Adaptive Bounding Box Uncertainties via Two-Step Conformal Prediction

Alexander Timans[1*], Christoph-Nikolas Straehle[2], Kaspar Sakmann[2], and Eric Nalisnick[1]

[1] UvA-Bosch Delta Lab, University of Amsterdam
[2] Bosch Center for AI, Robert Bosch GmbH

Abstract. Quantifying a model's predictive uncertainty is essential for safety-critical applications such as autonomous driving. We consider quantifying such uncertainty for multi-object detection. In particular, we leverage conformal prediction to obtain uncertainty intervals with guaranteed coverage for object bounding boxes. One challenge in doing so is that bounding box predictions are conditioned on the object's class label. Thus, we develop a novel two-step conformal approach that propagates uncertainty in predicted class labels into the uncertainty intervals of bounding boxes. This broadens the validity of our conformal coverage guarantees to include incorrectly classified objects, thus offering more actionable safety assurances. Moreover, we investigate novel ensemble and quantile regression formulations to ensure the bounding box intervals are adaptive to object size, leading to a more balanced coverage. Validating our two-step approach on real-world datasets for 2D bounding box localization, we find that desired coverage levels are satisfied with practically tight predictive uncertainty intervals.

Keywords: Object Detection · Conformal Prediction · Uncertainty

1 Introduction

Safety-critical applications in domains such as autonomous transportation [37, 69] and mobile robotics [36] benefit greatly from accurate estimates of the model's predictive uncertainty. Yet one obstacle to principled uncertainty quantification (UQ) for computer vision is the pervasive use of deep neural networks, which are often unamenable to traditional techniques for UQ. The framework of *Conformal Prediction* (CP) [2, 54, 65] enables a form of distribution-free UQ that is agnostic to the predictive model's structure, rendering it well-suited for such 'black-box' models.

In this work, we propose a CP framework designed to quantify predictive uncertainties in multi-object detection tasks with multiple classes (see Fig. 2). CP allows us to produce computationally cheap, *post-hoc* distribution-free prediction intervals, which come equipped with a coverage guarantee for the true

* Correspondence to <a.r.timans@uva.nl>

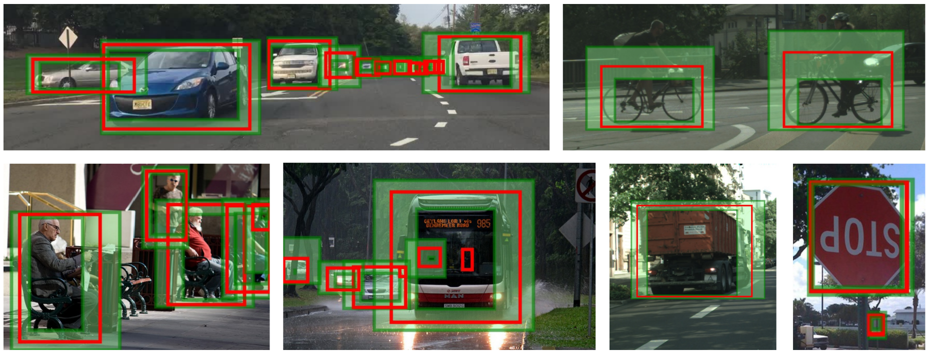

Fig. 1: Examples of our method for multiple classes on test images. True bounding boxes are in red, two-sided prediction interval regions are shaded in green. Produced uncertainty estimates come with a probabilistic coverage guarantee of the true boxes.

bounding boxes of previously unseen objects (of known classes). Specifically, we provide users with the following safety assurance: *"The conformal prediction interval covers the object's true bounding box with probability* $(1-\alpha)$ *for any known object class"*, where α is an acceptable margin of error. Such a guarantee can, *e.g.*, in the context of autonomous driving, help certify collision avoidance by steering clear of the outer interval bounds, or in the case of robot picking, enforce cautious handling by demarcating a reliable grasping zone via the inner bounds. We provide visual examples of our obtained intervals in Fig. 1 and § C.5.

Employing strategies based on ensembling and quantile regression, we ensure that the obtained intervals are adaptive to object size: they may grow or shrink in individual dimensions to account for object variability and prediction difficulty. A challenge to the desired assurance is that constructed intervals rely on the model's predicted class labels, which may be erroneous. We thus introduce an additional conformal step over the class labels, shielding against misclassification and ensuring that downstream coverage is satisfied. That is, our *two-step* conformal pipeline remains theoretically and empirically valid regardless of the underlying object detector's predictive performance for either class labels or box coordinates – the incurred costs are solely reflected in the obtained prediction interval sizes. In the experiments, we apply our methodology to multiple classes on several real-world 2D object detection datasets. We obtain bounding box prediction intervals that adhere to the desired guarantee, and are both adaptive and practically useful for downstream decision-making.

To summarize, our core contribution is an *end-to-end* framework for safe bounding box uncertainties which is *post-hoc*, efficient, and generalizable. In that process, we introduce several original concepts such as (i) ensemble and quantile CP adaptations for object detection, (ii) leveraging strong class-conditional guarantees for multi-class settings, and (iii) proposing a sequential two-step approach that propagates classification uncertainties forward.

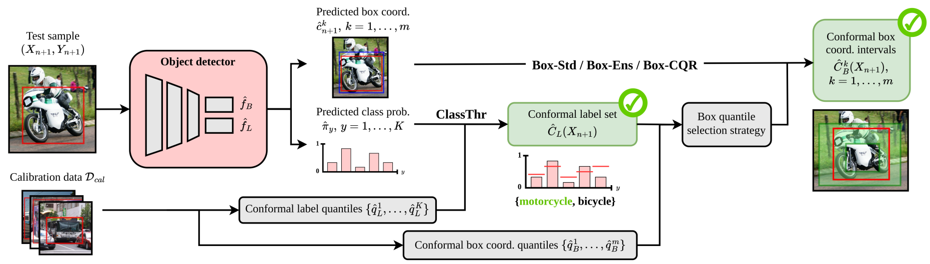

Fig. 2: A diagram of our proposed *two-step* conformal approach. We compute conformal quantiles for both class labels and box coordinates on calibration data following the CP framework. These are used on the predictions of a 'black-box' object detector for a new test sample to (1) form a conformal label set with guarantee (✓) which informs our box quantile choice, and (2) form a conformal prediction interval for the bounding box with guarantee (✓), providing a reliable predictive uncertainty estimate.

2 Background

We begin by providing background on conformal prediction and the desired coverage guarantees, and then relate our object detection setting to it.

2.1 Conformal prediction

We consider the most common setting of split CP [43], where we perform a single split to obtain hold-out calibration data $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^n \sim P_{XY}$, as opposed to alternative partitioning schemes [6,64]. If the general conformal procedure outlined in Algorithm 1 (deferred to § A.1) is followed, a coverage guarantee for an unseen test sample $(X_{n+1}, Y_{n+1}) \sim P_{XY}$ is provided in terms of a prediction set $\hat{C}(X_{n+1})$, where a finite-sample, distribution-free guarantee is given over the event of $\hat{C}(X_{n+1})$ containing Y_{n+1}.

That is, assuming the samples $\mathcal{D}_{cal} \cup \{(X_{n+1}, Y_{n+1})\}$ are exchangeable – a relaxation of the *i.i.d.* assumption – we obtain a probabilistic guarantee that

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha \tag{1}$$

for some tolerated miscoverage rate $\alpha \in (0, 1)$ [54]. The provided guarantee is *marginally* valid, since it holds on average across any sample (X_{n+1}, Y_{n+1}) and set $\mathcal{D}_{cal}$ drawn from some fixed distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. This is in contrast to the ideal scenario of *conditionally* valid coverage per input X_{n+1}, which has been shown to be impossible to achieve in a distribution-free manner [19,63]. However, recent work on in-between notions of conditionality such as *group-* [23,49] and *feature-conditional* [53] strive towards more granular guarantees.

In particular, *class-conditional* validity can be achieved by applying CP separately to samples from each class [11,52,55,65], yielding the following guarantee:

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) | Y_{n+1} = y) \geq 1 - \alpha \quad \forall y \in \mathcal{Y}, \tag{2}$$

where $\mathcal{Y} = \{1, \ldots, K\}$ are distinct class labels. The class-conditional guarantee in Eq. 2 is stronger and implies Eq. 1, in that we aim to control the miscoverage level for samples within *each class*. It also permits setting individual miscoverage levels $\{\alpha_y\}_{y \in \mathcal{Y}}$ per class if desired, and is robust to imbalances in class proportions [47,62]. Such a class-conditional guarantee is precisely what we aim to provide.

Classification and regression. Applying CP to a classification task yields conformal label *prediction sets* $\hat{C}_L(X_{n+1}) \subseteq \{1, \ldots, K\}$ as finite subsets of the K class labels, at a target miscoverage level α_L. For regression, the sets $\hat{C}_B(X_{n+1}) \subseteq \mathbb{R}$ take the form of *prediction intervals* (PIs) on the target domain, at a target miscoverage level α_B. Naturally, we have both $(\alpha_L, \alpha_B) \in (0, 1)^2$.

2.2 Object detection

We next formalize our multi-object detection setting. Consider an input image $X \in \mathbb{R}^{H \times W \times D}$, where H, W and D correspond to image height, width and depth. For each image in $\mathcal{D}_{cal}$ we also receive a set of tuples (c^1, c^2, c^3, c^4, l), where $(c^1, c^2, c^3, c^4) \in \mathbb{R}^4$ are the coordinates indicating an object's bounding box location in the image, and $l \in \{1, \ldots, K\}$ represents the object's class label.

Each tuple parameterizes an object, with a total of $O(X)$ true objects located in the image. For image X we thus have objects $\{(c^1, c^2, c^3, c^4, l)_j\}_{j=1}^{O(X)}$. Note that the model predicts $\hat{O}(X)$ objects, and it is possible that $O(X) \neq \hat{O}(X)$. We model every object as an individual sample for our CP procedures, *i.e.*, the same input image X can produce multiple calibration samples of shape $(X, (c^1, c^2, c^3, c^4, l)_j)$, where $j = 1, \ldots, O(X)$ denote the contained objects.

Object detection model. For our object detector $\hat{f}$, we define two separate output heads. The probabilistic classification head is defined as the map $\hat{f}_L : X \mapsto (\hat{\pi}_1, \ldots, \hat{\pi}_K)$, where $\hat{\pi}_y$ is the model's estimate of the true class probability π_y of some object in image X belonging to class y. The object's class label is then $l = \arg\max_{y \in \{1, \ldots, K\}} \hat{\pi}_y$. The bounding box regression head, denoted as $\hat{f}_B : X \mapsto (\hat{c}^1, \hat{c}^2, \hat{c}^3, \hat{c}^4)$, maps to an object's real-valued bounding box coordinates.

2.3 Conformal prediction for object detection

Given our multi-object detection setting, we consider a class-conditional CP approach to be particularly meaningful. It is sensible to only leverage information on detected objects of the same class, *e.g.*, class 'car', to construct PIs for new objects of that class. In contrast, a general marginal approach will unintuitively also employ information from unrelated classes, such as 'person' or 'bicycle'.

We apply CP to the bounding boxes on a per-coordinate basis, previously denoted (c^1, c^2, c^3, c^4). However, from now on let us consider the generalization to an arbitrary amount of coordinates c^k, $k = 1, \ldots, m$ [3]. If we consider the class label l within each group of objects belonging to a common class as fixed

[3] This easily permits extending our approach to higher-dimensional object parameterizations such as 3D bounding boxes.

(since the same label is shared), the response of an individual sample (X_i, Y_i) can be interpreted as a realization of the m coordinates only, *i.e.*, we define $Y_i := (c_i^1, \dots, c_i^m) \in \mathbb{R}^m$. The desired guarantee in Eq. 2 is re-interpreted as

$$\mathbb{P}\left(\bigcap_{k=1}^{m}\left(c_{n+1}^k \in \hat{C}_B^k(X_{n+1})\right) \mid l_{n+1} = y\right) \geq 1 - \alpha_B \quad \forall y \in \mathcal{Y}, \tag{3}$$

where components are indexed accordingly per specific coordinate dimension. For example, $\hat{C}_B^k(X_{n+1})$ is the k-th coordinate's prediction interval of an object of class y (its class label l_{n+1} matches y) located in image X_{n+1}. Applying CP per coordinate gives rise to multiple testing issues, which we address in § 4.1.

Practical limitations. Naively applying class-conditional CP to the box coordinates necessitates a correct class label prediction in order to satisfy validity. That is, for Eq. 3 to hold, a valid PI construction requires $\hat{l}_{n+1} = l_{n+1}$ for any considered class $y \in \mathcal{Y}$. We alleviate this practically limiting dependence on the model's classification ability using a conformal set-based classifier in § 5 (see also Fig. 2). However, we of course still rely on the model's general detection abilities: the provided guarantees only hold for true objects that are actually detected (true positives) and do not account for undetected objects (false negatives), as also noted by Andéol *et al.* [1,15]. Finally, the assumption on data exchangeability underlying CP requires P_{XY} to remain fixed, albeit recent works have explored CP under settings of mild or known distribution shifts [2,18].

3 Related work

Many existing approaches for uncertainty in bounding box regression leverage standard UQ techniques such as Bayesian inference [13,25,71], loss attenuation [22,29,30], or practical approximations like Monte Carlo Dropout [21,39,45,74] and Deep Ensembles [16,40,68]. These can require substantial modifications to the model architecture or training procedure, and do not provide a guarantee or statement of assurance about provided estimation quality. See Feng *et al.* [17] for a recent survey. A complementary line of work investigates the calibration of object detectors [27,42,46], which can benefit our approach by improving the underlying 'black-box' probabilistic model.

Conformal approaches have recently gained traction for computer vision and related tasks, with applications such as image classification [5,51], geometric pose estimation [73], or tracking and trajectory planning [34,35,41,58]. Yet, the domain remains comparatively unexplored given current surveys [2,18]. Specific attempts at principled UQ for bounding boxes include using the Probably Approximately Correct (PAC) framework to produce guarantees by composition of PAC sets at multiple modelling stages [32,44], or leveraging p-values and risk estimates obtained from concentration inequalities for related risk control [3,4,7]. Such works differ from our two-step conformal approach in several ways, such as (i) considering different vision tasks and using different data modalities, (ii) integrating CP into complex modelling pipelines that cease to be *post-hoc* and model-agnostic, or (iii) employing methods not based on conformal prediction.

Closest prior work. Conformal PIs for bounding boxes have been previously considered by Andéol *et al.* [1, 15]. However, a crucial limitation in their approach is that bounding box uncertainty is considered and evaluated for a *single class* and for *correctly classified* objects only. Thus only the simplest form of our guarantee in Eq. 3 is provided, since the class label is known *a priori* and therefore $\hat{l}_{n+1} = l_{n+1}$ trivially holds. This means that prevalent uncertainty in the class label predictions (which we address in § 5) is entirely ignored, making their approach unsuitable for settings with multiple interacting classes, such as autonomous driving. We also introduce several methodological improvements, such as (i) novel ensemble and quantile scoring functions for the bounding box setting, (ii) more informative two-sided intervals, and (iii) a multiple testing correction that exploits correlation structure between box coordinates, as opposed to more naive Bonferroni [15] or max-corrections [1].

4 Conformal methods for box coordinates

A key modelling decision in CP is the choice of scoring function $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ to compute the required nonconformity scores (see § A.1). We consider three choices of scoring function and PI construction for each box coordinate $k \in \{1, \dots, m\}$, which we outline next. Additional implementation details can be found in § B.

Standard conformal (Box-Std). We firstly consider the standard approach of employing regression residuals $s(\hat{f}_B(X), Y) = |\hat{c}^k - c^k|$ as scores [54]. The resulting PIs are constructed as $\hat{C}^k_B(X_{n+1}) = [\hat{c}^k_{n+1} - \hat{q}^k_B,\ \hat{c}^k_{n+1} + \hat{q}^k_B]$, where $\hat{q}^k_B$ denotes the computed conformal quantile for the k-th coordinate. While straightforward, this construction only permits for non-adaptive, fixed-width intervals. Andéol *et al.* [15] use this approach to construct their one-sided intervals.

Conformal ensemble (Box-Ens). In order to produce more adaptive intervals, we next consider normalized residual scores [31] of the form $s(\hat{f}_B(X), Y) = |\hat{c}^k - c^k| / \hat{\sigma}(X)$, where $\hat{\sigma}$ is some form of heuristic uncertainty estimate (*i.e.*, without guarantees) obtained from the underlying model. The resulting conformal PIs are constructed as $\hat{C}^k_B(X_{n+1}) = [\hat{c}^k_{n+1} - \hat{\sigma}(X_{n+1})\, \hat{q}^k_B,\ \hat{c}^k_{n+1} + \hat{\sigma}(X_{n+1})\, \hat{q}^k_B]$. By incorporating model uncertainty, the intervals can be re-scaled individually per coordinate to adapt their magnitude at test time. We can interpret this as an empirical conditioning on the particular test sample. We employ an ensemble of object detectors and quantify $\hat{\sigma}$ as the standard deviation of the ensemble's box coordinate predictions [28]. A joint coordinate prediction $\hat{c}^k$ is obtained from the ensemble via confidence-weighted box fusion [56].

Conformal quantile regression (Box-CQR). As an alternative adaptive method, we extend the approach of Conformal Quantile Regression (CQR) [50] to our setting. Additional regression heads $\hat{Q}_B$ are trained with a *quantile loss* alongside $\hat{f}_B$ to produce lower and upper quantile predictions $\hat{Q}_{\alpha_B/2}$ and $\hat{Q}_{1-\alpha_B/2}$ for the bounding box coordinates. Under regularity conditions, these predictors will asymptotically converge to the true conditional quantiles [12, 24], motivating their viability. Following CQR, we define the scores as $s(\hat{Q}_B(X), Y) = \max\{\hat{Q}_{\alpha_B/2}(X) - c^k,\ c^k - \hat{Q}_{1-\alpha_B/2}(X)\}$, and construct the conformal PIs as

$\hat{C}^k_B(X_{n+1}) = [\hat{Q}_{\alpha_B/2}(X_{n+1}) - \hat{q}^k_B,\ \hat{Q}_{1-\alpha_B/2}(X_{n+1}) + \hat{q}^k_B]$. The obtained interval ensures adaptivity through the use of its quantile predictions, which will differ in their distance relative to the mean coordinate prediction $\hat{c}^k_{n+1}$ per sample.

4.1 Multiple testing correction

Applying CP to each of the box coordinates $k = 1, \dots, m$ separately gives rise to multiple testing issues, since the conformal procedure can be interpreted from a hypothesis testing view as running m permutation tests on nonconformity in parallel [55,65]. This results in a guaranteed coverage of at most $(1 - m \cdot \alpha_B)$, as opposed to the desired rate of $(1 - \alpha_B)$ (see § A.3 for further details).

The naive Bonferroni correction [67] offers a possible remedy, since choosing $\alpha'_B = \alpha_B/m$ will satisfy target coverage. However, it is known to be overly conservative under positive dependency of the individual hypothesis [66], a reasonable assumption given that the coordinates parametrize an object's bounding box jointly. In fact, Bates *et al.* [8] assert that a set of conformal p-values exhibits positive dependency structure *a priori* since they are jointly *positive regression dependent on a subset* [9]. We leverage an alternate procedure by Timans *et al.* [61], which exploits correlation structure among box coordinates for a less conservative correction. Their max-rank procedure adapts the Westfall & Young [70] permutation correction to make it suitable for the setting of conformal prediction (see § A.4). While the use of a max-correction has been previously considered for CP [1,11], max-rank operates in the more robust scale-invariant rank space. In addition, it requires less compute than previously proposed copula-based testing corrections [38,60].

5 Class label prediction sets

In practice, the object detector may incorrectly predict an object's class label given our multi-class setting. This complicates a direct application of class-conditional CP to an object's bounding box at test time, since we need to correctly select the conformal quantiles $\hat{q}^k_B$, $k \in \{1, \dots, m\}$ to construct bounding box intervals that satisfy Eq. 3. This limits intially provided safety assurances to correctly classified objects only, *i.e.*, those where $\hat{l}_{n+1} = l_{n+1}$ successfully match.

To alleviate this restrictive dependence on the model's classification ability, our modelling pipeline in Fig. 2 introduces an additional conformal step which preceeds the bounding box construction. Specifically, we consider applying CP to the model's classifier head $\hat{f}_L$ to first generate class label prediction sets $\hat{C}_L(X_{n+1})$ with a guarantee on label containment. These are subsequently used to select our box coordinate quantiles, ensuring the validity of our provided guarantees in Eq. 3 is broadened to even include *incorrectly* classified objects.

We achieve this using another class-conditional CP approach on the class labels with a strict label coverage guarantee of 99% (*i.e.*, $\alpha_L = 0.01$). Thus, we approximate the condition $\hat{l}_{n+1} = l_{n+1}$ by effectively ensuring $l_{n+1} \in \hat{C}_L(X_{n+1})$. The resulting *two-step* sequential approach maintains validity regardless of the object detector's classification or bounding box regression performance. The incurred costs are reflected in the obtained prediction set and interval sizes only. Our experiments in § 6.3 demonstrate that even under these strong safety assurances, our approach provides actionably tight PIs. We follow with a description of the employed conformal label set method and related baselines.

5.1 Conformal class thresholding (ClassThr)

We propose using a class-conditional variant of the prediction set classifier introduced by Sadinle *et al.* [52], based on a similar conformal procedure as for the bounding boxes. Given our probabilistic classifier head $\hat{f}_L$, we define the scoring function $s(\hat{f}_L(X), y) = 1 - \hat{\pi}_y(X)$ for every class $y \in \mathcal{Y}$ to compute per-class conformal quantiles $\hat{q}_L^y$. The class label prediction set for a new object to classify is then given by $\hat{C}_L(X_{n+1}) = \{y \in \mathcal{Y} : \hat{\pi}_y(X_{n+1}) \geq 1 - \hat{q}_L^y\}$. The detailed procedure is given in Algorithm 2, deferred to § A.2. Importantly, class-conditional validity is ensured by comparing each class probability against its class-specific threshold on set inclusion. The class-conditional guarantee stated in Eq. 2 can now be provided similarly for the object detector's classification task as

$$\mathbb{P}(l_{n+1} \in \hat{C}_L(X_{n+1}) \,|\, l_{n+1} = y) \geq 1 - \alpha_L \quad \forall y \in \mathcal{Y}. \tag{4}$$

Impact on bounding box coverage guarantee. A class-conditional guarantee is enforced for the label prediction sets (Eq. 4), since only imposing the weaker marginal guarantee could invalidate the subsequent class-conditional box guarantee (Eq. 3). If for instance a class is systematically undercovered, we would fail to retrieve the correct box quantiles for some of its associated objects, propagating the undercoverage down-stream. Our approach instead enforces guarantees of equivalent strength. Observe that we perform two distinct conformal procedures in sequence, rendering the coverage guarantees conditionally independent. Thus the down-stream coverage effect for their two-step application is that

$$\begin{aligned}
&\mathbb{P}\left(l_{n+1} \in \hat{C}_L(X_{n+1}) \wedge \bigcap_{k=1}^{m} \left(c_{n+1}^k \in \hat{C}_B^k(X_{n+1})\right) \mid l_{n+1} = y\right) \\
&= \mathbb{P}\left(l_{n+1} \in \hat{C}_L(X_{n+1}) \,|\, l_{n+1} = y\right) \cdot \mathbb{P}\left(\bigcap_{k=1}^{m} \left(c_{n+1}^k \in \hat{C}_B^k(X_{n+1})\right) \mid l_{n+1} = y\right) \\
&\qquad \geq (1 - \alpha_L)(1 - \alpha_B) \quad \forall y \in \mathcal{Y}.
\end{aligned} \tag{5}$$

That is, a preceeding label coverage guarantee of $(1 - \alpha_L)$ will nominally only assure subsequent box coverage of $(1 - \alpha_L)(1 - \alpha_B)$. In our experiments we approximate $(1-\alpha_L)(1-\alpha_B) \approx (1-\alpha_B)$ by setting $\alpha_L = 0.01$, thus alleviating the down-stream coverage reduction[4]. Eq. 5 highlights that coverage trade-offs between objectives are possible depending on application-specific requirements. For example, a nominal box coverage of 90% can be achieved by choosing either $\alpha_L = 0.05, \alpha_B = 0.05$ or $\alpha_L = 0, \alpha_B = 0.1$.

[4] As we observe in § 6.3, enforcing $\alpha_L = 0$ leads to empirically inefficient box intervals.

	Label set		Box interval	
Method	Guarantee	Size	Guarantee	Size
Top	✗*	Single	✗†	Small
Naive	✗*	Small	✗†	Small
ClassThr	✓	Medium	✓	Medium
Full	✓	Large	✓	Large

Table 1: Provided *nominal* coverage guarantees and expected *empirical* prediction set/interval sizes for the considered label prediction set methods on the basis of both correctly and incorrectly classified objects. *Top and Naive provide a label guarantee 'for free' if $(1-\alpha_L)$ is below the classifier's accuracy level *for each class*. †Top and Naive provide a box guarantee for correctly classified objects only. Naive may also satisfy both guarantees under practically unattainable perfect model calibration.

5.2 Bounding box quantile selection

The obtained label prediction sets are subsequently used to select a valid box coordinate quantile for any interval construction following § 4. A natural quantile selection strategy is $\hat{q}_B^k = \max\{\hat{q}_B^{k,y}\}_{y\in\hat{C}_L(X_{n+1})}\,\forall k \in \{1,\ldots,m\}$, where $\hat{q}_B^{k,y}$ is the quantile of the k-th coordinate for class y. Using a max-operator on the label set is a valid but conservative approach, since all labels in the set are regarded as equally likely for every sample. Obtained box intervals thus tend to overcover, which could perhaps be alleviated with a different strategy, resulting in narrower PIs. Yet, we find even this straightforward selection to yield reasonably tight results. A hypothesis testing motivation for its use can be found in § A.5.

5.3 Label set baselines

We compare obtained label predictions sets via conformal class thresholding (ClassThr) to several reasonable alternatives, whose nominal guarantees and expected empirical set sizes are outlined in Tab. 1, and which we detail next.

Top singleton set (Top). We return label prediction sets that only consist of the highest probability class for every sample, *i.e.*, $\hat{C}_L(X_{n+1}) = \{y^* : \hat{\pi}_{y^*}(X_{n+1}) = \max_{y\in\{1,\ldots,K\}} \hat{\pi}_y(X_{n+1})\}$. This approach returns singleton sets, is void of nominal guarantees, and its empirical coverage relies fully on the classifier's accuracy. The distinction to our initial condition $\hat{l}_{n+1} = l_{n+1}$, which we refer to as **Oracle**, is subtle: instead of ensuring correct quantile selection, we permit the use of potentially wrong quantiles to construct the box intervals.

Density level set (Naive). Assuming a perfectly calibrated classifier such that $\hat{\pi}_y(X) = \pi_y(X)\ \forall y \in \mathcal{Y}$, the optimal prediction set is provided by density level sets. That is, we collect all labels sorted by descending $\hat{\pi}_y$ until we reach probability mass $(1-\alpha_L)$. Under this assumption, prediction sets will also approach conditional coverage for any $X \in \mathcal{X}$ [51,52]. While unattainable in practical settings where the classifier tends to be miscalibrated (*i.e.*, $\hat{\pi}_y(X) \neq \pi_y(X)$), it can be considered a theoretically motivated extension of the Top baseline.

Full domain set (Full). We consider taking the full set of possible class labels per sample, thus $\hat{C}_L(X_{n+1}) = \mathcal{Y}$ and $|\hat{C}_L(X_{n+1})| = K$. In combination with our quantile selection strategy this approach guarantees label coverage conditionally per sample, *i.e.*, it ensures $\alpha_L = 0$. However, this comes at the cost of overly inflated label sets whose size is expected to propagate to the box intervals.

We do not consider other popular conformal approaches for classification such as APS [51] or RAPS [5] since they aim to empirically improve conditional coverage under the requirements of a *marginal* guarantee – these advantages do not extend to a class-conditional setting as ours.

6 Experiments

For our experiments we primarily rely on pre-trained object detectors from `detectron2` [72], based on a Faster R-CNN architecture and trained on COCO [33]. We consider three datasets: COCO validation, Cityscapes [14] and BDD100k [75], which contain 2D bounding box annotations and are split into appropriate calibration and test sets. We run our two-step conformal procedure for a variety of classes, but focus reported results on a coherent set of object classes which exists across datasets: *person, bicycle, motorcycle, car, bus* and *truck* (see § B.3).

Since the images can contain multiple objects, we require a pairing mechanism between true and predicted bounding boxes. Following prior work [1, 15] we perform Hungarian matching [26] based on an intersection-over-union (IoU) threshold of 0.5. Throughout, we set $\alpha_L = 0.01, \alpha_B = 0.1$ for a target box coverage of $(1 - \alpha_L)(1 - \alpha_B) \approx 90\%$, and employ max-rank [61] for multiple testing correction. Results are averaged across multiple trials of data splitting. Additional results, including varying combinations of (α_L, α_B), are in § C [5].

6.1 Metrics

Our approaches are validated by assessing the key desiderata of CP via relevant metrics described below, which jointly capture the desired notions of 'reliable' uncertainty [2]. We denote the test set of size n_t as $\mathcal{D}_{test} = \{(X_j, Y_j)\}_{j=n+1}^{n+n_t}$.

Validity. We assess if nominal coverage guarantees are satisfied by verifying *empirical coverage*, which we define in generality as

$$Cov = \frac{1}{n_t} \sum_{j=n+1}^{n+n_t} \mathbb{1}[Y_j \in \hat{C}(X_j)], \tag{6}$$

where $\mathbb{1}[\cdot]$ is the indicator function, of form $\mathbb{1}[l_j \in \hat{C}_L(X_j)]$ for label prediction sets and $\mathbb{1}[\bigcap_{k=1}^{m}(c_j^k \in \hat{C}_B^k(X_j))]$ for box intervals. Note that Cov is a random quantity parametrized by an empirical coverage distribution, and will deviate from nominal coverage based on factors such as calibration set size $|\mathcal{D}_{cal}|$ [63].

[5] Our code is publicly available at https://github.com/alextimans/conformal-od.

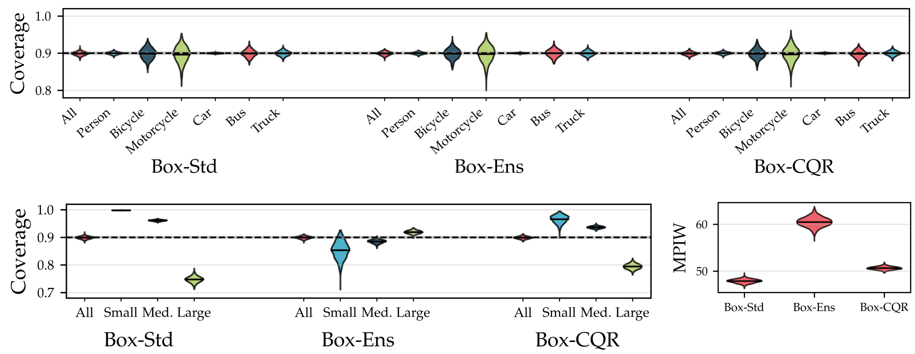

Fig. 3: *Top:* Empirical coverage levels marginally across all objects (All) and across objects from selected classes for the three bounding box methods (§ 4) on the BDD100k dataset. Target coverage is achieved both marginally and for individual classes. *Bottom:* Coverage levels are stratified by object size (Small, Medium, Large), showing that Box-CQR and in particular Box-Ens provide a more balanced empirical coverage across sizes. However, this comes at the cost of slightly larger intervals, as seen when comparing $MPIW$. We also visualize target coverage (**- - -**) and the marginal coverage distribution (▬). Displayed densities are results obtained over 1000 trials.

Adaptivity. To examine if target coverage is satisfied by an imbalance of over- and undercoverage across objects, similarly to [5,51] we verify empirical coverage by stratification, namely across object sizes. We follow the COCO challenge [6] and stratify across three sizes by bounding box surface area: small (Cov_S, area $\leq 32^2$), medium (Cov_M, area $\in (32^2, 96^2]$) and large (Cov_L, area $> 96^2$).

Efficiency. Obtained conformal prediction sets and intervals are desired to be as small as possible while still maintaining target coverage (*i.e.*, remaining valid). We define the *mean set size* for label prediction sets and *mean prediction interval width (MPIW)* for bounding box prediction intervals as

$$\frac{1}{n_t} \sum_{j=n+1}^{n+n_t} |\hat{C}_L(X_j)| \quad \text{and} \quad \frac{1}{n_t m} \sum_{j=n+1}^{n+n_t} \sum_{k=1}^{m} |\hat{C}_B^k(X_j)|. \tag{7}$$

That is, *mean set size* denotes the average number of labels in the obtained sets, while $MPIW$ expresses the average interval width in terms of image pixels.

Predictive performance. We also follow standard practice and validate model performance using object detection-specific metrics from the COCO challenge, in particular average precision across multiple IoU thresholds (see § C).

6.2 Comparison of bounding box methods

Empirical coverage levels stratified by class labels as well as object sizes for the three proposed bounding box methods are displayed in Fig. 3 for BDD100k.

[6] See `https://cocodataset.org/#detection-eval`.

Uncertainty method	Object detector	Two-sided box intervals *MPIW*	Two-sided box intervals *Cov*	One-sided box intervals *MPIW*	One-sided box intervals *Cov*
DeepEns	5× Faster R-CNN	12.31 ± 0.47	0.21 ± 0.01	74.15 ± 2.01	0.49 ± 0.01
GaussianYOLO	YOLOv3	7.00 ± 0.14	0.08 ± 0.01	87.07 ± 4.25	0.35 ± 0.01
Andéol *et al.* (Best)	Faster R-CNN	N/A		87.62 ± 1.79	0.91 ± 0.01
	YOLOv3	N/A		107.93 ± 4.85	0.92 ± 0.02
	DETR	N/A		82.21 ± 1.64	0.90 ± 0.01
	Sparse R-CNN	N/A		79.35 ± 1.78	0.91 ± 0.01
Box-Std (Ours)	Faster R-CNN	55.47 ± 2.97	0.88 ± 0.02	85.42 ± 1.99	0.88 ± 0.02
	YOLOv3	61.73 ± 3.66	0.88 ± 0.02	103.12 ± 3.95	0.88 ± 0.02
	DETR	45.34 ± 3.33	0.88 ± 0.02	80.57 ± 1.78	0.88 ± 0.01
	Sparse R-CNN	41.92 ± 2.16	0.89 ± 0.01	77.33 ± 1.72	0.89 ± 0.01

Table 2: We compare our simplest method **Box-Std** to Andéol *et al.*'s best results (see § B.5) across different object detectors (Faster R-CNN [72], YOLOv3 [48], DETR [10], Sparse R-CNN [59]) as well as deep ensembles (DeepEns [28]) and GaussianYOLO [13], two popular UQ approaches. The former satisfies coverage but is only designed for one-sided intervals, while the latter can heavily undercover in practice (marked ▬). Results are for COCO across classes and 100 trials, for target coverage $(1 - \alpha_B) = 0.9$. The key difference between various object detectors are the obtained interval widths ($MPIW$), which relate to predictive performance and are smaller for better models.

We see that target coverage of 90% is satisfied even per class, validating the class-conditional guarantees provided in Eq. 3. The visible coverage variations are explained by the differences in available calibration samples per class (see § A.6 and Tab. 3). We further observe that the fixed-width intervals of Box-Std may be large enough to cover small objects, but will fail to account for the magnitude of large ones, resulting in significant undercoverage. In contrast, the adaptive nature of Box-CQR and in particular Box-Ens via its scaling factor can better account for varying magnitudes, achieving higher coverage for large objects at a slight loss in efficiency due to a higher $MPIW$. Whilst coverage across small objects reduces somewhat, it now intuitively aligns with observed prediction difficulty (see Tab. 6). That is, objects which are more challenging to predict exhibit a higher variation and chance of miscoverage. We note that the improved coverage balance across object sizes is a purely empirical benefit of our adaptive designs – the employed conformal procedures only aim to guarantee target coverage per class and do *not* condition on object size.

Baseline comparisons. We further validate our conformal bounding box step by comparing to Andéol *et al.* [1,15], modifying our own two-sided interval methods to produce one-sided PIs, and evaluating efficiency via $MPIW$ (see § B.5 for details and their *box stretch* metric). Tab. 2 demonstrates that we achieve marginally tighter intervals even in their own, more restricted setting, while remaining equally valid. We also evaluate generated uncertainties via deep ensembles (DeepEns) [28] and GaussianYOLO [13], two popular UQ approaches for object detection. Results confirm the unreliability of produced uncertainties due to their *lack of guarantees*, as seen by severe undercoverage in both settings.

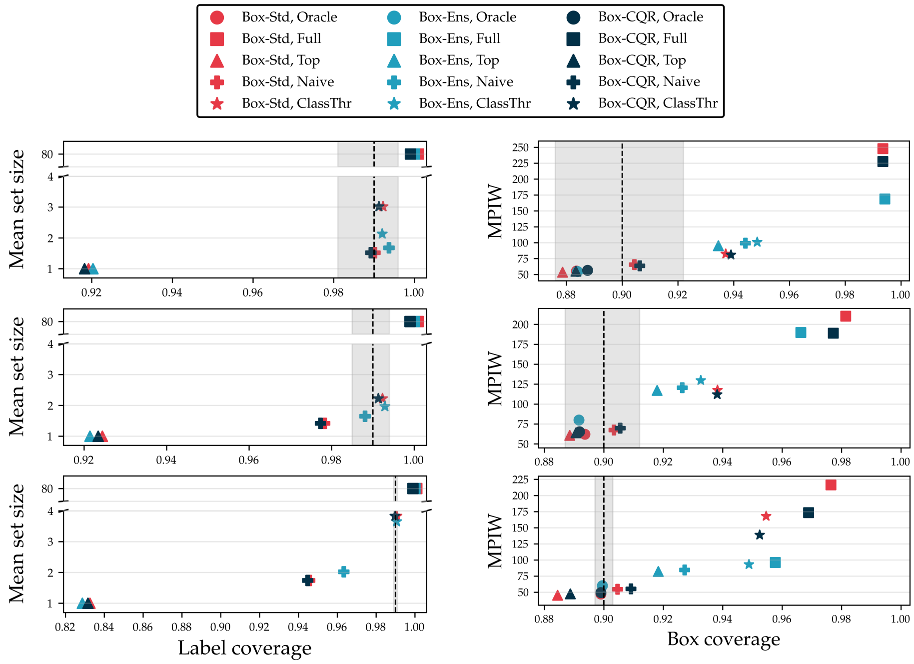

Fig. 4: Every combination of conformal label set and bounding box method is evaluated along two axes for COCO (*top row*), Cityscapes (*middle row*) and BDD100k (*bottom row*). On the vertical axis we display efficiency, *i.e.*, *mean set size* for label sets (*left column*) and $MPIW$ for box intervals (*right column*). On the horizontal axis we display empirical coverage levels. We also draw target coverage (- - -) and marginal coverage distributions ( ). In line with Tab. 1, approaches employing ClassThr or Full consistently achieve both label and box target coverage, at the cost of larger prediction sets/intervals. Results are averaged across classes and 100 trials.

Finally, a comparison of the max-rank correction to Bonferroni in § C.2 asserts that substantially tighter PIs can be obtained with our employed correction.

6.3 Results for the two-step approach

After having benchmarked our bounding box step, we next compare the full *two-step* approach by adding the preceeding conformal step for class labels via ClassThr, and comparing to the proposed label set baselines in § 5.3. Each method's nominal guarantee and expected efficiency is displayed in Tab. 1.

In line with expectations, we observe in Fig. 4 that only approaches using ClassThr or Full (the full label set) consistently achieve target coverage for both class labels and box intervals across all three datasets. Notably, while Full results in overly inflated interval widths due to its construction, ClassThr provides surprisingly efficient label sets (with *mean set size* ≤ 4) which propagate into reasonably tight box intervals. Differences also exist in the efficiency of the three bounding box methods, with Box-Ens performing notably better on BDD100k. Stratifying results by object (mis-)classification also confirms that wrongly classified objects tend to exhibit higher uncertainty (see Fig. 8). Oracle relies on knowing the correct quantile and thus does not require generating a label set. While it provides nominal guarantees assuming correct label prediction and empirically satisfies them with high efficiency (*i.e.*, small $MPIW$), the condition severely limits its practicality. Top consistently undercovers the true label, and also tends towards undercovering boxes as sample sizes increase. Naive is able to consistently maintain box coverage even though label coverage is violated, with surprisingly tight PIs. However, additional experiments in § C.1 showcase its sensitivity to model miscalibration, yielding it less robust than ClassThr. Yet, it is interesting that methods such as Top and Naive perform quite well empirically, even under void nominal guarantees.

Discussion and practitioner's choice. We end by discussing the choices in selecting a suitable conformal *two-step* approach a practitioner may want to consider. In terms of label set methods, we suggest the following: if the model classifier records a high accuracy and there exists a way to externally validate predicted class labels, then Top may be a highly efficient choice. If the model is strongly calibrated and only empirical assurances are sufficient, then Naive may be a suitable selection. However, if strong safety assurances with both nominal and empirical guarantees are desired, then ClassThr is the only safe choice. We highlight that improvement potential remains: obtained PIs using ClassThr tend to overcover, presumably due to our conservative quantile selection strategy. It may be indeed possible to obtain the same set of safety assurances with higher efficiency under a different strategy. For example, one may consider a weighted quantile construction on the basis of the classifier's confusion matrix.

Regarding the choice of bounding box method, we observe that Box-Std will be the most efficient if the object detection task contains only similarly sized objects. However, if objects vary substantially in size, an adaptive approach such as Box-Ens or Box-CQR will be more suitable. One may also consider designing a conformal approach which explicitly satisfies target coverage across other reasonable strata beyond classes, such as object sizes or shapes. A limiting factor to consider may be the size of available calibration data per partition.

7 Conclusion

We present a novel procedure to quantify predictive uncertainty for multi-object detection. We leverage CP to generate uncertainty intervals with a per-class coverage guarantee for new samples. Our proposed *two-step* conformal approach provides adaptive bounding box intervals with safety assurances robust to object misclassification. Addressing similar types of guarantees, the procedure can be extended to 3D bounding boxes, object tracking and other detection tasks in future work. Whilst improvements can be made to achieve even narrower intervals, our results are promisingly tight, paving the way for a safer deployment of vision-based systems in scenarios involving decision-making under uncertainty.

Acknowledgements

We thank members of the Bosch-UvA Delta Lab and anonymous reviewers for helpful discussions and feedback. This project was generously supported by the Bosch Center for Artificial Intelligence.

References

1. Andéol, L., Fel, T., De Grancey, F., Mossina, L.: Conformal prediction for trustworthy detection of railway signals. AI and Ethics (2024)
2. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning (2023)
3. Angelopoulos, A.N., Bates, S., Candès, E.J., Jordan, M.I., Lei, L.: Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. arXiv Preprint (arXiv:2110.01052) (2021)
4. Angelopoulos, A.N., Kohli, A.P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., Romano, Y.: Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging. International Conference on Machine Learning (2022)
5. Angelopoulos, A.N., Bates, S., Jordan, M., Malik, J.: Uncertainty sets for image classifiers using conformal prediction. International Conference on Learning Representations (2020)
6. Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: Predictive inference with the jackknife+. The Annals of Statistics (2021)
7. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. Journal of the ACM (JACM) (2021)
8. Bates, S., Candès, E., Lei, L., Romano, Y., Sesia, M.: Testing for outliers with conformal p-values. The Annals of Statistics (2023)
9. Benjamini, Y., Yekutieli, D.: The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics (2001)
10. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. European Conference on Computer Vision (2020)
11. Cauchois, M., Gupta, S., Duchi, J.C.: Knowing what You Know: Valid and validated confidence sets in multiclass and multilabel prediction. Journal of Machine Learning Research (2021)
12. Chaudhuri, P.: Global nonparametric estimation of conditional quantile functions and their derivatives. Journal of Multivariate Analysis (1991)
13. Choi, J., Chun, D., Kim, H., Lee, H.J.: Gaussian Yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. International Conference on Computer Vision (2019)
14. 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. Conference on Computer Vision and Pattern Recognition (2016)
15. de Grancey, F., Adam, J.L., Alecu, L., Gerchinovitz, S., Mamalet, F., Vigouroux, D.: Object Detection with Probabilistic Guarantees: A Conformal Prediction Approach. Computer Safety, Reliability, and Security (SAFECOMP) Workshops (2022)

16. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems (2020)
17. Feng, D., Harakeh, A., Waslander, S.L., Dietmayer, K.: A review and comparative study on probabilistic object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems (2021)
18. Fontana, M., Zeni, G., Vantini, S.: Conformal prediction: a unified review of theory and new challenges. Bernoulli (2023)
19. Foygel Barber, R., Candès, E.J., Ramdas, A., Tibshirani, R.J.: The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA (2020)
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.: On calibration of modern neural networks. International Conference on Machine Learning (2017)
21. Harakeh, A., Smart, M., Waslander, S.L.: Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. IEEE International Conference on Robotics and Automation (2020)
22. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. Conference on Computer Vision and Pattern Recognition (2019)
23. Jung, C., Noarov, G., Ramalingam, R., Roth, A.: Batch Multivalid Conformal Prediction. International Conference on Learning Representations (2023)
24. Koenker, R., Bassett, G.: Regression Quantiles. Econometrica (1978)
25. Kraus, F., Dietmayer, K.: Uncertainty estimation in one-stage object detection. IEEE Intelligent Transportation Systems Conference (2019)
26. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly (1955)
27. Kuppers, F., Kronenberger, J., Shantia, A., Haselhoff, A.: Multivariate confidence calibration for object detection. Conference on Computer Vision and Pattern Recognition Workshops (2020)
28. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems (2017)
29. Le, M.T., Diehl, F., Brunner, T., Knoll, A.: Uncertainty estimation for deep neural object detectors in safety-critical applications. IEEE International Conference on Intelligent Transportation Systems (2018)
30. Lee, Y., Hwang, J.w., Kim, H.I., Yun, K., Kwon, Y., Bae, Y., Hwang, S.J.: Localization uncertainty estimation for anchor-free object detection. European Conference on Computer Vision (2022)
31. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association (2018)
32. Li, S., Park, S., Ji, X., Lee, I., Bastani, O.: Towards PAC Multi-Object Detection and Tracking. arXiv Preprint (arXiv:2204.07482) (2022)
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. European Conference on Computer Vision (2014)
34. Lindemann, L., Cleaveland, M., Shim, G., Pappas, G.J.: Safe planning in dynamic environments using conformal prediction. IEEE Robotics and Automation Letters (2023)

35. Luo, R., Zhao, S., Kuck, J., Ivanovic, B., Savarese, S., Schmerling, E., Pavone, M.: Sample-efficient safety assurances using conformal prediction. International Workshop on the Algorithmic Foundations of Robotics (2022)
36. Lütjens, B., Everett, M., How, J.P.: Safe reinforcement learning with model uncertainty estimates. International Conference on Robotics and Automation (2019)
37. McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., Weller, A.: Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. International Joint Conference on Artificial Intelligence (2017)
38. Messoudi, S., Destercke, S., Rousseau, S.: Copula-based conformal prediction for multi-target regression. Pattern Recognition (2021)
39. Miller, D., Dayoub, F., Milford, M., Sünderhauf, N.: Evaluating merging strategies for sampling-based uncertainty techniques in object detection. International Conference on Robotics and Automation (2019)
40. Miller, D., Sünderhauf, N., Zhang, H., Hall, D., Dayoub, F.: Benchmarking sampling-based probabilistic object detectors. Conference on Computer Vision and Pattern Recognition Workshops (2019)
41. Muthali, A., Shen, H., Deglurkar, S., Lim, M.H., Roelofs, R., Faust, A., Tomlin, C.: Multi-agent reachability calibration with conformal prediction. IEEE Conference on Decision and Control (2023)
42. Neumann, L., Zisserman, A., Vedaldi, A.: Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. NeurIPS Workshop on Machine Learning for Intelligent Transportation Systems (2018)
43. Papadopoulos, H., Vovk, V., Gammerman, A.: Conformal Prediction with Neural Networks. 19th IEEE International Conference on Tools with Artificial Intelligence (2007)
44. Park, S., Bastani, O., Matni, N., Lee, I.: Pac confidence sets for deep neural networks via calibrated prediction. International Conference on Learning Representations (2020)
45. Peng, L., Wang, H., Li, J.: Uncertainty evaluation of object detection algorithms for autonomous vehicles. Automotive Innovation (2021)
46. Phan, B., Salay, R., Czarnecki, K., Abdelzad, V., Denouden, T., Vernekar, S.: Calibrating uncertainties in object localization task. NeurIPS Workshop on Bayesian Deep Learning (2018)
47. Podkopaev, A., Ramdas, A.: Distribution-free uncertainty quantification for classification under label shift. Uncertainty in Artificial Intelligence (2021)
48. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv Preprint (arXiv:1804.02767) (2018)
49. Romano, Y., Barber, R.F., Sabatti, C., Candès, E.J.: With Malice Towards None: Assessing Uncertainty via Equalized Coverage. Harvard Data Science Review (2020)
50. Romano, Y., Patterson, E., Candes, E.: Conformalized Quantile Regression. Advances in Neural Information Processing Systems (2019)
51. Romano, Y., Sesia, M., Candes, E.: Classification with Valid and Adaptive Coverage. Advances in Neural Information Processing Systems (2020)
52. Sadinle, M., Lei, J., Wasserman, L.: Least Ambiguous Set-Valued Classifiers With Bounded Error Levels. Journal of the American Statistical Association (2019)
53. Sesia, M., Romano, Y.: Conformal Prediction using Conditional Histograms. Advances in Neural Information Processing Systems (2021)
54. Shafer, G., Vovk, V.: A Tutorial on Conformal Prediction. Journal of Machine Learning Research (2008)

55. Shi, F., Ong, C.S., Leckie, C.: Applications of class-conditional conformal predictor in multi-class classification. International Conference on Machine Learning and Applications (2013)
56. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing (2021)
57. Steinwart, I., Christmann, A.: Estimating conditional quantiles with the help of the pinball loss. Bernoulli (2011)
58. Su, S., Han, S., Li, Y., Zhang, Z., Feng, C., Ding, C., Miao, F.: Collaborative multi-object tracking with conformal uncertainty propagation. IEEE Robotics and Automation Letters (2024)
59. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse R-CNN: End-to-end object detection with learnable proposals. Conference on Computer Vision and Pattern Recognition (2021)
60. Sun, S.H., Yu, R.: Copula conformal prediction for multi-step time series prediction. International Conference on Learning Representations (2023)
61. Timans, A., Straehle, C.N., Sakmann, K., Nalisnick, E.: A powerful rank-based correction to multiple testing under positive dependency. arXiv Preprint (arXiv:2311.10900) (2023)
62. Toccaceli, P., Gammerman, A.: Combination of inductive mondrian conformal predictors. Machine Learning (2019)
63. Vovk, V.: Conditional Validity of Inductive Conformal Predictors. Proceedings of the Asian Conference on Machine Learning (2012)
64. Vovk, V.: Cross-conformal predictors. The Annals of Mathematics and Artificial Intelligence (2015)
65. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005)
66. Vovk, V., Wang, B., Wang, R.: Admissible ways of merging p-values under arbitrary dependence. The Annals of Statistics (2022)
67. Vovk, V., Wang, R.: Combining p-values via averaging. Biometrika (2020)
68. Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S.: Data-uncertainty guided multi-phase learning for semi-supervised object detection. Conference on Computer Vision and Pattern Recognition (2021)
69. Watkins, L., Hamilton, D., Young, T.A., Zanlongo, S., Whitcomb, L.L., Spielvogel, A.R., Kobzik-Juul, B.: The roles of autonomy and assurance in the future of uncrewed aircraft systems in low-altitude airspace operations. Computer (2023)
70. Westfall, P.H., Young, S.S.: Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons (1993)
71. Wirges, S., Reith-Braun, M., Lauer, M., Stiller, C.: Capturing object detection uncertainty in multi-layer grid maps. IEEE Intelligent Vehicles Symposium (IV) (2019)
72. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
73. Yang, H., Pavone, M.: Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation. Conference on Computer Vision and Pattern Recognition (2023)
74. Yelleni, S.H., Kumari, D., Srijith, P., et al.: Monte carlo dropblock for modeling uncertainty in object detection. Pattern Recognition (2024)
75. 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. Conference on Computer Vision and Pattern Recognition (2020)