A Highly Accurate Feature Fusion Network For Vehicle Detection In Surveillance Scenarios

Jianqiang Wang
wangjian16@mails.tsinghua.edu.cn
Ya-Li Li
liyali13@tsinghua.edu.cn
Shengjin Wang
wgsgj@tsinghua.edu.cn

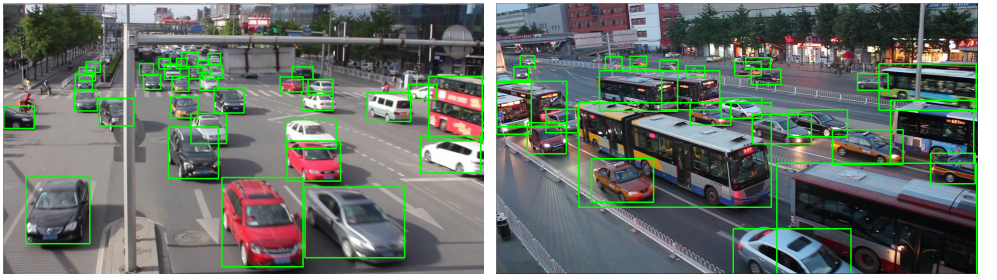
Department of Electronic Engineering
Tsinghua University
Beijing,China

Abstract

In this paper we present a novel vehicle detection method in traffic surveillance scenarios. This work is distinguished by three key contributions. First, a feature fusion backbone network is proposed to extract vehicle features which has the capability of modeling geometric transformations. Second, a vehicle proposal sub-network is applied to generate candidate vehicle proposals based on multi-level semantic feature maps. Finally, a head network is used to refine the categories and locations of these proposals. Benefits from the above cues, vehicles with large variation in occlusion and lighting conditions can be detected with high accuracy. Furthermore, the method also demonstrates robustness in the case of motion blur caused by rapid movement of vehicles. We test our network on DETRAC[21] benchmark detection challenge and it shows the state-of-the-art performance. Specifically, the proposed method gets the best performances not only at 4 different level: *overall*, *easy*, *medium* and *hard*, but also in *sunny*, *cloudy* and *night* conditions.

1 Introduction

With constant improvements of urban intelligent level, more and more traffic surveillance devices has been used. Locating vehicles from videos or images in traffic surveillance scenarios is not only an important research field in computer vision, but also meaningful applications in the real world. Based on the detected vehicles, further processes can be carried out, such as vehicle tracking, vehicle counting, extracting the license number, recognizing vehicle type, and so on. The basis of these further work is to detect vehicles as accurate as possible. However, there are still a fairly amount of challenges in vehicle detection from traffic surveillance scenarios. Occlusion is one of the most common interference factors and it reduces the detection performance when the traffic is crowded. Accuracy of a vehicle detector is also influenced by weather and lighting conditions. Moreover, fast moving vehicles lead to amounts of motion blur images. Furthermore, vehicle varies more widely in appearance, which comes from not only the diversity of vehicle types such as bus, truck, car



(a) (b)

Figure 1: Using our method, vehicles in the images with various sizes, types and severe occlusion can be accurately detected with a very high recall rate and no false alarm. (a) and (b) are two examples of detected results in DETRAC test set.

and so on, but also the various view angles from different orientations. This makes vehicle detection even more challenging.

In the past few years, some vehicle detectors have been designed based on different hand-crafted features such as Aggregate Channel Features (ACF)[1] or Deformable Part Model (DPM)[6]. Relying on the part-based features or rigid characteristic of vehicle, these methods can detect a few near complete vehicles, but they are incapable of detecting the ones which are heavily distracted by complex orientations, types, occlusion, and illumination.

Recently, deep convolutional neural network (CNN)[12] has been widely used for object classification and detection. Among kinds of variants of the feed-forward CNN based approaches, we can roughly divide them into two categories. One is the two stage detectors such as RCNN[9], Fast R-CNN[8], Faster R-CNN[13], R-FCN [4] and so on. They have a similar structure: the first stage proposes many plausible regions and the second stage further refines these regions. The other one is the one stage detectors [[7], [14],[17], [16]] which get rid of the proposal generation process and directly train a single stage end-to-end detector.

Vehicle detectors based on these networks have been proved to be more effective than traditional ones. But there is still many problems for improvement in these methods. Although the two stage detectors can generate higher quality detections than the single stage ones via a more computationally expensive head net, any missed vehicles in the first stage cannot be recovered in the following network. So it is very crucial for the two stage detectors to ensure a good recall rate during the proposal generation stage. Furthermore, looking for the response characteristics just from a single feature map layer, it is not enough to handle with vehicle detection in complicated situation. They are two main problems that we try to solve and to make vehicle detection more accurately.

In this paper, a novel vehicle detection network based on R-CNN architecture is proposed. The proposed method has a high capability of handling with vehicles in the traffic surveillance images with various sizes, types and severe occlusion. The result examples are shown in Fig.1. Our main contributions are summarized as follows: (1) A new feature extractor has been constructed. Inspired by an ingenious improvement which is called the deformable convolution [3], we add this mechanism to a feature fusion backbone network to improve the ability of modeling geometric transformations. This is easily restricted by the limited training samples. The new extractor has been proven competitive in feature extraction. (2) A vehicle proposal sub-network is applied to generate proposals. It can obtain higher recall rate than the previous method such as sliding window or the region proposal network (RPN) in faster RCNN. (3) An effective filtering mechanism in the process of train-

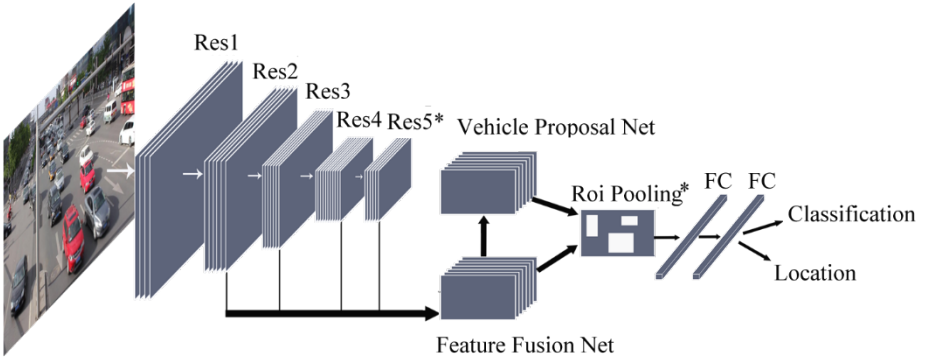


Figure 2: Overview of our vehicle detection network. Res{1, 2, 3, 4} are the building modules of ResNet-101 network. *Res5** denotes Res5 building module assigned with deformable convolution layers. *Roi Pooling** is a ROI pooling layer which can distribute the ROIs to the appropriate feature map to carry out the pooling operation. FC is the fully-connected layer with 2048 channels.

ing. In order to avoid that the vehicle detector harvests hard negatives from the areas where vehicles have not been annotated, we design a filtering mechanism to supervise vehicle proposals encoding processes. This makes training procedure focusing on vehicles with obvious features. Therefore, our network can be trained more effective.

We have tested our network on the DETRAC benchmark detection challenge. Furthermore, we have compared our method with all the submitted approaches and analyzed the comparison results. Our method ranks first in the detection leaderboard and gets the highest accuracy scores at 7 different evaluation metrics: overall, easy, medium, hard, sunny, night and cloudy. The detection results confirm that our network achieves the state-of-the-art performance.

The following of this paper is organized as follows. In Section 2, we explain the method in details. In Section 3, we introduce the experiments and present the results. In Section 4, we conclude this paper.

2 Framework

2.1 Overview

Our vehicle detection network is illustrated in Fig.2. ResNet-101 [10] is selected to extract vehicle features using five building modules [10]. We assign the last module with deformable convolution layers[3] at reasonable position. When a test image is fed into this network, feature maps are produced at multiple layers. Next, a feature fusion network fuses these feature maps and feeds them into vehicle proposal network which is applied to generate vehicle proposals. Based on the generated proposals, feature response of interest are found on the suitable feature map from the feature fusion network. Then, these regions of interest are pooled into the same size vectors and sent to the following network. Finally, the bounding boxes for different categories are produced in the head network. The entire framework is an end-to-end structure for vehicle detection. We will describe each individual network in details in the following stage.

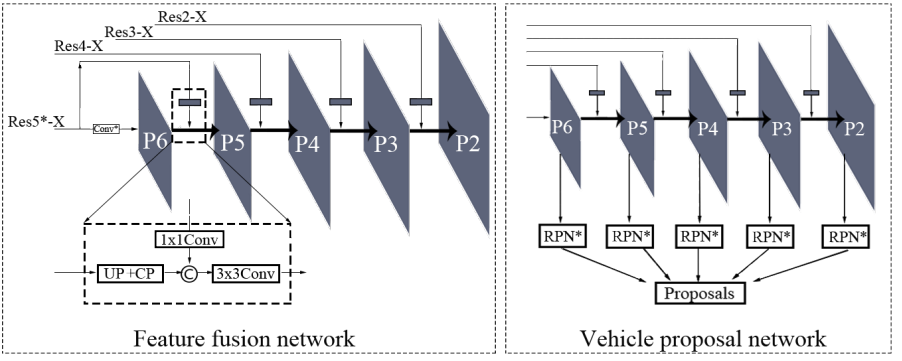


Figure 3: The left part is feature fusion network, *UP+CP* denotes the upsampling and the cropping operation. Res2-x is the last layer of the Res2 module. This is same for other modules. *Conv** is a 3×3 convolution layer which stride is 2 and © is the feature fusion operation. The right part is vehicle proposal network. *RPN** is the region proposal network specialized for vehicle.{P6, P5, P4, P3, P2} are the fused feature maps.

2.2 Backbone Convolution Network

For the requirements of rich feature representation, we choose the ResNet-101 as the basal convolution network, which has shown a strong capability of extracting feature in visual recognition tasks such as image classification[6], object detection[9] and semantic segmentation[15]. Considering that training this network from scratch costs a lot of time and need a large dataset, we take a strategy of loading the weights pre-trained on the ImageNet dataset[5]. This is a common practice in object detection networks [[13], [7]].

As discussed in [3], the capability of modeling geometric transformations of convolution neural networks is influenced by the inherently limitations. One of them is that all activation units in the same CNN layers have a same receptive field size. This is undesirable for high-level layers to encode the semantic over spatial locations, especially when object appearance and size varies widely in images. On the contrary, using convolution operation with position offset, the deformable convolution activation units have a more flexible receptive and alleviate the inherently limitation of CNNs. This is the reason why we embed the deformable convolution layers in the last building module of our basal convolution network. The specific operation is as follows. From the preceding feature maps, the 2D location offsets of each units are learned. Through adding a location offset to its corresponding unit coordinates, we can obtain a new unit coordinates. When a 3×3 filter kernel operates convolution, the unit in sampling region will be replaced by a new unit which is obtained by the method described above. Besides, channels of these layers are divided into 4 groups and the offsets are shared in different channels of each same group. Owing to that, the deformable convolution layer dose not introduce too many additional parameters to the network and enhances the capability of modeling geometric transformations.

At the same time, locating serious occlusion vehicles accurately in heavy traffic needs adequate vehicle boundary information extracted by the convolution layers. Although high-level convolution layers have adequate semantics to classify vehicle, they cannot afford to reserve the vehicle boundary characteristics since the resolution of feature maps gradually become weak by the pooling operation. On the contrary, the activation units in low-level feature maps have a smaller and better scope to localize object boundary. So we need both

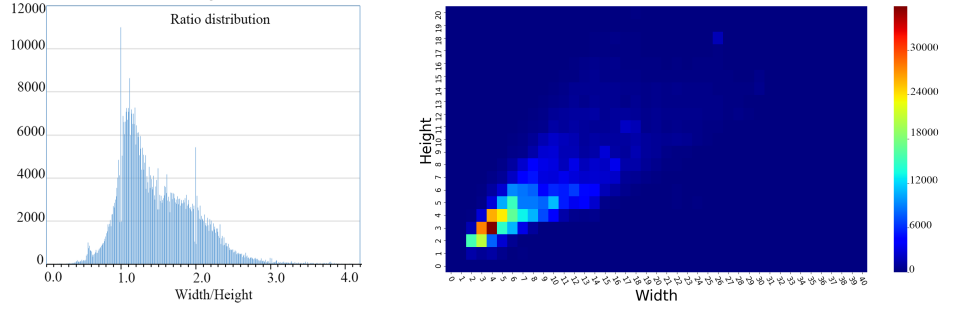


Figure 4: The left figure shows the aspect ratio distribution of the vehicles in DETRAC dataset. The right part is the size distribution heat map. Different colors represent the number of vehicles. The width and height are the 10 times smaller than the truth value.

the high-level semantics and high-resolution features maps for vehicle detection. Prior works [[13], [11]] have also shown the gain of merging and concatenating feature maps from different level layers. Inspired by [13], we design a feature fusion network to leverage the feature hierarchy and build new feature maps with both the high-level semantics and high-resolution feature information, which is shown in the left part of Fig.3. At first, we pick out feature maps of the last layers from Res{2, 3, 4, 5} modules and create new feature maps through implementing a 3×3 convolution operation of step size 2 on the last layers of Res5. Second, we use the 1×1 convolution filters to change the feature maps channels of these five groups into the same size of 256. Third, from high to low levels layers, two feature maps in same stage are fused by the method of adding elements in same position directly and then mixed via a 3×3 convolution filter. Besides, we operate the upsampling and cropping in feature maps {P6, P5, P4, P3} to confirm that different resolution feature maps are normalized into the same size. Through this process, the feature fusion network can produce multi-level feature maps with different resolutions and rich semantics.

2.3 Vehicle Proposal Network

Since vehicle is the rigid-body object, its shape has a certain distribution in traffic surveillance scenarios. On the other hand, as we analyzed in section 1, a high recall rate during the proposal generation stage is important for the two-stage detector. Considering that, a region proposal network specialized for vehicle is constructed. We analyze the training set of DETRAC dataset, which includes 84k images and 598,281 annotated vehicles. It is a large dataset that has strong representativeness of traffic surveillance. So we counted the distribution of vehicles in this dataset. As shown in Fig.4, the left part is vehicle aspect ratio distribution. We can easily find that there are two different peaks value of the histogram and most of the data is centered around them. The corresponding x-coordinate of the two peaks are approximate 1.0 and 2.0. It suggests that the majority of aspect ratios of the vehicles in this scenario are around 1.0 and 2.0. The other part in Fig.4 is the size distribution heat map which can roughly describe the vehicle width and height. All the scales in the graph is 10 times smaller than the truth value for the sake of demonstration. This heat map clearly shows that the vehicles distribute in the bottom left which represents small vehicles in images, but the whole distribution is a wide area. Taking the width range as an example, there are a lot of vehicles from 10 to 300 in width.

In view of the above analysis, we put forward a special vehicle proposal network to gen-

erate vehicle region proposals as shown in the right part of Fig.3. RPN* is a region proposal network which is similar to [18]. As the way it works, with 3×3 convolutional filters slides on the feature map, a certain number of vectors are produced, then 1×1 convolutional filters change the channels of these vectors to produce the proposals. In this process, many anchors are encoded in advance. We set the anchor ratios to 1.0 and 2.0 to match the aspect ratio distribution of the vehicles. Furthermore, we use five parallel RPN* to generate proposals at multi-level semantic feature maps and choose 5 base anchors with height of 16, 32, 64, 128, 256 pixels to cover the vehicle size distribution range. Each RPN* corresponds to a certain size anchor according to the resolution of feature maps. For example, the RPN* connected to the P3 layer has two anchors with 32 pixels height and 1.0, 2.0 aspect ratios. This is because that the P3 is second size in the fusion layers and it has appropriate resolution to capture second small proposals around 32 pixels. Due to this structure, our vehicle proposal network can generate proposals covering vehicle size distribution with only 10 anchors. To extent, it compensates for the time loss caused by complex structure. Through using multi-level fused feature maps and special anchors for vehicle in surveillance scenarios, high quality and sufficient proposals are generated and fed into the head network.

Furthermore, according to the data set, there are many ignored regions which are hard to be manually annotated or too far away from surveillance equipment. These ignored regions have been declared in the annotations. We assign these regions another label named "ignored" and send them into the network together with vehicle labels. This category is not the target to be trained but the supervision to eliminate the proposals located in or nearby the ignored regions. We set a hyper-parameter which is *Ignore Fraction*. When the anchors and proposals are generated in vehicle proposal network, they will be calculated the overlap with the "ignored" labeled region. And the ones with IOU larger than *Ignore Fraction* do not contribute the data to the training process. This handling mechanism is implemented to avoid that our vehicle detector harvests hard negatives from those areas and makes the network training procedure more effective and focused.

2.4 Head Network

The head network is constructed with a Roi Pooling* layer and two fully-connected layers. Different from the original ROI pooling operation applied by general detection of neural networks just like [18], we add a distribution mechanism which assign the ROIs to the appropriate feature map to carry out the pooling operation. We implement this function with the following formula:

$$p = \log_2 \left(\sqrt{roi\_w \times roi\_h}/x \right) \tag{1}$$

$$Feature\_id = \begin{cases} P2 & p \leq 2 \\ P3 & 2 < p \leq 3 \\ P4 & 3 < p \leq 4 \\ P5 & 4 < p \leq 5 \\ P6 & 5 < p \end{cases} \tag{2}$$

Here *roi_w* and *roi_h* denote width and height of the ROI, respectively. The denominator x comes from the *pooled w* or *pooled h* parameter which is the fixed width or height of this layer output. For example, x is 6 in our network because we set the *pooled w* to 6. The result of the first formula, p, is the number of pooling operation required. It is used to

find the corresponding feature map with an appropriate receptive field through the relational expression in Equation (2).This algorithm can make each position value of the pooled feature region retain the original information as far as possible instead of doing interpolation or rounding operation. The entire distribution mechanism is effective to ensure that the features of the regions of interest are sufficiently large before pooling operation. As a result, there are adequate information to feed into the following layers. Then, two fully-connected layers with 2048 channels are applied to refine the classes and the bounding boxes of these ROIs.

3 Experiments

3.1 Network Training

As stated above, our backbone convolution network is initialized with the pre-trained ImageNet model and the offsets in the deformable convolution layer are zero initialized. Other sub-networks are randomly initialized from a zero-mean Gaussian distribution with standard deviation of 0.01. Our network is trained end-to-end with stochastic gradient descent (SGD) using the entire training and validation set. The initial learning rate is 0.01, which is then divided by 10 at 50k iterations and again at 70k iterations. Our model is totally trained with 90k iterations. We use horizontal image flipping as the only form of data augmentation unless other noted. The weight decay is set to 0.0001 and the momentum is set to 0.9. The mini-batch size of the input images is one. For vehicle proposal network, we use the batch size of 512 proposals. Anchors that overlap any ground truth for more than 0.7 in intersection over union (IoU) are assigned positive. And those that overlap the ground truth for less than 0.3 in IoU are assigned as negative examples. We keep 2000 proposals using non-maximum suppression (NMS) with threshold 0.7 to eliminate redundant boxes. For the head network, the candidates that overlap the ground truth for $IoU \geq 0.5$ are assigned positive ones and the others are assigned negative. OHEM [19]is used to control loss reverse propagation. The standard cross entropy loss is used for classification and the standard smooth L_1 loss [8] is used for the bounding box regression.

Our model are trained on DETRAC detection dataset. The images have the same size of 960×540 and they are continuous frames from 60 different videos which are captured in different scenarios including *sunny*, *cloudy*, *rainy* and *night* scenarios.We split the training images into the training set with 56k images and the validation set with 28k images.

3.2 Control Experiments

A number of experiments have been performed to evaluate the effectiveness of the proposed network: (1) We examine how each component of our backbone network affects the detection accuracy. The results are listed in Table 1. The mean average precision (mAP) is produced by the controlled experiments on the validation set. DCL denotes the deformable convolution layer and VPN denotes the vehicle proposal net. Fusion+RPN is implemented by just using one resolution feature maps from feature fusion net to generate proposals. We choose the Faster RCNN as the baseline and add the improved components. Our method achieves considerable improvement, which validates the previous analysis. (2) We remove the second stage part of our network and train the proposal net on training set. The recall rate of region proposal stage using different number of fused feature maps on validation set are presented in Table 2. The IoU threshold of test procedure is set to 0.5. We take the RPN

Controlled Experiments	Overall(mAP)
Faster RCNN[13]	73.32
Res101+DCL+RPN	76.07
Res101+DCL+Fusion+RPN	85.64
Res101+DCL+Fusion+VPN	**88.93**

Table 1: Effects of components on the performance of our method.

Feature Map	Recall Rate (IOU=0.5)
Res4[10]	94.10
P4	96.32
P5+P4+P3	97.68
P5+P4+P3+P2	98.96
P6+P5+P4+P3+P2	**99.02**

Table 2: Recall rates of the proposal net in the first stage with different feature maps.

in Faster RCNN as the baseline and use Res4 as the input of the RPN. The results prove that generating vehicle proposals from multiple fused feature maps can achieve a higher recall rate.

3.3 Comparison with state-of-the-art

We compare our method with state-of-the-art vehicle detection approaches on the test set of DETRAC detection dataset. The results are shown in Table 3. Our method HAVD (high accurate vehicle detector) achieve the highest overall accuracy of 80.51% atop the leaderboard. Compared to HAT, which is the second in the leaderboard, our proposed network improves the overall accuracy by 2.4%. Besides, we achieve a significant overall improvement of 2.6% mAP over the GP-FRCNNm[7] and 10.6% mAP over R-FCN[4]. Notably, our method obtains the highest accuracy scores in 7 scenarios, which shows the effectiveness of vehicle detection and robustness to various scenarios. Fig.5 further compares the performance of different vehicle detection methods. It can be seen from the figure that our method with red precision-recall curve achieves better detection coverage as well as accuracy.

Method	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny
RCNN[8]	48.95	59.31	54.06	39.47	59.73	39.32	39.06	67.52
YOLO[17]	57.72	83.28	62.25	42.44	57.97	64.53	47.84	69.75
Faster RCNN2[13]	58.45	82.75	63.05	44.25	66.29	69.85	45.16	62.34
EB[20]	67.96	89.65	73.12	53.64	72.42	73.93	53.40	83.73
R-FCN1[4]	69.87	93.32	75.67	54.31	74.38	75.09	56.21	84.08
RTN	74.15	91.52	79.16	61.73	77.02	77.20	65.27	84.14
GP-FRCNNm[7]	77.96	92.74	82.39	67.22	83.23	77.75	**70.17**	86.56
HAT	78.64	93.44	83.09	68.04	86.27	78.00	67.97	88.78
Ours-HAVD	**80.51**	**94.48**	**86.13**	**69.02**	**87.28**	**82.30**	69.37	**89.71**

Table 3: Mean average precision (mAP) of the submitted approaches on the DETRAC test set.

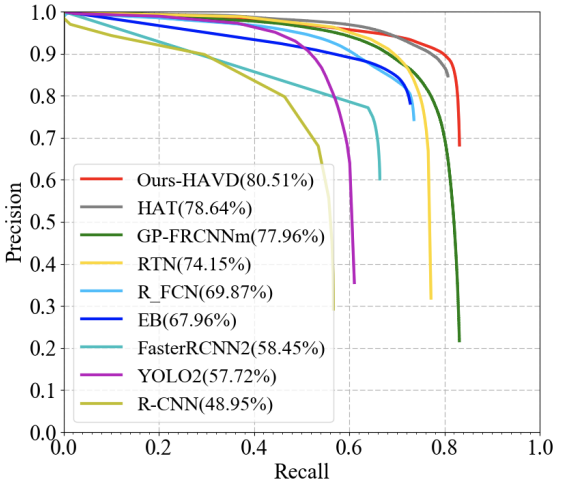
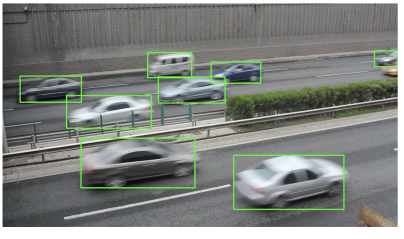
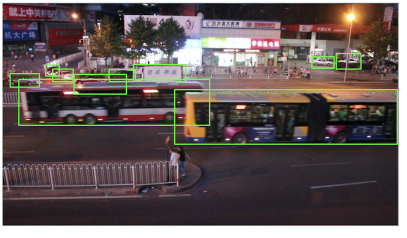


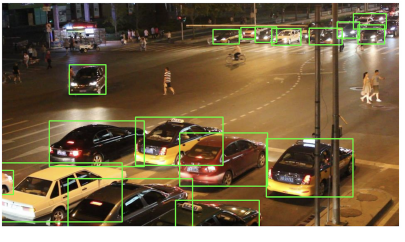
Figure 5: Comparison of precision-recall curves for different vehicle detection methods on the DETRAC test set.



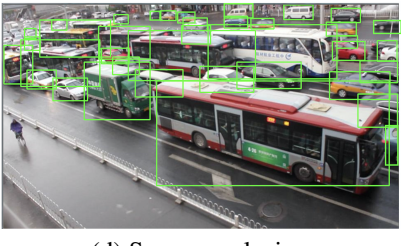
(a) Motion blur



(b) Large scale variance



(c) Night illumination condition



(d) Severe occlusion

Figure 6: Examples of detection results. Our method is capable of handling vehicles with severe occlusion, large scale variance and motion blur. It also performs well in night illumination condition.

4 Conclusion

In this paper, we propose a highly accurate vehicle detection network which can be applied in traffic surveillance scenarios. More specifically, a feature fusion backbone network for vehicle feature representation is firstly designed. Then a vehicle proposal net with high recall is applied to generate proposals. Lastly, a head network which is more reasonable to carry out pooling operation is proposed refine the proposals. Benefits from the designed network architecture and an effective filtering mechanism in the training procedure, the proposed

network is effective in feature representation and various vehicle detection. Experiments demonstrate that our method achieves the state-of-the-art performance on the challenging DETRAC detection dataset.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61771288, 61701277 and the state key development program in 13th Five-Year under Grant No. 2016YFB0801301.

References

[1] A. Alpher and J. P. N. Fotheringham-smythe. frobnication revisited. *Journal of Foo*, 2003.

[2] Sikandar Amin and Fabio Galasso. Geometric proposals for faster r-cnn. *AVSS*, 2017.

[3] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *ICCV*, 2017.

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *NIPS*, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained partbased models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[7] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv:1701.06659*, 2017.

[8] Ross Girshick. Fast r-cnn. *ICCV*, 2015.

[9] Ross Girshick, Jeff Donahue, revor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[11] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. *CVPR*, 2016.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[13] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *CVPR*, 2017.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *ECCV*, 2016.

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[16] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *CVPR*, 2016.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CVPR*, 2016.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 2015.

[19] A. Shrivastava, A. Gupta, and R. Girshick. Training regionbased object detectors with online hard example mining. *CVPR*, 2016.

[20] Li Wang, Yao Lu, Hong Wang, Yingbin Zheng, Hao Ye, and Xiangyang Xue. Evolving boxes for fast vehicle detection. *IEEE International Conference on Multimedia and Expo (ICME)*, 2017.

[21] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu. DETRAC: A new benchmark and protocol for multi-object detection and tracking. *arXiv:1511.04136*, 2015.