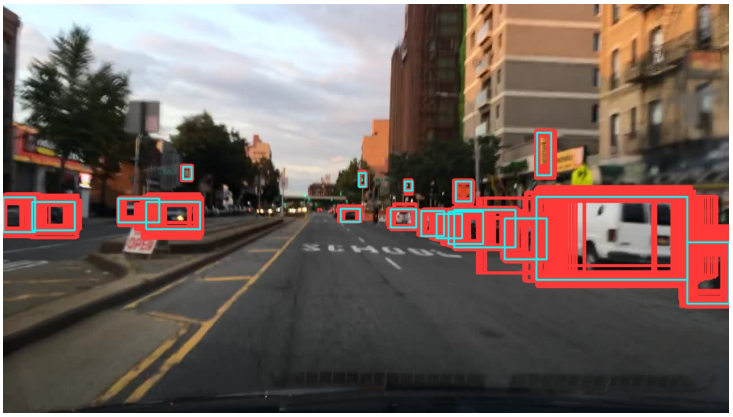
TÉCNICO LISBOA



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

Rui Miguel Campos Magalhães

Thesis to obtain the Master of Science Degree in

Aerospace Engineering

Supervisor: Prof. Alexandre José Malheiro Bernardino

Examination Committee

Chairperson: Prof. Paulo Jorge Coelho Ramalho Oliveira
Supervisor: Prof. Alexandre José Malheiro Bernardino
Member of the Committee: Prof. Chryssa Zerva

November 2022

Declaration

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of Universidade de Lisboa.

Acknowledgments

O trabalho realizado nesta tese fecha um dos ciclos mais difíceis e desafiantes da minha vida. Com muito sofrimento consegui superar obstáculos e cresci muito como pessoa. Chegar a este ponto seria impossível sem o apoio e carinho das pessoas que também fizeram parte desta minha jornada, que acreditaram em mim e estiveram lá para mim quando precisei. Estou eternamente grato a todos.

Obrigado Mãe, Pai e Cláudia por tudo o que me deram ao longo da minha vida. Tudo o que alcancei não seria possível sem vocês. Obrigado por acreditarem em mim e pelo amor incondicional.

Obrigado Patrícia, por todo o apoio, carinho e motivação que me deste ao longo destes anos. Estiveste sempre lá para mim mesmo nos momentos mais difíceis e foi a teu lado que saí deles.

Obrigado aos meus amigos por todas as memórias, risos, conversas estúpidas e noites passadas no aquário. Conheci-vos por causa do técnico, mas não saía do técnico sem vos ter conhecido.

Por fim, gostaria de agradecer ao ISR e ao professor Alexandre Bernardino pela oportunidade de realizar esta tese, assim como por toda a ajuda, disponibilidade e ideias inovadoras.

"Eu dei a César o que é de César, dêm o meu corpo à minha mãe."

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 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.

Keywords

Uncertainty Quantification; Object Detection ; Test-Time Augmentation ; Autonomous Driving.

Resumo

A condução autónoma depende de vários sistemas complexos para executar tarefas vitais. Uma tarefa essencial para a perceção do ambiente exterior ao carro é a deteção de objetos com base em imagens de câmaras. Avanços recentes em visão computacional tornaram a utilização de redes neurais convolucionais o estado da arte para tarefas de deteção de objetos. Em sistemas onde a segurança é vital, como na condução autónoma, é essencial medir a fiabilidade das deteções estimadas. A quantificação da incerteza associada a cada deteção pode proporcionar uma condução autónoma mais segura e mais fiável. Neste trabalho, investigamos abordagens inovadoras para quantificar incerteza em métodos de deteção de objetos, assim como exploramos melhorias nos atuais métodos utilizados. Propomos utilizar pela primeira vez o método de *Test-Time Augmentation* (*TTA*) para quantificação de incerteza aleatória na deteção de objetos. Para isso, desenvolvemos uma metodologia inovadora com recurso ao *TTA* e mostramos que é capaz de superar o desempenho dos métodos atuais de estado da arte, *Monte Carlo Dropout* e *Output Redundancy* na qualidade das distribuições e incertezas estimadas tanto para a tarefa de classificação como de regressão. São efetuados estudos para investigar a utilização de um critério para seleção de deteções e dois valores mínimos de *Intersection-over-Union* (*IOU*) diferentes para cada tarefa (classificação e regressão). Mostramos que é possível obter melhorias no desempenho dos métodos *Monte Carlo Dropout* e *Output Redundancy* através da aplicação de um critério de seleção ótimo e dois valores mínimos de *IOU* diferentes para cada tarefa separada. Mostramos que valores mais baixos de *IOU* geram os melhores resultados para a tarefa de regressão, enquanto que valores mais altos de *IOU* produzem os melhores resultados para a tarefa de classificação.

Palavras Chave

Quantificação de Incerteza; Deteção de Objetos ; Test-Time Augmentation ; Condução Autónoma.

Contents

List of Figures

List of Tables

Acronyms

AP	Average Precision
BSAS	Basic Sequential Algorithmic Scheme
BDD	Berkeley Deep Drive
FP	False Positives
FN	False Negatives
IOU	Intersection-over-Union
NLL	Negative Log Likelihood
NMS	Non-Maximum Supression
mAP	mean Average Precision
MUE	Minimum Uncertainty Error
MC	Monte Carlo
PR	Precision-Recall
PDQ	Probability-based Detection Quality
TP	True Positives
TTA	Test-Time Augmentation
UE	Uncertainty Error
UQ	Uncertainty Quantification

1

Introduction

Contents

In this chapter, an overview of the motivation and background behind the work developed in this dissertation is given first in Section 1.1. The objectives and contributions are explained in Section 1.2, and, lastly, a general outline of the whole document is given in Section 1.3.

1.1 Motivation and Background

Autonomous driving relies on various complex systems to perform essential tasks in the automated driving scenario. Different sensors can be used, such as Lidar, cameras and radar to perceive the environment that surrounds the vehicle. 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 [4], YOLO [5] and R-CNN [6].

A neural network object detector produces multiple bounding box predictions, each combined with a classification score. The bounding box is represented by four values that can vary between detectors and also between datasets. For example, the YOLOv5 object detector [7] represents bounding boxes by Height, Width, and (X,Y) coordinates of the center of the box. In Figure 1.1, a visualization of the YOLOv5 bounding box coordinate system is given.

Figure 1.1: Sample image from Coco Dataset, representing the YOLOV5 Bounding Box Coordinate System [1]

For the category of an object, classification scores (with values ranging from 0 to 1) are produced for each different category. The winning category class is the one with the highest classification score. As multiple bounding boxes are predicted for each object, an algorithm known as Non-Maximum Supression

(NMS) [8] is usually used to remove redundant detections. A practical example of the detection of a car with the predicted bounding box and classification score is given in Figure 1.2.

Figure 1.2: Example of a Car Detection

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 [9] 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 uncertainty associated to the location of the bounding box. Current state-of-the-art object detectors have no measure of how certain they are in their predictions [10]. It is not sufficient to rely on the classification score alone. In Figure 1.3, a practical example of a falsely detected object with a high confidence score is shown. In this situation, the car itself is detected in the bottom region of the image, where no objects should be detected.

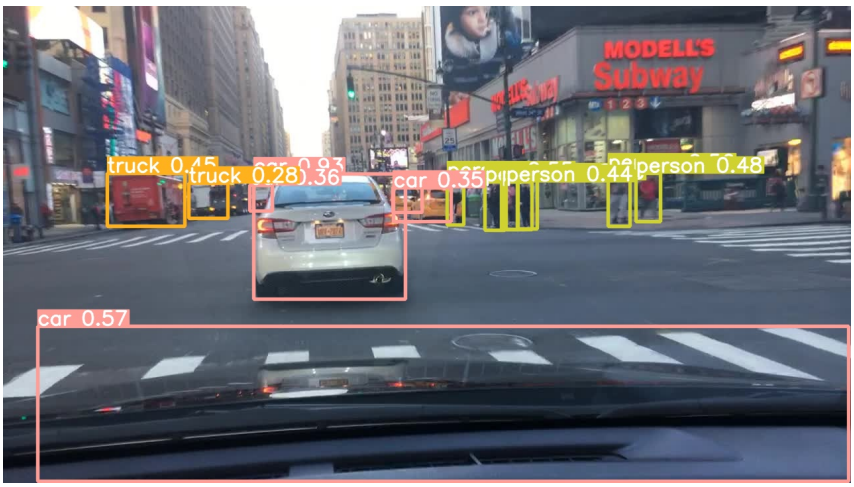


Figure 1.3: Example of high confidence for a false detection in the hood of the car

For safety-critical systems, such as autonomous driving, reliable predictions are required. Therefore, it is essential to measure how reliable the estimated output of the detector is. The lack of uncertainty

measurements could have drastic consequences if incorrect and uncertain detections are not identified.

Driving at night or in adverse weather conditions are examples of real-world situations where the sensors can have significantly reduced amounts of light and contrast available for visual perception, producing detections with high levels of uncertainty. Accurate quantification of the uncertainty associated with each detection can provide safer and more reliable autonomous driving.

Gal in [11] 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 [12]. Several methods have been developed for Uncertainty Quantification (UQ) in object detection: Direct Modeling [13–19], Monte Carlo (MC) Dropout [16, 18, 20–22], Output Redundancy [13] and Deep Ensembles [22]. So far, these methods have been the only ones used for the task of object detection, but various other methods have been employed in the literature to quantify uncertainty in other tasks [23].

An example of such a method is the Test-Time Augmentation (TTA), which was previously used to quantify aleatoric uncertainty in image segmentation [24–27] and image classification [28, 29], but has not yet been applied to object detection. In the TTA method, multiple augmented versions of each image are created and predictions are performed for each version, creating an ensemble of those predictions. By measuring how diverse the predictions are for each detected object, uncertainty estimations can be performed. Positional or color augmentations can be used. The chosen augmentations must take into account the domain of the image dataset and the task at hand.

For sampling-based methods, the predictions can be clustered into groups for each detected object, using Intersection-over-Union (IOU) as a measure of the spatial affinity between them [21]. IOU measures the overlap between two two-dimensional boxes.

1.2 Objectives and Contributions

Following the need for probabilistic object detectors for safer and more reliable autonomous driving, the main goal of this dissertation 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 of this dissertation, and these studies should hopefully lead to better decision making in the automotive automation industry.

Although this dissertation focuses on the domain of autonomous driving, it is important to note that the approaches developed in this work can also be applied to numerous other applications and domains. For example, the developed work can be used for drone or satellite images, as well as autonomous

aircraft in the aerospace domain.

The contributions in this dissertation are as follows:

- 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, for the domain of autonomous driving.
- We explore the use of two different IOU thresholds in the clustering step 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.

1.3 Outline of the Document

In Chapter 2 an overview of relevant state-of-the-art work and theoretical concepts is provided. In Chapter 3, the methods used in this work will be thoroughly explained and, taking them into account, Chapter 4 will detail the different experimental setups created for the investigation work in this dissertation. In Chapter 5 the results will be presented and a discussion regarding them will be provided. Chapter 6 will give the main conclusions from the work performed in this thesis, as well as mention possible future work to be developed.

2

State-of-the-art

Contents

In this chapter, an overview of the relevant theoretical concepts and related work is given for the development of this dissertation. The main focus relies on the methods of UQ for the object detection task, as it is the focal point of this work.

2.1 Uncertainty Quantification in Deep Learning

2.1.1 Epistemic and Aleatoric Uncertainty

According to the work of Yarin Gal [11], there are two types of uncertainties in deep learning: aleatoric and epistemic.

Epistemic, or model uncertainty, concerns the uncertainty related to the model itself, which is typically associated with a lack of sufficient data in the training dataset. For example, in the domain of autonomous driving, if a detector is trained only with images of driving situations during the day in sunny conditions, if an image taken at night is forward-passed through that detector, a higher value of epistemic uncertainty is expected in its detections.

Aleatoric, or data uncertainty, is associated with intrinsic randomness present in the data itself. In a practical sense, this type of uncertainty can be influenced by the quality of the sensors used (RGB cameras and Lidars with different resolutions provide images with higher or lower quality). Moreover, the situation captured by the sensors can have inherent uncertainty. For example, images with adverse weather conditions such as rain, snow, or fog will present higher aleatoric uncertainty in its predictions.

There are various methods for UQ, where some are able to quantify both types of uncertainty, while others only one or the other. Probabilistic object detectors are deep object detectors that can estimate the uncertainty associated with each detection. The work developed by Feng and Harakeh in [30], provides a great review of these types of detector and the work developed in this field.

2.1.2 Practical Methods for Uncertainty Estimation

Following the literature review of probabilistic object detectors in [30], the different methods used for uncertainty estimation in object detection are: Direct Modeling [13–19], MC Dropout [16, 18, 20–22], Output Redundancy [13] and Deep Ensembles [22]. The works mentioned here refer only to those that perform object detection for camera-based input images, since the work developed in this dissertation will focus only on that type of input data. However, it is important to note that other works have also been developed for UQ in object detection with Lidar sensor data [30].

2.1.2.A Direct Modeling

Direct Modeling is a method that requires modifications to the output layers of the deep object detector, as well as training with a new loss function. A particular probability distribution is assumed, and the network output layers are used to directly predict the parameters that describe that distribution. The modified network will be able to directly estimate aleatoric uncertainty for each detection on a single forward-pass. For the classification task, the categorical distribution is the most commonly estimated, and for the regression task, most of the works assume Gaussian distributions [13–19].

2.1.2.B Output Redundancy

Output Redundancy was developed by Le et al. in [13] 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. 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. It is a time-efficient method that only requires small modifications to the post-processing steps of an original object detector. However, its uncertainty estimates have been shown to be less accurate than the BayesOD direct modeling method [18].

2.1.2.C 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 [9] for UQ in image classification and regression tasks. Miller et al. extended this method for the object detection task for the first time in [20].

Dropout is a technique that randomly deactivates neurons in a layer, usually to avoid overfitting during the training process and typically being switched off for inference. However, 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 object detection task [20–22] and together with the direct modeling method in [16, 18] to estimate both aleatoric and epistemic uncertainty simultaneously.

2.1.2.D Deep Ensembles

Deep Ensembles is a method developed by Lakshminarayanan et al. [31] to model epistemic uncertainty. The technique consists of training an ensemble of networks, each with randomly shuffled training data and random initialization of network weights. Each network uses the same architecture, but is expected

to behave differently for the same input due to differences in training. The output of each network represents a sample, and the variability between the obtained samples will be used to estimate uncertainty. Miller et al. in [22] apply this method to the object detection task, concluding that it outperforms the MC Dropout technique. Again, as with the MC Dropout method, the Deep Ensembles technique can be used jointly with a Direct Modeling method to estimate both aleatoric and epistemic uncertainty.

2.1.3 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 [23]. One of these methods is TTA, which was previously used to quantify aleatoric uncertainty in image segmentation [24–27] and image classification [28, 29], 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, as seen in [24–29]. Positional or color augmentations are used and the chosen augmentations take into account the domain of the image dataset and the task at hand.

For the object detection task, TTA has been used in practical applications to improve mean Average Precision (mAP) and recall metrics [32]. 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, for the domain of autonomous driving.

2.1.4 Clustering Techniques and Affinity Measures

For Output Redundancy [13] and sampling-based methods (Deep Ensembles [22] and MC Dropout [16, 18, 20–22]), clustering of redundant detections is the established technique for uncertainty estimation in object detection.

Miller et al. [21] 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 [22], this technique will be treated as the current established clustering 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 [21], no work in the literature has 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.

2.1.5 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 the bounding box selection criterion [20–22]. After each forward-pass, NMS is performed, and only the resulting predictions from NMS are accumulated and used for uncertainty estimation.

However, as seen in the literature with the Output Redundancy method [13], all output predictions from a single forward-pass can be used for uncertainty estimation. We believe that limiting the number of output predictions by using NMS as a 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.

2.2 Evaluation of Predictive Distributions and Uncertainty

Feng et al. in [30] 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 mAP [15,17–19,21,33], Probability-based Detection Quality (PDQ) [18,19,22] and Minimum Uncertainty Error (MUE) [18,21].

According to Harakeh and Waslander in [34], proper scoring rules are essential for the evaluation of predicted distributions from probabilistic object detectors. With this in mind, Harakeh in [10] and in [30] uses Negative Log Likelihood (NLL) as a proper scoring rule to evaluate the quality of the predicted category and bounding box probability distributions.

This dissertation will follow the work developed in [10], also using NLL for the evaluation of predictive distributions, as well as MUE to assess the quality of the estimated uncertainty. The mAP metric [35] will also be computed since it is the standard evaluation metric in object detection.

3

Methodology

Contents

In this chapter, the methodology applied in this thesis is explained. One of the main contributions of the developed work is the use of the TTA method to measure uncertainty in the task of object detection for the first time. As this method was not created with the intention of quantifying uncertainty, some modifications and extensions to it were performed for that purpose.

Thus, this chapter presents a first overview of the original TTA method, followed by the proposed modifications that result in a novel TTA method to quantify uncertainty in object detection, as well as an explanation regarding how predictive distributions are estimated.

3.1 Test-Time Augmentation

The method of TTA is a sampling-based method that requires multiple runs of inference, where at each run, the input image is transformed by means of an augmentation. The performed augmentation generates a visual modification to the input image, which produces variability to the detector output predictions. By providing augmented versions of each image, objects of the original image can become easier to identify. Thus, the detector has a better chance of correctly identifying an object and producing higher quality detections, improving performance. This method has been used in [32] to improve accuracies in the YOLOv5 object detector.

In Figure 3.1, the original TTA pipeline is represented, with its various steps. For each original image, a fixed number of total runs are performed with increasing augmentation factors to accumulate predictions. After all runs are completed, the NMS algorithm is performed in the accumulated predictions to remove redundant detections and obtain the final detections. A detailed explanation of the NMS algorithm is provided in Section 3.1.3

3.1.1 Description of 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. A further explanation on the reasoning behind these choices and an illustration of each one of them will be given in Section 3.1.2.

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 actual total number of predictions produced for each image will vary based on the detector used.

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.

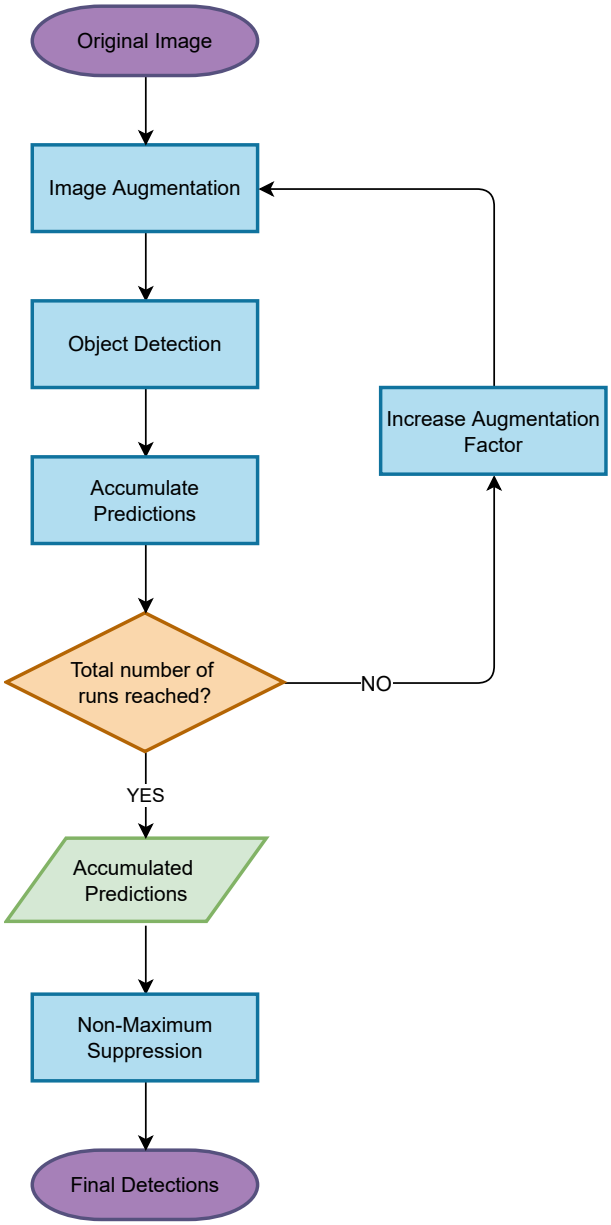


Figure 3.1: Original TTA Pipeline

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

3.1.2 Image Augmentations

There are different ways of performing augmentations in images, mainly with positional or color transformations. These should be chosen carefully, in order to preserve the quality and domain of the dataset used. For the work developed in this thesis, an autonomous driving dataset is utilized, which is a compilation of images of the road taken from a camera in the front of a vehicle while driving.

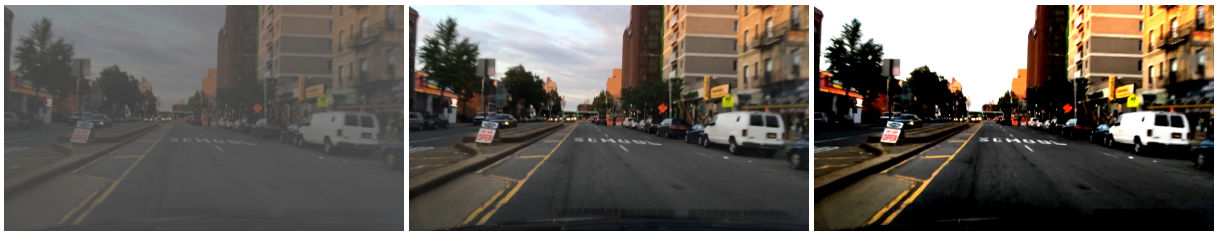
Color augmentations used in images from an autonomous driving dataset can produce augmented images that can look similar to real driving scenarios. Not only is it possible to create better image versions for improved predictions in the task of object detection, but it is also possible to recreate possible adverse conditions where objects are harder to identify. With that in mind, in this dissertation, only color augmentations were performed, in which the values of the pixels of each image were modified.

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

3.1.2.A Contrast 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.

In Figure 3.2, a visualization of two different contrast factors is provided, compared to the original image.



(a) Contrast Factor 0.3 **(b)** Original Image **(c)** Contrast Factor: 2

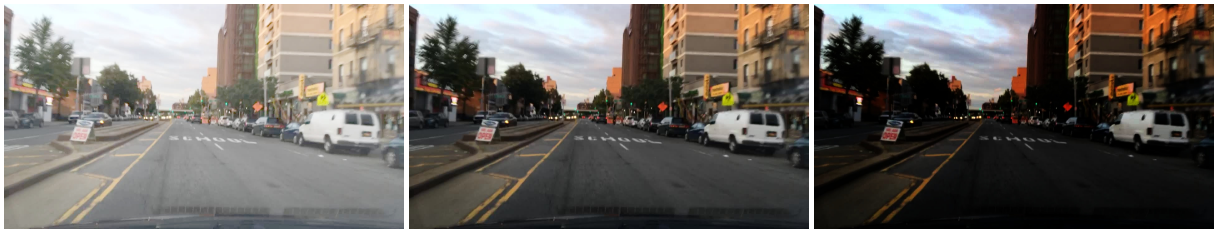
Figure 3.2: Augmentation by Contrast

3.1.2.B 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.

In Figure 3.3, two gamma augmentation factors are depicted in comparison to the original image. In Figure 3.3(a), we can see the possible effects of the gamma augmentation, where the bottom contours of the white van (on the right side of the image) become clearer and easier to identify compared to the original image.

In summary, higher gamma factors cause shadows to become darker, whereas lower gamma factors produce brighter shadows.



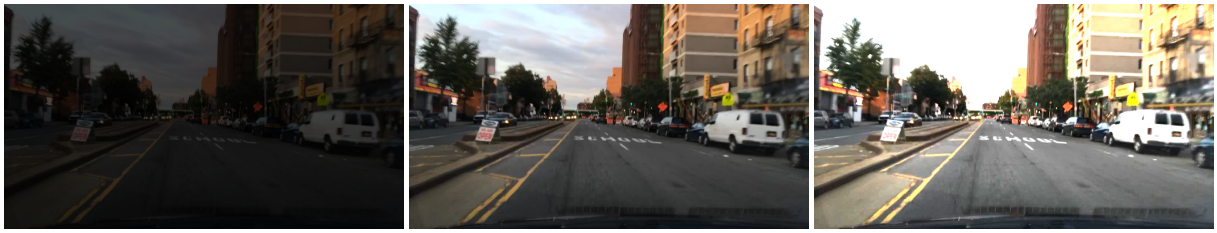
(a) Gamma Factor 0.4 **(b)** Original Image **(c)** Gamma Factor: 2

Figure 3.3: Augmentation by Gamma

3.1.2.C Brightness Augmentation

Altering brightness of an image will modify all pixels equally. When the 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.

Figure 3.4 shows the original image and two augmented versions with different brightness factors.



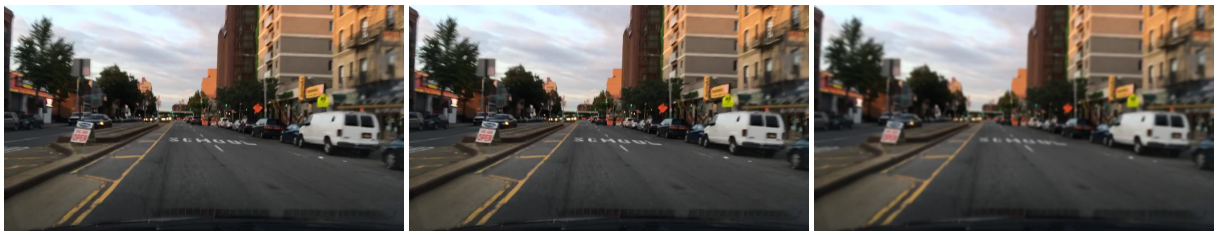
(a) Brightness Factor 0.3 **(b)** Original Image **(c)** Brightness Factor: 2

Figure 3.4: Augmentation by Brightness

3.1.2.D Gaussian Blur

A Gaussian blur will introduce 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.

In Figure 3.5, a visualization of two different Gaussian blur factors is provided, compared to the original image.



(a) Gaussian Blur Factor 0.1 **(b)** Original Image **(c)** Gaussian Blur Factor: 3

Figure 3.5: Augmentation by Gaussian Blur

3.1.3 Non-Maximum Supression and Intersection-over-Union

In order to explain the NMS technique, it is first needed to define what IOU is, also known as the Jaccard index. It is used in the evaluation and post-processing steps in object detectors.

IOU measures the overlap between two two-dimensional bounding boxes. As the name suggests, it is determined by dividing the area of intersection of two boxes by the area of their union. A visual representation of this concept can be seen in Figure 3.6. Typically, this score is used to evaluate the quality of a detection in terms of location in comparison to the ground truth target. Its value varies between 0 and 1, where 1 means a total overlap between the two boxes, and 0 having no overlap.

NMS is a technique used to select one single bounding box out of many overlapping boxes. For the task of object detection, a group of predicted bounding boxes is generated, usually with multiple redundant and overlapping detections for each object. With NMS, these redundant detections are removed and only one predicted detection per object is kept.

It is required to define some parameters for the NMS, mainly a minimum confidence score threshold and a minimum IOU threshold. The confidence score threshold will define the minimum confidence score that a prediction needs to have in order to be considered correct. The IOU threshold will be used to define whether two boxes belong to the same object in the image.

NMS starts by excluding all boxes that have not met the minimum classification score threshold.

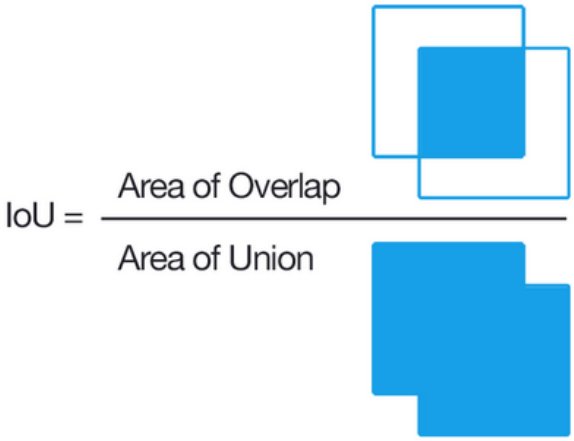


Figure 3.6: IOU representation [2]

Then, the top-scoring bounding box is selected and set aside as a final output. The IOU between this top scoring box and all the other bounding boxes is computed. The boxes that meet the minimum IOU score threshold are removed. In practical terms, these boxes are classifying the same object but have a lower classification score and should therefore be removed. These steps are repeated in the remaining boxes until there are no boxes left. The end result is a single bounding box detection per object class in the image.

3.2 TTA for Uncertainty Quantification in Object Detection

In this section, an overview is given of the modifications performed on the original TTA method to quantify aleatoric uncertainty for every final detection. As explained in Section 3.1, TTA is a sampling-based method in which multiple augmented versions of each image are created and predictions are performed for each version, creating an ensemble of those predictions. The performed augmentations produce variability to the detector output predictions. By measuring how diverse the predictions are for a given image, aleatoric uncertainty estimations can be performed, as seen in [24–29].

In this dissertation, we extend the original pipeline of TTA to quantify aleatoric uncertainty for the first time in object detection. In Figure 3.7, the novel pipeline can be visualized. The extended steps added to the original pipeline are largely based on the established merging strategy technique developed by Miller et al. [21].

3.2.1 Description of 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

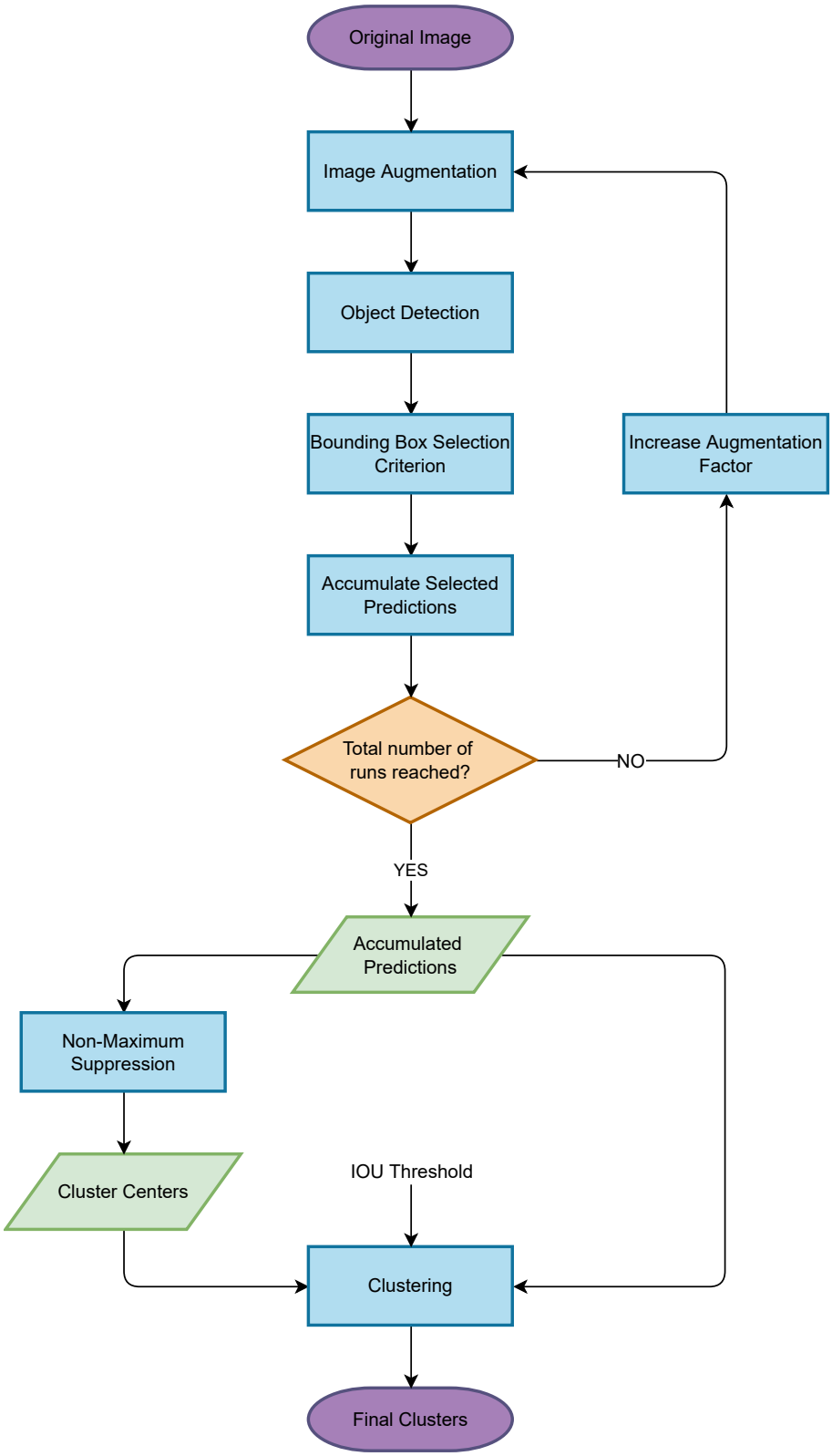


Figure 3.7: Novel TTA Pipeline for UQ

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. As defined in Section 2.1.5, a bounding box selection criterion is a technique or rule that decides which predictions are kept in each run. The criteria used in this dissertation will be based on the maximum classification score obtained for a prediction as it is an efficient way of removing background detections, keeping only the relevant ones that surround the objects in the image. 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 the runs. In the next step, the NMS algorithm is performed in 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. [21], which uses a BSAS with *Same Label* and IOU as affinity measures.

The BSAS is a basic clustering algorithm that sequentially groups detections that meet a minimum threshold for affinity. For each detection, if the affinity with any existing cluster center meets the minimum requirement, the detection joins the cluster.

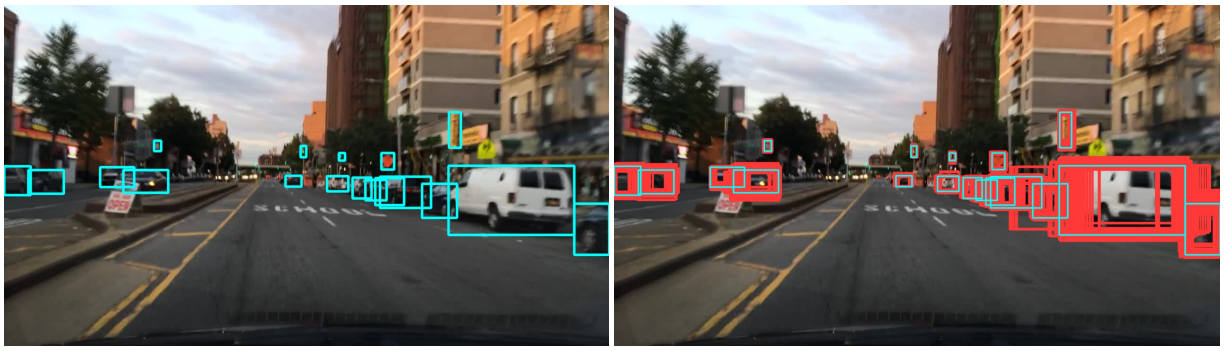
The *Same Label* affinity measure is a semantic affinity measure that requires a detection to have the same predicted class label as the cluster center. The IOU affinity measure is a spatial affinity measure, as explained in Section 3.1.3, that compares the location and shape between the detection and 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.

With these final clusters, we can estimate the predictive distributions for each detected object: a categorical distribution for the classification scores and a multivariate Gaussian distribution for the coordinates of each bounding box prediction. With these estimated distributions, we can measure the aleatoric uncertainty. The steps to estimate the distributions are explained in detail in Section 3.3.

In Figure 3.8, a practical example visualization of the final clusters for an image is given. The cluster

centers are represented in blue and the bounding boxes associated to each cluster in red.



(a) Cluster Centers **(b)** Final Clusters

Figure 3.8: Practical representation of the final clusters. For each cluster, a cluster center (in blue) is surrounded by the predictions belonging to that cluster (in red).

3.3 Distribution Estimation

For a given image, at the end of the novel TTA pipeline for UQ in object detection, each object will be associated with a cluster of predictions. With these clusters of predictions, we can estimate the predictive distributions.

3.3.1 Categorical Distribution

For each bounding box in a cluster, regardless of the object detector used, a vector of classification scores is associated with it. The classification score is a value between 0 and 1, which reflects the confidence of an object being of a certain class category (for example, a car, person, or bicycle). The length of the vector depends on the dataset in which the detector has been trained on and reflects the number of different classes that the object detector can identify.

Based on the work developed by Kendall and Gal [11] and Miller et al. [20], 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 [11, 20], 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.

$$q_i \approx s_i = \frac{1}{n} \sum_{j=1}^{n} s_j \tag{3.1}$$

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:

$$p(y_i|\mathcal{I}, \mathbb{D}) \sim Cat(k, q_i) \tag{3.2}$$

Each cluster of boxes that is associated with a single object will then have a final average vector with the classification scores for all the possible different classes. In practical terms, the cluster center is always the prediction with higher classification score. Therefore, averaging the scores with any other bounding box in the cluster will lower the classification score for the predicted class, which has the highest confidence value.

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.

By averaging the classification scores of the multiple bounding boxes of a cluster, it is possible to have an estimation of the uncertainty associated with that detection. If the uncertainty is low, the other bounding boxes of the cluster will have results similar to the cluster center, which in turn will result in the highest classification score staying with a high value associated to it and a distribution of scores concentrated on that class. On the other hand, if there is high uncertainty, by averaging the results of the bounding boxes, lower scores are expected to the predicted class and a larger distribution of values between all classes.

3.3.2 Multivariate Gaussian Distribution

Regarding the localization task, the object detector will predict four values for the coordinates that determine the location of the bounding box. Again, for each final detection, a group of bounding boxes is associated to it and an estimation of the uncertainty in localization is also possible.

Following the same work [11, 20], as in Section 3.3.1, 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:

$$\overline{b_i} = \frac{1}{n}\sum_{j=1}^{n} b_j, \quad \Sigma = \frac{1}{n}\sum_{j=1}^{n}(b_j - \overline{b_i})(b_j - \overline{b_i})^T \tag{3.3}$$

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:

$$p(z|\mathcal{I}, \mathbb{D}) \sim \mathcal{N}(\overline{b_i}, \Sigma) \tag{3.4}$$

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 to 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.

4

Experimental Setup

Contents

In this chapter, the experiments performed will be explained in detail, as well as the setup needed for the implementation. In this dissertation, a novel TTA pipeline is used in the domain of autonomous driving for UQ in the object detection task.

Different experiments are performed to investigate what are some of the optimal parameters to use in different steps of the TTA method, which can be applied to other sampling-based methods. We study the effect that different bounding box selection criteria can have on the quality of the estimated uncertainty and predicted distributions in Experiment 1. For Experiment 2, two different IOU minimum thresholds are used for each task (classification and regression), and their effect on performance is explored. The impact of different augmentations in the TTA method is investigated in Experiment 3, and finally a comparison of the novel TTA method with other state-of-the-art methods such as MC Dropout [16, 18, 20–22], and Output redundancy [13] is performed in Experiment 4.

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 was chosen, and for the dataset, the Berkeley Deep Drive (BDD) [36]. A detailed explanation of these two essential components will be given in Sections 4.1 and 4.2.

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 [7]. Furthermore, the main inspiration for code development came from the work of Ali Harakeh [37], 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 [38]. The experiments were performed with an NVIDIA GeForce RTX 2070 SUPER graphics processing unit, provided by ISR - Instituto de Sistemas e Robótica.

4.1 Object Detector

The YOLOv5 architecture was chosen for the experiments carried out in this dissertation as the deep object detector. There are two main types of object detectors: one-stage and two-stage [39]. The one-stage detectors prioritize speed over accuracy, whereas in the two-stage detectors, the opposite occurs. Due to the importance of real-time usage capability in the task of autonomous driving, a one-stage detector is preferred.

There are multiple one-stage object detector architectures to choose from. Some have already been used in the literature for the study of UQ, such as Retinanet [18, 30, 34], Single Shot MultiBox Detector [13, 20, 21] and YOLOv3 [14, 16]. YOLOv5 has not yet been used in the literature for works in the domain of UQ. It is the first architecture from the YOLO family to be developed in the Python programming language and Pytorch framework, with a large community of contributors actively managing and

improving it. On that note, it proved to be a reasonable choice, especially when one considers that its performances are among the best in terms of real-time object detection in the COCO dataset [40]. Meanwhile, during the development of this dissertation, newer versions of the YOLO family architecture were released, such as the YOLOv7 [41], which is the current state-of-the-art real-time object detector in the COCO dataset.

There are various versions of the YOLOv5 architecture, differing in their complexity (number of parameters of the network) and, therefore, performing with different speeds and accuracies. The smaller version YOLOv5s was chosen.

The intricacies behind the YOLOv5 architecture will not be explained in detail in this dissertation, since the main purpose is to study the method of TTA, which would remain the same for any of the existing object detector architectures. The interested reader is directed to the work developed in [42], where a detailed description of the components of the YOLOv5 architecture is given. Regardless, some aspects of this deep object detector need to be mentioned in order to have a sufficient understanding of it.

YOLOv5 uses anchor boxes to produce output object predictions. When an image is forward-passed through the YOLOv5 architecture, the image is divided into a grid with a fixed number of grid cells according to its resolution. Each cell of the grid has a fixed number of anchor boxes associated with it of various shapes and sizes. Therefore, assuming that the resolutions of the images from a dataset are equal, the final number of output predicted bounding boxes will be equal for every image. This occurs with the chosen dataset BDD, which will be explained in more detail in the next section.

Finally, it is important to note that the YOLOv5s model used for the experiments in this dissertation was pre-trained with the COCO dataset [35] and therefore can localize and classify 80 different classes.

4.2 Dataset

The dataset chosen is the BDD, a diverse driving dataset for heterogeneous multitask learning [36]. The dataset is composed of 100000 images taken while driving under a set of different conditions. The 100000 images are partitioned into three separate groups: Training set (70000 images), Test set (20000 images), and Validation set (10000 images).

As a pre-trained YOLOv5 architecture was used, the training set was not needed. Inference was performed on the images of the validation set since, apart from the training set, it is the only other partition with associated ground truth annotations for each image. Therefore, the test set is also not used.

The original BDD dataset identifies 10 different category classes. Following the work developed in [30, 37], 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*. Technically, the detector can accurately identify objects of this class in the image, but will classify them as *person*. This occurs because the detector did not learn to distinguish between a pedestrian and a person sitting on top of a bicycle or motorcycle.

In the domain of autonomous driving, the distinction between person and rider is, in fact, very important and in practice should not be neglected. It represents the difference between a pedestrian and a person, which should be considered as a part of a vehicle. For example, if there is a motorcyclist in the middle of the road, it is relevant that the detector distinguishes it from a pedestrian. Otherwise, it could cause the automated car to brake or change direction to avoid hitting the pedestrian, which would be an exaggerated reaction.

Taking this into account, to solve this issue we decided to consider the *rider* class equal to a *person* ground truth label. Although there will be no evaluation regarding the ability to distinguish a person from a rider, there is confirmation that the model is detecting a person on top of the bicycle/motorcycle, which is important and should not be considered as an incorrect detection.

Out of the approximately 120000 ground truth detections present in the 10000 images of the validation set, only around 700 are from the *rider* class, so even though there will be an impact of this decision, the impact is expected to be small.

4.3 Output of YOLOv5 in BDD

The YOLOv5 architecture receives an image as input and outputs a vector with the necessary information and predictions for that specific image. For the BDD images, the output vector is of dimensions (15120,85).

The value 15120 refers to the total number of anchor box predictions and depends on the resolution of the input image. Regarding the second dimension of the vector, which has 85 values, the first four values refer to the coordinates of the bounding box in the format (X,Y,Width,Height), where X and Y are the coordinates for the center of the bounding box and Width and Height refer to how wide and tall the bounding box is. This coordinate system has been shown previously in Figure 1.1, but is repeated here for the convenience of the reader in Figure 4.1.

The fifth value refers to the objectness score, a value that ranges from 0 to 1, representing a confidence score with respect to the presence of an object in that bounding box. Finally, the remaining 80 values refer to the 80 different classes that this architecture is able to localize and classify, since it was trained with the COCO dataset [35]. For each class, a value from 0 to 1 is assigned, which represents

Figure 4.1: Sample image from Coco Dataset, representing the YOLOv5 Bounding Box Coordinate System [1]

the confidence score that the object inside the bounding box is of that specific class. The final confidence score for each class is calculated by multiplying the objectness score by every class confidence score. Again, it is important to mention that of these 80 classes, only six are relevant for this dataset, namely: person, car, bus, truck, motorcycle, and bicycle.

To better understand the output of the YOLOv5 architecture, an example visualization is represented in Figure 4.2. With all 15120 bounding boxes, the underlying input image is almost unrecognizable (each bounding box is drawn red on top of the image). Hence, a display with only 10000 of those bounding boxes is represented for better visualization. We can see how the bounding boxes are placed at every possible location of each image to try and locate objects everywhere.

(a) Original Image **(b)** 10000 anchors

Figure 4.2: Output predictions from YOLOv5

4.4 Experiments

4.4.1 Experiment 1 - Bounding Box Selection Criteria and IOU Thresholds

As defined in Chapter 2, the technique or rule that decides which predictions are kept after each run is called a bounding box selection criterion. After each forward-pass in sampling-based methods, a bounding box selection criterion is applied and only certain predictions are accumulated to estimate the uncertainty. The current established technique uses NMS as the bounding box selection criterion and an IOU $= 0.95$ for the clustering step [20–22].

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 bounding box selection criterion, we propose an investigation of different IOU thresholds in the clustering step.

To conduct the proposed experiment and evaluate the results, some parameters must be fixed. The object detector chosen to perform the inference on the images is YOLOv5 and the chosen images are the test set taken from the BDD dataset. The number of runs performed in the TTA method is fixed at 10 and the augmentation used is the Gamma Augmentation, with gamma factors in the interval $[0.4; 2]$ in steps of $\frac{2-0.4}{10-1} = 0.178$.

Therefore, the gamma factors used are $[0.4, 0.578, 0.756, 0.934, 1.112, 1.29, 1.468, 1.646, 1.824, 2]$. The original image corresponds to a gamma correction equal to 1. We do not use the original image, but there are two factors, 0.934 and 1.112, which are very similar to the original image.

Different bounding box selection criteria will be used to accumulate different sets of bounding boxes after 10 runs. The chosen bounding box selection criterion will be based on the maximum confidence score obtained for a detection and will be compared to the established technique NMS. A visualization of the final accumulated bounding boxes with different criteria will be provided in Chapter 5. The different criteria chosen to be investigated 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

After the step of accumulating bounding boxes, NMS is used to obtain the final detections and based on these final detections, which will be considered cluster centers, the clustering step described in Chapter 3 is used. A bounding box is incorporated into a cluster only if it classifies the same class as the cluster center and only if it has an IOU with the cluster center above a certain threshold.

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: $[0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95]$.

Summarizing the different fixed and varying parameters for Experiment 1:

- Fixed Parameters: YOLOv5, BDD, Gamma Augmentation $[0.4; 2]$, 10 runs
- Varying Parameters:

  Bounding box selection criteria: $[\text{NMS}, 0.1, 0.01, 0.001, 0.0001]$

  IOU thresholds: $[0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95]$

4.4.2 Experiment 2 - Two Different IOU Thresholds for Regression and Classification

In object detection, two tasks are performed to detect an object. A regression task to predict the bounding box coordinates and a classification task to predict the category of the object. In the literature, the established technique [20–22] uses a single IOU threshold in the clustering step to group detections for every object. The defined group of detections is used to estimate uncertainty for both the regression and classification tasks.

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 and predictive distributions. 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.

In principle, to guarantee a correct measure of uncertainty in the regression task, smaller values of IOU should be used to ensure that all relevant bounding boxes are included in the clusters for the estimation of the distributions. Not only that, but higher IOU thresholds create clusters with only bounding boxes that are concentrated around the cluster centers, which could lead the cluster variance to be underestimated.

Inversely, in the classification task, smaller IOUs can lead to the inclusion of bounding boxes that have lower accuracies and possess confidence scores of very low orders of magnitude, which could cause biases in the final distributions. A practical visualization of different IOU thresholds in the remaining bounding boxes will be given in Chapter 5.

Taking this into account, the intervals of IOUs chosen for each task were:

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

As the main goal is to investigate multiple combinations of IOU thresholds for each task, all other parameters must be fixed. The only difference in the fixed parameters from Experiment 1 is the bounding box selection criterion, which is set at 0.01. Therefore, the object detector is YOLOv5, images are from the test set of the BBD dataset, 10 runs are performed, and in each run a gamma augmentation is performed with gamma factors in the interval $[0.4, 2]$ in steps of 0.178.

In summary, for experiment 2:

- Fixed Parameters: YOLOv5, BDD, Gamma Augmentation [0.4;2], 10 runs, 0.01 bounding box selection criterion
- Varying Parameters:

  IOU Regression: $[0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65]$

  IOU Classification: $[0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1]$

4.4.3 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.

With this in mind, 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. The number of runs used is equal to 10. By fixing these parameters, the experiment will focus only on the impact of each augmentation.

Four different augmentations will be tested. For each of these, an interval of possible augmentation factors was previously defined, as mentioned in Section 3.1.2. These intervals take into account the autonomous driving domain, making sure that any augmentation used will transform the images while still guaranteeing that they look realistic and are part of the dataset they were extracted from. As in Experiments 1 and 2 for the gamma augmentation, for each interval of possible factors, incremental factors are used 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.

4.4.4 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 in this experiment 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 the *Same Label* and IOU as affinity measures. The MC dropout will use NMS as a 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 [13].

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 [30], a dropout rate of 0.1 was inserted before the final convolutional layer of the YOLOv5 architecture. Consequently, in every run, 10% of the final layer weights will be randomly deactivated (zeroed out).

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 to the TTA method. This will provide a fairer comparison to the best performing TTA method, as well as giving a further look into ways of possibly improving the state-of-the-art methods. These modifications could occur in the bounding box selection criterion used (in case of Output Redundancy, the usage of one,

and for the MC dropout possibly a different one), as well as two different IOU thresholds for each task.

4.5 Evaluation

4.5.1 True Positives, False Positives and False Negatives

To evaluate object detection methods, it is important to partition predicted detections into different groups such as True Positives (TP), False Positives (FP) and False Negatives (FN).

In the literature, there are different rules for defining these partitions. In this dissertation we use the rules from the PASCAL VOC challenge [43], adding only an extra requirement for the considered TP detections, to account for the autonomous driving domain. A visualization of the rules and different partitions can be seen in Figure 4.3.

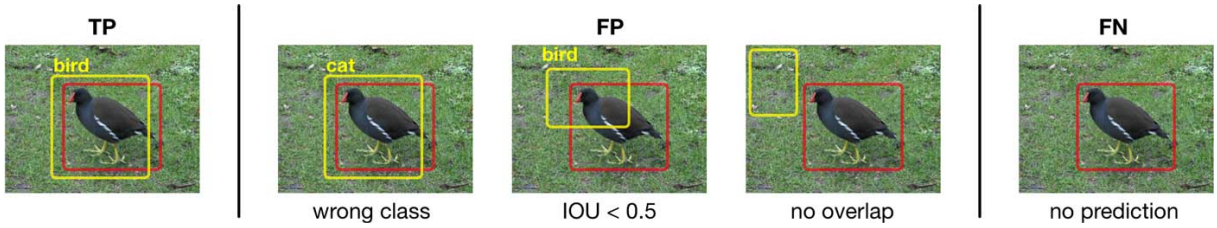


Figure 4.3: Rules for partitioning detections [3]

We note that in the PASCAL VOC challenge [43], only the requirement of an IOU above 0.5 is used to define a detection as a TP. Taking into account the autonomous driving domain, where safety is of the utmost importance, identifying an object with its correct location but failing to distinguish, for example, a pedestrian from a vehicle can have drastic consequences. Because of that, a detection will only be considered a TP if it performs well in both localization and classification tasks.

Therefore, 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.

4.5.2 Metrics for Evaluation

4.5.2.A Mean Average Precision

Based on the previous definition for TP, FP and FN, two metrics can be used to evaluate the quality of the predictions of an object detector: precision and recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{4.1}$$

One issue regarding the use of precision and recall as metrics to evaluate object detection is their dependence on the thresholds for the minimum confidence score and IOU. Everingham et al. [43] resolved the dependence on the confidence score threshold by defining Average Precision (AP), the area below the continuous Precision-Recall (PR) curve. The area is obtained by performing a numeric integration over a finite number of sample points on the PR curve [43]. By varying all confidence score thresholds between 0 and 1, the PR curve is built by estimating the precision and recall at each different threshold. The AP is estimated for each class, and, averaging the computed AP over all classes, the mAP is obtained. The issue of dependence with the IOU threshold was tackled in [35], by averaging the mAP for multiple IOU thresholds in the interval $[0.5, ..., 0.95]$, a metric known as the COCO mAP. COCO mAP was the metric utilized in this dissertation to evaluate detections, and therefore will be referred to simply as mAP for the remainder of this work. The maximum mAP achievable by a detector is 100%.

4.5.2.B Negative Log Likelihood

The NLL is an evaluation metric that allows the evaluation of the quality of the predicted distributions. The notation utilized to describe this metric in this dissertation is based on the work developed by Ali Harakeh in his Ph.D. thesis [10].

For the classification task, with y as the output of the neural network object detector, x the features from an input image and θ the parameters of the neural network object detector. Let $p(y|x;\theta)$ represent the estimated predictive categorical distribution, described as $Cat(p_1(x;\theta), ..., p_K(x;\theta))$, with $p_k(x;\theta) \in [0,1], \sum_{k=1}^{K} p_k(x;\theta) = 1$, where K refers to the number of possible different predicted classes.

Given Y as the associated ground truth target represented as a one-hot vector, a measurement of distribution quality can be given by equation 4.2.

$$\text{NLL}(p(y|x;\theta), Y) = \sum_{k=1}^{K} -Y_k \log p_k(x;\theta) \tag{4.2}$$

For the regression task, with z as the output of the neural network object detector, x the features from an input image and θ the parameters of the neural network object detector. Let $p(z|x;\theta)$ represent the es-

timated multivariate Gaussian distribution, with $p(z|x;\theta) = \mathcal{N}(\mu(x;\theta), \Sigma(x;\theta))$, where $\mu(x;\theta)$ represents the predicted vector of mean coordinates and $\Sigma(x;\theta)$ the predicted covariance matrix.

Given Z as the multivariate vector with the ground truth target coordinates, a measurement of the predictive distribution quality for the regression task can be given by

$$\text{NLL}(p(z|x,\theta), Z) = \frac{1}{2}(Z - \mu(x_n;\theta))^T \Sigma(x_n;\theta)^{-1}(Z - \mu(x_n;\theta)) + \frac{1}{2}\log\det\Sigma(x_n;\theta) \tag{4.3}$$

Since in order to calculate this metric a target value is needed, only the TP detections can be used to estimate it, as these are the only ones that are associated to a ground truth target. Therefore, only the quality of the predicted distributions of the TP will be evaluated. It is important to note that the measured NLL for the regression task can be influenced by the estimated categorical distribution, as the predicted class for each detection will depend on it and therefore impact the created TP and FP partitions. The best possible results for NLL by a detector occur when NLL is equal to 0.

4.5.2.C Minimum Uncertainty Error

The Uncertainty Error (UE) was first developed by Miller et al. [21] to evaluate probabilistic object detectors. UE measures the ability to distinguish between TP and FP detections based on an entropy threshold δ_{ent}.

$$UE(\delta_{ent}) = 0.5\frac{|H(TP) > \delta_{ent}|}{|TP|} + 0.5\frac{|H(FP) \leq \delta_{ent}|}{|FP|}, \tag{4.4}$$

where H is the measured entropy of the predictive distributions.

For the classification task, as defined in Chapter 3, a categorical distribution is assumed over all confidence scores and the entropy is computed by $H(q) = \sum_{i=1}^{k} -q_i \cdot log(q_i)$, where q is a vector with k different class probabilities. For the regression task, the entropy is calculated by $H(\mathcal{N}(\mu,\Sigma)) = \frac{1}{2}\ln det(2\pi e\Sigma)$, where $\mathcal{N}(\mu,\Sigma)$ is the predicted 4-dimensional multivariate Gaussian distribution for each detection.

Explaining equation 4.4 in more detail, $|H(TP) > \delta_{ent}|$ represents the number of TP detections above the entropy threshold and $|TP|$ the total number of TP detections. Whereas $|H(FP) \leq \delta_{ent}|$ is the number of FP below the entropy threshold and $|FP|$ the total number of FP detections. The UE is computed for all possible entropy thresholds δ_{ent}.

The UE varies between 0 and 0.5, where having a score of 0.5 makes it indifferent to use a threshold to separate true and false positive detections, and a score of 0 means that there is an entropy threshold value where it is possible to exactly separate those detections.

The best achievable UE by a detector over all possible entropy thresholds is called MUE. In [18, 21], MUE is used to compare probabilistic object detectors, and we will also use it in this dissertation.

5

Results and Discussion

Contents

In this chapter, the results obtained in the experiments performed are presented together with a relevant discussion and possible conclusions that can be drawn from the observed results.

5.1 Experiment 1 - Bounding Box Selection Criteria and IOU Thresholds

The main goal of Experiment 1 is to investigate 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.

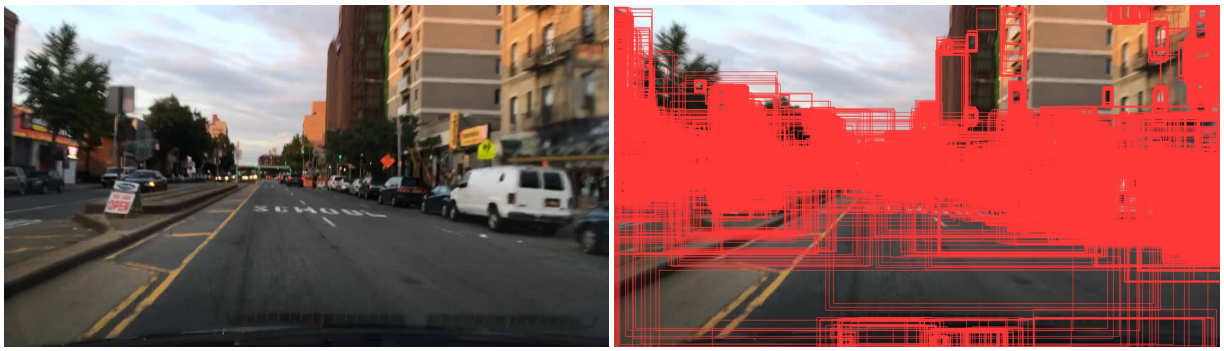
Before discussing the results obtained for Experiment 1, in Section 5.1.1 a practical visualization of the different bounding box selection criteria is given. Furthermore, Section 5.1.2 shows the effect of different IOU thresholds on the final clusters of bounding boxes for each detection.

5.1.1 Visualization of Different Bounding Box Selection Criteria

To further illustrate the impact of each bounding box selection criterion on the final accumulated detections, multiple figures were extracted for a single input image where the remaining accumulated bounding boxes after the 10 runs were drawn over it, as shown in Figure 5.1. The different criteria used were the resulting boxes from NMS and a minimum threshold for the bounding box maximum confidence score in the range: $[0.1, 0.01, 0.001, 0.0001]$.

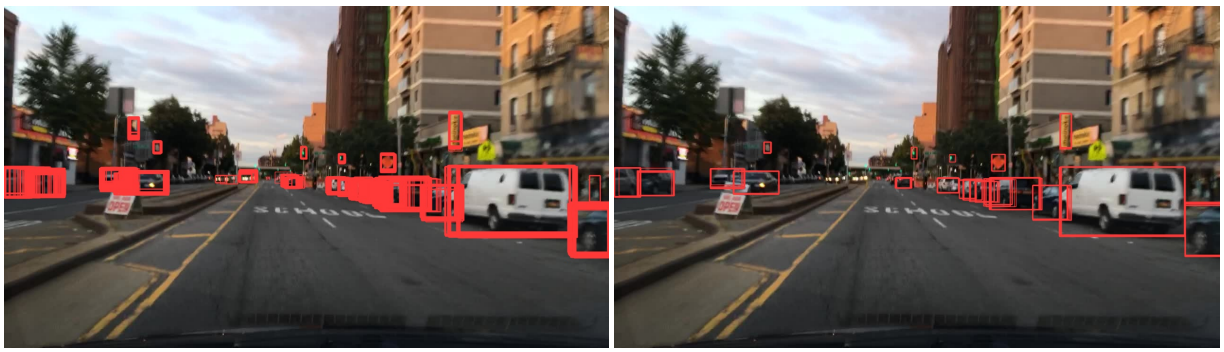
As the threshold for the confidence score gets stricter, fewer bounding boxes are kept, but with higher levels of predictive quality in both localization and classification. Bounding boxes that are located in background regions of the image, such as the sky, are removed. The bounding boxes that are kept are located in areas of the image where objects are more likely to be present.

Having a small amount of highly precise bounding boxes can cause the method to miss out on possible vital information for the measurement of the uncertainty associated with the final detections, especially in terms of location. At the same time, having too many bounding boxes can become problematic if these predictions cause major deviations on the estimated distributions of the clusters to which they will be attributed, since the values for their confidence scores will be of smaller orders of magnitude.



(a) Original **(b)** 0.0001

(c) 0.001 **(d)** 0.01



(e) 0.1 **(f)** NMS

Figure 5.1: Impact of different bounding box selection criteria on the final accumulated bounding boxes

5.1.2 Visualization of Different IOU Thresholds

In Figure 5.2, different IOU thresholds for the clustering step are depicted. In this example, the criterion used was 0.01 for the accumulated bounding boxes, as in Figure 5.1(d). With lower IOU values, each cluster has more bounding boxes associated with it. Higher IOU thresholds limit the number of bounding boxes in each cluster, where only those boxes with similar position and shape are grouped together.

(a) Original

(b) IOU: 0.15

(c) IOU: 0.35

(d) IOU: 0.55

(e) IOU: 0.75

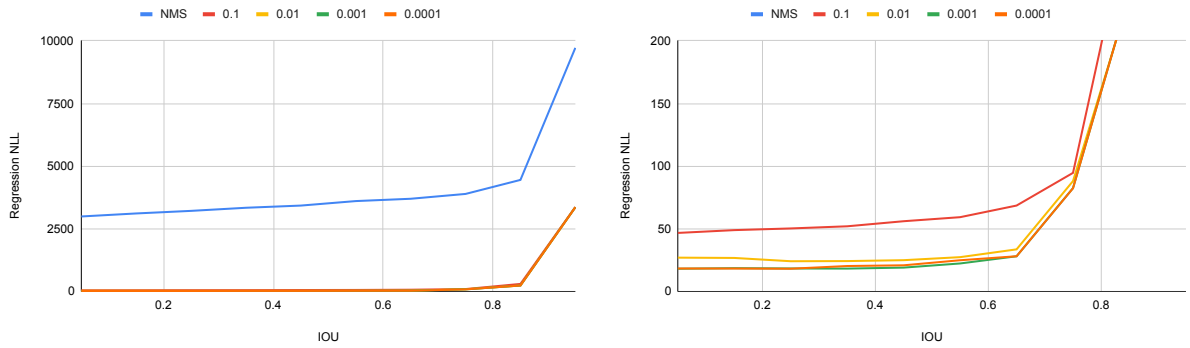
(f) IOU: 0.95

Figure 5.2: Impact of different IOU thresholds in the final clusters.

5.1.3 Results for Experiment 1 - Negative Log Likelihood

5.1.3.A Regression Task

The results obtained for the NLL of the regression task are shown in Figure 5.3.



(a) NLL Regression vs IOU

(b) NLL Regression vs IOU (zoomed in)

Figure 5.3: NLL for the regression task versus IOU for different bounding box selection criteria

As the IOU threshold values increase, the quality of the predicted distributions worsens. This is expected since higher IOUs will restrict the number of bounding boxes attributed to each final cluster, gathering only those with similar localization, as seen in Figure 5.2, which in turn can result in missing valuable information for the accurate estimation of the distributions. For example, detections with higher uncertainties will mistakenly portray a lower measured uncertainty, since the bounding boxes of lower quality will not be included. With lower IOU values, it is then possible to gather more relevant information to obtain higher quality predicted distributions.

Regarding the different bounding box selection criteria, NMS performs the worse, even when lower IOUs are used. One possible explanation is the fact that this technique filters out too many bounding boxes, only leaving the most accurate ones. For the other criteria, however, 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.

The best results for the task of localization are then obtained by selecting only the bounding boxes after each run with at least 0.0001, 0.001 or 0.01 for the classification scores. These are the criteria that provide the most accurate predicted distributions for the regression task.

5.1.3.B Classification Task

In Figure 5.4 the NLL results for the classification task are shown. Almost the inverse of what was previously discussed for the regression task is observed. With higher IOU values, better results are obtained

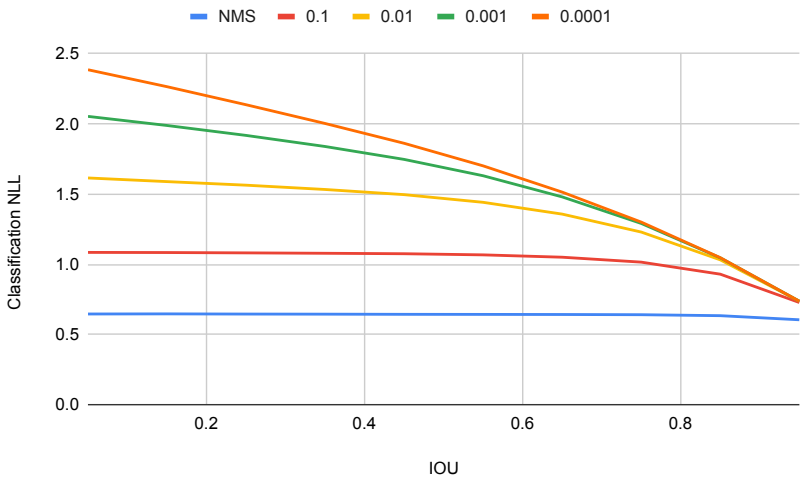


Figure 5.4: NLL for the classification task versus IOU for different bounding box selection criteria

for each bounding box selection criterion and the stricter the criterion used, the higher 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 predicted distributions worsen.

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. With lower IOU thresholds, more boxes will be used to estimate the categorical distribution. These extra boxes typically have lower confidence scores and can lead to lower quality distributions, especially when the confidence scores are of lower orders of magnitude.

The best results for the predicted distributions in the classification task are then obtained by using NMS, followed by a confidence score threshold of 0.1 until 0.0001.

In a broader view of the results obtained for the NLL, taking into account both the regression and classification tasks, one could argue for the use of different bounding box selection criteria for each task. NMS or 0.1 for the classification task and 0.0001, 0.001 or 0.01 in the regression task. However, having different criteria and keeping two different groups of bounding boxes at the same time to be post-processed (creation of clusters and calculation of distributions) during inference increases the complexity of the algorithms. With this in mind, a possible compromise can be made between the accuracy of the regression and the classification. The worst performers for the regression task were NMS and 0.1 and for the classification task 0.001 and 0.0001. A possible compromise between the performance of both tasks could then be to use the threshold of 0.01.

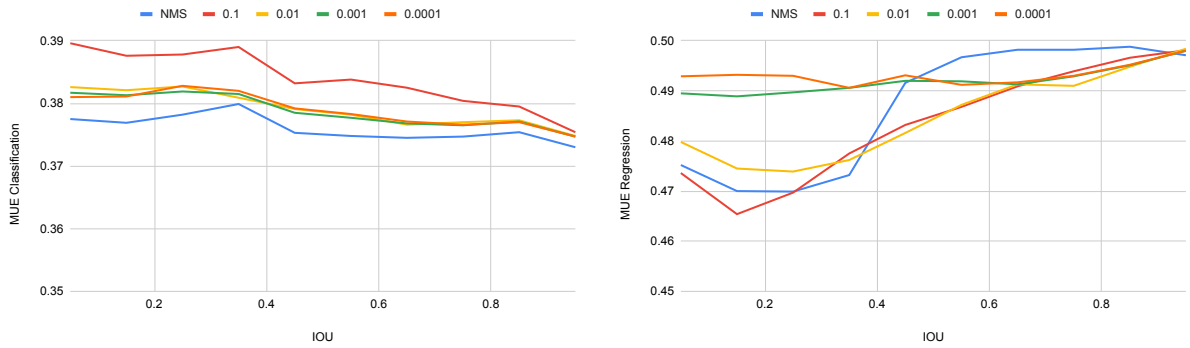
In a practical way, the bounding box selection criteria NMS and 0.1 provide very few final bounding

boxes that will leave out vital information for the regression task, regardless of the used IOU. On the other hand, having a very broad criterion will include too many bounding boxes, which can cause worse results in the classification task. However, in this situation, choosing a higher IOU can improve the results obtained.

One final observation regarding the IOU thresholds is that lower IOUs are preferred for the regression task while higher IOUs are preferred for the classification task. An interesting novelty solution that has never been previously performed in the literature is to use two different thresholds of IOU for each task, a smaller IOU for regression and a higher IOU for classification. This could lead to overall better results, and it will be further investigated in Experiment 2.

5.1.4 Results for Experiment 1 - Minimum Uncertainty Error

Analysing the results obtained for the MUE metric in Figure 5.5, it is possible to understand how the measured uncertainty can be used to distinguish TP from FP. Ideally, FP detections will have higher uncertainty values associated to them, in terms of localization or classification, which, in turn, can be used to eliminate them.



(a) MUE Classification vs IOU

(b) MUE Regression vs IOU

Figure 5.5: MUE versus IOU for different bounding box selection criteria

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 5.5(b), the worst performances occur with a bounding box selection criteria of 0.0001 and 0.001, while the other three have similar results. It is important to note here that the best values obtained are around the 0.15-0.25 IOU mark, and values below that actually tend to perform worse. Therefore, there is an optimal IOU value threshold, 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 5.5(a), the results improve with higher values of IOU. 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.

Again, taking these results into account, the trade-off mentioned in Section5.1.3.B continues to be reasonable, since the 0.01 criterion provides good performances both for the classification task and for the regression task.

5.2 Experiment 2 - Two Different IOU Thresholds for Regression and Classification

For Experiment 2, the bounding box selection criterion is fixed at 0.01, since the purpose of this experiment is to investigate the impact of using two different IOU thresholds for the regression and classification task. 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.

In principle, to ensure a correct measurement of uncertainty for the regression task, it is important to use smaller values of IOU so that all relevant bounding boxes are included. Not doing this could lead to perceive every single detection as having low uncertainty, since only the boxes that share high similarity in localization with the cluster center will be kept.

Taking this into account, the interval of IOUs chosen for the regression task were of smaller values: $[0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65]$.

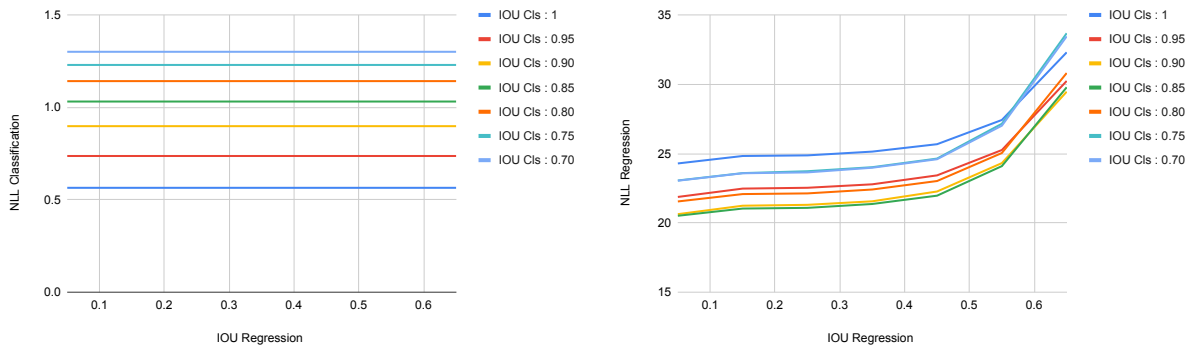
Inversely, for the classification task, smaller IOUs can cause the inclusion of low-accuracy bounding boxes (with confidence scores of very low orders of magnitude), which can cause biases on the final distributions.

Therefore, for the classification task, higher IOUs are preferred. The chosen interval values were as follows: $[0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1]$.

From now on, the IOU threshold chosen for the regression task will be mentioned as IOU_r and the IOU threshold chosen for the classification task as IOU_c.

In summary, for each object, two different clusters of detections are created to estimate the predictive distributions. A group of detections is created with the IOU_c threshold and will be used to estimate the categorical distribution for the classification task. The other group is created with the IOU_r threshold and will be used to estimate the multivariate Gaussian distribution for the regression task.

5.2.1 Results for Experiment 2 - Negative Log Likelihood



(a) NLL Classification Task

(b) NLL Regression Task

Figure 5.6: NLL in the regression and classification tasks for different combinations of IOU_r and IOU_c

For the classification task, in Figure 5.6(a), first 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. The worst result was obtained with the threshold at 0.70. The results are expected to worsen as the threshold value is lower.

Analyzing now Figure 5.6(b) with the results of the regression task, a different behavior is visible. For different IOU_c thresholds, improvements are obtained as the threshold IOU_r decreases. In a practical sense, this effect can be explained by the fact that, as more detections are collected for each object, the quality of the predicted distributions increases.

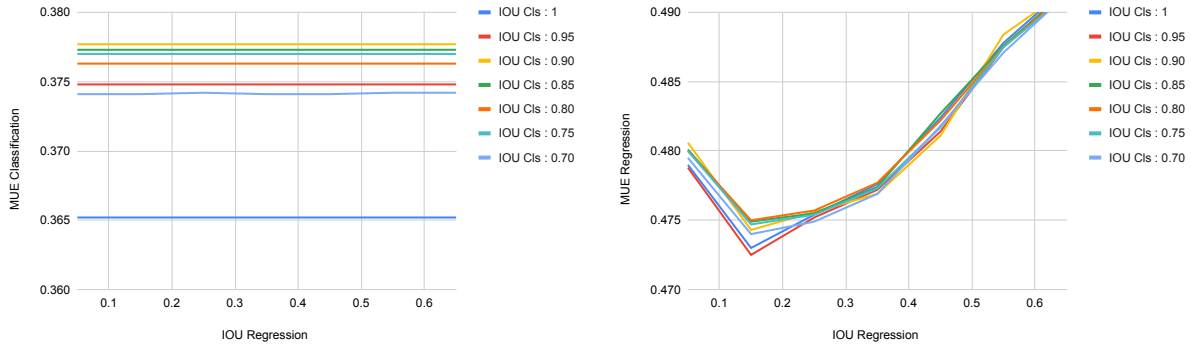
In the regression task, an impact of the IOU_c threshold can be observed. This happens because the IOU_c threshold can determine which detections are incorporated in the TP partition where the NLL will be evaluated. As the estimated categorical distributions vary with each IOU_c value, the predicted class for each detection can also vary, creating different TP and FP partitions.

The worst results in regression occurred when IOU_c equals 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. A 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.

In summary, the best possible result for NLL in the classification task is obtained with an IOU_c threshold of 1, regardless of the IOU_r threshold. However, for the regression task, the best results are obtained

when the IOU_c threshold is equal to 0.85 or 0.90 and the IOU_r threshold is equal to 0.05.

5.2.2 Results for Experiment 2 - Minimum Uncertainty Error



(a) MUE Classification Task

(b) MUE Regression Task

Figure 5.7: MUE in the regression and classification tasks for different combinations of IOU_r and IOU_c.

Regarding the quality of the measured uncertainty for the classification task, shown in Figure 5.7(a), 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 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, 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 a great improvement occurs, 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 5.7(b) and the regression task, all IOU_c thresholds perform very similarly, although there is improvement in performance for the IOU_c values of 1 and 0.95. The influence of the IOU_c in the regression task results can be explained by the impact that different IOU_c values have on the generated TP and FP partitions that will be used to evaluate the MUE metric.

The IOU_r threshold has a great impact on the performance obtained. There is an optimal threshold IOU_r at 0.15, thus, 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.

Taking into account all the results in this experiment for both tasks and metrics, to obtain the best possible performance with respect to MUE (regression and classification) and NLL (classification), the

best combination of thresholds is 0.15 for the IOU_r threshold and 1 for the IOU_c threshold. For the best NLL in the regression task , the best combination is 0.05 IOU_r and 0.90 or 0.85 IOU_c.

5.3 Experiment 3 - Different Augmentations for UQ in TTA

The main goal of Experiment 3 is to study the impact on performance of utilizing different color augmentations in the quality of the predicted distributions and measured uncertainty. With this in mind, the bounding box selection criterion will be fixed at 0.01 as in Experiment 2 and a IOU_c threshold of 1 and a IOU_r threshold of 0.15 will be used, as these were one of the combinations that were shown to provide better performance in Experiment 2.

Four augmentations were used: Gamma, Contrast, Brightness, and Gaussian Blur. We refer the reader to Section 3.1.2, where these transformations were previously shown.

For each of these augmentations, an interval of possible augmentation factors was defined previously, as mentioned in Chapter 3, taking into account the domain of the images and making sure that they still look realistic after the augmentations.

The intervals were defined as follows:

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

Table 5.1: Results Obtained for 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

Analyzing the results obtained in Table 5.1 with regard to mAP, contrast and brightness performed the best, followed by gamma augmentation, while Gaussian blur provided the worst performance. An important practical note to take also regarding the Gaussian blur is the fact that the speed of the method drastically decreases due to the nature of the augmentation, which includes convolutional operations across the whole image.

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 the 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.

Taking into account the results obtained for each of the augmentations, where the Gaussian blur performed the worst, and noticing how contrast, brightness, and gamma perform best for different tasks, one possible way of improving the performance even further is to combine different augmentations together.

On that note, more experiments were performed that combined the three best augmentations previously reported: gamma, contrast, and brightness. It could be possible that each augmentation is responsible for improvements in certain metrics, and with a combination of them, a better overall performance can be obtained. The augmentations were combined as pairs, where each augmentation was used for half of the total 10 runs (5 runs each).

Table 5.2: Results Obtained for different Augmentations

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

The combination of augmentations further improved some of the results previously obtained. Better performances were achieved for the mAP, NLL and MUE measured in the classification task. For the regression task however, no noticeable improvements were achieved.

The pair with the greatest improvement was the combination of brightness and contrast, which improved mAP, NLL and MUE in classification. This illustrates 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.

5.4 Experiment 4 - Comparative Study of UQ Methods

In Experiment 4, the main goal is to compare the performance of the novel TTA method with the performance of the current state-of-the-art methods (MC Dropout and Output Redundancy) as well as the baseline deterministic object detector.

It is important to note that the baseline deterministic YOLOv5 detector will only be compared in terms of the mAP metric, since this detector is not able to estimate uncertainty and predictive distributions.

The TTA method utilized here for comparison is the one with a bounding box selection criterion of 0.01, 0.15 for IOU_r and 1 for IOU_c, with contrast and brightness as augmentations, since this was the one with the overall best performance. The results for each method are shown in Table 5.3.

TTA performed better for all metrics in comparison to the current state-of-the-art methods and base-

Table 5.3: Results Obtained for each method

	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**

line, 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, MC Dropout uses NMS as the bounding box selection criterion, which was shown in Experiment 1 to be the worst performing criterion for estimating predictive distributions in the regression task.

Output Redundancy does not use a criterion and shows improvements in comparison to the MC Dropout. However, the results obtained by this method are still drastically worse than the ones achieved by the novel TTA method to estimate predictive distributions on the regression task.

Taking this into account, a further study was performed to investigate possible improvements to the MC Dropout and Output Redundancy methods by using the optimal parameters obtained from the experiments performed in this dissertation.

Each method will now apply a bounding box selection criterion of 0.01 and in the clustering step, the optimal combination IOU_r of 0.15 and IOU_c of 1. The obtained results are presented in Table 5.4.

Table 5.4: Results Obtained for each method. *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

As expected, the results 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 task was obtained, again proving the relevance of using the correct criteria and IOU threshold for each task.

The lower mAP scores of MC Dropout are expected, since the method itself causes the deactivation of a portion of the YOLOv5 architecture in each run, which will cause the quality of the detections to drop.

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

With this experiment, it is shown that the use of the TTA method can become a very good option for quantification of uncertainty in object detection in the domain of autonomous driving. Not only that, but already in use state-of-the-art methods could have their performances vastly improved by using two different IOU thresholds for each task and incorporating a bounding box selection criterion.

6

Conclusions and Future Work

Contents

The main goal of this dissertation was to research novel approaches to characterize uncertainty in deep learning detection methods for the autonomous driving domain. That goal was achieved by using a novel method of TTA for the first time to estimate aleatoric uncertainty in the object detection task. Not only that, but we showed that it achieves better performance than the current sample-based state-of-the-art method MC Dropout [16, 18, 20–22] and Output Redundancy [13].

Another goal was to improve current state-of-the-art methods, which was also achieved by introducing novel modifications to the uncertainty quantification pipeline. Improvements were shown in both MC Dropout and Output Redundancy methods by adding a bounding box selection criterion and using different IOU thresholds for each task.

6.1 Achievements

In this section, an overview of the main conclusions and accomplishments achieved with the experiments performed in this dissertation is listed:

- We show that the novel TTA method for UQ is able to outperform the baseline and state-of-the-art methods, MC Dropout and Output Redundancy, for all metrics except NLL in the classification task, where it obtained results similar to the MC Dropout.

- We show that the use of a bounding box selection criterion can improve the quality of the predictive distributions and the uncertainty measured. The use of different bounding box selection criteria for each task could further improve performance.

- The lower thresholds IOU are shown to generate the best results for the regression task, while the higher IOU thresholds produce the best results for the classification task, in both predictive distributions and measured uncertainties.

- We show that optimal results can be achieved with two different IOU thresholds for each task. The best combinations obtained were with $\{\text{IOU}_r = 0.15, \text{IOU}_c = 1\}$ and $\{\text{IOU}_r = 0.05, \text{IOU}_c = 0.90/0.85\}$

- Improvements in current state-of-the-art methods were obtained by using an optimal bounding box selection criterion and two different IOU thresholds for each task.

- For the autonomous driving dataset BDD, we show that contrast and brightness produced the best quality predictive distributions and uncertainty estimations and that the use of multiple augmentations together can further improve performance in the classification task.

6.2 Future Work

As TTA was used in this dissertation for the first time in object detection to quantify uncertainty, various other studies can be performed to further evaluate and possibly improve the performance of this method.

6.2.1 Evaluation under dataset shift

A dataset shift occurs when the training and test distributions are different. In [30], Feng et al. perform evaluations for different UQ methods with and without dataset shift. They train an object detector with data from the BDD [36] dataset and then evaluate its performance on test data from the same BDD dataset (evaluation without dataset shift) and with data from the KITTI [44] and Lyft [45] datasets (evaluation with dataset shift).

In this dissertation a pre-trained architecture in the COCO dataset [35] was used to perform experiments. A possible future investigation would be to perform experiments similar to those in [30] regarding dataset shifts. For that, the detector should be trained with custom data from an autonomous driving dataset and evaluated on unseen data from that same dataset and also from other autonomous driving datasets. Different performances are to be expected due to the changing of the meta-data of each dataset, such as resolution of the images, position of cameras or even locations of the captured images.

6.2.2 Comparison with non sampling-based methods

Referencing again the work by Feng et al. in [30], sampling-based methods (MC Dropout and Deep Ensembles) were compared with non-sampling based methods (Direct Modeling and Output Redundancy). In this dissertation, the novel TTA method is only compared to the MC Dropout and Output Redundancy methods. Further investigative work could be performed in order to compare TTA performance with Direct Modeling methods and Deep Ensembles. Not only that, but incorporating MC Dropout together with the TTA method could provide measurements of both aleatoric and epistemic uncertainty simultaneously.

6.2.3 Practical evaluation of measured uncertainty

Le et al. in [13] use the uncertainty associated to each detection to eliminate possible incorrect detections. In principle, it is expected that FP detections have higher values of uncertainty associated to them, while TP low values of uncertainty. By defining an entropy threshold, they reject detections based on their measured uncertainty and evaluate the number of TP and FP before and after the removal. A measured uncertainty of greater quality should cause the number of TP to remain the same while FP detections are removed.

A similar investigation can be performed for the TTA method in order to evaluate the quality of its estimated uncertainty. Additionally, a study between different types of measured uncertainty and their impact on the removal of incorrect detections could be performed. Metrics of uncertainty such as the total variance and generalized variance (from the covariance matrices), shannon entropy (from the categorical distributions) or a mix between these could provide greater insights on this topic.

Bibliography

[1] G. Jocher. (2022) Train custom data. Accessed 22-June-2022. [Online]. Available: https://github.com/ultralytics/yolov5/wiki/Train-Custom-Data

[2] A. Rosebrook. (2016) Intersection over union (iou) for object detection. Accessed 02-October-2022. [Online]. Available: https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

[3] M. Hollemans. One-shot object detection. Accessed 16-October-2022. [Online]. Available: https://www.twblogs.net/a/5c45cae4bd9eee35b3a72b98

[4] 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.

[5] 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.

[6] 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. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[7] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, K. Michael, J. Fang, imyhxy, Lorna, C. Wong, Z. Yifu), A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, P. Skalski, A. Hogan, M. Strobel, M. Jain, L. Mammana, and xylieong, "ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations," Aug. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7002879

[8] 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.

[9] 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. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[10] A. Harakeh, “Estimating and evaluating predictive uncertainty in deep object detectors,” Ph.D. dissertation, University of Toronto, 2021. [Online]. Available: https://tspace.library.utoronto.ca/bitstream/1807/108903/3/Harakeh_Ali_202111_PhD_thesis.pdf

[11] 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. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

[12] D. Hall, F. Dayoub, J. Skinner, H. Zhang, D. Miller, P. Corke, G. Carneiro, A. Angelova, and N. Suenderhauf, “Probabilistic object detection: Definition and evaluation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[13] 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.

[14] J. Choi, D. Chun, H. Kim, and H.-J. Lee, “Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 502–511.

[15] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, “Bounding box regression with uncertainty for accurate object detection,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2883–2892.

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

[17] Y. Lee, J. won Hwang, H. Kim, K. Yun, and J. Park, “Localization uncertainty estimation for anchor-free object detection,” *ArXiv*, vol. abs/2006.15607, 2020.

[18] 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.

[19] Y. He and J. Wang, “Deep mixture density network for probabilistic object detection,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 550–10 555.

[20] 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.

[21] 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.

[22] 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.

[23] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001081

[24] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219301961

[25] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, “Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation,” in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 61–72.

[26] M. Amiri, R. Brooks, B. Behboodi, and H. Rivaz, “Two-stage ultrasound image segmentation using u-net and test time augmentation,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, 04 2020.

[27] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, “Test-time augmentation for deep learning-based cell segmentation on microscopy images,” *Scientific Reports*, vol. 10, 03 2020.

[28] 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. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/2ba596643cbbbc20318224181fa46b28-Paper.pdf

[29] M. S. Ayhan and P. Berens, “Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks,” in *Medical Imaging with Deep Learning*, 2018. [Online]. Available: https://openreview.net/forum?id=rJZz-knjz

[30] 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.

[31] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” 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. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[32] G. Jocher. Yolov5 documentation - test-time augmentation (tta). Accessed 22-June-2022. [Online]. Available: https://docs.ultralytics.com/tutorials/test-time-augmentation/

[33] Y. Chen, L. Tai, K. Sun, and M. Li, “Monopair: Monocular 3d object detection using pairwise spatial relationships,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 090–12 099.

[34] A. Harakeh and S. L. Waslander, “Estimating and evaluating regression predictive uncertainty in deep object detectors,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YLewtnvKgR7

[35] 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.

[36] 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.

[37] A. Harakeh. (2021) A review and comparative study on probabilistic object detection in autonomous driving. Accessed 15-June-2022. [Online]. Available: https://github.com/asharakeh/pod_compare

[38] R. Magalhães. (2022) Uncertainty quantification with test-time augmentation. Accessed 29-August-2022. [Online]. Available: https://github.com/ruimagalhaes24/UQ_TTA_OD

[39] M. Carranza-García, J. Torres-Mateo, P. Lara-Benítez, and J. García-Gutiérrez, "On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data," *Remote Sensing*, vol. 13, no. 1, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/1/89

[40] Real-time object detection on coco. Accessed 23-October-2022. [Online]. Available: https://paperswithcode.com/sota/real-time-object-detection-on-coco

[41] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[42] R. Xu, H. Lin, K. Lu, L. Cao, and Y. Liu, "A forest fire detection system based on ensemble learning," *Forests*, vol. 12, p. 217, 02 2021.

[43] 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.

[44] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[45] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Level 5 perception dataset 2020," https://level-5.global/level5/data/, 2019.