Copyright © 2019 Tech Science Press CMES, vol.119, no.3, pp.559-576, 2019

# 3D Bounding Box Proposal for on-Street Parking Space Status Sensing in Real World Conditions

Yaocheng Zheng[1], Weiwei Zhang[1, *], Xuncheng Wu[1] and Bo Zhao[1, *]

**Abstract:** Vision-based technologies have been extensively applied for on-street parking space sensing, aiming at providing timely and accurate information for drivers and improving daily travel convenience. However, it faces great challenges as a partial visualization regularly occurs owing to occlusion from static or dynamic objects or a limited perspective of camera. This paper presents an imagery-based framework to infer parking space status by generating 3D bounding box of the vehicle. A specially designed convolutional neural network based on ResNet and feature pyramid network is proposed to overcome challenges from partial visualization and occlusion. It predicts 3D box candidates on multi-scale feature maps with five different 3D anchors, which generated by clustering diverse scales of ground truth box according to different vehicle templates in the source data set. Subsequently, vehicle distribution map is constructed jointly from the coordinates of vehicle box and artificially segmented parking spaces, where the normative degree of parked vehicle is calculated by computing the intersection over union between vehicle's box and parking space edge. In space status inference, to further eliminate mutual vehicle interference, three adjacent spaces are combined into one unit and then a multinomial logistic regression model is trained to refine the status of the unit. Experiments on KITTI benchmark and Shanghai road show that the proposed method outperforms most monocular approaches in 3D box regression and achieves satisfactory accuracy in space status inference.

**Keywords:** 3D object proposal, image processing and analysis, parking space detection, fully convolutional network, multinomial logistic regression model.

## 1 Introduction

Nowadays video surveillance gains a wide application in Intelligent Transportation System (ITS) thanks to the notable progress in machine vision techniques. Follow this tendency, a multitude of vision-based vehicle detection algorithms have been extensively exploited for on-street parking lot management, providing vehicle location, space information and vacant space guidance for drivers. Unfortunately, accurate and robust vehicle detection is intensely challenging due to the difficulties from illumination condition, various size, different viewpoints, truncation and occlusion. Especially in scenes such as on-street parking lot, vehicles are always linearly aligned and easily

[1] College of Mechanical Automotive Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China.

[*] Corresponding Author: Weiwei Zhang. Email: zwwsues@163.com.

occluded each other, posing high demands on the mounting height and angle of cameras. Furthermore, adjacent vehicles show severe overlap on the image plane due to a limited perspective of camera, thus the detection algorithm must be able to extract robust features to accurately distinguish their boundaries.

Recently impressive CNN-based works in 2D object detection [Liu, Anguelov, Erhan et al. (2015); Redmon and Farhadi (2018); Ren, He, Girshick et al. (2017)] are able to provide significant information and generate 2D box related to scenes content but does not allow to describe objects in 3D real world scene [Chabot, Chaouch, Rabarisoa et al. (2017)]. Additionally, typical vision-based methods achieve very low recall rate when objects are heavily occluded or truncated, presenting the bottleneck for state-of-the-art object detectors like Faster R-CNN [Ren, He, Girshick et al. (2017)] and SSD [Liu, Anguelov, Erhan et al. (2015)] on this benchmark. Hence, the representation of vehicle location should detailedly describe vehicle boundary and center. In this paper, we focus on 3D vehicle analysis from monocular images. For parking space sensing in the real world, vehicle localization and surrounding vehicles description jointly utilized with spatial interpretation are obviously indispensable. As an example, current location of vehicles on the road is required to infer the relationship between vehicles and spaces, which is used to determine whether vehicles are inside parking spaces or outside parking spaces. In another case, it can be empirically manifested that 2D vehicle detection is highly unreliable for space status inference because the exact location of vehicle is represented by the pixels in the image, thus determining the interaction between the vehicle and the specific space is arduous. Here we present an approach that, given a single image, provides exact vehicle localization, vehicle distribution map, parking space status and non-standard parking recognition, as illustrated in Fig. 1.

We proposed a novel parking space status inference scheme based on cascaded net, camera inverse projection matrix, and multinomial logistic regression model (MLRM). Composed of ResNet [Wu, Zhong and Liu (2016)] and Feature Pyramid Networks (FPN) [Lin, Dollár and Girshick (2017)], the cascaded net is utilized to extract multi-scale stereo features of vehicles and predict a set of high-quality 3D candidate boxes, dealing with truncation or occlusion adaptively. Here we introduce the 3D anchor to enforce the cascaded net to learn an excellent representation for different vehicle shapes with various length, width and height ratio. Inspired by the 3D templates in Chabot et al. [Chabot, Chaouch, Rabarisoa et al. (2017)], 3D anchors are clustered from the ground truth box in the source dataset according to diverse geometrical templates. In our method, proposals are generated at multi-scale feature maps based on five different types of 3D anchors. In the procedure of space status inference, MLRM is trained to model this process and complete inference by combining three adjacent parking spaces into one unit. This guarantees that the outputs will not be interfered by adjacent parked vehicles. Extensive experiments are conducted on two data sets, KITTI benchmark and a surveillance data set collected in Shanghai. Our data set is developed containing more than 7 thousands of labeled ground truth of both fully and partially observed vehicles on four distinctive roads across the scenes of various parking space.

The contributions of our work can be summarized as follows:

- First, we introduced 3D vehicle detection for a more distinct description of vehicle location on roadside parking lot, providing an intuitive way to infer pace status and non-standard parking.
- Second, we discovered that the key for improving 3D box generation by optimizing conventional 2D anchor to 3D anchor. This procedure can be efficiently implemented with the proposed cascaded network.

The rest of the article is organized as follows. Section 2 summarizes relevant research of parking space monitoring system. The architecture of the roadside monitoring system and the procedure of space status inference are presented in Section 3, including some details to address occlusion and other issues. Then, Section 4 evaluates our system through an experiment on a dedicated dataset and compares the system performance with other similar detection systems. Section 5 concludes our research of this paper.

## 2 Related work

### *2.1 Parking space sensing*

Currently numerous vision-based approaches gradually introduced for on-street parking space status inference. For instance, Yang et al. [Yang, Ju, Hsieh et al. (2017)] proposed a real-time parking space monitoring and guiding system based on image analysis, collecting information of space through event recorders embedded in cars on the roads. Sevillano et al. [Sevillano, Marmol and Fernandez-Arguedas (2014)] proposed a vacant on-street parking space detection system by combining the widely-deployed video surveillance camera networks and parking sensor networks together. Some other methods try to determine the vacant parking space by extracting the foreground and background information of the parking space [Pazos, Müller, Favre-Bulle et al. (2016); Almeida, Oliveira, Silva et al. (2013)]. However, false detection usually occurs in these methods due to challenges e.g. lighting variations and occlusion.

Huang et al. [Huang, Tai and Wang (2013)] proposed a plane-based 3D scene model composed of plentiful planar surfaces, which contribute to handling inter-object occlusion and perspective distortion. By means of scene layer, label layer, and observation layer, the plane-based Bayesian hierarchical framework is presented to address challenges of status inference and performance improvement. Meanwhile, Masaki [Masaki (1998)] identify parking space status by tracking and recording the trajectory of vehicle. Recently, a decentralized parking lot occupancy detection approach was proposed by Amato et al. [Amato, Carrara, Falchi et al. (2016)]. Based on a specifically designed deep Convolutional Neural Network (CNN), they integrated the entire system on a smart camera and obtained real-time parking space information subsequently. Similar to the method in Huang et al. [Huang, Tai and Wang (2013)], A Multilayer Inference Framework based on Bayesian network for vacant parking lot detection was presented in Huang et al. [Huang and VU (2017)]. The framework consists of four component layers: an image layer, a patch layer, a space layer, and a lot layer. It is able to extract high-level features from different patches of space and well address visual challenges such as lighting variations, casting shadows, and perspective distortion.

All of the above methods were designed to identify parking space status through vehicle detection algorithms. By extracting detailed properties from vehicles, such as edge, texture and color, these methods are capable of successfully sensing vacant spaces to a certain degree. Whereas, none of them has mentioned precise localization of vehicles and the recognition of non-standard parking. Based on this, the proposed system aims to address the aforementioned problems by exploiting 3D vehicle detection.

### *2.2 Vision-based approaches*

Accurate 3D object detection means considerable significance for ITS, and an ocean of vision or stereo or/and Lidar-based approaches were presented recently. The PointNet [Qi, Liu, Wu et al. (2017)] and RoarNet [Shin, Kwon and Tomizuka (2018)] conduct object proposals directly on the point clouds themselves, considering a series of points which lie within a frustum defined by a 2D object proposal through images. Whereas, Chen et al. [Chen, Kundu, Zhu et al. (2018)] introduced a novel 3D object proposal method by formulating the whole problem as minimizing an energy function that received detailed properties obtained from image and point cloud densities. Under the combination of image captured by the camera and point cloud data generated by the radar, these works enhance the sensing ability of the algorithm, and locate object accurately on the premise that the input is abundant enough although more computational cost.

Without stereo information, it is relatively difficult to generate accurate 3D boxes with monocular image only. Hence, numerous works dedicate to refine candidate boxes through deep features extracted from real world scene. Given a monocular image, they encode contextual information, ground plane, object size, location and semantic segmentation [Chen, Kundu, Zhang et al. (2016)]. In Xiang et al. [Xiang, Choi, Lin et al. (2015)], Xiang et al. proposed an original object description, 3DVP, that simultaneously transfers the key properties of vehicles such as 3D shape, orientation, and truncation. After that, SubCNN [Xiang, Choi, Lin et al. (2017)] is proposed through a novel CNN-based 3D bounding box proposal network that using subcategory information to guide the candidates generating process. Other work such as Mousavian et al. [Mousavian, Anguelov, Košecká (2017)] combined 2D bounding box with geometric constraints on translation and regress relatively stable 3D object properties. Recently Roddick et al. [Roddick, Kendall and Cipolla (2018)] introduced the orthographic feature transform, which provides a totally different view for the CNN to holistically reason about the spatial configuration of the scene.

## 3 Parking space status inference

In this paper, a prominent change by generating 3D bounding box with monocular image to infer space status and estimate vehicle pose (e.g., non-standard parking) through an elegant and effective solution with minimal computing overhead is proposed. As depicted in Fig. 1, the system is composed of two modules, i.e., 3D bounding box generation and space status inference. Firstly, the monocular image is passed through the cascaded net that output a series of 3D boxes, associated with 3D anchors (vehicle templates). The cascaded net architecture is detailed in Section 3.1. The second procedure is the space status inference which uses coordinates of box and artificially segmented space. This procedure creates the vehicle distribution map and completes inference with MLRM.

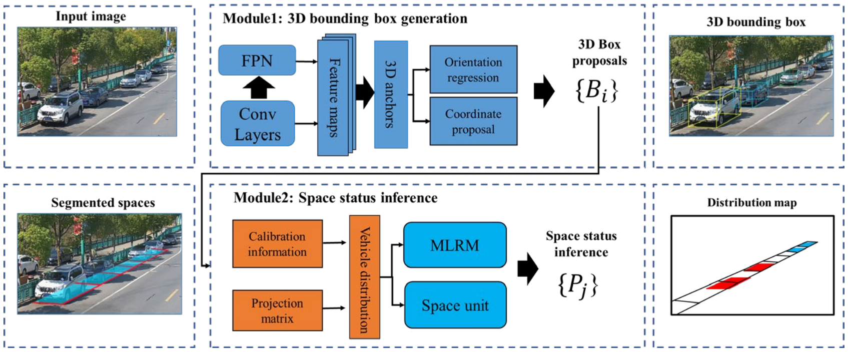


**Figure 1:** The pipeline of the proposed framework. The entire system captures monocular image and outputs space status. Given a single image, the 3D bounding box generation network regresses a set 3D box. Then, the parking spaces are segmented manually and the distribution map is reconstructed, and three adjacent parking spaces are combined into one unit to infer status of the unit

### *3.1 Network architecture*

3D bounding box describes the location of vehicle as well as their parking pose in spatial. Instead of exhaustive searching on traditional object proposal network, an end-to-end network based on Darknet-53 [Redmon and Farhadi (2018)] backbone is designed. In experiments, for a feature map with size of *N*×*N*, Darknet predicts 2D boxes at three scales so the tensor is *N*×*N*×[(4+1+B)×3] for the 4 bounding box offsets, 1 objectness prediction, B class predictions and 3 sampling anchors. In our method, the cascaded net predicts 5 candidates box according to five different 3D anchors, so the tensor is extended to *N*×*N*×[(4+1+1)×5] for the 4 3D box representation (will be elaborated next), 1 class predictions for vehicle.

The whole cascaded network architecture is thus a kind of fully convolutional network (FCN) [Shelhamer, Long and Darrell (2017)], as depicted in Fig. 2. The advantages of this approach is that it provides an opportunity to localize diverse scale vehicles more accurately by relocating the box on higher resolution map, as an end-to-end trainable approach it is also extremely faster than other two-stage approaches due to feature extraction and bounding box prediction are accomplished by one net. All the scale of input images is normalized as 448×448. The feed-forward convolutional net contains five different types of residual blocks and the last three blocks generate various size feature maps, which is transferred to the FPN. Reaping huge fruits from multi-scale feature map fusion of FPN, the network completes prediction on the feature maps of three different scales. For the prediction stage in the FPN with a size *N*×*N* feature map, it generates a confidence score of vehicle and shape offset relative to the anchor box coordinates. The loss function of the net is a sum of bounding box regression loss and corresponding confidence loss:

$$\begin{aligned} L(\{p_i\},\{t_i\}) &= L_{cls}(\{p_i\},\{p_i^*\}) + L_{reg}(\{t_i\},\{\mathrm{t}_i^*\}) \\ &= -\log p_{i_{p_i^*}} + \sum_i \mathrm{smooth}_L(t_i - \mathrm{t}_i^*) \end{aligned} \tag{1}$$

where

$$\text{smooth}_L(x)=\begin{cases}0.5x^2 & \text{if } |x|<1\\ |x|\text{-}0.5 & \text{otherwise}\end{cases} \tag{2}$$

where $i$ is the index of 3D anchor in a mini-batch, $p_i$ and $t_i$ are the predicted confidence of the $i$th anchor being a vehicle and the predicted coordinates of the vehicle. $p_i^*$ and $t_i^*$ are the associated ground truth class label and coordinates of the $i$th anchor.

In a feature map with size of *N*×*N*, each grid point will generate 5 proposal boxes (5 different anchors), such a 26*26 feature map will generate approximately 30 thousands boxes. Obviously, most boxes are negative samples, so it is necessary to add some restrictions to decrease the computational complexity of the net. Particularly, the ground plane estimated from monocular camera is utilized by placing 3D candidate boxes on it. Vehicle candidates are scored on the plane ground by extracting adjacency constraint, contour features and location priors. After this procedure, there are about 5000 proposal boxes left in one feature map.

When it comes to the way of 3D box representation, an inspiring work [Chen, Kundu, Zhu et al. (2018)] provides an outstanding method to represent each object proposal *B*, which is parametrized by a tuple (*x, y, z, θ, t*), where (*x, y, z*) is the box center, *θ* denotes the azimuth angle and *t* represent which vehicle template the vehicle belongs to.

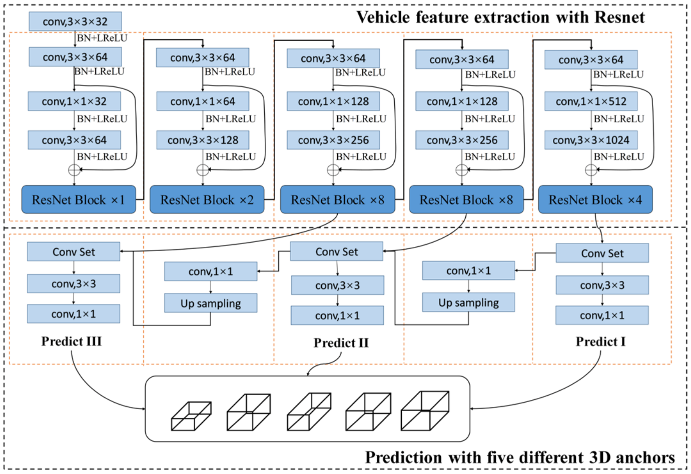


**Figure 2:** 3D bounding box generation network. Composed of two backbone of ResNet and FPN, this network predicts the offsets to anchor boxes of different scales, confidence and aspect ratios. In addition, five 3D anchors clustered from the ground truth of vehicles in the dataset are utilized as the reference to generate candidate box

### *3.2 World scene reconstruction*

With prior knowledge of real-world scene, it is possible to obtain a high-precision bounding box. Instead of using stereo information, a known ground plane assumption is considered. That, through the requirement that all bounding boxes lie on the ground plane, aerial bird-view map of spaces is able to be constructed with camera inverse-projection matrix provided.

Fortunately, the projection matrix is easy to deduce since the input is a monocular image and the angle between ground and image plane is known. According to the description of Hu et al. [Hu, Wu and Wang (2007)], camera parameters are estimated from just one view of five corresponding points based on geometric analysis. Then, the 3D world scene model needs to be reconstructed and each parking space is segmented artificially in real world scene subsequently. Camera projection matrix is derived from the combination of the intrinsic and extrinsic parameters finally.

The central point of the camera is defined as the origin of the 3D coordinate system, hence the image projection matrix $\mathbf{P} = \mathbf{KR}[\mathbf{I} \mid \mathbf{-C}]$ is achievable, where $\mathbf{C}$ is the camera positon in the 3D world defined as $3 \times 1$ vectors, where $\mathbf{I}$ is the identity matrix and $\mathbf{R}$ is rotation, and $\mathbf{K}$ is calibration. A vertex $\mathbf{X}$ is assumed to lie on the ground and it is projected into the camera. $\lambda\mathbf{x} = \mathbf{KR}\left(\mathbf{X} - \mathrm{X}\right) = [\mathbf{KR} \mid -\mathbf{KPX}]\begin{bmatrix}\mathbf{X} \\ \mathbf{1}\end{bmatrix}$ (3)

where $\lambda \in R$. Further information is essential to determine the exact position of vertex $\mathbf{X}$. Considering the inverse projection, it could be written as:

$$\lambda\mathbf{x} = \mathbf{KRX} - \mathbf{KRC} \tag{4}$$

$$\mathbf{KRC} + \lambda\mathbf{x} = \mathbf{KRX} \tag{5}$$

$$\mathbf{C} + \lambda\left(\mathbf{KR}\right)^{-1}\mathbf{x} = \mathbf{X} \tag{6}$$

Then the camera center can be computed from the image projection $\mathbf{P} = \mathbf{KR}[\mathbf{I} \mid \mathbf{-C}]$.

$$\mathbf{C} = -\left(\mathbf{KR}\right)^{-1}\ \mathbf{P}_4 \tag{7}$$

where $\mathbf{P}_4$ is the fourth column of $\mathbf{P}$. With the fore-mentioned ground plane equation, the bottom side of 3D box can be constructed. From the inverse projection of image point $\mathbf{x}$ the ray $\mathbf{I}_\mathrm{X} = \mathbf{C} + (\mathbf{KR})^{-1}\ \mathbf{x}$ is obtained. As mentioned before that $\mathbf{X}$ lies in the ground, it has to hold $\mathbf{n} \cdot \mathbf{X} + d = 0$. Plugging in for $\mathbf{X}$ and solving for $\lambda$ gives us

$$\mathbf{n} \cdot \left(\mathbf{C} + \boldsymbol{\lambda}\mathbf{I}_\mathrm{X}\ \right) + d = \mathbf{0} \tag{8}$$

$$\mathbf{n} \cdot \mathbf{C} + \lambda\mathbf{n} \cdot \mathbf{I}_\mathrm{X} + d = \mathbf{0} \tag{9}$$

$$\lambda = \frac{\mathbf{n} \cdot \mathbf{C} + d}{\mathbf{n} \cdot \mathbf{I}_\mathrm{X}} \tag{10}$$

Then the four bottom coordinates of bounding box could be constructed (only two points are calculated actually because the rest two points may be inferred from others).

### *3.3 Parking space status inference and non-standard parking recognition*

Given output of the second module, the space status can be preliminarily inferred based on where the center of bounding box falls in. However, this approach is not able to guarantee a high recall due to inter-object occlusions and non-standard parking behavior. As shown in Fig. 3, it is obvious that neighboring spaces are correlative because of partial overlap always happens. Simply determining the parking space status according to the relationship between vehicle center and parking space boundary may cause ignorance of non-standard parking, e.g., pressing parking line and exceeding the parking space. As shown in Fig. 3, the space is responsible for the corresponding vehicle if the center of vehicle's 3D bounding box falls into the space. Even though the status of single space may be incorrect, more robust feature can be achieved by combining detection result of three adjacent parking spaces. This model may infer the status appropriately but if the influence of adjacent space prior to the network is taken into account, it is easier to predict good inference.

To model the inference process, three adjacent parking spaces are combined into one unit and an MLRM is trained, where each space is defined into two states: parked and no-parked, which are denoted as P and N. Note that space is regard as a state of P if any part of one vehicle falls in. According to the description, there are two output nodes to indicate the possible status hypothesis $h$ of one space, that is $h \in \{h_i\}_{i=1\text{-}2}$. When it comes to three adjacent spaces, there is 8 output nodes to express one unit and these hypothesis is defined as $H \in \{H_j\}_{j=1\text{-}8}$, where $H_j$ denotes the $i$th status in the set {(P,P,P), (P,P,N), (P,N,P), (P,N,N), (N,P,N), (N,N,P), (N,N,N), (N,P,P)}. The softmax function is used as the output function and $M_f = (P_0, P_1, \cdots, P_M)$ is defined to represent the collection of all the features. Hence, $P = (H = H_j \mid M_f)$, the probability of the status hypothesis $H_j$ is given the input $M_f$, could be calculated by

$$P = (H_j \mid M_f) = P(H_j \mid P_0, P_1, \cdots, P_M) = \frac{\exp(\sum_{m=0}^{M} w_{m,j} \cdot P_m)}{\sum_{l=1}^{8} \exp(\sum_{m=0}^{M} w_{m,j} \cdot P_m)} \tag{11}$$

The weight set $\{w_{m,j}\}_{m=0-M, j=1-8}$ , is trained by the standard back-propagation learning process.

Non-standard parking usually brings numerous traffic safety hazards, especially in the scene of on-street parking. As illustrated in Fig. 4, an intuitive way could be provided to observe vehicle pose by projecting coordinates of 3D box. In order to cope with the location deviation of the vehicle box and simultaneously decrease the false detection rate, the actual space boundary is extended in a certain proportion (1.2 times longitudinally and 1.05 times transversely compared to the original scale) in the vehicle distribution map. Space is extended less in transverse direction for two reasons: (I) the vehicle is limited by the fence on the roadside; (II) it is more dangerous to go beyond the parking space transversely and occupy the road, there is no such concern in the longitudinal direction yet. When the vehicle exceeds the (region of interest) ROI area, it will be labeled as non-standard.

The recognition of non-standard parking may be interfered if vehicle is parking or leaving the space. Even if this situation rarely happened, the system should still distinguish it. In order to solve the misdetection caused by this situation, the same space will be matched with two photos captured in two adjacent periods. If one vehicle is labeled in the previous period, the framework will compare the position of this vehicle in the next cycle. Vehicles will be regarded as non-standard pose when detected in two periods.

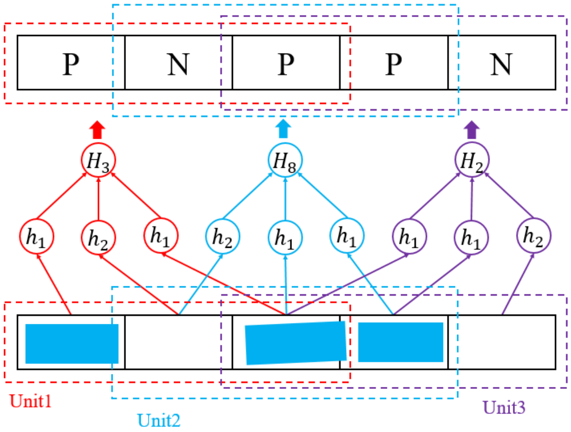


**Figure 3:** MLRM for parking space status inference. Three adjacent parking spaces are combined and then transferred to the MLRM. The final output of $H_j$ determines the space status, where parked is denoted as P and vacant is denoted as N respectively. Note that each space may have different output nodes according to different units. In general, these output nodes should be the same, but if there are different results, the most distributed output node will be retained as the status of the parking space

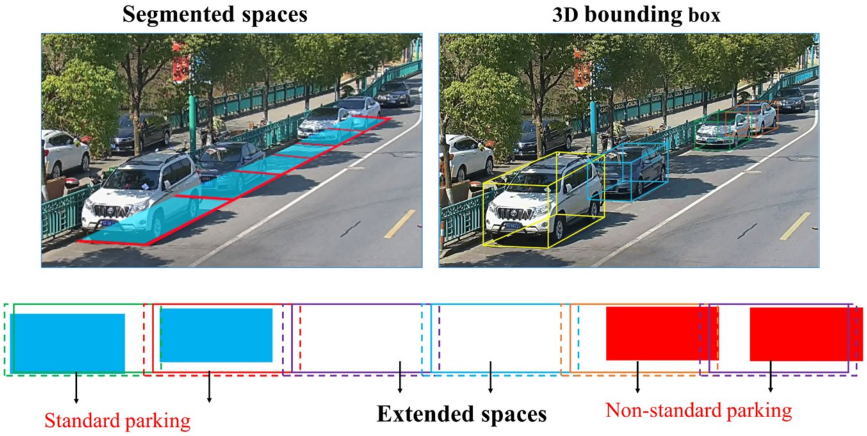


**Figure 4:** Parking space status inference and non-standard parking behavior recognition. The status of space depends on the output of MLRM. The scale of each space is extended 1.2 times longitudinally and 1.05 times transversely (dotted line outside the parking space). By constructing an aerial bird-view map of parking spaces, non-standard parking behavior is easy to be deduced

## 4 Experiments

Three experiments are conducted in this paper. The first experiment evaluated the performance of 3D box generation and compared with other state-of-the-art approaches. The network was trained on both KITTI and our dataset. The strategy of joint training allows the network to achieve high average precision and wide range of application. The second one is an assessment of non-standard parking detection and the final one examined the quality of parking space status inference on our data set.

### *4.1 Datasets*

The KITTI detection benchmark [Urtasun, Lenz and Geiger (2012)] contains 7481 training and 7518 test images, including three classes: Car, Pedestrian, and Cyclist. These images are convenient to initialize our network, however, experiments on the test set in Shanghai show that the performance of our system is not exceptional enough if only training on KITTI benchmark, since the application scenario is totally different. Hence collecting a new dataset is essential.

The original data were collected from the spherical camera mounted on roadside. Totally, 12 cameras are equipped and deployed at the same height. Detection often requires fine-grained visual information so the resolution is reached 1920*1080 pixels. Each camera rotates 360 degrees each minute and captures four images in a period. Obviously, there is no need to store each frame of the surveillance camera because parking cars will not change typically in several minutes. In order to get different types of pictures and reduce the amount of data, a group of photos was down-sampled every 20 minutes. Then, a dataset contains 8600 pictures was collected, among them 5,672 images were captured in day and 2928 images at night. All the images were labeled as the standard of KITTI and the occluded part of vehicle still need to be labeled.

### *4.2 Performance of 3D bounding box regression*

To demonstrate the robustness of 3D box generation net, the cascaded net was evaluated on the KITTI 3D vehicle detection benchmark and our dataset. The KITTI test set has not labeled cars, 2200 images were split from the training set as a validation set, containing 6450 labeled cars. The network was trained on the remaining images and then evaluated on the validation set.

3D bounding box regression was assessed using two metrics as in Chen et al. [Chen, Kundu, Zhu et al. (2017)]: Average Precision (AP 3D), using 3D bounding box overlap measure, and Average Localization Precision (ALP). Similar to calculating the IoU between 2D box and ground truth, AL3D calculates the IoU between ground rectangle of 3D box and ground truth counterpart. The ALP threshold is set as 0.7, means that detected 3D box location is correct if its overlap value is larger than the ALP threshold. The calculation of the ALP is computing the deviation of the predicted 3D location from the ground truth 3D location. Note that better performance of ALP is obviously more paramount than higher AP because the accurate calculation of the vehicle deviation is essential to eliminate detection error of space status.

The numerical comparison result of the two metrics is shown in Tab. 1 and Tab. 2. Obviously our method is capable of outperform most comparable monocular only method (i.e., AVOD [Ku, Mozifian, Lee et al. (2017)]) by a considerable margin on both metrics. The performance of our method on AP3D and ALP is very close to Mono3D [Chen, Kundu, Zhang et al. (2016)], the most accurate monocular approach across the board in 3D IOU. Note that the comparison of our method and 3DOP is unfair as the proposed method just use monocular image without additional depth information. Nevertheless, thanks to the computing power of the cascaded net and the computational advantages of monocular images, our approach has achieved great advantages in terms of speed. As is known to all, single-stage method such as SSD always suffers from the inherited problem that struggles to get the boxes perfectly aligned with the object. An important factor is the anchors generated from dimension clusters is not able to match all object geometry well, whereas, Fig. 5 shows that the framework optimizes this problem well by substituting clustered anchors with 3D anchors specifically clustered for specific object. This is a promising achievement, because it suggests that performance will be significantly improved with more reliable prediction template.

In Tab. 1, the computing time of 3D box generation frameworks is summarized, 3DOP [Chen, Kundu, Zhu et al. (2018)] takes about 1.2 s, MV3D takes 0.36 s on average and PointNet takes about 0.17 s to perform 3D bounding box regression on TITAN X. The proposed network with Darknet-53 takes in total 0.1 s on KITTI benchmark. Compared with previous works, the proposed approach exceeds in accuracy and advances in computing efficiency. Fig. 5 reports the detection performance of MV3D and the proposed network. It is robust enough to handle scale diversity, exhibiting the advantages in 3D bounding box prediction.

**Table 1:** 3D box regression performance on KITTI: Average Precision (AP 3D) and Average Localization Precision (ALP). Paper of methods: SubCNN [Xiang, Choi, Lin et al. (2017)]; 3DVP [Xiang, Choi, Lin et al. (2015)]; 3DOP [Chen, Kundu, Zhu et al. (2018)]; Mono3D [Chen, Kundu, Zhang et al. (2016)]; MV3D [Chen, Ma, Wan et al. (2017)]; VoxNet [Maturana and Scherer (2015)]; PointNet [Qi, Liu, Wu et al. (2017)]; AVOD [Ku, Mozifian, Lee et al. (2017)]; RoarNet [Shin, Kwon and Tomizuka (2018)]

| Proposals | Runtime | Type | KITTI | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Easy | | Moderate | | Hard | |
| | | | AP 3D | ALP (<1 m) | AP 3D | ALP (<1 m) | AP 3D | ALP (<1 m) |
| SubCNN | - | Mono | - | 39.28 | - | 31.04 | - | 25.96 |
| 3DVP | - | Voxel | 80.48 | 45.61 | 68.05 | 34.28 | 57.20 | 27.72 |
| 3DOP | 1.2 | Stereo | 77.50 | 64.89 | 56.79 | 47.34 | 50.84 | 42.20 |
| 3DOP | 1.2 | Hybrid | 89.49 | 82.16 | 81.21 | 75.44 | 74.32 | 69.27 |
| Mono3D | - | Mono | 86.54 | 79.31 | 72.45 | 70.48 | 45.68 | 54.67 |
| MV3D | 0.36 | Hybrid | 71.29 | 72.14 | 62.68 | 66.48 | 56.56 | 46.92 |
| VoxNet | 0.23 | | 71.09 | - | 62.35 | - | 55.12 | - |
| PointNet | 0.17 | | 81.20 | - | 70.39 | - | 62.19 | - |
| AVOD | 0.1 | Mono | 81.94 | - | 71.88 | - | 66.38 | - |
| RoarNet | 0.1 | Lidar | 83.71 | 78.86 | 73.04 | 69.84 | 59.61 | 65.14 |
| **Ours** | **0.11** | **Mono** | **87.26** | **80.75** | **73.46** | **73.45** | **54.31** | **66.06** |

**Table 2:** 3D box regression performance on our data set: average precision (AP 3D) and average localization precision (ALP)

| Method | Runtime | Type | AP 3D | ALP (<1 m) |
|---|---|---|---|---|
| RoarNet | 0.1 | Lidar | 87.89 | 79.56 |
| **Ours** | **0.11** | **Mono** | **88..53** | **82.74** |

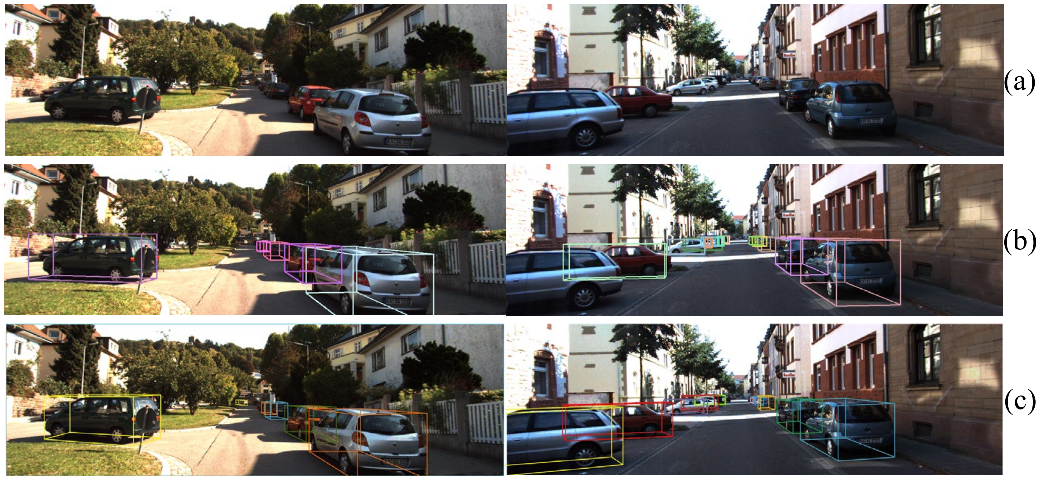


**Figure** 5**:** Comparative performance of MV3D and our method on 3D box generation. (a) The original images; (b)MV3D; (c) The proposed method. Notice that MV3D manages small vehicles well since the application of point cloud information. Although the proposed network slightly inferior to MV3D is small object, it achieves higher precision of bounding box location

### *4.3 Recognition of non-standard parking behavior*

To assess the detection accuracy of non-standard parking behavior recognition, the ground truth of 2474 parking cars was built manually, with 1862 standard vehicles and 612 non-standard vehicles. As shown in Fig. 7, vehicle pose is visualized by creating the vehicle distribution map of parking lot. For numerical comparison, the result is assessed by using false-negative rate (FNR), false-positive rate (FPR) and relative accuracy (ACC), the definition are follows:

$$\text{FPR}=\frac{\text{number of vacant spaces that are detected as occupied}}{\text{total number of occupied spaces}} \quad (12)$$

$$\text{FNR}=\frac{\text{number of occupied spaces that are detected as vacant}}{\text{total number of occupied spaces}} \quad (13)$$

$$\text{ACC}=\frac{\text{number of correct detection}}{\text{total number of occupied spaces}} \quad (14)$$

Tab. 3 reports the performance of non-standard parking behavior recognition on the testing dataset the proposed system is capable of obtaining 83.41% accuracy during the day and 71.54% at night when one camera monitors 32 parking spaces. Reasonably, the accuracy decreases with the increase of detection distance as depicted in Fig. 6, and since the far distance has a great influence on the accuracy of 3D detection, when the monitoring range is increased, the monitoring of irregular parking becomes increasing difficult.

**Table 3:** Performance of non-standard parking pose recognition

| Environment | Number of tested vehicles | | | Our method performance | | |
|---|---|---|---|---|---|---|
| | Standard | Non-standard | Total | FPR | FNR | ACC |
| Day | 1244 | 210 | 1454 | 7.89 | 8.94 | 83.41 |
| Night | 745 | 149 | 894 | 11.52 | 13.74 | 71.54 |

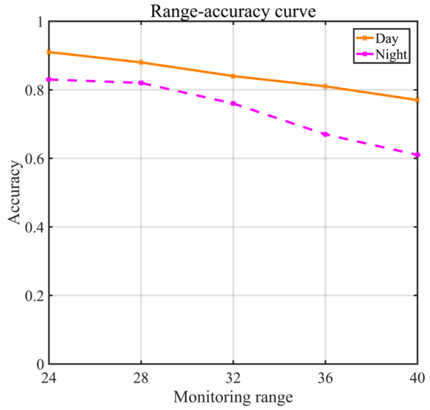


**Figure 6:** Accuracy curve of non-standard parking recognition

### *4.4 Assessment of space status inference*

For evaluating performance on on-street parking space status inference, five methods were compared in this domain: (i) Our method; (ii) Lixia's method [Wang and Jiang (2012)]; (iii) Huang's method [Huang and VU (2017)]; (iv) Wu's method [Wu, Huang, Wang et al. (2007)]; (v) Amato's method [Amato, Carrara, Falchi et al. (2016)]. All of these methods only use vision equipment and were presented in recent years. Experiments were performed using the two data set, one is our dataset collected from another two roads and another one is an open access dataset called CNRPark. Comparison was implemented using manually labeled dataset, on which spaces are labeled with two situations: parked and no-parked.

Fig. 8 shows the comparison result with the Receiver Operating Characteristic (ROC) curves on two datasets and Tab. 4 gives the statistical results on our dataset. It is obvious that the MLRM means a lot to the proposed method especially in complicated environment of the night. Even though the complicated environment makes the status inference challenging, our method is capable of achieving considerable performance. Considering the previous approach only utilize 2D intensity information and ignore the stereo information, it is sensitive to occluded, perspective distortion and camera view.

The reason why the proposed system achieves so high recall is that 3D box provides a large fault tolerance for parking space detection. Even if there is a slice of error in the vehicle localization, it won't affect the output of the entire system. In the long-distance vehicle detection, due to the lack of point cloud and deep information, the 3D box of the vehicle will have a large deviation, but within the specified monitoring range, the entire system is capable of maintaining high accuracy. From the experiments implemented in real road, one camera can maintain 98.50% accuracy while monitoring 36 parking spaces (18 spaces in each side).

**Table 4:** Detection result on our dataset of different environments

| Environment | Proposed method (with the MLRM) | | | Proposed method (without the MLRM) | | |
|---|---|---|---|---|---|---|
| | FPR | FNR | ACC | FPR | FNR | ACC |
| Night | 3.46 | 1.84 | 96.48 | 5.26 | 2.84 | 92.26 |
| Day | 2.16 | 0.59 | 98.50 | 3.45 | 1.74 | 96.24 |

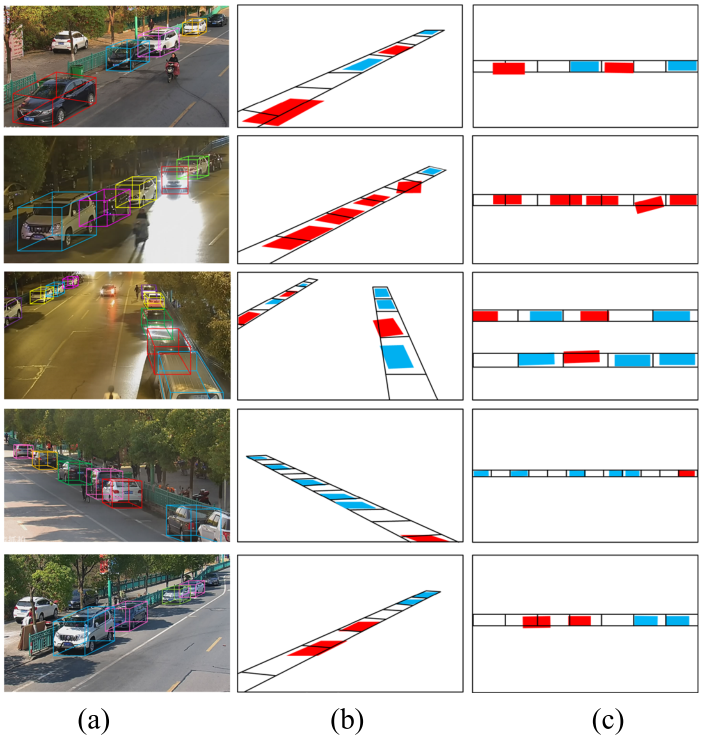
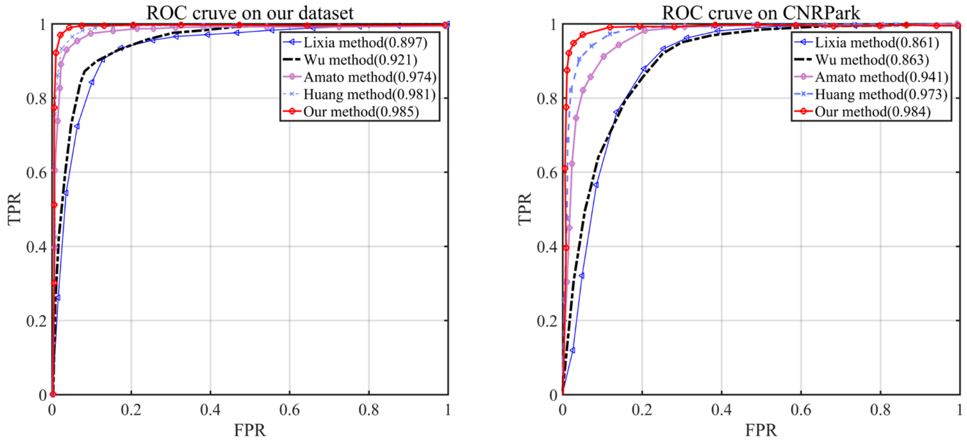


**Figure 7:** Visualization of parking space status inference and vehicle pose estimation. (a) 3D bounding box; (b) Space segmentation; (c) vehicle distribution map. Through the generation of high-quality 3D bounding box, the exact position of vehicles can be located. Then vehicles will be labeled as non-standard parking behavior and marked red if exceeding the ROI area



(a) (b)

**Figure 8:** Comparison of five different approaches. (a) ROC curve on our dataset; (b) ROC curve on CNRPark

## 5 Conclusion

In this paper, a novel on-street parking space detection system, based on 3D bounding box regression and 3D scene model reconstruction, is proposed to handle common challenges that come across in roadside parking space, especially inter-object occlusion. For vehicle 3D bounding box regression, an effective solution is developed that optimizes 2D anchor of Darknet to 3D anchor. The net is jointly trained on two datasets to conquer localization deviation. Then, high-quality 3D box is obtained and three adjacent spaces are combined into one unit on the aerial bird-view map to determine its status. Furthermore, the proposed method calculates the IoU between space edge and ground coordinates of 3D box to perform recognition of non-standard parking behavior, a dangerous pose in roadside that may cause accidents.

Several experiments in real environments were performed with this system demonstrate their functionality compliance and relatively good performance, comparable to that of the most recent works in the field. As a result, the proposed approach is worth to be widely used due to its efficiency, convenience, and capacity to relieve traffic congestion and save fuel.

**Acknowledgement:** 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) and in part by the Shanghai University of Engineering Science Innovation Fund for Graduate Students (No. 18KY0613).

## Disclosures

The author claims that there is no relevant financial interests and other potential conflicts of interest in the manuscript.

## References

**Almeida, P.; Oliveira, L. S.; Silva, E.; Britto, A.; Koerich, A.** (2013): Parking space detection using textural descriptors. *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3603-3608.

**Amato, G.; Carrara, F.; Falchi, F.; Gennaro, C.; Meghini, C. et al.** (2016): Deep learning for decentralized parking lot occupancy detection. *Expert Systems with Applications*, vol. 72, pp. 327-334.

**Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teulière, C.; Chateau, T.** (2017): Deep MANTA: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1827-1836.

**Chen, X. Z.; Kundu, K.; Zhang, Z. Y.; Ma, H. M.; Fidler, S. J. et al.** (2016): Monocular 3D object detection for autonomous driving. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2147-2156.

**Chen, X. Z.; Kundu, K.; Zhu, Y. K.; Ma, H. M.; Fidler, S. et al.** (2017): 3D object proposals using stereo imagery for accurate object class detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 5, pp. 1259-1272.

**Chen, X. Z.; Ma, H. M.; Wan, J.; Li, B.; Xia, T.** (2017): Multi-view 3D object detection network for autonomous driving. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6526-6534.

**He, K.; Zhang, X. Y.; Ren, S. Q.; Sun, J.** (2016): Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

**Huang, C. C.; Tai, Y. S.; Wang, S. J.** (2013): Vacant parking space detection based on plane-based bayesian hierarchical framework. *IEEE Transactions on Circuits & Systems for Video Technology*, no. 9, pp. 1598-1610.

**Huang, C. C.; VU, H. T.** (2017): Vacant parking space detection based on a multi-layer inference framework. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 9, pp. 2041-2054.

**Hu, Z. Y.; Wu, F. C.; Wang, L.** (2007): Multi-camera calibration with one-dimensional object under general motions. *11th IEEE International Conference on Computer Vision*, pp. 1-7.

**Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.** (2017): Joint 3D proposal generation and object detection from view aggregation. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1-8.

**Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B. et al.** (2017): Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936-944.

**Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. et al.** (2015): SSD: single shot multibox detector. *European Conference on Computer Vision*, pp. 21-37.

**Long, J.; Shelhamer, E.; Darrell, T.; Darrell, T.** (2014): Fully convolutional networks for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 39, no. 4, pp. 640-651.

**Masaki, I.** (1998): Machine-vision systems for intelligent transportation systems. *IEEE Intelligent Systems*, vol. 13, no. 6, pp. 24-31.

**Maturana, D.; Scherer, S.** (2015): VoxNet: a 3D convolutional neural network for real-time object recognition. *IEEE International Conference on Intelligent Robots and Systems*.

**Mousavian, A.; Anguelov, D.; Flynn, J.** (2017): 3D bounding box estimation using deep learning and geometry. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5632-5640.

**Pazos, N.; Müller, M.; Favre-Bulle, M.; Brandt-Dit-Grieurin, K.; Hüsser, O. et al**. (2016): Dynamic street-parking optimization. *IEEE 30th International Conference on Advanced Information Networking and Applications*, pp. 1020-1026.

**Qi, C. R.; Liu, W.; Wu, C.; Su, H.; Guibas, L. J.** (2017): Frustum pointnets for 3D object detection from RGB-D data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

**Redmon, J.; Farhadi, A.** (2018): YOLOv3: An incremental improvement. arXiv:1804.02767.

**Ren, S.; He, K.; Girshick, R.; Sun, J.** (2015): Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149.

**Roddick, T.; Kendall, A.; Cipolla, R.** (2018): Orthographic feature transform for monocular 3D object detection. arXiv:1811.08188.

**Sevillano, X.; Màrmol, E.; Fernandez-Arguedas, V.** (2014): Towards smart traffic management systems: vacant on-street parking spot detection based on video analytic. *17th International Conference on Information Fusion*.

**Shin, K.; Kwon, Y.P.; Tomizuka, M.** (2018): RoarNet: a robust 3D object detection based on region approximation refinement. arXiv:1811.03818.

**Urtasun, R.; Lenz, P.; Geiger, A.** (2012): Are we ready for autonomous driving? The KITTI vision benchmark suite. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361.

**Wang, L.; Jiang, D.** (2012): A method of parking space detection based on image segmentation and LBP. *4th International Conference on Multimedia and Security*, pp. 229-232.

**Wu, Q.; Huang, C. C.; Wang, S. Y.; Chiu, W. Y.; Chen, T. H.** (2007): Robust parking space detection considering inter-space correlation. *IEEE International Conference on Multimedia and Expo*, pp. 659-662.

**Xiang, Y.; Choi, W.; Lin, Y. Q.; Savarese, S.** (2015): Data-driven 3D voxel patterns for object category recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1903-1911.

**Xiang, Y.; Choi, W.; Lin, Y. Q.; Savarese, S.** (2017): Subcategory-aware convolutional neural networks for object proposals and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 924-933.

**Yang, C. F.; Ju, Y. H.; Hsieh, C. Y.; Lin, C. Y.; Tsai, M. H. et al.** (2017): iParking-a real-time parking space monitoring and guiding system. *Vehicular Communications*, vol. 9, pp. 301-305.