IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 March 2023, accepted 16 April 2023, date of publication 26 April 2023, date of current version 11 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3270447

RESEARCH ARTICLE

Assessing Distribution Shift in Probabilistic Object Detection Under Adverse Weather

MATHEW HILDEBRAND[1], (Member, IEEE), ANDREW BROWN[1], STEPHEN BROWN[1], (Senior Member, IEEE), AND STEVEN L. WASLANDER[2], (Senior Member, IEEE)

[1]The Edward S. Rogers Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada
[2]Institute for Aerospace Studies and Robotics Institute, University of Toronto, Toronto, ON M3H 5T6, Canada

Corresponding author: Steven L. Waslander (steven.waslander@utoronto.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

ABSTRACT Object detection is a safety-critical aspect of autonomous driving, allowing vehicles to identify moving objects in the scene for tracking, prediction and decision making. Current detectors, however, tend to provide point estimates for detected objects, which lack information on the variability of the prediction and how well it fits the model that produced the prediction. Proper uncertainty estimation can be incorporated into traditional object detection pipelines to produce a measure of uncertainty alongside traditional point estimate object predictions. In this work, uncertainty estimates are implemented for LiDAR and camera object detectors using Bayesian theory, and the resulting output distributions are assessed using signal detection theory to generate an uncertainty based classifier that can evaluate its own performance. The classifier can be used to track the ratio of false positive to true positive detections, defined as the *anomalous detections ratio*. Findings from this work indicate that this novel metric is responsive to degraded driving conditions including night time driving and lens obstructions for the RGB camera, while in LiDAR data, the metric is responsive to snowfall and simulated rain conditions. These results are focused on the classification and regression of vehicle objects, making use of the sizeable ground-truth sets for vehicles that are provided in publicly-available autonomous driving data sets.

INDEX TERMS Autonomous driving system, Bayesian neural network, probabilistic object detection, sensor fusion, uncertainty in object detection.

I. INTRODUCTION

The task of object detection is a core component of any autonomous driving system (ADS), feeding downstream processes that track, predict and plan for interactions with objects in the driving environment. Current state-of-the-art models have shown remarkable accuracy in object detection using sensors such as RGB cameras and LiDAR sensors. These systems provide a set of object predictions (recognition) and their location in the environment (localization), which can be used in the decision making process for autonomous vehicles.

A drawback to traditional object detection methods is that the recognition and localization tasks produce point estimates of detected objects, offering an estimated confidence value of the detection (0.0-1.0) alongside a set of bounding box coordinates enclosing the object. However, these outputs do not capture how well the predictions fit the prior evidence acquired in the training process in forming its confidence level and bounding box parameters. As driving conditions, viewpoints and occlusion rates vary extensively, a more detailed prediction of the uncertainty associated with detection outputs is needed and can alert the autonomy system of the presence of degraded sensor performance during operation.

One way to provide a predictive output distribution instead of a point estimate is to employ Bayesian neural networks (BNNs) [30] which learn weight distributions but are too computationally complex for modern detection architectures. Instead, BNNs have been approximated to form Bayesian object detectors in recent works [15], [21], [39], enabling

The associate editor coordinating the review of this manuscript and approving it for publication was Yang Tang.

networks to extract uncertainty alongside their predictions. Uncertainty can be viewed as a measure of how unsure a neural network is about its predictions. The field of probabilistic object detection is rapidly emerging as an approach to expand system robustness in an open world setting. Thus far, uncertainty has been used in the Deep Neural Network (DNN) itself, as a feedback mechanism to increase accuracy of single sensor detectors [15], [39] and in a fusion network feedback loop [10]. A recent survey [11] outlines the main approaches and performs a detailed comparison study on probabilistic object detection methods proposed to date.

In this work, uncertainty estimation is evaluated for LiDAR and RGB camera object detectors, using an approximation of the Bayesian neural network introduced by Kendall and Gal [21]. Evaluation is performed on the Kitti, Waymo, and Canadian Adverse Driving Conditions (CADC) datasets [13], [31], [35]. These datasets allow for uncertainty evaluation under varied weather scenarios. Using a reformulation of signal detection theory in combination with uncertainty estimation, a novel metric called the *anomalous detections ratio* (ADR) is developed, which measures the viability of using detection uncertainty to differentiate between true and false positives at inference time. The responsiveness of the metric in real world and simulated adverse weather conditions is analyzed, to determine if it can be used to assess degraded sensor performance in adverse conditions.

The main contributions of this work include:

- Implementation of a Bayesian object detector architecture for both LiDAR and RGB camera data, capable of estimating uncertainty.
- A comprehensive evaluation of uncertainty estimation techniques for camera and LiDAR object detectors across multiple datasets, degraded via augmentation techniques and naturally differing operating conditions.
- Development and evaluation of a novel metric, the *anomalous detections ratio* (ADR), that can indicate sensor performance degradation, with tests on real adverse conditions such as low-light driving, rain and snow.

More details about many of the contributions in this work can be found in [18]. The remainder of this paper is organized as follows: Section II briefly describes general related work, Section III outlines techniques used in this research to model uncertainty, Section IV presents the framework that we use for image detection, Section V describes the datasets that we use in our experiments, Section VI shows how we apply statistical methods to calculate our results, which are presented in Section VII, Section VIII provides concluding remarks, and Section IX proposes future work.

II. RELATED WORK

Modern object detection relies on convolutional neural networks, in either single-stage [29], [32] or two-stage region proposal architectures [16], [34]. In this work, we develop a two-stage Faster-RCNN [34] baseline detector that can be used with both camera and LiDAR data for probabilistic object detection evaluation under adverse weather conditions. In the second stage, objects are localized with a smooth L1 loss function and classified with a cross entropy function, producing point estimate predictions of bounding boxes. *Probabilistic object detection* was made tractable through the approximations to BNNs introduced in Kendall and Gal's work [21]. It proposed that two types of uncertainty can be extracted using Bayesian deep learning techniques: *aleatoric uncertainty*, which captures noise inherent in the input to a neural network, and *epistemic uncertainty*, which captures noise within the model itself. The authors model aleatoric uncertainty through *loss attenuation*, and epistemic uncertainty through Monte-Carlo *dropout sampling*.

Liu et al. build on Kendall and Gal's work by proposing an ensemble of object detectors to extract epistemic uncertainty, rather than dropout [26]. While this method avoids many of the pitfalls of Monte-Carlo dropout uncertainty estimation, such as incompatibility with batch normalization layers during training, it is computationally expensive. Recent work by Harakeh et al. [15] has implemented both epistemic and aleatoric uncertainty estimation in a single stage image detection network. The authors increase detection accuracy by incorporating uncertainty into a detection clustering stage, similar to and in-place of, non-maximum suppression. This work shows that uncertainty can be used in a feedback loop to increase accuracy.

Similar work by Zhang et al. extracts and evaluates uncertainty using a LiDAR sensor instead of an RGB camera [39]. The authors evaluate uncertainty extensively against metrics such as IoU (*overlap*) and object distance. The authors also claim that simply including uncertainty measurement into the object detector can result in a boost in robustness and accuracy. Further work by Feng et al. [10] leverages uncertainty in a fusion network that combines the RGB camera and LiDAR. The authors use uncertainty estimation to increase robustness to LiDAR noise in the region proposal stage, resulting in a boost in baseline detection accuracy. They also evaluate the effects of temporal misalignment between sensors and how uncertainty can significantly boost detection accuracy in these scenarios. These works present a powerful idea of leveraging uncertainty in a fusion network, but do not extract uncertainty for both sensor modalities and only model aleatoric uncertainty.

In our work, both aleatoric and epistemic uncertainty are estimated for both classification and regression tasks in a two-stage probabilistic object detector. It is designed to be compatible with both LiDAR and RGB camera sensors and is trained on a range of datasets and conditions. Thus, we can evaluate and compare the extracted uncertainty between the two sensors to better understand if it is an informative measure, specifically in adverse weather conditions.

III. UNCERTAINTY ESTIMATION

Bayesian modelling provides a method to estimate uncertainty by transforming traditional point estimate predictions into full probability distributions. A Bayesian modelling

approach treats model parameters as random variables that are described by a "prior" distribution over the weights $p(\theta)$. A posterior distribution of the parameters $p(\theta|\mathcal{D})$ can be obtained from training the machine learning model on data $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$, by multiplying the network output, modelled as the likelihood $P(\mathcal{Y}|\mathcal{X}, \theta)$ by the prior $p(\theta)$ and dividing by the marginal probability $P(\mathcal{Y}|\mathcal{X})$, forming Bayes theorem. However, in practice the marginal probability cannot be evaluated analytically, and the posterior must be approximated with a surrogate distribution $p(\theta|\mathcal{X}, \mathcal{Y}) \approx q^*(\theta)$ using Kullback-Leibler (KL) divergence [21].

Likewise, a predictive posterior $p(y_*|x_*, \mathcal{D})$ is computed by integrating the predictive likelihood of an outcome with the approximated posterior $q^*(\theta)$. This is also approximated in practice, by either using ensembles of ML models or multiple stochastic runs through one ML model by enabling Monte-Carlo dropout sampling at inference.

In the interest of brevity, details about the methods that we use to model epistemic and aleatoric uncertainty, for localization and classification detections, are not included here. A detailed discussion of these methods can be found in [18], [12], and [39].

IV. NETWORK ARCHITECTURE

As illustrated in Fig. 1, a Faster R-CNN network architecture was chosen for both the LiDAR and RGB camera object detectors, with pre-processing to allow a common neural network architecture for both sensor modalities. Resnet-101 [17] and FPN [25] backbones are used with ROI-Align pooling on 2D image and bird's-eye view (BEV) LiDAR data for the 2D and 3D detection tasks, respectively. A baseline implementation from Chen and Gupta was used [7]. Major modifications include the added support for various datasets (Kitti, Waymo, CADC) and support for the LiDAR based 3D object detector. Uncertainty estimation support was added, following the approach in [21].

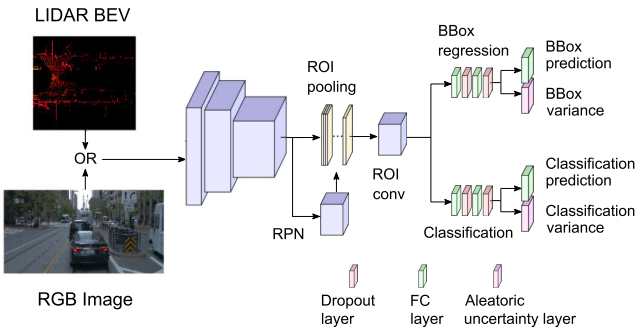


FIGURE 1. Proposed detection framework.

V. EXPERIMENTS

A. DATASETS

In this work, datasets were chosen to act in a complementary way, to test different aspects of the object detectors, as well as provide varying adverse weather conditions.

TABLE 1. AP results for Kitti.

Configuration	Image AP			LiDAR BEV AP			LiDAR 3D AP		
	Easy	Med	Hard	Easy	Med	Hard	Easy	Med	Hard
Base	0.89	0.82	0.71	0.20	0.23	0.18	0.30	0.32	0.30
Base +	0.85	0.83	0.75	0.60	0.58	0.53	0.40	0.42	0.35
Base + +	**0.94**	0.84	0.76	0.79	0.73	0.68	0.59	0.52	0.49
Faster-RCNN	0.83	0.88	0.72	-	-	-	-	-	-
YOLO-V4	0.91	**0.94**	**0.80**	-	-	-	-	-	-
BirdNet	-	-	-	0.54	0.89	0.57	0.27	0.41	0.25
PointPillars	-	-	-	**0.86**	**0.90**	**0.82**	**0.74**	**0.82**	**0.69**

Dataset selection was restricted to those that 1) contain synchronized and calibrated LiDAR and RGB camera data, and 2) include accurate and extensive ground truths for training purposes. We note that in some recent works [27], [28], improvements in object detection have been shown by forcing bounding boxes to be on a ground plane. We have not included such enhancements in our object detectors at this time, for consistency when comparing our results to other successful 3D object detection methods like [4], [34], and [23]. These comparison are provided in Tables 1 and 2, which are described below.

1) KITTI

The Kitti Vision Benchmark Suite [13] is considered a pioneering dataset that popularized the use of synchronized *frames* of data between a GPS, IMU, LiDAR and set of RGB cameras [13]. Kitti contains 7,481 synchronized frames, which each have associated labeling information. In this work, Kitti is used as a baseline test to ensure nominal accuracy of the object detectors used in our experiments.

Mean average precision results from training on the Kitti dataset for each sensor modality are given in Table 1. Ablation is performed by training a baseline detector for both sensors (*Base*), alongside the detector with standard augmentation applied (*Base* +) and finally with uncertainty estimation enabled as well (i.e. loss attenuation and dropout) (*Base* ++). For the RGB (Image) detector, results are shown for our work, Faster-RCNN [34], and YOLO-V4 [4]. Our LiDAR detector results are compared against similar two stage BEV LiDAR object detectors, PointPillars [23] and BirdNet [1].

2) WAYMO

The Waymo dataset contains over 230,000 annotated data frames using five LiDAR sensors, five cameras, a GPS and an IMU [35]. Waymo includes over 1000 scenes of data, each containing approximately 20 seconds of a drive, annotated at 10 Hz. Each scene is labeled for various environmental conditions, such as day, night, twilight, rainy, or sunny.

In this work, the various labeled weather scenarios enable analysis of detector uncertainty in real world conditions. Specifically, the Waymo dataset allows for the analysis of the effects of changing illumination conditions for the RGB camera through day time and night time scenes, while the rain and sun dataset split can be used to analyze the effects of rain on uncertainty-based metrics.

TABLE 2. AP results for Waymo.

Configuration	Image AP		LiDAR BEV AP		LiDAR 3D AP	
	Level 1	Level 2	Level 1	Level 2	Level 1	Level 2
Base	0.42	0.40	0.50	0.47	0.33	0.31
Base +	0.43	0.41	0.55	0.54	0.39	0.37
Base + +	0.54	0.52	**0.71**	**0.68**	0.54	0.51
Faster-RCNN	0.35	0.31	-	-	-	-
YOLO-V3 + Faster-RCNN	0.57	0.51	-	-	-	-
RW-TSDet	**0.84**	**0.77**	-	-	-	-
SECOND	-	-	-	-	0.50	0.43
PointPillars	-	-	-	-	0.55	0.49
HorizonLiDAR3D	-	-	-	-	**0.83**	**0.78**

Relative to Kitti, it is more difficult to achieve high AP results with the Waymo dataset, due to an increase in ground truth density and variety of objects in the frame. This includes an increase in the distance of object labels, resulting in a larger variation of bounding box size in the image perspective and a higher number of vehicles at a far distance (with few points) in the LiDAR BEV perspective. As shown in Table 2 the RGB camera and LiDAR detectors are compared against similar two stage detector architectures (including YOLO-V3+Faster-RCNN [33], RW-TSDet [19], SECOND [37], and HorizonLiDAR3D [9]) to ensure that our detectors are performing nominally. The state-of-the-art results for recent Waymo-tested object detectors are included, which outperform our detectors. These methods have gone through significant architecture optimization, such as anchor free detection, Cascade R-CNN and multi-scan aggregation [9], [19].

3) CANADIAN ADVERSE DRIVING CONDITIONS

The Canadian Adverse Driving Conditions Dataset [31] contains approximately 7,000 frames of data across 75 scenes featuring various adverse weather scenarios. Each frame contains captured data from eight cameras, one LiDAR, a GPS unit and an IMU [31]. Rather than focus on night/day driving like Waymo, CADC aims to capture and label data across various snowy conditions experienced in the harsh Canadian winter. CADC includes labels for five snowfall levels: none, light, medium, heavy and extreme. Also, scenes are labeled to include information such as road snow cover camera lens occlusion from snow.

In this work, CADC is leveraged to analyze uncertainty metrics involving snowy conditions on the LiDAR object detector. Also, the effects of obstructions (snowflakes) on an RGB Camera lens can be analyzed to understand if obstructions degrade camera performance significantly. This is useful, as obstructions are common and can come from many sources including snow, dirt or even a broken lens.

No benchmarks are available for the CADC dataset, and no released works discuss AP for the same split used. Thus, AP results are released as a baseline, shown in Table 3.

TABLE 3. AP results for CADC.

Configuration	Image 2D AP	LiDAR BEV AP	LiDAR 3D AP
Base	0.39	0.31	0.11
Base + Augmentation	0.40	0.45	0.25
Base + Augmentation + Uncertainty	**0.48**	**0.48**	**0.30**

VI. MODELING STATISTICS

The approach taken in this work is to directly model uncertainty estimates via the added network output parameters $g^{\boldsymbol{w}}(\boldsymbol{x_i})$ and sample variance of the T Monte-Carlo samples. Although this approach does not allow for uncertainty estimates to be physically interpretable, less noise is introduced into the distribution as it is directly modelled. Furthermore, uncertainty estimates do not need to be physically interpretable in this work, as degraded sensor states can be detected through relative changes in the distributions of uncertainty with the use of modeling statistics.

To model uncertainty for degraded state detection, aleatoric and epistemic uncertainty estimates are independently extracted for every detection during inference. Epistemic uncertainty is extracted by computing sample variance of the T Monte-Carlo samples of predicted regression values and classification logit scores, while aleatoric uncertainty is extracted by computing the mean of the T samples of the additional network output parameters $g^{\boldsymbol{w}}(\boldsymbol{x_i})$. For the regression task, the sample variance values obtained are modelled as the independent diagonal elements I of a multi-variate Gaussian covariance matrix. Similarly for the classification task, the sample variance is also used to directly estimate uncertainty from the noise in the logit score samples.

Special care must be taken when formulating statistical models of uncertainty, because the output of the neural network, which is described by the predictive likelihood function $p(y_*|x_*, \mathcal{D})$, can form any arbitrary distribution. This is due to the Gaussian prior placed over each weight $p(w_i) \sim \mathcal{N}(0, \sigma^2)$, meaning the output distribution is a sum of a large number of Gaussian distributions.

Due to the arbitrary nature of the distributions of uncertainty produced by the network, non-parametric techniques are leveraged to model distributions of uncertainty estimates, such as *Kullback-Leibler* (*KL*) divergence and multivariate *Kernel Density Estimators* (*KDEs*) [8], [36]. Using KDEs, a nominal distribution of uncertainty values can be modelled, when each sensor is in a non-degraded state. Then, the estimated uncertainty density from the nominal distribution at run-time can be used to predict if a detection belongs to (i.e. is classified as) the distribution of false positives (FP) or true positives (TP), similar to the averaged precision (*AP*) metric. An example of KDEs is shown in Fig.2. It depicts distributions of variance for an uncertainty parameter; true positive uncertainty is shown in blue, with the corresponding KDE in orange, and false positive distribution in green, with its KDE in red.

A. DETECTION THEORY

Signal detection theory (SDT) was originally formulated to analytically find the optimal signal strength at which to

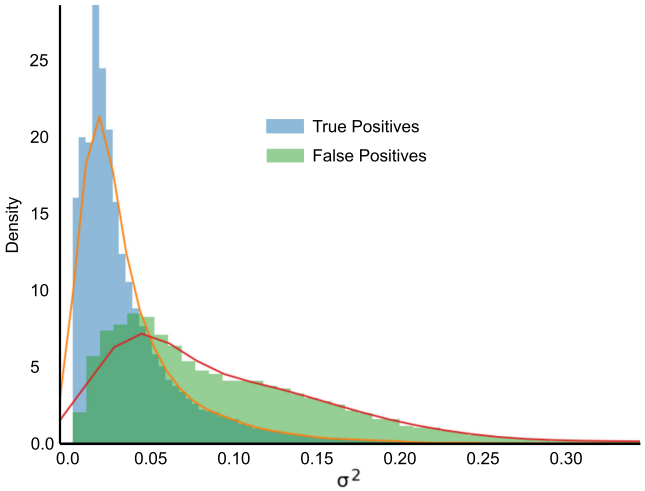


FIGURE 2. Example histogram of true positives (*blue*) and false positives (*green*) of a uncertainty parameter. Univariate KDEs are estimated over the TP split (*orange*) and FP split (*red*), respectively.

differentiate radar returns from two different types of objects, A and B. In this problem, two Gaussian probability distributions of signal return strength are defined, one belonging to the return for a type A object, the other for type B. An incoming return signal can be compared to both distributions and a decision can be made about what the object is likely to be. Any arbitrary point β can be chosen to represent the decision point between classifying the type of object represented by the return. Transferring this concept to object detection for vehicles and uncertainty estimation is straightforward, as the distributions are defined as uncertainty estimates belonging to False Positives (FP) and True Positives (TP).

The performance of the classifier can be evaluated by classifying a set of detections from the validation set as true positives or false positives based on the decision point β. The associated detection can then be compared to the ground truth from the validation set, to determine if the detection truly was a false positive or true positive. If the object has a significant overlap with any ground truth, its signal is denoted as TP ($IoU \geq 0.7$), or if it has little to no overlap with a vehicle ($IoU < 0.7$), the signal is denoted as FP. Using the combination of the two classifications, we can define four types of probabilities: *hit*, *miss*, *false alarm* (FA) and *correct rejection* (CR). For example, a *hit* represents the probability of a FP detection being correctly distinguished; it represents the chance that a true anomaly is classified as such. Thus, from leveraging SDT, an analytical understanding of uncertainty measures as a metric can be gained.

We can calculate the probabilities of correctly predicting TP and FP detections on the SDT classifier by integrating the conditional area of the distribution, either $(-\infty, \beta)$ or (β, ∞) to produce P_{HIT}, P_{MISS}, P_{FA}, and P_{CR}. Only P_{HIT} and P_{FA} need to be evaluated using Equations 1 and 2, as the other two are complements of the first pair ($P_{\text{MISS}} = 1 - P_{\text{HIT}}, P_{\text{CR}} = 1 - P_{\text{FA}}$).

$$P_{\text{HIT}} = \int_{\beta}^{\infty} p_{\text{FP}}(x)dx \tag{1}$$

$$P_{\text{FA}} = \int_{\beta}^{\infty} p_{\text{TP}}(x)dx \tag{2}$$

To understand the effects of moving the threshold value β on the classifier performance, it is swept across a range of values, producing a plot of P_{HIT} as a function of P_{FA}. The generated plot is called the *receiver operating characteristic* (*ROC*) curve. The informativeness of the classifier can be quantified by computing the *area under the curve* (AUC), which is a commonly-used metric for machine-learning applications [5].

To extend SDT from Gaussian distributions to non-parametric density functions, the decision boundary β is re-defined as a minimum threshold parameter. An incoming detection can be classified with a response of good detection or anomaly by computing the difference in densities between KDEs at the query point x_*, representing the estimated uncertainty associated with the detection. The detection is classified as a true positive if the TP density function exceeds the FP density function plus some threshold β, as seen in Equation 3.

$$TP := p_{TP}(x) > p_{FP}(x) + \beta \tag{3}$$

VII. RESULTS

This section evaluates uncertainty as a metric for discriminating between true- and false-positive detections. Results are given for the Kitti, Waymo, and CADC datasets.

A. COMPARING UNCERTAINTY FOR TP AND FP DETECTIONS

To produce uncertainty results we use an inference run over a *validation set* of the data. For each detection we check the bounding box against any corresponding ground truth. If the detection's IoU metric is ≥ 0.7 then it is recorded as a true-positive detection (*TP*), else false-positive (*FP*). We model the distribution of the uncertainty (variance) produced by this experiment (see Figure 2) for *TP* and *FP* detections and use KL divergence to quantify their differences.

For the 2D image-based object detector, uncertainty is considered for several combinations of regression parameters: the bounding box center (x_c, y_c), size (l, w), and all four (x_c, y_c, l, w). For the LiDAR detector, parameter combinations include: center of the bounding cuboid (x_c, y_c, z_c), size (l, w, h), heading (r_y), and all seven $(x_c, y_c, z_c, l, w, h, r_y)$.

For classification, both detectors are simplified to a binary classifier containing the foreground (*fg*) and background (*bg*) class. The *fg* is defined as the "Vehicle/Car" and the *bg* class forms the complement.

A visualization comparing *TP* and *FP* uncertainties is given in the form of a contour plot in Fig. 3. It shows, using Waymo data, a distribution of TP *blue* and FP *red* aleatoric regression uncertainty for the parameters (l, w). The figure

TABLE 4. Regression results for various parameters.

		Image			LiDAR			
		x_c, y_c	l, w	All	r_y	x_c, y_c, z_c	l, w, h	All
Kitti	Ale.	1.5	1.5	**1.8**	1.3	1.5	1.5	**1.6**
	Epi.	1.0	0.9	**1.3**	0.4	0.5	0.5	**0.8**
Waymo	Ale.	3.1	3.3	**4.3**	2.9	3.5	3.2	**4.9**
	Epi.	3.0	3.0	**3.4**	0.2	2.0	1.3	**2.5**
CADC	Ale.	2.0	2.0	**2.3**	0.9	1.6	1.7	**2.8**
	Epi.	1.6	1.4	**1.8**	0.2	0.6	0.5	**1.0**

TABLE 5. Classification results for various parameters.

		Image			LiDAR		
		fg	bg	All	fg	bg	All
Kitti	Ale.	0.8	0.9	**1.5**	0.7	0.3	**1.2**
	Epi.	1.0	1.05	**1.2**	0.6	0.6	**0.7**
Waymo	Ale.	2.9	0.3	**3.3**	2.7	0.7	**2.8**
	Epi.	**2.2**	**2.2**	**2.2**	2.0	**2.2**	**2.2**
CADC	Ale.	0.8	0.6	**1.2**	0.9	0.6	**1.2**
	Epi.	1.4	1.5	**1.6**	**0.9**	**0.9**	**0.9**

indicates that the distributions are dissimilar. Compared using KL divergence, these distributions produce a score of 3.3. The same measures were taken using epistemic regression uncertainty, which produced a KL divergence score of 3.0.

Results for various regression parameters are summarized in Table 4 and for classification parameters in Table 5. For each dataset there is one row showing aleatoric results (Ale.) and one row for epistemic (Epi.). The left side of the tables show results for the 2D image sensor, and the right side corresponds to LiDAR. While various levels of KL divergence are shown for the parameters, the column labelled "All" indicates that the combination of all parameters within an uncertainty measure results in the largest divergence. This suggests that the multi-dimensional KDE that models the distributions is able to leverage information from each uncertainty parameter in a complementary fashion.

The results in the "All" column in Table 4 are on average 1.8× larger for aleatoric uncertainty in comparison to epistemic uncertainty. Similarly, in Table 5 this factor is 1.3×. The tables show that in most cases epistemic uncertainty gives the smallest KL divergence score for most parameters, indicating it is the least informative measure.

Sensor Comparisons: Care must be taken when analyzing KL divergence scores from one sensor to another, given that the regression tasks may operate in a different number of dimensions. But the classification task can be compared for relative uncertainty values between the LiDAR and image detectors. On average, aleatoric classification uncertainty gives a 1.1× larger KL divergence for the image detector relative to LiDAR over all datasets. For epistemic classication uncertainty the image detector gives a 1.5× larger divergence. These results indicate that the image detector uncertainty estimates may be more informative than their LiDAR counterparts.

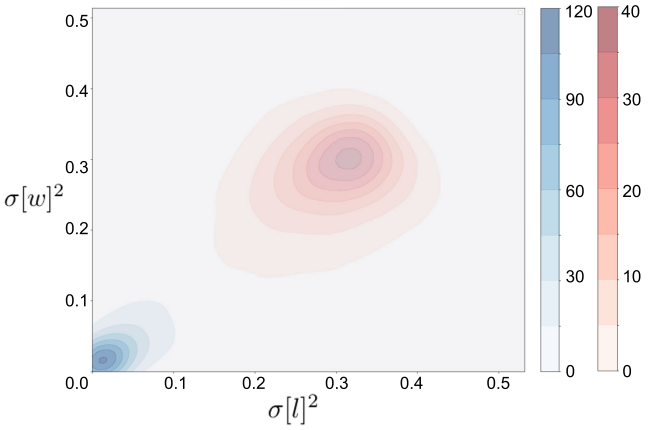


FIGURE 3. Contour plot of aleatoric regression uncertainty for the image detector (*l*, *w*) on Waymo. True positive distribution in *blue*, false positive distribution in *red*.

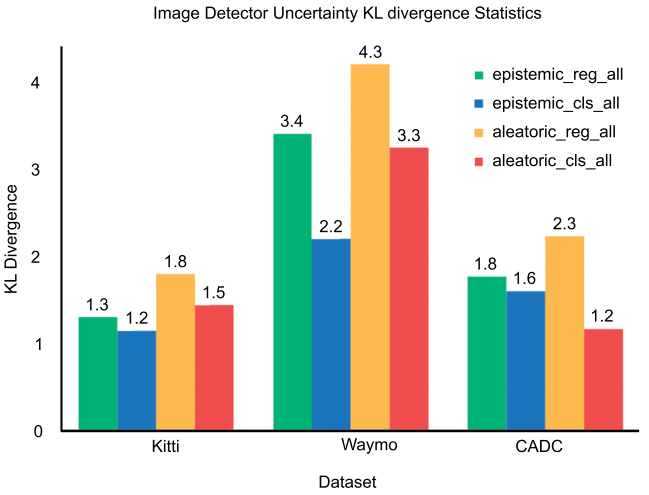


FIGURE 4. Image detector KL divergence for all tasks.

Dataset Comparisons: Examination of the results across the three datasets shows that Waymo produces the highest KL divergence scores. This may be a result of the increase in the diversity of training data for Waymo compared to the other datasets. Even epistemic uncertainty, which according to Kendall and Gal, should be reducible with more data [21], is significantly higher for Waymo. An illustration of these results is provided in Fig. 4. Each colored bar corresponds to a column labelled "All" for the 2D image sensor from either Table 4 or 5. For example the column identified as *epistemic_cls_all* is epistemic classification uncertainty and *aleatoric_reg_all* is aleatoric regression uncertainty.

B. RECEIVER OPERATING CHARACTERISTIC

The experimental results given above indicate that we can use non-parametric KDEs to model uncertainty distributions of TP and FP detections, and use SDT to distinguish them. We wish to utilize this method to form a classifier of TP and FP detections that can be used during inference. This classifier needs to achieve a high performance that yields

a desired probability of TP, and an acceptable probability of FP detections. The receiver operating characteristic (Section VI-A) provides such a metric that can describe the performance of our uncertainty-based classifier.

1) EXPERIMENTAL SETUP

Following the procedure described earlier, an inference run is performed over the validation split of a dataset, collecting output detections and associated uncertainty values into a results database. The database comprises 15, 000 detections drawn from the validation run. We compare detections to ground truth objects and discriminate between FP ($IoU < 0.7$) and TP ($IoU \geq 0.7$). An N-dimensional multi-variate KDE is used to generate two density distributions over the TP/FP split. For the classification task $N = 2$, corresponding to the foreground (vehicle) and background class (complement). For the regression task, a multivariate KDE is generated of dimension $N = 4$ for the image detector and $N = 7$ for LiDAR. Additionally, *total* predictive uncertainty is included, which is up to an $N = 18$ dimensional KDE (for LiDAR), containing all parameters for aleatoric and epistemic regression and classification.

Using equation 3, each detection can be classified as TP or FP based on a density threshold β. By performing a *sweep* over a range of thresholds a receiver operating curve (ROC) can be obtained. The area under the curve (AUC) can be calculated by using discrete integration, giving an estimate of the performance for each uncertainty parameter combination. Furthermore, we can pick a desired ROC operating point via the threshold value (β_{opt}) which is set by choosing a maximum acceptable FA rate on the ROC curve; for this work the probability FA = 0.1 is selected, to create a conservative anomaly classifier. An example ROC curve generated using the Waymo dataset for the 2D image detector can be seen in Fig. 5. This result includes filtering of detections using a 0.5 confidence threshold, because the uncertainty values of low confidence detections would not be indicative of sensor performance.

2) ANALYSIS

Complete AUC results over all datasets for the image detector are shown in Table 6, and for LiDAR in Table 7. For each dataset, results are shown for three confidence levels: 0.01, 0.5, and 0.7. These results demonstrate that for each sensor and dataset **at least one metric has a *good* AUC** (≥0.8) **at the highest confidence threshold** (0.7). This indicates high metric informativeness, even with a significant threshold applied, which cannot simply be filtered out without reducing recall.

Tables 6 and 7 indicate that the combination of all uncertainties generates the most informative classifier in most cases, with aleatoric bounding box uncertainty occasionally outperforming it. However, not all metrics appear highly informative, especially at high confidence thresholds. Uncertainty parameters like epistemic classification uncertainty demonstrate the highest drop in performance with increasing confidence threshold, with near random guess performance at 0.7 confidence threshold for the Kitti trained LiDAR detector.

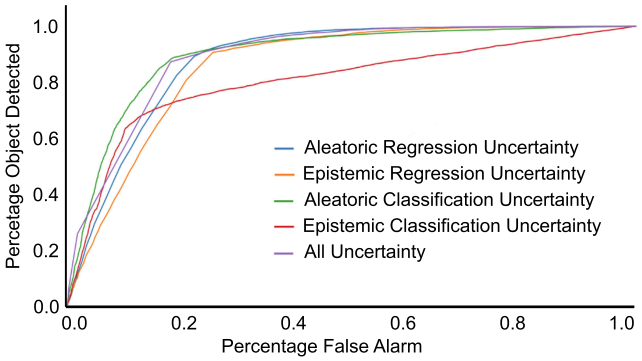


FIGURE 5. ROC curve for the image detector on Waymo with 0.5 confidence threshold.

TABLE 6. Image detector ROC AUC metric.

Dataset	Confidence threshold	Aleatoric regression	Epistemic regression	Aleatoric classification	Epistemic classification	All
Kitti	0.01	0.90	0.86	0.87	0.85	**0.92**
	0.5	0.85	0.78	0.81	0.75	**0.87**
	0.7	0.83	0.76	0.79	0.71	**0.85**
Waymo	0.01	0.91	0.89	0.87	0.91	**0.94**
	0.5	0.89	0.86	0.89	0.80	**0.92**
	0.7	0.88	0.84	0.86	0.64	**0.89**
CADC	0.01	0.92	0.90	0.85	0.90	**0.9**
	0.5	**0.85**	0.80	0.80	0.78	0.80
	0.7	0.83	0.77	0.77	0.73	**0.78**

TABLE 7. LiDAR detector ROC AUC metric.

Dataset	Confidence threshold	Aleatoric regression	Epistemic regression	Aleatoric classification	Epistemic classification	All
Kitti	0.01	0.91	0.81	0.83	0.78	**0.92**
	0.5	0.84	0.71	0.74	0.57	**0.85**
	0.7	0.81	0.66	0.73	0.54	0.83
Waymo	0.01	0.97	0.87	0.97	0.95	**0.98**
	0.5	**0.88**	0.73	0.85	0.72	0.85
	0.7	**0.85**	0.71	0.81	0.66	0.81
CADC	0.01	0.90	0.80	0.85	0.82	**0.91**
	0.5	0.79	0.70	0.71	0.64	**0.81**
	0.7	0.78	0.69	0.69	0.60	**0.80**

A comparison between the AUC metric and the KL divergence scores reveals a positive correlation between the two measures. This is indicated by identical ranking order of magnitude between the two measures, which is generally (from lowest to highest): epistemic classification, epistemic regression, aleatoric classification and finally aleatoric regression uncertainty. Although this general order has some notable exceptions like the CADC Image Detector, the KL divergence scores reflect this as well. Thus, many of the same trends identified in KL divergence between datasets, sensors and metrics hold with the AUC metric.

C. ADVERSE CONDITIONS

This section presents test cases under adverse conditions, including rain, fog, and snow, which is the primary motivation

for our uncertainty metric analysis. After filtering results using a .5 confidence threshold, the performance of uncertainty metrics is quantified by computing the ratio of false positives to true positives $\frac{FP}{TP}$, using the SDT classifier created from the nominal KDEs of uncertainty outputs. This metric is defined as the *Anomalous Detections Ratio* (ADR). To compute ADR, the uncertainty values produced for each detection are used as input values to the nominal KDEs of TP and FP detections described in Section VII-B1, above. Each of these nominal KDEs produces a likelihood value corresponding to the current detection's uncertainties, and Equation 3 classifies this detection as being either TP or FP. No ground truths are required to compute ADR, since it is estimated using only the uncertainty likelihoods that are stored in the nominal KDEs. Thus, in a actual usage scenario ADR can be measured in real-time, and a running average of ADR provides a mechanism to assess whether a sensor is currently working well and producing a high ratio of correct detections, or whether that sensor is more likely to be producing detections that have a relatively-low likelihood of being correct.

The datasets are augmented by generating additional test cases using simulated distortions, including random dropout of pixels at 20% and 40%, simulated lens obstruction, and fog [20]. We use augmentation methods akin to those discussed and evaluated in [3] and [38]. For the LiDAR detector, a point cloud rain simulator is implemented that can distort point clouds using estimated power loss and back scatter equations measured in mm/h [14]. In this work, the test cases include relatively common rain rates, defined as light: 1 mm/h, medium: 3 mm/h and heavy: 5 mm/h.

1) WAYMO IMAGE DISTORTIONS

Using the Waymo dataset with RGB camera four splits on the validation dataset are tested: day, night, rainy, and the full set (as a control test). After performing an inference run over each split, the average precision is computed over the two levels of difficulty. Waymo includes a (20 second) sequence of images labelled *rainy*, illustrated by Fig. 7, in which image features are distorted due to water droplets and spray.

We produced a detection classifier from the aleatoric regression uncertainty measure to classify detections as TP or FP. Aleatoric regression uncertainty is chosen over all uncertainties as findings indicate a comparable performance but reduced computational cost for this measure. Table 8 indicates that ADR is highly responsive to adverse conditions that give reduced AP. Night driving (Fig 6) and driving in the rain with an obscured camera lens (Fig 7) report increases in the ADR of 1.4× and 2.1× respectively. This result suggests that uncertainty metrics can be used to detect a degraded state, as ADR increases when AP decreases.

We performed the following simulated degradation tests: pixel dropout at 20% (Fig 8) and 40% (Fig. 9), fog (Fig 11), and lens obstruction (dirt spatter) (Fig 10). The results indicate a clear inverse correlation between ADR and AP across the board of simulated tests. As an example, 20% dropout results in a 1.5× increase in ADR and 12% decrease in AP. Fog simulation also reports a relatively large (1.4×) increase in ADR and decrease (19%) in AP, which is attributable to fog partially obscuring ground truth objects in some frames. The full set of results can be viewed in Table 9.

TABLE 8. Waymo image adverse conditions.

Metric	All	Day	Night	Rain
AP L1	0.54	0.54	0.51	0.48
AP L2	0.52	0.52	0.51	0.48
ADR	**1.07**	**1.02**	**1.47**	**2.26**

TABLE 9. Waymo image simulated adverse conditions.

Metric	20% Dropout	40% Dropout	Fog	Spatter
AP L1	0.42	0.35	0.35	0.40
AP L2	0.39	0.31	0.32	0.38
ADR	**1.16**	**1.46**	**1.39**	**1.16**

FIGURE 6. Image - Waymo - a) day test case. b) night test case. White: Ground truth boxes. Green-to-blue: more uncertain detections.

FIGURE 7. Image - Waymo - rain test case. White: Ground truth boxes. Green-to-blue: more uncertain detections.

FIGURE 8. Image - Waymo - 20% dropout test case. White: Ground truth boxes. Green-to-blue: more uncertain detections.

FIGURE 9. Image - Waymo - 40% dropout test case. White: Ground truth boxes. Green-to-blue: more uncertain detections.

FIGURE 10. Image - Waymo - spatter test case. White: Ground truth boxes. Green-to-blue: more uncertain detections.

FIGURE 11. Image - Waymo - fog test case. White: Ground truth boxes. Green-to-blue: more uncertain detections.

2) WAYMO LiDAR DISTORTIONS

Using the Waymo dataset and the LiDAR detector, the same four splits are investigated. These tests use all of the same methods as described above, apart from the signal classification using a 3D IoU metric, rather than in 2D. In an important test of asymmetric sensor distortion, the night-driving split is tested and compared between the LiDAR and camera sensors. While the camera reports a nearly $\sim$1.5$\times$ increase in the ADR with a $\sim$3% AP drop, the LiDAR sensor reports relatively small changes in both ADR and AP (Table 10). This result supports the idea introduced by Bijelic et al. that asymmetric sensor distortion can result in AP drops in one detection modality, while not affecting another [2].

The rain simulation algorithm is used to augment the dataset with light (1 mm/h), moderate (3 mm/h) and heavy (5 mm/h) rain conditions, by distorting the LiDAR point clouds. The AP and ADR results can be seen in Table 11, while example LiDAR pseudo images distorted with rain can be seen in Fig. 12.

Similarities are found between the image dropout distortion tests and the LiDAR rain simulation tests, as it requires a significant decrease in AP to report a large increase in ADR. The findings indicate that ADR is efficacious in detecting a degraded rain state: the heaviest rain rate 5 mm/h reports a $\sim$4.2$\times$ increase in the ADP, with a $\sim$30% drop in AP.

TABLE 10. Waymo LiDAR adverse conditions.

Metric	All	Day	Night	Rain
AP L1	**0.54**	**0.50**	**0.52**	**0.53**
AP L2	0.51	0.49	0.51	0.51
ADR	0.16	0.15	0.12	0.09

TABLE 11. Waymo LiDAR simulated rain.

Metric	1 mm/h	3 mm/h	5 mm/h
AP L1	**0.43**	0.26	0.22
AP L2	0.40	0.24	0.20
ADR	0.12	**0.62**	**0.76**

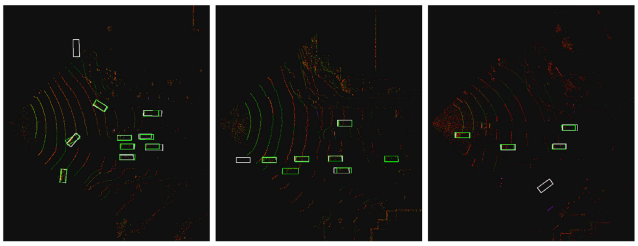


FIGURE 12. LiDAR - Waymo - simulated rain test case. White: Ground truth boxes. Green-to-blue: more uncertain detections. a) no rain b) 1mm/h c) 3mm/h d) 5mm/h.

TABLE 12. CADC LiDAR snowfall adverse conditions.

Metric	All	None, Light	Medium, Heavy, Extreme
AP	**0.30**	**0.34**	**0.28**
ADR	0.15	0.12	0.17

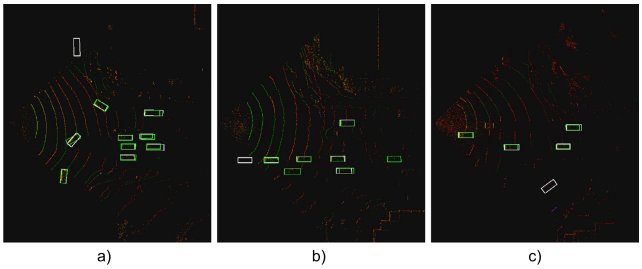


FIGURE 13. LiDAR - CADC - Snowfall test white: Ground truth boxes. Green: detections. a) light snowfall b) medium snowfall c) heavy snowfall.

TABLE 13. CADC LiDAR adverse conditions.

Metric	All	Clear Lens	Obstructed Lens
AP	0.48	0.54	0.32
ADR aleatoric_reg	**0.66**	**0.69**	**0.45**
ADR aleatoric_cls	0.20	0.19	0.21

3) CADC LiDAR DISTORTIONS

This section evaluates the various adverse snowfall conditions in the *Canadian Adverse Driving Conditions* (CADC) dataset. Nominal KDEs are generated using a split that removes the various snowfall rates (Light, Medium, Heavy, Extreme, as visualized in Fig. 13) from the dataset. Next, two splits are used in a comparison test, to investigate the relationship between AP and ADR in snowfall conditions. The first split contains frames from None and Light levels of snowfall, while the second half contains Medium, Heavy and Extreme. For this test the all uncertainty metric is chosen, as it is the only metric that scores above an 0.8 AUC at a 0.5 confidence threshold. Findings indicate that LiDAR experiences degraded performance in higher snowfall conditions, losing approximately 7% AP. Also, ADR reports a relatively large increase of 1.4$\times$ when comparing the two splits, indicating that the metric is responsive to snowfall-related

sensor degradations. The full set of results are shown in Table 12.

VIII. DISCUSSION AND CONCLUSION

In this paper, an object detector supporting the RGB camera and LiDAR sensors is developed using a Faster R-CNN architecture as a baseline framework. Then, using the approximate modelling of a Bayesian neural network, the baseline object detector is modified to generate estimates of four types of uncertainties: aleatoric regression, epistemic regression, aleatoric classification and epistemic classification. The focus of our work is on uncertainty estimation, and in this context our network architecture with uncertainty estimation reports AP results that are comparable to previous results from the literature.

We perform a comprehensive analysis of the four types of estimated uncertainty over three datasets, *Kitti, CADC, Waymo*, for our two sensor types. As a preliminary analysis, KL divergence is used as a disparity measure between the sets of uncertainty values associated with true positive and false positive detections. This is followed by an investigation into the performance of uncertainty estimation as a discriminative classifier using signal detection theory (SDT), with its performance measured by the AUC metric. Last, a novel metric is presented in this work, defined as the *anomalous detections ratio* (ADR). We evaluate the responsiveness of this novel ADR metric by using cases that involve adverse weather and simulated sensor degradation.

Findings from the KL divergence and AUC metric analysis indicate that the two measures are highly correlated, as they follow similar trends across uncertainty measures, sensor types and datasets. Aleatoric regression uncertainty is found to be significantly more discriminative between false positive and true positive detections when compared to other uncertainty types. In general, epistemic uncertainty demonstrates the poorest results; but this could be due to a lack of variability in our Monte-Carlo output samples, since dropout was enabled only in the last few layers of the network. The image detector is found to produce higher KL divergence and AUC metric scores than the LiDAR detector, indicating that the 2D RGB image is a more informative input or perspective than a LiDAR BEV pseudo-image. Waymo is observed to consistently give the highest KL divergence and AUC metric scores, indicating that highly varied data in large datasets results in more informative uncertainty measures. However, regardless of the trends indicated, the results from the SDT analysis demonstrate that uncertainty measures are informative when used as false positive and true positive classifiers, producing a minimum AUC of 0.8 for all sensors and datasets at a confidence threshold of 0.7, with at least one uncertainty measure.

Our results with the RGB camera object detector show that the novel ADR metric is highly responsive to degraded conditions such as night time driving and obscured images in the rain, produced from the Waymo dataset. In both test cases, drops in AP correspond to increases in ADR, indicating an inverse correlation between the two metrics. Further testing with simulated image distortions such as dropout, fog and lens spatter also demonstrate a similar relationship between AP and ADR, but appear less sensitive than real-world adverse weather conditions. Results using the LiDAR object detector indicate that ADR is responsive to degraded snowfall conditions in the CADC dataset split chosen, as well as simulated rain conditions with the Waymo dataset. This is evidenced by an overall drop in AP and increase in ADR, suggesting the two are inversely correlated for the LiDAR object detector as well.

As stated above, the emphasis of this work is on being able to leverage the ADR method in real-time as a metric for assessing whether a sensor is currently working well, or not. In this context, while absolute detection accuracy is not as important as the ability to produce the ADR metric as a running average, we need to support a sufficiently-high detection rate for real-time usage. We have supported this goal in our approach that uses a two-stage backbone network, rather than just a single stage. Thusly, we are able to ameliorate much of the possible deleterious effects on performance by computing the Monte-Carlo samples needed for uncertainty estimation in only the second-half (refinement stage) of the network, not in the network backbone. Of course, the absolute performance of any object detector will be highly dependent on the compute-hardware chosen, such as CPU, GPU, or custom accelerator. In addition to the Faster-RCNN network that we have used, it would be of interest to test the ADR metric with other two-stage networks; however, there should be little variation in performance between the two-stage backbones when controlling for the number of parameters in the backbones and inference time if training on large scale datasets, such as Waymo and the like.

Overall, the novel ADR metric demonstrates a responsiveness to real world and simulated degraded performance test cases, suggesting that uncertainty estimation can be an informative measure for assessing sensor performance degradation. However, there is a significant amount of work remaining when it comes to ensuring the safety of autonomous vehicles on the road today and in the future. Current datasets are just beginning to enable research into the realm of adverse conditions, and more high quality data is needed to enable a deeper analysis. The applications space of uncertainty estimation still remains largely unexplored, especially in the context of sensor fusion, which is discussed below.

IX. FUTURE WORK

As an extension of the in-depth analysis of uncertainty estimates in object detectors, we propose that this work is the beginning of a research stream, which focuses on autonomous driving in adverse conditions with multiple sensors. Two possible avenues of future work are proposed below.

A. FUSION NETWORKS

In this paper, uncertainty estimation is performed for sensors in independent domains. However, in real-world multi-sensor

systems, an autonomous driving system (ADS) must reduce all of its inputs down to a single set of decisions made to control the vehicle. Thus, future work must be performed to combine, or *fuse*, information between sensors and perception systems. Combining fusion with uncertainty estimation could result in a flexible fusion architecture, capable of dynamically fusing sensors. This is especially relevant when considering adverse conditions, as asymmetric sensor distortion can degrade fixed fusion architecture performance if the distortion was not experienced during training. Thus, we present two flexible fusion approaches, involving late fusion and feature fusion, respectively.

1) LATE FUSION

One approach to flexible fusion is to combine detections in a late fusion voting based scheme. In this approach, independently produced detections can be matched between sensor modalities using the IoU metric and then combined using uncertainty. This approach could increase the robustness of the object detector if driven with uncertainty, as detection voting can be weighted with uncertainty metrics, such as the inverse of the ADR. However, the current detector configuration does not support 3D object detection with the late fusion technique, as the image detector only operates in 2D.

To support late fusion, 3D object detection with the RGB camera must be implemented, in one of two suggested approaches. In the first approach, LiDAR point cloud data can be transformed to the image domain, where a sparse depth encoding over the image can be obtained. Next, depth completion can be applied, even with a simple algorithm such as linear interpolation, to obtain an RGB-D image. This RGB-D image is capable of natively storing 3D information in its depth channel, just like the LiDAR point cloud. An alternative approach to adding 3D support for RGB image object detection is to use stereo camera sensors to predict in 3D, much like existing works Stereo R-CNN [24].

2) FEATURE FUSION

Feature based fusion is another approach that can be leveraged to extend this work into the fusion domain. The suggested approach is to follow the feature-level fusion architecture that was introduced by works such as MV3D and AVOD [6], [22]. These works leverage a two stage architecture (i.e. Faster R-CNN) to fuse data at the region proposal layer, where each proposal is a cropped set of features. In the original fusion implementation, these cropped proposals or ROIs, are fused by performing element-wise addition between the features produced from each sensor's feature extractor, before being presented to the second stage detector. Uncertainty estimation can also be implemented at the region proposal stage. Adding uncertainty into the fusion method could out-perform basic element-wise addition, as cropped ROIs from sensors in degraded states will introduce unwanted noise into the fused ROI. Specifically, asymmetric sensor distortion scenarios (such as night time driving), could result in poor performance with element wise addition, as an RGB camera would add unwanted noise to the nominally operating LiDAR sensor.

As future work, two possible techniques are proposed to perform feature-level fusion with uncertainty estimates extracted from the region proposal layer. The first involves element-wise multiplication of ROIs with a mask tensor, which is known as the *attention masking* technique. Uncertainty can be used to drive the values of this mask tensor, acting to attenuate feature information in high uncertainty scenarios. Alternatively, uncertainty estimates for every ROI can be reduced to a scalar value, either through averaging operations or a simple multi-layer perceptron. This scalar value can be used as a uniform attention mask across the entire ROI, before element wise feature level fusion is applied. This technique can be thought of as a region proposal mixture of experts method, where each expert is a different sensor type.

REFERENCES

[1] J. Beltran, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. D. L. Escalera, "BirdNet: A 3D object detection framework from LiDAR information," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3517–3523.

[2] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11679–11689.

[3] M. Bijelic, T. Gruber, and W. Ritter, "Benchmarking image sensors under adverse weather conditions for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1773–1779.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[5] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.

[6] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6526–6534.

[7] X. Chen and A. Gupta, "An implementation of faster RCNN with study for region sampling," 2017, *arXiv:1702.02138*.

[8] Y. Chen, "A tutorial on kernel density estimation and recent advances," *Biostatist. Epidemiol.*, vol. 1, no. 1, pp. 161–187, 2017.

[9] Z. Ding, Y. Hu, R. Ge, L. Huang, S. Chen, Y. Wang, and J. Liao, "1st place solution for Waymo open dataset challenge 3D detection and domain adaptation," 2020, *arXiv:2006.15505*.

[10] D. Feng, Y. Cao, L. Rosenbaum, F. Timm, and K. Dietmayer, "Leveraging uncertainties for deep multi-modal object detection in autonomous driving," 2020, *arXiv:2002.00216*.

[11] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 9961–9980, Aug. 2022.

[12] Y. Gal, "Uncertainty in deep learning," Ph.D. thesis, Dept. Eng., Univ. Cambridge, Cambridge, U.K., Sep. 2016.

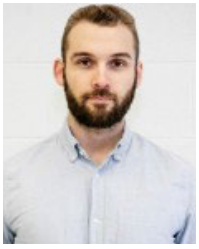
[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[14] C. Goodin, D. Carruth, M. Doude, and C. Hudson, "Predicting the influence of rain on LiDAR in ADAS," *Electronics*, vol. 8, no. 1, p. 89, Jan. 2019.

[15] A. Harakeh, M. Smart, and S. L. Waslander, "BayesOD: A Bayesian approach for uncertainty estimation in deep object detectors," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 87–93.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *2017 IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[18] K. M. Hildebrand, "Uncertainty measurement as a sensor performance metric in adverse conditions," M.S. thesis, Dept. Elect. Comput. Eng., Univ. Toronto, Toronto, ON, Canada, 2020.
[19] Z. Huang, Z. Chen, Q. Li, H. Zhang, and N. Wang, "1st place solutions of Waymo open dataset challenge 2020 2D object detection track," 2020, *arXiv:2008.01365*.
[20] A. B. Jung, K. Wada, J. Crall, S. Tanaka, and J. Graving. (2020). *Imgaug*. [Online]. Available: https://github.com/aleju/imgaug
[21] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 5580–5590.
[22] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–8.
[23] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12697–12705.
[24] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7644–7652.
[25] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[26] J. Liu, J. Paisley, M.-A. Kioumourtzoglou, and B. Coull, "Accurate uncertainty estimation and decomposition in ensemble learning," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, pp. 8952–8963.
[27] T. Liu and Y. Liu, "Moving camera-based object tracking using adaptive ground plane estimation and constrained multiple kernels," *J. Adv. Transp.*, vol. 2021, pp. 1–15, Jul. 2021.
[28] T. Liu, Y. Liu, Z. Tang, and J.-N. Hwang, "Adaptive ground plane estimation for moving camera-based 3D object tracking," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2017, pp. 1–6.
[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and C. A. Berg, "SSD: Single shot multibox detector," in *Proc. ECCV*, Oct. 2016, pp. 21–37.
[30] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Cham, Switzerland: Springer, 2012.
[31] M. Pitropov, D. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnecki, and S. Waslander, "Canadian adverse driving conditions dataset," 2020, *arXiv:2001.10117*.
[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[33] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. IPS*, 2015, pp. 1–12.
[35] P. Sun, "Scalability in perception for autonomous driving: Waymo open dataset," 2019, *arXiv:1912.04838*.
[36] S. Weglarczyk, "Kernel density estimation and its application," in *Proc. ITM Web Conf.*, vol. 23, 2018, Art. no. 00037.
[37] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.
[38] N. Zhang, L. Zhang, and Z. Cheng, "Towards simulating foggy and hazy images and evaluating their authenticity," in *Proc. Int. Conf. Neural Inf. Process.*, Guangzhou, China, Nov. 2017, pp. 405–415.
[39] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for LiDAR 3D vehicle detection," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3266–3273.



MATHEW HILDEBRAND (Member, IEEE) received the B.S. degree in electrical and electronics engineering from Carleton University and the M.A.Sc. degree in computer engineering from the University of Toronto, in 2020. He is currently a Pixel IP Design Engineer with Apple.

ANDREW BROWN received the B.A.Sc. degree in computer engineering from Queen's University, Kingston, ON, Canada, in 2021. He is currently pursuing the M.A.S. degree with the University of Toronto, performing research on natural language-processing.

STEPHEN BROWN (Senior Member, IEEE) received the B.A.Sc. degree in electrical engineering from the University of New Brunswick and the M.A.Sc. and Ph.D. degrees in electrical engineering from the University of Toronto. He joined the University of Toronto, as a Faculty Member, in 1992, where he is currently a Professor with the Department of Electrical and Computer Engineering. He is also the Director of the FPGA Academic Programs, Intel Corporation. He is the coauthor of more than 150 scientific research articles and two other textbooks: *Fundamentals of Digital Logic with Verilog Design* and *Field-Programmable Gate Arrays*. His research interests include field-programmable VLSI technology, CAD algorithms, computer architecture, and applications of machine learning. He received the Canadian Natural Sciences and Engineering Research Council's 1992 Doctoral Prize for the Best Ph.D. thesis in Canada, and the New Brunswick Governor-General's 1985 Award for the highest academic standing in the Faculty of Engineering. He received many awards for excellence in teaching electrical engineering, computer engineering, and computer science courses.

STEVEN L. WASLANDER (Senior Member, IEEE) received the B.Sc.E. degree from Queen's University, in 1998, and the M.S. and Ph.D. degrees in aeronautics and astronautics from Stanford University, in 2002 and 2007, respectively. He joined the University of Waterloo, in 2008, where he founded and directed the Waterloo Autonomous Vehicle Laboratory (WAVELab). He is a leading authority on autonomous aerial and ground vehicles, including multirotor drones and autonomous driving vehicles, simultaneous localization and mapping (SLAM), and multi-vehicle systems. In 2018, he joined the Institute for Aerospace Studies (UTIAS), University of Toronto, and founded the Toronto Robotics and Artificial Intelligence Laboratory (TRAILab). His innovations were recognized by the Ontario Centres of Excellence Mind to Market Award for the best industry/academia collaboration (2012, with Aeryon Labs), and the Best Paper Award and the Best Poster Award from the Computer and Robot Vision Conference, in 2018. His work on autonomous vehicles has resulted in the Autonomoose, the first autonomous vehicle created with a Canadian University to drive on public roads.