2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

BayesOD: A Bayesian Approach for Uncertainty Estimation in Deep Object Detectors

Ali Harakeh[1], Michael Smart[2], and Steven L. Waslander[1]

***Abstract*— When incorporating deep neural networks into robotic systems, a major challenge is the lack of uncertainty measures associated with their output predictions. Methods for uncertainty estimation in the output of deep object detectors (DNNs) have been proposed in recent works, but have had limited success due to 1) information loss at the detectors non-maximum suppression (NMS) stage, and 2) failure to take into account the multitask, many-to-one nature of anchor-based object detection. To that end, we introduce BayesOD, an uncertainty estimation approach that reformulates the standard object detector inference and Non-Maximum suppression components from a Bayesian perspective. Experiments performed on four common object detection datasets show that BayesOD provides uncertainty estimates that are better correlated with the accuracy of detections, manifesting as a significant reduction of 9.77%-13.13% on the minimum Gaussian uncertainty error metric and a reduction of 1.63%-5.23% on the minimum Categorical uncertainty error metric. Code will be released at https://github.com/asharakeh/bayes-od-rc.**

I. Introduction

Due to their high level of performance, deep object detectors have become standard components of perception stacks for safety critical tasks such as autonomous driving [1], [2], [3] and automated surveillance [4]. Therefore, the quantification of how trustworthy these detectors are for subsequent modules, especially in safety critical systems, is of utmost importance. To encode the level of confidence in an estimate, a meaningful and consistent measure of uncertainty should be provided for every detection instance (see Fig. 1).

Two important goals must be met to create a meaningful uncertainty measure. First, the robotic system should be capable of using the uncertainty measure to fuse an object detector's output with prior information from different sources [5] to connect sequences of detections over time and increase detection and tracking performance as a result. Second and most importantly, the robotic system should be able to use its own estimates of detection uncertainty to reliably identify incorrect detections, including those resulting from *out of distribution instances*, where object categories, scenarios, textures, or environmental conditions have not been seen during the training phase [5].

Two sources of uncertainty can be identified in any machine learning model. *Epistemic* or model uncertainty is the uncertainty in the model's parameters, usually as a result of the confusion about which model generated the training data, and can be explained away given enough representative training data points [6]. *Aleatoric* or observation uncertainty results from the stochastic nature of the observed input, and persists in network output despite expanded training on additional data [7].

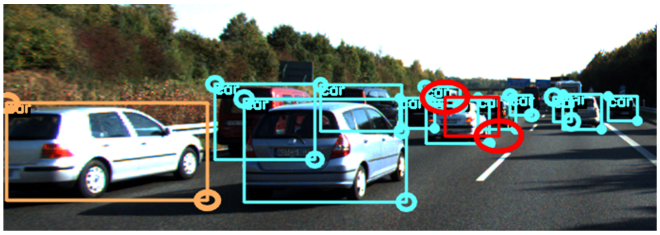

Fig. 1. The output from BayesOD, demonstrated on a test image frame from the KITTI Dataset [8]. Three levels of trust (**teal**: highly reliable, **orange**: slightly reliable and **red**: unreliable) are determined based on thresholds of the Gaussian entropy provided by BayesOD. All bounding boxes are shown with the 95% confidence ellipse of their top-left and bottom-right corners.

Methods to estimate both uncertainty types in DNNs have been recently proposed by Kendal *et al*. [7], with applications to pixel-wise perception tasks. Recent methods [9], [10], [11], [12], [13], [14], [15] extended Kendal's work [7] to object detection, but fail to consider the multi-task, many-to-one nature of the object detection task. To that end, we introduce BayesOD, a framework designed to estimate the uncertainty in both bounding box and category of detected object instances. This paper offers the following contributions:

- We provide a Bayesian treatment for every step of the neural network inference procedure, allowing the incorporation of anchor-level and object-level priors in closed form.
- We replace standard non-maximum suppression (NMS) with Bayesian inference, allowing the detector to retain all predicted information for **both the bounding box and the category** of a detected object instance.
- We perform comprehensive experiments to quantify the quality of the estimated uncertainty on four commonly used 2D object detection datasets, COCO, Pascal VOC, Berkeley Deep Drive and Kitti. We show that BayesOD provides a significant reduction of $9.77\% - 13.13\%$ on the minimum Gaussian uncertainty error metric, a reduction of $1.63\% - 5.23\%$ on the minimum Categorical uncertainty error metric, and an increase of $0.07\% - 3.00\%$ on the probabilistic detection quality over the next best method from current state of the art.

[1]Ali Harakeh and Steven L. Waslander are with The Institute For Aerospace Studies (UTIAS), University of Toronto, Toronto, Canada, ali.harakeh@utoronto.ca, steven.w@utias.utoronto.ca

[2] Michael Smart is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, Canada, michael.smart@uwaterloo.ca

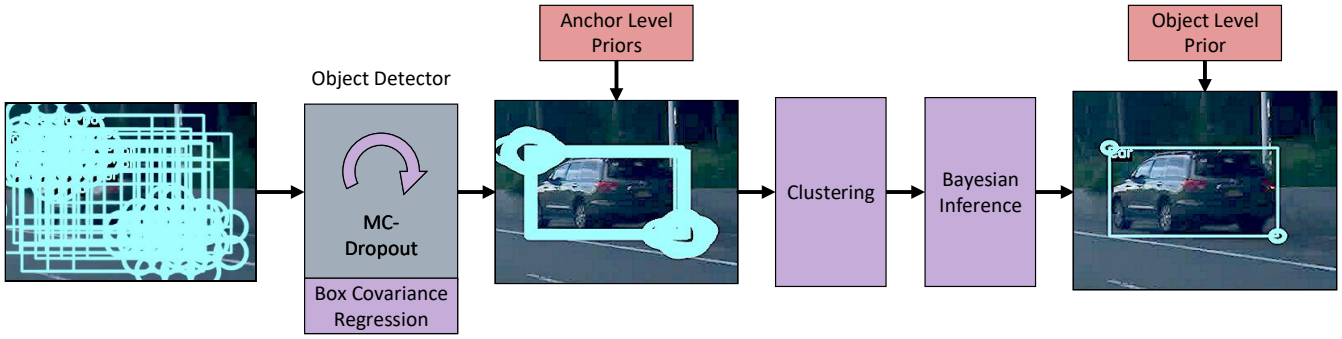

Fig. 2. The different stages of estimation employed in BayesOD, demonstrated on a test image frame from the BDD Dataset [16]. The additions by BayesOD to a standard object detector (**grey**) are shown in **purple**. Prior information is shown in **red**. **Left:** Prior bounding boxes. **Middle:** Object detector results after processing the prior boxes and incorporating anchor-level non-informative priors. **Right:** Final detection results after clustering and Bayesian Inference. Box corner covariance is visualized as in Fig. 1

II. Related Work

A. Deep Neural Networks For Object Detection

The object detection problem requires the estimation of both the category to which an object belongs, and its spatial location and extent, often expressed as the tightest fitting bounding box. The majority of state of the art object detectors in 2D [17] or in 3D [1], [2], [3] follow a standard algorithm, which maps a scene representation to object instances. Since the number of object instances in the scene is usually unknown a priori, the procedure begins with a densely sampled grid of prior object bounding boxes, referred to as *anchors* [18], [19], where the object detector provides a category and a bounding box estimate for each anchor element. Since multiple anchors can be mapped to a single bounding box in space, redundant outputs are eliminated through Non-Maximum Suppression. BayesOD builds on the RetinaNet 2D object detector [19].

B. Uncertainty Estimation In Deep Object Detectors

To account for epistemic uncertainty, Bayesian Neural Networks [20] usually apply a prior distribution over their parameters $\boldsymbol{\theta}$ to compute a posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ over the set of all possible parameters given the training dataset $\mathcal{D}$. A marginal distribution can then be computed for any prediction as:

$$p(\hat{\boldsymbol{y}}_i|\mathbf{x}_i,\mathcal{D}) = \int_{\theta} p(\hat{\boldsymbol{y}}_i|\mathbf{x}_i,\mathcal{D},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}, \quad (1)$$

where $\mathbf{x}_i$ is the input, and $\hat{\boldsymbol{y}}_i$ is the output of the neural network. Unfortunately, the calculation of the integral in Eq. (1) is usually intractable due to the non-linear activation function between consecutive layers [21]. Tractable approximations can be derived through Monte-Carlo integration by using ensemble methods [22] or Monte Carlo (MC) Dropout [6].

To estimate the epistemic uncertainty in the output of deep object detectors, Miller *et al*. [9] directly applies MC Dropout, treating the deep object detector as a **black box**. Uncertainty is then estimated as sample statistics from spatially correlated detector outputs. Subsequent work [10] studied the effect of various correlation and merging algorithms on the quality of the estimated uncertainty measures from the black box method in [9]. The black box method is shown to provide weakly correlated estimates for bounding box uncertainty, mainly because it observes the output bounding box after NMS, where most of the information from redundant predictions has already been removed.

Kendall *et al*. [7] provides one of the first works to address the estimation of aleatoric uncertainty for computer vision tasks. For regression tasks, a log likelihood loss is used to estimate heteroscedastic aleatoric uncertainty, written for every regression target as:

$$L_{reg}(\mathbf{x},\boldsymbol{\theta}) = \frac{1}{2\sigma(\mathbf{x},\boldsymbol{\theta})^2}||\mathbf{y}-f(\mathbf{x},\boldsymbol{\theta})||_2^2 + \frac{1}{2}\log\sigma(\mathbf{x},\boldsymbol{\theta})^2, \quad (2)$$

where $\mathbf{x}$ is the input to, and $f(\mathbf{x},\boldsymbol{\theta})$ is the output from the neural network. Furthermore $\mathbf{y}$ is the ground truth regression target, $||.||_2$ is the L_2 norm, $\boldsymbol{\theta}$ are the neural network parameters, and $\sigma(\mathbf{x},\boldsymbol{\theta})$ is the **estimated** output variance.

Le *et al*. [15] directly apply the formulation in Eq. (2) to estimate the diagonal elements of the covariance matrix of the bounding box output from object detectors. Such methods are referred to as **sampling free** and require only a single run of the deep object detector to estimate uncertainty. The estimated variance in Eq. (2) has also been used in [11], [12], [14] to increase average precision, by incorporating it in the non-maximum suppression stage, while disregarding the quality of the output uncertainty. The proposed sampling free methods assume a diagonal covariance matrix and still use NMS to eliminate low scoring predictions, reducing the quality of their estimated uncertainty for both objects' bounding box and category.

Le *et al*. [15] estimate aleatoric uncertainty in deep object detectors by exploiting **anchor redundancy**, where multiple per-anchors predictions map to the same object. These predictions are clustered using spatial affinity *before NMS*, and uncertainty measures are estimated using the cluster associated with every output prediction. Finally, a straightforward extension of [7] is typically used to perform **joint estimation of epistemic and aleatoric** uncertainty in deep object detectors [13], [23], while still employing NMS to eliminate rather than fuse information from redundant anchors.

Unlike each of the existing methods, BayesOD replaces

NMS with Bayesian inference significantly improving the quality of its uncertainty estimates. In addition, BayesOD is the first method to tackle fusion of the category from redundant output anchors, as well as to provide a multivariate extension of Eq. (2) to estimate the aleatoric uncertainty of objects' bounding boxes.

III. A Bayesian Formulation For Object Detection:

Throughout this section, the bounding box of an object, represented by its top left and bottom right corners, is denoted as $\mathcal{B}$, whereas its category, represented by a one-hot vector, is denoted as $\mathcal{S}$. The index i is used to signify a variable related to the i^{th} anchor in the anchor grid. Variables not indexed with i represent inference output clustered over several anchors. Finally, predictions provided by the neural network are denoted with a $\hat{.}$ operator.

A. Computing The Per-Anchor Gaussian Posterior:

Computing the uncertainty in the estimated per-anchor bounding box: Following [7] and using MC-Dropout as a tractable approximation of the integral in Eq. (1), the sufficient statistics of the Gaussian marginal probability distribution describing the estimated per-anchor bounding box $\hat{\mathcal{B}}_i \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i))$ can be derived as:

$$\boldsymbol{\mu}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_i, \boldsymbol{\theta}_t) \tag{3}$$

$$\Sigma_e(\mathbf{x}_i) = \frac{1}{T}\left(\sum_{t=1}^{T} f(\mathbf{x}_i, \boldsymbol{\theta}_t) f(\mathbf{x}_i, \boldsymbol{\theta}_t)^\intercal\right) - \boldsymbol{\mu}(\mathbf{x}_i)\boldsymbol{\mu}(\mathbf{x}_i)^\intercal, \tag{4}$$

where T is the number of times MC-Dropout sampling is performed, and $f(\mathbf{x}_i, \boldsymbol{\theta}_t)$ is the bounding box regression output of the neural network for the t^{th} MC-Dropout run. The covariance matrix, Σ_e, captures the epistemic uncertainty in the estimated bounding box $\hat{\mathcal{B}}_i$.

Eq. (3) is sufficient to compute the output mean of the per-anchor bounding box $\hat{\mathcal{B}}_i$. However, Eq. (4) still needs to account for the aleatoric component of uncertainty, where the final per-anchor output covariance $\Sigma(\mathbf{x}_i)$ can be approximated as:

$$\Sigma(\mathbf{x}_i) = \Sigma_e(\mathbf{x}_i) + \frac{1}{T}\sum_{t=1}^{T} \Sigma_a(\mathbf{x}_i, \boldsymbol{\theta}_t). \tag{5}$$

To estimate the full covariance matrix $\Sigma_a(\mathbf{x}_i)$, a novel multivariate log likelihood regression loss is derived as:

$$L_{mv}(\mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{2}(f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^\intercal \Sigma_a(\mathbf{x}_i, \boldsymbol{\theta})^{-1}(f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i) + \frac{1}{2}\log\det\Sigma_a(\mathbf{x}_i, \boldsymbol{\theta}), \tag{6}$$

where $\Sigma_a(\mathbf{x}_i, \boldsymbol{\theta})$ is the predicted per-anchor aleatoric covaraince matrix, $f(\mathbf{x}_i, \boldsymbol{\theta})$ is the predicted per-anchor bounding box, and $\mathbf{y}_i$ is the associated regression target. However, the loss in Eq. (6) is found to be numerically unstable. Furthermore, there are no guarantees on the positive definiteness of the predicted covariance matrix $\Sigma_a(\mathbf{x}_i, \boldsymbol{\theta})$. Using the LDL decomposition of $\Sigma_a(\mathbf{x}_i, \boldsymbol{\theta}) = L(\mathbf{x}_i, \theta)D(\mathbf{x}_i, \theta)L(\mathbf{x}_i, \theta)^\intercal$, in conjunction with the Cauchy-Schwarz inequality, a numerically stable surrogate loss function is derived as:

$$L_{mv}(\mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{2}||L(\mathbf{x}_i, \boldsymbol{\theta})^{-1}||_F^2 ||D(\mathbf{x}_i, \boldsymbol{\theta})^{-\frac{1}{2}}(f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)||_2^2 + \frac{1}{2}tr(\log D(\mathbf{x}_i, \boldsymbol{\theta})), \tag{7}$$

where $L(\mathbf{x}_i, \boldsymbol{\theta})$ is a lower triangular matrix with ones for its diagonal entries, and $D(\mathbf{x}_i, \boldsymbol{\theta})$ is a diagonal matrix. The loss function in Eq. (7) is a numerically stable upper bound of the one in Eq. (6) and can guarantee the positive definiteness of $\Sigma_a(\mathbf{x}_i, \boldsymbol{\theta})$ by predicting positive values for the diagonal elements of $D(\mathbf{x}_i, \boldsymbol{\theta})$ through standard activation functions. The final output distributions after incorporating both epistemic and aleatoric covariance estimates are plotted as bounding boxes in the middle image of Fig. 2.

Incorporating per-anchor bounding box priors: The per-anchor bounding box prior is usually defined based on the training dataset $\mathcal{D}$ as $p(\mathcal{B}|\mathbf{x}_i) \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$. The per-anchor posterior distribution describing the bounding box $\mathcal{B}$ can then be written as:

$$p(\mathcal{B}|\mathbf{x}_i, \mathcal{D}, \hat{\mathcal{B}}_i) \propto p(\hat{\mathcal{B}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{B})p(\mathcal{B}|\mathbf{x}_i, \mathcal{D}). \tag{8}$$

$p(\hat{\mathcal{B}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{B})$ is a Gaussian likelihood function described by the sufficient statistics $[\boldsymbol{\mu}(\mathbf{x}_i), \Sigma(\mathbf{x}_i)]$ in equations Eq. (3) and Eq. (5). The sufficient statistics can be computed through the multivariate Gaussian conjugate update, as:

$$\Sigma'(\mathbf{x}_i) = (\Sigma_0^{-1} + \Sigma(\mathbf{x}_i)^{-1})^{-1} \tag{9}$$

$$\boldsymbol{\mu}'(\mathbf{x}_i) = \Sigma'(\mathbf{x}_i)(\Sigma_0^{-1}\boldsymbol{\mu}_0 + \Sigma(\mathbf{x}_i)\boldsymbol{\mu}(\mathbf{x}_i)). \tag{10}$$

The choice of anchor priors depends on the application, and whether object information is actually available a priori. Since no useful bounding box information is available from our 2D training datasets, a non-informative prior, visually shown in the left image of Fig. 2, is chosen for $\mathcal{B}$ following [24].

B. Computing The Per-Anchor Categorical Posterior:

Computing the uncertainty in the estimated per-anchor category: Since the neural network outputs the parameters of a Categorical distribution rather than one-hot categorical samples, the parameters for the Categorical marginal conditional probability distribution $\hat{\mathcal{S}}_i \sim Cat([\hat{p}_1 \dots \hat{p}_K])$ can be computed as:

$$\hat{p}_k = \frac{1}{T}\sum_{t=1}^{T} \text{SoftMax}(g(\mathbf{x}_i, \boldsymbol{\theta}_t))_k, \tag{11}$$

where SoftMax(.) is the soft max function, and $g(\mathbf{x}_i, \boldsymbol{\theta}_t)_k$ is the output *logit* of the k^{th} category, estimated at the t^{th} MC-Dropout run of the neural network. No explicit

treatment of the aleatoric classification uncertainty is performed, since it is already contained within the estimated parameters $[\hat{p}_1 \dots \hat{p}_K]$ [14].

Incorporating per-anchor category priors: For the object category, a Dirichlet distribution is set as a prior over the parameters $\mathcal{P}$ of the categorical distribution $Cat(\mathcal{P})$ generating $\mathcal{S}$, instead of incorporating a prior distribution directly over the category $\mathcal{S}$. The posterior distribution of the categorical parameters can be written as:

$$p(\mathcal{P}|\mathbf{x}_i, \mathcal{D}, \hat{\boldsymbol{Z}}_i) \propto p(\hat{\boldsymbol{Z}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{P}) p(\mathcal{P}|\mathbf{x}_i, \mathcal{D}), \tag{12}$$

where $\mathcal{P}$ is the set of updated parameters $[p'_1, \dots, p'_K]$, and $\hat{\boldsymbol{Z}}_i = [\hat{\boldsymbol{z}}_1, \dots, \hat{\boldsymbol{z}}_H]$ are H **i.i.d.** samples from $Cat([\hat{p}_1, \dots, \hat{p}_K])$. Since the likelihood function $p(\hat{\boldsymbol{Z}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{P})$ is a categorical distribution, the prior distribution $p(\mathcal{P}|\mathbf{x}_i, \mathcal{D})$ is chosen to be a Dirichlet distribution allowing a Dirichlet posterior to be computed in closed form as:

$$\begin{aligned} p(\mathcal{P}|\mathbf{x}_i, \mathcal{D}, \hat{\boldsymbol{Z}}_i) &\propto \prod_{k=1}^{K} p_k^{\alpha_k - 1} \prod_{h=1}^{H} \prod_{k=1}^{K} p_k^{\hat{z}_{hk}=1} \\ &= Dir(\alpha'_1, \dots, \alpha'_K), \end{aligned} \tag{13}$$

where $\hat{z}_{hk}$ is the element in instance $\hat{\boldsymbol{z}}_h$ corresponding to category k, and $[\alpha'_k = \alpha_k + \sum_{h=1}^{H} \hat{z}_{hk} \;\forall\; k = 1, \dots, K]$ are the inferred parameters of the Dirichlet posterior distribution. The per-anchor categorical posterior distribution can be written as:

$$p(\mathcal{S}|\mathbf{x}_i, \mathcal{D}, \boldsymbol{Z}_i) = Cat([p'_1, \dots, p'_K]), \tag{14}$$

where p'_k is the mean of the Dirichlet posterior distribution [24] in Eq. (13) written as:

$$p'_k = \frac{\alpha'_k}{\sum_{j=1}^{K} \alpha'_j}.$$

Similar to the prior used for the per-anchor bounding box, we choose a non-informative Dirichlet prior for the per-anchor category following [24]. Although non-informative, the prior still serves an essential purpose by allowing the derivation of a Dirichlet posterior in Eq. (13), which will allow the fusion of information from multiple clustered categorical variables in the next section.

C. Bayesian Inference as a Replacement to NMS:

Similar to NMS, BayesOD clusters per-anchor outputs from the neural network using spatial affinity. However, all elements in the cluster are then combined regardless of their classification score during inference. Greedy clustering is chosen as it provides adequate performance when compared to standard NMS, while maintaining computational efficiency. For better performing but slower clustering algorithms, see [10].

For the remainder of this section, we will continue the derivation for a single anchor cluster containing M anchors. The anchor with the highest categorical score is considered the cluster's center, is indexed by 1, and is described with the posterior distributions in Eq. (8) and Eq. (12). The rest of the cluster members are assumed to be measurement outputs from the neural network described by the states $\hat{\mathcal{S}}_i$ and $\hat{\mathcal{B}}_i$, and are used to update the bounding box and category of the cluster center. Specifically, the final posterior distribution describing an object's bounding box is:

$$\begin{aligned} p(\mathcal{B}|\mathcal{X}, \mathcal{D}, [\hat{\mathcal{B}}_1, \dots, \hat{\mathcal{B}}_M]) &\propto p(\mathcal{B}|\mathbf{x}_1, \mathcal{D}, \hat{\mathcal{B}}_1) \prod_{i=2}^{M} p(\hat{\mathcal{B}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{B}) \\ &= \mathcal{N}(\boldsymbol{\mu}''(\mathcal{X}), \Sigma''(\mathcal{X})), \end{aligned} \tag{15}$$

where $\mathcal{X}$ is the set of inputs $[\mathbf{x}_i \mid i = 1 \dots M]$ corresponding to the M cluster members, $p(\mathcal{B}|\mathbf{x}_1, \mathcal{D}, \hat{\mathcal{B}}_1)$ is the per-anchor posterior distribution of the cluster center, and $p(\mathcal{B}|\mathbf{x}_1, \mathcal{D}, \hat{\mathcal{B}}_1) \prod_{i=2}^{M} p(\hat{\mathcal{B}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{B})$ is the likelihood derived through a conditional independence assumption of the $\hat{\mathcal{B}}_i$ of the cluster members given $\mathcal{B}$. The sufficient statistics of Eq. (15) can be estimated in closed form as:

$$\Sigma''(\mathcal{X}) = \left(\sum_{i=1}^{M} \Sigma'(\mathbf{x}_i)^{-1} \right)^{-1} \tag{16}$$

$$\boldsymbol{\mu}''(\mathcal{X}) = \Sigma''(\mathcal{X}) \left(\sum_{i=1}^{M} \Sigma'(\mathbf{x}_i)^{-1} \boldsymbol{\mu}'(\mathbf{x}_i) \right), \tag{17}$$

where $\boldsymbol{\mu}'(\mathbf{x}_i), \Sigma'(\mathbf{x}_i)$ are the sufficient statistics of the per anchor posterior distribution derived in Eq. (8).

To arrive at the final posterior distribution describing the category $\mathcal{S}$, a similar analysis can be performed to update the sufficient statistics $\mathcal{P}$ of the cluster center with categorical measurements $[\hat{\boldsymbol{Z}}_2, \dots, \hat{\boldsymbol{Z}}_m]$ of the rest of the cluster members. Specifically, the posterior probability of $\mathcal{P}$ can be derived as:

$$\begin{aligned} p(\mathcal{P}|\mathbf{x}_i, \mathcal{D}, [\hat{\boldsymbol{Z}}_1, \dots, \hat{\boldsymbol{Z}}_M]) &\propto p(\mathcal{P}|\mathbf{x}_1, \mathcal{D}, \hat{\boldsymbol{Z}}_1) \prod_{i=2}^{M} p(\hat{\boldsymbol{Z}}_i|\mathbf{x}_i, \mathcal{D}, \mathcal{P}) \\ &= Dir(\alpha''_1, \dots, \alpha''_K) \end{aligned} \tag{18}$$

where $\alpha''_k = \alpha'_k + \sum_{i=2}^{M} \sum_{h=1}^{H} \hat{z}_{ihk} \;\forall\; k = 1 \dots K$, and the categorical measurements $[\hat{\boldsymbol{Z}}_2, \dots, \hat{\boldsymbol{Z}}_m]$ are assumed to be **i.i.d**. In summary, α''_k is derived by updating the per-anchor Dirichlet posterior distribution in (12) of the cluster center with index $i = 1$ with categorical measurements $\boldsymbol{Z}_2, \dots, \boldsymbol{Z}_M$ from all cluster members. The final categorical distribution describing the state $\mathcal{S}$ is then:

$$p(\mathcal{S}|\mathcal{X}, \mathcal{D}, [\hat{\boldsymbol{Z}}_1, \dots, \hat{\boldsymbol{Z}}_M]) = Cat(p''_1, \dots, p''_K), \tag{19}$$

where $[p''_1, \dots, p''_K]$ is computed as the mean of the posterior distribution in Eq. (18):

$$p''_k = \frac{\alpha''_k}{\sum_{j=1}^{K} \alpha''_j}, \tag{20}$$

Note that every member of the cluster contributes to the estimation of the final bounding box and category states of the object. Furthermore, the output distributions for both the category and bounding box can be updated with object-level priors using the same equations presented in sections III-A and III-B. The final output from BayesOD is shown as the rightmost image in Fig. 2.

IV. Experiments and Results

To show the effectiveness of BayesOD in comparison to the state of the art, it is applied to the problem of 2D object detection in image space. The evaluation is based on four commonly used datasets:

- **Berkeley Deep Drive 100K Dataset (BDD)** [16] road scene dataset, with $80K$ frames used according to the official $70K/10K$ training/validation split. Models trained on BDD are also tested on $7,481$ frames of **KITTI** [8]. Both datasets contain 7 common road scene object categories.
- **MS COCO** [25] dataset, with $223K$ frames that contain instances from 81 different object categories, and an official $118K/5K$ training/testing split. Models trained on COCO are also tested on $5,823$ frames from **Pascal VOC** [26], which shares 20 object categories with the COCO dataset.

Models used for testing are not allowed to observe instances from the KITTI or Pascal VOC datasets.

All baseline uncertainty estimation methods used in comparison are integrated into the inference process of RetinaNet [19], trained using the regression loss function in Eq. (2) to estimate a diagonal bounding box covariance matrix. Full aleatoric covariance matrix results are provided through a second RetinaNet model, trained using the proposed regression loss in Eq. (7). For additional information on RetinaNet's training procedure and hyperparamters, see [19].

A. Evaluation Metrics

Three evaluation metrics are used to quantify the performance of uncertainty estimation methods in comparison to BayesOD. For performance on the detection task, we use the **Mean Average Precision (mAP)** [25], [26], [16], [8] at 0.5 IOU. The maximum mean average precision achievable by a detector is 100%.

The **Minimum Uncertainty Error (MUE)** [10] at 0.5 IOU is used to determine the ability of the detector's estimated uncertainty to discriminate true positives from false positives. The lowest MUE achievable by a detector is 0%. We define the Gaussian MUE (GMUE) when the Gaussian entropy is used, Categorical MUE (CMUE) when the Categorical entropy is used. Finally, we average the GMUE and CMUE over all categories in a testing dataset to arrive to a single value, the Mean (Gaussian or Categorical) MUE (mGMUE or mCMUE).

Finally, we use the newly proposed **Probability Based Detection Quality (PDQ)** [27] to jointly quantify the bounding box and category probability assigned to true positives by the detector. The highest PDQ achievable by a detector is 100%, where the PDQ increases as the distributions assigned to a detection better match those of the ground truth instance. For detailed information on the three evaluation metrics, we refer the reader to the [26], [10], [27].

B. Comparison With State of The Art Methods:

BayesOD is compared against four approaches representing the state of the art methods for uncertainty estimation methods used for object detection. The four approaches are referred to as: *Black Box* [9], [10], *Sampling Free* [15], [14], *Anchor Redundancy* [15], and *Joint Aleatoric Epistemic* [13]. BayesOD, Black Box, and Joint Aleatoric Epistemic use 10 stochastic runs of MC-Dropout, while Sampling Free and Anchor Redundancy use only one non-stochastic run. As such, BayesOD, Black Box, and Joint Aleatoric Epistemic run at a similar frame rate, approximately $4\times$ slower than Sampling Free and Anchor Redundancy. The affinity threshold used for clustering in all methods was set to the 0.5 IOU, similar to that used for NMS in RetinaNet. The number of categorical samples H in Eq. (12) is empirically set to 30.

Table I shows the results of evaluating the four methods in comparison to BayesOD, on the four testing datasets. BayesOD is seen to outperform all four methods on mAP when tested on the BDD, COCO and PASCAL VOC datasets by a margin of $0.57\% - 1.7\%$ over the second best method, but is outperformed on the KITTI dataset by $\sim 1.5\%$ when using the Sampling Free and Anchor Redundancy methods. Such reduction in performance on KITTI is noted with all methods using MC-Dropout, implying that MC-Dropout might hurt mAP performance in cases where the testing dataset is semantically different than the training dataset.

Similarly, BayesOD also outperforms all four methods on PDQ when tested on the BDD, KITTI and COCO datasets by a margin of $0.07\% - 3.00\%$ over the second best method. BayesOD is outperformed on the PASCAL VOC dataset by 0.95% when using the sampling free method. Considering the performance only on PDQ, it cannot be determined if a method is assigning lower probability values to false positives.

On the other hand, the mGMUE/mCMUE are capable of providing a quantitative measure of how well the estimated uncertainty can be used to separate correct and incorrect detections [10]. BayesOD provides a significant reduction of $9.77\% - 13.13\%$ in mGMUE over the next best method on all four testing datasets. Combined with BayesOD's performance on the PDQ metric, it can be inferred that BayesOD not only assigns adequate probability to true positives, but also assigns a lower probability to false positives when compared to true positives. Finally, when comparing mCMUE, BayesOD provides a reduction between $1.63\% - 5.23\%$ over the next best method on all four datasets.

C. Ablation Studies:

Table II shows the results of the mAP, PDQ, mGMUE, and mCMUE for the ablation studies performed on the COCO dataset. The results of the full BayesOD framework can

Training Dataset	Testing Dataset	Method	mAP(%) ↑	PDQ Score(%) ↑	mGMUE(%) ↓	mCMUE(%) ↓
BDD	BDD	Sampling Free	36.59	33.97	44.19	28.46
		Black Box	36.43	32.46	47.63	30.45
		Anchor Redundancy	32.92	29.57	48.56	35.58
		Joint Aleatoric-Epistemic	36.84	29.57	46.35	28.28
		BayesOD	**38.14**	**36.79**	**34.42**	**24.85**
BDD	Kitti	Sampling Free	64.78	29.24	46.70	20.67
		Black Box	62.96	32.26	49.23	22.27
		Anchor Redundancy	**64.83**	29.57	48.56	35.58
		Joint Aleatoric-Epistemic	62.96	29.57	46.35	28.28
		BayesOD	63.34	**35.26**	**30.06**	**15.58**
COCO	COCO	Sampling Free	31.89	22.43	40.39	25.76
		Black Box	33.71	21.87	45.26	28.68
		Anchor Redundancy	29.94	17.63	43.74	31.13
		Joint Aleatoric-Epistemic	32.68	23.08	42.90	26.51
		BayesOD	**35.41**	**23.15**	**30.23**	**24.13**
COCO	Pascal VOC	Sampling Free	54.94	**14.18**	49.49	29.63
		Black Box	54.67	12.77	48.90	29.42
		Anchor Redundancy	51.56	13.06	48.67	39.64
		Joint Aleatoric-Epistemic	55.43	11.62	49.99	30.14
		BayesOD	**56.00**	13.23	**36.36**	**24.19**

TABLE I

THE RESULTS OF THE EVALUATION OF *Sampling Free* [14], [15], *Black Box* [9], [10], *Anchor Redundancy* [15], AND JOINT ALEATORIC-EPISTEMIC [13], [23] STATE OF THE ART METHODS COMPARED TO BAYESOD.

be seen in experiment #1. By analyzing the results of the ablation studies, the following claims are put forth:

Learning the off-diagonal elements of the covariance matrix provides slightly better uncertainty estimates for the objects' bounding box. To support this claim, RetinaNet is trained using the original log likelihood loss in Eq. (2) instead of the proposed multivariate loss in Eq. (7). The results of BayesOD using this original loss formulation are shown in experiment #2. When compared to the full system, an increase of 0.48% is observed in mGMUE. Although the improvement is not substantial, the new proposed loss avoids an explicit independence assumption and allows the neural network to learn to drive the off-diagonal elements of the covariance matrix towards 0 if needed.

Aleatoric uncertainty provides a more discriminative uncertainty estimate for the objects' bounding box over epistemic uncertainty estimated from MC-Dropout. To support this claim BayesOD is implemented without the update step in Eq. (5), to use only the per-anchor sample variance computed from multiple stochastic runs of MC-Dropout. The results, presented in experiment #3, show an increase of 5.65% and 2.34% is observed in the mGMUE and mCMUE respectively. Note however that this conclusion is specific to MC-Dropout, and might not be valid for alternative epistemic uncertainty estimation mechanisms.

#	Experiment	mAP(%) ↑	PDQ Score(%) ↑	mGMUE(%) ↓	mCMUE(%) ↓
1	Full System	**35.41**	**23.15**	30.23	**24.13**
2	Diagonal Covariance	34.77	22.64	30.69	25.25
3	Epistemic Only	34.15	22.62	35.88	26.47
4	Aleatoric Only	34.12	22.67	**28.95**	25.60
5	Standard NMS	34.70	22.65	43.19	25.10

TABLE II

THE RESULTS OF ABLATION STUDIES PERFORMED ON BAYESOD USING THE COCO DATASET FOR TRAINING AND TESTING.

To provide better insight on the effect of epistemic uncertainty from MC-Dropout on the full system, experiment #4 is performed by using BayesOD with a single inference run, and without any epistemic uncertainty estimation mechanism. The results show a decrease in mGMUE of 6.93% over experiment #3, and 1.28% over the full system, further cementing the conclusion that MC-Dropout might not be a good method to estimate epistemic uncertainty in deep object detectors.

Greedy Non-Maximum Suppression is detrimental to the discriminative power of the uncertainty in the objects' bounding box. To support this claim, the elimination scheme of NMS is selected to retain only cluster centers, while discarding the remaining cluster members. The results presented in experiment #5 show a large increase of 12.96% mGMUE when compared to the full system. We conclude that merging information from all cluster members into the final object estimate is essential for proper quantification of bounding box uncertainty by a neural network.

V. CONCLUSION

This paper presents BayesOD, a Bayesian approach for estimating the uncertainty in the output of deep object detector. Experiments using BayesOD show that replacing NMS with Bayesian inference and explicitly incorporating full *aleatoric* covariance matrix estimation allows for a much more meaningful estimated category and bounding box uncertainty in deep object detectors. This work aims to pave the path for future research directions that would use BayesOD for active learning, exploration, as well as object tracking. Future work will study the effect of informative priors originating from multiple detectors, temporal information, and different sensors on the perception capabilities of a robotic system.

REFERENCES

[1] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[2] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Trupti M Pandit, PM Jadhav, and AC Phadke. Suspicious object detection in surveillance videos for security applications. In *Inventive Computation Technologies (ICICT), International Conference on*, 2016.

[5] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

[7] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, 2017.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[9] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[10] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. *arXiv preprint arXiv:1809.06006*, 2018.

[11] Gregory P Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K Wellington. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12677–12686, 2019.

[12] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2888–2897, 2019.

[13] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

[14] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Leveraging heteroscedastic aleatoric uncertainties for robust real-time lidar 3d object detection. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

[15] Michael Truong Le, Frederik Diehl, Thomas Brunner, and Alois Knol. Uncertainty estimation for deep neural object detectors in safety-critical applications. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018.

[16] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[17] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, 2015.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. In *Network: Computation In Neural Systems*, pages 469–505, 1995.

[21] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NuerIPS)*, pages 6402–6413. 2017.

[23] Florian Kraus and Klaus Dietmayer. Uncertainty estimation in one-stage object detection. *arXiv preprint arXiv:1905.10296*, 2019.

[24] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[27] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic Object Detection: Definition and Evaluation. *arXiv e-prints*, page arXiv:1811.10800, Nov 2018.