Towards Building Self-Aware Object Detectors via Reliable Uncertainty Quantification and Calibration

Kemal Oksuz, Tom Joy, Puneet K. Dokania
Five AI Ltd., United Kingdom
{kemal.oksuz, tom.joy, puneet.dokania}@five.ai

Abstract

*The current approach for testing the robustness of object detectors suffers from serious deficiencies such as improper methods of performing out-of-distribution detection and using calibration metrics which do not consider both localisation and classification quality. In this work, we address these issues, and introduce the **S**elf **A**ware **O**bject **D**etection (SAOD) task, a unified testing framework which respects and adheres to the challenges that object detectors face in safety-critical environments such as autonomous driving. Specifically, the SAOD task requires an object detector to be: robust to domain shift; obtain reliable uncertainty estimates for the entire scene; and provide calibrated confidence scores for the detections. We extensively use our framework, which introduces novel metrics and large scale test datasets, to test numerous object detectors in two different use-cases, allowing us to highlight critical insights into their robustness performance. Finally, we introduce a simple baseline for the SAOD task, enabling researchers to benchmark future proposed methods and move towards robust object detectors which are fit for purpose. Code is available at: https://github.com/fiveai/saod.*

1. Introduction

The safe and reliable usage of object detectors in safety critical systems such as autonomous driving [10,65,73], depends not only on its accuracy, but also critically on other robustness aspects which are often only considered in addition or not all. These aspects represent its ability to be robust to domain shift, obtain well-calibrated predictions and yield reliable uncertainty estimates at the image-level, enabling it to flag the scene for human intervention instead of making unreliable predictions. Consequently, the development of object detectors for safety critical systems requires a thorough evaluation framework which also accounts for these robustness aspects, a feature lacking in current evaluation methodologies.

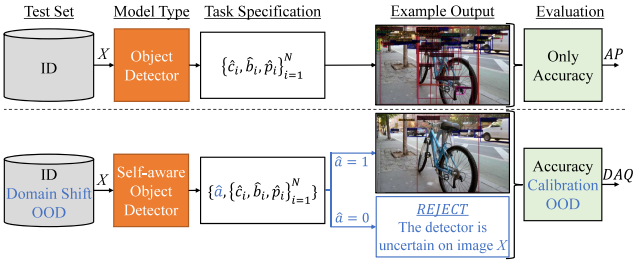


Figure 1. (Top) The vanilla object detection task vs. (Bottom) the self-aware object detection (SAOD) task. Different from the vanilla approach; the SAOD task requires the detector to: predict $\hat{a} \in \{0, 1\}$ representing whether the image X is accepted or not for further processing; yield accurate and calibrated detections; and be robust to domain shift. Accordingly, for SAOD we evaluate on ID, domain-shift and OOD data using our novel DAQ measure. Here, $\{\hat{c}_i, \hat{b}_i, \hat{p}_i\}^N$ are the predicted set of detections.

Whilst object detectors are able to obtain uncertainty at the *detection-level*, they do not naturally produce uncertainty at the *image-level*. This has lead researchers to often evaluate uncertainty by performing out-of-distribution (OOD) detection at the detection-level [13,21], which cannot be clearly defined. Thereby creating a misunderstanding between OOD and in-distribution (ID) data. This leads to an improper evaluation, as defining OOD at the detection level is non-trivial due to the presence of known-unknowns or background objects [12]. Furthermore, the test sets for OOD in such evaluations are small, typically containing around 1-2K images [13,21].

Moreover, as there is no direct access to the labels of the test sets and the evaluation servers only report accuracy [18, 43], researchers have no choice but to use small validation sets as testing sets to report robustness performance, such as calibration and performance under domain shift. As a result, either the training set [11,59]; the validation set [13, 21]; or a subset of the validation set [36] is employed for cross-validation, leading to an unideal usage of the dataset splits and a poor choice of the hyper-parameters.

Finally, prior work typically focuses on only one of: calibration [35, 36]; OOD detection [13]; domain-shift [45,

68, 71, 72]; or leveraging uncertainty to improve accuracy [5, 9, 20, 23, 70], with no prior work taking a holistic approach by evaluating all of them. Specifically for calibration, previous studies either consider classification calibration [35], or localisation calibration [36], completely disregarding the fact that object detection is a joint problem.

In this paper, we address the critical need for a robust testing framework which evaluates object detectors thoroughly, thus alleviating the aforementioned deficiencies. To do this, we introduce the Self-aware Object Detection (SAOD) task, which considers not only accuracy, but also calibration using our novel Localisation-aware Expected Calibration Error (LaECE) as well as the reliability of image-level uncertainties. Furthermore, the introduction of LaECE addresses a critical gap in the literature as it respects both *classification* and *localisation* quality, a feature ignored in previous methods [35, 36]. Moreover, the SAOD task requires an object detector to either perform reliably or reject images outside of its training domain.

We illustrate the SAOD task in Fig. 1, which not only evaluates the accuracy, but also the calibration and performance under OOD or domain-shifted data. We can also see the functionality to reject an image, and to only produce detections which have a high confidence; unlike for a standard detector which has to accept every image and produce detections. To summarise, our main contributions are:

- We introduce the SAOD task, which evaluates: accuracy; robustness to domain shift; ability to accept of reject an image; and calibration in a unified manner. We further construct large datasets totaling 155K images and provide a simple baseline for future researchers to benchmark against.
- We explore how to obtain image-level uncertainties from any object detector, enabling it to reject the entire scene for the SAOD task. Through our investigations, we discover that object detectors are inherently strong OOD detectors and provide reliable uncertainties.
- Finally, we define the LaECE as a novel calibration measure for object detectors in SAOD, which requires the confidence of a detector to represent both its *classification* as well as its *localisation* quality.

2. Notations and Preliminaries

Object Detection Given that the set of M objects in an image X is represented by $\{b_i, c_i\}^M$ where $b_i \in \mathbb{R}^4$ is a bounding box and $c_i \in \{1, \dots, K\}$ its class; the goal of an object detector is to predict the bounding boxes and the class labels for the objects in X, $f(X) = \{\hat{c}_i, \hat{b}_i, \hat{p}_i\}^N$, where $\hat{c}_i, \hat{b}_i, \hat{p}_i$ represent the class, bounding box and confidence score of the ith detection respectively and N is the number of predictions. Conventionally, the detections are obtained in two steps, $f(X) = (h \circ g)(X)$ [6, 42, 61, 66]: where $g(X) = \{\hat{b}_i^{raw}, \hat{p}_i^{raw}\}^{N^{raw}}$ is a deep neural network predicting raw detections with bounding boxes $\hat{b}_i^{raw}$ and predicted class distribution $\hat{p}_i^{raw}$. Given these raw-detections $h(\cdot)$ applies post-processing to obtain the final detections[1]. In general, $h(\cdot)$ comprises removing the detections predicted as background; Non-Maximum-Suppression (NMS) to discard duplicates; and keeping useful detections, normally achieved through top-k survival, where in practice $k = 100$ for COCO dataset [43].

Evaluating the Performance of Object Detectors Average Precision (AP) [15, 18, 43], or the area under the precision-recall (PR) curve, has been the common performance measure of object detection. Though widely accepted, AP suffers from the following three main drawbacks [58]. First, it only validates true-positives (TPs) using a localisation quality threshold, completely disregarding the continuous nature of localisation. Second, as an area-under-curve (AUC) measure, AP is difficult to interpret, as PR curves with different characteristics can yield the same value. Also, AP rewards a detector that produces a large number of low scoring detections than actual objects in the image, which becomes a significant issue when relying on top-k survival as shown in Fig. 1. App. D includes details.

Alternatively, the recently proposed Localisation-Recall-Precision Error (LRP) [53, 58] combines the number of TP, false-positive (FP), false-negative (FN), denoted by $\mathrm{N_{TP}}$, $\mathrm{N_{FP}}$, $\mathrm{N_{FN}}$, respectively, as well as the Intersection-over-Union (IoU) of TPs with the objects that they match with:

$$\frac{1}{\mathrm{N_{FP}} + \mathrm{N_{FN}} + \mathrm{N_{TP}}} \left(\mathrm{N_{FP}} + \mathrm{N_{FN}} + \sum_{\psi(i)>0} (1 - \mathrm{lq}(i)) \right) \quad (1)$$

where $\mathrm{lq}(i) = \frac{\mathrm{IoU}(\hat{b}_i, b_{\psi(i)}) - \tau}{1 - \tau}$ is the localisation quality with τ being the TP assignment threshold, $\psi(i)$ is the index of the object that a TP i matches to; else i is a FP and $\psi(i) = -1$. LRP can be decomposed into components providing insights on: the localisation quality; the precision; and the recall error. Besides, low-scoring detections are demoted by the term $\mathrm{N_{FP}}$ in Eq. (1). Thus, LRP arguably alleviates the aforementioned drawbacks of AP.

3. An Overview to the SAOD Task

For object detectors to be deployed in safety critical systems it is imperative that they perform in a robust manner. Specifically, we would expect the detector to be aware of situations when the scene differs substantially from the training domain and to include the functionality to flag the scene for human intervention. Moreover, we also expect that the confidence of the detections matches the performance, referred to as calibration. With these expectations in mind, we characterise the crucial elements needed to evaluate and

[1] for probabilistic detectors [5, 19–21, 23], $\hat{b}_i^{raw}$ follows a probability distribution mostly of the form $g(X) = \{\mathcal{N}(\mu_i, \Sigma_i), \hat{p}_i^{raw}\}^{N^{raw}}$, where Σ_i is either a diagonal [5, 23] or full covariance matrix [20]

perform the SAOD task. Specifically, the SAOD task requires an object detector to:

- Have the functionality to reject a scene based on its image-level uncertainties through a binary indicator variable $\hat{a} \in \{0, 1\}$.
- Produce detection-level confidences that are calibrated in terms of classification *and* localisation.
- Be robust to domain-shift.

For brevity, and to enable future researchers to adopt the SAOD framework, the explicit practical specification for Self-aware Object Detectors (SAODETs) is

$$f_A(X) = \{\hat{a}, \{\hat{c}_i, \hat{b}_i, \hat{p}_i\}^N\}, \tag{2}$$

where $\hat{a} \in \{0, 1\}$ implies if the image should be accepted or rejected and that the predicted confidences $\hat{p}_i$ are calibrated.

Evaluation Datasets As the SAOD emulates challenging real-life situations, the evaluation needs to be performed using large-scale test datasets. Unlike previous approaches on OOD detection using around 1-2K OOD images [13,21] for testing or calibration methods [36] relying on 2.5K ID test images, our test set totals to 155K individual images for each of our two use-cases when combining ID and OOD data. Specifically, we construct two test datasets, where each $\mathcal{D}_{\text{Test}}$ in our case is the *union* of the following datasets:

- $\mathcal{D}_{\text{ID}}$ ($45K$ Images): ID dataset with images containing the same foreground objects as were present in $\mathcal{D}_{\text{Train}}$.
- $\mathcal{T}(\mathcal{D}_{\text{ID}})$ ($3 \times 45K$ Images): domain-shift dataset obtained by applying transformations to the images from $\mathcal{D}_{\text{ID}}$, which preserve the semantics of the image.
- $\mathcal{D}_{\text{OOD}}$ ($110K$ Images): OOD dataset with images that do not contain any foreground object from $\mathcal{D}_{\text{ID}}$. These images tend to include objects not present in $\mathcal{D}_{\text{Train}}$.

We present exact splits in Tab. 1 for object detection in General and Autonomous Vehicles (AV) use-cases (refer App. A for further details). Collected from a different dataset, our $\mathcal{D}_{\text{ID}}$ differs from $\mathcal{D}_{\text{Train}}$, but is still semantically similar; which is reflective of a challenging real-word scenario, as domains change over time and scenes differ in terms of appearance. For $\mathcal{T}(\mathcal{D}_{\text{ID}})$, we apply ImageNet-C style corruptions [25] to $\mathcal{D}_{\text{ID}}$, where for each image we randomly choose one of 15 corruption types (fog, blur, noise, etc.) at severity levels 1, 3 and 5 as is common in practice [21]. Then, we expect that for a given input $X \in \mathcal{D}_{\text{Test}}$, a SAODET makes the following decisions:

- if $X \in \mathcal{D}_{\text{ID}} \cup \mathcal{T}(\mathcal{D}_{\text{ID}})$ for corruption severities 1 and 3, 'accept' the input and provide *accurate and calibrated* detections. Penalize a rejection.
- if $X \in \mathcal{T}(\mathcal{D}_{\text{ID}})$ at corruption severity 5, provide the choice to 'accept' and evaluate but do not *not* penalize a 'rejection' as the transformed images might not contain enough cues to perform object detection reliably.
- if $X \in \mathcal{D}_{\text{OOD}}$, 'reject' the image and provide *no* detections as, by design, the predictions would be wrong. An 'accept' should be penalized in this case.

Table 1. Our dataset splits for SAOD. We design test sets for COCO [43] and nuImages [4] as ID data (train & val). We exploit Objects365 [63] and BDD100K [73] for $\mathcal{D}_{\text{ID}}$ and $\mathcal{T}(\mathcal{D}_{\text{ID}})$, and use Objects365, iNaturalist [27] and SVHN [50] for $\mathcal{D}_{\text{OOD}}$.

Dataset	$\mathcal{D}_{\text{Train}}$	$\mathcal{D}_{\text{Val}}$	$\mathcal{D}_{\text{Test}}$		
			$\mathcal{D}_{\text{ID}}$	$\mathcal{T}(\mathcal{D}_{\text{ID}})$	$\mathcal{D}_{\text{OOD}}$
SAOD-Gen	COCO$^{(train)}$	COCO$^{(val)}$	Obj45K	Obj45K-C	SiNObj110K-OOD
SAOD-AV	nuImages$^{(train)}$	nuImages$^{(val)}$	BDD45K	BDD45K-C	SiNObj110K-OOD

Models In terms of evaluating SAOD on common object detectors, it would prove useful at this point to introduce the models used in our investigation. We mainly exploit a diverse set of four object detectors:

1. Faster R-CNN (F-RCNN) [61] is a two-stage detector with a softmax classifier
2. Rank & Sort R-CNN (RS-RCNN) [55] is another two-stage detector but with a ranking-based loss function and sigmoid classifiers
3. Adaptive Training Sample Selection (ATSS) [77] is a common one-stage baseline with sigmoid classifiers
4. Deformable DETR (D-DETR) [79] is a transformer-based model, again using sigmoid classifiers

We also evaluate two probabilistic detectors with a diagonal covariance matrix minimizing the negative log likelihood [23] (NLL-RCNN) or energy score [21] (ES-RCNN), allowing us to obtain uncertainty estimates for localisation. Please see App. B for the training details of the methods as well as their accuracy on $\mathcal{D}_{\text{Val}}$, $\mathcal{T}(\mathcal{D}_{\text{Val}})$, $\mathcal{D}_{\text{ID}}$ and $\mathcal{T}(\mathcal{D}_{\text{ID}})$.

As we have now outlined clear requirements for a SAODET, it is natural to ask how well the aforementioned object detectors perform under these requirements. We will extensively investigate this by first introducing a simple method to extract image-level uncertainty enabling the acceptance or rejection of an image in Sec. 4; evaluate the calibration and provide methods to calibrate such detectors in Sec. 5; before finally providing a complete analysis of them using the SAOD framework in Sec. 6.

4. Obtaining Image-level Uncertainty

As there is no clear distinction between background and an OOD object unless each pixel in $\mathcal{D}_{\text{Train}}$ is labelled [12], evaluating uncertainties of detectors is nontrivial at detection-level. Thus, different from prior work [13, 21] conducting OOD detection at detection-level, we evaluate the uncertainties on image-level OOD detection task. Thereby aligning the evaluation and the definition of an OOD image. Please see App. C.1 for further discussion.

Practically, one method to accept or reject an image is to obtain an estimate of uncertainty at the image-level through a function $\mathcal{G} : \mathcal{X} \rightarrow \mathbb{R}$ and a threshold $\bar{u} \in \mathbb{R}$, where the image is accepted if $\mathcal{G}(X) < \bar{u}$ and $\hat{a} = 1$; and rejected vice-versa. We take this approach when constructing our

Table 2. AUROC scores (in %) for image-level uncertainties when aggregating through different methods, where we use the uncertainty score of $1 - \hat{p}_i$ for the detections. Here, top-m refers to the average of the lowest m uncertainties for the detections. As we can see, using the most certain detections performs better. Bold and underline are best and second best respectively.

Dataset	Detector	sum	mean	top-5	top-3	top-2	min
SAOD-Gen	F-RCNN	20.9	84.1	93.4	<u>94.1</u>	**94.4**	93.8
	RS-RCNN	85.8	85.8	94.3	**94.8**	**94.8**	93.5
	ATSS	66.2	86.3	93.8	**94.2**	<u>94.0</u>	92.6
	D-DETR	85.2	85.2	94.4	**94.7**	<u>94.6</u>	93.3
SAOD-AV	F-RCNN	27.1	84.1	96.4	<u>97.3</u>	**97.4**	96.0
	ATSS	18.8	92.2	**97.7**	<u>97.6</u>	97.3	95.7

Table 3. AUROC scores (in %) of different detection-level uncertainty estimates. Classification-based uncertainties perform better compared to localization and $1 - \hat{p}_i$ performs generally the best.

Dataset	Detector	Classification			Localisation		
		$\mathrm{H}(\hat{p}_i^{raw})$	DS	$1 - \hat{p}_i$	$\lvert\Sigma\rvert$	$\mathrm{tr}(\Sigma)$	$\mathrm{H}(\Sigma)$
SAOD Gen	F-RCNN	92.6	89.7	**94.1**	N/A	N/A	N/A
	RS-RCNN	93.7	30.0	**94.8**	N/A	N/A	N/A
	ATSS	**94.3**	36.9	94.2	N/A	N/A	N/A
	D-DETR	93.9	73.8	**94.4**	N/A	N/A	N/A
	NLL-RCNN	92.4	89.0	**94.1**	87.6	87.5	87.7
	ES-RCNN	92.8	89.9	**94.1**	85.0	85.2	86.4
SAOD AV	F-RCNN	**97.3**	96.0	**97.3**	N/A	N/A	N/A
	ATSS	97.2	97.1	**97.6**	N/A	N/A	N/A

baseline and now specifically outline the method to do so.

Obtaining Image-level Uncertainties This can be achieved through aggregating the detection-level uncertainties. We hypothesise that there is implicitly enough uncertainty information in the detections to produce image-level uncertainty, they just need to be extracted and aggregated in an appropriate way. In terms of the extraction, we can obtain detection level uncertain through: the uncertainty score ($1 - \hat{p}_i$); the entropy of the predictive classification distribution of the raw detections ($\mathrm{H}(\hat{p}_i^{raw})$); and Dempster-Shafer [14,62] (DS). In addition, for probabilistic detectors, we can extract uncertainty from Σ by taking the: determinant, trace, or entropy of the multivariate normal distribution [49]. In terms of the aggregation strategy, given the uncertainties for the detections after top-k survival, we let $\mathcal{G}$ either take their: sum, mean, minimum, or their mean of the m smallest uncertainty values, i.e. the most certain top-m detections. For further details, please see App. C. Whilst these strategies are simple, as we will now show, they provide a suitable method to obtain image-level uncertainty, enabling effective performance on OOD detection, a common task for evaluating uncertainty quantification.

To do this, we evaluate the Area-under ROC Curve (AUROC) score between the uncertainties of the data from $\mathcal{D}_{\text{ID}}$ and $\mathcal{D}_{\text{OOD}}$ and display the results in Tab. 2; which

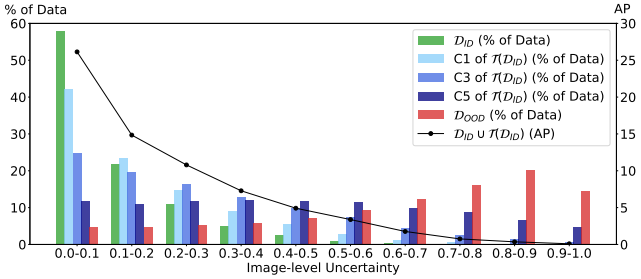


Figure 2. The distribution of image-level uncertainties obtained from F-RCNN (SAOD-Gen) on $\mathcal{D}_{\text{ID}}$, different severities 1, 3, 5 (C1, C3, C5) of $\mathcal{T}(\mathcal{D}_{\text{ID}})$ and $\mathcal{D}_{\text{OOD}}$ vs. the accuracy in COCO-Style AP in % (AP in short). App. C includes more examples.

Table 4. AUROC scores (in %) on subsets of $\mathcal{D}_{\text{OOD}}$. In all cases, near-OOD (Obj365) has a lower AUROC than far-OOD (SVHN).

Dataset	Detector	near to far OOD			all OOD
		Obj365	iNat	SVHN	
SAOD Gen	F-RCNN	83.7	97.6	**99.8**	94.1
	RS-RCNN	85.6	97.8	**99.8**	94.8
	ATSS	84.5	97.4	99.5	94.2
	D-DETR	85.7	**98.1**	99.4	94.4
SAOD AV	F-RCNN	**95.2**	97.4	98.8	97.3
	ATSS	95.0	97.3	99.7	**97.7**

shows that high AUROC scores are obtained when $\mathcal{G}$ is formed by considering up to the mean(top-5) detections, with the mean(top-3) aggregation strategy of $1-\hat{p}_i$ performs the best. This highlights that the detections with lowest uncertainty in each image provide useful information to reliably estimate image-level uncertainty. We believe the poor performance for mean and sum stem from the fact that there are typically too many noisy detections (up to $k = 100$) for only a few objects in the image. We further provide assurance that $1 - \hat{p}_i$ is the most appropriate method to extract detection-level uncertainty in Tab. 3, where we can see that $1-\hat{p}_i$ obtains higher AUROC scores compared to $\mathrm{H}(\hat{p}_i^{raw})$ and DS. We also note that classification uncertainties (except DS) perform consistently better than localisation ones for probabilistic detectors. We believe one of the reasons for that is the classifier is trained using both the proposals matching and not matching with any object, preventing the detector from becoming over-confident everywhere.

How Reliable are these Image-level Uncertainties? Though the aforementioned results show that the image-level uncertainties are effective, we now see how reliable these uncertainties are in practice. For this, we first evaluate the detectors on different subsets of our SiNObj110K OOD set. Tab. 4 shows that for all detectors, the AUROC score is lower for near-OOD subset (Obj365) than for far-OOD (iNat and SVHN) and is consistently very high for far-OOD subsets (up to 99.8 on SVHN).

We then consider the uncertainties of $\mathcal{D}_{\text{ID}}$, $\mathcal{T}(\mathcal{D}_{\text{ID}})$ and $\mathcal{D}_{\text{OOD}}$ by plotting histograms of the image-level uncertain-

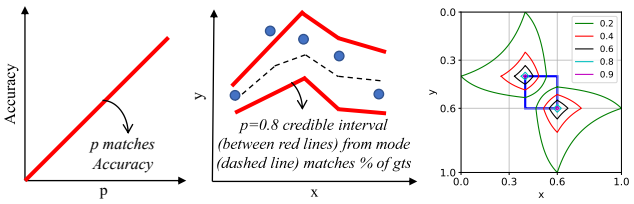
ties in 10 equally-spaced bins in the range of $[0, 1]$. In Fig. 2 we see that the uncertainties from $\mathcal{D}_{\text{ID}}$ have a significant amount of mass in the smaller bins and vice versa for $\mathcal{D}_{\text{OOD}}$, moreover the uncertainties get larger as the severity of corruption increases. We also display AP (black line), where it can be clearly seen that as the uncertainty increases AP decreases, implying that the uncertainty reflects the performance of the detector. Thereby suggesting that *the image-level uncertainties are reliable and effective*. As already pointed out, this conclusion is not necessarily very surprising, since the classifiers of object detectors are generally trained not only by proposals matching the objects but also by a very large number of proposals not matching with any object, which can be ~ 1000 times more [57]. This composition of training data prevents the classifier from becoming drastically over-confident for unseen data, enabling the detector to yield reliable uncertainties.

Thresholding Image-level Uncertainties For our SAOD baseline, we can obtain an appropriate value for $\bar{u}$ through cross-validation. Ideally, this will require a validation set including both ID and OOD images, but unfortunately $\mathcal{D}_{\text{Val}}$ consists of only ID images. However, given that in this case our image-level uncertainty is obtained by aggregating detection-level uncertainties, the images which have detections with high uncertainty will produce high image-level uncertainty and vice-versa. Using this fact, if we remove the ground-truth objects from the images in $\mathcal{D}_{\text{Val}}$, the resulting image-level uncertainties should be high. We leverage this approach to construct a *pseudo* OOD dataset out of $\mathcal{D}_{\text{Val}}$, by replacing the pixels inside the ground-truth bounding boxes with zeros, thereby removing them from the image and enabling us to cross-validation.

As for the metric to cross-validate $\bar{u}$ against, we observe that existing metrics such as: AUC metrics are unsuitable to evaluate binary predictions, F-Score is sensitive to the choice of the positive class [60] and TPR@0.95[2] [13,24] requires a fixed threshold. As an attractive candidate, Uncertainty Error [46] computes the arithmetic mean of FP and FN rates. However, the arithmetic mean does not heavily penalise choosing $\bar{u}$ on extreme values, potentially leading to the situation where $\hat{a} = 1$ or $\hat{a} = 0$ for all images. To address this, we instead leverage the harmonic mean, which is sensitive to these extreme values. Specifically, we define the Balanced Accuracy (BA) as the harmonic mean of TP rate (TPR) and FP rate (FPR), addressing the aforementioned issue and enabling us to use it to obtain a suitable $\bar{u}$.

5. Calibration of Object Detectors

Accepting or rejecting an image is only one component of the SAOD task, in situations where the image is accepted SAOD then requires the detections to be calibrated. Here we define calibration as the alignment of performance and confidence of a model; which has already been extensively studied for the classification task [8,17,34,47,52,69]. However, existing work which studies the calibration properties of an object detector [35, 36, 48, 51] is limited. For object detection, the goal is to align a detector's confidence with the quality of the joint task of classification and localisation (regression). Arguably, it is not obvious how to extend merely classification-based calibration measures such as Expected Calibration Error (ECE) [17] for object detection. A straightforward extension would be to replace the accuracy in such measures by the precision of the detector, which is computed by validating TPs from a specific IoU threshold. However, this perspective, as employed by [35], does not account for the fact that two object detectors, while having the same precision, might differ significantly in terms of localisation quality.



(a) Classifier [17] (b) Regressor [33] (c) Detector

Figure 3. (a) Calibrated classifier; (b) Calibrated Bayesian regressor, where empirical and predicted CDFs match; (c) Loci of constant IoU boundary, e.g. any predicted box with top-left and bottom-right corners obtained from within the green loci has an IoU > 0.2 with the blue box. The detector is calibrated if its confidence matches the classification *and* the localisation quality.

Hence, as one of the main contributions of this work, we consider the calibration of object detectors from a fundamental perspective and define Localisation-aware Calibration Error (LaECE) which accounts for the joint nature of the task (classification and localisation). We further analyse how calibration measures should be coupled with accuracy in object detection and adapt common post hoc calibration methods such as histogram binning [74], linear regression, and isotonic regression [75] to improve LaECE.

5.1. Localisation-aware ECE

To build an intuitive understanding and to appreciate the underlying complexity in developing a metric to quantify the calibration of an object detector, we first revisit its subtasks and briefly discuss what a calibrated classifier and a calibrated regressor correspond to. For the former, a classifier is calibrated if its confidence matches its accuracy as illustrated in Fig. 3(a). For calibrating Bayesian regressors, there are different definitions [33, 37, 38, 64]. One notable definition [33] requires aligning the predicted and the empirical cumulative distribution functions (cdf), implying $p\%$ credible interval from the mean of the predictive distribution should include $p\%$ of the ground truths for all

[2] Which is the FPR for a fixed threshold set when TPR=0.95.

$p \in [0, 1]$ (Fig. 3(b)). Extending this definition to object detection is nontrivial due to the increasing complexity of the problem. For example, a detection is represented by a tuple $\{\hat{c}_i, \hat{b}_i, \hat{p}_i\}$ with $\hat{b}_i \in \mathbb{R}^4$, which is not univariate as in [33]. Also, this definition to align the empirical and predicted cdfs does not consider the regression accuracy explicitly, and therefore not fit for our purpose. Instead, we take inspiration from an alternative definition that aims to directly align the confidence with the regression accuracy [37, 38].

To this end, without loss of generality, we use IoU as the measure of localisation quality for the detection boxes. Therefore, broadly speaking, if the detection confidence score $\hat{p}_i = 0.8$, then the localisation task is calibrated (ignoring the classification task for now) if the average localisation performance (IoU in our case) is 80% over the entire dataset. To demonstrate, following [56] we plot the loci for fixed values of IoU in Fig. 3(c). In this example, considering the blue-box to be the ground-truth, $\hat{p}_i = 0.2$ implies that a detector is calibrated if the detection box lie on the 'green' loci corresponding to IoU $= 0.2$.

Focusing back onto the joint nature of object detection, we say that an object detector $f : X \mapsto \{\hat{c}_i, \hat{b}_i, \hat{p}_i\}^N$ is calibrated if the classification and the localisation performances *jointly* match its confidence $\hat{p}_i$. More formally,

$$\underbrace{\mathbb{P}(\hat{c}_i = c_i|\hat{p}_i)}_{\textit{Classification perf.}} \underbrace{\mathbb{E}_{\hat{b}_i \in B_i(\hat{p}_i)}[\text{IoU}(\hat{b}_i, b_{\psi(i)})]}_{\textit{Localisation perf.}} = \hat{p}_i, \forall \hat{p}_i \in [0, 1] \quad (3)$$

where $B_i(\hat{p}_i)$ is the set of TP boxes with the confidence score of $\hat{p}_i$, and $b_{\psi(i)}$ is the ground-truth box that $\hat{b}_i$ matches with. Note that in the absence of localisation quality, the above calibration formulation boils down to the standard classification calibration definition.

For a given $B_i(\hat{p}_i)$, the first term in Eq. (3), $\mathbb{P}(\hat{c}_i = c_i|\hat{p}_i)$, is the ratio of the number of correctly-classified to the total number of detections, which is simply the precision. Whereas, the second term represents the average localisation quality of the boxes in $B_i(\hat{p}_i)$.

Following the approximations used to define the well-known ECE, we use Eq. (3) to define LaECE. Precisely, we discretize the confidence score space into $J = 25$ equally-spaced bins [17, 34], and to prevent more frequent classes to dominate the calibration error, we compute the average calibration error for each class separately [34, 47]. Thus, the calibration error for the c-th class is obtained as

$$\text{LaECE}^c = \sum_{j=1}^{J} \frac{|\hat{\mathcal{D}}_j^c|}{|\hat{\mathcal{D}}^c|} \left| \bar{p}_j^c - \text{precision}^c(j) \times \bar{\text{IoU}}^c(j) \right|, \quad (4)$$

where $\hat{\mathcal{D}}^c$ denotes the set of all detections, $\hat{\mathcal{D}}_j^c \subseteq \hat{\mathcal{D}}^c$ is the set of detections in bin j and $\bar{p}_j^c$ is the average of the detection confidence scores in bin j for class c. Furthermore, $\text{precision}^c(j)$ denotes the precision of the j-th bin for c-th class and $\bar{\text{IoU}}^c(j)$ the average IoU of TP boxes in bin j. Then, LaECE is computed as the average of LaECE^c over all the classes. We highlight that for the sake of better accuracy the recent detectors [2, 23, 28–30, 39, 40, 44, 54, 55, 67, 76] tend to obtain $\hat{p}_i$ by combining the classification confidence with the localisation confidence (e.g., obtained from an auxiliary IoU prediction head), which is very well aligned with our LaECE formulation, enforcing $\hat{p}_i$ to match with the joint performance in Eq. (4).

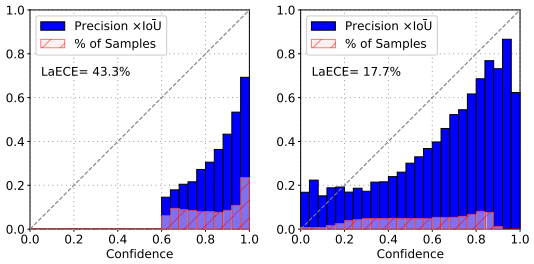
Reliability Diagrams We also produce reliability diagrams to provide insights on the calibration properties of a detector (Fig. 4(a)). To obtain a reliability diagram, we first obtain the performance, measured by the product of precision and IoU (Eq. (4)), for each class over bins and then average the performance over the classes by ignoring the empty bins. Note that if a detector is perfectly calibrated with $\text{LaECE} = 0$, then all the histograms will lie along the diagonal in the reliability diagram since $\text{LaECE}^c = 0$. Similar to classification, if the performance tends to be lower than the diagonal, then the detector is said to be over-confident as in Fig. 4(a), and vice versa for an under-confident detector. Please see Fig. A.14 for more examples.

5.2. Impact of Top-k Survival on Calibration

Top-k survival, a critical part of the post-processing step, selects k detections with the highest confidence in an image. The value of k is typically significantly larger than the number of objects, for example, $k = 100$ for COCO where an average of only 7.3 ground-truth objects exist per image on the val set. Therefore, the final detections may contain numerous low-scoring noisy detections. In fact, ATSS on COCO val set, for example, produces 86.4 detections on average per image after postprocessing, far more than the average number of objects per image.

Since these extra noisy detections do not impact on the widely used AP, most works do not pay much attention to them, however, as we show below, they do have a negative impact on the calibration metric. Thus, this may mislead a practitioner in choosing the wrong model when it comes to calibration quality.

We design a synthetic experiment to show the impact of low-scoring noisy detections on AP and calibration (LaECE). Specifically, if the number of final detections is less than k in an image, we insert "dummy" detections into the remaining space. These dummy detections are randomly assigned a class $\hat{c}_i$, $\hat{p}_i = 0$, and only one pixel to ensure that they do *not* match with any object. Hence, by design, they are "perfectly calibrated". As shown in Fig. 5(a), though these dummy detections have *no* impact on the AP (mathematical proof in App. D), they do give an impression that the model becomes more calibrated (lower LaECE) as k increases. Therefore, considering that extra noisy detections are undesirable in practice, *we do not advocate top-k survival*, instead, we motivate the need to select a detection confidence threshold $\bar{v}$, where detections are rejected if



(a) Base model (b) Calibrated by LR

Figure 4. Reliability diagrams of F-RCNN on $\mathcal{D}_{ID}$ for SAOD-Gen before and after calibration.

their confidence is lower than $\bar{v}$.

An appropriate choice of $\bar{v}$ should produce a set of thresholded-detections with a good balance of precision, recall and localisation errors[3]. In Fig. 5(b), we present the effect of $\bar{v}$ on LRP, where the lowest error is obtained around 0.30 for ATSS and 0.70 for F-RCNN, leading to an average of 6 detections/image for both detectors, far closer to the average number of objects compared to using $k = 100$. Consequently, to obtain $\bar{v}$ for our baseline, we use LRP-optimal thresholding [53, 58], which is the threshold achieving the minimum LRP for each class on the val set.

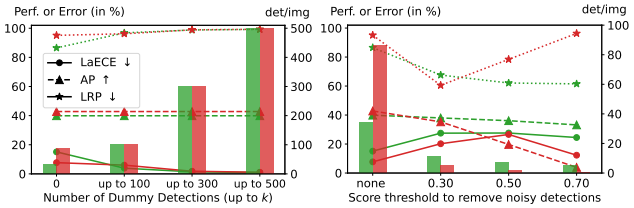
5.3. Post hoc Calibration of Object Detectors

For our baseline, given that LaECE provides the calibration error of the model, we can calibrate an object detector using common calibration approaches from the classification and regression literature. Precisely, for each class, we train a calibrator $\zeta^c : [0,1] \rightarrow [0,1]$ using the input-target pairs $(\{\hat{p}_i, t_i^{cal}\})$ from $\mathcal{D}_{\text{Val}}$, where t_i^{cal} is the target confidence. As shown in App D, LaECE for bin j reduces to

$$\left| \sum_{\substack{\hat{b}_i \in \hat{\mathcal{D}}_j^c \\ \psi(i)>0}} \left(t_i^{cal} - \text{IoU}(\hat{b}_i, b_{\psi(i)})\right) + \sum_{\substack{\hat{b}_i \in \hat{\mathcal{D}}_j^c \\ \psi(i)\leq 0}} t_i^{cal} \right|. \quad (5)$$

Consequently, we seek t_i^{cal} which minimises this value assuming that $\hat{p}_i$ resides in the jth bin. In situations where the prediction is a TP ($\psi(i) > 0$), Eq. (5) is minimized when $\hat{p}_i = t_i^{cal} = \text{IoU}(\hat{b}_i, b_{\psi(i)})$ and conversely, if $\psi(i) \leq 0$, it is minimised when $\hat{p}_i = t_i^{cal} = 0$. We then train linear regression (LR); histogram binning (HB) [74]; and isotonic regression (IR) [75] models with such pairs. Tab. 5 shows that these calibration methods improve LaECE in five out of six cases, and in the case where they do not improve (ATSS on SAOD-Gen), the calibration performance of the base model is already good. Overall, we find IR and LR perform better than HB and consequently we employ LR for SAODETs since LR performs the best on three detectors. Fig. 4(b) shows an example reliability histogram after applying LR, indicating the improvement to calibration.

[3]Using properly-thresholded detections is in fact similar to the Panoptic Segmentation, which is a closely-related task to object detection [31, 32]



(a) Adding dummy detections (b) Thresholding detections

Figure 5. Red: ATSS, green: F-RCNN, histograms present det/img using right axes, results are on COCO val set with 7.3 objects/img. **(a)** Dummy detections decrease LaECE (solid line) artificially with no effect on AP (dashed line). LRP (dotted line), on the other hand, penalizes dummy detections. **(b)** AP is maximized with more detections (threshold 'none') while LRP Error benefits from properly-thresholded detections. (refer App. D)

6. Baseline SAODETs and Their Evaluation

Using the necessary features developed in Sec. 4 and Sec. 5, namely, obtaining: image-level uncertainties, calibration methods as well as the thresholds $\bar{u}$ and $\bar{v}$, we now show how to convert standard detectors into ones that are self-aware. Then, we benchmark them using the SAOD framework proposed in Sec. 3 whilst leveraging our test datasets and LaECE.

Baseline SAODETs To address the requirements of a SAODET, we make the following design choices when converting an object detector into one which is self aware: The hard requirement of predicting whether or not to accept an image is achieved through obtaining image-level uncertainties by aggregating uncertainty scores. Specifically, we use mean(top-3) and obtain an uncertainty threshold $\bar{u}$ through cross-validation using pseudo OOD set approach (Sec. 4). We only keep the detections with higher confidence than $\bar{v}$, which is set using LRP-optimal thresholding (Sec. 5.2). To calibrate the detection scores, we use linear regression as discussed in Sec. 5.3. Thus, we convert all four detectors that we use (Sec. 3) into ones that are self-aware, prefixed by a SA in the tables. For further details, please see App. E.

The SAOD Evaluation Protocol The SAOD task is a robust protocol unifying the evaluation of the: (i) reliability of uncertainties; (ii) the calibration and accuracy; (iii) and performance under domain shift. To obtain quantitative values for the above, we leverage the Balanced Accuracy (Sec. 4) for (i). For (ii) we evaluate the calibration and accuracy using LaECE (Sec. 5) and the LRP [53] respectively, but combine them through the harmonic mean of $1 - \text{LRP}$ and $1 - \text{LaECE}$ on $X \in \mathcal{D}_{\text{ID}}$, which we define as the In-Distribution Quality (IDQ). Similarly, for (iii) we compute the IDQ for $X \in \mathcal{T}(\mathcal{D}_{\text{ID}})$, denoted by IDQ_{T}, but with the principal difference that the detector is flexible to accept or reject severe corruptions (C5) as discussed in Sec. 3. Considering that all of these features are crucial in a safety-critical application, a lack of performance in one

Table 5. Effect of post hoc calibration on LaECE and LRP (in %). ✗: Uncalibrated, HB: Histogram binning, IR: Isotonic Regresssion, LR: Linear Regression. ATSS, combining localisation and classification confidences using multiplication as in our LaECE (Eq. (4)), performs the best on both datasets before/after calibration. Aligned with [47], uncalibrated F-RCNN using cross-entropy loss performs the worst.

Dataset	SAOD-Gen																SAOD-AV							
Detector	F-RCNN				RS-RCNN				ATSS				D-DETR				F-RCNN				ATSS			
Calibrator	✗	LR	HB	IR	✗	LR	HB	IR	✗	LR	HB	IR	✗	LR	HB	IR	✗	LR	HB	IR	✗	LR	HB	IR
LaECE	43.3	17.7	18.6	**16.9**	32.0	17.4	19.6	**17.2**	**15.7**	16.8	18.7	16.7	15.9	**15.7**	17.7	15.9	26.5	**9.8**	10.2	10.2	16.8	**9.0**	9.7	9.7
LRP	74.7	74.7	74.7	74.7	73.6	73.6	73.6	73.6	74.0	74.0	74.1	74.0	71.9	71.9	71.9	71.9	73.5	73.5	73.5	73.5	70.6	70.6	70.6	70.6

Table 6. Evaluating SAODETs. With higher BA and IDQs, SA-D-DETR achieves the best DAQ on SAOD-Gen. For SAOD-AV datasets, SA-ATSS outperforms SA-F-RCNN thanks to its higher IDQs. Bold: SAODET achieves the best, values are in %.

	Self-aware Detector	DAQ↑	$\mathcal{D}_{\text{OOD}}$ BA↑	$\mathcal{D}_{\text{ID}}$ IDQ↑	LaECE↓	LRP↓	$\mathcal{T}(\mathcal{D}_{\text{ID}})$ IDQ↑	LaECE↓	LRP↓	$\mathcal{D}_{\text{Val}}$ LRP↓	AP↑
Gen	SA-F-RCNN	39.7	87.7	38.5	17.3	74.9	26.2	18.1	84.4	59.5	39.9
Gen	SA-RS-RCNN	41.2	**88.9**	39.7	17.1	73.9	27.5	**17.8**	83.5	58.1	42.0
Gen	SA-ATSS	41.4	87.8	39.7	16.6	74.0	27.8	18.2	83.2	58.5	42.8
Gen	SA-D-DETR	**43.5**	**88.9**	**41.7**	**16.4**	**72.3**	**29.6**	17.9	**81.9**	**55.9**	**44.3**
AV	SA-F-RCNN	43.0	**91.0**	41.5	9.5	73.1	28.8	7.2	83.0	54.3	55.0
AV	SA-ATSS	**44.7**	85.8	**43.5**	**8.8**	**71.5**	**30.8**	**6.8**	**81.5**	**53.2**	**56.9**

Table 7. Ablation study by removing: LRP-Optimal thresholding (Sec. 5.2) for $\bar{v} = 0.5$; LR calibration (Sec. 5.3) for uncalibrated model; and image-level threshold $\bar{u}$ (Sec. 4) for the threshold corresponding to $\text{TPR} = 0.95$.

$\bar{v}$	LR	$\bar{u}$	DAQ↑	BA↑	LaECE↓	LRP↓	LaECE_{T}↓	LRP_{T}↓
			36.0	83.2	42.7	76.2	44.1	84.7
✓			36.5	83.2	41.7	**74.8**	43.9	84.7
✓	✓		39.1	83.2	**17.2**	**74.8**	**18.1**	84.7
✓	✓	✓	**39.7**	**87.7**	17.3	74.9	**18.1**	**84.4**

them needs to be heavily penalized. To do so, we introduce the Detection Awareness Quality (DAQ), a unified performance measure to evaluate SAODETs, constructed as the the harmonic mean of BA, IDQ and IDQ_{T}. The resulting DAQ is a higher-better measure with a range of $[0, 1]$.

Main Results Here we discuss how our SAODETs perform in terms of the aforementioned metrics. In terms of our hypotheses, the first evaluation we wish observe is the effectiveness of our metrics. Specifically, we observe in Tab. 6 that a lower LAECE and LRP lead to a higher IDQ; and that a higher BA, IDQ and IDQ_{T} lead to a higher DAQ, indicating that the constructions of these metrics is appropriate. To justify that they are reasonable, we observe that typically more complex and better performing detectors (DETR and ATSS) outperform the simpler F-RCNN, indicating that these metrics reflect the quality of the object detectors.

In terms of observing the performance of these self-aware variants, we can see that while recent state-of-the-art detectors perform very well in terms of LRP and AP on $\mathcal{D}_{\text{Val}}$, their performance drops significantly as we expose them to our $\mathcal{D}_{ID}$ and $\mathcal{T}(\mathcal{D}_{ID})$ which involves domain shift, corruptions and OOD. We would also like to note that the best DAQ corresponding to the best performing model SA-D-DETR still obtains a low score of 43.5% on the SAOD-Gen dataset. As this performance does not seem to be convincing, extra care should be taken before these models are deployed in safety-critical applications. Consequently, our study shows that a significant amount of attention needs to be provided in building self-aware object detectors and effort to reduce the performance gap needs to be undertaken.

Ablation Analyses To test which components of the SAODET contribute the most to their improvement, we perform a simple experiment using SA-F-RCNN (SAOD-Gen). In this experiment, we systematically remove the LRP-optimal thresholds; LR calibration; and pseudo-set approach and replace these features, with a detection-score threshold of 0.5; no calibration; and a threshold corresponding to a TPR of 0.95 respectively. We can see in Tab. 7 that as hypothesized, LRP-optimal thresholding improves accuracy, LR yields notable gain in LAECE and using pseudo-sets results in a gain for OOD detection. In App. E, we further conduct additional experiments to (i) investigate the effect of $\bar{u}$ and $\bar{v}$ on reported metrics and (ii) how common improvement strategies for object detectors affect DAQ.

Evaluating Individual Robustness Aspects We finally note that our framework provides the necessary tools to evaluate a detector in terms of reliability of uncertainties, calibration and domain shift. Thereby enabling the researchers to benchmark either a SAODET using our DAQ measure or one of its individual components. Specifically, (i) uncertainties can be evaluated on $\mathcal{D}_{\text{ID}} \cup \mathcal{D}_{\text{OOD}}$ using AUROC or BA (Tab. 2); (ii) calibration can be evaluated on $\mathcal{D}_{\text{ID}} \cup \mathcal{T}(\mathcal{D}_{\text{ID}})$ using LAECE (Tab. 5); and (iii) $\mathcal{D}_{\text{ID}} \cup \mathcal{T}(\mathcal{D}_{\text{ID}})$ can be used to test detectors developed for single domain generalization [68, 72].

7. Conclusive Remarks

In this paper, we developed the SAOD task, which requires detectors to obtain reliable uncertainties; yield calibrated confidences; and be robust to domain shift. We curated large-scale datasets and introduced novel metrics to evaluate detectors on the SAOD task. Also, we proposed a metric (LAECE) to quantify the calibration of object detectors which respects both classification *and* localisation quality, addressing a critical shortcoming in the literature. We hope that this work inspires researchers to build more reliable object detectors for safety-critical applications.

References

[1] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *The IEEE European Conference on Computer Vision (ECCV)*, 2020. 27

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6

[3] François Bourgeois and Jean-Claude Lassalle. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of ACM*, 14(12):802–804, 1971. 16

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 15

[5] Qi Cai, Yingwei Pan, Yu Wang, Jingen Liu, Ting Yao, and Tao Mei. Learning a unified sample weighting network for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 31

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv*, 1906.07155, 2019. 19

[8] Jiacheng Cheng and Nuno Vasconcelos. Calibrating deep neural networks by pairwise constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[9] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. Active learning for deep object detection via probabilistic modeling. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 15

[11] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross B. Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. arXiv e-prints:2102.01066, 2021. 1

[12] Akshay Raj Dhamija, Manuel Günther, Jonathan Ventura, and Terrance E. Boult. The overlooked elephant of object detection: Open set. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1, 3, 20

[13] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022. 1, 3, 5, 17, 19, 20, 23

[14] Ayers Edward, Sadeghi Jonathan, Redford John, Mueller Romain, and Dokania Puneet K. Query-based hard-image retrieval for object detection at test time. *arXiv*, 2209.11559, 2022. 4

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 2, 24, 25

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 15

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. 5, 6

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 25

[19] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Suenderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[20] Ali Harakeh, Michael H. W. Smart, and Steven L. Waslander. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. *IEEE International Conference on Robotics and Automation (ICRA)*, 2020. 2

[21] Ali Harakeh and Steven L. Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 17, 19, 20, 23

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 31

[23] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 6, 23

[24] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, 2022. 5

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. 3, 17

[26] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *The IEEE European Conference on Computer Vision (ECCV)*, 2012. 27

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 3

[28] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[29] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *The European Conference on Computer Vision (ECCV)*, 2018. 6

[30] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *The European Conference on Computer Vision (ECCV)*, 2020. 6

[31] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[32] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[33] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning (ICML)*, 2018. 5, 6

[34] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 5, 6

[35] Fabian Kuppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1, 2, 5

[36] Fabian Kuppers, Jonas Schneider, and Anselm Haselhoff. Parametric and multivariate uncertainty calibration for regression and object detection. In *Safe Artificial Intelligence for Automated Driving Workshop in The European Conference on Computer Vision*, 2022. 1, 2, 3, 5

[37] Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, pages 393–412, 2020. 5, 6

[38] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors (Basel)*, 22 (15):5540–5550, 2022. 5, 6

[39] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[40] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

[41] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 19

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(2):318–327, 2020. 2

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *The European Conference on Computer Vision (ECCV)*, 2014. 1, 2, 3, 14, 24, 25

[44] Ji Liu, Dong Li, Rongzhang Zheng, Lu Tian, and Yi Shan. Rankdetnet: Delving into ranking constraints for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–273, June 2021. 6

[45] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. In *NeurIPS Workshop on Machine Learning for Autonomous Driving*, 2019. 1

[46] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *International Conference on Robotics and Automation (ICRA)*, 2019. 5

[47] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15288–15299. Curran Associates, Inc., 2020. 5, 6, 8, 31

[48] Muhammad Akhtar Munir, Muhammad Haris Khan, M. Saquib Sarfraz, and Mohsen Ali. Towards improving calibration in object detection under domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5

[49] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. 4, 21

[50] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 3

[51] Lukás Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *NIPS MLITS Workshop on Machine Learning for Intelligent Transportation System*, 2018. 5

[52] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[53] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (LRP): A new performance metric for object detection. In *The European Conference on Computer Vision (ECCV)*, 2018. 2, 7, 24

[54] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. A ranking-based, balanced loss function unifying classification and localisation in object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 6

[55] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Rank & sort loss for object detection and instance segmentation. In *The International Conference on Computer Vision (ICCV)*, 2021. 3, 6, 23, 31

[56] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Generating positive bounding boxes for balanced training of object detectors. In *IEEE Winter Applications on Computer Vision (WACV)*, 2020. 6

[57] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2020. 5

[58] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. One metric to measure them all: Localisation recall precision (lrp) for evaluating visual detection tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2, 7, 24, 27, 33

[59] Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. On model calibration for long-tailed object detection and instance segmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2529–2542. Curran Associates, Inc., 2021. 1

[60] Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. Regmixup: Mixup as a regularizer can surprisingly improve accuracy and out distribution robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149, 2017. 2, 3, 19, 23

[62] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 4, 20

[63] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3, 14

[64] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. 5

[65] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 15

[66] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. SparseR-CNN: End-to-end object detection with learnable proposals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[67] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6

[68] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection, 2023. 1, 8

[69] Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[70] Shaoru Wang, Jin Gao, Bing Li, and Weiming Hu. Narrowing the gap: Improved detector training with noisy location annotations. *IEEE Transactions on Image Processing*, 31:6369–6380, 2022. 2

[71] Xin Wang, Thomas E Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E Gonzalez, and Trevor Darrell. Robust object detection via instance-level temporal cycle confusion. *International Conference on Computer Vision (ICCV)*, 2021. 1

[72] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 8

[73] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 15

[74] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001. 5, 7

[75] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002. 5, 7

[76] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object de-

tector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[77] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 19, 23, 31

[78] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 31

[79] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 19, 23