Aleatoric Uncertainty with Test-Time Augmentation for Object Detection in Autonomous Driving

Rui Miguel Campos Magalhães
rui.magalhaes@tecnico.ulisboa.pt
Instituto Superior Técnico, Lisboa, Portugal
November 2022

***Abstract*—Autonomous driving relies on various complex systems to perform essential tasks in the automated driving scenario. A key task for environment perception is camera-based object detection. Recent advances in the field of computer vision have made the use of deep convolutional neural networks the state-of-the-art for object detection tasks. For safety-critical systems, such as autonomous driving, it is essential to measure how reliable the estimated output detections are. Accurate quantification of the uncertainty associated with each detection can provide safer and more reliable autonomous driving. In this work, we research novel approaches to characterize uncertainty in deep learning detection methods, as well as explore improvements to the current state-of-the-art methods. We propose the first Test-Time Augmentation (TTA) method for estimating aleatoric uncertainty in object detection. For this purpose, we develop a novel TTA pipeline and show that it is capable of outperforming the current state-of-the-art methods, Monte Carlo (MC) Dropout and Output Redundancy, in the quality of predicted distributions and estimated uncertainty for both the classification and regression task. Studies are carried out to investigate the use of a bounding box selection criterion and two different Intersection-over-Union (IOU) thresholds for each task (classification and regression). We show that improvements in performance in the MC Dropout and Output Redundancy methods can be obtained by applying an optimal bounding box selection criterion and two different IOU thresholds for each separate task. The lower IOU thresholds, used in the clustering step, are shown to generate the best results for the regression task, while the higher IOU thresholds produce the best results for the classification task.**

***Index Terms*—Uncertainty Quantification, Object Detection, Test-Time Augmentation, Autonomous Driving**

I. Introduction

Autonomous driving relies on various complex systems to perform essential tasks in the automated driving scenario. A key task for autonomous driving systems is camera-based object detection, which allows the detection of road users, such as vehicles and pedestrians, essential for a safe driving environment.

Object detection in camera-based inputs is the task of predicting both the position and type of multiple objects in a given image. For the past few years, advances in the field of computer vision have made the use of deep convolutional neural networks the state-of-the-art for the object detection task, with algorithms such as SSD [1], YOLO [2] and R-CNN [3].

A neural network object detector produces multiple bounding box predictions, each combined with a classification score. The shape and position of the bounding box are represented by four predicted coordinate values. For the category of an object, classification scores (with values ranging from 0 to 1) are produced for each different category. As multiple bounding boxes are predicted for each object, an algorithm known as Non-Maximum Supression (NMS) [4] is usually used to remove redundant detections.

The classification score for each predicted bounding box is often mistakenly treated as a confidence probability, when it is actually just a normalized network output. Gal in [5] noted that networks tend to yield overconfident predictions to falsely detected objects. Not only that, the classification scores are used as detection scores without considering the localization uncertainty. Current state-of-the-art object detectors have no measure of how certain they are in their predictions [6].

It is not sufficient to rely on the classification score alone. For safety-critical systems, such as autonomous driving, an accurate quantification of the uncertainty associated with each detection can provide safer and more reliable autonomous driving.

Gal in [7] defined two types of uncertainty in deep learning: aleatoric and epistemic. Epistemic, or model uncertainty, concerns the uncertainty related to the model itself, and aleatoric, or data uncertainty, is associated with intrinsic randomness present in the data.

Object detectors that can measure the associated uncertainty for each detection are known in the literature as Probabilistic Object Detectors. Several works have been developed in the field of Uncertainty Quantification (UQ) in object detection. We refer the reader to the work by Feng et al. [8], where a review of such detectors was performed.

Following the need for probabilistic object detectors for safer and more reliable autonomous driving, the main goal of this work is to research novel approaches to characterize uncertainty in deep learning detection methods. The improvement of current state-of-the-art methods in this field is also another objective, and these studies should hopefully lead to better decision making in the automotive automation industry.

This dissertation offers the following contributions:

- We propose the first TTA method for the estimation of aleatoric uncertainty in object detection. For this purpose, a novel TTA pipeline was developed.
- We evaluate how different color augmentations impact the quality of the estimated uncertainty and predictive

distributions with the TTA, in the domain of autonomous driving.
- We explore the use of two different IOU thresholds for each task (classification and regression) and its effects on the quality of the estimated uncertainties and distributions.
- We study the effect that different bounding box selection criteria (which define what boxes are used for uncertainty estimation) have on the quality of the estimated uncertainty and predictive distributions.

II. Theoretical Background

A. Practical Methods for Uncertainty Estimation

Following the literature review of probabilistic object detectors in [8], the different methods used for the task of uncertainty estimation in object detection are: Direct Modeling, MC Dropout, Output Redundancy and Deep Ensembles. As only the MC Dropout and Output Redundancy are used in this dissertation, solely an explanation for those methods is provided.

*1) **Output Redundancy**:* Output Redundancy was developed by Le et al. in [9] to model aleatoric uncertainty without sampling by iterating all detection proposals. This technique replaces the NMS post-processing step with spatial clustering of redundant output detections, with IOU as an affinity measure between detections. IOU measures the overlap between two two-dimensional boxes. Then, by computing the sample mean and variance among the members of each cluster, category and bounding box uncertainty estimates can be generated that describe each output detection.

*2) **Monte Carlo Dropout**:* The MC Dropout is a sampling-based method that uses dropout layers to model epistemic uncertainty. It was developed by Gal and Ghahramani in [5] for UQ in image classification and regression tasks. Miller et al. extended this method for the object detection task for the first time in [10].

Dropout is a technique that randomly deactivates neurons in a layer. By retaining dropout layers while testing, multiple forward passes can be performed to generate different output predictions by stochastically deactivating neurons for each forward-pass. The variance of the predictions can be used to estimate epistemic uncertainty. In the literature, it has been used for the object detection task by Miller et al. in several works [10]–[12] to estimate epistemic uncertainty, as well as jointly used with the direct modeling method in [13], [14] to estimate both aleatoric and epistemic uncertainty simultaneously.

B. Test-Time Augmentation

The aforementioned methods are those used so far in the literature for UQ in object detection. However, different UQ methods have also been used in deep learning for other tasks [15]. One of these methods is TTA, which was previously used to quantify aleatoric uncertainty in image segmentation [16] and image classification [17], mainly in the biomedical domain.

The TTA method is an application of data augmentation to the test dataset. Multiple augmented versions of each image are created, and predictions are performed for each version, creating an ensemble of those predictions. Tipically, the method is used to improve predictive results. However, by measuring how diverse the predictions for a given image are, aleatoric uncertainty estimations can be performed [16], [17].

For the object detection task, TTA has been used in practical applications to improve mAP and recall metrics [18]. By providing augmented versions of each image, the detector has a better chance of correctly identifying an object and, thus, improving performance.

However, TTA has not yet been applied to UQ for object detection in the literature. One of the main contributions of this dissertation is the use of TTA for the first time to estimate aleatoric uncertainty in object detection. Studies will also be performed to evaluate how different color transformations impact the quality of the estimated uncertainty in the domain of autonomous driving.

C. Clustering Techniques and Affinity Measures

For Output Redundancy [9] and sampling-based methods (Deep Ensembles [12] and MC Dropout [10]–[14]), clustering of redundant detections is the established technique for uncertainty estimation in object detection.

Miller et al. [11] evaluated various clustering techniques and affinity measures. They found that a Basic Sequential Algorithmic Scheme (BSAS) with IOU and *Same Label* as affinity measures produced the highest uncertainty quality. As in [12], this technique will be treated as the current established merging strategy for the remainder of this dissertation.

Using IOU as a spatial affinity measure, a minimum IOU score must be met between detections for them to be clustered together. Although experiments with various IOU minimum thresholds were performed in [11], no works in the literature have explored the use of two different IOU minimum thresholds for each separate task (regression and classification). In this dissertation, we will use two different IOU minimum thresholds for each task and explore its impact on the quality of the estimated uncertainties.

D. Bounding Box Selection Criteria

After each forward-pass in sampling-based methods, from the total output predictions, only some are kept. The kept predictions from each run are accumulated and used to estimate the uncertainty. For the remainder of this dissertation, the technique or rule that decides which predictions are kept after each run is called a bounding box selection criterion.

The current established technique uses NMS as a criterion [10]–[12]. After each forward-pass, NMS is performed, and only the resulting predictions from NMS are accumulated and used for uncertainty estimation.

However, all output predictions from a single forward-pass can be used for uncertainty estimation [9]. We believe that limiting the number of output predictions by using NMS as a bounding box selection criterion can result in missing valuable information to estimate uncertainty. As no other works have been developed regarding the usage of different bounding box selection criteria, in this dissertation we propose to study the effect that different criteria can have on the quality of the estimated uncertainty.

E. Evaluation of Predictive Distributions and Uncertainty

Feng et al. in [8] remark that there is little agreement on how to evaluate probabilistic object detectors in the literature. There are a range of different metrics used for each method, the most used being Mean Average Precision (mAP), Probability-based Detection Quality (PDQ) and Minimum Uncertainty Error (MUE).

This dissertation will follow the work developed in [6], also using Negative Log Likelihood (NLL) for the evaluation of predictive distributions, as the well as Minimum Uncertainty Error (MUE) to assess the quality of the estimated uncertainty. The Mean Average Precision (mAP) metric will also be computed since it is the standard evaluation metric in object detection.

III. Methodology

In this section, we firstly present an overview of the original TTA method pipeline, followed by the proposed modifications that result in a novel TTA pipeline to quantify uncertainty in object detection, as well as an explanation regarding how predictive distributions are estimated.

A. Original Test-Time Augmentation Pipeline

In Figure 1, the original TTA pipeline is represented.

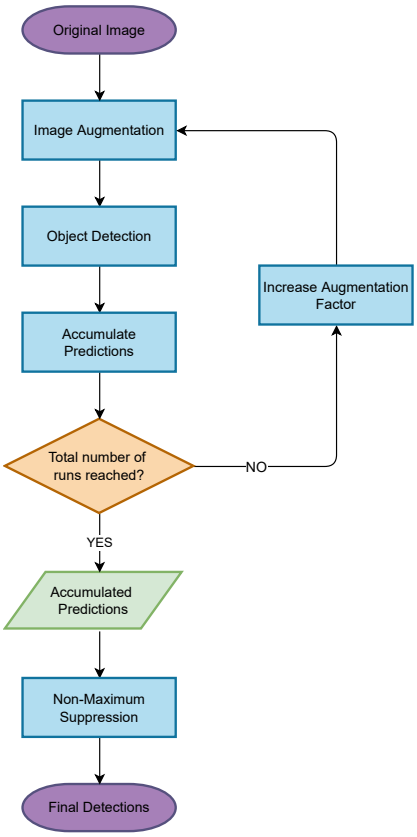

Fig. 1. Original TTA Pipeline

The first step in the pipeline is the image augmentation. The chosen augmentations used in this dissertation are contrast, gamma, brightness, and Gaussian blur.

In the second step, after the augmentation has been performed, the resulting augmented image will be provided as an input to the object detector, which will output predictions that describe possible objects in the image.

The third step in the pipeline is to accumulate all the predictions. Then, if the total number of runs has not yet been reached, the augmentation factor is increased, and the first three steps are performed again, resulting in more accumulated predictions.

When the total number of runs is reached, a group of accumulated predictions is formed from all of the runs. The final step is to perform the NMS algorithm on the accumulated predictions, which will remove redundant predictions and output the final object detections of the image.

B. Image Augmentations

There are different ways of performing augmentations in images, mainly with positional or color transformations. These should be chosen carefully to preserve the quality and domain of the dataset used.

In this dissertation, only color augmentations were performed, in which the values of the pixels in each image were modified.

For each augmentation, the range of parameter variation was chosen by visualizing the different transformations in the images from the used dataset and choosing the values that would not distort the image to a point where it was no longer realistic.

1) ***Constrast Augmentation****:* Increasing the contrast of an image will cause the brighter regions to become brighter and the dark regions of the image to become darker. Decreasing contrast will result in a smaller difference between the bright and dark regions of an image.

2) ***Gamma Augmentation****:* The gamma correction directly impacts the shadows in an image. With lower gamma factors, shadows become brighter, which can be beneficial to better define the contours of objects in an image. Higher gamma factors cause shadows to become darker.

3) ***Brightness Augmentation****:* Altering the brightness of an image will modify all pixels equally. When brightness is increased, both the dark and bright areas become brighter. Vice versa, decreasing the brightness will cause both bright and dark areas to become darker.

4) ***Gaussian Blur****:* A Gaussian blur will introduce a blur to the original image. In an autonomous driving setting, the occurrence of motion blur in images is very common due to car motion. This phenomenon is characterized by having a larger incidence of blur in the direction of the car's movement than in other directions. Although Gaussian blur is not able to completely reproduce motion blur, this augmentation will be used to provide a closer representation of images where blur is present.

C. Test-Time Augmentation Pipeline for Uncertainty Quantification in Object Detection

In this section, an overview of the modifications performed on the original TTA method is given. We extend the original pipeline of TTA to quantify aleatoric uncertainty for the first time in object detection.

In Figure 2, the novel pipeline can be visualized. By measuring how diverse the predictions are for a given image, aleatoric uncertainty estimations can be performed, as seen in [16], [17]. The extended steps added to the original pipeline are largely based on the established merging strategy technique developed by Miller et al. [11].

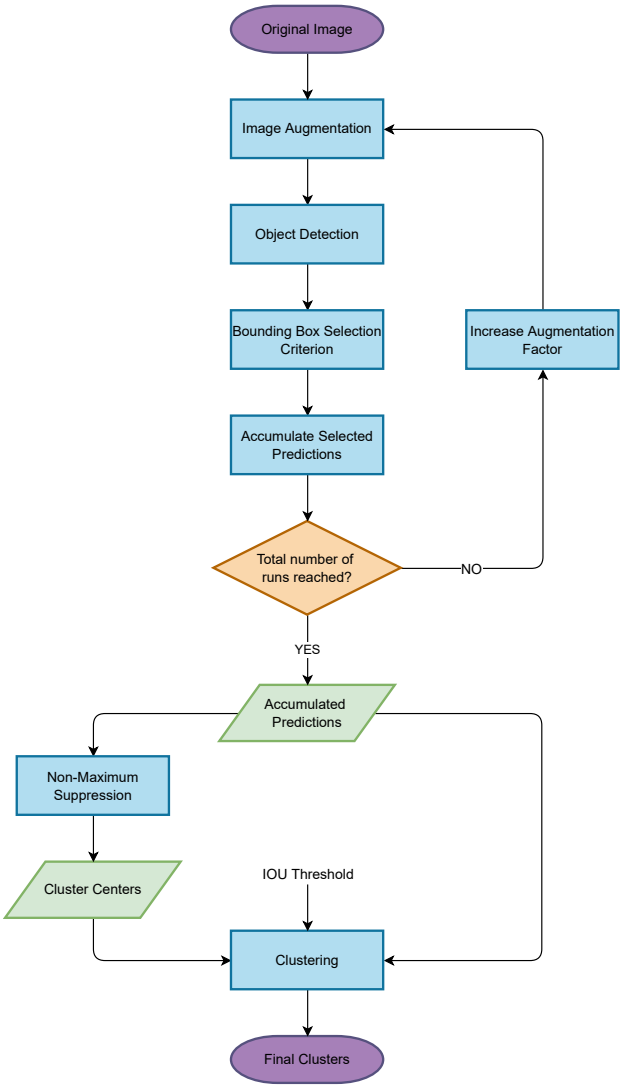

Fig. 2. Novel TTA Pipeline for UQ

The first two steps of the novel pipeline remain the same as those of the original pipeline. First, an image augmentation is performed. In the second step, the resulting augmented image will be sent as an input to the object detector, which will output predictions that describe possible objects in the image.

The third step is the first major difference between these two pipelines, where a bounding box selection criterion step is performed on the output predictions. The criteria used in this dissertation will be based on the maximum classification score obtained for a prediction. For example, with a bounding box selection criterion of 0.01, only predictions with a maximum classification score of at least 0.01 are kept.

The remaining predictions from the bounding box selection criterion step are accumulated. Then, as in the original pipeline, if the total number of runs has not yet been reached, the augmentation factor is increased, and the first four steps are performed again, resulting in more accumulated selected predictions.

When the total number of runs is reached, a group of accumulated predictions is formed from all of the runs. In the next step, the NMS algorithm is performed on the accumulated predictions, as in the original pipeline. Redundant predictions are removed, and the final object detections for the image are obtained.

The final step is the second major difference between the original and novel pipelines. Using the final detections as cluster centers, a clustering step is performed that will associate each final detection with a group of predictions. This clustering step is based on the established merging strategy technique developed by Miller et al. [11], which uses a BSAS with *Same Label* and IOU as affinity measures.

BSAS is a basic clustering algorithm that sequentially groups detections that meet a minimum affinity threshold. For each detection, if the affinity with any existing cluster center meets the minimum requirement, the detection joins the cluster. The *Same Label* affinity measure is a semantic affinity measure that requires a detection to have the same predicted class label as the cluster center. The IOU affinity measure is a spatial affinity measure that compares the location and shape between the detection and the cluster center.

Thus, in the last step of the novel pipeline, for each cluster center, the IOU score is calculated with all the accumulated predictions. Every prediction that predicts the same class label as the cluster center and also meets a minimum defined IOU threshold score with the cluster center joins the cluster. The output of this step will be the final clusters of predictions for every object present in the image.

D. Distribution Estimation

With the final clusters, we can estimate the predictive distributions for each detected object and measure the aleatoric uncertainty.

1) ***Categorical Distribution****:* Based on the work developed by Kendall and Gal [7] and Miller et al. [10], a single forward-pass generates a set of individual detections, each with bounding box coordinates b and a classification score vector s. Denoting these detections as $D_i = \{s_i, b_i\}$ for a single pass, performing multiple forward-passes creates a larger set $\mathbb{D} = \{D_1, ..., D_n\}$. These detections will be paired in groups of clusters and as per [7], [10], the vector class of probabilities q_i (for each final cluster) can be approximated by averaging all classification score vectors s_i of the bounding boxes that belong to that cluster.

$$q_i \approx s_i = \frac{1}{n}\sum_{j=1}^{n} s_j \tag{1}$$

Given an image $\mathcal{I}$ and the set of detections $\mathbb{D}$, the previous equation gives us an approximation of the probability of the class label y_i for a detected object in image $\mathcal{I}$ given all predicted detections in $\mathbb{D}$, which is described as a Categorical distribution parameterized by q_i, with k different classes:

$$p(y_i|\mathcal{I}, \mathbb{D}) \sim Cat(k, q_i) \tag{2}$$

For each detected object, the uncertainty in the classification task can be measured by computing the entropy $H(q_i) = \sum_{i=1}^{k} -q_i \cdot log(q_i)$. If a detection has high uncertainty associated to it, it is expected that the classification scores for each class are more evenly distributed in terms of mass probability (which causes a higher value of entropy) as well as a lower maximum score.

*2) **Multivariate Gaussian Distribution**:* Following the same work [7], [10], as in Section III-D1, the distribution over the bounding box coordinates can also be approximated by computing a covariance matrix Σ and averaging over the bounding box vectors b_i of all grouped detections in every single cluster:

$$\overline{b_i} = \frac{1}{n}\sum_{j=1}^{n} b_j, \quad \Sigma = \frac{1}{n}\sum_{j=1}^{n}(b_j - \overline{b_i})(b_j - \overline{b_i})^T \tag{3}$$

Therefore, given an image $\mathcal{I}$ and the set of detections $\mathbb{D}$, an approximation of the ground truth target coordinates z for a detected object in the image $\mathcal{I}$ given all predicted detections in $\mathbb{D}$ is described as a multivariate Gaussian distribution, parameterized with $\overline{b_i}$ and Σ:

$$p(z|\mathcal{I}, \mathbb{D}) \sim \mathcal{N}(\overline{b_i}, \Sigma) \tag{4}$$

For each detected object, the uncertainty in localization can be measured by computing the entropy $H(\mathcal{N}(\overline{b_i}, \Sigma)) = \frac{1}{2}\ln det(2\pi e\Sigma)$. The uncertainty is then correlated with the amount of variability, in terms of location, of the predicted coordinates of the bounding boxes present in each final cluster around an object. If all boxes are situated very similarly in the same locations, a low value of uncertainty will be measured, while if each box location varies greatly for a single object, a high value of uncertainty will be attributed to that detection.

IV. Experimental Set-Up

In this dissertation, four different experiments were performed, which are detailed in Section IV-B. To implement the proposed studies, a dataset with images from the autonomous driving domain and a deep object detector had to be chosen. For the object detector, the YOLOv5 architecture [19] was chosen, and for the dataset, the Berkeley Deep Drive (BDD) [20], a diverse driving dataset for heterogeneous multitask learning.

The algorithms were developed using the Python programming language and mainly the Pytorch framework. The code developed in this work was built on top of the existing code for the YOLOv5 detector [19]. Furthermore, the main inspiration for code development came from the work of Ali Harakeh [21], especially the state-of-the-art implementation of methods such as Output Redundancy, MC Dropout, and the evaluation metrics. The fully developed code for the experiments performed in this dissertation is presented in [22]. The experiments were performed with an NVIDIA GeForce RTX 2070 SUPER GPU, provided by ISR - Instituto de Sistemas e Robótica.

A. Object Detector and Dataset

The smaller version YOLOv5s was chosen. The object detector used for this dissertation was pre-trained with the COCO dataset [23] and therefore can localize and classify 80 different classes. The original BDD dataset identifies 10 different category classes.

Following the work developed in [8], [21], the only categories used are the following seven: *car*, *person*, *bus*, *truck*, *motorcycle*, *bicycle* and *rider*. Of the 80 different classes that the pre-trained YOLOv5s can classify, only 6 correspond to the classes in the BDD dataset, where the missing class is *rider*. To solve this issue, we decided to consider the *rider* class equal to a *person* ground truth label.

B. Experiments

Some parameters are fixed and equal to all experiments. The object detector chosen to perform the inference on the images is YOLOv5s and the chosen images are the test set taken from the BDD dataset. The number of runs performed in the TTA method is also always fixed at 10.

For Experiments 1 and 2, only the gamma augmentation is performed, with gamma factors in the interval $[0.4; 2]$ in steps of $\frac{2-0.4}{10-1} = 0.178$.

*1) **Experiment 1 - Bounding Box Selection Criteria and IOU Thresholds**:* In Experiment 1 we propose to study the effect that different bounding box selection criteria can have on the quality of the estimated uncertainty and predictive distributions. Additionally, for each criterion, we propose an investigation of different IOU thresholds in the clustering step.

The chosen bounding box selection criteria will be based on the maximum confidence score obtained for a detection and will be compared to the established technique NMS [10]–[12]. The different criteria chosen to investigate are as follows.

- NMS - Perform NMS on the detections of each particular run and only accumulate those remaining
- Detections with maximum Confidence Score > 0.1
- Detections with maximum Confidence Score > 0.01
- Detections with maximum Confidence Score > 0.001
- Detections with maximum Confidence Score > 0.0001

The IOU threshold used in the clustering step will also be investigated. Therefore, for each bounding box selection criterion, a range of different IOU thresholds will be utilized.

To have a general understanding of its impact, the values chosen to be studied for the IOU threshold are: $[0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95]$.

*2) **Experiment 2 - Two Different IOU Thresholds for Regression and Classification**:* For Experiment 2, we propose the use of two different IOU thresholds for classification and regression to explore its impact on the quality of the estimated uncertainties. Therefore, for the same object, two different groups of bounding boxes are used to estimate the uncertainty for each task. We hope to find an optimal combination of IOU thresholds that maximizes performance.

The intervals of IOUs chosen for each task were as follows:

- Regression task: $[0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65]$.
- Classification task: $[0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1]$.

For this experiment, the bounding box selection criterion is fixed at 0.01.

*3) **Experiment 3 - Different Augmentations for UQ in TTA**:* In Experiment 3 we propose to evaluate how different color transformations impact the quality of the predicted distributions and the quality of the measured uncertainty for the domain of autonomous driving.

The bounding box selection criterion will be fixed at 0.01 as in Experiment 2 and now a combination of IOUs will be chosen and fixed for the classification and regression task. The actual values to be used will be those that show improvements in performance in Experiment 2.

Four different augmentations will be tested, each with an interval of possible augmentation factors that are incremented in each run.

The augmentations performed were as follows:

- Gamma: $[0.4; 2]$
- Contrast: $[0.3; 2]$
- Brightness: $[0.3; 2]$
- Gaussian blur: $[0.1; 3]$

After evaluating the results for each augmentation, a second study will be performed. We propose to investigate whether the combination of different augmentations can further improve the performance obtained.

From the initial four augmentations, we will choose the three best performing ones to combine. These three augmentations will be combined in pairs, producing three possible combinations. As 10 runs are performed, one augmentation will be used for 5 runs, while the other augmentation will be used for the remaining runs. As each augmentation is only used 5 times, the augmentation factor interval will now be split into 5 uniform portions, instead of the previous 10.

*4) **Experiment 4 - Comparative Study of UQ Methods**:* In Experiment 4 we propose to investigate the performance (in both classification and regression tasks) of the novel TTA method, in comparison to the state-of-the-art methods (MC Dropout and Output Redundancy) and the baseline deterministic YOLOv5 object detector. A second objective is to study possible improvements to the aforementioned state-of-the-art methods by modifying some of their parameters with more optimal ones that were obtained with the discoveries from the previous experiments.

The TTA method will utilize the best performing parameters found in Experiments 1,2 and 3. These parameters include a bounding box selection criterion from Experiment 1, a combination of two IOU thresholds for each task from Experiment 2 and the augmentations from Experiment 3.

This experiment will be divided into two sections. In the first section, the best TTA method will be compared with the baseline and state-of-the-art methods, using the established state-of-the-art parameters in the literature.

Therefore, for both the MC Dropout and Output Redundancy, the clustering algorithm used will be a BSAS with *Same Label* and $IOU = 0.95$ as affinity measures [11]. The MC dropout will use NMS as the bounding box selection criterion, and the Output Redundancy, as it only performs one run, will use all 15120 bounding boxes, without a bounding box selection criterion [9].

To provide a fairer comparison, the MC dropout will also perform 10 runs of inference for each image, just as the TTA. Following the work of Feng et al. in [8], a dropout rate of 0.1 was inserted before the final convolutional layer of the YOLOv5 architecture.

For the second section of the experiment, changes to the parameters of the MC Dropout and Output Redundancy will be performed, taking into account the optimal results already incorporated into the TTA method. These modifications could occur in the bounding box selection criterion used as well as different IOU thresholds for each task.

C. Evaluation

*1) **True Positives (TP), False Positives (FP) and False Negatives (FN)**:* In this dissertation, we use the rules from the PASCAL VOC challenge [24], adding only an extra requirement for the considered TP detections, to account for the autonomous driving domain.

A detection is considered a TP if its IOU score with the ground truth target is greater than 0.5 and, at the same time, if its predicted class corresponds to the ground truth target class.

A detection is considered a FP if it does not meet the minimum required threshold 0.5 of IOU with any ground truth target, or if it meets the minimum IOU threshold but the predicted class is different from the ground truth target label.

If multiple detections meet the requirements of a TP with the same ground truth (IOU greater than 0.5 and the correct class), only the detection with the highest confidence score will be considered as the TP. Any other detection is considered a FP if it does not meet the requirements of being a TP with any other ground truth.

FN are all ground truth target boxes that did not have any detection correctly attributed to them as TP.

*2) **Metrics for Evaluation**:* The metrics used for the evaluation in this work are based on the work in [6]. Three evaluation metrics are used to quantify the performance of the methods. For performance in the detection task, we use mAP [23]. The maximum mAP achievable by a detector is 100%. The MUE [11] is used to determine the ability to distinguish between TP and FP detections using the detector's estimated uncertainty. The lowest MUE achievable by a detector is 0. Finally, the NLL is used to evaluate the estimated predictive distributions for both classification and regression tasks for TP detections. The best results for NLL by a detector occur when NLL is equal to 0.

V. Experimental Results and Discussion

A. Experiment 1 - Bounding Box Selection Criteria and IOU Thresholds

*1) **Results for Experiment 1 - Negative Log Likelihood**:* The results obtained in Experiment 1 for NLL in the regression task are shown in Figures 3 and 4.

As the IOU threshold values increase, the quality of the predicted distributions worsens. Higher IOUs will restrict the number of bounding boxes attributed to each final cluster, gathering only those with similar localization, which in turn

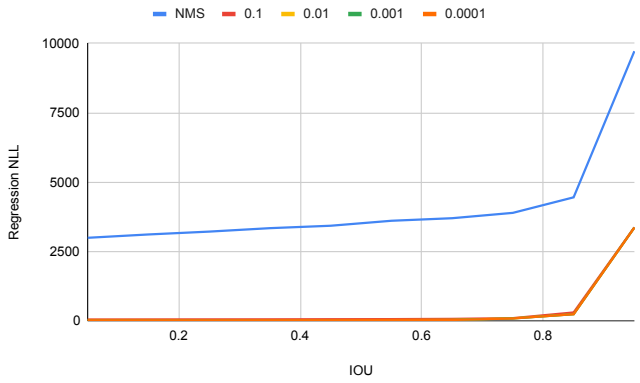

Fig. 3. Results for Experiment 1 - NLL Regression Task vs. IOU for different bounding box selection criteria

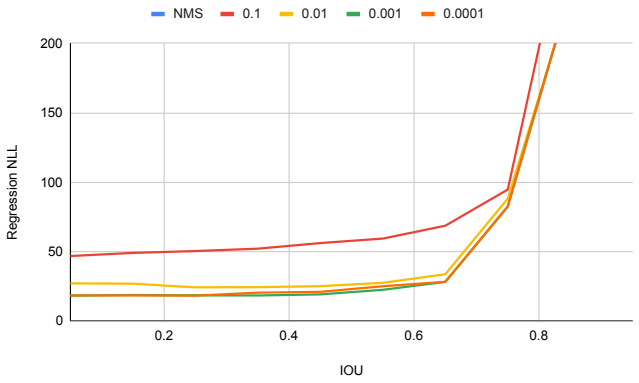

Fig. 4. Results for Experiment 1 - NLL Regression Task vs. IOU (zoomed in) for different bounding box selection criteria

can result in missing valuable information for the accurate estimation of the distributions.

Regarding the different bounding box selection criteria, NMS performs the worst, even when lower IOUs are used. One possible explanation is that this technique filters out too many bounding boxes, only leaving the most accurate ones. However, for the other criteria, similar results were obtained, with only an exception for the 0.1 criterion, which performs slightly worse in comparison to the 0.01, 0.001 and 0.0001 criteria.

In Figure 5, the NLL results in the classification task for Experiment 1 are shown. For the classification task, almost the

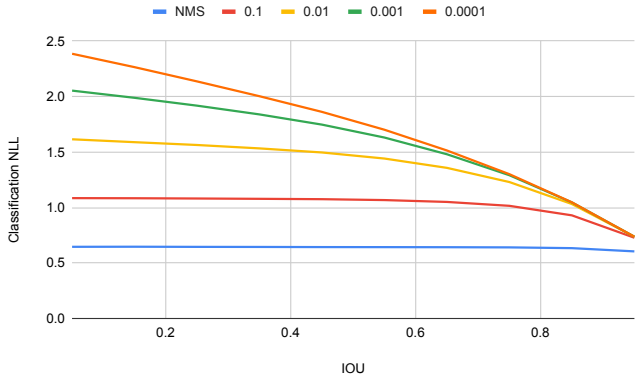

Fig. 5. Results for Experiment 1 - NLL Classification Task vs. IOU for Different Bounding Box Selection Criteria

inverse of what was previously discussed for the regression task is observed. With higher IOU values, better results are obtained for each criterion, and the stricter the criterion used, the higher the quality of the estimated distributions. There is also an evident bias caused by broader criteria and lower quality detections, which can be mitigated by higher values of IOU. This is evident from the fact that for lower IOU values, where a large number of bounding boxes are clustered for each object, the quality of the predicted distributions worsens.

One possible extrapolation from these results is that the fewer the number of bounding boxes used to compute the mean of the classification scores, the better the final predictions.

In a broader view of the results obtained for the NLL in Experiment 1, taking into account both the regression and classification tasks, one could argue for the use of different bounding box selection criteria for each task, with NMS or 0.1 for the classification task and 0.0001, 0.001 or 0.01 in the regression task.

Furthermore, lower IOUs are preferred for the regression task, while higher IOUs are preferred for the classification task.

*2) **Results for Experiment 1 - Minimum Uncertainty Error**:* The results for the MUE in both the regression and classification tasks are shown in Figure 6 and Figure 7, respectively.

The results obtained continue to demonstrate a similar remark previously given in the NLL section: the higher the IOU threshold, the better the results for the estimated uncertainty in the classification task, but the worse the results for the regression task.

For the regression task, in Figure 6, the worst performances occur with a bounding box selection criteria of 0.0001 and 0.001, while the other three have similar results. There is an optimal IOU value threshold around the 0.15-0.25 IOU mark, and adding more bounding boxes to the cluster (by further lowering the IOU) can indeed worsen the distributions created. One possible explanation for this is, for example, the introduction of bounding boxes that belong to nearby objects.

For the classification task, in Figure 7, the results improve with higher IOU values. The bounding box selection criteria perform very similarly, except for the 0.1 criterion, which is the worst of them, and NMS, which is the best of them.

B. Experiment 2 - Two Different IOU Thresholds for Regression and Classification

In this section, the IOU threshold used for the regression task will be mentioned as IOU_r and the IOU threshold used for the classification task as IOU_c.

*1) **Results for Experiment 2 - Negative Log Likelihood**:* The results for the NLL in regression and classification tasks are shown in Figures 8 and Figure 9, respectively.

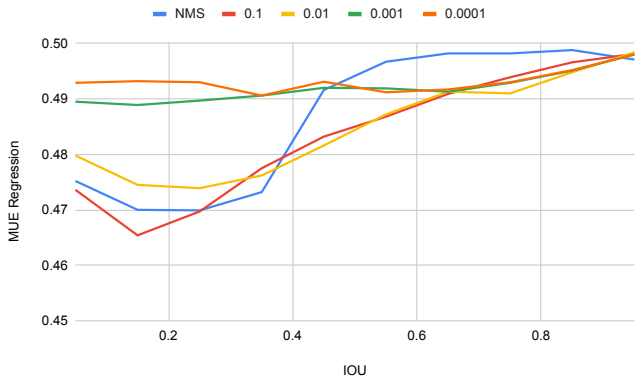

Fig. 6. Results for Experiment 1 - MUE Regression Task vs. IOU for different bounding box selection criteria

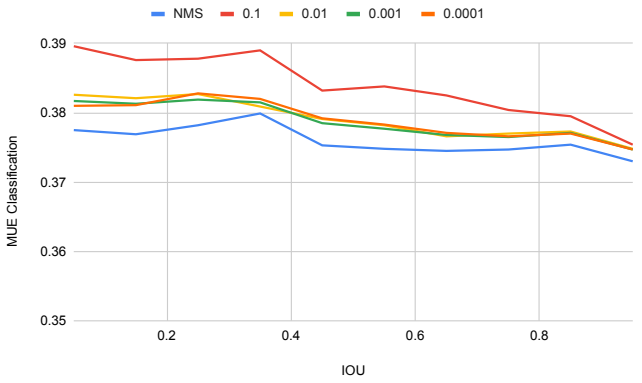

Fig. 7. Results for Experiment 1 - MUE Classification Task vs. IOU for different bounding box selection criteria

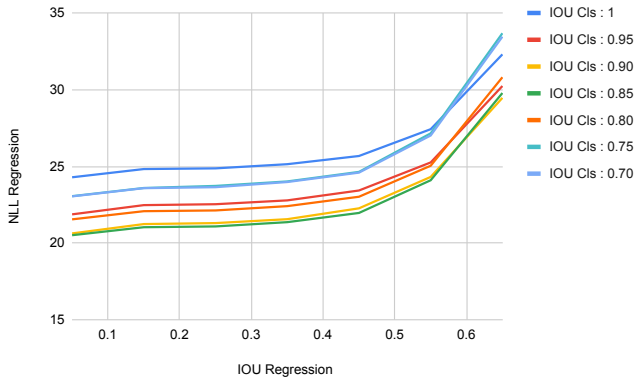

Fig. 8. Results for Experiment 2 - NLL Regression Task for different combinations of IOU_r and IOU_c

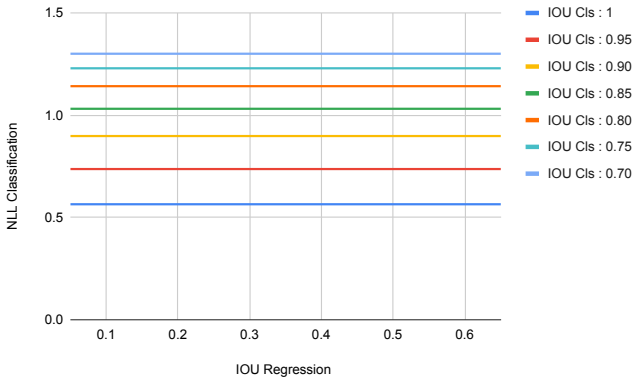

Fig. 9. Results for Experiment 2 - NLL Classification Task for different combinations of IOU_r and IOU_c

For the classification task, a clear observation can be made: NLL depends only on IOU_c. Varying IOU_r has no effect on the final value of NLL.

For different IOU_c values, the higher the threshold, the better the results obtained. The best situation is to use a IOU_c value equal to 1, which is equivalent in practice to only using the confidence scores of the best detection, the cluster center.

Analyzing now Figure 8 with the results of the regression task, a different behavior is visible. For different thresholds IOU_c, improvements are obtained as the threshold IOU_r decreases. In a practical sense, as more detections are collected for each object, the quality of the predicted distributions increases.

The worst results in regression occurred when IOU_c is equal to 1. As the threshold IOU_c drops, the results continue to improve up to the threshold mark of 0.85. However, lowering it more than that causes the opposite effect: thresholds of 0.80, 0.75 and 0.70 have increasingly worse results. Therefore, the aggregation of an increasing number of bounding boxes to calculate the categorical distribution can help the performance in the regression task to some extent. With an IOU_c threshold lower than 0.85, the results of the regression task worsen.

*2) **Results for Experiment 2 - Minimum Uncertainty Error**:* Regarding the quality of the measured uncertainty for the classification task, shown in Figure 10, the uncertainty error is only affected by the IOU_c threshold, similarly to the previous case. Any changes to the IOU_r threshold have no effect on the final results. Upon analysis of the effect of

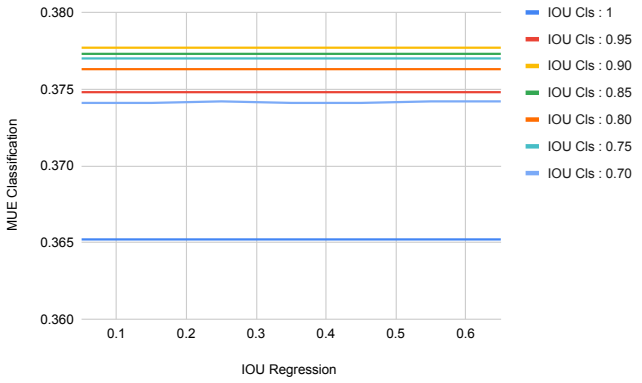

Fig. 10. Results for Experiment 2 - MUE Classification Task for different combinations of IOU_r and IOU_c

varying the IOU_c thresholds, no broad conclusion can be reached. As the threshold goes down, indeed as before, the results worsen until the 0.90 threshold is reached. However,

TABLE I
RESULTS FOR EXPERIMENT 3 - COMBINATION OF AUGMENTATIONS

	mAP (%) ↑	NLL Cls ↓	NLL Reg ↓	MUE Cls ↓	MUE Reg ↓
Gamma + Brightness	16.9	0.5576	**23.1218**	0.3575	0.4756
Gamma + Contrast	16.9	0.5520	25.2416	0.3605	**0.4726**
Contrast + Brightness	**17**	**0.5493**	26.1155	**0.3537**	0.4753

after that, there is a slight increase in performance with the 0.85 and 0.80 marks, a decrease with the 0.75 threshold, and at the 0.70 mark there is a great improvement, where the second-best performance is recorded.

Regardless of these fluctuations, a clear observation can be made. The best result is obtained with an IOU_c of 1, with a clear performance gap when compared to the other thresholds that perform rather similarly.

Now, referring to Figure 11 and the regression task, all IOU_c thresholds perform very similarly, although there is a noticeable improvement in performance for the IOU_c values of 1 and 0.95.

For this task, however, inversely to the classification task, the IOU_r threshold has a great impact on the performance obtained. There is an optimal threshold IOU_r at 0.15. Therefore, to obtain the best results for the MUE in the regression task, the thresholds to be used are 0.15 for IOU_r and 1 or 0.95 for IOU_c.

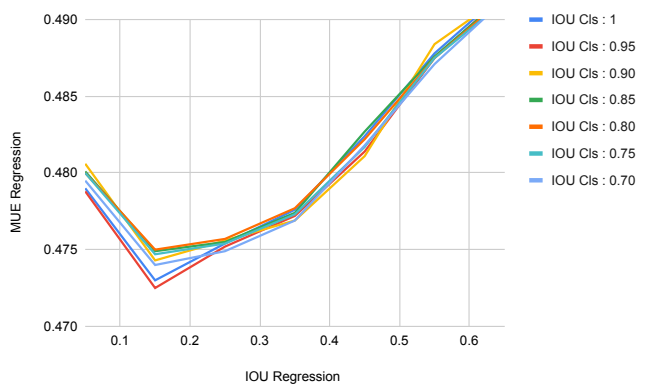

Fig. 11. Results for Experiment 2 - MUE Regression Task for different combinations of IOU_r and IOU_c

C. Experiment 3 - Different Augmentations for UQ in TTA

Analyzing the results obtained for Experiment 3 in Table II with regard to mAP, contrast and brightness performed the best, followed by gamma augmentation, while Gaussian blur provided the worst performance.

The contrast augmentation also provided the best performance for the predicted distributions in the classification task, while brightness achieved the best predictive distributions for the regression task and the best uncertainty estimation in classification. For the uncertainty measured in the regression task, gamma augmentation achieved the best performance, with contrast and brightness only slightly behind with similar results.

More experiments were performed that combined in pairs the three best augmentations previously reported: gamma, contrast, and brightness. The results for the combined augmentations are shown in Table I.

TABLE II
RESULTS FOR EXPERIMENT 3 - DIFFERENT AUGMENTATIONS

	mAP (%) ↑	NLL Cls ↓	NLL Reg ↓	MUE Cls ↓	MUE Reg ↓
Gamma	16.7	0.5646	24.8237	0.3625	**0.4730**
Contrast	**16.9**	**0.5555**	30.0253	0.3644	0.4764
Brightness	**16.9**	0.5601	**20.0892**	**0.3601**	0.4761
Gaussian Blur	16.5	0.5628	22.1136	0.3631	0.4802

The combination of augmentations with the greatest improvement was brightness and contrast, with improved mAP, NLL, and MUE in classification. For the regression task however, no noticeable improvements were achieved.

The improvements illustrate the fact that perhaps each of these augmentations provides better detections in a specific manner, and using them together can give the best out of both augmentations.

D. Experiment 4 - Comparative Study of UQ Methods

In Table III, the results for the first section of Experiment 4 are shown. TTA performed better for all metrics in comparison to the current state-of-the-art methods and baseline, with an exception for the NLL in the classification task, where it obtained only slightly worse results than the MC Dropout.

The most noticeable performance difference occurred in the predicted distributions for the regression task, which can be explained by the fact that MC Dropout and Output Redundancy use 0.95 as the IOU threshold in the clustering step, which was previously observed in Experiment 2 not to be the optimal threshold to use for the regression task. Furthermore, Output Redundancy does not apply a bounding box selection criterion and MC Dropout uses NMS as a criterion, which was shown in Experiment 1 to be the worst performing criterion for estimating predictive distributions in the regression task.

In TableIV, results are shown for the modified state-of-the-art methods with optimal parameters. As expected, performance improved greatly, especially for the regression task, since now an optimal value of IOU_r of 0.15 was used, as well as an optimal bounding box selection criterion. Still, TTA performed best in terms of mAP, NLL in regression and MUE in classification. With the improved MC Dropout, now the best performance for the MUE in the regression was obtained, again proving the relevance of using the correct criteria and IOU threshold for each task.

The increase in mAP with the TTA in comparison to the Output Redundancy and the Baseline can be explained by the fact that by performing multiple augmentations and inferences for each image, it is possible to visualize objects in a clearer way that would not be possible with the original image, giving

TABLE III
RESULTS FOR EXPERIMENT 4 - COMPARISON STUDY BETWEEN METHODS

	mAP (%) ↑	NLL Cls ↓	NLL Reg ↓	MUE Cls ↓	MUE Reg ↓
Baseline (Deterministic)	16.2	-	-	-	-
Output Redundancy	16.3	0.6771	8230.3852	0.3690	0.4889
Monte Carlo Dropout	14.5	**0.5483**	29697.7105	0.3543	0.4784
Test-Time Augmentation	**17**	0.5493	**26.1155**	**0.3537**	**0.4753**

TABLE IV
RESULTS FOR EXPERIMENT 4 - COMPARISON STUDY BETWEEN METHODS. *IOU_r OF 0.15 AND IOU_c OF 1; 0.01 CRITERION

	mAP (%) ↑	NLL Cls ↓	NLL Reg ↓	MUE Cls ↓	MUE Reg ↓
Baseline (Deterministic)	16.2	-	-	-	-
Output Redundancy*	16.3	0.5734	52.6753	0.3665	0.4780
Monte Carlo Dropout*	14.3	**0.5413**	34.2484	0.3620	**0.4706**
Test-Time Augmentation	**17**	0.5493	**26.1155**	**0.3537**	0.4753

the detector a better chance of correctly identifying an object and thus improving the quality of the final detections.

VI. CONCLUSIONS

In this dissertation, we show that the TTA method can become a very good option for UQ in object detection in the domain of autonomous driving. Furthermore, already in use state-of-the-art methods could have their performances greatly improved by using two different IOU thresholds for each task and incorporating a bounding box selection criterion. The lower IOU thresholds are shown to generate the best results for the regression task, while the higher IOU thresholds produce the best results for the classification task.

REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[4] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3, 2006, pp. 850–855.

[5] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059.

[6] A. Harakeh, "Estimating and evaluating predictive uncertainty in deep object detectors," Ph.D. dissertation, University of Toronto, 2021.

[7] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[8] D. Feng, A. Harakeh, S. L. Waslander, and K. C. J. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 9961–9980, 2022.

[9] M. T. Le, F. Diehl, T. Brunner, and A. Knol, "Uncertainty estimation for deep neural object detectors in safety-critical applications," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3873–3878.

[10] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–7, 2018.

[11] D. Miller, F. Dayoub, M. Milford, and N. Sünderhauf, "Evaluating merging strategies for sampling-based uncertainty techniques in object detection," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 2348–2354, 2019.

[12] D. Miller, N. Sunderhauf, H. Zhang, D. Hall, and F. Dayoub, "Benchmarking sampling-based probabilistic object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[13] F. Kraus and K. Dietmayer, "Uncertainty estimation in one-stage object detection," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 53–60.

[14] A. Harakeh, M. Smart, and S. L. Waslander, "Bayesod: A bayesian approach for uncertainty estimation in deep object detectors," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 87–93.

[15] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[16] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[17] I. Kim, Y. Kim, and S. Kim, "Learning loss for test-time augmentation," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4163–4174.

[18] G. Jocher. Yolov5 documentation - test-time augmentation (tta). [Online]. Available: https://docs.ultralytics.com/tutorials/test-time-augmentation/

[19] G. J. et al. (2022, Aug.) ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations.

[20] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[21] A. Harakeh. (2021) A review and comparative study on probabilistic object detection in autonomous driving. [Online]. Available: https://github.com/asharakeh/pod_compare

[22] R. Magalhães. (2022) Uncertainty quantification with test-time augmentation. [Online]. Available: https://github.com/ruimagalhaes24/UQ_TTA_OD

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[24] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.