Probability-based Detection Quality (PDQ): A Probabilistic Approach to Detection Evaluation

David Hall[1,2], Feras Dayoub[1,2], John Skinner[1,2], Peter Corke[1,2], Gustavo Carneiro[1,3], Niko Sünderhauf[1,2]
[1]Australian Centre for Robotic Vision
[2]Queensland University of Technology (QUT), [3]University of Adelaide
[2]{d20.hall, feras.dayoub, j6.skinner, peter.corke, niko.suenderhauf}@qut.edu.au
[3]gustavo.carneiro@adelaide.edu.au

Abstract

We propose a new visual object detector evaluation measure which not only assesses detection quality, but also accounts for the spatial and label uncertainties produced by object detection systems. Current evaluation measures such as mean average precision (mAP) do not take these two aspects into account, accepting detections with no spatial uncertainty and using only the label with the winning score instead of a full class probability distribution to rank detections. To overcome these limitations, we propose the probability-based detection quality (PDQ) measure which evaluates both spatial and label probabilities, requires no thresholds to be predefined, and optimally assigns ground-truth objects to detections. Our experimental evaluation shows that PDQ rewards detections with accurate spatial probabilities and explicitly evaluates label probability to determine detection quality. PDQ aims to encourage the development of new object detection approaches that provide meaningful spatial and label uncertainty measures.

1. Introduction

Visual object detection provides answers to two questions: *what* is in an image, and *where* is it? Embedded in a robot or an autonomous system such as a driverless car, object detection is part of a complex, embodied, active, and goal driven system, which provides crucial information that ultimately result in actions and interaction of the robot with its environment [1, 7]. In such a scenario, mistakes in object detection can lead to catastrophic outcomes that not only risk the success of the robot's mission, but potentially endanger human lives [11].

For safe and trusted operation in robots or autonomous systems in general, object detectors must express meaningful *uncertainty* information for both the reported labels (*what?*) and bounding boxes (*where?*) [18]. This would enable treating object detectors as yet another sensor within the established and trusted framework of Bayesian information fusion [16]. However, while state-of-the-art object detectors report at least an *uncalibrated* indicator of uncertainty via label distributions or label scores [5, 12], they currently do *not* report spatial uncertainty. As a result, evaluating the quality of the label uncertainty or spatial uncertainty is not within the scope of typical benchmark measures and competitions.

Much of the progress in computer vision, machine learning, and object detection in particular, has been driven by meaningful and carefully designed datasets, competitions, and evaluation measures [3, 9, 17]. A well-designed evaluation measure can provide direction and guidance for a research community to evolve their approaches towards impactful goals.

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

This paper proposes a new probability-based detection quality (PDQ) measure that jointly evaluates both label and spatial probabilities of object detectors unlike current detection measures such as mean average precision (mAP) [9]. Unlike mAP, PDQ rewards detections with accurate spatial probabilities, explicitly evaluates label probability to determine detection quality, uses no predefined thresholds to determine successful detections, and provides optimal assignment of detetions to ground-truth objects. We demonstrate this with an in-depth analysis of PDQ's behaviour and we compare it to mAP, highlighting core differences and similarities.

2. Related Work

For the past decade, detection algorithms have predominantly been evaluated using average precision (AP) or variants based upon it. Average precision was introduced for the PASCAL-VOC challenge in 2007 to replace measuring the area under the ROC curve [3]. Since that time, it became the standard evaluation measure in the PASCAL-VOC challenge and is the basis for analysis in many other works examining object detection [2, 8, 9, 10, 15, 17]. AP determines if a detection adequately describes a ground-truth object based upon the predicted class and the intersection over union (IoU) between detection and ground-truth bounding boxes. Using this, a precision-recall (PR) curve is generated and precisions at 11 levels of recall are extracted and then averaged to provide a final score. The critical factor in AP, which has been the focus for most of its adaptations, is the IoU threshold which must be reached before a detection can be considered a true positive. Originally, in order to compensate for errors in the ground-truth labelling, the IoU threshold set by the PASCAL-VOC challenge was generous at 0.5.

The highly influential ImageNet challenge [17], which can be directly linked with the rise of deep learning for object detection [4], uses a variant to the AP used by PASCAL-VOC where detection of small objects are treated differently. Rather than discounting small objects like PASCAL-VOC, ImageNet instead utilised a variable IoU threshold based upon the width and height of the object being observed. This enable detections which were off by up to 5 pixels on average to still be counted as valid detections. This variable threshold was used unless it exceeded the default threshold of 0.5.

Another variation created for the COCO detection challenge [9], is the mean average precision (mAP). This variation calculates the precision values at 101 steps along a PR curve, for each class, for a range of IoU thresholds between 0.5 and 0.95, and then averages all of these precision values to determine overall performance. This provided a harsher critique of detection systems, requiring a higher level of spatial accuracy to get a high score. This is the measure we compare to within our later experiments.

To address shortcomings they found in AP-based techniques, Oksuz et al. introduced localization recall precision (LRP) [13]. The key issues they identified with AP were its inability to distinguish between very different PR curves and it not allowing for direct measurement of bounding box localization accuracy. These limitations were addressed by splitting the metric into three parts examining: localization, false positives, and false negatives – allowing for a deeper understanding of the localization errors caused by detection systems.

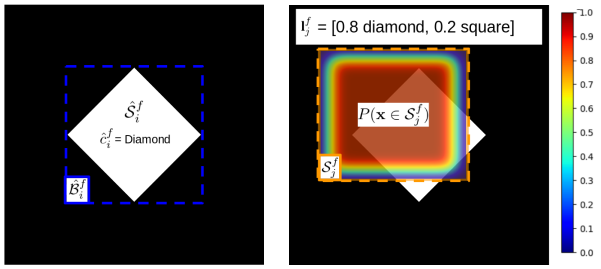
While these measures successfully build upon the foundation of the AP algorithm, their base principles and assumptions limit their effectiveness for evaluating object detection systems adequately for robotics and computer vision applications where the expression of accurate probabilities is important. Fundamentally, each measure is based on generating PR curves, assuming that every detection is entirely correct or incorrect. This has entrenched the notion of using only maximum class label confidences to rank detections, IoU thresholding, and ignoring spatial uncertainty. We argue that this leads to sub-optimal detection assignment, a tuneable definition of success, and promotes detection systems that don't estimate spatial uncertainty. To the best of our knowledge, there are no detection evaluation measures which are free of all these limitations.

3. Probability-based Detection Quality (PDQ)

Addressing the aforementioned limitations with AP-based measures, we introduce the major contribution of this paper: the probability-based detection quality (PDQ) measure which utilizes both spatial and label probabilities in determining the quality of detections. 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 tunable parameters. Furthermore, PDQ is based on an approach that provides optimal assignment of detections to ground-truth objects, incorporating both the label and spatial attributes of the detections in this assignment.

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 $\hat{c}_i^f$. Any dataset evaluated using PDQ must provide pixel-level annotations and class labels for every object evaluated.

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 $\mathbf{l}_j^f$. All detections must provide both a spatial probability function and a full label probability distribution in order to be evaluated with PDQ. Whilst the spatial probability requirement is novel and challenging, later in this paper we define a novel variant to traditional bounding boxes that provides a method for estimating these probabilities. A visualisation of both ground-truth objects and detections is provided in Figure 1.



(a) ground-truth object (b) detection

Figure 1: Example of a ground-truth object (a) and a detection (b) as defined in our work. Ground-truth object (a) consists of a segmentation mask $\hat{\mathcal{S}}_i^f$ shown by white pixels, a bounding box $\hat{\mathcal{B}}_i^f$ shown as blue box, and a class label $\hat{c}_i^f$ which here is diamond. Detection object (b) consists of a probability function $P(\mathbf{x} \in \mathcal{S}_j^f)$ which is shown here as a heatmap of values between 0 and 1, a segmentation mask $\mathcal{S}_j^f$ shown as all pixels within the orange box, and a probability distribution across all classes $\mathbf{I}_j^f$ which here provides probabilities for diamond and square classes.

Data: a dataset of $f = 1 \dots N_F$ 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

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 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. Algorithm 1 summarises the overall PDQ calculation. In the following, we detail each of the involved steps and both quality measures.

3.1. 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:

$$\text{pPDQ}(\mathcal{G}_i^f, \mathcal{D}_j^f) = \sqrt{Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f) \cdot Q_L(\mathcal{G}_i^f, \mathcal{D}_j^f)}. \quad (1)$$

Next we describe the details for each quality measure.

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 calculated 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:

$$Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f) = \exp(-(L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f) + L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f)), \quad (2)$$

where $Q_S(\mathcal{G}_i^f, \mathcal{D}_j^f) \in [0, 1]$. The spatial quality in (2) 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.

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

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, unlike AP-based methods using bounding-box IoUs, as illustrated in Figure 2.

Background Loss The background loss term L_{BG} penalises any probability mass the detection incorrectly assigned to pixels outside the ground truth bounding box. It

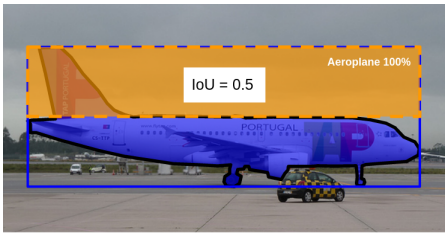


Figure 2: Example of a detection of an aeroplane, shown as an orange box, a ground-truth box, shown as a blue line, and a ground-truth segmentation mask, shown as a blue-coloured region with a black border. At an IoU threshold of 0.5, AP-based methods consider this detection entirely correct despite detecting 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 is 3.64×10^{-6}. Best viewed in colour.

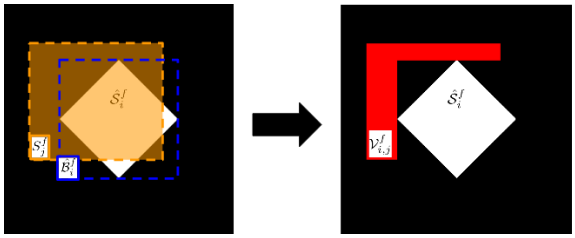


Figure 3: Example visualisation of how the background evaluation region $V_{i,j}^f$ (red region) is defined for a given detection region $\mathcal{S}_j^f$ (orange box) trying to describe a given ground-truth region $\hat{\mathcal{S}}_i^f$ (white diamond) which is encompassed by bounding box $\hat{\mathcal{B}}_j^f$ (blue box). Best viewed in colour.

is formally defined as

$$L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f) = -\frac{1}{|\hat{\mathcal{S}}_i^f|} \sum_{\mathbf{x} \in \mathcal{V}_{ij}^f} \log((1 - P(\mathbf{x} \in \mathcal{S}_j^f))), \quad (4)$$

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 3. 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.

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:

$$Q_L(\mathcal{G}_i^f, \mathcal{D}_j^f) = \mathbf{l}_j^f(\hat{c}_i^f). \quad (5)$$

3.2. 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, 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 (1) between possible detection-object pairings to determine the optimal assignment through the Hungarian algorithm [6]. This provides the optimal assignment between two sets of information which produces 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 missing (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.

3.3. 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.

4. Probabilistic Bounding Boxes (PBoxes)

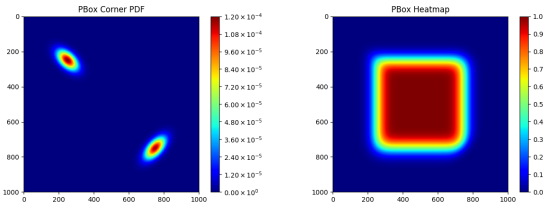
PDQ is intentionally designed to evaluate the spatial uncertainty of object detectors. However, current detectors which use bounding boxes (BBoxes) for object localisation do not express spatial uncertainty. To provide a measure of spatial uncertainty suitable for evaluation under PDQ, we propose expressing detections as probabilistic bounding boxes (PBoxes).

In comparison to standard BBoxes, PBoxes do not define their corners with absolute certainty. Instead, PBoxes express their top-left and bottom-right corners as Gaussian distributions depicting where the box corners that describe the object may exist. Formally, the PBox for the j-th detection in the f-th frame $\mathcal{B}_j^f = \{\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\}$ where $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ are the mean vectors and covariance matrices for the multivariate normals describing the top-left and bottom-right corners of the PBox $(\mathcal{N}_0, \mathcal{N}_1)$, respectively. An example of these corners is shown in Figure 4a.

For PBox detections $P(\mathbf{x} \in \mathcal{S}_j^f)$ is the probability that the pixel would exist within a standard BBox generated by sampling the PBox distributions. This representation provides high probabilties for pixels at the centre of the box and gradually decreases the probabilty towards the edge of the box. This probability distribution is in comparison to the naive approach for standard BBoxes, where the spatial probability is 1 for each pixel within the box has and 0 for all other pixels. The formalised procedure for calculating $P(\mathbf{x} \in \mathcal{S}_j^f)$ for PBoxes as used in our work can be found in the Appendix. An example of a probability heatmap for a PBox is shown in Figure 4b.

5. Evaluation of PDQ Traits

We demonstrate the key characteristics of PDQ compared with mAP. We show how each of them responds to different types of imperfect detections. Specifically, we highlight how PDQ rewards detections which provide an accurate estimation of spatial probabilities and how it explicitly evaluates label probability when determining detection quality. We also show that PDQ and mAP provide similar responses when faced with false detections and missed ground-truth objects. Further comparisons between PDQ and mAP on different forms of error can be found in the Appendix.



(a) PBox corner Gaussians (b) PBox heatmap

Figure 4: Visualization of the Gaussian corners (a) and probability heatmap (b) of a probabilistic bounding box where $\boldsymbol{\mu}_0 = [250, 250]$, $\boldsymbol{\mu}_1 = [750, 750]$, $\boldsymbol{\Sigma}_0 = \left[\begin{smallmatrix} 1600 & 800 \\ 800 & 1500 \end{smallmatrix}\right]$, and $\boldsymbol{\Sigma}_1 = \left[\begin{smallmatrix} 1600 & -1000 \\ -1000 & 1800 \end{smallmatrix}\right]$

5.1. Accurate Spatial Probabilities

PDQ rewards detections that provide accurate spatial probabilities based on the output of the detector. To demonstrate this point quantitatively, we evaluated detections on the COCO dataset using simulated detectors which produce PBox detections based upon known ground-truth objects. For each ground-truth object, we create one PBox detection meant to describe it. The location of the top-left and bottom-right corners for the created PBox's detection are sampled from multivariate Gaussians centred at the corresponding corners of the ground-truth object's bounding box. Each detection is provided with perfect label predictions matching their ground-truth object.

PDQ is designed to reward detections which accurately predict the actual spatial uncertainties inherent to the detector. We therefore expect that PDQ scores will be maximised when the detector's PBoxes model their spatial probabilities by Gaussian corners which match the distribution that is used to generate them.

To observe this, we vary the covariance matrices used to produce corner locations (actual covariance) and the reported covariance matrices of the PBox's Gaussian corners (predicted covariance). Covariance matrices are modelled as spherical Gaussians with a variance which is varied to simulate different detector behaviour.

Experiments were run on 100 images from the COCO dataset for simulated detectors with varied actual and predicted covariances. Tests on these 100 images were repeated 20 times to accommodate for randomisation in corner location sampling, and we report the average PDQ and mAP scores across all 20 tests. We plot the results from

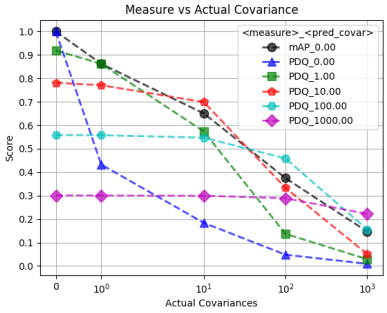


Figure 5: Test for examining the effect of accurate spatial probabilities. Plot shows mAP and PDQ scores Vs actual covariance. We see that for each level of actual covariance, the highest ranking detector under PDQ is the one where the predicted covariance (pred_covar) matches the actual covariance of the detector. We also see that mAP (black circles plot) is more lenient on detections with no spatial uncertainty than PDQ under the same conditions (blue triangles plot). Best viewed in colour.

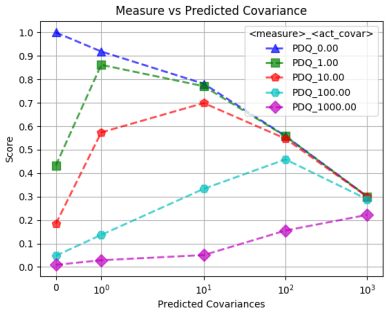


Figure 6: Test for examining the effect of accurate spatial probabilities. Plot shows PDQ scores Vs predicted covariance. We see that the PDQ peaks when predicted covariance matches the actual covariance (act_covar). Best viewed in colour.

these tests, examining mAP and PDQ scores versus actual covariance and versus predicted covariance in Figures 5 and 6, respectively.

From these plots we observe three important points. Firstly, in Figure 5 we observe that mAP is more lenient than PDQ on standard BBoxes which have no spatial uncertainty. PDQ scores for such BBoxes drop below 0.5 when actual covariance is only 1.0 whilst mAP is approximately 0.85. This rapid decline in PDQ is caused by the spatial loss incurred by entirely missing foreground pixels and detecting background pixels as foreground without any uncertainty. Secondly, we note from Figure 5 that for each level of actual covariance, the highest ranked detector is the one which provides an accurate prediction of that covariance. This is reinforced when analysing Figure 6 as we see that for each level of actual covariance, PDQ will peak when the actual covariance matches the predicted covariance. Examining the response of PDQ when adjusting the predicted covariance as shown here can also be used to find the best estimate of a known detector's covariance as we shall demonstrate in Section 6. These plots show that unlike mAP, PDQ rewards systems that provide accurate representations of spatial probabilities.

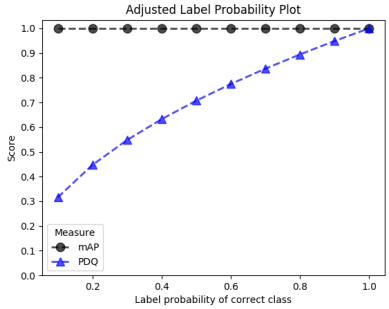


Figure 7: Evaluation of the effect of varying true-class probability on mAP and PDQ scores when the true-class always has the highest class probability. We observe that PDQ declines while mAP is unaffected. Best viewed in colour.

5.2. Label Probability

Unlike mAP, PDQ explicitly evaluates label probability to determine detection quality. To show this, we perform two experiments on detections that are perfectly spatially aligned, describing every pixel for a single object in a single image. For the first test, we vary the reported label probability for the correct class whilst ensuring that it always remains the dominant class in the probability distribution of the BBox. For the second test, we consider a two-class scenario, varying the correct class label probability in the same way as before but where the correct class is not guaranteed to be the dominant class. The plots for these results are shown in Figures 7 and 8.

Here we observe that label probability has no direct effect on mAP beyond dictating which class is the maximum class. In mAP, label probability is only used for ranking possible detection matches, and does not explicitly measure the probability of the given object's class. For mAP, only the ranking of label probabilities matters and not the explicit values. This is why in both plots, mAP is either 0 or 1 depending on which class is the maximum class. In comparison, PDQ has a similar response to the detections, regardless of which class is the maximum class. PDQ scores are adjusted according to the probability given to the correct class rather than simply assigning a "right" or "wrong" label.

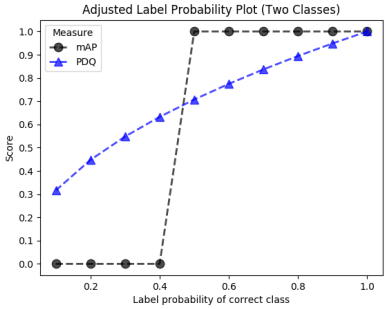


Figure 8: Evaluation of the effect of varying true-class probabilities on mAP and PDQ scores in a two-class scenario. We observe that mAP is binary, always producing scores of 0 or 1, whereas PDQ provides gradual change similarly to that seen in Figure 7. Best viewed in colour.

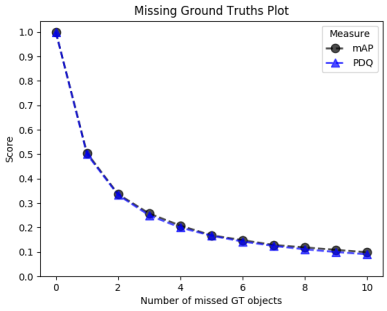


Figure 9: Evaluation of the effect of missing ground-truth objects on evaluation scores. We observe that mAP and PDQ respond the same to missed ground-truth objects.

5.3. Missing Ground-truth Objects

PDQ and mAP have the same response to missed ground-truth objects. To show this, 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 it has a perfect detection both in terms of spatial and label probabilities. Results for mAP and PDQ for this scenario are visualised in Figure 9.

Examining these results, we observe consistency between PDQ and mAP when evaluating missed ground-truth objects, despite their other differences. This finding is reinforced in the Appendix.

5.4. False Detections

PDQ and mAP have the same response to false detections. To demonstrate this, we test 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 in the top-left corner of the image.

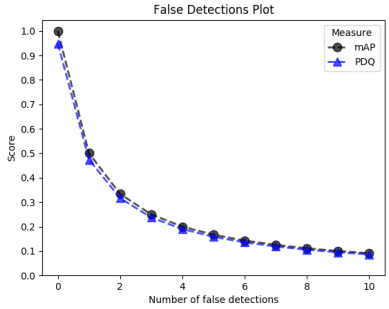


Figure 10: Evaluation of the effect of false detections on evaluation scores. We observe that both mAP and PDQ respond the same to false detections.

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 an edge case for mAP which is explained in the Appendix. We plot the resultant mAP and PDQ scores in Figure 10.

Here we again observe consistency between the mAP and PDQ responses to false detections despite their differences in formulation. Variations between PDQ and mAP are caused by the lower label confidence for the correct detection which are known to effect PDQ. The only differences observed between the two are for edge cases which disrupt mAP which are explored in the Appendix.

5.5. Summary

We have shown through controlled experiments that PDQ addresses key limitations of mAP whilst maintaining consistency in its response to false detections and missed ground-truth objects. We demonstrated that unlike mAP, PDQ appropriately includes spatial probability in its evaluations, rewarding detections with accurate spatial probabilities which are consistent with actual location variances. We also identified that PDQ explicitly evaluates label probability rather than using it to simply determine the rank order at which detections are considered entirely correct or incorrect. Despite these differences, we demonstrated that PDQ responds in the same manner as mAP for false detections and missed ground-truth objects. This measure should be utilised and built upon so that future detection systems are evaluated in a probabilistic environment which is more appropriate for computer vision and robotics applications in more realistic and unstructured environments.

6. PDQ Evaluation on Existing Detectors

Whilst current state-of-the-art detection systems are not designed to output accurate spatial probabilities, PDQ can still be used to analyse such systems. We transform BBox detection outputs into PBoxes by modelling the BBox cor-

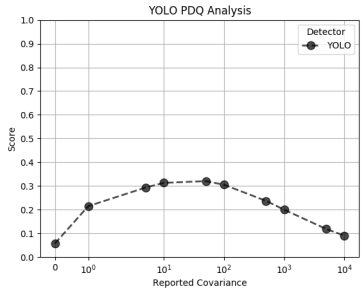


Figure 11: Demonstration of PDQ score on YOLOv3 detector outputs. YOLO BBoxes are transformed into PBoxes with a fixed corner variance. We observe that peak performance for YOLO occurs when the variance is 50 pixels and the best PDQ score achieved is 0.32.

ners as Gaussians with a given spherical covariance matrix. This covariance matrix is kept the same for each corner across all detection outputs. As was done in Section 5.1, the variance of this matrix can be adjusted to examine the resultant effect on PDQ score and find the ideal level of variance for the detector. To demonstrate this, we evaluate the output from a YOLOv3 [14] detector on 5000 COCO images and examine the PDQ score as the variance is adjusted in Figure 11.

Here we observe that, as in our previous experiments, there is an ideal predicted variance that provides peak PDQ performance. For YOLOv3, this ideal variance is seen to be approximately 50 pixels, achieving the highest PDQ score of 0.32. However, this PDQ score is achieved using only an approximation of the corner variances. Future work should strive to move away from traditional BBox detection methods and develop approaches that provide accurate estimations of a detections spatial uncertainty as these will provide the best possible PDQ results.

7. Conclusions

Detection algorithms have long been evaluated in a world of certainty where detections are always either correct or incorrect provided that they meet a given set of thresholds. Whilst this has been the standard approach for some time using evaluation measures such as mAP to evaluate performance, they do not incorporate spatial uncertainties. To overcome this limitation, we introduced the probability-based detection quality (PDQ) measure which incorporates both spatial and label uncertainty, as well as optimal assignment of detections to ground-truth objects. We defined how PDQ is calculated and compared it to mAP, highlighting the key differences and similarities between them.

We showed that PDQ penalises over-confident bounding box detections when their spatial accuracy is not perfect and it rewards detections with accurate representations of spatial probabilities. We showed also that PDQ incorporates label accuracy for more than ranking individual detections, providing a score based upon the probability of the detection label. PDQ is an effective measure for both spatial and label probabilities that determines the overall quality of detections in a manner more appropriate for robotics and computer vision applications than current state-of-the-art evaluation measures.

Acknowledgements

This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision under project CE140100016, and supported by a Google Faculty Research Award to Niko Sünderhauf.

References

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

[2] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[3] 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.

[4] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] T. Namba and Y. Yamada. Risks of deep reinforcement learning applied to fall prevention assist by autonomous mobile robots in the hospital. *Big Data and Cognitive Computing*, 2(2):13, 2018.

[12] 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.

[13] 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.
[14] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, Apr. 2018.
[15] 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*, pages 91–99, 2015.
[16] 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.
[17] 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.
[18] 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.

Appendix Overview

This appendix contains the supplementary material and analysis that was not included in the main paper due to space restraints.

This supplementary material is organized as follows:

1. Extended PDQ Detection Error Analysis - Providing reinforcement and further insight about how PDQ and mAP respond to different errors.
2. Analysis of mAP Edge Cases - Outlining an edge case that causes unintuitive behaviour in mAP which is not present in PDQ
3. Probabilistic Bounding Boxes (BBoxes) - Providing formal procedure for calculating spatial probabilities for PBox detections
4. Definition of mAP - Describing in further detail how mAP is calculated.

A. Extended PDQ Detection Error Analysis

We show results for an extended analysis of PDQ under different types of detection error to reinforce the findings of the main paper and provide additional insights.

Spatial Confidence

We evaluate perfectly aligned BBox detections which have varying values of spatial probability for every pixel therein. Whilst not a realistic type of detection, it allows for easy examination of the reaction of mAP and PDQ to spatial probability variations. The results are shown in Figure 12

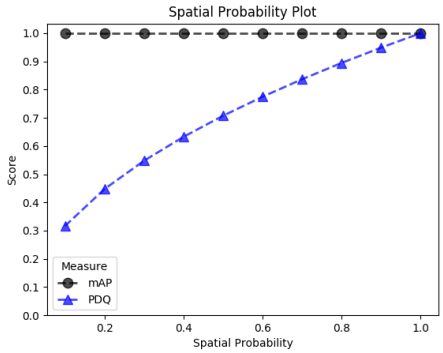


Figure 12: Evaluation of the effect of spatial probability on a perfectly aligned BBox. We see that PDQ, unlike mAP, is effected by spatial probability changes.

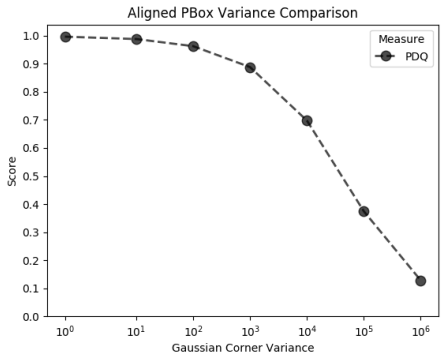


Figure 13: Plot showing the effect on PDQ of increasing variance, and by extension uncertainty, on perfectly aligned PBoxes. We see that for perfectly aligned detections, the score goes down the more uncertain the detection is.

This experiment shows that PDQ is affected by adjusted spatial probabilities whereas mAP consistently considers the provided output to be perfect. This reinforces our statements in the main paper about how PDQ, unlike mAP evaluates spatial probability.

Analysis of PBox covariance on PDQ

To visualise the effect of increasing spatial uncertainty on PDQ using PBoxes, we performed a simple test using perfectly aligned detections 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 with variance which is varied to observe the effect of increased uncertainty. The results of this test are shown in Figure 13. We see a decline in PDQ with increased variance and uncertainty.

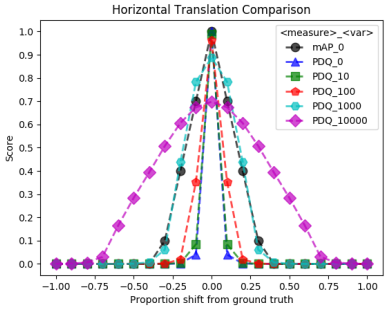


Figure 14: Evaluation of the effect of translation on mAP and PDQ scores. X-axis shows proportional shift of detection box either to the left (negative) or right (positive). Variance (var) refers to the variance of corner Gaussians for PBox detection. BBox is used when var is zero. We see mAP is very lenient to detections with no uncertainty compared to PDQ and that PDQ is more lenient the more uncertain the detector is. Best viewed in colour.

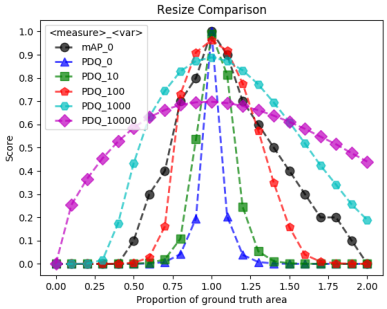


Figure 15: Evaluation of the effect of scaling on mAP and PDQ scores. X-axis shows the proportional size of the detection to the ground-truth object. Variance (var) refers to the variance of corner Gaussians for PBox detection. BBox is used when var is zero. We see mAP is very lenient to detections with no uncertainty compared to PDQ and that PDQ is more lenient the more uncertain the detector is. Best viewed in colour.

Detection Misalignment

We perform two experiments to analyse responses to misaligned detections. The first observes 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 variance. The results from this test are shown in Figure 14.

Here, we see that when there is a BBox with no spatial uncertainty, PDQ strongly punishes any deviation from the ground-truth with scores dropping close to zero after only a 10% shift. This is in strong comparison to mAP 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. We see that as uncertainty increases, PDQ provides increased leniency, however, the highest score attainable drops suggesting PDQ requires accurate detections with accurate spatial probabilities as was said within the main paper.

We see a similar effect when creating detections that are too big or too small for the object being detected. 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 15.

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 mAP which treats 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.

Label Quality

Reinforcing our findings about how, unlike mAP, PDQ explicitly measure label quality, we test on real-world COCO data[9] using the same setup as the experiments shown in Section 5.1 of the main paper. To perform 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 16. This reinforces what had been seen previously, that mAP is not explicitly effected by label probability, except when the maximum label confidence does not belong to the correct class.

Missed Ground-truth Objects

We reinforce our finding that mAP and PDQ react in the same manner to missed ground-truth objects using the same simulated detector setup used in Section 5.1 of the main paper. To show this we use a variable missed object rate defining the probability that a detection is generated for

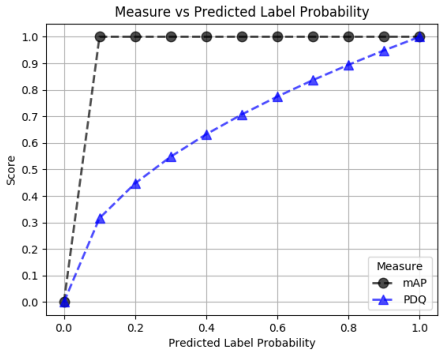


Figure 16: Effects of adjusting label confidences on mAP and PDQ when label probability for the correct class is adjusted using simulated detections on 100 images from the COCO dataset. We see that mAP is unaffected as long as the correct class is the class with highest probability in the label distribution. PDQ by comparison decreases alongside the label probability.

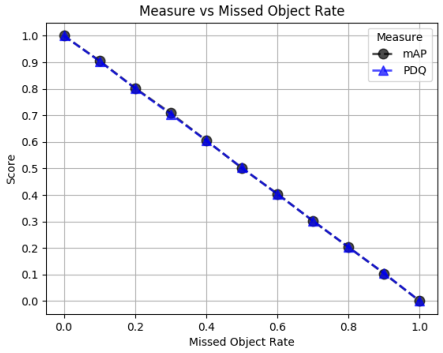


Figure 17: Effects on PDQ and mAP when an increased proportion of ground-truth objects are missed using simulated detections on the COCO dataset. We see the response from both is the same.

each ground-truth object. This was done for perfectly spatially aligned BBox detections. The final results of this test can be seen in Figure 17. This again reinforces the findings of the main paper, that PDQ and mAP respond in the same manner to missed ground-truth objects.

B. Analysis of mAP Edge Cases

When analysing mAP, we identified edge cases which provide unintuitive outputs when faced with false or duplicate detections. Whilst not representative of how mAP is meant to act, these show another fragility to mAP which PDQ does not suffer from.

In the first scenario depicting this edge case, we consider detecting a single object in a single image where at least one detection is perfectly aligned with the ground-truth object with high label probability and multiple extra false detections either perfectly overlap, or are scattered randomly throughout the image. We tested this using a toy scenario where every detection was perfectly aligned with the ground-truth object and every deteciton had 100% label confidence, gradually increasing the number of duplicate detections. We observed that PDQ responded as would be expected for an increasing number of false detections, whereas mAP gave 100% accuracy at all times. This is shown visually in Figure 18.

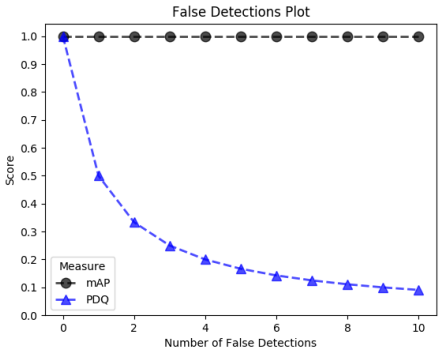


Figure 18: Duplications test results showing mAP and PDQ values when number of perfect duplicate detections (false positives) are added. True positive detection is ordered before false detections and so all subsequent false positives are ignored by mAP, providing perfect 100% score despite false positives. PDQ is unaffected by this edge case

This edge case breaks mAP due to how the PR curve for this scenario is generated and utilised. As is explained later, 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 detecitons are added to the analysis, the result will result in continually increasing recall. If at any point the recall becomes perfect, any further false detections become ignored in the evaluation of mAP. This would be the same whether detections are perfect duplicates or located randomly within the image. The dependence on an assumed structure for a PR curve causes mAP to break under this edge case. These problems from smoothing the PR curve are amplified when used on small datasets due to the large recall that is attained per TP detection. This problem is why we attain the result for mAp shown in Figure 18. As the first detection used to generate the PR curve matches the one object examined, the subsequent detections are not counted after the PR curve smoothing. This is not a new problem with mAP and such behaviour has been outlined in past works [14]. Examining Figure 18, we see that PDQ behaves as expected in this scenario.

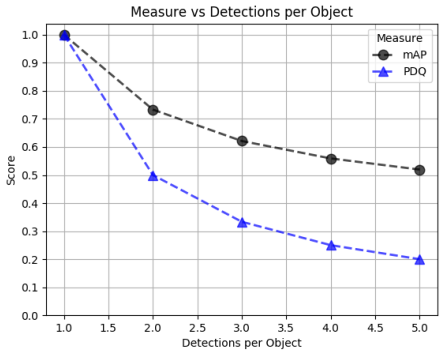


Figure 19: Duplications test results on COCO data showing mAP and PDQ values when number of perfectly aligned duplicate detections (false positives) are added. Each duplicate FP is added directly after their corresponding true positive detection. Calculated precision becomes incorrect for some classes because of this causing mAP to be higher than expected. PDQ is unaffected by this edge case.

Extending this analysis, in a second experiment examining the effect of duplicate detections per ground-truth object, we generated simulated detections based upon known ground-truth objects for 100 images in the COCO dataset [9]. Here, every detection provided 100% probability of being the correct class and was perfectly spatially aligned. It is expected that for such an experiment, the result for mAP would be reciprocal in nature (e.g. when there are 2 detections per object the score will be 1/2). However, this is not what we observed by our results as shown in Figure 19. What we see from this figure, is that the mAP provides scores higher than expected, whereas our PDQ measure follows the nature we would expect from such an experiment.

Again, this issue with mAP is caused by the smoothing of the PR curve. In our scenario, the ordering of detections is such that each correct detection of a given object is followed immediately by its duplicates. Due to the process where only the maximum precision is taken at every value of recall, each FP is evaluated only after a TP detection has been generated. This can lead to scenarios, for example, where 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 the fact that there are three FPs present at this level of recall. It should be noted again that these issues with mAP are more pronounced the smaller the set of data being analysed. We again see that PDQ does not suffer from this same unusual behaviour.

In summary, there are edge cases where mAP responds in a manner which does not fit what mAP means to measure. These issues result from the assumptions made when generating PR curves for mAP, become more pronounced on small datasets, and ares not present when using PDQ.

C. Probabilistic Bounding Boxes (PBoxes)

Here we provide the formal process for calculating the probability function $P(\mathbf{x} \in \mathcal{S}_j^f)$ for BBox and PBox detections. Usually, a BBox detection defines a set of box coordinates $\mathcal{B}_j^f = \{\mathbf{x}_0, \mathbf{x}_1\}$ where $\mathbf{x}_0$ and $\mathbf{x}_1$ define the pixel coordinates of the top-left and bottom-right corners of the BBox. For such a box, we dictate that $P(\mathbf{x} \in \mathcal{S}_j^f)$ will be uniform for all pixels within the confines of this box. By comparison, PBoxes do not define precisely where their top-left and bottom-right corners exist, instead defining each as a Gaussian distribution for where the top-left and bottom-right corners might exist.

Using the definition for PBox corners outlined in the main paper, we define the probability function for a detection described by a PBox as follows.

$$P(\mathbf{x} \in \mathcal{S}_j^f) = P((0,0) \leq \mathbf{x}_0 \leq \mathbf{x}) \times P(\mathbf{x} \leq \mathbf{x}_1 < (w,h))$$

where here, $\mathbf{x}_0$ and $\mathbf{x}_1$ represent top-left and bottom-right corners governed by the multivariate distributions $\mathcal{N}_0$ and $\mathcal{N}_1$ respectively, and (w,h) is the width and height of the image. We define that for two hypothetical pixels ($\mathbf{x}_a$, $\mathbf{x}_b$), $\mathbf{x}_a \leq \mathbf{x}_b$ if both the x and y coordinates of $\mathbf{x}_a$ are less than or equal to the x and y coordinates of $\mathbf{x}_b$ respectively. The reverse of this is applied for $\mathbf{x}_a \geq \mathbf{x}_b$. We assume a coordinate frame where the origin is at the top-left corner of the image with positive x and y values to the right and below that point. This is calculated for every pixel in the frame and then, for simplicity, we dictate the final set of detection pixels by thresholding the probabilities such that $\mathcal{S}_j^f = \{\mathbf{x} | P(\mathbf{x} \in \mathcal{S}_j^f) > \tau\}$. The threshold τ was chosen as the probability of being outside three standard deviations from a Gaussian distribution (0.00135).

D. Definition of mAP

For the sake of completeness, here we define mean average precision (mAP) as used by the COCO detection challenge [9]. 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

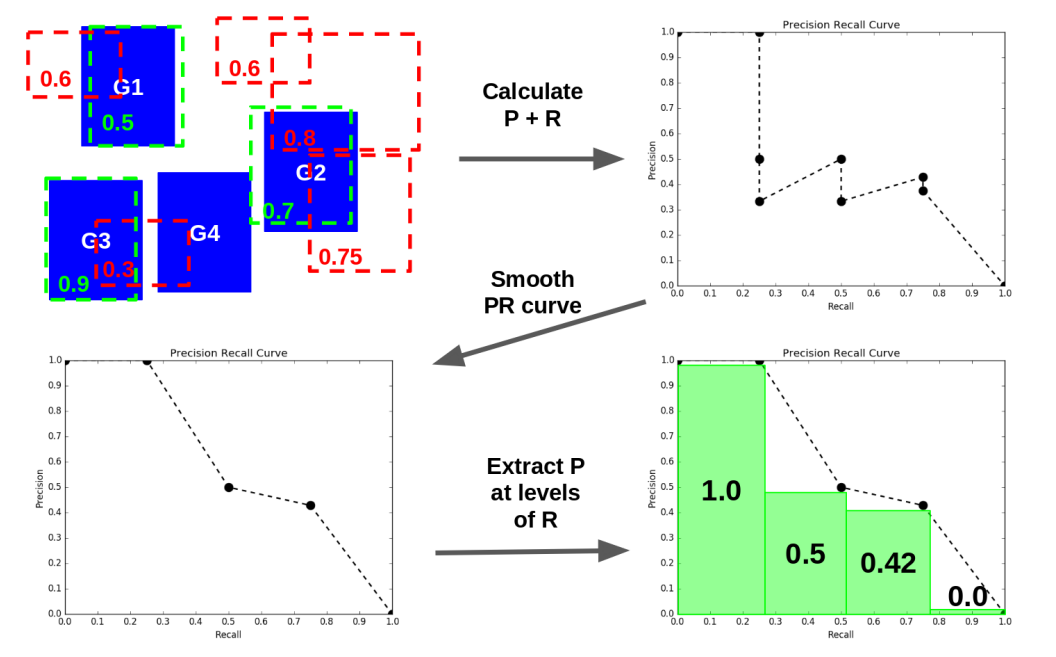


Figure 20: Process for extracting precision values from a PR curve for a given object class at a given threshold. The top-left shows the example scenario with ground-truth objects shown as blue boxes, true-positive detections shown as green BBoxes, and false-positive detections shown as red BBoxes. Numbers 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 all values of recall. Best viewed in colour.

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

where $\hat{\mathcal{B}}_i^f \cap \mathcal{B}_j^f$ is the intersection of the ground-truth and detection bounding boxes and $\hat{\mathcal{B}}_i^f \cap \mathcal{B}_j^f$ 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. 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 values at 101 levels of recall between zero and one is recorded. The maximum precision is used to avoid "wiggles" in the PR curve, resulting in a smoothed PR curve. Note that if a given level of recall is never reached, precision at that level of recall is treated as zero, so as to handle false negatives

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**
    $\mathcal{G}_*^f = \text{argmax}_{\mathcal{G}_i^f} IoU(\mathcal{G}_i^f, \mathcal{D}_j^f)$ **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

(FNs). This process on a simple scenario is outlined visually in Figure 20. This is 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.