boxes using small convolutional filters applied to different scales from feature maps of different scales, and explicitly separate predictions by aspect ratio. This framework showed much faster and comparably performance with other methods. R-FCN [16] proposed region-based detector which is a fully convolutional with almost all computation shared on the entire image. To achieve this goal, the position-sensitive score maps is designed to address a dilemma between translation-invariance in image classification and translation-variance in object detection. The MS-CNN [6] consists of a proposal sub-network and a detection sub-network. In the proposal sub-network, detection is performed at multiple output layers, so that receptive fields match objects of different scales. Another interesting work is YOLO [17], which outputs object detections within a 7x7 grid. This network runs at 40 fps, but with some compromise of detection accuracy. Most of these deep learning models target general object detection including vehicle. To better handle the detection problem of vehicles in complex conditions, Zhou et al. [18] proposed a fast vehicle proposal network (FVPN) for vehicle-like objects extraction and an attribute learning network (ALN) aiming to verify each proposal and infer each vehicle's pose, color and type simultaneously. These two nets are jointly optimized so that abundant latent knowledge learned from the ALN can be exploited to guide FVPN training. Dong et al. [19] proposed a vehicle type classification method using a semi-supervised convolutional neural network from vehicle frontal-view images. In addition, in order to capture rich and discriminative information of vehicles, sparse Laplacian filter learning is introduced to obtain the filters of the network with large amounts of unlabeled data. In [20], the authors proposed two context-aware structural descriptors, termed as a context-aware difference sign transform feature and context-aware difference magnitude transform feature. Hu et al. [21] proposed a scale-insensitive convolutional neural network for fast detecting vehicles with a large variance of scales. Context-aware RoI pooling is designed to maintain the contextual information and original structure of small-scale objects, and a multi-branch decision network is introduced to minimize the intra-class distance of features.

3. PROPOSED APPROACH

Figure 1 illustrates the overall framework of the proposed approach. As shown in Figure 1, VGG-16 architecture is first adopted to generate base convolution layers from input image. The enhanced multi-scale feature map generation module is then designed to create the enhanced feature map which improves the resolution of small vehicle and simultaneously includes more semantic information. The region proposal network is adopted to generate object proposals from enhanced feature map. RoI pooling layer is used to transfer each proposal and contextual regions of each object proposal to fixed size feature map. Due to the different scales of the feature representations generated by the RoI pooling layer, this paper concatenates these feature representations along the channel axis to form a concatenated feature representation. Finally, the final fused feature vector is fed into the classifier to classify proposals into vehicle and background class and adjust the bounding box for each of detected vehicle. Details of proposed framework are explained at the following sections.

3.1 Enhanced Multi-Scale Feature Map Generation

First, this study extracts convolutional features based on VGG-16 network [1]. VGG-16 network has 16 weight layers in its original form. VGG-16 is a simpler architecture model, since it is not using much hyper parameters. It always uses 3 x 3 filters with stride of 1 in convolution layer and uses same padding in pooling layers 2 x 2 with stride of 2. Figure 2 shows the architecture of the enhanced feature map generation module. The first number in the labels such as 1 and 2 represents the associated hidden layer in VGG-16 architecture, and the second number represents the ID of the convolution layer in a hidden layer.

In deep learning-based object detection, the higher resolution feature maps in lower convolution layers could better describe the characteristics of the small-scale objects, and the lower resolution feature maps in higher convolution layers could better describe the characteristics of the large-scale objects. To detect vehicle at different sizes, this paper conducts several experiments to compare the performance of different convolution layers. According to results from extensive experiments, it is found that the output of convolution layer Conv3-3 is the most suitable for localization of small vehicle because it possesses smaller receptive fields and higher resolution compared with Conv4 and Conv5. However, when used as a feature map, Conv3-3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

leads to poor detection performance because it contains less semantic information. Thus, this paper designs an enhanced multi-scale feature map generation module that improves the resolution of small vehicle and simultaneously includes more semantic information, which improves the performance of the region proposals.

As shown in Figure 2, this paper first uses the multi-scale deconvolution operation to up-sample the output feature map of the deeper convolution layers, including Conv4-3 and Conv5-3 layers. Notably, this deconvolution operation is different from the original up-sampling operation as it provides a set of parameters by which to learn a nonlinear up-sampling of the features in the deep layers. Then, the output features of different layers, including the output feature at Conv3-3, the output feature after the first deconvolution layer and the output feature after the second deconvolution layer are assembled to create the concatenated feature. The feature maps of different convolution layers have a different scale of value, and feature values from the shallower layer are generally larger than them from the deeper layer. Directly concatenating them easily leads to the smaller values being dominated by the larger values. Thus, this paper adds L2 normalization before the concatenation operation. L2 normalization can effectively keep the feature values from the different convolution layer on the same scale. For each pixel vector $i = (i_1, i_2, i_3, \dots, i_k)$ in the concatenated feature map, L2 normalization is defined as follow:

$$\hat{\imath} = \frac{i}{\|i\|_2} = \frac{i}{(\sum_{l=1}^{k} |i_l|^2)^{1/2}} \quad (1)$$

where $\hat{\imath}$ represents the normalized vector; $\|i\|_2$ represents the L2 normalization of i; k represents the number of channels.

Finally, a 1x1 pointwise convolution is used to compress the number of channels within the concatenated feature and create the final fused feature map.

3.2 Enhanced Region Proposal Network (RPN) and Detection Network

The RPN [2] takes the fused convolution feature map generated by the multi-scale feature map generation module as input to create a set of anchor boxes. An anchor is centered at the sliding window, and is associated with a scale and aspect ratio. Based on the size of vehicle in images, this paper uses three scales and three aspect ratios for each anchor, yielding 9 anchors at each sliding position. More specific, the scales are set at 64, 128, 256 and the aspect ratios are set at 0.5, 1, 2. The RPN then takes all the anchor boxes and outputs two different outputs for each of the anchors. The first one is objectness score, which means the probability that an anchor is an object. The second output is the bounding box regression for adjusting the anchors to better fit the object. The RPN is implemented efficiently in a fully convolutional way, using the fused convolutional feature map returned by multi-scale feature map generation module as an input. First, a convolutional layer with 512 channels and 3x3 kernel size is used and then two parallel convolutional layers using a 1x1 kernel size are used, whose number of channels depends on the number of anchors per point. Since anchors usually overlap, proposals end up also overlapping over the same object. Non-Maximum Suppression (NMS) is used to solve the issue of duplicate proposals. The proposal whose region overlaps a ground truth region more than 70% is regarded as a positive proposal. Otherwise, it is regarded as a negative proposal. After applying NMS, the top 256 proposals sorted by score are keep for next stage.

The fused feature map in the RPN is intended to improve the resolution and semantic information for small vehicle detection. The contextual information drawn from the neighborhood of the object proposal can provide important cues for object classification. Thus, this paper proposes an enhanced method to leverage contextual information for the object proposals. Let p_i is an object proposal at (x_i, y_i) in the fused convolution feature map. Let w_i, h_i represents the width and the height of this proposal. This paper defines two contextual regions for each object proposal, which are cropped from the fused feature map at two different scales. The two contextual regions for proposal p_i are defined as follows:

$$p'_{i1} = (x_i, y_{i,} 1.2w_i, 1.2h_i) \quad (2)$$

$$p'_{i2} = (x_i, y_{i,} 1.5w_i, 1.5h_i) \quad (3)$$

Then, the object proposal p_i and its contextual regions, including p'_{i1}, p'_{i2}, are fed into the RoI pooling layer. RoI pooling layer takes a list of regions with different sizes to create a list of corresponding feature maps with a fixed size by using max pooling operation. Fixed size feature maps are needed for the classifier in order to classify them into a fixed number of classes. In this paper, each feature representation has a fixed size of 7x7x256.

In addition, due to the different scales of the three feature representations generated by the RoI

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pooling layer, this paper adopts the L2 normalization after each representation. Then, the three feature representations after L2 normalization are concatenated along the channel axis to form a concatenated feature representation. Finally, a 1x1 convolution is employed to compress the number of channels within the concatenated feature from 7x7x3x256 to 7x7x256.

At the final stage, the final fused feature vector is fed into the classifier. The classifier has two different goals: Classify proposals into vehicle and background class and adjust the bounding box for each of detected vehicle. The proposed classifier has two fully connected layers, a box classification layer and a box regression layer. The first fully connected layer has two outputs, which are fed into the softmax layer to compute the confidence probabilities of being vehicle and background. The second fully connected layer with linear activation functions regresses the bounding boxes of vehicle. All convolutional layers are followed by a batch normalization layer and a ReLU layer.

4. EXPERIMENTAL RESULTS

In order to compare the effectiveness of the proposed method with other methods on vehicle detection, this paper conducts experiments on the KITTI dataset [3]. KITTI dataset is the most used dataset for evaluating vehicle detection method. The proposed method is implemented on a Window system machine with Intel Core i7 8700 CPU, NVIDIA GTX 1080 GPU and 8 GB of RAM. TensorFlow is adopted for implementing deep CNN frameworks, and OPENCV library is used for real-time processing.

4.1 Dataset and Evaluation Metrics

KITTI dataset [3] is the most widely used dataset for evaluating vehicle detection approaches. The KITTI dataset includes 7481 images for training with available ground-truth and 7518 images for testing with no available ground-truth. Images in this dataset include various scales of vehicles in different scenes and conditions and were divided into three difficulty-level groups: easy, moderate, and hard. If the bounding boxes size was larger than 40 pixels, a completely unshielded vehicle was considered to be an easy object, if the bounding boxes size was larger than 25 pixels but smaller than 40 pixels, a partially shielded vehicle was considered as a moderate object, and a vehicle with the bounding boxes size smaller than 25 pixels and an invisible vehicle that was difficult to see with the naked eye were considered as hard objects. Figure 3 shows example images in the KITTI dataset.

For evaluation metrics, this paper uses the average precision (AP) and intersection over union (IoU) metrics [4] to evaluate the performance of the proposed method in all three difficulty level groups of the KITTI dataset. These criteria have been used to assess various object detection algorithms. The IoU are defined as follow:

$$IoU(b_1, b_2) = \frac{area(b_1 \cap b_2)}{area(b_1 \cup b_2)} \quad (4)$$

where b_1 represents the bounding box of the proposal, b_2 represents the bounding box of the ground truth. The IoU is set to 0.7 in this paper, which means only the overlap between the detected bounding box and the ground truth bounding box greater than or equal to 70% is considered as a correct detection. Average precision evaluates the accuracy of detection from the perspective of recall rate and precision rate. It can be used to analyze the detection performance of a single category.

4.2 Detection Results

To evaluate the effectiveness of the proposed method, this paper compares the detection results of the proposed method with the results of recent state-of-the-art object detection methods on the KITTI dataset, including Faster R-CNN [2], Single Shot MultiBox Detector [5] (SSD), and MS-CNN [6]. Faster R-CNN framework contains two stages: region proposal network and object detection network. While Fast R-CNN algorithm [7] is based on the selective search algorithm, the Faster R-CNN introduces the region proposal network, which has improved over the traditional methods. The SSD framework combines region proposals and region classifications in a single stage. The core of SSD is predicting category scores and box offsets for a fixed set of default bounding boxes using small convolutional filters applied to different scales from feature maps of different scales, and explicitly separate predictions by aspect ratio. MS-CNN extends the detection over multiple scales of feature layers, which produces good detection performance improvement.

Figure 4 shows examples of detection results of the proposed method on the KITTI test dataset. As shown in Figure 4, the proposed approach can detect vehicle in difficult environments effectively. Furthermore, the proposed method can detect small vehicle and avoid producing multiple bounding boxes for one vehicle. Figure 5 presents example images in which some vehicles are not correctly detected. The main challenges of the vehicle

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

detection in KITTI dataset come from the heavy occlusion or truncation of the vehicles. Moreover, other external factors such as illumination change and cluttered background can affect the accuracy of the proposed method.

Table 1 shows the detection results of the proposed method and other state-of-the-art deep CNN-based object detectors on the KITTI test dataset. As shown in Table 1, the performance of the proposed method outperforms both Faster R-CNN and SSD framework in Easy and Moderate group. More specific, compared with Faster R-CNN, the performance of the proposed method is improved by 1.74% in Easy group and 5.21% in Moderate group. Compared with SSD, the performance of the proposed method is improved by 5.75% in Easy group and 17.15% in Moderate group. With Hard group, the proposed method achieves nearly as performance as Faster R-CNN. For the inference speed, the proposed framework takes 0.16 second to process an image while Faster R-CNN takes up to 1.7 second. SSD is the fastest framework with only 0.03 second, but SSD shows worse performance than other methods. Thus, the proposed approach meets the real-time detection standard and can be applied to the road driving environment of actual vehicles. Results in Table 1 show that MS-CNN achieves the best detection results. However, MS-CNN is slower than the proposed method. The proposed method achieves nearly as performance as MS-CNN while being faster and simpler.

5. CONCLUSIONS

This paper proposes a multi-scale deep learning framework that employs the two-stage fusion strategy for vehicle detection. The proposed framework integrates multiple levels of convolution feature and multiple levels of contextual information. At the detection stage, the region proposals are generated from the fused feature map with sufficient information. This paper designs the enhanced module that fuses different convolution layers by using deconvolution and normalization operations. At the classification stage, a fused feature map is created for the fully connected layer and the multi-scale contextual regions is designed to exploit the surrounding information for a given object proposal. Experimental results on the KITTI dataset show the proposed method's superiority for detecting small vehicle, and it achieved nearly as performance as other state-of-the-art methods. In future work, this paper focus on other deep learning architecture for improving detection accuracy.

REFERENCES:

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *NIPS*, 2015.

[2] Ren, S., He, K., Girshick, R. et al., "Faster r-cnn: Towards real-time object detection with region proposal networks", *Advances in neural information processing systems*, 2015, pp. 91–99.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite", *Proc. CVPR*, Jun. 2012, pp. 3354–3361.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge", *Int. J. Comput. Vis.*, vol. 88, no. 2, Sep. 2009, pp. 303–338.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector", *European conference on computer vision*, 2016, pp. 21–37, Springer.

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

[7] R. Girshick, "Fast R-CNN", *IEEE International Conference on Computer Vision (ICCV),* 2015.

[8] B. Li, T. Wu, and S.-C. Zhu, "Integrating context and occlusion for car detection by hierarchical and-or model," in *Proc. ECCV*, 2014, pp. 652–667.

[9] T.Wu, B. Li, and S.-C. Zhu, "Learning and-or model to represent context and occlusion for car detection and viewpoint estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1829–1843, Sep. 2016.

[10] Z. Chen, T. Ellis, and S. A. Velastin, "Vehicle detection, tracking and classification in urban traffic," in *Proc. ITSC*, Sep. 2012, pp. 951–956.

[11] J. Cui, F. Liu, Z. Li, and Z. Jia, "Vehicle localisation using a single camera," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2010, pp. 871–876.

[12] Slimani Ibtissam, Zaarane Abdelmoghit, Hamdoun A., Issam Atouf, "Traffic surveillance system for vehicle detection using discrete wavelet transform", *Journal of Theoretical and Applied Information Technology*, Vol. 96, 2018, pp. 5905-5917.

[13] H. T. Niknejad, A. Takeuchi, S. Mita, and D. McAllester, "On-road multivehicle tracking

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

using deformable object model and particle filter with improved likelihood estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 748–758, Jun. 2012.

[14] J.-W. Hsieh, L.-C. Chen, and D.-Y. Chen, "Symmetrical SURF and its applications to vehicle detection and vehicle make and model recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 6–20, Feb. 2014.

[15] J. J. Yebes, L. M. Bergasa, R. Arroyo, and A. Lázaro, "Supervised learning and evaluation of KITTI's cars detector with DPM," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 768–773.

[16] Y. Li, J. Dai, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. NIPS*, 2016, pp. 379–387.

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

[18] Y. Zhou, L. Liu, L. Shao, and M. Mellor, "DAVE: A unified framework for fast vehicle detection and annotation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 278–293.

[19] Z. Dong, Y. Wu, M. Pei and Y. Jia, "Vehicle Type Classification Using a Semisupervised Convolutional Neural Network," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2247-2256, Aug. 2015.

[20] X. Yuan, X. Cao, X. Hao, H. Chen and X. Wei, "Vehicle Detection by a Context-Aware Multichannel Feature Pyramid," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1348 1357, July 2017.

[21] X. Hu et al., "SINet: A Scale-Insensitive Convolutional Neural Network for Fast Vehicle Detection," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1010-1019, March 2019.

[22] S. Sivaraman and M. M. Trivedi, "A General Active-Learning Framework for On-Road Vehicle Recognition and Tracking," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 267-276, June 2010.

[23] A. Mukhtar, L. Xia and T. B. Tang, "Vehicle Detection Techniques for Collision Avoidance Systems: A Review," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2318-2338, Oct. 2015.

[24] A. Makris, M. Perrollaz and C. Laugier, "Probabilistic Integration of Intensity and Depth Information for Part-Based Vehicle Detection," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1896-1906, Dec. 2013.

[25] S. Sivaraman and M. M. Trivedi, "Vehicle Detection by Independent Parts for Urban Driver Assistance," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1597-1608, Dec. 2013.

[26] E. Ohn-Bar and M. M. Trivedi, "Learning to Detect Vehicles by Clustering Appearance Patterns," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2511-2521, Oct. 2015.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

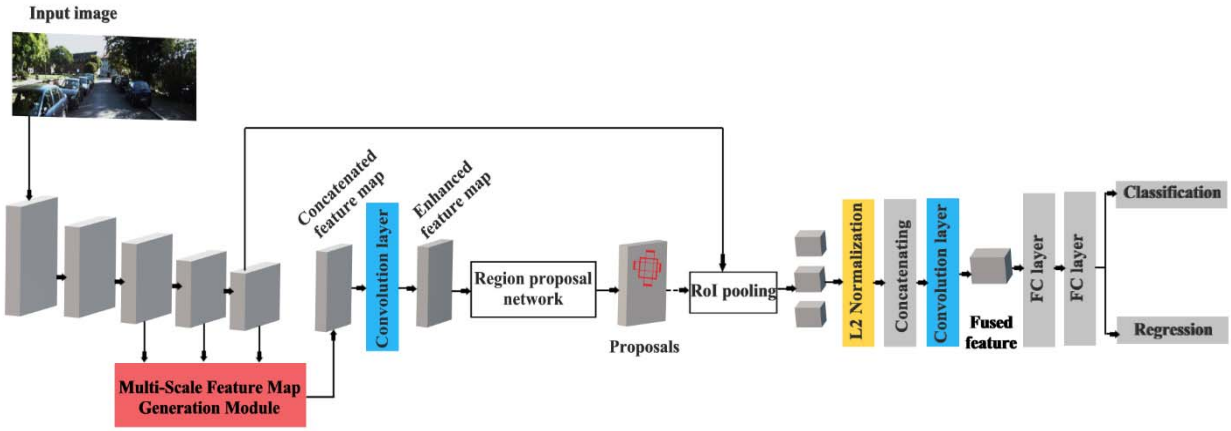

Figure 1: The Overall Framework of The Proposed Approach.

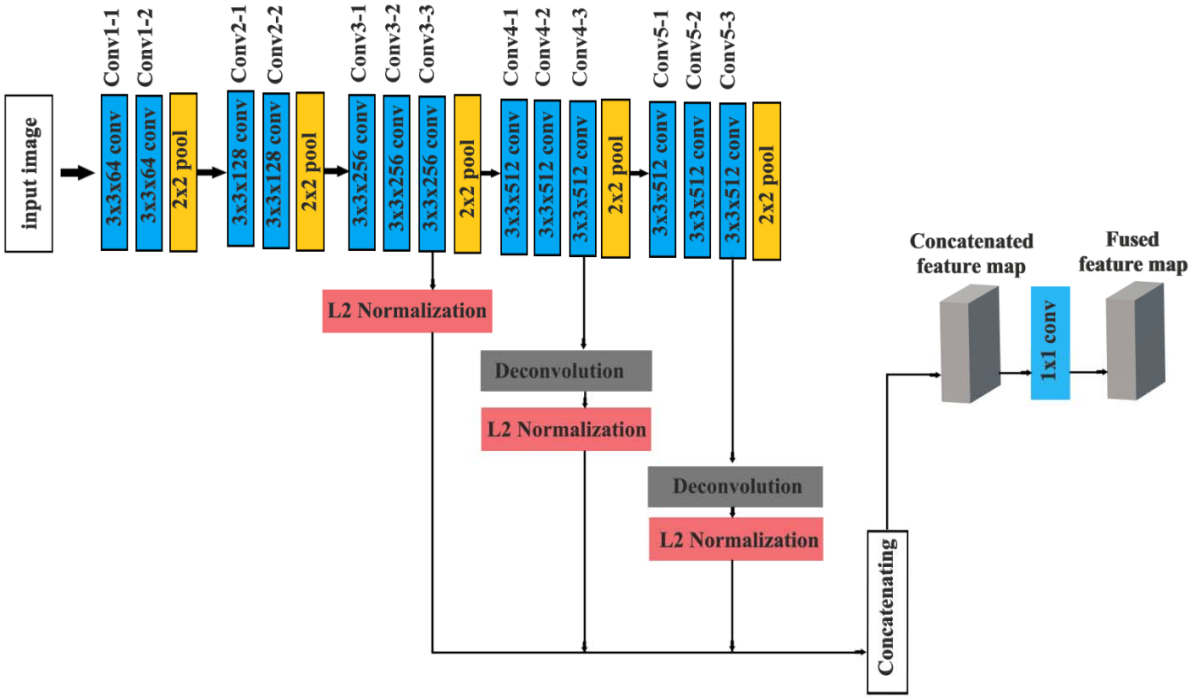

Figure 2: The Architecture of The Enhanced Feature Map Generation Module.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 3: Example Images in The KITTI Dataset.

Journal of Theoretical and Applied Information Technology
31st December 2019. Vol.97. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 4: Detection Results of The Proposed Method on The KITTI Test Dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 5: Undetected Vehicle Due to Small Size, Heavy Occlusion or Truncation of The Vehicles.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Detection Results of The Proposed Method and Other Methods.

Method	Difficulty-level groups			Inference time (second)
	Easy (%)	Moderate (%)	Hard (%)	
SSD [5]	83.89	67.17	59.09	0.03
Faster R-CNN [2]	87.90	79.11	79.19	1.7
MS-CNN [6]	90.46	88.83	74.76	0.3
Proposed method	89.64	84.32	79.03	0.16