Probabilistic Object Detection: Definition and Evaluation

David Hall[1,2] Feras Dayoub[1,2] John Skinner[1,2] Haoyang Zhang[1,2] Dimity Miller[1,2]
Peter Corke[1,2] Gustavo Carneiro[1,3] Anelia Angelova[4] Niko Sünderhauf[1,2]
[1]Australian Centre for Robotic Vision
[2]Queensland University of Technology [3]University of Adelaide [4]Google Brain
[2]{d20.hall, feras.dayoub, j6.skinner, haoyang.zhang.acrv,
d24.miller, peter.corke, niko.suenderhauf}@qut.edu.au
[3]gustavo.carneiro@adelaide.edu.au [4]anelia@google.com

Abstract

We introduce Probabilistic Object Detection, the task of detecting objects in images and accurately quantifying the spatial and semantic uncertainties of the detections. Given the lack of methods capable of assessing such probabilistic object detections, we present the new Probability-based Detection Quality measure (PDQ). Unlike AP-based measures, PDQ has no arbitrary thresholds and rewards spatial and label quality, and foreground/background separation quality while explicitly penalising false positive and false negative detections. We contrast PDQ with existing mAP and moLRP measures by evaluating state-of-the-art detectors and a Bayesian object detector based on Monte Carlo Dropout. Our experiments indicate that conventional object detectors tend to be spatially overconfident and thus perform poorly on the task of probabilistic object detection. Our paper aims to encourage the development of new object detection approaches that provide detections with accurately estimated spatial and label uncertainties and are of critical importance for deployment on robots and embodied AI systems in the real world.

1. Introduction

Visual object detection provides answers to two questions: *what* is in an image and *where* is it? State-of-the-art approaches that address this problem are based on deep convolutional neural networks (CNNs) that localise objects by predicting a bounding box, and providing a class label with a confidence score, or a full label distribution, for every detected object in the image [27, 37, 38]. The ability of deep CNNs to quantify epistemic and aleatoric uncertainty [19] has recently been identified as paramount for deployment in safety critical applications, where the perception and decision making of an agent has to be trusted [1, 19, 43, 49].

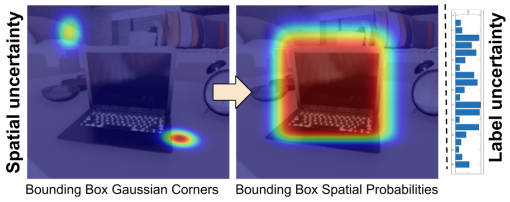


Figure 1: In contrast to conventional object detection, *probabilistic* object detections express semantic *and* spatial uncertainty. Our probabilistic object detections represent object locations as probabilistic bounding boxes where corners are modelled as 2D Gaussians (left) used to express a spatial uncertainty over the pixels (centre). Semantic uncertainty is represented as full label probability distributions (right).

While state-of-the-art object detectors have limited capability to express epistemic and aleatoric uncertainty about the class label through the confidence score or label distribution [14, 15, 33, 44, 48], uncertainty about the spatial aspects of the detection is currently not at all quantified. Furthermore, none of the existing benchmarks using average precision (AP) as the basis for their evaluation [7, 8, 20, 23, 26, 40] can evaluate how well detectors quantify spatial and semantic uncertainties.

We introduce **Probabilistic Object Detection**, the task of detecting objects in images while accurately quantifying the spatial and semantic uncertainties of the detections. Probabilistic Object Detection poses a key challenges that goes beyond the established conventional object detection: the detector must quantify its *spatial uncertainty* by reporting *probabilistic* bounding boxes, where the box corners are modelled as normally distributed. As illustrated in Figure 1, this induces a spatial probability distribution over the image for each detection. The detector must also reliably quantify its *semantic uncertainty* by providing a full probability distribution over the known classes for each detection.

To evaluate how well detectors perform on this chal-

lenging task, we introduce a new evaluation measure, **Probability-based Detection Quality** (PDQ). In contrast to AP-based measures, PDQ explicitly evaluates the reported probability of the true class via its *Label Quality* component. Furthermore, PDQ contains a *Spatial Quality* term that evaluates how well a detection's spatial probability distribution matches the true object.

Unlike existing measures such as mAP [26] and moLRP [34], PDQ jointly evaluates spatial and label uncertainty quality, foreground and background separation quality, and the number of true positive (correct), false positive (spurious), and false negative (missed) detections. Importantly, PDQ does not rely on fixed thresholds or tuneable parameters, and provides optimal assignments of detections to ground-truth objects. Although PDQ has been primarily developed to evaluate *new* types of probabilistic object detectors that are designed to quantify spatial and semantic uncertainties, PDQ can also evaluate conventional state-of-the-art, non-probabilistic detectors.

As we show in Section 7, current conventional detection methods perform poorly on the task of probabilistic object detection due to spatial over-confidence and are outperformed by a recently proposed probabilistic object detector that incorporates Monte Carlo Dropout into a VGG16-based Single Shot MultiBox Detector (SSD) [29].

In summary, our contributions include defining the challenging new task of probabilistic object detection, introducing the new evaluation measure PDQ, evaluating current object detectors, and showing for the first time that novel probabilistic object detectors achieve better performance on this new task, that is highly relevant for applications such as robotics or embodied AI.

2. Motivation

Object detection embedded in a robot or autonomous system, such as a self-driving car, is part of a complex, active, goal-driven system. In such a scenario, object detection provides crucial perception information that ultimately determines the performance of the robot in its environment. Mistakes in object detection could lead to catastrophic outcomes that not only risk the success of the robot's mission, but potentially endanger human lives [1, 2, 24, 32, 35, 49].

For safe and trusted operation in robots or autonomous systems, CNNs must express meaningful *uncertainty* information [1, 14, 15, 19, 43, 49]. Object detectors will have to quantifying uncertainty for both the reported labels and bounding boxes, which would enable them to be treated as yet another sensor within the established and trusted framework of Bayesian information fusion [39, 47]. However, while state-of-the-art object detectors report at least an *uncalibrated* indicator of label uncertainty via label distributions or label scores [14, 15, 33, 44, 48], they currently do *not* report spatial uncertainty. As a result, evaluating the quality of the label or spatial uncertainties is not within the scope of typical benchmark measures and competitions [7, 8, 20, 23, 26, 40].

We argue in favour of accurate quantification of spatial and semantic uncertainties for object detectors in computer vision and robotics applications. Our work builds on this idea by creating a measure that will guide research towards developing detection systems that can operate effectively within a robot's sensor fusion framework.

3. Related Work

Conventional Object Detection: Object detection is a fundamental task in computer vision and aims to localise each instance of certain object classes in an image using a bounding box. The typical output from an object detection system is a set of bounding boxes with a class label score [5, 9, 50]. Since the advent of convolutional neural networks (CNNs) [21], object detection has experienced impressive progress in terms of accuracy and speed [4, 12, 13, 25, 27, 37, 38]. Nonetheless, current overconfident object detection systems fail to provide spatial and semantic uncertainties, and as a result, can be a source of risk in various vision and robotics applications. The probabilistic object detection task introduced by this paper requires that object detectors estimate the spatial and semantic uncertainty of their detections.

Uncertainty Estimation: To improve system robustness and accuracy or avoid risks, quantifying uncertainty has become popular in many vision tasks. Kendall et al. [18] propose a Bayesian model that outputs a pixel-wise semantic segmentation with a measure of model uncertainty for each class. In [19], the authors propose to model the aleatoric and epistemic uncertainties for the pixel-wise semantic segmentation and depth regression tasks, and argue that epistemic uncertainty is important for safety-critical applications and training with small data sets. Kampffmeyer et al. [17] propose a model that estimates pixel-wise classification uncertainty in urban remote sensing images – they argue that the estimated uncertainty can indicate the correctness of pixel labelling. Miller et al. [29, 30] estimate both spatial and classification uncertainties for object detection and use the uncertainty to accept or reject detections under open-set conditions. Nair et al. [31] provide four different voxel-based uncertainty measures for their 3D lesion segmentation system to enable a more complete revision by clinicians. In [45] an uncertainty map for super-resolution of diffusion MR brain images is generated to enable a risk assessment for the clinical use of the super-resolved images. In [42], the authors build an ensemble of predictors to estimate the uncertainty of the centre of nuclei in order to produce more accurate classification results. All the methods above, except the last one [42], estimate uncertainty based on the Monte Carlo (MC) dropout technique [10, 11]. The

papers above provide evidence that it is important to estimate uncertainty for various vision tasks. Most of the proposed methods, except [29, 42], deal with pixel-wise classification.We argue that it is essential to capture the uncertainty of object detectors as motivated in Section 2.

Performance Measures: For the past decade, detection algorithms have predominantly been evaluated using average precision (AP) or variants thereof. Average precision was introduced for the PASCAL VOC challenge [8] in 2007 to replace measuring the area under the ROC curve. It is the average of the maximum precision values at different recall values. These use a pre-defined threshold for the intersection over union (IoU), typically 0.5, defining a true positive detection. This is calculated and averaged across all classes. Since then, AP has become the standard evaluation measure in the PASCAL VOC challenge and is the basis for many other works examining object detection [4, 25, 26, 27, 38, 40]. Most recently, a variation of AP was created which averages AP over multiple IoU thresholds (varying from 0.5 to 0.95 in intervals of 0.05) [26]. This averaging over IoUs rewards detectors with better localisation accuracy. In this work we refer to this measure as mean average precision (mAP) to distinguish it from AP despite mAP typically referring to averaging AP over all classes.

AP-based measures have biased the community to develop object detectors with high recall rate and localisation precision, but these measures have several weaknesses [3, 16, 36]. They rely on fixed IoU thresholds which can lead to overfitting for certain IoU thresholds – the negative consequence is that a small change in the thresholds can cause abrupt score changes. Additionally, these measures use the label score as the detection ranking evidence, without considering the spatial quality, which can lead to sub-optimal detection assignment. In our work, we propose the new evaluation measure PDQ to evaluate both label and spatial qualities of object detections, without using any fixed thresholds and relying on an optimal assignment of detection to ground-truth objects based on both spatial and label qualities.

Oksuze et al. [34] propose the Localisation Recall Precision (LRP) metric to overcome two main deficiencies of mAP: the inability to distinguish different precision-recall (PR) curves, and the lack of a direct way to measure bounding box localisation accuracy. When used for analysing multi-class detectors, the mean optimal LRP (moLRP) is used. Comparing to mAP, moLRP is also based on PR curves but measures localisation quality, false positive rate and false negative rate at some optimal label threshold for each class. The localisation quality is represented by the IoU between the detection and the ground-truth object, scaled by the IoU threshold being used to plot the PR curves. In contrast, our PDQ measure estimates the spatial uncertainty through probabilistic bounding boxes and evaluates how well the detection bounding box's spatial probability distribution coincides with the true object.

4. Probabilistic Object Detection

Probabilistic Object Detection is the task of detecting objects in an image, while accurately quantifying the spatial and semantic uncertainties of the detections. Probabilistic Object Detection thus extends conventional object detection, and makes the quantification of uncertainty an essential part of the task and its evaluation.

Probabilistic Object Detection requires a detector to provide for each known object in an image:

- a categorical distribution over all class labels, and
- a bounding box represented as $\mathcal{B} = (\mathcal{N}_0, \mathcal{N}_1) = (\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1))$ such that $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariances for the multivariate Gaussians describing the top-left and bottom-right corner of the box.

From this probabilistic box representation $\mathcal{B}$, we can calculate a probability distribution P over all pixels (u', v'), such that $P(u', v')$ is the probability that the pixel is contained in the box:

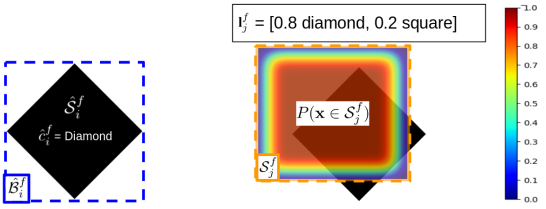
$$P(u', v') = \iint_{0,0}^{v', u'} \mathcal{N}_0(u, v)\, du\, dv \iint_{v', u'}^{H, W} \mathcal{N}_1(u, v)\, du\, dv,$$

where H, W is the height and width of the image. This is illustrated in Fig. 1, with Gaussians over two corners illustrated on the left, and the resulting distribution $P(u', v')$ in the centre.

The evaluation of each detection focuses on the probability value assigned to the true class label, and the spatial probability mass from $P(u', v')$ assigned to the ground truth object *vs*. the probability mass assigned to the background. Since existing measures for conventional object detection such as mAP [26] or moLRP [34] are not equipped to evaluate the probabilistic aspects of a detection, we introduce a novel evaluation measure for Probabilistic Object Detection in the following section.

5. Probability-based Detection Quality (PDQ)

This section introduces the major technical contribution of our paper: the probability-based detection quality (PDQ) measure which evaluates the quality of detections based on spatial and label probabilities. Unlike AP-based measures, our approach penalises low spatial uncertainty when detecting background as foreground, or when detecting foreground as background, and explicitly evaluates the label probability in calculating detection quality. PDQ has no thresholds or tuneable parameters that can redefine the conditions of success. Furthermore, PDQ is based on an approach that provides optimal assignment of detections to



(a) ground-truth object (b) detection

Figure 2: In our notation, a ground-truth object (a) consists of a segmentation mask $\hat{\mathcal{S}}_i^f$ (black), a bounding box $\hat{\mathcal{B}}_i^f$ (blue box), and a class label $\hat{c}_i^f$ which here is *diamond*. A detection (b) consists of a probability density function $P(\mathbf{x} \in \mathcal{S}_j^f)$ (illustrated as a heatmap), a segmentation mask $\mathcal{S}_j^f$ (all pixels within the orange box), and a probability distribution across all classes $\mathbf{l}_j^f$, which here provides probabilities for diamond and square classes.

Data: a dataset of $f = 1 \ldots N_F$ frames with detections $\mathcal{D}_j^f$ and ground-truths $\mathcal{G}_i^f$
forall frames in the dataset **do**
  forall pairs $(\mathcal{G}_i^f, \mathcal{D}_j^f)$ **do**
    calculate spatial quality $Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f)$
    calculate label quality $Q_L(\mathcal{G}_i^f, \mathcal{D}_j^f)$
    calculate $\text{pPDQ}(\mathcal{G}_i^f, \mathcal{D}_j^f) = \sqrt{Q_S \cdot Q_L}$
  end
  Based on the $\text{pPDQ}(.)$ computed between all pairs, find optimal assignment between detections and ground-truth objects, yielding optimal pPDQ for frame f.
end
Combine frame-wise optimal pPDQs into an overall PDQ measure.

Algorithm 1: PDQ Evaluation Process

ground-truth objects, incorporating both the label and spatial attributes of the detections in this assignment.

A reference implementation of PDQ will be made available on github (link withheld for double-blind review).

Notation We write the i-th ground-truth object in the f-th frame (image) as the set $\mathcal{G}_i^f = \{\hat{\mathcal{S}}_i^f, \hat{\mathcal{B}}_i^f, \hat{c}_i^f\}$, comprising a segmentation mask defined by a set of pixels $\hat{\mathcal{S}}_i^f$, a set of bounding box corners $\hat{\mathcal{B}}_i^f$ fully encapsulating all pixels in $\hat{\mathcal{S}}_i^f$, and a class label $\hat{c}_i^f$.

We define the j-th detection in the f-th frame as the set $\mathcal{D}_j^f = \{P(\mathbf{x} \in \mathcal{S}_j^f), \mathcal{S}_j^f, \mathbf{l}_j^f\}$, comprising a probability function that returns the spatial probability that a given pixel is a part of the detection (regardless of class prediction) $P(\mathbf{x} \in \mathcal{S}_j^f)$, a set of pixels with a non-zero $P(\mathbf{x} \in \mathcal{S}_j^f)$ which we refer to as the detection segmentation mask $\mathcal{S}_j^f$, and a label probability distribution across all possible class labels $\mathbf{l}_j^f$. A visualisation of both ground-truth objects and detections is provided in Figure 2.

Requirements PDQ requires pixel-accurate ground-truth annotations for the segmentation mask $\hat{\mathcal{S}}_i^f$. Such annotations can be easily obtained from simulated environments [6, 41] and also from datasets containing only bounding box annotations by considering all pixels within a box part of the segmentation mask. PDQ can evaluate probabilistic detectors that provide bounding boxes with Gaussian corners as defined in Section 4, or conventional detectors by assuming $P(\mathbf{x} \in \mathcal{S}_j^f) = 1 - \epsilon$ for all pixels inside the respective bounding box and ϵ outside, for a small $\epsilon > 0$.

Overview PDQ evaluates both the *spatial* and *label* quality of a detector. It is therefore based on a combination of a spatial quality measure Q_S and a label quality measure Q_L. Both are calculated between all possible pairs of detections and ground-truth objects within a single frame. We define the geometric mean between these two quality measures as the pairwise PDQ (pPDQ), and use it to find the optimal assignment between all detections and ground-truth objects within an image. The optimal pPDQ measures are then combined into an overall PDQ measure for the whole dataset. However, many of these intermediate results can also be recorded and analysed for a more detailed breakdown of performance. Algorithm 1 summarises the overall PDQ calculation. In the following, we detail each of the involved steps and both quality measures.

5.1. Spatial Quality

The spatial quality Q_S measures how well a detection $\mathcal{D}_j^f$ captures the spatial extent of a ground-truth object $\mathcal{G}_i^f$, and takes into account the spatial probabilities for individual pixels as expressed by the detector.

Spatial quality Q_S comprises two loss terms, the foreground loss L_{FG} and the background loss L_{BG}. Spatial quality is defined as the exponentiated negative sum of the two loss terms, as follows:

$$Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f) = \exp(-(L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f) + L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f)), \tag{1}$$

where $Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f) \in [0, 1]$. The spatial quality in (1) is equal to 1 if the detector assigns a spatial probability of 1 to all ground-truth pixels, while not assigning any probability mass to pixels outside the ground-truth segment. This behaviour is governed by the two loss terms explained below.

Foreground Loss The foreground loss L_{FG} is defined as the average negative log-probability the detector assigns to the pixels of a ground-truth segment.

$$L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f) = -\frac{1}{|\hat{\mathcal{S}}_i^f|} \sum_{\mathbf{x} \in \hat{\mathcal{S}}_i^f} \log(P(\mathbf{x} \in \mathcal{S}_j^f)), \tag{2}$$

where, as defined above, $\hat{\mathcal{S}}_i^f$ is the set of all pixels belonging to the i-th ground-truth segment in frame f, and $P(\cdot)$ is the spatial probability function that assigns a probability value to every pixel of the j-th detection. The foreground loss is minimised if the detector assigns a probability value of one to every pixel of the ground-truth segment, in which case $L_{FG} = 0$. It grows without bounds otherwise.

Notice that L_{FG} intentionally ignores pixels that are inside the ground-truth bounding box $\hat{\mathcal{B}}_i^f$ but are *not* part of the ground-truth segment $\hat{\mathcal{S}}_i^f$. This avoids treating the detection of background pixels as critically important in the case of irregularly shaped objects when pixel-level annotations are available, unlike AP-based methods using bounding-box IoUs, as illustrated in Figure 3.

Background Loss The background loss term L_{BG} penalises any probability mass that the detector incorrectly assigned to pixels outside the ground-truth bounding box. It is formally defined as

$$L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f) = -\frac{1}{|\hat{\mathcal{S}}_i^f|} \sum_{\mathbf{x} \in \mathcal{V}_{ij}^f} \log((1 - P(\mathbf{x} \in \mathcal{S}_j^f))), \quad (3)$$

which is the sum of negative log-probabilities of all pixels in the set $\mathcal{V}_{ij}^f = \{\mathcal{S}_j^f - \hat{\mathcal{B}}_i^f\}$, i.e. pixels that are part of the detection, but not of the true bounding box. A visualisation of this evaluation region $\mathcal{V}_{i,j}^f$ is shown in Figure 4. Note that we average over $|\hat{\mathcal{S}}_i^f|$ rather than $|\mathcal{V}_{i,j}^f|$ to ensure that foreground and background losses are scaled equivalently, measuring the loss incurred per ground-truth pixel the detection aims to describe. The background loss term is minimised if all pixels outside the ground-truth bounding box are assigned a spatial probability of zero.

5.2. Label Quality

While spatial quality measures how well the detection describes *where* the object is within the image, label quality Q_L measures how effectively a detection identifies *what* the object is. We define Q_L as the probability estimated by the detector for the object's ground-truth class. Note that this is irrespective of whether this class is the highest ranked in the detector's probability distribution. Unlike with mAP, this value is explicitly used to influence detection quality rather than just for ranking detections regardless of actual label probability. We define label quality as:

$$Q_L(\mathcal{G}_i^f, \mathcal{D}_j^f) = \mathbf{l}_j^f(\hat{c}_i^f). \quad (4)$$

5.3. Pairwise PDQ (pPDQ)

The pairwise PDQ (pPDQ) between a detection $\mathcal{D}_j^f$ and a ground-truth object $\mathcal{G}_i^f$ in frame f is the geometric mean of the spatial quality and label quality measures Q_S and Q_L:

$$\text{pPDQ}(\mathcal{G}_i^f, \mathcal{D}_j^f) = \sqrt{Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f) \cdot Q_L(\mathcal{G}_i^f, \mathcal{D}_j^f)}. \quad (5)$$

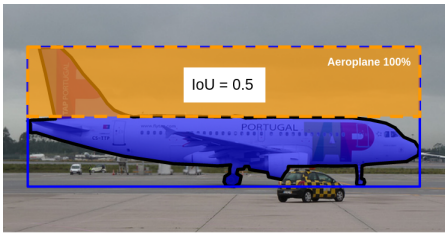


Figure 3: Example of a detection of an aeroplane (orange box), a ground-truth box (blue line), and a ground-truth segmentation mask, (blue-coloured region with black border). At an IoU threshold of 0.5, AP-based methods consider the orange detection entirely correct, despite covering only 16% of the plane's pixels. There is no correlation between the bounding box overlap analysed and the content within the bounding box. By comparison, PDQ penalises this detection heavily for only detecting this small portion without any spatial uncertainty. The pPDQ for this detection containing no spatial uncertainty is 3.64×10^{-6}.

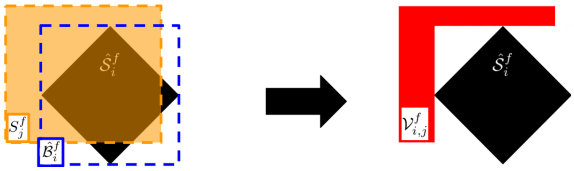


Figure 4: PDQ defines the background evaluation region $\mathcal{V}_{i,j}^f$ (red) as the set of pixels that are part of the detection $\mathcal{S}_j^f$ (orange), but not of the true bounding box $\hat{\mathcal{B}}_j^f$ (blue).

Using the geometric mean requires both components to have high values for a high pPDQ score, and is zero if either component reaches zero. Notice that it is also possible to use a weighted geometric mean for applications where the spatial or label quality component is more important.

5.4. Assignment of Optimal Detection-Object Pairs

It is important that, for every frame, each detection is matched to, at most, one ground-truth object and vice versa. This is also done for mAP, but it utilises a greedy assignment process based upon label confidence ranking, rather than ensuring that the optimal assignment takes into account both the spatial and label aspects of the detection. To mitigate this problem, we use our proposed pPDQ score in (5) between possible detection-object pairings to determine the optimal assignment through the Hungarian algorithm [22]. This provides the optimal assignment between two sets of information which produce the best total pPDQ score.

Using assignments from the Hungarian algorithm, we store the pPDQs for all non-zero assignments in the f-th frame in a vector $\mathbf{q}^f = [q_1^f, q_2^f, q_3^f, ..., q_{N_{TP}^f}^f]$ where N_{TP}^f is the number of non-zero (true positive) assignments within

the f-th frame. Note that these "true positive" detections are not ones which are considered 100% accurate as is done for AP-based measures. Instead these are detections which, even marginally, describe the ground-truth object they are matched with and provide a non-zero pPDQ. If the pPDQ from an optimal assignment is zero, there is no association between the ground-truth object and detection. This occurs when either a ground-truth object is undetected (false negative) or a detection does not describe an object (false positive). We also record the number of false negatives and false positives for each frame, expressed formally as N_{FN}^f and N_{FP}^f respectively, to be used in our final evaluation. After obtaining $\mathbf{q}^f$, N_{TP}^f, N_{FN}^f, and N_{FP}^f for each frame, the PDQ score can be calculated.

5.5. PDQ Score

The final PDQ score across a set of ground-truth objects $\mathcal{G}$ and detections $\mathcal{D}$ is the total pPDQ for each frame divided by the total number of TPs, FNs and FPs assignments across all frames. This can be seen as the average pPDQ across all TPs, FNs and FPs observed, which is calculated as follows:

$$PDQ(\mathcal{G}, \mathcal{D}) = \frac{1}{\sum_{f=1}^{N_F} N_{TP}^f + N_{FN}^f + N_{FP}^f} \sum_{f=1}^{N_F} \sum_{i=1}^{N_{TP}^f} \mathbf{q}^f(i), \tag{6}$$

where $\mathbf{q}^f(i)$ is the pPDQ score for the i-th assigned detection-object pair in the f-th frame. This final PDQ score provides a consistent, probability-based measure, evaluating both label and spatial probabilities, that can determine how well a set of detections has described a set of ground-truth objects without the need for thresholds to determine complete success or failure of any given detection.

6. Evaluation of PDQ Traits

The previous section introduced PDQ, a new measure to evaluate the performance of detectors for *probabilistic* object detection. PDQ has been designed with one main goal in mind: it should reward detectors that can accurately quantify both their spatial and label uncertainty. In this section, we are going to demonstrate that this goal has been met, by showing PDQ's behaviour in controlled experiments. We show the most critical experiments here and more are provided in supplementary material.

PDQ Rewards Accurate Spatial Uncertainty We perform a controlled experiment on the COCO 2017 validation dataset with a simulated object detector. For every ground truth object with true bounding box corners $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{x}}_1$, the detector generates a detection with bounding box corners sampled as $\mathbf{x}_0 \sim \mathcal{N}(\hat{\mathbf{x}}_0, \hat{\mathbf{\Sigma}})$ and $\mathbf{x}_1 \sim \mathcal{N}(\hat{\mathbf{x}}_1, \hat{\mathbf{\Sigma}})$, with $\hat{\mathbf{\Sigma}} = \text{diag}(\hat{s}^2, \hat{s}^2)$. We vary the value of $\hat{s}^2$ throughout the experiments and refer to $\hat{s}^2$ as the detector's *true* variance.

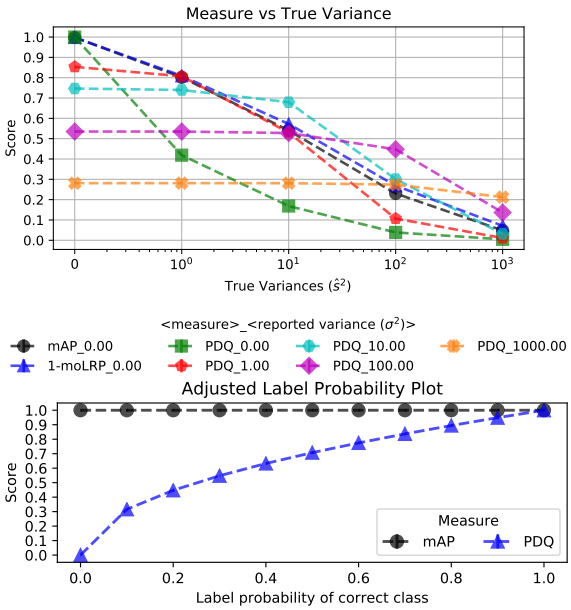


Figure 5: Top: PDQ rewards detectors that accurately evaluate their true spatial uncertainty. Bottom: PDQ explicitly evaluates label uncertainty, in contrast to mAP. See Section 6 for explanation of the experiments.

Independent of the value of $\hat{s}^2$, the simulated detector expresses spatial uncertainty for each probabilistic detection with a different variance σ^2, which we refer to as the *reported* variance. Each detection is assigned probability 1.0 for the *true* label, corresponding to perfect classification.

When varying the values of $\hat{s}^2$ and σ^2 and evaluating the resulting detections, PDQ should reward when $\hat{s}^2$ is similar to σ^2, i.e. when the *reported* spatial uncertainty is close to the *true* spatial uncertainty that was used to sample the detection corners. When both *reported* and *true* spatial uncertainty are equal, PDQ should reach its peak performance. This would indicate that PDQ does indeed reward the accurate estimation of spatial uncertainty.

Figure 5 confirms this conjecture. We repeated the experiment described above 20 times, evaluating on all objects in the 5,000 images of the COCO 2017 validation set. Every line corresponds to a detector with a different *reported* variance σ^2. The *true* variance $\hat{s}^2$ varies along the x axis. We can see that the for each value of $\hat{s}^2$, simulated detectors with $\sigma^2 = \hat{s}^2$ give the best performance.

PDQ Explicitly Evaluates Label Uncertainty We perform a controlled experiment in a simulated scenario where a single object is detected by a single detection with perfect spatial quality. We vary the detection's reported label probability for the correct class and ensure that it always remains the dominant class in the probability distribution. The resulting PDQ and mAP scores are compared in Figure 5. We observe that PDQ is affected by the label probability of the

correct class via its label quality term. This is in contrast to mAP which uses label probability only to determine the dominant class and for ranking detection matches.

7. Evaluation of Object Detectors

In this section we evaluate a number of state-of-the-art conventional detectors and the recently proposed probabilistic object detector MC-Dropout SSD [30] that is based on Monte Carlo Dropout. We compare the ranking of all tested detectors using PDQ and its components, as well as the established measures mAP and moLRP [34], and discuss our most important observations and gained insights.

7.1. Experimental Set-up

Evaluated Detectors The state-of-the-art conventional object detectors evaluated were SSD-300 [27], YOLOv3 [37], FasterRCNN with ResNet backbone (FRCNN R) [51], FasterRCNN with ResNet backbone and feature pyramid network (FPN) (FRCNN R+FPN) [28], and FasterRCNN with ResNeXt backbone and FPN (FRCNN X+FPN) [28]. To evaluate these conventional detectors with PDQ, we set $P(\mathbf{x} \in \mathcal{S}_j^f) = 1 - \epsilon$ for all pixels $\mathbf{x}$ inside the provided standard bounding box, and ϵ for all pixels outside, when performing the calculations in equations (2) and (3), with $\epsilon = 10^{-14}$ to avoid infinite loss.

In addition to conventional object detectors, we evaluate a probabilistic MC-Dropout object detector based on the work by Miller et al. [29, 30]. We follow the established implementation [29], where Monte Carlo Dropout [10] is utilised in a SSD-300 object detector [27] with two dropout layers inserted and activated during both training and testing. Each image is tested with 20 forward passes through the network with randomised dropout masks to obtain samples of detections. The recommended merging strategy was used to cluster these samples [29], namely a BSAS clustering method [46] with spatial affinity IoU and label affinity 'same label' (we found an IoU threshold of 0.7 performed better than the 0.95 threshold recommended in [29]). Final probabilistic detections were obtained by averaging sample label probability distributions and estimating $\mathcal{N}_0$ and $\mathcal{N}_1$ from the average and covariance of sample bounding boxes.

We furthermore modify a FasterRCNN with ResNeXt backbone and feature pyramid network to approximate probabilistic detections. We achieve this by the following process: for every detection surviving the normal non-maximum suppression, we find all of the suppressed detections that have an IoU of above 0.75 with the surviving detections and cluster them (including the survivors). We then calculate the Gaussian corner mean and covariances of each cluster, weighted by the detection's winning label confidences. We denote this method as probFRCNN in Table 1.

Evaluation Protocol and Datasets Evaluation was performed on the 5,000 images of the MS COCO 2017 validation set [26], after all detectors have been trained or fine-tuned on the 2017 training dataset. During the evaluation, we ignored all detections with the winning class label probability below a threshold τ. We compare the effect of this process for $\tau = 0.5$ and 0.05.

7.2. Insights

Table 1 presents the results of our evaluation, comparing PDQ and its components with mAP and moLRP. From these results we observe the following:

1. PDQ exposes the performance differences between probabilistic and non-probabilistic object detectors. When evaluating using mAP or moLRP, both SSD-300 [27] and FasterRCNN with ResNeXt and FPN [28], and their respective probabilistic variants (MC-DropoutSSD [29] and probFRCNN) show very similar performance. However, evaluating with PDQ reveals their performance differences in terms of probabilistic object detection: both probabilistic variants perform significantly better than their non-probabilistic counterparts. This is especially true for their overall spatial quality and its foreground and background quality components. Comparing probFRCNN with MC-DropoutSSD, we found that probFRCNN achieved a higher PDQ score, benefiting from its more accurate base network.

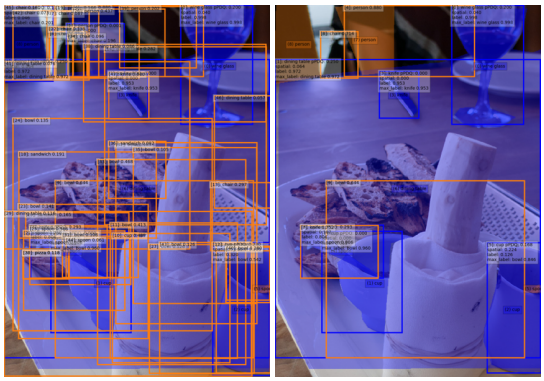
2. PDQ reveals differences in spatial and label quality. Since PDQ comprises meaningful components, it allows a detailed analysis of how well detectors perform in terms of spatial and label quality. For example, in Table 1 we observe that the YOLOv3 detector achieves the highest label quality (95.8%/92.8% for $\tau = 0.5/0.05$), but the worst spatial quality (6.2%/5.1%) of all tested detectors. This gives important insights into worthwhile directions of future research, suggesting YOLO can be more trusted to understand *what* an object is than other detectors but is less reliable in determining precisely *where* that object is.

3. Probabilistic localisation performance of existing object detectors is weak. Spatial quality in PDQ measures how well detectors probabilistically localise objects in an image. Conventional object detectors assume full confidence in their bounding box location and achieve low spatial qualities between 5.1% and 17.6%, indicating they are spatially overconfident. Since conventional object detectors have comparatively high label qualities, we conclude that for probabilistic object detection tasks where spatial uncertainty estimation is important, improving the localisation performance and the estimation of spatial uncertainty has the biggest potential of improving performance.

4. PDQ does not obscure false positive errors. Unlike mAP and moLRP, PDQ explicitly penalises a detector for spurious (false positive) detections, as well as for missed

Table 1: PDQ-based Evaluation of Probabilistic and Non-Probabilistic Object Detectors. Legend: mLRP = 1 − moLRP, Sp = Spatial Quality, Lbl = Label Quality, FG = Foreground Quality ($\exp(-L_{FG})$), BG = Background Quality ($\exp(-L_{BG})$, TP = True Positives, FP = False Positives, FN = False Negatives. pPDQ, Sp, Lbl, FG and BG averaged over all TP.

Approach (τ)	mAP (%)	mLRP (%)	PDQ (%)	pPDQ (%)	Sp (%)	Lbl (%)	FG (%)	BG (%)	TP	FP	FN
probFRCNN (0.5)	35.5	32.2	**28.4**	**56.7**	**45.0**	90.7	**77.8**	**60.7**	23,434	10,016	13,347
MC-Dropout SSD (0.5) [29]	15.8	15.6	12.8	47.3	39.9	74.0	73.1	57.3	10,510	**2,165**	26,271
MC-Dropout SSD (0.05) [29]	19.5	16.6	1.3	26.1	27.3	35.9	60.1	46.2	24,843	461,074	11,938
SSD-300 (0.5) [27]	15.0	14.3	3.9	18.1	9.7	80.2	57.5	25.1	8,999	4,746	27,782
SSD-300 (0.05) [27]	19.3	16.0	0.6	9.7	6.4	40.2	38.1	32.3	21,961	324,067	14,820
YOLOv3 (0.5) [37]	29.7	30.8	5.7	14.6	6.2	**95.8**	52.2	20.4	17,390	7,728	19,391
YOLOv3 (0.05) [37]	30.1	27.7	3.3	12.2	5.1	92.8	44.6	22.9	23,447	50,074	13,334
FRCNN R (0.5) [51]	32.8	29.1	6.7	19.1	10.3	88.8	62.2	23.6	19,930	20,044	16,851
FRCNN R (0.05) [51]	34.3	29.1	3.0	17.1	9.5	78.5	57.8	25.1	23,081	93,141	13,700
FRCNN R+FPN (0.5) [28]	34.6	31.2	11.8	27.1	16.9	86.5	60.6	35.7	22,537	14,706	14,244
FRCNN R+FPN (0.05) [28]	37.0	30.4	4.2	23.1	15.8	69.5	54.4	38.7	29,326	123,511	7,455
FRCNN X+FPN (0.5) [28]	37.4	**32.7**	11.9	27.9	17.6	88.2	60.8	36.8	24,523	20,444	12,258
FRCNN X+FPN (0.05) [28]	**39.0**	32.1	4.4	24.8	16.7	74.4	55.6	39.1	**29,922**	130,009	**6,859**



(a) FRCNN X+FPN (0.05) (b) FRCNN X+FPN (0.5)

Figure 6: Visualisation of all TPs (blue segmentation mask and corresponding BBox), FPs (orange BBox), and FNs (orange segmentation mask) as defined by PDQ for FRCNN X+FPN with $\tau = 0.5$ (a) and $\tau = 0.05$ (b). We see here that a lower τ leads to far more FPs that are strongly penalised by PDQ but are largely ignored under mAP.

(false negative) detections. We observe that decreasing the label threshold τ and consequently massively increasing the number of false positive detections (see Fig. 6 for an example) actually increases mAP, and does not tend to affect moLRP much. In contrast, PDQ scores decrease significantly. PDQ is designed to evaluate systems for application in real-world systems and does not filter detections based on label ranking or calculating the optimal threshold τ. It involves *all* reported detections in its analysis.

8. Conclusions and Future Work

We introduced Probabilistic Object Detection, a challenging new task that is highly relevant for domains where accurately estimating the spatial and semantic uncertainties of the detections is of high importance such as embodied AI (such as robotics, autonomous systems, driverless cars), and medical imaging. To foster further research in this direction, we introduced the probability-based detection quality (PDQ) measure which explicitly evaluates both spatial and label uncertainty.

PDQ is not meant to *replace* mAP, but to *complement* it. Both evaluation measures are designed for two *different* problems. While mAP has been the established performance measure for conventional object detection, we developed PDQ specifically for the new task of *probabilistic* object detection.

After evaluating a range of object detectors, including the first emerging probabilistic object detector in Section 7, we are confident that PDQ is a useful performance measure that can guide and inform the research of even better probabilistic object detectors in the future. In future work we will investigate how to train object detectors to directly optimise for PDQ by incorporating it into the training loss function. The concept of probabilistic object detection can also be easily extended to Probabilistic *Instance Segmentation* where each pixel would contain a probability of belonging to a certain object instance, along with a label distribution.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv*

preprint arXiv:1606.06565, 2016.

[2] D. CireşAn, U. Meier, J. Masci, and J. Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural networks*, 32:333–338, 2012.

[3] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation?. In *BMVC*, volume 27, page 2013. Citeseer, 2013.

[4] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[6] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[7] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[8] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[10] Y. Gal and Z. Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[11] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[12] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[15] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[16] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012.

[17] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–9, 2016.

[18] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[19] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[20] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, volume 1, page 4, 2012. 00312.

[22] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2:83–97, 1955.

[23] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[24] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[28] F. Massa and R. Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [Insert date here].

[29] D. Miller, F. Dayoub, M. Milford, and N. Sünderhauf. Evaluating Merging Strategies for Sampling-based Uncertainty Techniques in Object Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[30] D. Miller, L. Nicholson, F. Dayoub, M. Milford, and N. Sünderhauf. Dropout Sampling for Robust Object Detection in Open-Set Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[31] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–663. Springer, 2018.

[32] T. Namba and Y. Yamada. Risks of deep reinforcement learn-

ing applied to fall prevention assist by autonomous mobile robots in the hospital. *Big Data and Cognitive Computing*, 2(2):13, 2018.

[33] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[34] K. Oksuz, B. Can Cam, E. Akbas, and S. Kalkan. Localization recall precision (lrp): A new performance metric for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 504–519, 2018.

[35] C. Otte. Safe and interpretable machine learning: Amethodological review. In C. Moewes and A. Nürnberger, editors, *Computational Intelligence in Intelligent Data Analysis*, pages 111–122, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[36] J. Pont-Tuset and F. Marques. Supervised evaluation of image segmentation and object proposal techniques. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1465–1478, 2016.

[37] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, Apr. 2018.

[38] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[39] C. Richter, W. Vega-Brown, and N. Roy. Bayesian learning for safe high-speed navigation in unknown environments. In *Robotics Research*, pages 325–341. Springer, 2018.

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015.

[41] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[42] K. Sirinukunwattana, S. e. A. Raza, Y. Tsang, D. Snead, I. Cree, and N. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35:1–1, 02 2016.

[43] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018.

[44] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *Proceedings of International Conference on Learning Representations*, 2014.

[45] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.

[46] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Second Edition*. 2nd edition, 2003.

[47] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT press, 2005.

[48] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, June 2011.

[49] K. R. Varshney and H. Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5 3:246–255, 2017.

[50] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[51] J. Yang, J. Lu, D. Batra, and D. Parikh. A Faster Pytorch Implementation of Faster R-CNN. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017.

Appendix Overview

In this appendix we provide supplementary material and analysis that was not included in the main paper due to space restraints. This appendix is organized as follows:

A. PDQ Qualitative Examples.

B. Evaluation of PDQ traits.

C. Traditional Measures Obscuring False Positives.

D. Definition of mAP.

A. PDQ Qualitative Examples

We provide qualitative results for detectors tested on COCO data in Section 7 of the main paper. Specifically, in this section we visualise results from SSD-300 [27], YOLOv3 [37], Faster RCNN with ResNext backbone and a feature pyramid network (FRCNN X+FPN) [28], and the probabilistic MC-Dropout SSD detector based on the work by Miller et al. [29, 30]. Unless otherwise stated, results shown are for detectors using a label confidence threshold of 0.5.

Using the detection-object pairing assignment from PDQ as outlined in section 5.4 of the main paper, we are able to provide visualisations outlining the true positives (TPs), false positives (FPs) and false negatives (FNs) present in a given image, as was done in Figure 6 of the main paper. In these visualisations we show TPs as blue segmentation masks and boxes, FPs as orange boxes, and FNs as orange segmentation masks. We also provide a way to visualise spatially probabilistic detections using **ellipses** in the top-left and bottom-right corners, showing the contours of the Gaussian corners at distances of 1, 2 and 3 standard deviations. For conventional detectors, there are no ellipses as they provide no spatial uncertainty. Because we know the optimal assignment, as mentioned in the main paper, we can extract pairwise quality scores between TPs. In our visualisations we provide pPDQ, spatial quality and label quality scores for all TP detections in a text box at the top-left corner of the detection box.

Using visualisations of this form enables us to qualitatively reinforce some of the findings from the main paper in the following three subsections. Firstly, we see again how the number of false positives under PDQ increases with lower label confidence thresholds (despite such detections getting higher mAP scores). Secondly, we get to observe the effect of spatial uncertainty estimation and how this effects spatial quality scores for different detections. Thirdly, we can visually show the high label quality but poorer localisation achieved by YOLOv3 when compared to FRCNN X+FPN.

A.1. Increased False Positives with Lower Label Confidence Threshold

Reinforcing the finding of the main paper, we show more examples for FRCNN X+FPN with label confidence thresholds of 0.5 and 0.05 respectively in Figure 7. Note that because these images are rather cluttered, we omit the detailed quality information beyond the detection's maximum class label. We see that the number of FPs (orange boxes) increases dramatically when the label confidence threshold is lowered to 0.05.

A.2. Spatial Uncertainty Estimation

We show some examples from the MC-Dropout SSD detector to highlight the effect that spatial uncertainty has on both spatial quality and overall pPDQ in Figures 8 and 9.

Figure 8 shows the effect that spatial uncertainty estimation has on the spatial quality of PDQ. In Figure 8a we see the spatial quality vary between three people based upon uncertainty estimation. The left-most person has the poorest spatial quality as the box misses part of his entire arm, goes too far below their feet, and yet has very little spatial uncertainty in it's detection, scoring a spatial quality of only 28.5%. This is in comparison to the right-most person who has a detection with some uncertainty to the top, left, and right of the box, matching where there is the most error in the detection itself. This leads to a much higher spatial quality of 88.4%.

In Figure 8b, we see that simply adding spatial uncertainty is not enough to guarantee a good score and a TP detection. We see the bottom of the detection box for the human is over-confident, leading to a FP detection. Finally, in Figure 8c, we see that the box around the laptop is nearly perfect and yet the right-most edge has high uncertainty. By comparison, we see the person in the picture has a poorer base bounding box but appears to have a more reasonable estimate of it's uncertainties. Comparing spatial quality scores, we see that despite it's better base bounding box, the spatial quality of the laptop is only 65% compared the person's spatial quality of 87.5%. This drop in spatial quality is due to the high spatial uncertainty expressed by the laptop detection.

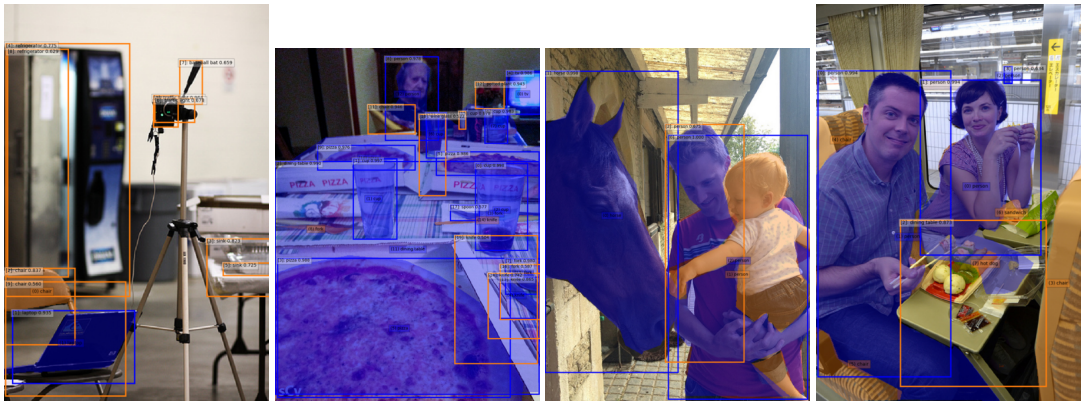
A.3. MC Dropout Vs SSD

In the main paper, we showed that MC-Dropout SSD was able to achieve higher spatial quality, and by extension pPDQ, than conventional detectors. We show this visually in Figure 9, comparing detections from MC-Dropout SSD to those of SSD-300. Neither has tight detections around the person or umbrella, but SSD-300 boxes visually appear tighter. However, SSD-300 detections are over-confident, expressing no spatial uncertainty and attaining spatial quality up to only 3.8% found on the person. In comparison, we see MC-Dropout SSD detections expressing uncertainty that coincides with the innacuraccies of the detection. This provides a spatial quality of up to 62.7% found on the person. Better pPDQ scores are seen for both objects with MC-Dropout.

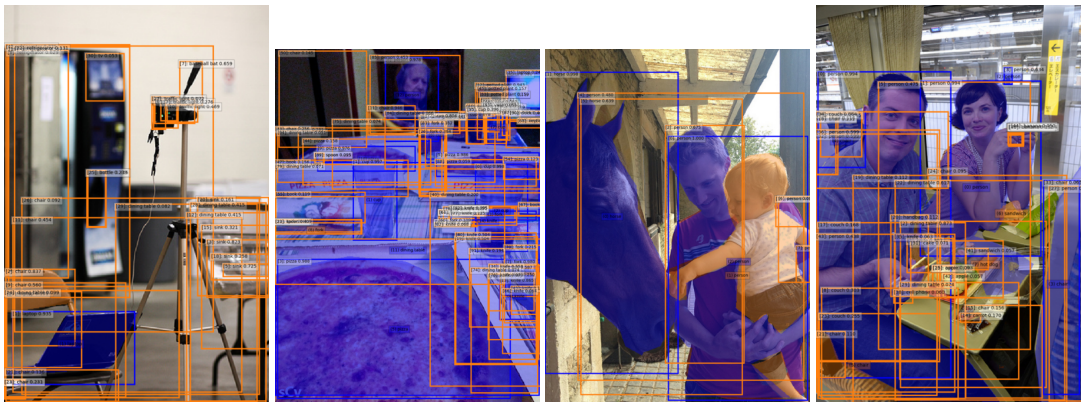
A.4. YOLO Label Vs Spatial Quality

In the experiments from the main paper, we showed that YOLOv3 achieves high label quality but comparatively low spatial quality when compared with other detectors such as FRCNN X+FPN. In Figure 10 we visually compare YOLOv3 and FRCNN X+FPN results to qualitatively confirm this observation.

Examining Figure 10, we see that in the left image YOLOv3 produces higher confidence detections for chair and hotdog than FRCNN X+FPN, but because their detections are over-confident and have poorer localisation, they are treated as FPs rather than TPs. On the right, we see a more confident pizza detection from YOLOv3 but a poorer box localisation leading to spatial quality of 0.1% compared to the 13.9% spatial quality of FRCNN X+FPN (0.5). This supports the observation from the main paper that YOLOv3 can have higher label quality than FRCNN detectors but tends to have a lower spatial quality due to poorer localisation.

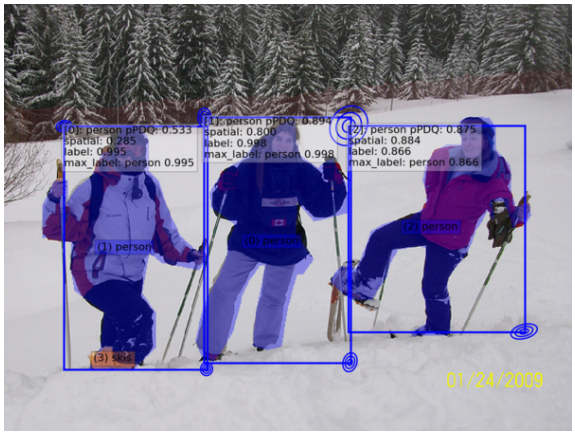


(a) 0.5 Label Threshold

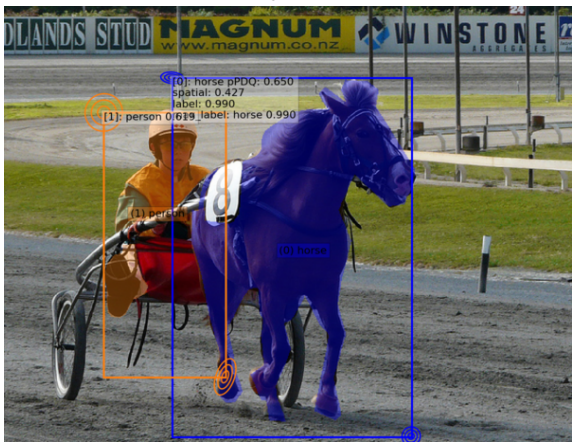


(b) 0.05 Label Threshold

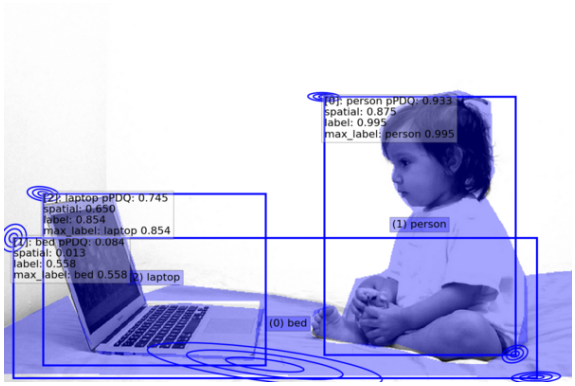
Figure 7: Detections from FRCNN X+FPN at label confidence thresholds of 0.5 (a) and 0.05 (b) as evaluated by PDQ. We see more false positives (orange boxes) under PDQ with 0.05 despite 0.05 giving higher mAP scores as shown in the main paper.



(a) general

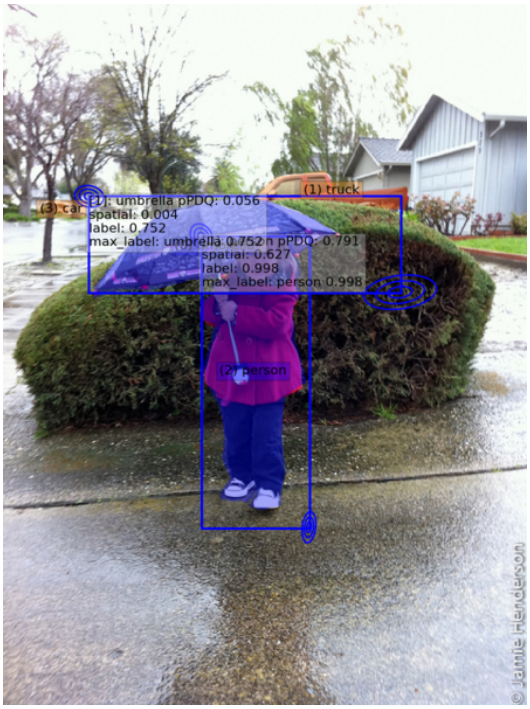


(b) over-confident



(c) under-confident

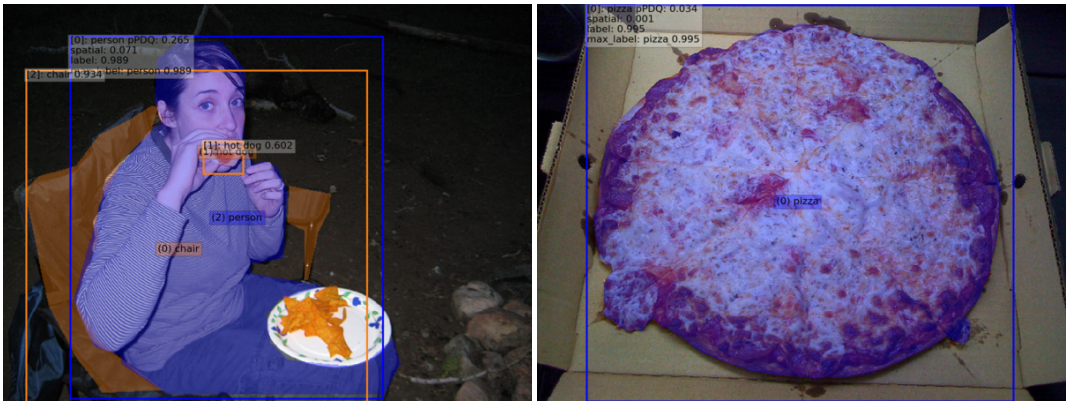
Figure 8: Visualisation of MC-Dropout SSD detections as analysed by PDQ. Ellipses represent spatial uncertainty. In (a) we see a general case where individuals have better or worse spatial quality dependant on uncertainty estimation. In (b) we see a detection with uncertainty which is still over-confident and misses the person. In (c) we see an under-confident detection around the laptop.



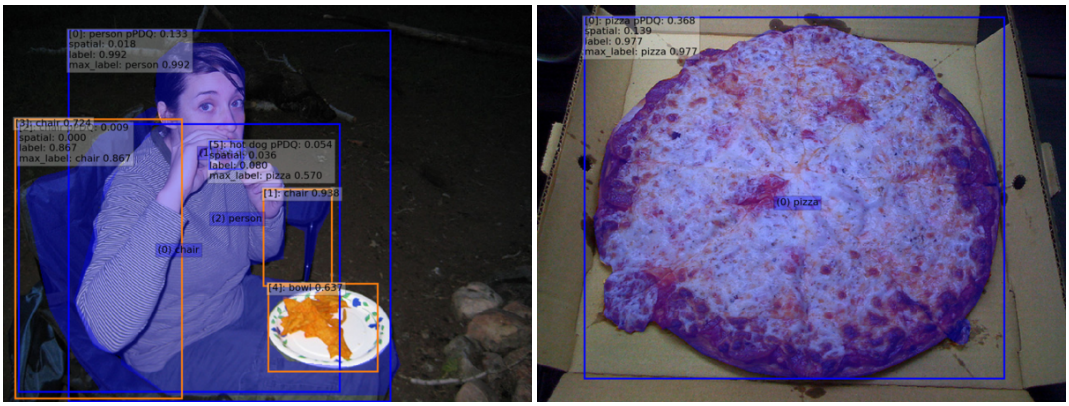
(a) MC-Dropout SSD

(b) SSD-300

Figure 9: Comparison of MC-Dropout SSD to SSD-300. SSD-300 is shown to be spatially over-confident leading to low scores despite tighter boxes.



(a) YOLOv3



(b) FRCNN X+FPN

Figure 10: Visualisation of YOLOv3 detections compared with FRCNN X+FPN.

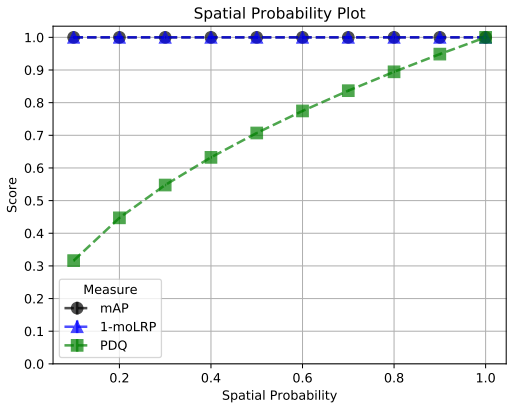


Figure 11: Evaluation of the effect of spatial probability on a perfectly aligned BBox. We see that unlike existing object detection measures, PDQ is effected by spatial probability changes.

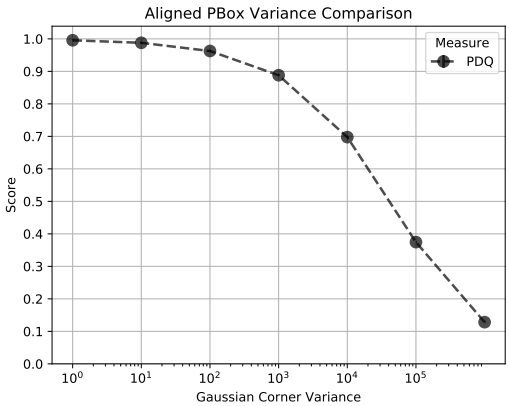


Figure 12: Plot showing the effect on PDQ of increasing variance, and by extension uncertainty, on perfectly aligned PBoxes. We see that for perfectly aligned detections, the score goes down the more uncertain the PBox detection is.

B. Evaluation of PDQ Traits

We demonstrate the characteristics of PDQ when compared with existing measures (mAP [26] and moLRP [34]) when responding to different types of imperfect detections, expanding upon what was covered in the main paper. Specifically, we examine the effect of spatial uncertainty, detection misalignment, label quality, missing ground-truth objects, and duplicate/false detections. Throughout, we refer to standard detections with no spatial uncertainty as bounding box (BBox) detections and probabilistic detections with spatial uncertainty as probabilistc bounding box (PBox) detections.

B.1. Spatial Uncertainty

We examine the effect of spatial uncertainty on BBox and PBox detections respectively.

BBox Spatial Uncertainty We evaluate a perfectly aligned BBox detection which has varying values of spatial probability for every pixel therein. Whilst not a realistic type of detection, it allows for easy examination of the response from existing measures and PDQ to spatial probability variations. The results are shown in Figure 11

This experiment shows that PDQ is gradually reduced by decreasing spatial certainty, whereas mAP and moLRP consistently consider the provided output to be perfect as they are not designed to measure uncertainty.

PBox Spatial Uncertainty To examine the effect of increasing spatial uncertainty on PDQ using PBoxes, we perform a test using a perfectly aligned PBox detection on a single object. We consider a simple square-shaped 500 x 500 object centred in a 2000 x 2000 image. PBox corner Gaussians are spherical and located at the corners of the object they are detecting. PBox reported variance for the corner Gaussians is varied to observe the effect of increased uncertainty. The results of this test are shown in Figure 12. We see a decline in PDQ with increased uncertainty demonstrating how PDQ penalises under-confidence.

B.2. Detection Misalignment

We perform two experiments to analyse responses to misaligned detections. These are translation error and scaling error.

Translation Error We observe the effect of translation errors by shifting a 500 x 500 detection left and right past a 500 x 500 square object centred within a 2000 x 2000 image. This is tested both using BBoxes, and PBoxes with spherical Gaussian corners of varying reported variance (BBoxes equivalent to reported variance of zero). The results from this test are shown in Figure 13.

Here, we see that PDQ strongly punishes any deviation from the ground-truth for BBoxes with no spatial uncertainty. In some cases PDQ drops close to zero after only a 10% shift. This is in strong comparison to mAP and moLRP which, while decreasing, does so at a far slower rate despite high confidence being supplied to incorrectly labelled pixels. As a shift of 10% is quite large for a 500 x 500 square, PDQ does not provide such leniency in its scoring until variance is 1000, at which point it closely follows the results of mAP and moLRP. We see that as uncertainty increases, PDQ provides increased leniency, however, the highest score attainable drops reinforcing the idea that PDQ requires accurate detections with accurate spatial probabilities as stated within the main paper.

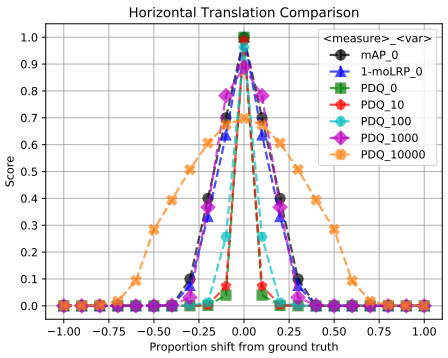


Figure 13: Evaluation of the effect of translation on mAP, moLRP, and PDQ scores. X-axis shows proportional shift of detection box either to the left (negative) or right (positive). Variance (*var*) refers to the variance of corner Gaussians of the PBox detections. BBox is used when *var* is zero. We see mAP and moLRP are lenient to BBox detections with no uncertainty when compared to PDQ and that PDQ is more lenient the more uncertain the detector is.

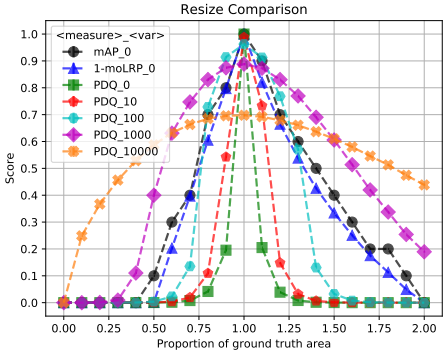


Figure 14: Evaluation of the effect of scaling on mAP, moLRP and PDQ scores. X-axis shows the proportional size of the detection to the ground-truth object. Variance (*var*) refers to the variance of corner Gaussians of PBox detections. BBox is used when *var* is zero. We see mAP and moLRP are lenient to detections with no uncertainty compared to PDQ and that PDQ is more lenient the more uncertain the detector is.

Scaling Error Using the same experimental setup as the translation tests, rather than translating detections, we keep detections centred around the square object and adjust the corner locations such that the area of the square generated by them is proportionally bigger or smaller than the original object. The results from this are shown in Figure 14.

This reinforces the findings of the translation tests, showing how PDQ strongly punishes over-confidence or under-confidence in spatial uncertainty. When there is greater deviation in box size, PDQ is more lenient when the uncertainty is higher. We do not see this same response from

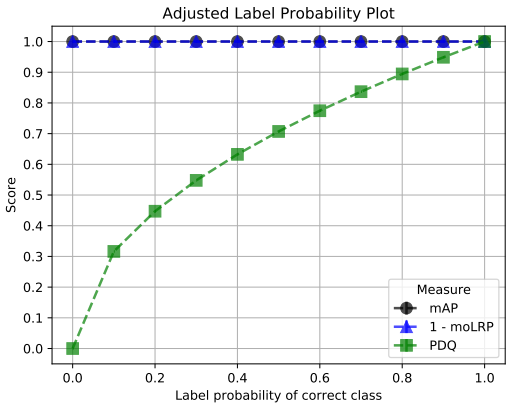


Figure 15: Effects of adjusting label confidences on mAP, moLRP, and PDQ when label probability for the correct class is adjusted using simulated detections on the COCO 2017 validation dataset. We see that existing measures are unaffected as long as the correct class is the class with highest probability in the label distribution. PDQ by comparison decreases with the label probability.

mAP and moLRP which treat standard BBoxes with high confidence in a similar manner to PDQ on PBoxes with variance of 100. We see from both this and the translation test that PDQ rewards boxes with high predicted variance when the actual variance of the box is high. This reinforces the finding of the main paper which states that PDQ requires accurate estimates of spatial uncertainty.

B.3. Label Quality

As demonstrated in the main paper, PDQ explicitly measures label quality, unlike existing measures. We performed an additional test on the COCO 2017 validation data[26] using simulated detectors beyond that done in Section 6 of the main paper. In this test, we set the label confidence for the correct class of each simulated detection to a given value and evenly distribute the remaining confidence between all possible other classes. The results from this experiment when using perfectly aligned BBox simulated detections are shown in Figure 15. This reinforces what had been seen previously, that existing measures are not explicitly effected by label probability, except when the maximum label confidence does not belong to the correct class.

B.4. Missed Ground-truth Objects

We provide the results of two experiments that show that PDQ and existing measures perform the same when ground-truth objects are missed. The first experiment is a simplified scenario where we add an increasing number of small 2 x 2 square objects around the edge of a single image with one large ground-truth object within it. In this image, only the large ground-truth object is ever detected and the detec-

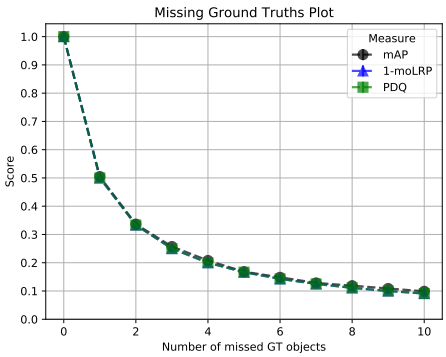


Figure 16: Evaluation of the effect of missing ground-truth objects on evaluation scores in simplified scenario. We observe that all measures respond the same to missed ground-truth objects.

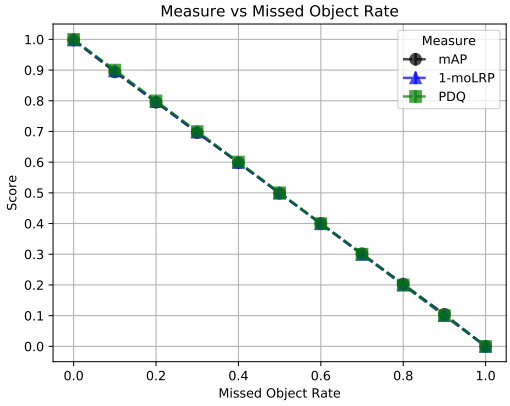


Figure 17: Evaluation of the effect of missing an increased proportion of ground-truth objects on COCO 2017 validation dataset images. We see the response from all measures is the same.

tion is spatially and semantically perfect. Results for mAP, moLRP, and PDQ for this scenario are visualised in Figure 16. The second experiment is performed on the COCO 2017 validation data using simulated detectors as done previously. Here we define a missed object rate for all detectors which dictates the probability that a detection is generated for the given ground-truth object. This was done for perfectly spatially aligned BBox detections and results can be seen in Figure 17.

Both experiments show that, despite their other differences, mAP, moLRP, and PDQ respond the same to missed ground-truth objects (FNs).

B.5. False Detections

We provide the results of a simplified scenario to show that, excluding edge cases that will be discussed in Section 8, mAP, moLRP, and PDQ respond almost the same to false positive detections. To demonstrate this, we test

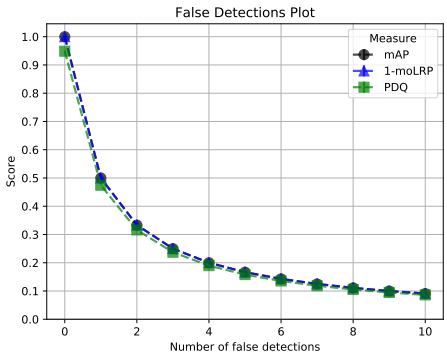


Figure 18: Evaluation of the effect of false detections on evaluation scores. We observe that generally, all measures respond the same to false detections.

a scenario where a single object in a single image is provided with a single perfectly spatially aligned detection and an increasing number of small 2 x 2 detections around the edge of the image. The correct detection always has a label probability of 0.9 and all subsequent detections have a label probability of 1.0 so as to avoid edge cases for mAP explained and discussed in Section 8. We plot the resultant mAP, moLRP, and PDQ scores in Figure 18.

Here we again observe consistency between the mAP, moLRP, and PDQ responses to false detections despite their differences in formulation. Variations between PDQ and the other measures are caused by the lower label confidence for the correct detection which is known to effect PDQ. While the responses here are almost identical, we have identified situations wherein mAP and moLRP obscure FP detections and lessen their impact.

C. Traditional Measures Obscuring False Positives.

In the main paper, we describe how mAP and moLRP are able to obscure the impact of FPs present in the detections presented for evaluation. To support these statements, we produce some simplified scenarios designed to demonstrate unintuitive outputs from mAP and moLRP when given FP detections. Whilst not representative of how these measures are meant to act, they show unusual behavior for testing deployed detectors that PDQ does not share. We do this through multiple test scenarios.

C.1. Duplicate 100% Confident Detections

In the first scenario, we consider detecting a single object in a single image where there is an increasing number of perfectly-aligned, 100% confidence detections of that single object. Results of this scenario are shown in Figure 19. We observed that PDQ and moLRP penalised the additional FP detections, whereas mAP gave 100% accuracy at all times.

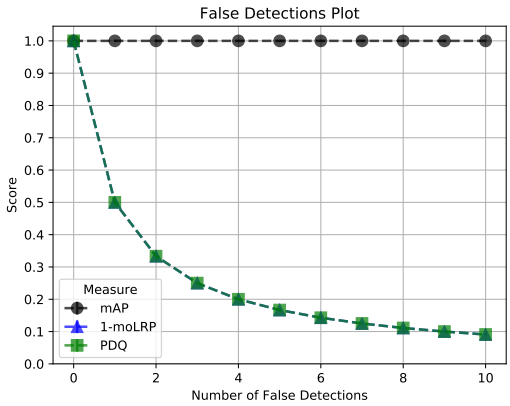


Figure 19: Duplications test results showing mAP, moLRP, and PDQ values when perfect duplicate FP detections are added in a one-object scenario. The TP detection is evaluated before the FPs, causing subsequent FPs to be ignored by mAP. PDQ and moLRP respond as expected, penalising FPs.

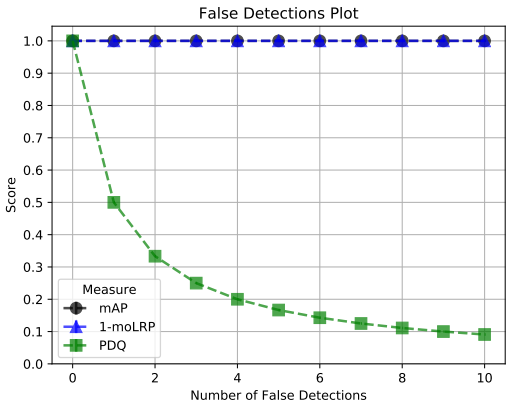


Figure 20: False detection test where all FP detections have slightly lower label confidence than the TP detection (90% Vs 100%). Both mAP and moLRP are shown to treat this as perfect detection output.

This edge case breaks mAP due to how the PR curve for this scenario is generated and utilised. As is explained later in Section 8, the PR curve used for mAP uses the maximum precision at each level of recall to provide a smooth PR curve. However, through this approach, it is assumed that as detections are added to the analysis, the result will be continually increasing recall. Once the recall becomes perfect, or reaches some maximum value, any further false detections are ignored. Here, as all detections have 100% confidence and perfectly overlap the ground truth, the first detection is treated as the TP and all others are ignored. The same effect would occur regardless of whether detections are perfect duplicates or located randomly within the image, as long as the TP is ordered first in confidence order (or in input order in the case of ties, see section 8). This is why we attain the result for mAP shown in Figure 19.

This is not a new problem with mAP, and such behaviour caused by relative ranking has been outlined in past works [37]. In comparison to this, moLRP and PDQ respond as expected to an increasing number of FP detections. This is because both explicitly measure the number of false positives or the false positive rate from the detector output. While robust to this first scenario, our second scenario shows that moLRP can also respond to false positive detections in the same unintuitive manner as mAP.

C.2. False Detections with Lower Confidence

Here, we consider a single image with a single object which is detected by a BBox detection of perfect spatial and semantic quality. In addition to this, we introduce an increasing number of small false detections with label confidence 90% around the border of the image. The results from this scenario are shown in Figure 20.

We observe in this scenario that mAP and moLRP both consider the results as perfect, regardless of the number of FPs, while PDQ penalises the increasing number of FP detections. This mAP result comes from the same relative ranking issues as outlined in the previous scenario (Section 8). The moLRP result, on the other hand, has changed due to the optimal thresholding done as part of the algorithm [34]. The moLRP score is designed to show the best possible performance of the detector if the best label confidence threshold for each class is chosen. Choosing an ideal threshold above 0.9, the performance of the detector becomes perfect, despite the high-confidence false positive detections. This trait of moLRP is beneficial for testing the ideal performance of a detector and for tuning a detector's final output. However, as stated in the main paper this is not beneficial for testing systems to be applied in real-world applications, which cannot choose the optimal threshold on-the-fly during operation. In contrast, PDQ does no such filtering and does not obscure false positive detections.

C.3. Duplicate Detections on COCO Data

Scenario 3 extends scenario 1 (Section 8) from a single image to examine duplicate detections on the COCO 2017 validation data [26]. Again, every detection provided 100% probability of being the correct class and was perfectly spatially aligned. The detections are ordered such that all detections for a given object occur before the detections of the following object. For example, if the number of duplicates is three, the order of detections would be three detections of object A followed by three detections of object B and so on. See Section 8 for why ordering is important. It is expected that for such an experiment, the result for all evaluation measures would be reciprocal in nature (i.e. when there

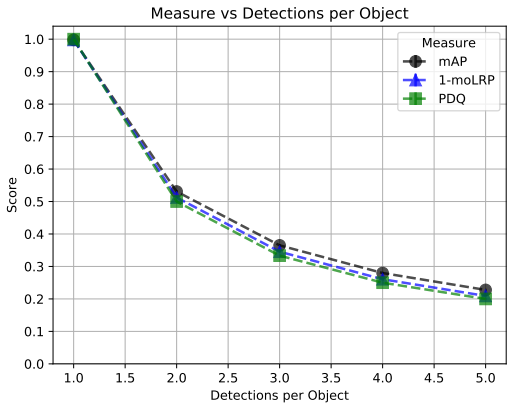


Figure 21: Test results on COCO 2017 validation data comparing scores when a number of perfectly aligned duplicate FP detections are added. Each duplicate FP is ordered directly after their corresponding TP detection. Due to smoothing of the PR curve, calculated precision becomes higher than expected for some classes at different levels of recall, causing mAP to be higher than expected. Other measures remain relatively unaffected.

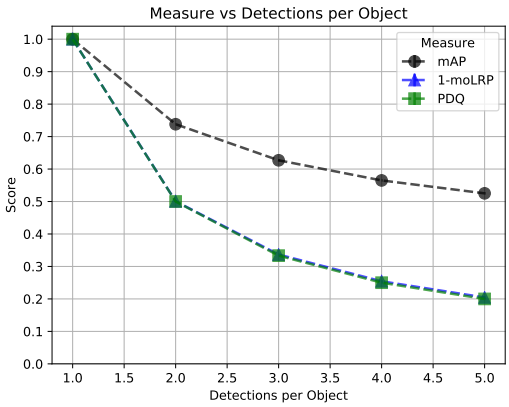


Figure 22: Duplication test results such as done for Figure 21 on subset of 100 images from COCO 2017 validation data. This shows heightened mAP scores from those shown Figure 21 demonstrating increased unintuitive behaviour from mAP as the dataset gets smaller.

are 2 detections per object the score will be 1/2). However, this is not exactly what we observed by our results as shown in Figure 21. What we see from this figure, is that the mAP provides scores slightly higher than expected, whereas PDQ and moLRP measures more closely follow the expected outcomes from such an experiment.

Again, this issue with mAP is caused by the smoothing of the PR curve outlined in Section 8 and the ordering of our detections. As described in Section 8, mAP takes the maximum precision at each of its 101 sample recall values. Additional FPs decrease precision, but don't affect the recall, and so are ignored. As a simplified example, if two detections are given for every object, the recorded precision after 3 objects have been correctly detected is not 0.5 but rather 0.6 as three TPs have been evaluated to only two FPs, despite three FPs being present at this level of recall. This can cause small discrepancies to occur and is the reason for mAP's unusual performance. As we see in the following scenario, this is a problem which increases in severity with small datasets.

C.4. Duplicate Detections on Subset of COCO Data

In the fourth scenario, we increase the severity of the mAP error found in the previous scenario (Section 8). We do this by testing on a subset of the full 5,000 COCO images previously used, evaluating on only the first 100 images. We show these results in Figure 22.

Here we see that the mAP scores are far higher at than expected for each level of detections per object, an exaggeration of the effect in Section 8. This occurs because the smaller number of ground truth instances results in fewer possible measurable recall values. As precision is recorded at 101 set levels of recall, and (as established in Section 8) FPs are obscured until a new measured level of recall is reached, the FPs remain obscured for more recorded levels of recall. Correspondingly, there are fewer total detections at each recorded level of recall, making the number of obscured FPs relatively more significant. This means that more of the recorded maximum precision values are higher, leading to a higher mAP score.

This can ultimately result in the extreme case discussed in Section 8. We observe then that the issues caused by the obfuscation of FPs under mAP increases as the number of samples tested gets smaller. Again, we note that both moLRP and PDQ do not suffer from this issue, as they explicitly measure FPs.

C.5. Duplicate Detections with Lower Confidence on COCO Data

Reinforcing our findings in Sections 8 and 8, we show again that moLRP, while sometimes avoiding pitfalls present in mAP, can still obscure false positive detections through optimal thresholding. In this scenario, we ensure that only the first detection has label confidence of 100% and all subsequent duplicate detections have label confidence of 90%. The results of this test are shown in Figure 23. As expected, PDQ continues to treat the false positives as significant whilst mAP and moLRP both consider the detection output as perfect.

C.6. Summary

In summary, we have demonstrated extreme scenarios showing that both mAP and moLRP can obscure false posi-

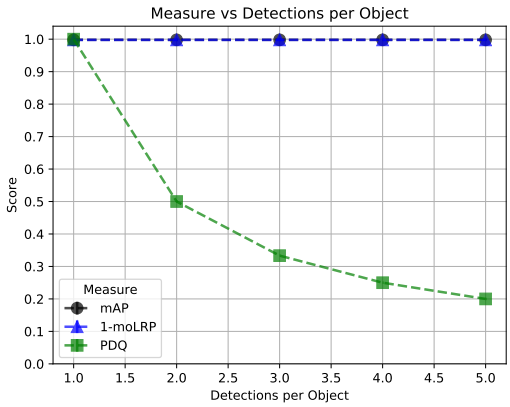


Figure 23: Duplication test results on COCO 2017 validation data where all FP duplicate detections have lower label confidence than the TP detection (90% Vs 100%). Unlike PDQ, both mAP and moLRP are shown to treat this as perfect detection output.

tive detections under different conditions leading to unintuitive results. These issues result from the assumptions made when generating and using PR curves for mAP and optimal thresholding for moLRP. As stated in the main paper, this unintuitive nature is inappropriate behavior for evaluating detectors meant for real-world deployment. We show that PDQ is unaffected by such scenarios, reinforcing the findings of the main paper.

D. Definition of mAP

For the sake of completeness and to aid in understanding the behaviour shown in Section 8, here we define mean average precision (mAP) as used by the COCO detection challenge [26]. Each detection provides a bounding box (BBox) detection location ($\mathcal{B}_j^f$) and a confidence score for its predicted class s_j^f. For each detection in the f-th frame of a given class, mAP assigns detections to ground-truth objects of that same class. Each detection is defined as as either a true positive (TP) if it is assigned to a ground-truth object, or a false positive (FP) if it is not. Detections for each class are ranked by confidence score and assigned to ground-truth objects in a greedy fashion if an intersection over union (IoU) threshold τ is reached. IoU is calculated as follows

$$IoU(\hat{\mathcal{B}}_i^f, \mathcal{B}_j^f) = \frac{area(\hat{\mathcal{B}}_i^f \cap \mathcal{B}_j^f)}{area(\hat{\mathcal{B}}_i^f \cup \mathcal{B}_j^f)}, \tag{7}$$

Data: a dataset of $f = 1 \dots N_F$ frames with detections $\mathcal{D}^f = \{\mathcal{B}_j^f, s_j^f\}_{j=1}^{N_D^f}$ and ground truths $\mathcal{G}^f = \{\hat{\mathcal{B}}_i^f\}_{i=1}^{N_G^f}$ for each frame for a given class $\hat{c}$
Let $\mathcal{U}$ be the set of unmatched objects
forall frames in the dataset **do**
  order detections by descending order of s_j^f
  forall detections in frame **do**
    $\mathcal{G}_*^f = \text{argmax}_{\mathcal{G}^f} IoU(\mathcal{G}_i^f, \mathcal{D}_j^f)$ **if** $IoU(\mathcal{G}_*^f, \mathcal{D}_j^f) > \tau$ and $\mathcal{G}_*^f \in \mathcal{U}$ **then**
      $z_j^f = 1$
      $\mathcal{U} = \mathcal{U} - \mathcal{G}_*^f$
  end
end
Return $\mathbf{z} = [z_1^1, z_2^1, \dots, z_{N_D^{N_F}}^{N_F}]$

Algorithm 2: mAP Detection Assignment

where $\hat{\mathcal{B}}_i^f \cap \mathcal{B}_j^f$ is the intersection of the ground-truth and detection bounding boxes and $\hat{\mathcal{B}}_i^f \cup \mathcal{B}_j^f$ is their union. The assignment process is summarized by Algorithm 1 and results in an identity vector $\mathbf{z}$ which describes for each detection, whether it is a TP or FP with values of 1 or 0 respectively.

After the assignment process is conducted for all images, a precision-recall (PR) curve is computed from the ranked outputs of the given class. Precision and recall are calculated for each detection as it is "introduced" to the evaluation set in order of highest class confidence (and then in submission order in the event of confidence ties). Precision is defined as the proportion of detections evaluated that were true positives, and recall is defined as the proportion of ground-truth objects successfully detected. After generating the PR curve for the given class, the maximum precision is recorded for 101 levels of recall uniformly spaced between zero and one. The maximum precision is used to avoid "wiggles" in the PR curve, resulting in a smoothed PR curve. If no precision has been measured for a given level of recall, the precision at the next highest measured level of recall is recorded. Maximum precision at recall values above the highest reached are 0, to handle false negatives (FNs). This process on a simple scenario is outlined visually in Figure 24. This is process repeated for every evaluated class and at multiple values of τ. The average of all recorded precision values across all IoU thresholds, classes, and recall levels, provides the final mAP score.

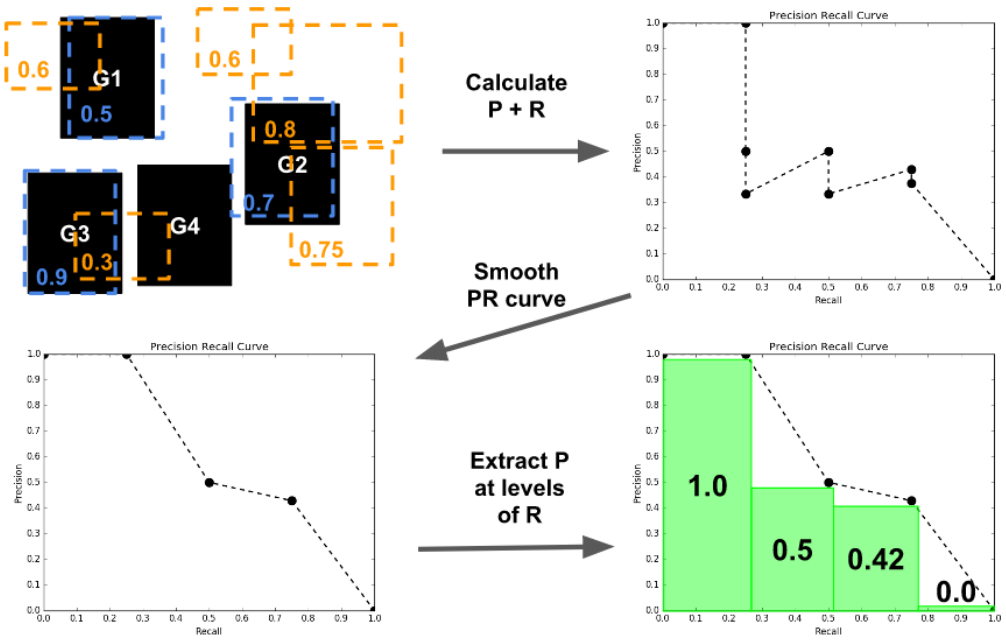


Figure 24: Process for extracting precision values from a PR curve for a given object class at a given threshold. The top-left shows the example scenario with ground-truth objects shown as black boxes, true-positive detections shown as light blue BBoxes, and false-positive detections shown as orange BBoxes. Numbers within the boxes represent label confidence. Top-right figure shows PR curve generated as each detection is added in order of decreasing label confidence. Bottom-left figure shows the effect of smoothing the PR curve by only taking the maximum precision values. Bottom-right shows the precision values extracted for a given range of recall values examined. Note that 101 samples are made across different levels of recall. Best viewed in colour.