Towards a Venn-Abers Calibration Method for Object Detectors

Bruce Cyusa Mukama **Soundouss Messoudi** **Sylvain Rousseau** **Sébastien Destercke**
Université de Technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France

Abstract

Object detection locates and classifies a variable number of objects in image. The dedicated architectures typically output a very large number of candidate bounding boxes with confidence scores that can be used to filter-out the false positive predictions. However, these confidence scores are not well calibrated by default: they can be inappropriate for safety-critical applications. To obtain well calibrated confidence scores, we transform these scores into Venn-Abers predictors and the resulting imprecise probabilities enable a principled human-in-the-loop decision process for rejecting the false positive outputs. We evaluate this method on two autonomous driving benchmarks: the Kitti and nuScenes datasets. The results show that the empirical calibration error is reduced, but the accuracy gap is not completely closed.

1 Introduction

Since the number of objects in an image is usually unknown beforehand, object detectors are trained to predict a large number of bounding boxes, to indicate candidate object locations, and to assign a confidence score $\widehat{\rho}_i \in [0,1]$ to each predicted bounding box $\widehat{\boldsymbol{b}}_i$. In modern architectures [Wang et al., 2024, Shi et al., 2020, Yin et al., 2021], these confidence scores are meant to indicate both the quality of the predicted object class $\widehat{c}_i$ and the quality of the predicted bounding box $\widehat{\boldsymbol{b}}_i$. Among other purposes, the scores can be used to filter out duplicated predictions via non-maximum suppression [Hosang et al., 2017]. However, as discussed by [Kumar et al., 2019, Kuppers et al., 2020], these confidence scores are not well calibrated by default: they can be overestimated or underestimated. Therefore, thresholding them to filter the outputs of the detector can lead to missing some objects or to hallucinating (creating) others. This study focuses on the rejection of inaccurate predictions.

To improve the calibration of these object detection confidence scores, the loss function and the architecture of the object detector can be suitably modified [Munir et al., 2023, Schröder et al., 2025, Schubert et al., 2021], but the output scores can also be post-processed. For this purpose, the traditional post-hoc probability calibration methods [Kumar et al., 2019] such as isotonic regression [Robertson et al., 1991], Platt scaling [Platt et al., 1999], and histogram binning [Zadrozny and Elkan, 2002] can be applied, but the predicted bounding box parameters can also be considered to improve the detection scores [Oksuz et al., 2023, Kuppers et al., 2020]. This problem has also been studied beyond object detection. In binary classification, [Gupta et al., 2020] showed that perfect calibration can be achieved by binning approaches, asymptotically, i.e., with an infinitely large calibration dataset. [Vovk et al., 2015] introduced a generalization of isotonic regression and binning that achieves perfect calibration with a finite size calibration dataset: inductive Venn-Abers predictors (IVAPs). To the best of our knowledge, IVAPs have not been applied to reject the false positive outputs of object detectors reliably. However, other conformal prediction (CP) methods [Vovk et al., 2022] have been applied to find suitable thresholds for these detection scores, for other purposes. For instance, [Mukama et al., 2025, Timans et al., 2024] used conformal thresholds to infer a (small) set of classes that contains the ground truth object class with a user-specified probability. Similarly, [Seferis et al., 2023, Copley et al., 2024, Andéol and Masséna, 2025] determined thresholds that keep the rate of missed objects under a user-specified level. In view of the aforementioned, in this study, we apply IVAPs to the scores that are output by object detectors, to obtain

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco. Copyright 2026 by the author(s).

well-calibrated confidence scores. Additionally, we derive a principled (human-in-the-loop) decision process that allows object detectors to reject inaccurate predictions or defer ambiguous cases to an oracle.

This article is organized as follows: Section 2 presents object detection. In Section 3 we apply IVAPs to object detection. Section 4 presents the experiments and the results. Section 5 summarizes this study.

2 Object detection

To annotate images, each target object is assigned a class label $c_i \in \{1, \ldots, C\}$ and located by drawing a "ground truth" bounding box $\boldsymbol{b}_i$ around its pixels, with $C < \infty$ as the number of possible categories, e.g., "Car", "Pedestrian", "Cyclist". In 2D, the ground truth bounding box $\boldsymbol{b}_i$ can be parameterized by the coordinates of its center point, its width, and its height. Afterwards, these annotations are used to train the object detector, via deep learning [Qian et al., 2022].

At test time, the object detector outputs a fixed (large) number of boxes. Each predicted bounding box $\widehat{\boldsymbol{b}}_i$ is associated with a predicted confidence score $\widehat{\rho}_i \in [0, 1]$. This score can be interpreted as the probability $P(t_i = 1)$ that the box $\widehat{\boldsymbol{b}}_i$ is a true positive. A predicted bounding box $\widehat{\boldsymbol{b}}_i$ is a true positive if the associated class prediction $\widehat{c}_i$ is accurate and if this box overlaps enough the ground truth box $\boldsymbol{b}_i$, given a threshold (IoU_{th}) of the intersection over union (IoU) [Qian et al., 2022]:

$$t_i = 1 \iff (\widehat{c}_i = c_i) \wedge \Big(\text{IoU}(\widehat{\boldsymbol{b}}_i, \boldsymbol{b}_i) \geq \text{IoU}_{\text{th}}\Big) \quad (1)$$

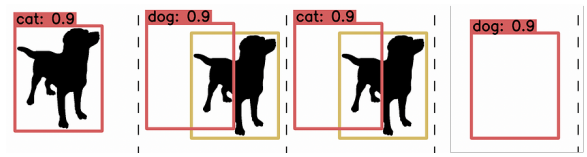
with $t_i \in \{0, 1\}$ and

$$\text{IoU}(\widehat{\boldsymbol{b}}_i, \boldsymbol{b}_i) = \frac{|\widehat{\boldsymbol{b}}_i \cap \boldsymbol{b}_i|}{|\widehat{\boldsymbol{b}}_i \cup \boldsymbol{b}_i|}, \quad (2)$$

where $|\widehat{\boldsymbol{b}}_i \cap \boldsymbol{b}_i|$ is the size of the spatial region at the intersection of the boxes $\widehat{\boldsymbol{b}}_i$ and $\boldsymbol{b}_i$, and $|\widehat{\boldsymbol{b}}_i \cup \boldsymbol{b}_i|$ is the size of the spatial region that is formed by their union. At test time, the ground truth classes and boxes are unavailable, but the inaccurate predictions (illustrated by Figure 1) could be reliably identified and rejected if the scores $\widehat{\rho}_i$ were well calibrated.

3 IVAPs-based object detection

In this study, we apply IVAPs [Vovk et al., 2015] to calibrate the confidence scores $\widehat{\rho}_i$ that are output by object detectors. We start from a set of labeled images $\mathcal{I}$. For each image in $\mathcal{I}$, we denote by A its set of ground truth boxes and by B the set of bounding boxes that are predicted when this image is processed



$\widehat{c}_i \neq c_i$ | $\text{IoU} \leq \text{IoU}_{\text{th}}$ | Both wrong | Hallucinated

Figure 1: illustrations [Bolya et al., 2020] of ground truth boxes (gold) and false positive predictions (red), and confidence scores (black). The texts at the bottom of the boxes describe the type of error. In the two rightmost cases, both $\widehat{c}_i$ and $\widehat{\boldsymbol{b}}_i$ are inaccurate.

by the object detector. For each box in A, we use the user-specified threshold IoU_{th} and the Hungarian matching algorithm [Kuhn, 1955] to find the box in B that verifies Equation (1), and maximizes the IoU score in Equation (2). Thus, by processing all the images in the set $\mathcal{I}$, we collect two finite subsets of the calibration dataset: $\mathcal{D}_+ = \{(\widehat{\boldsymbol{b}}_i, \boldsymbol{b}_i, \widehat{\rho}_i) \mid t_i = 1\}$ and $\mathcal{D}_- = \{(\widehat{\boldsymbol{b}}_i, \emptyset, \widehat{\rho}_i) \mid t_i = 0\}$. In other words, we draw inspiration from [Schubert et al., 2021] and consider the rejection of false positive predictions as a meta-classification problem. More specifically, we assign the meta-label $t_i = 1$ to the boxes in the set $\mathcal{D}_+$ and the meta-label $t_i = 0$ to the boxes in the set $\mathcal{D}_-$. In this way, we carefully collect the proper calibration dataset $\mathcal{D}_{\text{Cal}} = \{(\widehat{\rho}_i, t_i)\}_{i=1}^n$. Note that the meta-label $t_i \in \{0, 1\}$ (i.e., "false positive" or "true positive") should not be confused with the category $c_i \in \{1, \ldots, C\}$ (e.g., "Car", "Pedestrian", "Cyclist") in the classification sub-task of object detection.

For any (new) unlabeled prediction $\widehat{\boldsymbol{b}}_{n+1}$ with the score $\widehat{\rho}_{n+1}$, we infer the well calibrated confidence or probability interval $[\underline{\rho}_{n+1}, \overline{\rho}_{n+1}]$ as follows. We fit a first isotonic regression $\overline{g}$ on the values $\{(\widehat{\rho}_i, t_i)\}_{i=1}^{n+1}$ and the unlabeled box $\widehat{\boldsymbol{b}}_{n+1}$ is processed by considering the possibility that the (unknown) label of this new box is $t_{n+1} = 1$. The map $\overline{g}$ is modeled as a non-decreasing piece-wise linear function that minimizes the following loss function [De Leeuw et al., 2010]:

$$\mathcal{L}_{n+1} = \sum_{i=1}^{n+1} (t_i - \overline{g}(\widehat{\rho}_i))^2. \quad (3)$$

By considering the alternative possibility for the (unknown) label of the new box, i.e., $t_{n+1} = 0$, we similarly fit a second isotonic regression $\underline{g}(\widehat{\rho}_i)$ on the values $\{(\rho_i, t_i)\}_{i=1}^{n+1}$. To infer the interval $[\underline{\rho}_{n+1}, \overline{\rho}_{n+1}]$, we set the lower bound $\underline{\rho}_{n+1}$ to $\underline{g}(\widehat{\rho}_{n+1})$ and $\overline{\rho}_{n+1}$ to $\overline{g}(\widehat{\rho}_{n+1})$.

Theorem 1 *When the confidence scores in the calibration dataset $\mathcal{D}_{Cal} = \{(\widehat{\rho}_i, t_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d.) w.r.t. the confidence*

score $\widehat{\rho}_{n+1}$ of the new predicted box $\widehat{\boldsymbol{b}}_{n+1}$, the IVAPs bounds $[\underline{\rho}_{n+1}, \overline{\rho}_{n+1}]$ provide the following almost sure (a.s.) statistical guarantees:

$$P\left(\widehat{c}_{n+1} = c_{n+1},\ \mathrm{IoU}(\widehat{\boldsymbol{b}}_{n+1}, \boldsymbol{b}_{n+1}) \geq \mathrm{IoU}_{th}\right) = P(t_{n+1} = 1) \in [\underline{\rho}_{n+1}, \overline{\rho}_{n+1}] \ \ a.s., \quad (4)$$

The above is a rephrasing of the Proposition (1) in [Vovk et al., 2015]. Thus, for the formal proofs, we refer the readers to [Vovk et al., 2015].

Regarding the efficiency of this method, the IVAPs add a reasonable computational load to the underlying object detector. Even though a new pair of functions $\underline{g}$ and $\overline{g}$ must be determined for each new prediction $i \in \{n+1, n+2, \ldots\}$, the dedicated PAVA [De Leeuw et al., 2010] algorithm is efficient. More specifically [Vovk et al., 2015], the functions $\underline{g}$, $\overline{g}$ can be initially fit in $O(n \log n)$ time, for the first unlabeled prediction ($i = n+1$), and adapted afterwards to the other unlabeled predictions ($i > n+1$) in $O(\log n)$ time. The tightness of these intervals depends on the size n of the calibration dataset, but also on the quality of the underlying binary classifier.

Furthermore, a scalar confidence score $\widetilde{\rho}_{n+1}$ can be deduced from the inferred probability interval [Vovk et al., 2015]:

$$\widetilde{\rho}_{n+1} = \frac{\overline{\rho}_{n+1}}{1 - \underline{\rho}_{n+1} + \overline{\rho}_{n+1}}. \quad (5)$$

Thus, we can predict the meta-label $\widehat{t}_{n+1} = 1$ when $\widetilde{\rho}_{n+1} > 0.5$, otherwise $\widehat{t}_{n+1} = 0$. However, the key to reliable false positive rejections is the uncertainty in the (raw) confidence score $\widehat{\rho}_i$. In fact, the width of each interval $[\underline{\rho}_{n+1}, \overline{\rho}_{n+1}]$ indicates the uncertainty in the underlying object detector. Therefore, given an abstention (or precision) threshold ρ_δ, the detector can justifiably abstain from predicting the meta-label $\widehat{t}_i$ when $(\overline{\rho}_{n+1} - \underline{\rho}_{n+1}) \geq \rho_\delta$, and defer such hard decisions to stronger experts (humans). Deferring increases the trustworthiness of the system.

4 Experiments

4.1 Implementation details

4.1.1 Datasets

We use the Kitti dataset [Geiger et al., 2013], a small popular autonomous driving benchmark, and nuScenes [Caesar et al., 2020], one of the most diverse and large autonomous driving benchmarks. On both datasets, we only predict "Car" objects. We use the PV-RCNN neural network [Shi et al., 2020] on Kitti and CenterPoint [Yin et al., 2021] on nuScenes. On both datasets, we train with 50% of the annotated images. The remaining set of images is randomly split two halves: one half for calibration and the other half for testing. Following Kitti's evaluation protocol [Geiger et al., 2013], the bounding boxes are matched with the threshold $\mathrm{IoU}_{\mathrm{th}} = 0.7$.

4.1.2 Baselines

"NN+HIL" designates our method in Section 3. We compare it to three other methods: the "NN" baseline which denotes the raw confidence scores $\widehat{\rho}_i$, the "NN+Platt" baseline which fits a logistic regression on the raw scores, and the "NN+Isot" baseline which fits a single isotonic regression on the raw scores. All the baselines use the binary classification threshold 0.5 (i.e., the argmax rule) to predict the meta-labels $\widehat{t}_i$.

4.1.3 Evaluation metrics

We evaluate the meta-classification accuracy (AC) and the abstention rate (AB):

$$\mathrm{AC} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{\widehat{t}_j = t_j\}, \quad (6)$$

$$\mathrm{AB} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{(\overline{\rho}_j - \underline{\rho}_j) \geq \rho_\delta\}, \quad (7)$$

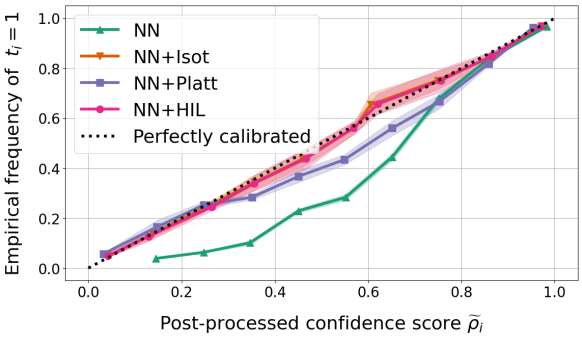
with $j \in \{1, \ldots, m\}$ as the index of the example in the test set. $\mathbb{1}\{S\}$ equates to 1 if the statement S is verified and to 0 otherwise. t_j is the right meta-label of the predicted box $\widehat{\boldsymbol{b}}_j$ (according to Equation (1)) and $\widehat{t}_j$ is the predicted meta-label. We also evaluate the expected calibration error rate [Kumar et al., 2019, Naeini et al., 2015]:

$$\mathrm{ECE} = \sqrt{\sum_{k=1}^{10} \frac{|\mathscr{B}_k|}{m} \left(\mathrm{prob}\left(\mathscr{B}_k\right) - \mathrm{pos}\left(\mathscr{B}_k\right)\right)^2}, \quad (8)$$

where the bins $\mathscr{B}_k$ are obtained by dividing the $[0, 1]$ interval into 10 equal-width ranges, as in [Naeini et al., 2015]. $\mathrm{pos}(\mathscr{B}_k)$ is $\frac{1}{|\mathscr{B}_k|} \sum_{j \in \mathscr{B}_k} t_j$, the ratio of true positives in the bin $\mathscr{B}_k$ of size $|\mathscr{B}_k|$. Similarly, $\mathrm{prob}(\mathscr{B}_k)$ is $\frac{1}{|\mathscr{B}_k|} \sum_{j \in \mathscr{B}_k} \widetilde{\rho}_j$, the average of the post-processed confidence scores $\widetilde{\rho}_j$ in $\mathscr{B}_k$.

By repeating the random split that separates the calibration data from the test data, we repeat the experiments 10 times and report the metrics in the format "average ± standard deviation". The best results are highlighted in boldface.

(a) The reliability curves on Kitti



(b) The reliability curves on nuScenes

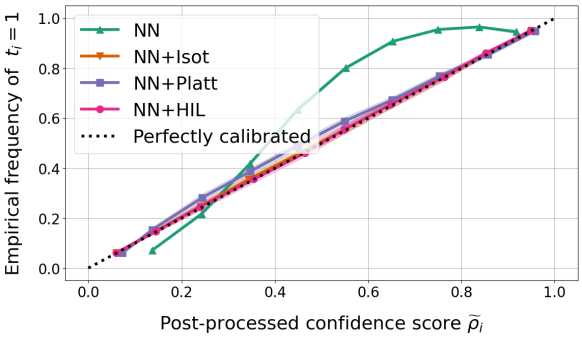


Figure 2: the reliability curves of the tested baselines.

4.1.4 Results

As detailed by Table 1, all the tested calibration methods improve the accuracy (AC) of the base model and their empirical calibration errors (ECE) are smaller on the (large) nuScenes dataset than on the (small) Kitti dataset. In fact, the calibration dataset on nuScenes is more than 17 times larger than the one on Kitti. Even though the ECE(s) of "NN+HIL" and "NN+Isot" remain very close on both datasets, our "NN+HIL" method has the advantage of being provably well calibrated with a finite-size calibration set, under the i.i.d. assumption in Theorem 1. Moreover, "NN+HIL" successfully increases the accuracy (AC) on both datasets through our principle abstention mechanism.

Ideally, deferring to a human should close the accuracy (AC) gap, but our "NN+HIL" method only reduces it. On both datasets, even though the abstention rates (AB) of "NN+HIL" are close to the accuracy gaps of the base model "NN", the accuracy gap of "NN+HIL" is merely reduced by these abstentions: it is not completely closed. This can be explained by the fact that our principle abstention mechanism only considers the width of the well calibrated IVAP interval.

Regarding the main hyperparameters, Table 2 shows that the abstention rate and the accuracy decrease as the abstention threshold ρ_δ increases. It also shows that accuracy gain depends on the content of the calibration dataset. Table 3 is obtained by calibrating with a portion of $\mathcal{D}_{\text{Cal}}$, with $\rho_\delta = 1$, and they show that larger calibration sets improve the empirical calibration error and the accuracy. These improvements also depend on the content of the calibration set.

Table 1: the performances of the tested baselines.

Dataset	Baseline	AC	AB	ECE
Kitti	NN	0.8841 ± 0.0026	**0.0000 ± 0.0000**	0.1028 ± 0.0034
	NN+HIL	**0.9135 ± 0.0057**	0.1290 ± 0.0535	**0.0091 ± 0.0053**
	NN+Isot	0.9028 ± 0.0023	**0.0000 ± 0.0000**	0.0093 ± 0.0051
	NN+Platt	0.9005 ± 0.0027	**0.0000 ± 0.0000**	0.0396 ± 0.0052
nuScenes	NN	0.8519 ± 0.0021	**0.0000 ± 0.0000**	0.1316 ± 0.0022
	NN+HIL	**0.8746 ± 0.0111**	0.1455 ± 0.1741	**0.0074 ± 0.0038**
	NN+Isot	0.8666 ± 0.0013	**0.0000 ± 0.0000**	**0.0074 ± 0.0039**
	NN+Platt	0.8662 ± 0.0015	**0.0000 ± 0.0000**	0.0200 ± 0.0045

Table 2: the influence of the abstention threshold ρ_δ.

	Kitti		nuScenes	
ρ_δ	AB	AC	AB	AC
0.004	0.2750 ±0.0440	**0.9366 ±0.0082**	0.5243 ±0.0467	**0.8968 ±0.0023**
0.006	0.2201 ±0.0499	0.9249 ±0.0078	0.5116 ±0.0467	0.8957 ±0.0025
0.010	0.1636 ±0.0420	0.9174 ±0.0060	0.5041 ±0.0447	0.8948 ±0.0022
0.020	0.0650 ±0.0605	0.9081 ±0.0048	0.4761 ±0.0485	0.8936 ±0.0023
0.100	**0.0000 ±0.0000**	0.9029 ±0.0021	**0.0000 ±0.0000**	0.8666 ±0.0012

Table 3: the influence of the size of the calibration set $\mathcal{D}_{\text{Cal}}$. We calibrate with a varying portion of $\mathcal{D}_{\text{Cal}}$.

	Kitti		nuScenes	
Used data	ECE	AC	ECE	AC
10% of $\mathcal{D}_{\text{Cal}}$	0.0239 ± 0.0107	0.9016 ± 0.0034	0.0140 ± 0.0063	0.8664 ± 0.0013
20% of $\mathcal{D}_{\text{Cal}}$	0.0162 ± 0.0061	0.9021 ± 0.0019	0.0097 ± 0.0055	0.8665 ± 0.0013
40% of $\mathcal{D}_{\text{Cal}}$	0.0133 ± 0.0050	0.9024 ± 0.0021	0.0091 ± 0.0059	0.8665 ± 0.0013
80% of $\mathcal{D}_{\text{Cal}}$	0.0098 ± 0.0046	**0.9029 ± 0.0022**	0.0086 ± 0.0048	**0.8666 ± 0.0012**
100% of $\mathcal{D}_{\text{Cal}}$	**0.0091 ± 0.0051**	**0.9029 ± 0.0021**	**0.0074 ± 0.0036**	**0.8666 ± 0.0012**

5 Conclusion

In this article, we apply IVAPs to calibrate the confidence scores of object detectors and to reject the false positive predictions reliably. This method can output an interval that almost surely contains the actual probability that a newly predicted bounding box is a true positive. This method justifiably defers the harder decisions to humans (oracles). The experiments show that this method can reduce the empirical calibration error more than the traditional probability calibration methods, but its performances decrease with smaller datasets and it relies on the assumption that the predictions are i.i.d. Future works may eliminate this assumption by learning the dependency structures that may actually exist between the predicted confidence scores $\{\widehat{\rho}_i\}_{i=1}^{n+1}$ due to confounding variables, limited data, and spurious correlations [Xiao et al., 2021].

References

[Andéol and Masséna, 2025] Andéol, L. and Masséna, T. (2025). Sequential conformal risk control for safe railway signaling detection. *Proceedings of Machine Learning Research*, 266:1–4.

[Bolya et al., 2020] Bolya, D., Foley, S., Hays, J., and Hoffman, J. (2020). Tide: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, pages 558–573. Springer.

[Caesar et al., 2020] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

[Copley et al., 2024] Copley, V., Finlay, G., and Hiett, B. (2024). The uncertain object: Application of conformal prediction to aerial and satellite images. In *The 13th Symposium on Conformal and Probabilistic Prediction with Applications*, pages 73–89. PMLR.

[De Leeuw et al., 2010] De Leeuw, J., Hornik, K., and Mair, P. (2010). Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32:1–24.

[Geiger et al., 2013] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

[Gupta et al., 2020] Gupta, C., Podkopaev, A., and Ramdas, A. (2020). Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723.

[Hosang et al., 2017] Hosang, J., Benenson, R., and Schiele, B. (2017). Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515.

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

[Kumar et al., 2019] Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. *Advances in neural information processing systems*, 32.

[Kuppers et al., 2020] Kuppers, F., Kronenberger, J., Shantia, A., and Haselhoff, A. (2020). Multivariate confidence calibration for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 326–327.

[Mukama et al., 2025] Mukama, B. C., Messoudi, S., Destercke, S., and Rousseau, S. (2025). Copula-based conformal prediction for prioritized heterogeneous multi-task learning. *Pattern Recognition*, page 112347.

[Munir et al., 2023] Munir, M. A., Khan, M. H., Khan, S., and Khan, F. S. (2023). 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*, pages 11474–11483.

[Naeini et al., 2015] Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*.

[Oksuz et al., 2023] Oksuz, K., Joy, T., and Dokania, P. K. (2023). Towards building self-aware object detectors via reliable uncertainty quantification and calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9263–9274.

[Platt et al., 1999] Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

[Qian et al., 2022] Qian, R., Lai, X., and Li, X. (2022). 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130:108796.

[Robertson et al., 1991] Robertson, T., Wright, F., and Dykstra, R. (1991). Order restricted statistical inference. *Journal of Applied Econometrics*, 6(1):105–107.

[Schröder et al., 2025] Schröder, C., Schlüter, M.-R., and Lienkamp, M. (2025). Calibrating the full predictive class distribution of 3d object detectors for autonomous driving. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 187–194. IEEE.

[Schubert et al., 2021] Schubert, M., Kahl, K., and Rottmann, M. (2021). Metadetect: Uncertainty quantification and prediction quality estimates for object detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

[Seferis et al., 2023] Seferis, E., Fraunhofer, I., Burton, S., Kollias, S., and Cheng, C.-H. (2023). Statistical guarantees for safe 2d object detection post-processing. In *International Conference on Computer Vision 2023*.

[Shi et al., 2020] Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., and Li, H. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538.

[Timans et al., 2024] Timans, A., Straehle, C.-N., Sakmann, K., and Nalisnick, E. (2024). Adaptive bounding box uncertainties via two-step conformal prediction. In *European Conference on Computer Vision*.

[Vovk et al., 2022] Vovk, V., Gammerman, A., and Shafer, G. (2022). *Algorithmic Learning in a Random World*. Springer Nature.

[Vovk et al., 2015] Vovk, V., Petej, I., and Fedorova, V. (2015). Large-scale probabilistic predictors with and without guarantees of validity. *Advances in Neural Information Processing Systems*, 28.

[Wang et al., 2024] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024). Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011.

[Xiao et al., 2021] Xiao, K. Y., Engstrom, L., Ilyas, A., and Madry, A. (2021). Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*.

[Yin et al., 2021] Yin, T., Zhou, X., and Krähenbühl, P. (2021). Center-based 3d object detection and tracking. *CVPR*.

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002). 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.