On Calibration of Object Detectors: Pitfalls, Evaluation and Baselines

Selim Kuzucu*, Kemal Oksuz*, Jonathan Sadeghi, and Puneet K. Dokania

Five AI Ltd., United Kingdom
{selim.kuzucu2, kemal.oksuz, jonathan.sadeghi, puneet.dokania}@five.ai

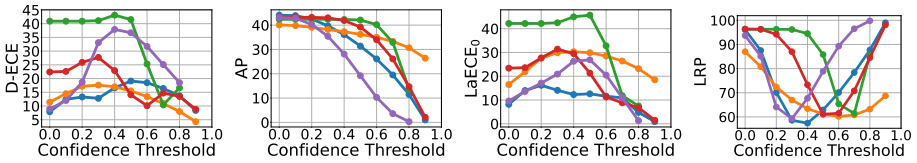
Abstract. Reliable usage of object detectors require them to be calibrated—a crucial problem that requires careful attention. Recent approaches towards this involve (1) designing new loss functions to obtain calibrated detectors by training them from scratch, and (2) post-hoc Temperature Scaling (TS) that learns to scale the likelihood of a trained detector to output calibrated predictions. These approaches are then evaluated based on a combination of Detection Expected Calibration Error (D-ECE) and Average Precision. In this work, via extensive analysis and insights, we highlight that these recent evaluation frameworks, evaluation metrics, and the use of TS have notable drawbacks leading to incorrect conclusions. As a step towards fixing these issues, we propose a principled evaluation framework to jointly measure calibration and accuracy of object detectors. We also tailor efficient and easy-to-use post-hoc calibration approaches such as Platt Scaling and Isotonic Regression specifically for object detection task. Contrary to the common notion, our experiments show that once designed and evaluated properly, post-hoc calibrators, which are extremely cheap to build and use, are much more powerful and effective than the recent train-time calibration methods. To illustrate, D-DETR with our *post-hoc* Isotonic Regression calibrator outperforms the recent *train-time* state-of-the-art calibration method Cal-DETR by more than 7 D-ECE on the COCO dataset. Additionally, we propose improved versions of the recently proposed Localization-aware ECE and show the efficacy of our method on these metrics. Code is available at: `https://github.com/fiveai/detection_calibration`.

Keywords: Calibration · Object Detection · Performance Evaluation

1 Introduction

Object detectors have been widely-used in a variety of safety-critical applications related to, but not limited to, autonomous driving [6, 9, 13, 14, 61, 65] and medical imaging [22, 28, 29, 64]. In addition to being accurate, their confidence estimates should also allow characterization of their error behaviour to make them reliable. This feature, known as calibration, can enable a model to provide valuable information to subsequent systems playing crucial role in making

*Equal contributions. SK contributed during his internship at Five AI Oxford team.



(a) D-ECE **(b)** AP **(c)** LaECE$_0$ **(d)** LRP

Fig. 1: The performance of different detectors over operating confidence thresholds on COCO *minitest*. **Orange**: Faster R-CNN, **Green**: RS R-CNN, **Purple**: ATSS, **Red**: PAA, **Blue**: D-DETR. All measures are lower better except AP. It is not trivial to identify an operating threshold and compare detectors, especially when the common evaluation [30, 41–43, 52], combining D-ECE for calibration and AP for accuracy, is used. Instead, we use LaECE$_0$ and Localisation-Recall-Precision Error (LRP).

safety-critical decisions [4, 24, 26, 37, 39]. Despite its importance, calibration of detectors is a relatively underexplored area in the literature and requires significant attention. Therefore, in this work, we focus on different aspects of the evaluation framework that is now being adopted by most recent works building calibrated detectors and discuss their pitfalls and propose fixes. Additionally, we tailor the well-known post-hoc calibration methods to improve the calibration of a given object detector (trained) with minimal effort.

Naturally, practitioners prefer detectors that perform well in terms of *both accuracy and calibration*, which we will refer to as joint performance. However, unlike classification, choosing the best performing model is non-trivial for object detection. This is because different detectors commonly yield detection sets with varying cardinalities for the same image, and this difference in population size is shown to affect the joint performance evaluation [48]. Furthermore, when object detectors are used in practice, an operating threshold is normally chosen [4, 24, 26, 31, 36, 37, 39], and the choice of this threshold directly influences a detector's performance. Thus, comparing the performance of a detector in terms of calibration or accuracy over different operating thresholds as well as with different detectors is not straightforward, as illustrated in Fig. 1.

We assert that a framework for joint evaluation should follow certain basic principles. Firstly, the detectors should be evaluated on a thresholded set of detections to align with their practical usage. While doing so, the evaluation framework will require a principled *model-dependent threshold selection* mechanism, as the confidence distribution of each detector can differ significantly [47]. Secondly, the calibration should enforce the confidence scores to provide *unambiguous and fine-grained information about the detection quality*. For example, if the confidence score represents the localisation quality of a detection, this provides more fine-grained information than representing whether the object is detected or not. Thirdly, *the datasets should be properly-designed* for evaluation. That is, the training, validation (val.) and in-distribution (ID) test splits should be sampled from the same underlying distribution, and additionally, the domain-shifted test splits — which are crucial for safety-critical applications — should

Table 1: Principles of joint performance evaluation of object detectors in terms of accuracy and calibration, and whether existing evaluation approaches violate them.

Principles of Joint Evaluation	D-ECE-style [30, 41–43, 52]	LaECE-style [48]	CE-style [55]	*Ours*
Model-dependent threshold selection	✗	✓	✗	✓
Unambiguous & fine-grained confidence scores	✗	✗	✗	✓
Properly-designed datasets	✗	✗	✗	✓
Properly-trained detectors & calibrators	✗	✓	✗	✓

be included. Finally, *baseline detectors and calibration methods must be trained properly*, as otherwise the evaluation might provide misleading conclusions.

There are three approaches for jointly evaluating accuracy and calibration:

- *D-ECE-style* [30, 41–43, 52] thresholds the detections commonly from a confidence of 0.30 to compute Detection Expected Calibration Error (D-ECE) and use top-100 detections from each image for Average Precision (AP),
- *LaECE-style* [48] enforces the detectors to be thresholded properly, and combine Localisation-aware Expected Calibration Error (LaECE) with LRP [47],
- *CE-style* [55] thresholds the detections from a confidence score of 0.50 to obtain Calibration Error (CE) and AP.

As summarized in Tab. 1, these evaluations do not adhere to the basic principles mentioned above. To exemplify, D-ECE-style evaluation — the most common evaluation approach [30, 41–43, 52] — uses different operating thresholds for calibration and accuracy, which does not align well with the practical usage of detectors. Also, using a fixed threshold (i.e., 0.30) for all detectors artificially promotes certain detectors. To illustrate, while D-ECE-style evaluation ranks the green detector as the worst in Fig. 1(a), the green one yields the best D-ECE at 0.70. Besides, as shown in Fig. 1(b), AP is maximized at the confidence of 0 (leading to too many detections with low confidences) for all the detectors, and thus AP cannot be used to obtain a proper operating threshold [47, 48]. In terms of conveying fine-grained information, D-ECE aims to align confidence with the precision only, which effectively ignores the localisation quality of the detections, a crucial performance aspect of object detection. Finally, this type of evaluation also has limitations in terms of dataset splits and the chosen baselines as we explore in Sec. 3.

Having proper baseline calibration methods is also essential to monitor the progress in the field. Recently proposed train-time calibration methods commonly employ an auxiliary loss term to regularize the confidence scores during training [30, 41–43, 52]. Such methods are shown to be effective against the Temperature Scaling (TS) [15], which is used as the only post-hoc calibration baseline. Post-hoc calibrators are obtained on a held-out val. set, and hence can easily be applied to any off-the-shelf detector. Despite their potential advantages, unlike for classification [15, 20, 23, 38, 56, 62, 69], post-hoc calibration methods have not been explored for object detection sufficiently [30, 48].

In this paper, we introduce a joint evaluation framework which respects the aforementioned principles (Tab. 1), and thus address the critical drawbacks of existing evaluation approaches. That is, we first define LaECE_0 and LaACE_0,

as novel calibration errors, each of which aims to align the detection confidence scores with their localisation qualities. Thus, the detectors respecting LaECE_0 and LaACE_0 provide informative confidence estimates about their behaviours. We measure accuracy using LRP [45,47], which requires a proper combination of false-positive (FP), false-negative (FN) and localisation errors. Thereby requiring the detectors to be properly-thresholded as shown by the bell-like curves in Fig. 1(d). Also, we design three datasets with different characteristics, and introduce Platt Scaling (PS) as well as Isotonic Regression (IR) as *highly effective* post-hoc calibrators tailored to object detection. Our main contributions are:

- We identify various quirks and assumptions in state-of-the-art (SOTA) methods in quantifying miscalibration of object detectors and show that they, if not treated properly, can provide misleading conclusions.
- We introduce a framework for joint evaluation consisting of properly-designed datasets, evaluation measures tailored to practical usage of object detectors and baseline post-hoc calibration methods. We show that our framework addresses the drawbacks of existing approaches.
- In contrast to the literature, we show that, if designed properly, post-hoc calibrators can significantly outperform the SOTA training time calibration methods. To illustrate, on the common COCO benchmark, D-DETR with our IR calibrator outperforms the SOTA Cal-DETR [43] significantly: (i) by more than 7 points in terms of D-ECE and (ii) ~ 4 points in terms of our challenging LaECE_0.

2 Background and Notation

Object Detectors and Evaluating their Accuracy Denoting the set of M objects in an image X by $\{b_i, c_i\}^M$ where $b_i \in \mathbb{R}^4$ is a bounding box and $c_i \in \{1, \dots, K\}$ is its class; an object detector produces the bounding box $\hat{b}_i$, the class label $\hat{c}_i$ and the confidence score $\hat{p}_i$ for the objects in X, i.e., $f(X) = \{\hat{c}_i, \hat{b}_i, \hat{p}_i\}^N$ with N being the number of predictions. During evaluation, each detection is first labelled as a true-positive (TP) or a FP using a matching function $\psi(\cdot)$ relying on an Intersection-over-Union (IoU) threshold τ to validate TPs. We assume $\psi(i)$ returns the index of the object that a TP i matches to; else i is a FP and $\psi(i) = -1$. Then, AP [11, 16, 33], the common accuracy measure, corresponds to the area under the Precision Recall (PR) curve. Though widely-used, AP has been criticized recently from different aspects [5, 45, 47, 50, 58]. To illustrate, AP is maximized when the number of detections increases [48] as shown in Fig. 1(b). Therefore, AP does not help choosing an operating threshold, which is critical for practical deployment. As an alternative, LRP [45, 47] combines the numbers of TP, FP, FN with the localisation error of the detections, which are denoted by N_{TP}, N_{FP}, N_{FN} and $\mathcal{E}_{loc}(i) \in [0, 1]$ respectively:

$$\text{LRP} = \frac{1}{\text{N}_{\text{FP}} + \text{N}_{\text{FN}} + \text{N}_{\text{TP}}} \left(\text{N}_{\text{FP}} + \text{N}_{\text{FN}} + \sum_{\psi(i)>0} \mathcal{E}_{loc}(i) \right). \tag{1}$$

Unlike AP, LRP requires the detection set to be thresholded properly as both FPs and FNs are penalized in Eq. (1).

Evaluating the Calibration of Object Detectors The alignment of accuracy and confidence of a model, termed calibration, is extensively studied for classification [8, 15, 27, 40, 44, 63]. That is, a classifier is *calibrated* if its accuracy is p for the predictions with confidence of p for all $p \in [0, 1]$. For object detection, [30] extends this definition to enforce that the confidence matches the precision of the detector, $\mathbb{P}(\hat{c}_i = c_i | \hat{p}_i) = \hat{p}_i, \forall \hat{p}_i \in [0, 1]$, where $\mathbb{P}(\hat{c}_i = c_i | \hat{p}_i)$ is the precision. Then, discretizing the confidence space into J bins, D-ECE is

$$\text{D-ECE} = \sum_{j=1}^{J} \frac{|\hat{\mathcal{D}}_j|}{|\hat{\mathcal{D}}|} \left| \bar{p}_j - \text{precision}(j) \right|, \tag{2}$$

where $\hat{\mathcal{D}}$ and $\hat{\mathcal{D}}_j$ are the set of all detections and the detections in the j-th bin, and $\bar{p}_j$ and precision(j) are the average confidence and the precision of the detections in the j-th bin. Alternatively, considering that object detection is a joint task of classification and localisation, LaECE [48] aims to match the confidence with the product of precision and average IoU of TPs. Also, to prevent certain classes from dominating the error, LaECE is introduced as a class-wise measure. Using superscript c to refer to each class and $\bar{\text{IoU}}^c(j)$ as the average IoU of $\hat{\mathcal{D}}_j^c$, LaECE is defined as:

$$\text{LaECE} = \frac{1}{K} \sum_{c=1}^{K} \sum_{j=1}^{J} \frac{|\hat{\mathcal{D}}_j^c|}{|\hat{\mathcal{D}}^c|} \left| \bar{p}_j^c - \text{precision}^c(j) \times \bar{\text{IoU}}^c(j) \right|. \tag{3}$$

Calibration Methods in Object Detection The existing methods for calibrating object detectors can be split into two groups: *(1) Training-time calibration approaches* [41–43, 52, 55] regularize the model to yield calibrated confidence scores during training, which is generally achieved by an additive auxiliary loss. *(2) Post-hoc calibration methods* use a held-out val. set to fit a calibration function that maps the predicted confidence to the calibrated confidence. Specifically, TS [15] is the only method considered as a baseline for recent training time methods [41, 42, 52, 55]. As an alternative, IR [1–3, 66] is used within a limited scope for a specific task called Self-aware Object Detection [48]. Furthermore, its effectiveness neither on a wide range of detectors nor against existing training-time calibration approaches has yet been investigated.

3 Analysis of the Common D-ECE-style Evaluation

D-ECE-style evaluation is the most common evaluation approach adopted by several methods [30, 41–43, 52]. For that reason, here we provide a comprehensive analysis of this evaluation approach and analyse the LaECE-style and CE-style evaluations in App. B. Our analyses below are based on the principles outlined in Sec. 1 and Tab. 1, and show that all approaches have notable drawbacks.

1. Model-dependent threshold selection. As AP is obtained using the top-100 detections and D-ECE is computed on detections thresholded above 0.30, D-ECE-style evaluation uses two different detection sets. This inconsistency is not reflective of how detectors are used in practice. Also, we observe that a fixed threshold of 0.30 for evaluating the calibration induces a bias for certain detectors. To illustrate, we compare the performance of different calibration methods over different thresholds in Fig. 2, where Cal-DETR [43] performs the best only for the threshold 0.30 and the post-hoc TS significantly outperforms it on all other thresholds. Therefore, this method of evaluation is sensitive to the choice of threshold, leading to ambiguity on the best performing method.

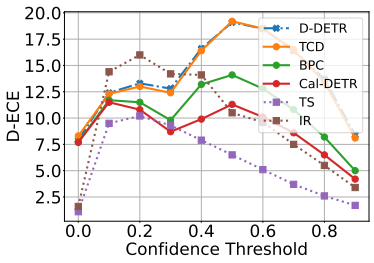


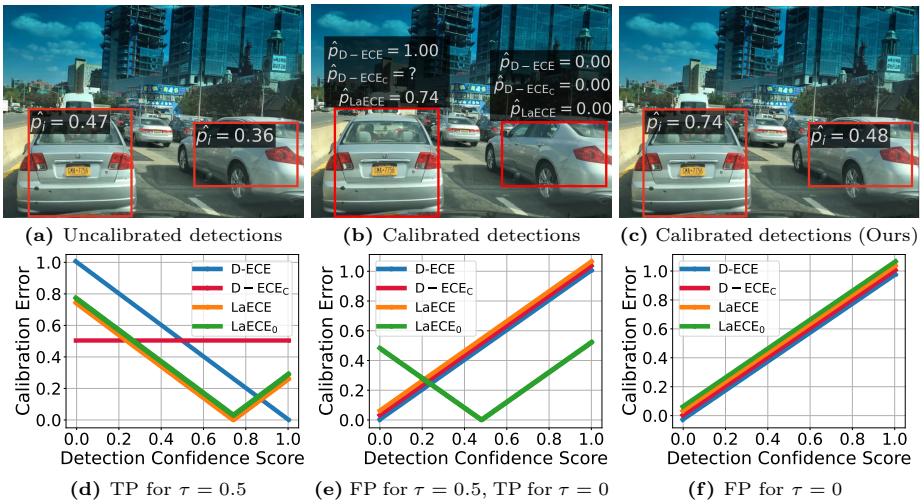
Fig. 2: Comparison of calibration methods in terms of D-ECE on COCO *mini-test* using D-DETR [71]. Post-hoc TS and IR calibrators are obtained on a subset of Objects365 [60] as in D-ECE-style evaluation.

2. Fine-grained confidence scores. Manipulating Eq. (2), we show in App. B that D-ECE for the j-th bin can be expressed as,

$$\left| \sum_{\hat{b}_i \in \hat{\mathcal{D}}_j, \psi(i) > 0} (\hat{p}_i - 1) + \sum_{\hat{b}_i \in \hat{\mathcal{D}}_j, \psi(i) = -1} \hat{p}_i \right|. \tag{4}$$

Eq. (4) implies that D-ECE is minimized when the confidence scores $\hat{p}_i$ of TPs are 1 and those of FPs are 0, which is also how the prediction-target pairs are usually constructed to train post-hoc TS [30, 41–43, 52]. Even if the detector is perfectly calibrated for these binary targets, the confidence scores do not provide information about localisation quality as illustrated by binary-valued $\hat{p}_{\text{D-ECE}}$ for both detections in Fig. 3(b). Also, Popordanoska et al. [55] utilise D-ECE in a COCO-style manner, that is they average D-ECE over different TP validation IoU thresholds similar to COCO-style AP [33]. However, we observe that this way of using D-ECE can promote ambiguous confidence scores. As an example, given two IoU thresholds τ_1 and τ_2, a detection $\hat{b}_i$ with $\tau_1 \leq \text{IoU}(\hat{b}_i, b_{\psi(i)}) < \tau_2$ is a TP for τ_1 but a FP for τ_2. Thus, given Eq. (4), it follows that $\hat{b}_i$ has contradictory confidence targets for τ_1 and τ_2. This is illustrated in Fig. 3(d) in which D-ECE_C (red line) remains constant regardless of the confidence. Thus, using D-ECE (or another calibration measure) in this way should be avoided.

3. Properly-designed datasets. In the literature, the val. set to obtain the post-hoc calibrator is typically taken from a different dataset than the ID dataset [41–43, 52]. Specifically, the post-hoc calibrators are obtained on a subset from from Objects365 [60] and BDD100K [65] for the models trained with COCO [33] and Cityscapes [9] respectively. Hence, as expected, a different dataset inevitably induces domain shift, affecting the performance of the post-hoc calibrator [51]. To show that, following existing approaches, we obtain an IR calibrator [48] on



(a) Uncalibrated detections **(b)** Calibrated detections **(c)** Calibrated detections (Ours)

(d) TP for $\tau = 0.5$ **(e)** FP for $\tau = 0.5$, TP for $\tau = 0$ **(f)** FP for $\tau = 0$

Fig. 3: A pictorial comparison of the different calibration errors. (a) Uncalibrated detections of D-DETR on an image from [65]. The detections on the left and right have IoUs of 0.74 and 0.48 with the objects. (b) Calibrated detections in terms of D-ECE and LaECE using $\tau = 0.50$, and D-ECE$_C$, COCO-style D-ECE as in [55]. D − ECE$_C$ =? as calibration error does not have a global minimum as shown in (d). (c) Calibrated detections in terms of LaECE$_0$ and LaACE$_0$ in which confidence matches IoU. (d-f) Calibration errors for different types of detections, for which LaACE$_0$ behave the same as LaECE$_0$, hence excluded for clarity. App. B presents the details.

Objects365 and compare it with the one obtained on the ID val. set in terms of D-ECE-style evaluation. Tab. 2 shows that the latter IR now outperforms (i) the former one by $\sim$ 11 D-ECE and (ii) SOTA Cal-DETR [43] by 7.4 D-ECE, showing the importance of dataset design for proper evaluation.

4. Properly-trained detectors and calibrators. Though Cityscapes is commonly used in the literature [41–43, 52], the models trained on this dataset follow COCO-style training. Specifically, D-DETR [71] was trained on Cityscapes for only 50 epochs though the training set of Cityscapes is $\sim 40\times$ smaller than that of COCO (3K vs. $\sim$ 118K). We now tailor the training of D-DETR for Cityscapes by (i) $4\times$ longer training considering the smaller training set and (ii) increasing the training image scale considering the original resolution following [7, 9]. We kept all other hyperparameters as they are for both Cal-DETR and D-DETR, and App. B presents the details. Tab. 3 shows that, once trained in this setting, D-DETR is more accurate and calibrated than Cal-DETR.

4 A Framework for Joint Evaluation of Object Detectors

We now present our evaluation approach that respects to the principles in Sec. 1.

Table 2: Effect of using domain-shifted val. set on IR calibrator. Results are reported on COCO-*minitest*. Val. set is N/A for uncalibrated D-DETR and training time calibration method Cal-DETR.

Method	Val set	D-ECE	AP ↑
D-DETR	N/A	12.8	44.1
Cal-DETR	N/A	8.7	44.4
IR	Objects365	14.2	44.1
IR	COCO	**1.3** (+7.4)	44.1

Table 3: COCO training settings are commonly adopted while training D-DETR on Cityscapes. When trained with larger images and longer, DETR performs slightly better than Cal-DETR.

Method	Training Style	D-ECE	AP↑
D-DETR [43]	COCO	13.8	26.8
Cal-DETR [43]	COCO	8.4	28.4
Cal-DETR	Cityscapes	4.0	34.9
D-DETR	Cityscapes	**2.9**	**36.1**

4.1 Towards Fine-grained Calibrated Detection Confidence Scores

Calibration refers to the alignment of accuracy and confidence of a model. Therefore, for an object detector to be calibrated, its confidence should respect both classification and localisation accuracy. We discussed in Sec. 3 that D-ECE, as the common calibration measure, only considers the precision of a detector, thereby ignoring its localisation performance (Eq. (2)). LaECE [48], defined in Eq. (3) as an alternative to D-ECE, enforces the confidence scores to represent the product of precision and average IoU of TPs. Thus, LaECE considers IoUs of only TPs, and effectively ignores the localisation qualities of detections if their IoU is less than the TP validation threshold $\tau > 0$. We assert that this selection mechanism based on IoU unnecessarily limits the information conveyed by the confidence score. We illustrate this on the right car in Fig. 3(b) for which LaECE requires a target confidence of 0 ($\hat{p}_{\text{LaECE}} = 0$) as its IoU is less than $\tau = 0.50$. However, instead of conveying a 0 confidence and implying no detection, representing its IoU by $\hat{p}_i$ provides additional information. Hence, we propose using $\tau = 0$, in which case the calibration criterion of LaECE reduces to,

$$\mathbb{E}_{\hat{b}_i \in B_i(\hat{p}_i)}[\text{IoU}(\hat{b}_i, b_{\psi(i)})] = \hat{p}_i, \forall \hat{p}_i \in [0, 1], \tag{5}$$

where we define $\text{IoU}(\hat{b}_i, b_{\psi(i)}) = 0$ for FPs when $\tau = 0$, $B_i(\hat{p}_i)$ is the set of boxes with the confidence of $\hat{p}_i$ and $b_{\psi(i)}$ is the ground-truth box that $\hat{b}_i$ matches with. To derive the calibration error for Eq. (5), we follow LaECE by using $J = 25$ equally-spaced bins and averaging over class-wise errors and define,

$$\text{LaECE}_0 = \frac{1}{K} \sum_{c=1}^{K} \sum_{j=1}^{J} \frac{|\hat{\mathcal{D}}_j^c|}{|\hat{\mathcal{D}}^c|} \left| \bar{p}_j^c - \bar{\text{IoU}}^c(j) \right|, \tag{6}$$

where $\hat{\mathcal{D}}^c$ and $\hat{\mathcal{D}}_j^c$ denote the set of all detections and those in jth bin respectively, $\bar{p}_j^c$ is the average confidence score and $\bar{\text{IoU}}^c(j)$ is the average IoU of detections in the j-th bin for class c, and the subscript 0 refers to the chosen τ which is 0. Furthermore, similar to the classification literature [40, 44], we define Localisation-aware Adaptive Calibration Error (LaACE) using an adaptive binning approach in which the number of detections in each bin is equal. In order

Table 4: Datasets for evaluating object detection and instance segmentation methods.

Type	Train set	Val set	ID test set	Domain-shifted test set
Common Objects	COCO train	COCO minival	COCO minitest	COCO minitest-C, Obj45K
Autonomous Driving	CS train	CS minival	CS minitest	CS minitest-C, Foggy-CS
Long-tailed Objects	LVIS train	LVIS minival	LVIS minitest	LVIS minitest-C

to capture the model behaviour precisely, we adopt the extreme case in which each bin has only one detection, resulting in an easy-to-interpret measure which corresponds to the mean absolute error between the confidence and the IoU,

$$\text{LaACE}_0 = \frac{1}{K}\sum_{c=1}^{K}\sum_{i=1}^{|\hat{\mathcal{D}}^c|}\frac{1}{|\hat{\mathcal{D}}^c|}\left|\hat{p}_i - \text{IoU}(\hat{b}_i, b_{\psi(i)})\right|. \tag{7}$$

As we show in App. C, LaECE_0 and LaACE_0 are both minimized when $\hat{p}_i = \text{IoU}(\hat{b}_i, b_{\psi(i)})$ for all detections, which is also a necessary condition for LaACE_0. Hence, as illustrated on the right car in Fig. 3(c) and (e), LaECE_0 and LaACE_0 requires conveying more fine-grained information compared to other measures.

4.2 Model-dependent Thresholding for Proper Joint Evaluation

In practice, object detectors employ an operating threshold to preferably output only TPs with high recall. However, AP as the common performance measure does not enable cross-validating such a threshold as it is maximized when the recall is maximized despite a drop in precision [47, 48]. This can be observed in Fig. 1(b) where AP consistently decreases as the confidence threshold increases. Alternatively, LRP (Eq. 1) prefers detectors with high precision, recall and low localisation error as illustrated by the bell-like curves in Fig. 1(d). This is because, unlike AP, LRP severely penalises detectors with low recall or precision, making it a perfect fit for our framework. As a result, we consider LaECE_0 and LaACE_0 with LRP and require each model to be thresholded properly.

4.3 Properly-designed Datasets

We curate three datasets summarized in Tab. 4: (i) COCO [33] including common daily objects; (ii) Cityscapes [9] with autonomous driving scenes; and (iii) LVIS [16], a challenging dataset focusing on the calibration of long-tailed detection. For each dataset, we ensure that train, val. and ID test sets are sampled from the same distribution, and include domain-shifted test sets. As these datasets do not have public labels for test sets, we randomly split their val. sets into two as minival and minitest similar to [17, 48, 49]. In such a way, we provide ID val. sets to enable obtaining post-hoc calibrators and the operating thresholds properly. For domain-shifted test sets, we apply common corruptions [21] to the ID test sets, and include Obj45K [48, 60] and Foggy Cityscapes [59] as more realistic shifts. Our datasets also have mask annotations and hence they can be used to evaluate instance segmentation methods. App. C includes further details.

4.4 Baseline Post-hoc Calibrators Tailored to Object Detection

It is essential to develop post-hoc calibration methods tailored to object detection, which has certain differences from the classification task. Existing methods [41–43, 52, 55] use only TS as a baseline without considering the peculiarities of detection. Specifically, a single temperature parameter T is learned to adjust the predictive distribution while the confidence $\hat{p}_i$ is commonly assumed to be a Bernoulli random variable [30]. However, PS, fitting both a scale and a shift parameter, is the widely-accepted calibration approach when the underlying distribution is Bernoulli [15, 54]. Also, how to construct a useful subset of the detections to train the post-hoc calibrators has not been explored. To address these shortcomings, we present (i) Platt Scaling in which the bias term makes a notable difference in the performance, and (ii) Isotonic Regression by modeling the calibration as a regression problem. Before introducing them, we now present an overview on how we determine the set of detections to train the calibrators.

Overview We obtain post-hoc calibrators on a held out val. set using the detections that are similar to those seen at inference to prevent low-scoring detections from dominating the training of the calibrator. To do so, we cross-validate a calibration threshold $\bar{u}^c$ for each class c and train a class-specific calibrator $\zeta^c : [0,1] \rightarrow [0,1]$ using the detections with higher scores than $\bar{u}^c$. Still, as $\zeta^c(\cdot)$ changes the confidence scores, we need another threshold $\bar{v}^c$, as the operating threshold, to remove the redundant detections after calibration. Following the accuracy measure, we cross-validate $\bar{u}^c$ and $\bar{v}^c$ using LRP. As for inference time, for the i-th detection $(\hat{p}_i, \hat{b}_i, \hat{c}_i)$, if $\hat{p}_i \geq \bar{u}^{\hat{c}_i}$, it survives to the calibrator and then $\hat{p}_i^{cal} = \zeta^{\hat{c}_i}(\hat{p}_i)$. Finally, if $\hat{p}_i^{cal} \geq \bar{v}^{\hat{c}_i}$, the i-th detection is an output of the detector. Alg. A.1 and A.2 provide the details. For $\zeta^c(\cdot)$, we prefer monotonically increasing functions in order not to affect the ranking of the detections significantly and to keep their accuracy as we detail below.

Distribution Calibration via Platt Scaling Assuming that $\hat{p}_i$ is sampled from Bernoulli distribution $\mathcal{B}(\cdot)$, we aim to minimize the Negative Log-Likelihood (NLL) of the predictions on the target distribution $\mathcal{B}(\text{IoU}(\hat{b}_i, b_{\psi(i)}))$ using PS [54]. Accordingly, we recover the logits, and then shift and scale the logits to obtain the calibrated probabilities $\hat{p}_i^{cal}$,

$$\hat{p}_i^{cal} = \sigma(a\sigma^{-1}(\hat{p}_i) + b), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid and $\sigma^{-1}(\cdot)$ is its inverse, as well as $a \geq 0$ and b are the learnable parameters. We derive the NLL for the ith detection in App. C as

$$-(\text{IoU}(\hat{b}_i, b_{\psi(i)}) \log(\hat{p}_i^{cal}) + (1 - \text{IoU}(\hat{b}_i, b_{\psi(i)})) \log(1 - \hat{p}_i^{cal})). \tag{9}$$

Please note that Eq. (9), which is in fact the cross-entropy loss, is minimized if $\hat{p}_i^{cal} = \text{IoU}(\hat{b}_i, b_{\psi(i)})$ when LaECE_0 and LaACE_0 are minimized. We optimize Eq. (9) via the second-order optimization strategy L-BFGS [35] following [30].

Confidence Calibration via Isotonic Regression As an alternative perspective, $\hat{p}_i$ can also be directly calibrated by modelling the calibration as a regression task. To do so, we construct the prediction-target pairs $(\{\hat{p}_i, \text{IoU}(\hat{b}_i, b_{\psi(i)})\})$ on the held-out val. set and then fit an IR model using scikit-learn [53].

Table 5: Comparison with SOTA methods in terms of other evaluation measures on COCO [33]. LRP is reported on LRP-optimal thresholds obtained on val. set. AP is reported on top-100 detections. τ is taken as 0.50. All measures are lower-better, except AP. **Bold:** the best, underlined: second best. PS: Platt Scaling, IR: Isotonic Regression.

Cal. Type	Method	Calibration (thr. 0.30) D-ECE	Calibration (thr. 0.30) LaECE	Calibration (LRP thr.) D-ECE	Calibration (LRP thr.) LaECE	Accuracy LRP	Accuracy AP ↑
Uncal.	D-DETR [71]	12.8	13.2	15.0	12.1	66.3	44.1
Training Time	MbLS [34]	15.6	16.3	18.7	15.8	65.9	44.3
	MDCA [19]	12.2	13.5	14.3	12.6	66.4	43.8
	TCD [42]	12.4	13.1	14.4	12.3	66.6	44.0
	BPC [41]	9.8	13.1	11.4	12.3	66.8	43.6
	Cal-DETR [43]	8.7	12.9	9.7	11.8	66.0	44.4
Post-hoc (Ours)	PS for D-ECE	**0.9**(+7.8)	16.3	**2.4**(+7.3)	15.8	66.3	44.1
	PS for LaECE	11.0	11.5(+1.4)	9.4	10.1(+1.7)	66.3	44.1
	IR for D-ECE	1.3(+7.4)	15.7	2.6(+7.1)	15.3	66.2	44.1
	IR for LaECE	10.2	**8.9**(+4.0)	9.3	**8.2**(+3.6)	66.3	43.7

Adapting Our Approach to Different Calibration Objectives Until now, we considered post-hoc calibrators for LaECE_0 and LaACE_0 though in practice different measures can be preferred. Our post-hoc calibrators can easily be adapted for such cases by considering the dataset design and optimisation criterion. To illustrate, for D-ECE-style evaluation, the calibration dataset is to be class-agnostic where the detections are thresholded from 0.30 with prediction-target pairs for IR as $(\{\hat{p}_i, 0\})$ and $(\{\hat{p}_i, 1\})$ for FPs and TPs respectively.

5 Experimental Evaluation

We now show that our post-hoc calibration approaches consistently outperform training time calibration methods by significant margins (Sec. 5.1) and that they generalize to any detector and can thus be used as a strong baseline (Sec. 5.2).

5.1 Comparing Our Baselines with SOTA Calibration Methods

Here, we compare PS and IR with recent training-time calibration methods considering various evaluation approaches. As these training-time methods mostly rely on D-DETR, we also use D-DETR with ResNet-50 [18]. We obtain the detectors of training time approaches trained with COCO dataset from their official repositories, whereas we incorporate Cityscapes into their official repositories and train them using the recommended setting in Tab. 3.

Comparison on Other Evaluation Approaches Before moving on to our evaluation approach, we first show that our PS and IR outperform all existing training time methods on existing evaluation approaches. For that, we consider D-ECE and the LaECE from $\tau = 0.5$ by including two different evaluation settings for each: (i) the detection set is obtained from the fixed threshold of 0.30 following the convention [30, 41–43], and (ii) the operating thresholds are cross-validated using LRP. Following their standard usage, we use 10 and 25

Table 6: Comparison with SOTA calibration methods on Common Objects using our proposed evaluation. Our gains (green/red) are reported for IR compared to the best existing approach. **Bold:** the best, underlined: second best in terms of calibration.

Calibration Type	Method	COCO *minitest* (ID)			COCO-C (Domain Shift)			Obj45K (Domain Shift)		
		$LaECE_0$	$LaACE_0$	LRP	$LaECE_0$	$LaACE_0$	LRP	$LaECE_0$	$LaACE_0$	LRP
Uncalibrated	D-DETR [71]	12.7	27.1	57.3	14.6	28.7	71.5	16.4	35.8	72.0
Training-time	MbLS [34]	16.5	30.3	56.8	16.8	31.1	71.8	17.3	37.1	71.6
	MDCA [19]	13.1	27.2	57.5	14.5	28.7	71.8	16.6	35.6	72.2
	TCD [42]	13.0	26.7	57.6	14.6	28.3	71.9	**16.3**	35.5	71.7
	BPC [41]	12.4	25.5	57.7	14.1	27.1	72.1	17.3	34.5	72.0
	Cal-DETR [43]	11.6	24.6	56.2	13.8	26.4	70.6	18.8	35.3	71.1
Post-hoc (Ours)	Platt Scaling	9.6	23.5	57.3	12.8	25.6	71.5	17.0	33.7	72.0
	Isotonic Regression	**7.7** (+3.9)	**23.1** (+1.5)	57.2	**10.7** (+3.1)	**25.3** (+1.1)	71.5	17.2 (-0.9)	**33.3** (+1.2)	72.0

Table 7: Comparison with SOTA on Autonomous Driving using our proposed evaluation. Our gains (green/red) are reported for IR compared to the best existing approach. **Bold:** the best, underlined: second best in terms of calibration.

Calibration Type	Method	Cityscapes *minitest* (ID)			Cityscapes-C (Domain Shift)			Foggy Cityscapes (Domain Shift)		
		$LaECE_0$	$LaACE_0$	LRP	$LaECE_0$	$LaACE_0$	LRP	$LaECE_0$	$LaACE_0$	LRP
Uncalibrated	D-DETR [71]	20.3	26.0	57.2	21.4	**25.6**	80.2	18.5	22.3	69.4
Training-time	TCD [42]	16.8	31.7	59.2	23.2	32.4	81.6	24.4	33.8	71.6
	BPC [41]	23.8	31.8	64.9	28.1	33.3	83.7	24.7	30.9	73.8
	Cal-DETR [43]	21.3	25.3	56.9	23.0	26.4	80.8	20.0	23.2	71.0
Post-hoc (Ours)	Platt Scaling	9.6	**23.3**	57.2	17.7	26.2	80.2	11.3	21.6	69.4
	Isotonic Regression	**9.0** (+7.8)	23.7 (+1.6)	56.8	**16.4** (+5.0)	25.8 (-0.2)	80.5	**10.0** (+8.5)	**21.2** (+1.1)	69.5

bins to compute D-ECE and LaECE respectively. We optimize PS and IR by considering the calibration objective as described in Sec. 4.4. Tab. 5 shows that PS and IR outperform SOTA Cal-DETR significantly by more than 7 D-ECE and up to 4 LaECE on COCO *minitest*. Please note that *all previous approaches are optimized for D-ECE thresholded from* 0.30*, in terms of which our PS yields only* 0.9 *D-ECE improving the SOTA by* 7.8. Tab. 5 also suggests that post-hoc calibrators perform the best when the calibration objective is aligned with the measure. App. D shows that our observations generalize to Cityscapes.

Common Objects Setting We now evaluate detectors using our evaluation approach. Tab. 6 shows that IR and PS share the top-2 entries on almost all test subsets by preserving the accuracy (LRP) of D-DETR. Specifically, our gains on ID set and COCO-C are significant, where IR outperforms Cal-DETR by around $3-4$ $LaECE_0$ and $1.0-1.5$ $LaACE_0$. As for Obj45K, the challenging test set with natural shift, IR and PS improve $LaACE_0$ but perform slightly worse in terms of $LaECE_0$. This is an expected drawback of post-hoc approaches when the domain shift is large as they are trained only with ID val. set [51].

Table 8: Comparison with TS using D-DETR. **Bold:** the best, underlined: second best. ✗ : domain-shifted val. set is used to obtain thresholds and calibrators, decreasing the accuracy (red font). Bias term only exists for PS (b in Eq. (8)), thus N/A for IR.

Method	Ablations on Dataset		Ablations on Model		COCO *minitest*			Cityscapes *minitest*		
	ID Val. Set	Threshold	Class-wise	Bias Term	LaECE_0	LaACE_0	LRP	LaECE_0	LaACE_0	LRP
Temperature Scaling	✗				12.3	**20.8**	61.5	21.0	25.9	60.3
(Current Baseline)	✓				12.5	23.1	57.3	20.9	26.3	57.2
Ablations on	✓	✓			11.3	24.8	57.3	13.3	25.5	57.2
Temperature	✓		✓		12.4	22.9	57.3	23.1	27.5	57.2
Scaling	✓	✓	✓		10.6	24.2	57.3	12.7	24.6	57.2
Platt Scaling (Ours)	✓	✓	✓	✓	9.6	23.5	57.3	9.6	**23.3**	57.2
Isotonic Regression (Ours)	✓	✓	✓	N/A	**7.7**	23.1	57.2	**9.0**	23.7	56.8

Table 9: Calibrating and evaluating different object detectors. We use Common Objects setting and report the results on COCO *minitest*. * denotes the detectors in Fig. 1. **Bold:** the best, underlined: second best for calibration.

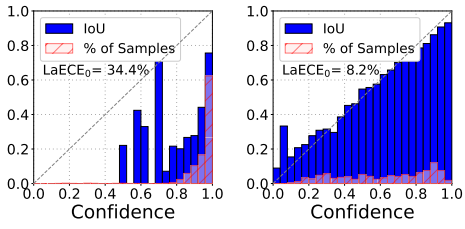
Type	Detector	Backbone	Uncalibrated			Platt Scaling			Isotonic Regression			AP ↑
			LaECE_0	LaACE_0	LRP	LaECE_0	LaACE_0	LRP	LaECE_0	LaACE_0	LRP	
One-Stage	PAA [25]*	R50	15.9	28.1	59.7	9.7	24.3	59.7	**7.7**	**23.8**	59.7	43.2
	ATSS [70]*	R50	19.1	34.0	59.5	10.3	24.7	59.5	**8.5**	**24.1**	59.5	43.1
	GFL [32]	R50	13.7	28.5	59.3	10.3	24.5	59.3	**8.3**	**24.0**	59.3	43.0
	VFNet [68]	R50	13.9	25.8	57.7	10.7	25.1	57.7	**8.3**	**24.6**	57.7	44.8
Two-Stage	Faster R-CNN [57]*	R50	27.0	29.9	60.4	10.4	23.8	60.4	**8.6**	**23.5**	60.4	40.1
	RS R-CNN [46]*	R50	19.7	28.9	58.7	10.2	23.5	58.7	**8.1**	**23.0**	58.8	42.4
DETR-like	D-DETR [71]*	R50	12.7	27.1	57.3	9.6	23.5	57.3	**7.7**	**23.1**	57.2	44.1
	UP-DETR [10]	R50	34.4	35.2	55.8	10.0	22.6	55.8	**8.2**	**22.2**	55.9	42.9
	DINO [67]	R50	13.6	26.9	53.6	10.6	23.5	53.6	**8.9**	**22.8**	53.6	50.4
OVOD	GLIP [31]	Swin-T	13.0	25.3	49.0	9.2	22.4	49.0	**7.7**	**21.8**	49.0	55.7
	G. DINO [36]	Swin-T	13.8	27.5	46.9	8.9	21.9	46.9	**7.7**	**21.3**	47.0	58.3
SOTA	Co-DETR [72]	Swin-L	10.8	23.0	41.5	8.6	20.2	41.5	**6.7**	**19.3**	41.6	64.5
	EVA [12]	ViT(EVA)	17.4	21.2	41.2	8.6	20.3	41.2	**7.1**	**19.9**	41.2	64.5
	MoCaE [49]	N/A	10.6	21.4	40.7	8.9	20.4	40.7	**7.3**	**19.9**	40.7	65.0

Autonomous Driving Setting Tab. 7 shows that our approaches consistently outperform all training time calibration approaches on this setting as well. Specifically, our gains are very significant ranging between 5.0-8.5 LaECE_0 compared to the SOTA Cal-DETR, further presenting the efficacy of our approaches.

Comparison with Existing Temperature Scaling Baseline and Ablations Tab. 8 compares TS for different design choices as well as with our PS and IR. Please note that ✗ corresponds to the baseline setting used in the recent approaches [41–43, 52] that employ Objects365 [60] and BDD100K [60] as domain-shifted val. sets for obtaining the calibrator. Due to this domain shift, the accuracy of TS degrades by up to 4 LRP, in red font, as the operating thresholds obtained on these val. sets do not generalize to the ID set; showing that it is crucial to use an ID val set. In ablations, thresholding the detections and class-wise calibrators generally improves the performance of TS and a more notable gain is observed once the bias term is used in PS. *Our PS outperforms TS baseline obtained on ID val. set by* ~ 3 LaECE_0 *on COCO and* 11.4 LaECE_0 *on Cityscapes.* Finally, IR performs on par or better compared to PS.

5.2 Calibrating and Evaluating Different Detection Methods

Another benefit of our post-hoc calibrators is that they generalize to *any* object detector, thus they can be reliably used as baselines. To show that, we calibrate 14 different detectors with a diverse set of architectures using PS and IR in Tab. 9. The results suggest that both IR and PS perform better than uncalibrated detectors. IR consistently outperforms PS as it fits multiple piece-wise linear functions while PS learns only two parameters (Eq. (8)). Specifically, IR decreases the range of LaECE_0 from $10.6-34.4$ to $6.7-8.9$ on COCO *minitest* by preserving the accuracy, making it a solid baseline. For further insights, Fig. 4 provides the reliability diagrams of the overconfident UP-DETR, of which IR significantly improves its calibration. As for SOTA, MoCaE performs the best in terms of accuracy with 40.7 LRP while Co-DETR has the best calibration with 6.7 LaECE_0 and 19.3 LaACE_0, which the future work should aim to surpass. App. D.3 includes our results on our Long-tailed Objects setting (including object detection and instance segmentation), showing the effectiveness of our post-hoc calibrators on this challenging setting as well.



(a) Uncalibrated **(b)** Calibrated by IR

Fig. 4: The reliability diagrams of UP-DETR.

The performance measures we use in our evaluation framework are also easily interpretable. For accuracy, LRP is a weighted combination of its FP, FN and localisation error components (App. A), which, as an example, are 10.1, 22.7 and 18.8 respectively for Co-DETR [72] calibrated with IR. Also considering $\text{LaACE}_0 = 19.3$, one can easily infer that: *once deployed with the operating thresholds determined by our framework, Co-DETR finds* 77.3% *of the objects with* 89.9% *precision and* 81.2% *IoU where* 19.3% *is the mean absolute error of the confidence to represent IoU*. We believe these intuitive measures will enable practicioners to make better decision when deploying object detectors.

6 Conclusions

The progress in a field heavily relies on the evaluation tools and the baselines used. In this paper, we showed that existing evaluation tools for calibration as well as the baseline post-hoc calibrators for object detectors have significant drawbacks. We remedied that by introducing an evaluation framework including baseline post-hoc calibrators tailored to object detection. Our experiments suggested that, once evaluated and designed properly, the post-hoc calibrators significantly outperform all existing training-time calibrators. This implies the need for research to develop better calibration techniques for object detection, for which, we believe, our evaluation framework will be an essential pillar.

References

1. Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E.: An empirical distribution function for sampling with incomplete information. The annals of mathematical statistics pp. 641–647 (1955)
2. Barlow, R.E., Brunk, H.D.: The isotonic regression problem and its dual. Journal of the American Statistical Association **67**(337), 140–147 (1972)
3. Best, M.J., Chakravarti, N.: Active set algorithms for isotonic regression; a unifying framework. Mathematical Programming **47**(1), 425–439 (1990)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE (Sep 2016). https://doi.org/10.1109/icip.2016.7533003, http://dx.doi.org/10.1109/ICIP.2016.7533003
5. Bolya, D., Foley, S., Hays, J., Hoffman, J.: Tide: A general toolbox for identifying object detection errors. In: The IEEE European Conference on Computer Vision (ECCV) (2020)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
7. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv **1906.07155** (2019)
8. Cheng, J., Vasconcelos, N.: Calibrating deep neural networks by pairwise constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
10. Dai, Z., Cai, B., Lin, Y., Chen, J.: Unsupervised pre-training for detection transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1–11 (2022). https://doi.org/10.1109/tpami.2022.3216514, http://dx.doi.org/10.1109/TPAMI.2022.3216514
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision (IJCV) **88**(2), 303–338 (2010)
12. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
14. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics **37**(3), 362–386 (Nov 2019). https://doi.org/10.1002/rob.21918, http://dx.doi.org/10.1002/rob.21918

15. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017)
16. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
17. Harakeh, A., Waslander, S.L.: Estimating and evaluating regression predictive uncertainty in deep object detectors. In: International Conference on Learning Representations (ICLR) (2021)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
19. Hebbalaguppe, R., Prakash, J., Madan, N., Arora, C.: A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16081–16090 (June 2022)
20. Hekler, A., Brinker, T.J., Buettner, F.: Test time augmentation meets post-hoc calibration: Uncertainty quantification under real-world conditions. Proceedings of the AAAI Conference on Artificial Intelligence **37**(12), 14856–14864 (Jun 2023). https://doi.org/10.1609/aaai.v37i12.26735, https://ojs.aaai.org/index.php/AAAI/article/view/26735
21. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019)
22. Jin, C., Udupa, J.K., Zhao, L., Tong, Y., Odhner, D., Pednekar, G., Nag, S., Lewis, S., Poole, N., Mannikeri, S., Govindasamy, S., Singh, A., Camaratta, J., Owens, S., Torigian, D.A.: Object recognition in medical images via anatomy-guided deep learning. Medical Image Analysis **81**, 102527 (2022). https://doi.org/https://doi.org/10.1016/j.media.2022.102527, https://www.sciencedirect.com/science/article/pii/S1361841522001748
23. Joy, T., Pinto, F., Lim, S.N., Torr, P.H., Dokania, P.K.: Sample-dependent adaptive temperature scaling for improved calibration. Proceedings of the AAAI Conference on Artificial Intelligence **37**(12), 14919–14926 (Jun 2023). https://doi.org/10.1609/aaai.v37i12.26742, https://ojs.aaai.org/index.php/AAAI/article/view/26742
24. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis **65**, 101759 (2020). https://doi.org/https://doi.org/10.1016/j.media.2020.101759, https://www.sciencedirect.com/science/article/pii/S1361841520301237
25. Kim, K., Lee, H.S.: Probabilistic anchor assignment with iou prediction for object detection. In: The European Conference on Computer Vision (ECCV) (2020)
26. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
27. Kumar, A., Liang, P.S., Ma, T.: Verified uncertainty calibration. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32 (2019)

28. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., Wang, Y., Koohbanani, N.A., Jahanifar, M., Tajeddin, N.Z., Gooya, A., Rajpoot, N., Ren, X., Zhou, S., Wang, Q., Shen, D., Yang, C.K., Weng, C.H., Yu, W.H., Yeh, C.Y., Yang, S., Xu, S., Yeung, P.H., Sun, P., Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Smedby, O., Wang, C., Chidester, B., Ton, T.V., Tran, M.T., Ma, J., Do, M.N., Graham, S., Vu, Q.D., Kwak, J.T., Gunda, A., Chunduri, R., Hu, C., Zhou, X., Lotfi, D., Safdari, R., Kascenas, A., O'Neil, A., Eschweiler, D., Stegmaier, J., Cui, Y., Yin, B., Chen, K., Tian, X., Gruening, P., Barth, E., Arbel, E., Remer, I., Ben-Dor, A., Sirazitdinova, E., Kohl, M., Braunewell, S., Li, Y., Xie, X., Shen, L., Ma, J., Baksi, K.D., Khan, M.A., Choo, J., Colomer, A., Naranjo, V., Pei, L., Iftekharuddin, K.M., Roy, K., Bhattacharjee, D., Pedraza, A., Bueno, M.G., Devanathan, S., Radhakrishnan, S., Koduganty, P., Wu, Z., Cai, G., Liu, X., Wang, Y., Sethi, A.: A multi-organ nucleus segmentation challenge. IEEE Transactions on Medical Imaging **39**(5), 1380–1391 (2020). https://doi.org/10.1109/TMI.2019.2947628
29. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Transactions on Medical Imaging **36**(7), 1550–1560 (2017). https://doi.org/10.1109/TMI.2017.2677499
30. Kuppers, F., Kronenberger, J., Shantia, A., Haselhoff, A.: Multivariate confidence calibration for object detection. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
31. Li, L.H., Zhang, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
32. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: The European Conference on Computer Vision (ECCV) (2014)
34. Liu, B., Ayed, I.B., Galdran, A., Dolz, J.: The devil is in the margin: Margin-based label smoothing for network calibration. In: CVPR (2022)
35. Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Math. Program. **45**(1-3), 503–528 (1989), http://dblp.uni-trier.de/db/journals/mp/mp45.html#LiuN89
36. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
37. Lu, Y., Lu, C., Tang, C.K.: Online video object detection using association lstm. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2363–2371 (2017). https://doi.org/10.1109/ICCV.2017.257
38. Ma, X., Blaschko, M.B.: Meta-cal: Well-controlled post-hoc calibration by ranking. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 7235–7245. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/ma21a.html
39. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image

segmentation. IEEE Transactions on Medical Imaging **39**(12), 3868–3878 (Dec 2020). https://doi.org/10.1109/tmi.2020.3006437, http://dx.doi.org/10.1109/TMI.2020.3006437

40. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 15288–15299. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf
41. Munir, M.A., Khan, M.H., Khan, S., Khan, F.S.: Bridging precision and confidence: A train-time loss for calibrating object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11474–11483 (June 2023)
42. Munir, M.A., Khan, M.H., Sarfraz, M., Ali, M.: Towards improving calibration in object detection under domain shift. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 38706–38718. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/fcd812a51b8f8d05cfea22e3c9c4b369-Paper-Conference.pdf
43. Munir, M.A., Khan, S., Khan, M.H., Ali, M., Khan, F.: Cal-DETR: Calibrated detection transformer. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=4SkPTD6XNP
44. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
45. Oksuz, K., Cam, B.C., Akbas, E., Kalkan, S.: Localization recall precision (LRP): A new performance metric for object detection. In: The European Conference on Computer Vision (ECCV) (2018)
46. Oksuz, K., Cam, B.C., Akbas, E., Kalkan, S.: Rank & sort loss for object detection and instance segmentation. In: The International Conference on Computer Vision (ICCV) (2021)
47. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: One metric to measure them all: Localisation recall precision (lrp) for evaluating visual detection tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021)
48. Oksuz, K., Joy, T., Dokania, P.K.: Towards building self-aware object detectors via reliable uncertainty quantification and calibration. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
49. Oksuz, K., Kuzucu, S., Joy, T., Dokania, P.K.: Mocae: Mixture of calibrated experts significantly improves object detection. arXiv preprint arXiv:2309.14976 (2023)
50. Otani, M., Togashi, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Optimal correction cost for object detection evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21107–21115 (2022)
51. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
52. Pathiraja, B., Gunawardhana, M., Khan, M.H.: Multiclass confidence and localization calibration for object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
54. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. **10** (06 2000)
55. Popordanoska, T., Tiulpin, A., Blaschko, M.B.: Beyond classification: Definition and density-based estimation of calibration in object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 585–594 (January 2024)
56. Rahimi, A., Mensink, T., Gupta, K., Ajanthan, T., Sminchisescu, C., Hartley, R.: Post-hoc calibration of neural networks by g-layers. arXiv preprint arXiv:2006.12807 (2020)
57. Ren, S., He, K., Girshick, R., Sun, J.: 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)
58. Rezatofighi, H., Nguyen, T.T.D., Vo, B., Vo, B., Savarese, S., Reid, I.D.: How trustworthy are the existing performance evaluations for basic vision tasks? arXiv e-prints:2008.03533 (2020)
59. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (Sep 2018), https://doi.org/10.1007/s11263-018-1072-8
60. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
61. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
62. Wang, D.B., Feng, L., Zhang, M.L.: Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 11809–11820. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf
63. Wang, D.B., Feng, L., Zhang, M.L.: Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
64. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. arXiv preprint arXiv:1710.01766 (2017)
65. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: 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)
66. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)
67. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)

68. Zhang, H., Wang, Y., Dayoub, F., Sünderhauf, N.: Varifocalnet: An iou-aware dense object detector. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
69. Zhang, J., Yao, W., Chen, X., Feng, L.: Transferable post-hoc calibration on pretrained transformers in noisy text classification. Proceedings of the AAAI Conference on Artificial Intelligence **37**(11), 13940–13948 (Jun 2023). https://doi.org/10.1609/aaai.v37i11.26632, https://ojs.aaai.org/index.php/AAAI/article/view/26632
70. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: 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)
71. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable {detr}: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2021)
72. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)