Probabilistic Object Detection: Definition and Evaluation - supplementary material

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

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

1. *PDQ Qualitative Examples.*
2. *Evaluation of PDQ traits.*
3. *Traditional Measures Obscuring False Positives.*
4. *PDQ Evaluation Without Ground-Truth Segmentation Masks*
5. *Definition of mAP.*

1. PDQ Qualitative Examples

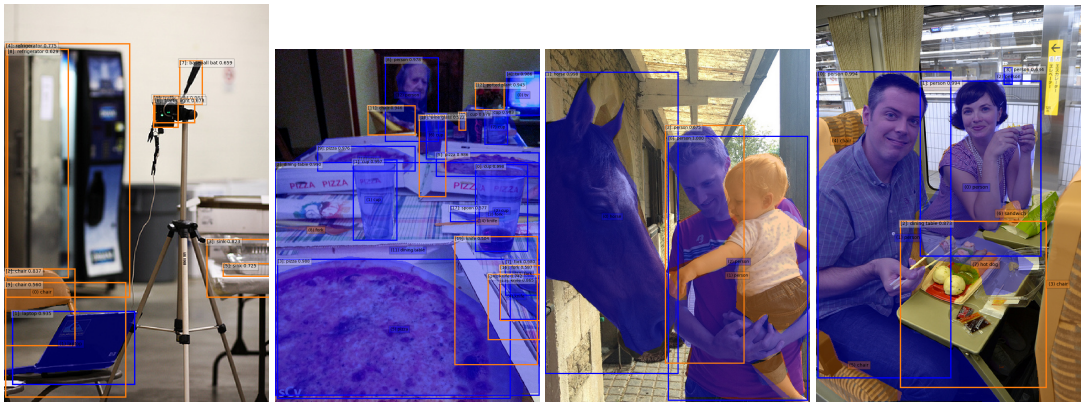
We provide qualitative results for detectors tested on COCO data in Section 7 of the main paper. Specifically, in this section we visualise results from SSD-300 [2], YOLOv3 [7], Faster RCNN with ResNext backbone and a feature pyramid network (FRCNN X+FPN) [3], and the probabilistic MC-Dropout SSD detector based on the work by Miller et al. [4, 5]. 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.

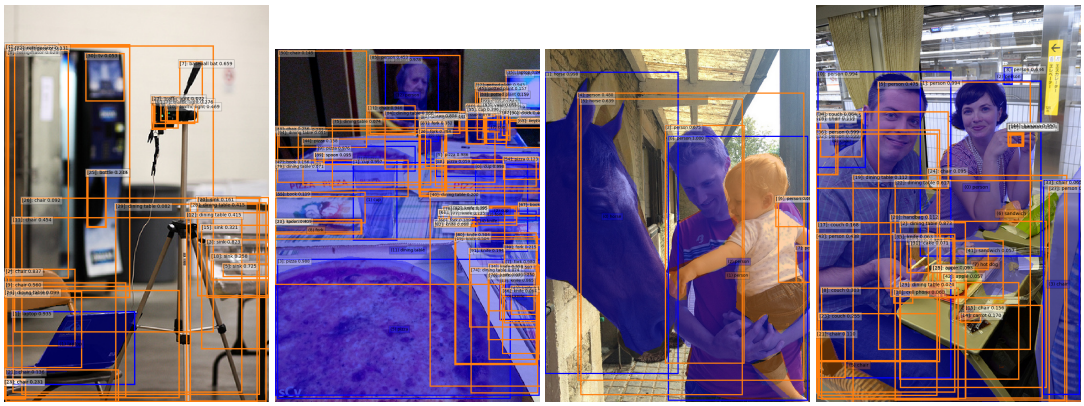
Using visualisations of this form enables us to qualitatively reinforce some of the findings from the main paper in the following three subsections. Firstly, we see again how the number of false positives under PDQ increases with lower label confidence thresholds (despite such detections getting higher 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.

1.1. 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 1. 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) 0.5 Label Threshold



(b) 0.05 Label Threshold

Figure 1: 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.

1.2. 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 2 and 3.

Figure 2 shows the effect that spatial uncertainty estimation has on the spatial quality of PDQ. In Figure 2a 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 2b, 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 2c, 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.

1.3. 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 3, 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.

1.4. 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 4 we visually compare YOLOv3 and FRCNN X+FPN results to qualitatively confirm this observation.

Examining Figure 4, 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.

2. Evaluation of PDQ Traits

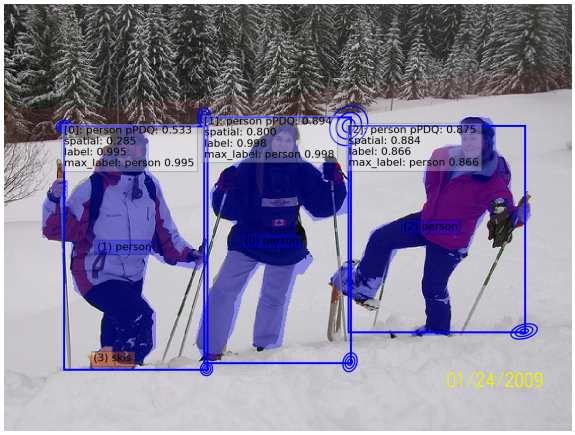
We demonstrate the characteristics of PDQ when compared with existing measures (mAP [1] and moLRP [6]) when responding to different types of imperfect detections, expanding upon what was covered in the main paper. Specifically, we examine the effect of spatial uncertainty, detection misalignment, label 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.

2.1. Spatial Uncertainty

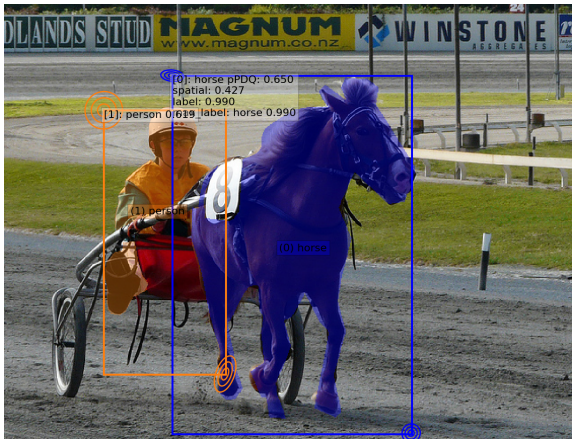
We examine the effect of spatial uncertainty on BBox and PBox detections respectively.

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 5

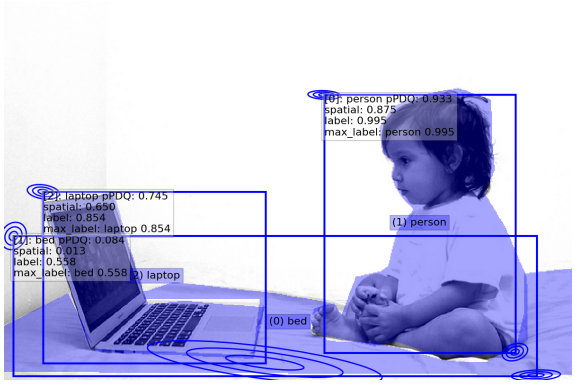
This experiment shows that PDQ is gradually reduced by decreasing spatial certainty, whereas mAP and moLRP



(a) general

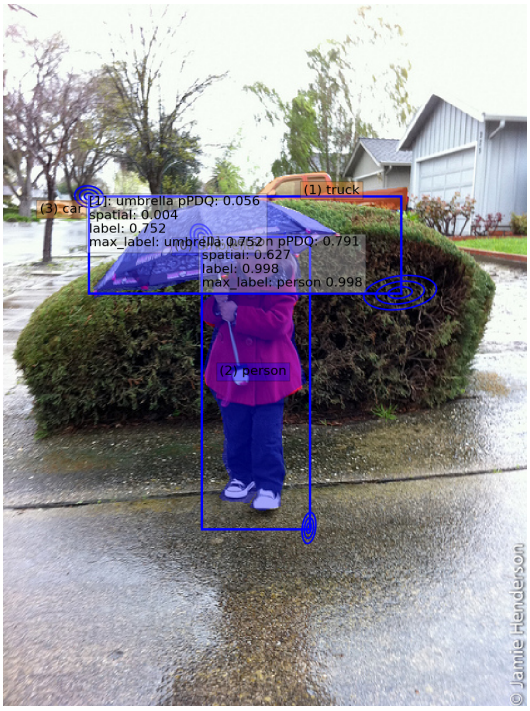


(b) over-confident



(c) under-confident

Figure 2: 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 3: Comparison of MC-Dropout SSD to SSD-300. SSD-300 is shown to be spatially over-confident leading to low scores despite tighter boxes.

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 6. We see a decline in PDQ with increased uncertainty demonstrating how PDQ penalises under-confidence.

2.2. Detection Misalignment

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

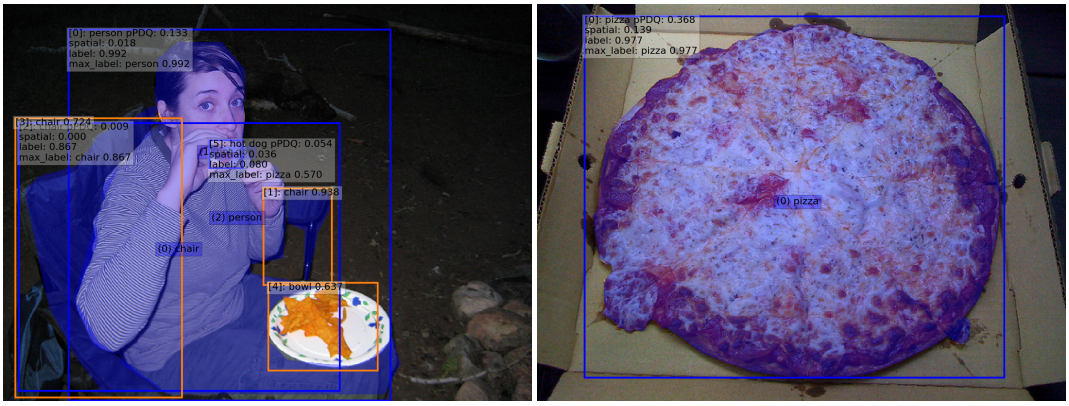
Translation Error We observe the effect of translation errors by shifting a 500 x 500 detection left and right past a 500 x 500 square object centred within a 2000 x 2000 image. This is tested both using BBoxes, and PBoxes with spherical Gaussian corners of varying reported variance (BBoxes equivalent to reported variance of zero). The results from this test are shown in Figure 7.

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.

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 8.

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

(a) YOLOv3



(b) FRCNN X+FPN

Figure 4: Visualisation of YOLOv3 detections compared with FRCNN X+FPN.

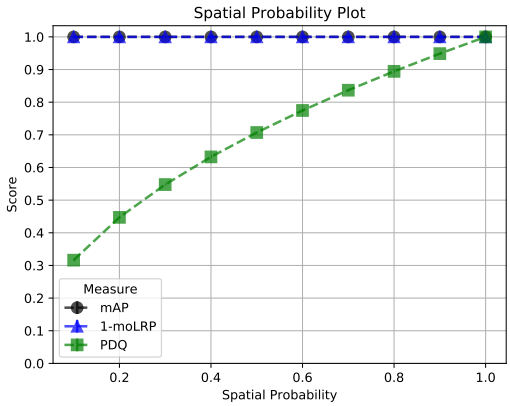


Figure 5: Evaluation of the effect of spatial probability on a perfectly aligned BBox. We see that unlike existing object detection measures, PDQ is effected by spatial probability changes.

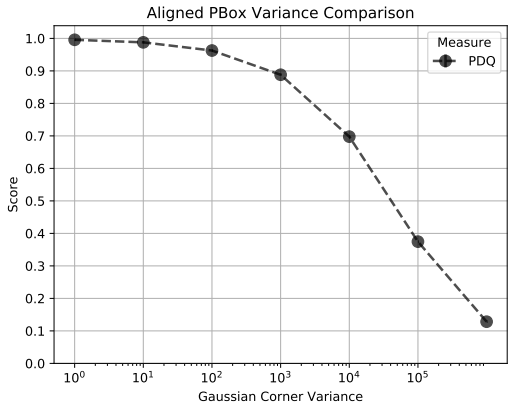


Figure 6: Plot showing the effect on PDQ of increasing variance, and by extension uncertainty, on perfectly aligned PBoxes. We see that for perfectly aligned detections, the score goes down the more uncertain the PBox detection is.

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

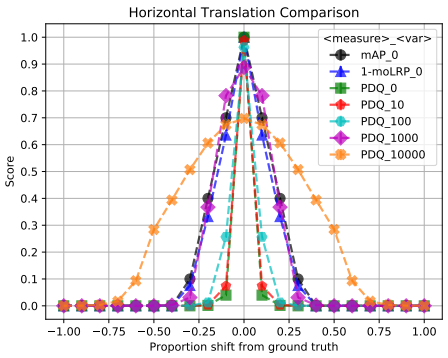


Figure 7: Evaluation of the effect of translation on mAP, moLRP, and PDQ scores. X-axis shows proportional shift of detection box either to the left (negative) or right (positive). Variance (*var*) refers to the variance of corner Gaussians of the PBox detections. BBox is used when *var* is zero. We 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.

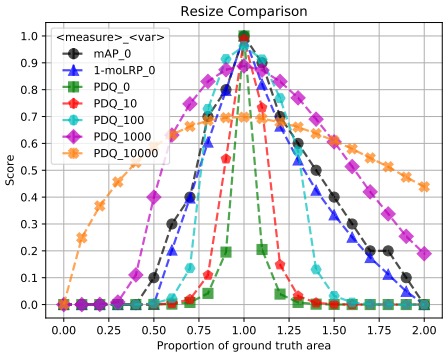


Figure 8: Evaluation of the effect of scaling on mAP, moLRP and PDQ scores. X-axis shows the proportional size of the detection to the ground-truth object. Variance (*var*) refers to the variance of corner Gaussians of PBox detections. BBox is used when *var* is zero. We see mAP and moLRP are lenient to detections with no uncertainty compared to PDQ and that PDQ is more lenient the more uncertain the detector is.

accurate estimates of spatial uncertainty.

2.3. 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[1] 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

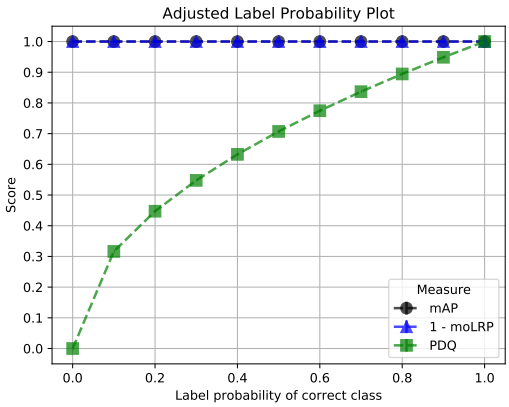


Figure 9: Effects of adjusting label confidences on mAP, moLRP, and PDQ when label probability for the correct class is adjusted using simulated detections on the COCO 2017 validation dataset. We see that existing measures are unaffected as long as the correct class is the class with highest probability in the label distribution. PDQ by comparison decreases with the label probability.

when using perfectly aligned BBox simulated detections are shown in Figure 9. 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.

2.4. 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 detection is spatially and semantically perfect. Results for mAP, moLRP, and PDQ for this scenario are visualised in Figure 10. 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 11.

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

2.5. False Detections

We provide the results of a simplified scenario to show that, excluding edge cases that will be discussed in Section 3, mAP, moLRP, and PDQ respond almost the same

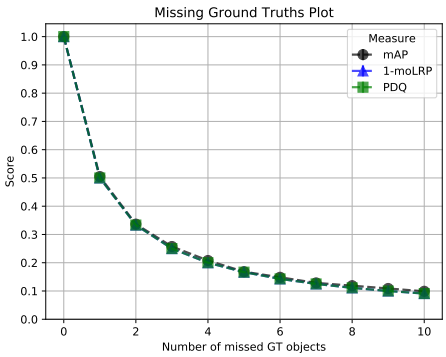


Figure 10: Evaluation of the effect of missing ground-truth objects on evaluation scores in simplified scenario. We observe that all measures respond the same to missed ground-truth objects.

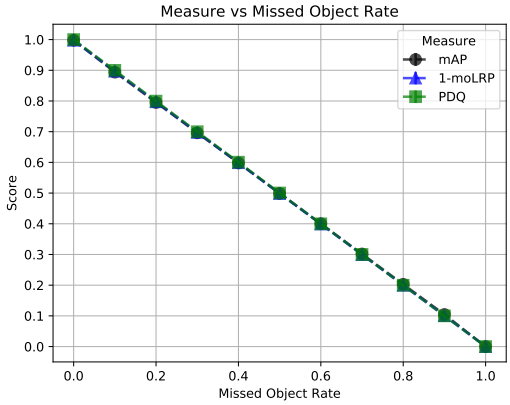


Figure 11: Evaluation of the effect of missing an increased proportion of ground-truth objects on COCO 2017 validation dataset images. We see the response from all measures is the same.

to false positive 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 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 3. We plot the resultant mAP, moLRP, and PDQ scores in Figure 12.

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.

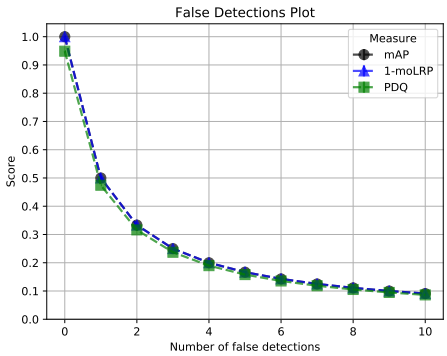


Figure 12: Evaluation of the effect of false detections on evaluation scores. We observe that generally, all measures respond the same to false detections.

3. 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.

3.1. 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 13. We observed that PDQ and moLRP penalised the additional FP detections, whereas mAP gave 100% accuracy at all times.

This edge case breaks mAP due to how the PR curve for this scenario is generated and utilised. As is explained later in Section 5, 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 5). This is why we attain the result for mAP shown in Figure 13.

This is not a new problem with mAP, and such be-

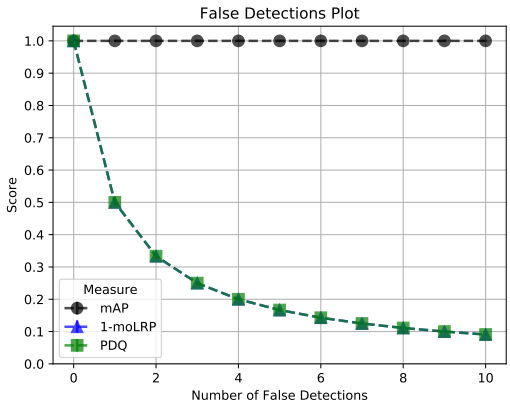


Figure 13: Duplications test results showing mAP, moLRP, and PDQ values when perfect duplicate FP detections are added in a one-object scenario. The TP detection is evaluated before the FPs, causing subsequent FPs to be ignored by mAP. PDQ and moLRP respond as expected, penalising FPs.

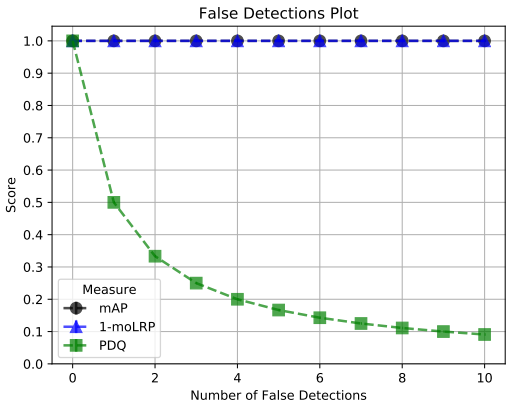


Figure 14: False detection test where all FP detections have slightly lower label confidence than the TP detection (90% Vs 100%). Both mAP and moLRP are shown to treat this as perfect detection output.

haviour caused by relative ranking has been outlined in past works [7]. 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.

3.2. 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 14.

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 3.1). The moLRP result, on the other hand, has changed due to the optimal thresholding done as part of the algorithm [6]. 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.

3.3. Duplicate Detections on COCO Data

Scenario 3 extends scenario 1 (Section 3.1) from a single image to examine duplicate detections on the COCO 2017 validation data [1]. 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 5 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 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 15. 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 5 and the ordering of our detections. As described in Section 5, 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

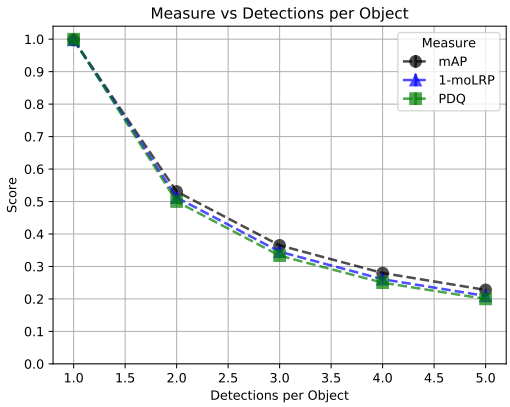


Figure 15: Test results on COCO 2017 validation data comparing scores when a number of perfectly aligned duplicate FP detections are added. Each duplicate FP is ordered directly after their corresponding TP detection. Due to smoothing of the PR curve, calculated precision becomes higher than expected for some classes at different levels of recall, causing mAP to be higher than expected. Other measures remain relatively unaffected.

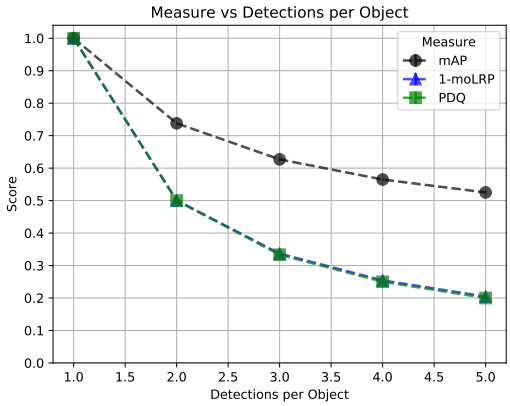


Figure 16: Duplication test results such as done for Figure 15 on subset of 100 images from COCO 2017 validation data. This shows heightened mAP scores from those shown Figure 15 demonstrating increased unintuitive behaviour from mAP as the dataset gets smaller.

mAP's unusual performance. As we see in the following scenario, this is a problem which increases in severity with small datasets.

3.4. 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 3.3). 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 16.

Here we see that the mAP scores are far higher at than

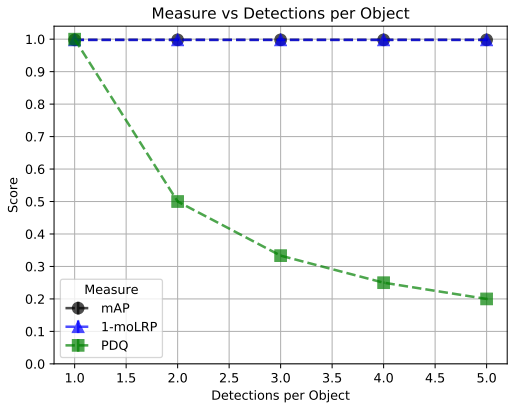


Figure 17: Duplication test results on COCO 2017 validation data where all FP duplicate detections have lower label confidence than the TP detection (90% Vs 100%). Unlike PDQ, both mAP and moLRP are shown to treat this as perfect detection output.

expected for each level of detections per object, an exaggeration of the effect in Section 3.3. 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 3.3) 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 3.1. 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.

3.5. Duplicate Detections with Lower Confidence on COCO Data

Reinforcing our findings in Sections 3.3 and 3.4, 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 17. As expected, PDQ continues to treat the false positives as significant whilst mAP and moLRP both consider the detection output as perfect.

3.6. Summary

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

4. PDQ Evaluation Without Ground-Truth Segmentation Masks

In the main paper we allude to the fact that PDQ can be used without pixel-perfect segmentation ground-truth. Here we reinforce this point by modifying the experiments from the main paper (Section 7) using the COCO bounding-box annotations as ground-truth rather than the segmentation mask annotations within our PDQ analysis. This enables us to observe the difference in PDQ scores between the two types of ground-truth. When using bounding-box annotations, all pixels within the bounding-box are considered part of the object's segmentation mask $\mathcal{S}_j^f$. Results from the bounding-box experiments alongside the average differences in scores between the original results from Table 1 of the main paper are summarized in Table 1.

These tests show that, although unideal, using bounding-boxes as ground-truth segmentation masks does not drastically change PDQ. The change yields only an absolute decrease in PDQ scores of 1.1% and does not change PDQ rank order significantly. We can see the change in PDQ is caused by the need to detect previously irrelevant pixels, decreasing foreground quality by an average of 27.7%. We see that the use of the bounding-box ground-truth also increases background quality by an average of 11.8% and this can be attributed to the higher number of foreground pixels used to scale background loss in Equation 3 of the main paper. Overall we can conclude that while there will be some changes to PDQ score when using the unideal case of bounding-boxes as ground-truth annotation masks, these changes are not drastic and PDQ is suitable for use in cases where only such annotations are available.

5. Definition of mAP

For the sake of completeness and to aid in understanding the behaviour shown in Section 3, here we define mean average precision (mAP) as used by the COCO detection challenge [1]. 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

$$IoU(\hat{\mathcal{B}}_i^f, \mathcal{B}_j^f) = \frac{area(\hat{\mathcal{B}}_i^f \cap \mathcal{B}_j^f)}{area(\hat{\mathcal{B}}_i^f \cup \mathcal{B}_j^f)}, \tag{1}$$

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.

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 1: mAP Detection Assignment

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

Table 1: PDQ-based Evaluation of Probabilistic and Non-Probabilistic Object Detectors using bounding-box ground-truth. 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. Final row shows average difference between bounding-box ground-truth and segment ground-truth with (+) and (-) indicating score increases and decreases when using bounding-box ground-truth respsectively.

Approach (τ)	mAP (%)	mLRP (%)	PDQ (%)	pPDQ (%)	Sp (%)	Lbl (%)	FG (%)	BG (%)	TP	FP	FN
probFRCNN (0.5)	35.5	32.2	**27.0**	**52.6**	**40.8**	90.2	54.4	**77.8**	23,809	9,641	12,972
MC-Dropout SSD (0.5) [4]	15.8	15.6	12.4	43.9	36.2	73.3	**55.9**	67.7	10,892	**1,783**	25,889
MC-Dropout SSD (0.05) [4]	19.5	16.6	1.3	26.1	23.2	33.3	50.1	49.3	27,797	458,120	8,984
SSD-300 (0.5) [2]	15.0	14.3	2.7	11.4	3.9	79.5	27.7	30.1	9,768	3,977	27,013
SSD-300 (0.05) [2]	19.3	16.0	0.4	5.8	2.2	41.8	19.9	29.1	23,318	322,710	13,463
YOLOv3 (0.5) [7]	29.7	30.8	4.8	10.0	2.8	**95.7**	21.5	34.0	20,145	4,973	16,636
YOLOv3 (0.05) [7]	30.1	27.7	2.7	8.1	2.2	93.0	18.8	33.9	27,499	46,022	9,282
FRCNN R (0.5) [8]	32.8	29.1	7.2	17.3	7.8	88.3	31.6	39.4	22,505	17,469	14,276
FRCNN R (0.05) [8]	34.3	29.1	3.2	15.3	7.1	78.2	30.0	38.5	26,235	89,987	10,546
FRCNN R+FPN (0.5) [3]	34.6	31.2	7.9	16.7	7.1	86.2	21.8	54.3	23,827	13,416	12,954
FRCNN R+FPN (0.05) [3]	37.0	30.4	2.8	2.8	6.5	69.8	20.4	53.2	30,942	121,895	5,839
FRCNN X+FPN (0.5) [3]	37.4	**32.7**	8.0	17.2	7.4	87.9	21.8	55.6	25,937	19,030	10,844
FRCNN X+FPN (0.05) [3]	**39.0**	32.1	2.9	15.1	6.9	74.6	20.6	54.5	**31,578**	128,353	**5,203**
Avg. Diff. from Original	0	0	-1.1	-6.5	-5.5	-0.3	-27.7	+11.8	+1873	-1873	-1873

above the highest reached are 0, to handle false negatives (FNs). This process on a simple scenario is outlined visually in Figure 18. 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.

References

[1] 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. 3, 6, 8, 10

[2] 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. 1, 11

[3] 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]. 1, 11

[4] 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. 1, 11

[5] 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. 1

[6] 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. 3, 8

[7] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, Apr. 2018. 1, 8, 11

[8] 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. 11

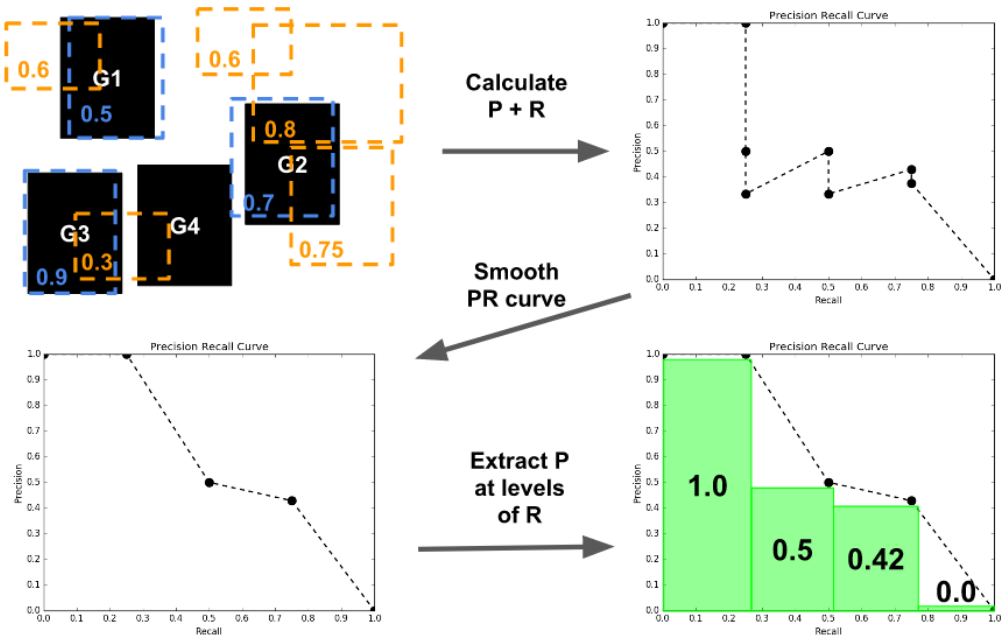


Figure 18: Process for extracting precision values from a PR curve for a given object class at a given threshold. The top-left shows the example scenario with ground-truth objects shown as black boxes, true-positive detections shown as light blue BBoxes, and false-positive detections shown as orange 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.