

Leveraging Heteroscedastic Aleatoric Uncertainties for Robust Real-Time LiDAR 3D Object Detection

Di Feng¹, Lars Rosenbaum¹, Fabian Timm¹, Klaus Dietmayer²

Abstract—We present a robust real-time LiDAR 3D object detector that leverages heteroscedastic aleatoric uncertainties to significantly improve its detection performance. A multi-loss function is designed to incorporate uncertainty estimations predicted by auxiliary output layers. Using our proposed method, the network ignores to train from noisy samples, and focuses more on informative ones. We validate our method on the KITTI object detection benchmark. Our method surpasses the baseline method which does not explicitly estimate uncertainties by up to nearly 9% in terms of Average Precision (AP). It also produces state-of-the-art results compared to other methods while running with an inference time of only 72 ms. In addition, we conduct extensive experiments to understand how aleatoric uncertainties behave. Extracting aleatoric uncertainties brings almost no additional computation cost during the deployment, making our method highly desirable for autonomous driving applications.

I. INTRODUCTION

A robust and accurate object detection system using on-board sensors (e.g. camera, LiDAR, Radar) is crucial for the road scene understanding of autonomous driving. Among different sensors, LiDAR can provide us with accurate depth information, and is robust under different illumination conditions such as daytime and nighttime. These properties make LiDAR indispensable for safe autonomous driving. The recent Uber’s autonomous driving fatal tragedy could have been avoided, if the LiDAR perception system had robustly detected the pedestrian, or had timely informed the human driver to trigger the emergency braking because it was uncertain with the driving situation [1].

Recently, deep learning approaches have brought significant improvement to the object detection problem [2]. Many methods have been proposed that use LiDAR point clouds [3]–[11] or fuse them with camera images [12]–[19]. However, they only give us deterministic bounding box regression and use softmax scores to represent recognition probability, which does not necessarily represent uncertainties in the network [20]. In other words, they do not provide detection confidence regarding classification and localization. For a robust perception system, we need to explicitly model the network’s uncertainties.

¹ Di Feng, Lars Rosenbaum, Fabian Timm are with Robert Bosch GmbH, Corporate Research, Driver Assistance Systems and Automated Driving, 71272 Renningen, Germany.

² Klaus Dietmayer is with Institute of Measurement, Control and Microtechnology, Ulm University, 89081 Ulm, Germany.

The video to this paper can be found in the attachment.

Towards this goal, in this work we build a probabilistic 2-stage-based object detector from LiDAR point clouds by introducing heteroscedastic aleatoric uncertainties - the uncertainties that represent sensor observation noises and vary with the data input. The method works by adding auxiliary outputs to model the aleatoric uncertainties, and training the network with a robust multi-loss function. In this way, the network learns to focus more on informative training samples and ignore the noisy ones. We call our method **PROD** (Probabilistic Real-time Object Detector). Our contributions can be summarized as follows:

- We model heteroscedastic aleatoric uncertainties in a 3D object detection network using LiDAR point clouds.
- We show that by leveraging aleatoric uncertainties, the network produces state-of-the-art results and significantly increases the average precision up to 9% compared to the baseline method without any uncertainty estimations.
- We systematically study how the aleatoric uncertainties behave. We show that the uncertainties are related with each other and are influenced by multiple factors such as detection distance, occlusion, softmax score, and orientation.

In the sequel, we first summarize related works in Sec. II, and then describe our proposed method in Sec. III in detail. Sec. IV shows the experimental results regarding the improvement of object detection performance by leveraging aleatoric uncertainties and understanding how the uncertainties behave. Sec. V summarizes the work and discusses future research. The video of this work is provided as supplementary material.

II. RELATED WORKS

In the following, we summarize methods for LiDAR-based object detection in autonomous driving and uncertainty quantification in deep neural networks.

A. LiDAR-based Object Detection

Many works process the LiDAR information directly from point clouds [3], [5], [7], [14], [15], [17]. For example, Zhou *et al.* [5] propose a voxel feature encoding layer to handle 3D point clouds. Li [3] employs a 3D fully convolutional neural network for discretized point clouds to predict an objectness map and a 3D bounding box map. Other works project 3D point clouds onto a 2D plane and use 2D convolutional

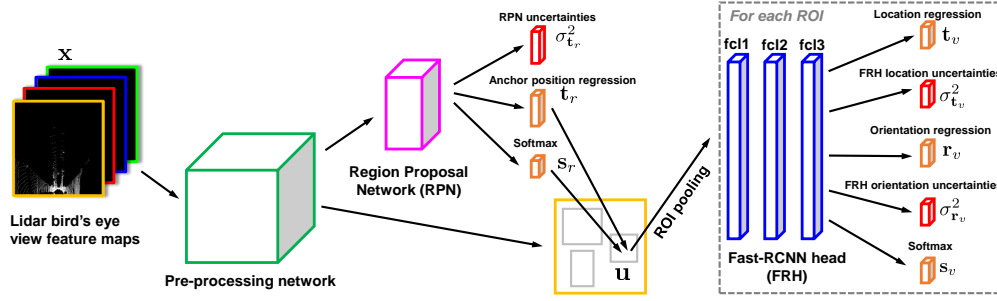


Fig. 1: Our proposed LiDAR 3D object detector (PROD). The network regresses the heteroscedastic aleatoric uncertainties in RPN and FRH.

network to process these LiDAR feature maps. They can be represented by front-view cylindrical images [4], [12], [19], [21], camera-coordinate images [8], [22], [23], or bird's eye view (BEV) map [24]–[27]. Besides LiDAR, an autonomous driving car is usually equipped with other sensors such as cameras or radar sensors. Therefore, it is natural to fuse them for more robust and accurate object detection. For instance, Chen *et al.* [12] use LiDAR BEV map to generate region proposals, and fuse the regional features from LiDAR BEV and front-view maps, as well as camera images for 3D car detection. Qi *et al.* [15] propose to generate 2D object bounding boxes by an image detector, and use the regional point clouds within these bounding boxes for object detection.

B. Uncertainty Quantification in Deep Neural Networks

There are two types of uncertainties we can model in neural networks. *Epistemic* uncertainty shows the model's uncertainty when describing the training dataset. It can be quantified through variational inference [28], sampling technique [20], [29] or ensemble [30], and has been applied to active learning [31], [32], image semantic segmentation [33], [34], camera location estimation [35] or open-dataset object detection problems [36]. *Aleatoric* uncertainty, on the other hand, models the observation noises of input data. It has been modeled by Laplacian distribution or Gaussian distribution for camera image semantics and geometry predictions [37], [38]. Recently, we explicitly model and compare the epistemic and aleatoric uncertainties in an object detector [24]. We have shown that epistemic uncertainties are related to detection accuracy, whereas aleatoric uncertainties are influenced by observation noises.

III. METHODOLOGY

In this section, we present our method which leverages heteroscedastic aleatoric uncertainties for robust LiDAR 3D object detection. We start with illustrating our network architecture, which is followed by a description of how to model the uncertainties. We end the section with a description of our proposed robust multi-loss function.

A. Network Architecture

1) *LiDAR Point Clouds Transformation*: In this work, LiDAR point clouds are encoded in 2D bird's eye view (BEV) feature maps as network inputs. These feature maps include height and density information generated by projecting the 3D point clouds on the 2D grid [13].

2) *Two-stage Object Detector*: The network architecture is shown in Fig. 1. We follow the two-stage object detection network proposed in [39]. The LiDAR BEV feature maps are fed into a pre-processing network based on VGG16 to extract high-level LiDAR features. After the pre-processing layers, a region proposal network (RPN) produces 3D region of interests (ROIs) based on pre-defined 3D anchors at each pixel on the feature map. The RPN consists of two task-specific fully connected layers, each with 256 hidden units. A ROI is parametrized by $[x_r, y_r, z_r, w_r, l_r, h_r, s_r]$, with x_r and y_r indicating the ROI position in the bird's eye view plane, z_r its height, w_r, l_r, h_r its dimensions, and s_r the softmax objectness score. The anchor's dimensions are determined by clustering the training samples with two clusters. Its height is based on the LiDAR's height above the ground plane. Similar to [39], the RPN regresses the region proposals by the normalized offsets denoted as $\mathbf{t}_r = [\Delta x_r, \Delta y_r, \Delta z_r, \Delta w_r, \Delta l_r, \Delta h_r]$ and predicts the objectness score s_r by a softmax layer.

The Fast-RCNN head (FRH) with three fully connected hidden layers (2048 units for each layer) is designed to fine-tune ROIs generated by RPN. It produces multi-task outputs, i.e. the softmax probability s_v , 3D bounding box location and orientation. We encode the location with four corner method introduced in [13]: $\mathbf{t}_v = [\Delta x_v^{(1)} \dots \Delta x_v^{(4)}; \Delta y_v^{(1)} \dots \Delta y_v^{(4)}; \Delta h_v^{(1)}, \Delta h_v^{(2)}]$, with Δx_v and Δy_v being the relative position of a bounding box corner in x, y axes of the LiDAR coordinate frame, and $\Delta h_v^{(1)}, \Delta h_v^{(2)}$ being the heights offsets from the ground plane. We also encode the orientation as $\mathbf{r}_v = [\cos(\theta), \sin(\theta)]$, with θ being the object orientation in BEV. As explained in [13], explicitly modeling the orientation can remedy angle wrapping problem, resulting in a better object detection performance.

B. Modeling Heteroscedastic Aleatoric Uncertainties

As introduced in Sec. I, heteroscedastic aleatoric uncertainties indicate data-dependent observation noises in LiDAR

sensor. For example, a distant or occluded object should yield high aleatoric uncertainties, since there are a few LiDAR points representing them. In our proposed robust LiDAR 3D object detector, we extract aleatoric uncertainties in both RPN and FRH.

Let us denote an input LiDAR BEV feature image as \mathbf{x} , and a region of interest produced by RPN as \mathbf{u} . We also refer to y_r and y_v as the classification labels for RPN and FRH outputs respectively, with $y_r, y_v = 1$ indicating “object” and 0 “background” class. We denote \mathbf{M} as the network weights and $\mathbf{f}^{\mathbf{M}}(\cdot)$ as the noise-free outputs of the object detection network.

We use multivariate Gaussian distributions with diagonal covariance matrices to model the observation likelihood for the anchor position $p(\mathbf{t}_r|\mathbf{f}^{\mathbf{M}}(\mathbf{x}))$, 3D bounding box location $p(\mathbf{t}_v|\mathbf{f}^{\mathbf{M}}(\mathbf{u}))$ and orientation $p(\mathbf{r}_v|\mathbf{f}^{\mathbf{M}}(\mathbf{u}))$:

$$\begin{aligned} p(\mathbf{t}_r|\mathbf{f}^{\mathbf{M}}(\mathbf{x})) &= \mathcal{N}(\mathbf{f}^{\mathbf{M}}(\mathbf{x}), \Sigma_{\mathbf{t}_r}), & \Sigma_{\mathbf{t}_r} &= \text{diag}(\sigma_{\mathbf{t}_r}^2), \\ p(\mathbf{t}_v|\mathbf{f}^{\mathbf{M}}(\mathbf{u})) &= \mathcal{N}(\mathbf{f}^{\mathbf{M}}(\mathbf{u}), \Sigma_{\mathbf{t}_v}), & \Sigma_{\mathbf{t}_v} &= \text{diag}(\sigma_{\mathbf{t}_v}^2), \\ p(\mathbf{r}_v|\mathbf{f}^{\mathbf{M}}(\mathbf{u})) &= \mathcal{N}(\mathbf{f}^{\mathbf{M}}(\mathbf{u}), \Sigma_{\mathbf{r}_v}), & \Sigma_{\mathbf{r}_v} &= \text{diag}(\sigma_{\mathbf{r}_v}^2), \end{aligned} \quad (1)$$

where $\sigma_{\mathbf{t}_r}^2$, $\sigma_{\mathbf{t}_v}^2$ and $\sigma_{\mathbf{r}_v}^2$ refer to the observation noise vectors, and each element of which indicates an aleatoric uncertainty scalar corresponding to an element in \mathbf{t}_r , \mathbf{t}_v and \mathbf{r}_v .

For classification tasks, the observation likelihood $p(y_r|\mathbf{f}^{\mathbf{M}}(\mathbf{x}))$ and $p(y_v|\mathbf{f}^{\mathbf{M}}(\mathbf{u}))$ can be represented by softmax functions:

$$\begin{aligned} p(y_r|\mathbf{f}^{\mathbf{M}}(\mathbf{x})) &= s_r = \text{softmax}(\mathbf{f}^{\mathbf{M}}(\mathbf{x})), \\ p(y_v|\mathbf{f}^{\mathbf{M}}(\mathbf{u})) &= s_v = \text{softmax}(\mathbf{f}^{\mathbf{M}}(\mathbf{u})). \end{aligned} \quad (2)$$

Here, we do not explicitly model the aleatoric classification uncertainties, as they are self-contained from the softmax scores which follow the categorical distribution.

The uncertainty scores $\sigma_{\mathbf{t}_r}^2$, $\sigma_{\mathbf{t}_v}^2$, $\sigma_{\mathbf{r}_v}^2$ can be obtained by adding auxiliary output layers in the object detection network. To increase numerical stability and consider the positivity constraints, we use $\log \sigma^2$ for regression. The regression outputs of RPN that model the aleatoric uncertainties can be formulated as $\hat{\mathbf{t}}_r = [\mathbf{t}_r, \log \sigma_{\mathbf{t}_r}^2]$, and for FRH they are $\hat{\mathbf{t}}_v = [\mathbf{t}_v, \log \sigma_{\mathbf{t}_v}^2]$ and $\hat{\mathbf{r}}_v = [\mathbf{r}_v, \log \sigma_{\mathbf{r}_v}^2]$.

C. Robust Multi-Loss Function

We incorporate aleatoric uncertainties $\sigma_{\mathbf{t}_r}^2$, $\sigma_{\mathbf{t}_v}^2$ and $\sigma_{\mathbf{r}_v}^2$ in a multi-loss function for training our object detector via:

$$\begin{aligned} \mathcal{L}(\mathbf{M}, \sigma_{\mathbf{t}_r}^2, \sigma_{\mathbf{t}_v}^2, \sigma_{\mathbf{r}_v}^2) &= \underbrace{\frac{1}{2\sigma_{\mathbf{t}_r}^2} \mathcal{L}(\mathbf{M}, \mathbf{t}_r) + \log \sigma_{\mathbf{t}_r}^2}_{\text{RPN regression loss}} + \underbrace{\mathcal{L}(\mathbf{M}, y_r)}_{\text{RPN classification loss}} \\ &+ \underbrace{\frac{1}{2\sigma_{\mathbf{t}_v}^2} \mathcal{L}(\mathbf{M}, \mathbf{t}_v) + \log \sigma_{\mathbf{t}_v}^2}_{\text{FRH location loss}} + \underbrace{\mathcal{L}(\mathbf{M}, y_v)}_{\text{FRH classification loss}} \\ &+ \underbrace{\frac{1}{2\sigma_{\mathbf{r}_v}^2} \mathcal{L}(\mathbf{M}, \mathbf{r}_v) + \log \sigma_{\mathbf{r}_v}^2}_{\text{FRH orientation loss}}, \end{aligned} \quad (3)$$

where we use smooth L_1 loss for $\mathcal{L}(\mathbf{M}, \mathbf{t}_r)$, $\mathcal{L}(\mathbf{M}, \mathbf{t}_v)$ and $\mathcal{L}(\mathbf{M}, \mathbf{r}_v)$, and cross entropy loss for $\mathcal{L}(\mathbf{M}, y_r)$ and $\mathcal{L}(\mathbf{M}, y_v)$, similar to [38]. Modeling auxiliary uncertainties in this multi-loss function has two effects. First, an uncertainty score σ^2 can serve as a relative weight to a sub-loss. Thus, optimizing relative weights enables the object detector to balance the contribution of each sub-loss, allowing the network to be trained more easily. Second, aleatoric uncertainties can increase the network robustness against noisy input data. For an input sample with high aleatoric uncertainties, i.e. the sample is noisy, the model decreases the residual regression loss $\mathcal{L}(\mathbf{M}, \mathbf{t}_r), \mathcal{L}(\mathbf{M}, \mathbf{t}_v), \mathcal{L}(\mathbf{M}, \mathbf{r}_v)$ because $\frac{1}{2\sigma^2}$ becomes small. Conversely, the network is encouraged to learn from the informative samples with low aleatoric uncertainty by increasing the residual regression loss with larger $\frac{1}{2\sigma^2}$ term.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

1) *Dataset and Input Representation:* We evaluate the performance of our proposed method on the “Car” category from the KITTI object detection benchmark [40]. We use the LiDAR point cloud within the range $length \times width \times height = [0, 70] \times [-40, 40] \times [0, 2.5]$ meters in the LiDAR coordinate frame. The point clouds are discretized into 5 height slices along the z axis with 0.5 meters resolution and the length and width are discretized with 0.1 meters resolution, similar to [13]. After incorporating a density feature map, the input LiDAR point clouds are represented by the feature maps with size $700 \times 800 \times 6$.

2) *Implementation Details:* The RPN network and the Fast-RCNN head are trained jointly in an end-to-end fashion. The background anchors are ignored for the RPN regression loss. An anchor is assigned to be “Car” class when its Intersection over Union (IoU) with ground truth in the BEV is larger than 0.5 and “background” if it is below 0.3. Anchors that are neither “Car” nor “background” do not contribute to the learning. Besides, we apply Non-Maximum Suppression (NMS) with the threshold 0.8 on the region proposals to reduce redundancy, keeping 1024 proposals during the training process and 300 during the test time. To train the Faster R-CNN head, a ROI is labeled to be positive when its 2D IoU overlap with ground truth in BEV is larger than 0.65 and negative when it is less than 0.55 [13]. We train the network with Adam optimizer. The learning rate is initialized as 10^{-4} and decayed exponentially for every 30000 steps with a decay factor 0.8. We also use Dropout and l_2 regularization to prevent over-fitting. We first train the network without aleatoric uncertainties for 30000 steps and then use the robust multi-loss function (Eq. 3) for another 90000 steps. We find that the network converges faster following this training strategy.

Method	$AP_{3D}(\%)$			$AP_{BEV}(\%)$		
	Easy	Moderate	Hard	Easy	Moderate	Hard
3D FCN [3]	-	-	-	69.94	62.54	55.94
MV3D (BV+FV) [12]	66.77	52.73	51.31	85.82	77.00	68.94
PIXOR [27]	-	-	-	81.70	77.05	72.95
VoxelNet [5]	77.47	65.11	57.73	89.35	79.26	77.39
Baseline	70.46	55.48	54.43	84.94	76.70	69.03
<i>Ours</i>	73.92	62.63	56.77	85.91	76.06	68.75

TABLE I: Comparison with state-of-the-art methods on KITTI *test* set.

Method	$AP_{3D}(\%)$			$AP_{BEV}(\%)$		
	Easy	Moderate	Hard	Easy	Moderate	Hard
VeloFCN [4]	15.20	13.66	15.98	40.14	32.08	30.47
MV3D (BV+FV) [12]	71.19	56.60	55.30	86.18	77.32	76.33
F-PointNet (LiDAR) [15]	69.50	62.30	59.73	-	-	-
PIXOR [27]	-	-	-	86.79	80.75	76.60
VoxelNet [5]	81.97	65.46	62.85	89.60	84.81	78.57
Baseline	71.50	63.71	57.31	86.33	76.44	69.72
<i>Ours</i>	78.81	65.89	65.19	87.03	77.15	76.95

TABLE II: Comparison with state-of-the-art methods on KITTI *val* set.

B. Comparison with State-of-the-art Methods

We first compare the performance of our proposed network PROD (*Ours*) with the baseline method which does not explicitly model aleatoric uncertainties, as well as other state-of-the-art methods (see Tab. I and Tab. II). For a fair comparison, **we only consider LiDAR-only methods**. We use 3D Average Precision for 3D detection (AP_{3D}), and bird’s eye view Average Precision for 3D localization (AP_{BEV}). The AP values are calculated at Intersection Over Union IOU=0.7 threshold introduced in [40] unless mentioned otherwise.

Tab. I shows the performance on KITTI *test* set. The baseline method performs similarly to MV3D(BV+FV) network. **By leveraging aleatoric uncertainties, PROD significantly improves the baseline method up to 7.15% for AP_{3D}** , and produces results comparable to PIXOR [27] and VoxelNet [5]. Tab. II shows the detection performance on KITTI *val* set with the same train-val split introduced in [12]. **By modeling aleatoric uncertainties, PROD improves the baseline method up to nearly 9%**. It also outperforms all other methods in AP_{3D} for moderate and hard settings. The experiments show the effectiveness of our proposed method.

C. Ablation Study

We then conduct an extensive study regarding on where to model the uncertainties, network speed and memory, as well as a qualitative analysis. We use the train-val split introduced in [12] for evaluations.

1) *Where to Model Aleatoric Uncertainties*: In this experiment we study the effectiveness of modeling aleatoric uncertainties in different networks of our LiDAR 3D object detector. In this regard, we train another two detectors that only capture uncertainties either in RPN (Aleatoric RPN) or FRH (Aleatoric FRH) and compare their 3D detection performance with the baseline method (Baseline) and our proposed method (*Ours*) that models uncertainties in both RPN and FRH. Tab. III illustrates the AP values and their improvements for Baseline on KITTI easy, moderate, and hard settings. We find that modeling the aleatoric uncertainties in

Method	$AP_{3D}(\%)$		
	Easy	Moderate	Hard
Baseline	71.50	63.71	57.31
Aleatoric RPN	72.92 (+1.42)	63.84 (+0.13)	58.61 (+1.30)
Aleatoric FRH	81.07 (+9.57)	65.51 (+1.80)	65.09 (+7.78)
<i>Ours</i>	78.81 (+7.31)	65.89 (+2.18)	65.19 (+7.88)

TABLE III: Ablation study - Comparison of 3D car detection performance for easy, moderate, and hard settings on KITTI *val* set.

Method	$AP_{3D}(\%)$			
	0-20 (m)	20-35 (m)	35-50 (m)	50-70 (m)
Baseline	72.42	78.96	57.87	26.17
Aleatoric RPN	80.86 (+8.44)	79.72 (+0.76)	65.44 (+7.57)	30.54 (+4.37)
Aleatoric FRH	79.10 (+6.68)	83.89 (+4.93)	61.98 (+4.11)	29.67 (+3.50)
<i>Ours</i>	80.78 (+8.36)	84.75 (+5.79)	66.81 (+8.94)	34.07 (+7.90)

TABLE IV: Ablation study - Comparison of 3D car detection performance for data within different distances from KITTI *val* set. We use IOU=0.5 threshold in this experiment.

either RPN, FRH or both can improve the detection performance, while modeling the uncertainties in both networks brings the largest performance gain in moderate and hard settings. Furthermore, we evaluate the detection performance on different LiDAR ranges, shown by Tab. IV. Again, our method consistently improves the detection performance compared to Baseline. Modeling the uncertainties in both RPN and FRH shows highest improvement between ranges 20 – 70 meters, indicating that aleatoric uncertainties in both networks can compensate each other. As we will demonstrate in the following section (Sec. IV-D), our proposed network handles cars from easy, moderate, hard, near-range or long-range settings differently. It learns to adapt to noisy data, resulting in improved detection performance.

2) *Runtime and Number of Parameters*: We use the runtime and the number of parameters to evaluate the computational efficiency and memory requirement. Tab. V shows the results of PROD relative to the baseline network. **We only need additional 2 ms and 26112 parameters to predict aleatoric uncertainties during inference, showing the high efficiency of our proposed method.**

Method	Number of parameters	Runtime	
		Data pre-processing	Inference
Baseline	36192171	97 ms	70 ms
<i>Ours</i>	+26112 (+0.07%)	+0 ms	+2 ms (+2.86%)

TABLE V: Comparison of runtime and the number of parameters. The runtime is measured by averaging the predictions over all KITTI *val* set on a TITAN X GPU.

3) *Qualitative Analysis*: Fig. 2 shows some exemplary results. In general, our proposed method can detect cars with higher recall, **especially at long distance (e.g. Fig. 2(b)) or in highly-occluded regions** (e.g. Fig. 2(c)). However, the network also tends to mis-classify objects with car-like shapes, e.g. in Fig. 2(d) the network incorrectly predicts the fences on the bottom-left side as a car. Such failures could be avoided by fusing the image features from vision cameras.

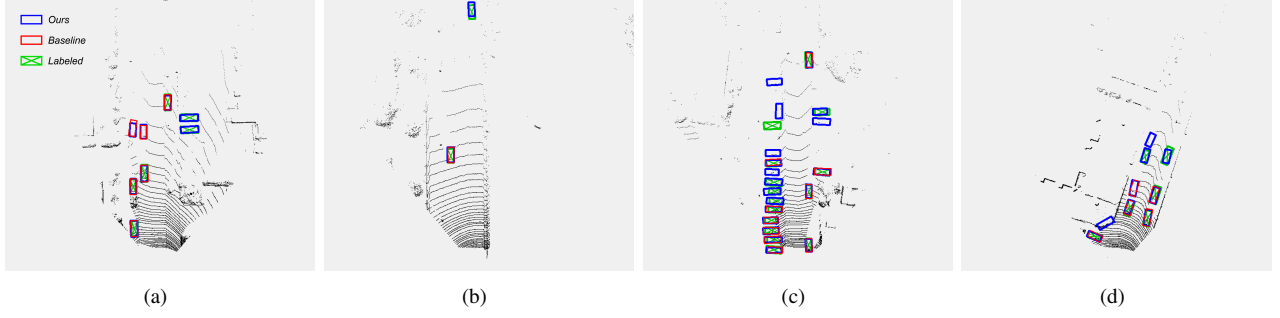


Fig. 2: Some car detection examples on KITTI *val* set within the range $length \times width = [0, 70] \times [-40, 40]$ meters. The LiDAR points which are out of camera front view are not evaluated. The predictions from our proposed network (*Ours*) are visualized in blue, the predictions from baseline method in red, and the labeled samples in green. Better view by magnificence.

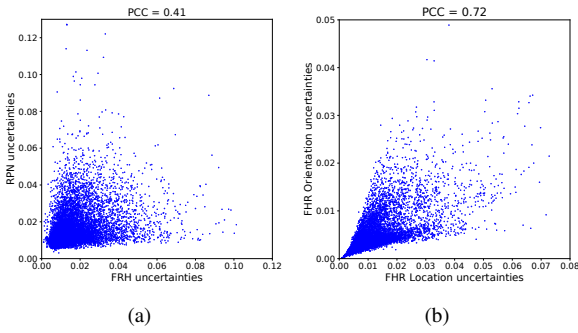


Fig. 3: (a) The distribution of uncertainties in FRH (x axis) and in RPN (y axis). Each scatter represents a prediction. (b) The distribution of uncertainties for FRH location (x axis) and orientation prediction (y axis).

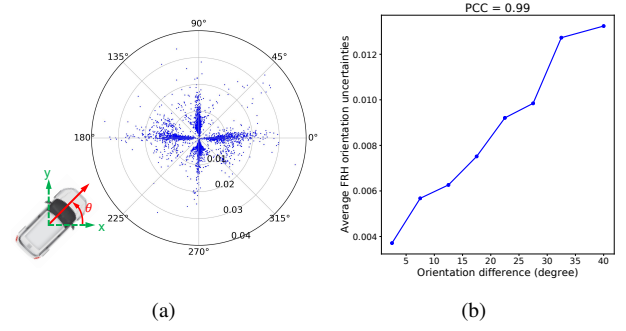


Fig. 4: (a) The prediction distribution of FRH orientation uncertainties w.r.t angle values θ (angular axis). (b) The average orientation uncertainties with orientation difference between predicted angles and the nearest base angles.

D. Understanding Aleatoric Uncertainties

We finally conduct comprehensive experiments to understand how aleatoric uncertainties behave. We use the scalar *Total Variance* (TV) to quantify aleatoric uncertainties introduced in [24]. A large TV score indicates high uncertainty. We also use Pearson Correlation Coefficient (PCC) to measure the linear correlation. We study RPN uncertainties and FRH uncertainties. The RPN uncertainties indicate observation noises when predicting anchor positions, whereas the FRH uncertainties indicate the noises for final bounding box predictions. It consists of FRH location uncertainties for the bounding box regression and FRH orientation uncertainties for heading predictions. We evaluate all predictions with a score larger than 0.5 from the KITTI *val* set unless mentioned otherwise.

1) *Relationship Between Uncertainties*: Fig. 3(a) and Fig. 3(b) show the prediction distribution of FRH uncertainties with RPN uncertainties, as well as FRH location uncertainties with orientation uncertainties, respectively. The uncertainties are highly correlated with each other, indicating that our detector has learned to represent LiDAR observation noises at different sub-networks and prediction outputs.

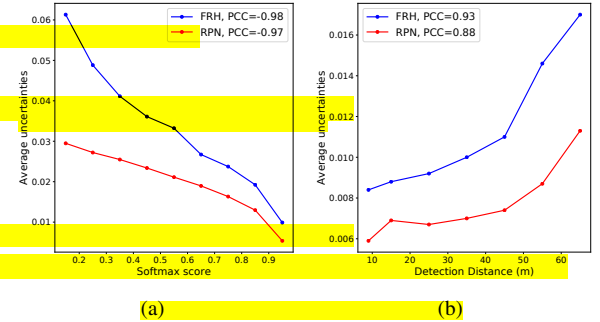


Fig. 5: (a) Average RPN and FRH uncertainties with the increasing softmax scores for anchor and final object classification. (b) Average uncertainties with detection distance.

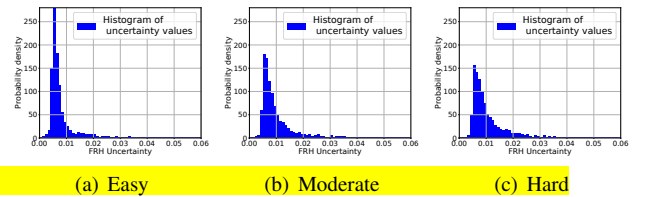


Fig. 6: Histogram of FRH uncertainties for easy, moderate, and hard settings in the KITTI *val* set.

Log(TV): -2.92 Softmax score: 0.87 Distance: 7.5 m Occ: fully visible Orientation: 92.1°	Log(TV): -2.74 Softmax score: 0.99 Distance: 15.7 m Occ: fully visible Orientation: 91.6°	Log(TV): -2.50 Softmax score: 0.99 Distance: 19.5 m Occ: fully visible Orientation: 89.8°	Log(TV): -2.25 Softmax score: 0.99 Distance: 22.4 m Occ: partly occluded Orientation: 269.2°	Log(TV): -2.00 Softmax score: 0.80 Distance: 22.4 m Occ: fully visible Orientation: 272.0°	Log(TV): -1.74 Softmax score: 0.67 Distance: 14.1 m Occ: difficult to see Orientation: 0.8°	Log(TV): -1.50 Softmax score: 0.55 Distance: 22.4 m Occ: partly occluded Orientation: 181.8°	Log(TV): -1.25 Softmax score: 0.55 Distance: 18.6 m Occ: partly occluded Orientation: 89.1°	Log(TV): -0.98 Softmax score: 0.56 Distance: 49.2 m Occ: fully visible Orientation: 269.7°

Fig. 7: Exemplary detections with the total variance of uncertainties being distributed equally from -3 to -1 at log scale ($\log(TV)$). The aleatoric uncertainty is reducing from left to right. The 3D bounding boxes are predicted from PROD and are projected onto camera coordinate for visualization purpose only. We also show the softmax score, object distance, level of occlusion (fully visible, partly occluded and difficult to see) as well as orientation for each detection.

2) *Orientation Uncertainties*: Fig. 4(a) illustrates the prediction distribution of FRH orientation uncertainties (radial axis) w.r.t. angle values θ (angular axis) in the polar coordinate. Most predictions lie at four base angles, i.e. $\theta = 0^\circ, 90^\circ, 180^\circ, 270^\circ$. Fig. 4(b) shows the average orientation uncertainties with orientation difference between predicted angles and the nearest base angles. They are highly correlated with $PCC=0.99$, showing that the network produces high observation noises when predicting car headings that are different from the base angles.

3) *Relationship Between Softmax Score, Detection Distance and Uncertainties*: In Fig. 5(a) we plot the average RPN and FRH uncertainties with the increasing softmax scores for anchor and final object classification. We find a strong negative correlation between them. As introduced by Eq. 2, the softmax scores can be regarded as aleatoric uncertainties for classification. This indicate that our network has learned to adapt the uncertainties in regression tasks (i.e. anchor position, bounding box location and orientation) to that in the classification, i.e. the uncertainties increase as the softmax score reduces. Fig. 5(b) shows that the average RPN and FRH uncertainties become larger as detection distance increases. This is because the LiDAR sensor measures fewer reflections from a distant car, which yields high observation noises.

4) *Uncertainty Distribution for Easy, Moderate, and Hard Settings*: We finally evaluate the FRH uncertainty distribution for easy, moderate, and hard objects, demonstrated in Fig. 6. The uncertainty distributions vary: for easy setting there are more predictions with lower uncertainties, whereas for hard objects which have larger occlusions, truncations, and detection distances, the network estimates higher uncertainties. The result indicates that the network has learned to treat objects from these three settings differently.

5) *Qualitative Observations*: In Fig. 7 we visualize nine exemplary detections whose uncertainties are equally distributed at log scale, ranging from -3 to -1. We observe: (1). The uncertainties are influenced by occlusion, detection distance, orientation as well as softmax score, as discussed above. (2) The network shows higher aleatoric uncertainties if there are fewer points around the car. The results show that our network has captured reasonable LiDAR observation noises.

V. DISCUSSION AND CONCLUSION

We have presented our robust LiDAR 3D object detection network called PROD that leverages heteroscedastic aleatoric uncertainties to significantly improve detection performance. Trained with a multi-loss function which incorporates aleatoric uncertainties, PROD learns to adapt to noisy data and increases the average precision up to 9%, producing state-of-the-art results on KITTI object detection benchmark. Our method only requires to modify the cost function and output layers with only 2 ms additional inference time. This makes our method suitable for real-time autonomous driving applications.

We have qualitatively analyzed how PROD learns to deal with noisy data in Sec. IV-C. The network tends to predict high uncertainties for detections that are highly occluded (Fig. 7), far from the ego-vehicle (Fig. 5(b)), with different orientations from the base angles (Fig. 4(b)) or with low objectness score (Fig. 5(a)). Therefore, our network learns less from noisy samples during the training process by penalizing the multi-loss objective with the $\frac{1}{2\sigma^2}$ terms (Eq. 3). Conversely, the network is encouraged to learn more from the informative training samples with more LiDAR reflections. In this way, its robustness against noisy data is enhanced, resulting in improved detection performance for data from easy, moderate, hard settings (Tab. III) or at different distances (Tab. IV). Note that this effect is because we explicitly model the observation noises rather than an ad-hoc solution.

Compared to *Focal Loss* [41] which incorporates prediction probability in the loss function to tackle the positive-negative sample imbalance problem, our proposed robust multi-loss function works in an opposite way: it depreciates “outliers” training samples with high aleatoric uncertainties, and encourages the network to learn from those with small errors. It is an interesting future work to investigate how a network behaves with Focal Loss and our proposed method using different driving datasets.

In the future, we intend to improve PIXOR [27] and VoxelNet [5] by incorporating aleatoric uncertainties. We also intend to investigate the influence of object reflectance and bounding box encodings on uncertainty estimations with our own test vehicles.

ACKNOWLEDGMENT

We thank Zhongyu Lou his suggestions and inspiring discussions. We also thank Xiao Wei for reading the script.

REFERENCES

- [1] "Preliminary report highway: Hwy18mh010," National Transportation Safety Board (NTSB), Tech. Rep., 05 2018.
- [2] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art," *arXiv preprint arXiv:1704.05519*, 2017.
- [3] B. Li, "3d fully convolutional network for vehicle detection in point cloud," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 1513–1518.
- [4] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," *Robotics: Science and Systems*, 2016.
- [5] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," *arXiv preprint arXiv:1711.06396*, 2017.
- [6] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection," in *Robotics: Science and Systems*, vol. 1, 2015, p. 5.
- [7] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1355–1361.
- [8] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, "Depthcn: vehicle detection using 3d-lidar and convnet," in *IEEE International Conference on Intelligent Transportation Systems*, 2017.
- [9] K. Minemura, H. Liau, A. Monrroy, and S. Kato, "Lmnet: Real-time multiclass object detection on cpu using 3d lidars," *arXiv preprint arXiv:1805.04902*, 2018.
- [10] Y. Zeng, Y. Hu, S. Liu, J. Ye, Y. Han, X. Li, and N. Sun, "Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3434–3440, Oct 2018.
- [11] S. Yu, T. Westfechtel, R. Hamada, K. Ohno, and S. Tadokoro, "Vehicle detection and localization on bird's eye view elevation images using convolutional neural network," in *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, Oct 2017, pp. 102–109.
- [12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [13] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander, "Joint 3d proposal generation and object detection from view aggregation," *arXiv preprint arXiv:1712.02294*, 2017.
- [14] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," *arXiv preprint arXiv:1711.10871*, 2017.
- [15] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," *arXiv preprint arXiv:1711.08488*, 2017.
- [16] X. Du, M. H. Ang Jr, S. Karaman, and D. Rus, "A general pipeline for 3d detection of vehicles," *arXiv preprint arXiv:1803.00387*, 2018.
- [17] X. Du, M. H. Ang, and D. Rus, "Car detection for autonomous vehicle: Lidar and vision fusion approach through deep learning framework," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 749–754.
- [18] D. Matti, H. K. Ekenel, and J.-P. Thiran, "Combining lidar space clustering and convolutional neural networks for pedestrian detection," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.
- [19] A. Pfeuffer and K. Dietmayer, "Optimal sensor data fusion architecture for object detection in adverse weather conditions," in *Proceedings of International Conference on Information Fusion*, 2018, pp. 2592 – 2599.
- [20] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016.
- [21] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," *arXiv preprint arXiv:1710.07368*, 2017.
- [22] T. Kim and J. Ghosh, "Robust detection of non-motorized road users using deep learning on optical and lidar data," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 271–276.
- [23] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing lidar and images for pedestrian detection using convolutional neural networks," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2198–2205.
- [24] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection," *arXiv preprint arXiv:1804.05132*, 2018.
- [25] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast lidar-based road detection using fully convolutional neural networks," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 1019–1024.
- [26] J. Beltran, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. de la Escalera, "Birdnet: a 3d object detection framework from lidar information," *arXiv preprint arXiv:1805.01195*, 2018.
- [27] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [28] G. E. Hinton and D. Van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in *Proceedings of the sixth annual conference on Computational learning theory*. ACM, 1993, pp. 5–13.
- [29] A. Graves, "Practical variational inference for neural networks," in *Advances in neural information processing systems*, 2011, pp. 2348–2356.
- [30] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," in *Advances in neural information processing systems*, 2016, pp. 4026–4034.
- [31] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International Conference on Machine Learning*, 2017, pp. 1183–1192.
- [32] W. H. Beluch, T. Genewein, A. Nrnberger, and J. M. Khler, "The power of ensembles for active learning in image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [33] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*. IEEE, 2016, pp. 680–688.
- [34] A. Kendall, V. Badrinarayanan, , and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.
- [35] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocation," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 4762–4769.
- [36] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.
- [37] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, 2017, pp. 5580–5590.
- [38] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.
- [41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection (best student paper award)," in *International Conference on Computer Vision (ICCV), Venice, Italy, 2017, oral*.