IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Part-Aware Region Proposal for Vehicle Detection in High Occlusion Environment

WEIWEI ZHANG [1], YAOCHENG ZHENG [1], QIAOMING GAO [2], ZEYANG MI [1]

[1] College of Mechanical Automotive Engineering, Shanghai University of Engineering Science, 333Long Teng Road, Shanghai, China, 201620
[2] School of Mechanical and Transportation Engineering, Guangxi University of Science and Technology, Guangxi 545006, China

Corresponding author: Yaocheng Zheng (e-mail: 15851723002@163.com).

This work was supported in part by National Natural Science Foundation of China (No. 51805312); in part by Shanghai Sailing Program (No.18YF1409400); in part by Training and Funding Program of Shanghai College young teachers (No. ZZGCD15102); in part by Scientific Research Project of Shanghai University of Engineering Science (No. 2016-19); in part by the Shanghai University of Engineering Science Innovation Fund for Graduate Students (No. 18KY0613) and in part by Science and Technology Commission of Shanghai Municipality (No. 19030501100).

ABSTRACT Visual-based vehicle detection has been extensively applied for autonomous driving systems and Advanced Driving Assistant Systems however, it faces great challenges as a partial observation regularly happens owing to occlusion from infrastructure or dynamic objects or a limited vision field. This paper presents a two-stage detector based on Faster R-CNN for high occluded vehicle detection, in which we integrate a part-aware region proposal network to sense global and local visual knowledge among different vehicle attributes. That entails the model simultaneously generating partial-level proposals and instance-level proposals at the first stage. Then, different parts belong to the same vehicle are encoded and reconfigured into a compositional entire proposal through a Part Affinity Fields, allowing the model to generate integral candidates and mitigate the impact of occlusion challenge to the utmost extent. Extensive experiments conducted on KITTI benchmark exhibit that our method outperforms most machine-learning-based vehicle detection methods and achieves high recall in the severely occluded application scenario.

INDEX TERMS Deep convolutional neural network, occlusion handling, region proposal network, vehicle detection.

I. INTRODUCTION

Vehicle detection occupies a significant position in computer vision field with various applications, such as Intelligent Transportation System (ITS), autonomous driving, and traffic safety, which is committed to generating a series of bounding boxes enclosing vehicle instances on an image. Recently impressive works concerning object detection [1]-[5] are driven by the deep feature automatically extracted from deep convolutional neural networks (CNNs), which are able to generate 2D boxes related to scenes based on bounding box regression techniques.

However, vehicle detection is intensely challenging due to the difficulties from illumination condition, perspective distortion, viewpoints variations, especially ubiquitous occlusion which further aggravate the inter-object structural variations. As an example, 35.8% of the annotated vehicles are occluded by other vehicles or objects on the KITTI dataset [6]. Fig.1 displays several examples from KITTI captured at the same camera view angles, showing serious occlusion from other vehicles and truncation from static infrastructure. Conventional approaches aim at narrowing the gap between the predicted bounding box and its designated ground truth merely [7]-[8], rarely considering the occlusion occurred among different vehicle semantics parts. Thus, these detectors are sensitive to the rigorous threshold of non-maximum suppression (NMS) in the crowded traffic scenes, wherein filling with inter-object occlusion that increases the difficulty in vehicle localization. To that end, Tian et.al. [9] design a vehicle detection grammar to handle partial occlusion, which predicts semantic structure information of vehicle with visibility prediction by applying the trained grammars into the network. Whereas, further vehicle division after the semantic part proposals have generated can't make the initial sampling network learn more suitable feature for occlusion representation and guarantee a robust expression of vehicle viewpoint difference. Furthermore, as depicted in [10], guaranteeing 100% recall of the region proposal output is virtually impossible under the constraint that IoU should surpass 0.7. That will severely impact the performance of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2929432, IEEE Access

following detection because some troublesome cases must be tackled with high-quality proposals absence.

In this paper, we propose an alternative occlusion-focus region proposal algorithm based on the Faster R-CNN detection framework [11] to effectively mitigate the impact of undesirable occlusion. Specifically, to decrease the false localizations triggered by inter-object occluded or truncated vehicles, we expect the region proposal network is capable of concentrating on global features and meanwhile allocating attention on local properties, thus enforcing the pipeline to localize vehicles compactly to the corresponding instances, irrespective of vehicle parts is visible or not. Inspired by the bottom-up object detection strategy, we design a new region proposal network, termed part-aware RPN, not only to enable the detector to quickly capture unique characteristics of vehicles under various occlusions and viewpoints, but also to narrow the gap between the proposal and ground truth. The part-aware RPN replace the original RPN at the first stage of Faster R-CNN module, which integrates the prior semantics information of vehicle with occluded component prediction into the network. That is, the vehicle region is first partitioned into four independent parts and the net generates proposals under each part's prediction as well as the entire vehicle's prediction onto the feature maps into variable-scale box candidates. Subsequently, the learned Part Affinity Fields (PAFs) [12] is utilized to construct variable-length feature vectors and reconfigure each part belongs to the same vehicle for the final candidate generation.

In addition, great majority detection frameworks employ the probability scores from the CNNs benchmark to perform NMS to refine bounding boxes. However, probability scores above a certain level are not strongly related to the reliability of box proposals [13]. The most important reason is that the detector is trained to distinguish objects from background rather than sort the threshold of intersection-over-union (IoU). Therefore, we propose to use part-aware NMS to refine final results, e.g., through a cascaded method, NMS is performed on the instance candidates and the partial candidates of vehicle with different thresholds respectively. As long as not all of these candidates share a high degree of overlap, then it is possible to infer whether both the bounding boxes belong to the same car.

Extensive experiments are conducted on KITTI vehicle dataset, to demonstrate the superiority of the proposed method, especially for high occluded scenes. Notably, the partial perception of part-aware RPN enhances the probability of hitting the vehicle significantly without burdening heavy extra computational cost. The main contributions of this work are summarized as follows.

- We design a novel part-aware RPN to replace the original RPN of Faster R-CNN module to integrate the prior semantics information of vehicle with global and local properties.
- We propose a part-aware NMS, which performs NMS sequentially on the vehicle candidates and corresponding part candidates in a cascaded manner, eliminating miss detection of different vehicles under high IoU.

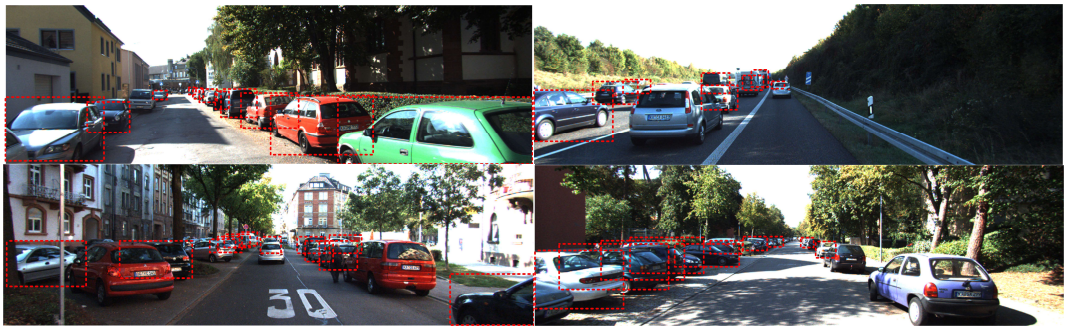


FIGURE 1. Example of vehicles captured with different occlusion and truncation on KITTI. Occluded or truncated vehicles are labeled with red dotted frames.

II. RELATED WORKS

Due to the fundamental requirement for the application of Advanced Driving Assistant System (ADAS) and increasing of video surveillance, numerous vision-based vehicle detection algorithms have been extensively exploited. Consequently, devoting investigation into the effective and efficient capacity of traffic vehicle detectors is achieving more academic significance and more practical importance.

The earliest object-detection system [14] used a cascaded strategy that obtains real-time performance with relatively impressive accuracy, which has been widely adopted as the backbone of sliding window pipelines for vehicle detection [15] and pedestrian detection [16]. Another robust object detection algorithm called part-based model which is parameterized by the geometric model and semantic properties of each sub-part perceiving geometric relationships of parts was proposed, some excellent examples like deformable part-based model (DPM) [17], [18]. DPM is similar to an extension of histogram of oriented gradients (HOG) that first calculates the gradient direction histogram, and then exploits a star-structured framework for severely occluded object detection by representing high variable objects.

Recently impressive vision-based detectors [19]-[22] have demonstrated the superior ability of deep CNNs. The capacity of automatic feature extraction allows these pipelines to dominate the top rank benchmarks in diverse vision applicants. Among these, OverFeat [23] trains several fully-connected layers to sense coordinates of box for the localization task that assumes a single object. Then multiple class-specific objects detection is completed by transferring features from fully-connected layers to a convolutional layer. Ren et al. [24] proposed R-CNN framework that combines selective search [25] for object proposal generation and CNN for proposals classification and regression. The two-stage strategy obtained excellent performance and is applicable to numerous object detectors. Whereas, the accuracy of R-CNN largely depends on the performance of proposal generation and is vulnerable in more complex environments. Faster R-CNN [11] advanced this pipeline by optimizing the original selective search with a Region Proposal Network (RPN). The learning attention mechanism of RPN guides the detector where to look and increases the computational efficiency by sharing convolutional features with the detector. Subsequently, Mask R-CNN [26] fine-tuned Faster R-CNN benchmark for object instance segmentation. In the procedure of prediction, it regresses candidates at multiple feature maps, allowing the net to perceive multi-scale objects with various receptive fields. This alternative scheme is also exploited in SSD [27] and Text Boxes [28]. Considering the redundant calculation of computing multiple feature maps independently, FPN [29] developed a top-down framework with lateral connections that cascades low-level feature maps with high-level feature maps at different scales. A multitude of detectors such as YOLOv3 [30] have implemented this tactic to detect objects across a range of scales.

All of these models concentrate on multiple object detection. It is necessary to design exclusive models for vehicles on the basis of the specific characteristics of cars. Reference [31] proposed an approach of learning reconfigurable parts from and-or models for vehicle detection, taking several vehicle-related elements into account. Then, an optimized and-or model was presented in [32] to explicitly represent occlusion configurations and context properties by modeling diverse combinations of part visibility. In [10], Chu et al. proposed a unique way of RoI representation that predicts the offset direction between RoI boundary and the relative ground box and completed ultimate detection through a multi-task learning framework. Furthermore, a novel vehicle region proposal algorithm that utilizes shapes of super-pixels to generate bounding box paradigm and score it with a graph-based algorithm was proposed in [33]. In response to the inherent scale-sensitivity of convolutional neural networks in vehicle detection task, SINet [34] was presented with a combination of context-aware RoI pooling and multi-branch decision network, concentrating on vehicle detection of multiple scales. PC-CNN [35] tends to fuse 2D detection pipelines with point clouds to percept 3D properties for accurate 3D bounding box generation. More recently, [36] proposed an original object description, 3D Voxel Pattern (3DVP), which simultaneously transfers the key properties of vehicles such as 3D shape, orientation, and truncation. The model seeks 3DVPs in a data-driven manner, and trains a set of dedicated detectors for a 3DVPs dictionary. In [9], the vehicle is represented by a two-layer grammar model that segments vehicle into semantic parts gradually. The joint training of specific grammar entails the model addressing the problem of occlusion to a certain of degree. In this paper, we jointly employ the entire vehicle and sub-parts classification as the semantic region proposals to enhance CNN-based detection performance.

III. PART-AWARE VEHICLE DETECTION

Reliably detecting a vehicle requires conquering a variety of challenges, including lighting variations, perspective distortions, especially occlusion and truncation among driving cars. Thus, the part-aware RPN is specially designed for inter-object occlusion addressing by generating high-quality proposals. In this section, we begin with a brief overview of the entire architecture, followed by specific details.

A. OVERVIEW OF OUR PROPOSED METHOD

In this paper, the key idea for occlusion handling is to partition vehicle into several independent parts rather than expressing it as a single box as previous object-detection methods do and then integrate parts share the same vehicle pattern. Whereas CNN-based frameworks aren't likely to detect an occluded vehicle completely due to the hidden parts, our model is capable of capturing the underlying properties from visible parts of a vehicle to correct the proposal of vehicle at the beginning.

Our proposed detector follows the adaptive Faster R-CNN detection framework [11] for vehicle detection, with the dedicatedly designed part-aware RPN (Section 3.2), and the relative part-aware NMS (Section 3.3). Generally, Faster R-CNN is composed of two modules, i.e., the first RPN module and the second Fast R-CNN module. The RPN module aims at generating proposals and inferring the Fast R-CNN module the exact location to sense, reaching the purpose of classifying object classes and regressing the accurate locations, based on proposals.

A pipeline of the proposed detector is manifested in Fig.2. The detector takes a color image of size $w \times h$ as input, and produces a series of 2D bounding boxes and corresponding vehicle's partial boxes. Firstly, the features of image are extracted through a convolutional network (initialized by VGG-16 [37] and fine-tuned), generating several feature maps $\mathbf{F}$ that are input to part-aware RPN. Secondly, the part-aware RPN simultaneously predicts instance-based candidates and part-based candidates, from which a batch of 2D confidence maps S are generated. Then, the candidates of vehicles and candidates of parts, along with the original

image features **F**, are concatenated and utilized to generate refined predictions. Finally, all these vehicle proposals are parsed by the Part-Aware NMS to reduce redundant boxes.

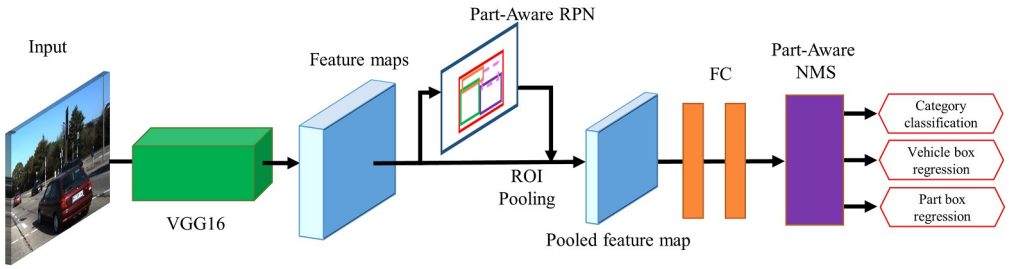


FIGURE 2. The architecture of the proposed part-aware detector. The VGG16 is utilized as the basic feature extraction network. Given an input image, we first exploit part-aware RPN to generate one instance-based proposal (shown in red) and several part-based proposals, including top part (shown in orange), front part (shown in green), side part (shown in purple), back part (shown in lavender), for each vehicle. Note that the back part is indicated by a dashed box because the front part and the back part can't appear in the image simultaneously. Then each proposal is flowed to the part-aware NMS to refine the final output.

B. PART-AWARE RPN

To accurately generate region proposals, we simultaneously predict part-based candidates and instance-based candidates, as depicted in Fig.3. The part-aware RPN is composed of two branches, the top branch shown in orange, predicts candidate boxes, generates confidence maps S and reconfiguration map, and the bottom branch, shown in blue, predicts the vector fields Q using PAFs [12] that encodes part-to-part or part-to-instance connection, where C represents convolution layer for feature extraction.

The common way of object annotation is framing the vehicle's largest circumscribed rectangle on the 2D imagery, including the occluded or truncated parts. However, such method isn't able to ensure that the network could perceive the visual information and occlusion information effectively, and thus, detection robustness can't be guaranteed. Instead, we assume a strongly supervised setting in which at training time we have ground truth bounding box annotations not only for full instances, but for a fixed set of semantic parts as well. Consequently, manually increasing annotation of instances, parts and part occlusions on real images is necessary to improve the expression of occlusion configurations and multi-car configurations. A car is split into four semantic parts as manifested in the Fig.4(b) with different colors. Those are the front of vehicle S^F, the side of vehicle S^S, the top of vehicle S^T, the back of vehicle S^B. Note that on 2D imagery one vehicle may compose no more than three semantic parts and different vehicles may contain unequal annotated parts due to the distinct viewpoint. Additionally, the left and right parts of vehicle usually exhibit similar features, and the two parts do not appear simultaneously in the image, hence we represent both sides in one way. Then a large number of part examples are generated by annotating each visible or invisible part. At the same time, the bounding box of entire vehicle is also necessary.

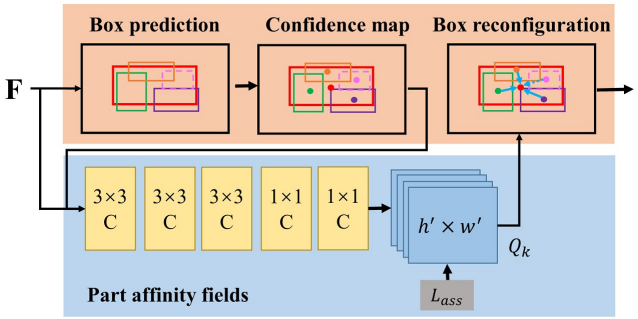
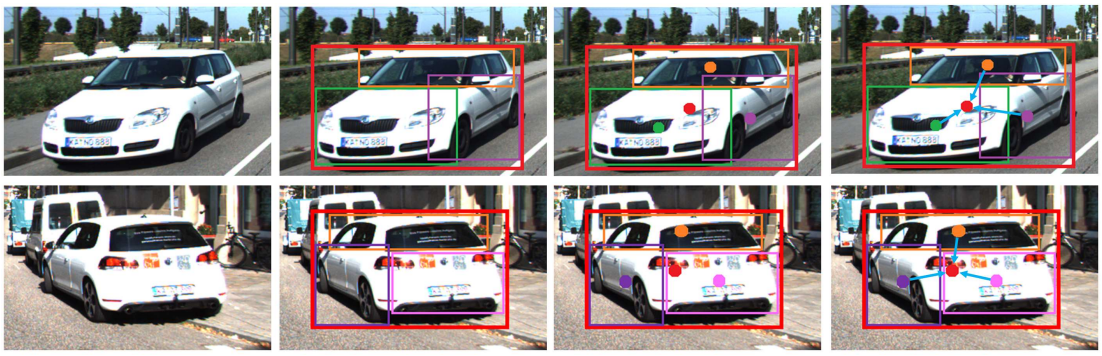


FIGURE 3. Architecture of part-aware RPN. First, the proposal prediction and confidence map of different parts are predicted. Then the PAFs constructs association vector fields Q_k that store the spatial association of different parts as specific layers convolutional weights, from which we could obtain exact vehicle proposals.



(a) (b) (c) (d)

Figure.4. Vehicle semantic parts segmentation and association. (a) Original image. (b) Semantic parts segmentation. (c) Confidence maps generation. (d) Vehicle parts reconfiguration. The vehicle above and the vehicle below contain different types of semantic parts, the vehicle above is represented with a top part, a front part and a side part, while the vehicle below is represented with a top part, a back part and a side part.

Given these annotations of vehicles, at training time all vehicles and each class of parts are initially treated as independent categories. Specially, the ground truth confidence map S is generated at the center of box by producing a fixed-size circle $r=(w_p \times h_p)/50$, where w_p and h_p are the width and height of proposal box respectively, as depicted in Fig.4(c). Each confidence map is an explicit 2D representation of the belief that a particular vehicle part or vehicle instance occurs at each pixel location. Ideally, each independent part of vehicle will generate one confidence map if the corresponding part has been predicted.

Given a series of proposed vehicle parts, assembling them to form full-vehicle boxes with an unknown number of vehicles is ambiguous. We need a confidence measure of the association for detected vehicle parts, i.e., that they belong to the same vehicle. Inspired by PAFs, we design our part association net to learn the connection of different parts. It is a reimplementation of PAFs with an identical block structure. We show the PAFs in figure 3, where it preserves both location and orientation information across the regions of vehicle through convolutional operation. The part affinity is a vector field for each vehicle, also shown in Fig.4(d): for each pixel in the area belonging to a particular vehicle, a ξ D vector encodes the directions that point from one part of the vehicle to the others is predicted, where $\xi \in \{2,\ 3,\ 4\}$. Each vehicle has a corresponding affinity field joining its associated parts.

The set $S=(S_1^F, S_2^F, ..., S_X^F;\ S_1^B, S_2^B, ..., S_Y^B;\ S_1^S, S_2^S, ..., S_Z^S;\ S_1^T, S_2^T, ..., S_N^T;\ S_1^V, S_2^V, ..., S_M^V)$ has $X+Y+Z+N+M$ confidence maps, where $S_x^F \in \mathbf{R}^{w \times h}$, $x \in \{1...X\}$, $S_y^B \in \mathbf{R}^{w \times h}$, $y \in \{1...Y\}$, $S_z^S \in \mathbf{R}^{w \times h}$, $z \in \{1...Z\}$, $S_n^T \in \mathbf{R}^{w \times h}$, $n \in \{1...N\}$ and $S_m^V \in \mathbf{R}^{w \times h}$, $m \in \{1...M\}$ represent the front part, the back part, the side part, the top part and the entire vehicle confidence map respectively. The set $Q=(Q_1, Q_2, ..., Q_C)$ has C vector fields, where $Q_c \in \mathbf{R}^{w \times h \times d}$, $c \in \{1...C\}$, $d \in \{2,\ 3,\ 4\}$ represents the vector fields, in which $d \in \{2,\ 3,\ 4\}$ indicates that different vehicles may consist of diverse parts due to the perspective variations. Let $G_{j\delta,k}$, $\delta \in \{1,\ 2,\ 3,\ 4\}$ be the ground truth positions of parts j_δ vehicle k in the image. We define the ground truth part affinity vector field, Q_k^*, at an image point p as

$$Q_k^*(p)=\begin{cases} v & \text{if } p \text{ on vehicle } k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here, $v=\sum_{\lambda=1}^{\delta}(x_{j(\lambda+1),k}-x_{j\lambda,k})/\parallel x_{j(\lambda+1),k}-x_{j\lambda,k} \parallel_2$ is the unit vector in the direction of several semantic proposals. The ground truth PAFs is achieved by averaging the affinity fields of all vehicles in this domain,

$$Q^*(p)=\frac{1}{n(p)}\sum_k Q_k^*(p) \tag{2}$$

where $n_c(p)$ denotes the number of non-zero vectors at image point p across all k vehicles. Specifically, for several part proposals d_{j1}, we sample the predicted PAFs, Q_k along the line segment to calculate the confidences in their association:

$$E=\int_0^{u=1} Q_k(p(u)) \cdot \sum_{\lambda=1}^{\delta} \frac{d_{j\lambda+1}-d_{j\lambda}}{\parallel d_{j\lambda+1}-d_{j\lambda} \parallel_2} du \tag{3}$$

where $p(u)$ represents the location of two vehicle parts $d_{j\delta}$ and $d_{j(\delta+1)}$.

$$p(u)=(1-u)d_{j\delta}+ud_{j(\delta+1)} \tag{4}$$

In experiments, the integral is approximated by sampling and summing uniformly-spaced values of u.

In this way, we add one more task of part-aware PRN to the pipeline, and the loss function of the part-aware RPN can be written as follows:

$$\begin{aligned} &L_{RPN}(\{p_i\},\{p_i^*\},\{Q_i\},\{Q_i^*\})=L_{cls}(\{p_i\},\{p_i^*\})+ \\ &\alpha \cdot L_{reg}(\{p_i^*\},\{t_i\},\{t_i^*\})+L_{ass}(\{Q_i\},\{Q_i^*\}) \end{aligned} \tag{5}$$

where i is the index of anchor in a mini-batch, p_i, t_i and Q_i are the predicted confidences of the ith anchor being a vehicle or a part, the predicted coordinates of the vehicle or part and the predicted part affinity vector field, p_i^*, t_i^* and Q_i^* are the associated ground truth class label, coordinates of

the ith anchor and ground truth part affinity vector field, α is the hyper-parameters used to balance the three loss terms, $L_{cls}(\{p_i\},\{p_i^*\})$ is the classification loss, $L_{reg}(\{p_i^*\},\{t_i\},\{t_i^*\})$ is the box regression loss, and $L_{ass}(\{Q_i\},\{Q_i^*\})$ is the association loss, which is defined as:

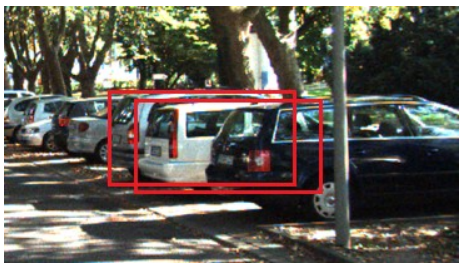
$$L_{ass}(\{Q_i\},\{Q_i^*\}) = \sum_p W(p)\cdot \| Q_c^i(p) - Q_c^*(p) \|_2^2 \qquad (6)$$

Where W is a binary mask with $W(p)$=0 when the annotation is missing at an image location p.

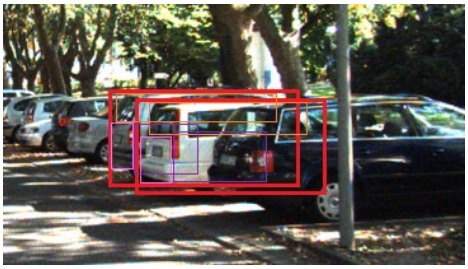
C. PART-AWARE NMS

In complex traffic scenes, such high overlap between cars may severely harm vehicle detection performance. As an example, in Fig.5(a), two cars are close to each other and the IoU is above 0.8. Although the basic pipeline is able to locate them and assign them high scores, the standard post-processing step NMS will filter the lower scored bounding box. Given a high threshold of the NMS, both two bounding boxes may be preserved. However, the recall of the detection results will decrease seriously.

To solve this dilemma, we introduce the part-aware NMS tactic. In our part-aware RPN, both the part candidates and vehicle candidates are predicted. Even if the two cars have a large degree of IoU of the vehicle box, a lower degree of IoU between the corresponding parts box must exist, except the two prediction frames share the same car. Our part-aware NMS exploits a cascaded pipeline. First, the standard NMS is exploited for the vehicle bounding boxes with a relatively loose threshold like 0.65. Then, bounding boxes corresponding to different categories of parts will implement NMS in turn, and all partial bounding boxes of all the culled bounding boxes will be eliminated. For example, NMS is first performed on the side candidate boxes, and then NMS is performed on the back candidate boxes… With the proposed part-aware NMS, the precision and recall of the detection results can achieve a satisfactory balance.



(a)



(b)

Figure 5. Part-aware NMS implantation. (a) Two boxes with IoU larger than 0.7. (b) Exploiting NMS on different parts.

IV. EXPERIMENTS

In this section, the proposed detector is validated on the KITTI dataset and compared with other state-of-the-art methods. Vehicle detection on KITTI faces a fundamental challenge that a majority of vehicles are truncated or occluded. A total of 35.8% of vehicles suffer from occlusion or incomplete display in the training set within 28742 annotated samples. The test annotations of KITTI benchmark is not available, hence we randomly split the KITTI trainset into the training and testing subsets followed by [31]. Furthermore, for the training of part-aware RPN, we added the labels of the vehicle parts to the original label of the data set.

A. IMPLEMENTATION DETAILS

Following Faster R-CNN benchmark, pre-trained ImageNet model—VGG16 is utilized to initialize the pipeline, which is implemented on most advanced frameworks [27]. The VGG16 network is trained on the single-scale set, the last max-pooling layer is replaced by the RoI pooling layer to pool the feature maps of each object proposal into fixed resolution, i.e. 14×14. The final fully-connected layer and softmax are replaced with two sibling fully-connected layers. Both the part-aware RPN and vehicle detection network are trained with Stochastic Gradient Descent (SGD) with a momentum of 0.9, and weight decay of 0.0005. We use the anchors as same as Faster R-CNN with three different scales. Each mini-batch contains 60 randomly sampled proposals, of which 20 positive proposals are with IoU with the ground truth box larger than 0.5, and the remaining 40 proposals which have IoU with the ground-truth bounding box less than 0.5 are treated as negative training instances. In the loss function, we use α =1 to balance classification loss, regression loss and association loss.

The pipeline is implemented on the open-source Tensorflow platform [38]. We trained the whole network on a single NVIDIA TITAN X GPU with 12GB memory. During training, the shared convolutional layer in VGG16 maintains initial parameters. The standard NMS is exploited before vehicle parts association with a fixed threshold at 0.65, which decreases the number of proposals to 3000 approximately. In the following, the parameters of remaining layers updated with an initial learning rate of 0.002.

B. PART-AWARE RPN EVALUATION ON RECALL

Table I presents the region proposal performance in terms of recall on the KITTI test dataset. The whole set is split into three subsets (Easy, Moderate, Hard) on the basis of difficulties among occlusion, truncation and object size. Performance of the Edge Boxes [39] and the RPN [12] algorithms are comparable to our method on the easy set.

Despite, when it comes to more challenge scale variation, occlusion of objects and more complex background on the moderate and hard set, RPN and Edge Box fall a lot. We also fine-tuned CRAFT [40] and GA-RPN [41] on KITTI data set and compared its region proposal performance with our method. These approaches are optimized and improved on the basis of RPN. CRAFT is expert in proposal generation on easy and moderate subcategory while with more time consummation. GA-RPN is capable of addressing scale-variation and maintaining high computational efficiency. Obviously, part-aware RPN is capable of outperforming other bottom-up region proposal methods on the hard subcategory, exhibiting that the part-aware RPN can accurately localize the objects based on the deep features combined of local and global properties. In Fig.6, some examples of proposals generation upon RPN, CRAFT and our method are visualized.

TABLE I
SEVERAL REGION PROPOSAL METHOD PERFORMANCES OF RECALL ON KITTI DATASET

Method	Time	Easy	Moderate	Hard
EB [39]	0.25	81.40	65.95	53.86
RPN [11]	0.09	84.24	68.91	58.25
CRAFT [40]	1.32	87.01	76.04	67.21
GA-RPN [41]	0.13	84.56	75.04	65.36
Part-Aware RPN	0.12	85.29	76.54	68.54

C. RESULTS ON KITTI DATASET

In this section, the improvement of our method brings to the detection task is demonstrated. The mAP with different IoU thresholds was adopted for evaluation. We carefully fine-tuned a Faster R-CNN vehicle detector on KITTI. As reported in Table II, our detector performs much better than the original Faster R-CNN implementation. Since both Faster R-CNN and our method share the same initialized VGG-16 network as the backbone network, the achievement is largely attributed to the direct consequence of the novel optimized part-aware RPN architecture actually. Note that the detector outperforms the backbone by more than 6% if the threshold of IoU is increased to 0.8, which strongly validates the robustness of our method in generating high-quality object boxes. Meanwhile, Fig.7 illustrates some detection examples between the two methods on KITTI test dataset. It's obvious that occluded cars which are formidable to perceive occupy a majority of failure cases for detection. Actually, faster R-CNN achieve low recall rate when vehicles are heavily occluded or truncated, presenting the bottleneck for state-of-the-art object detector regarding generating high-quality bounding box. We can tackle this problem well with the adoption of part-aware RPN.

Figure 6. Examples of RPN proposals (top row), CRAFT proposals (middle row) and our method proposals (bottom row).

Figure 7. **Comparison between Faster-R-CNN and our method. Top row: results of Faster R-CNN, failed to localize vehicles with heavy occlusion. bottom row: results of our method, outperform the backbone with IoU higher than 0.7 to the ground truth.**

Table III shows the numerical comparison results on the three sub-categories, where we demonstrate the effectiveness of multi-components performance, part-aware RPN, multi-part localization and part-aware NMS are all impactful design. Qualitative results of several pipelines are shown in Fig.9. Particularly, for those vehicles belong to moderate and hard level, our method performs better with the efficiency of fore-mentioned components. Among all the published methods, our method ranks on top on the basis of hard level. It is worth noticing that our model is comparable with RVCNN [10], CRAFT [40], F-PointNet [42], Mono3D [43] and Deep3DBox [44]. Our method slightly outperforms SubCNN [45] on the moderately and hard categories, but a bit inferior on the easy category. Although our method obtains better rank than Deep3Dbox on the hard category, it is inferior on the easy and moderate subsets. Since Deep MANTA [46] utilizes more landmark information, it is able to regress more accurate bounding box when vehicles have good visual effect, hence the highest precision is achieved on the easy subcategory. Similarly, considering the deep information utilized in Deep3DBox, vehicles are easier to located especially. However, by exploiting part information of vehicles explicitly, our method exhibits the promised improvement on the hard subcategory when vehicles are occluded heavily compared with Deep MANTA and Deep3DBox. Comparing Faster R-CNN, our method exceeds it on the moderate and hard set with a considerable margin (seven percent and eight percent respectively), while on the easy category, we obtain small improvement (1.42 percent). These results infer that vehicle detection in complex scenes has large potential improvement with more reliable deep features, and the proposed algorithm can address occlusion effectively. Precision-recall curves on the KITTI dataset of the moderate category is illustrated in Fig.8. Meanwhile, some detection examples of different scenes on KITTI test set upon our method is provided in Fig.10.

TABLE II
PERFORMANCE ON THE KITTI DATASET FOR DIFFERENT IOU THRESHOLDS

Method	0.6	0.65	0.7	0.8
Faster R-CNN [11]	92.44	90.55	89.26	77.14
OURs	93.56	91.36	90.55	83.15

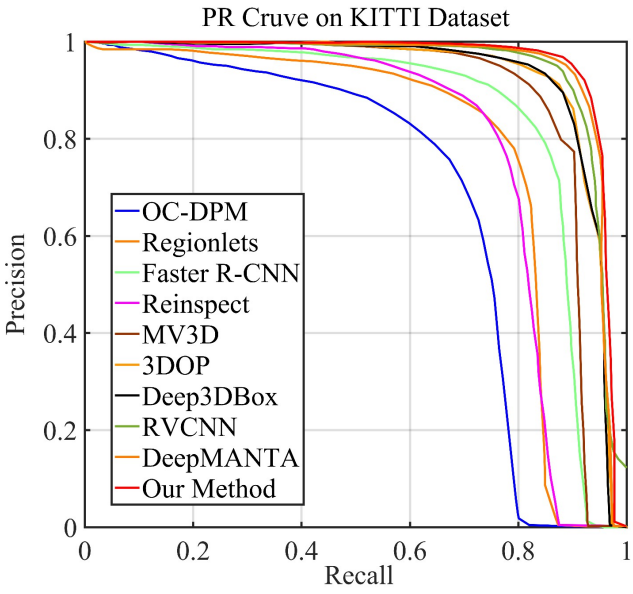


Figure 8. **Precision-recall curves on the KITTI test set of the moderate category.**

TABLE III
VEHICLE DETECTION PERFORMANCE ON KITTI TEST SET WITH IOU OF 0.7

Method	Time	Easy	Moderate	Hard
3DVP [36]	40	87.46	75.77	65.38
MV3D [47]	0.36	89.11	87.67	79.54
PC-CNN-V2 [35]	0.5	90.79	90.15	87.58
SINet+ [34]	0.3	90.51	89.73	77.82
SINet_VGG [34]	0.2	90.60	89.56	78.19
SINet_PVA [34]	0.11	90.44	89.08	75.85
Aston-EAS [51]	0.24	90.49	89.64	77.95
F-ConvNet [52]	0.47	90.44	89.79	80.66
RefineNet [48]	0.22	89.88	79.17	66.38
Faster R-CNN [11]	2	86.71	81.84	71.12
F-PointNet [42]	0.17	90.78	90.00	80.80
3DOP [49]	3	93.04	88.64	79.10
RRC [22]	3.6	90.61	90.23	87.44
Mono3D [43]	4.2	92.33	88.66	78.96
MS-CNN [50]	0.2	90.03	89.02	76.11
CRAFT [40]	4.5	90.76	90.00	81.83
UberATG-MMF [53]	0.08	91.82	90.17	88.54
TuSimple [54]	1.6s	90.77	90.33	82.86
SJTU-HW [55]	0.85	90.81	90.08	79.98
SubCNN [45]	2	90.81	89.04	79.27
Deep3DBox [44]	1.5	92.98	89.04	77.17
Deep MANTA [46]	0.7	95.77	90.03	80.62
Our model	2.1	90.21	89.01	80.72

Figure 9. Comparison of four different methods on KITTI test set. From top to bottom: SINet, CRAFT, F-PointNet and our method. It is obviously that when the vehicles are truncated or the occlusion is serious, our method is capable of locating them more reliably and regressing relatively accurate bounding boxes.

Figure 10. Detection results of our method on KITTI testing set.

D. ABLATION STUDY

To investigate the behavior of the proposed method, we conducted two ablation studies on part-aware RPN and part-aware NMS separately. First, we disentangle the RPN's influence on training the Faster R-CNN detection network. For this purpose, we trained a Faster R-CNN model by using

part-aware RPN with different vehicle part combinations. We fixed this detector and evaluated the detection mAP by changing the number of vehicle parts combined at test-time on the moderate category. As depicted in Table IV, for the threshold of 0.7, the result of associating four semantic parts and one vehicle instance outperforms the results using three or fewer parts. Note that when only the instance-level box is used, the entire detector is equivalent to Faster R-CNN. Specially, it is 7.1% higher than one part of vehicle box and 1.8% higher than three parts of vehicle box, top part, front part. The loss in mAP is mainly because of the inconsistency between the training/testing proposals. This result serves as the baseline for the following comparisons.

To reveal the importance of part-aware NMS in occlusion handling, especially for accurate localization, the performance comparison between two types of NMS is summarized in Table V, adopting a part-aware NMS with four semantics parts achieves better performance than original NMS, despite the two types roughly carried on the same set. Specially, part-aware NMS outperforms original NMS with 2.5% on the hard set, which strongly suggests that a significant part of the success of our architecture comes from its ability to occlusion conquering.

TABLE IV
CONTRIBUTION OF EACH PART OF VEHICLE

Vehicle box	Side part	Top part	Front part	Back part	mAP(moderate)
✓					81.84
✓	✓				84.35
✓	✓	✓			87.21
✓	✓	✓	✓		88.07
✓	✓	✓	✓	✓	89.01

TABLE V
CONTRIBUTION OF PART-AWARE NMS.

Method	Easy	Moderate	Hard
NMS	90.01	88.26	78.26
Part-aware NMS	90.21	89.01	80.72

V. CONCLUSION

In this paper, we developed a novel vehicle detection algorithm focus on occlusion and truncation handling based on vehicle part-based proposals generation and PAFs-based combination algorithm. The problem of occlusion in complex backgrounds could be addressed effectively by enforcing the pipeline to simultaneously explore the part-based and instance-based features. Experiments that evaluate the performance of the region proposal algorithms were carried on the KITTI datasets at three difficulty levels. Compared with the state-of-the-art vehicle detection approaches, the proposed method indeed improves the performance of accuracy especially when vehicles are heavily occluded.

REFERENCES

[1] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-Shot Refinement Neural Network for Object Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[2] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 2999-3007, 2017.

[3] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu, "CoupleNet: Coupling Global Structure with Local Parts for Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[4] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[5] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel Feature Pyramid Network for Object Detection," in *The European Conference on Computer Vision (ECCV)*, 2018.

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 3354-3361, 2012.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904-1916, 2015.

[8] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable Object Detection using Deep Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2155-2162, 2013.

[9] B. Tian, M. Tang, and F. Y. Wang, "Vehicle detection grammars with partial occlusion handling for traffic surveillance," *Transp. Res. Part C Emerg. Technol.*, vol. 56, pp. 80-93, 2015.

[10] W. Chu, Y. Liu, C. Shen, D. Cai, and X. S. Hua, "Multi-task vehicle detection with region-of-interest voting," *IEEE Trans. Image Process.*, vol.27, no. 1, pp. 432-441 2018.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, 2017

[12] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[13] S. Liu, C. Lu, and J. Jia, "Box aggregation for proposal decimation: Last mile of object detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2569-2577, 2015.

[14] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vis.*, 2004.

[15] E. Ohn-Bar and M. M. Trivedi, "Learning to Detect Vehicles by Clustering Appearance Patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2511-2521, Mar. 2015.

[16] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3361-3369, Dec. 2015.

[17] D. Forsyth, "Object detection with discriminatively trained part-based models," *Computer (Long. Beach. Calif).*, vol.47, pp.6-7, 2014.

[18] S. Sivaraman and M. M. Trivedi, "Vehicle detection by independent parts for urban driver assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1597 - 1608 , June 2013.

[19] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen, "Deep network cascade for image super-resolution," in

European Conference on Computer Vision., pp.49-64, 2014.

[20] X. Wang, W. Zhang, X. Wu, L. Xiao, Y. Qian, and Z. Fang, "Real-time vehicle type classification with deep convolutional neural networks," *J. Real-Time Image Process.*, vol. 16, no. 1, pp. 5–14, 2019.

[21] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[22] J. Ren *et al.*, "Accurate single stage detector using recurrent rolling convolution," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 752-760, 2017.

[23] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," 2013.

[24] R. Girshick, J. Donahue, T. Darrell, U. C. Berkeley, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012.

[25] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, pp. 154-171, 2013.

[26] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[27] W. Liu *et al.*, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*, pp. 21-37, 2016.

[28] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2016.

[29] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 936-944, 2017.

[30] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, 2018.

[31] B. Li, T. Wu, and S. C. Zhu, "Integrating context and occlusion for car detection by hierarchical and-or model," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014.

[32] 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. 39, no. 9, pp.1829-1843, 2016.

[33] X. Yuan, S. Su, and H. Chen, "A graph-based vehicle proposal location and detection algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 12, pp. 3282-3289, 2017.

[34] X. Hu et al., "SINet: A Scale-Insensitive Convolutional Neural Network for Fast Vehicle Detection," IEEE Trans. Intell. Transp. Syst., 2019.

[35] X. Du, M. H. Ang, S. Karaman, and D. Rus, "A General Pipeline for 3D Detection of Vehicles," in Proceedings - IEEE International Conference on Robotics and Automation, 2018.

[36] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D Voxel Patterns for object category recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1903-1911, 2015.

[37] C. Chung *et al.*, "Very Deep Convolutional Networks for Large-Scale Image Recognization," in *ICLR,* 2018.

[38] M. Abadi *et al.*, "TensorFlow : A System for Large-Scale Machine Learning," *arXiv preprint arXiv: 1605.08695* 2016.

[39] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, pp. 391-405, 2014.

[40] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "CRAFT objects from images," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016.

[41] J Wang et al., "Region Proposal by Guided Anchoring," arXiv:1901.03278v2, 2019.

[42] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D Object Detection from RGB-D Data," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.

[43] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D Object Detection for Autonomous Driving," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2147-2156, 2016.

[44] A. Mousavian, D. Anguelov, J. Košecká, and J. Flynn, "3D bounding box estimation using deep learning and geometry," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 5632-5640, 2017.

[45] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-Aware convolutional neural networks for object proposals & detection," in *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, 2017.

[46] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 1827-1836, 2017.

[47] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 6526-6534, 2017.

[48] O. B. Eshed, R. N. Rajaram, and M. M. Trivedi, "RefineNet: Iterative refinement for accurate object localization," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, pp. 1528–1533, Nov. 2016.

[49] X. Chen *et al.*, "3D Object Proposals for Accurate Object Class Detection," in *Advances in Neural Information Processing Systems (NIPS)*, 2015

[50] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[51] J. Wei, J. He, Y. Zhou, K. Chen, Z. Tang, and Z. Xiong, "Enhanced Object Detection With Deep Convolutional Neural Networks for Advanced Driving Assistance," IEEE Trans. Intell. Transp. Syst., 2019.

[52] Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection

[53] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-Task Multi-Sensor Fusion for 3D Object Detection," in IEEE Conference on Computer Vision and Pattern Recognition, Proceedings, CVPR, 2019.

[54] F. Yang, W. Choi, and Y. Lin, "Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016.

[55] S. Zhang et al., "LED: LOCALIZATION-QUALITY ESTIMATION EMBEDDED DETECTOR," IEEE International Conference on Image Processing 2018, 2018.