Object Detection With Probabilistic Guarantees

Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, David Vigouroux

Object Detection With Probabilistic Guarantees: a Conformal Prediction Approach

Florence de Grancey[1], Jean-Luc Adam[2], Lucian Alecu[3], Sébastien Gerchinovitz[4,5], Franck Mamalet[5], and David Vigouroux[5]

[1] Thales AVS France SAS, Toulouse, France
[2] Renault, Toulouse, France
[3] Continental, Toulouse, France
[4] Institut de Mathématiques de Toulouse, Toulouse, France
[5] IRT Saint Exupéry, Toulouse, France

firstname.surname@irt-saintexupery.com

Abstract. Providing reliable uncertainty quantification for complex visual tasks such as object detection is of utmost importance for safety-critical applications such as autonomous driving, tumor detection, etc. Conformal prediction methods offer simple yet practical means to build uncertainty estimations that come with probabilistic guarantees. In this paper we apply such methods to the task of object localization and illustrate our analysis on a pedestrian detection use-case. Throughout the paper we highlight both theoretical and practical implications of our analysis.

Keywords: Object detection · Conformal prediction · Uncertainty quantification.

1 Introduction

Recent works in object detection show a great variety of models and approaches. Among the most notable we can mention: RCNN [13], Fast-RCNN [12], RetinaNet [22], FPN [21], YOLO and its several versions [28–30], SSD [24] or DETR [7].

Despite their impressive success observed on various benchmarks, many challenges remain ahead. For critical systems, several additional guarantees shall be provided to avoid catastrophic consequences: in an autonomous vehicle, a pedestrian mislocated by the system could be hurt or killed; in a cancer detection system, several cancer cells missed by the object detector could not be treated. To ensure the safety of the user, the uncertainty of the location of the object to detect should be quantified, allowing to create safeguards around the object.

The main challenge consists in providing *reliable* uncertainty quantification of their prediction errors. While many object detection models compute so-called confidence scores which can be interpreted as basic estimators of uncertainty, they are often unreliable (i.e. over or under-estimating the true uncertainty).

Another difficulty stems from the complex interplay between the classification-type errors and the localization-type errors of the object detectors. In addition, the risks associated with each type of error are application-dependent.

For safety-related applications, one may seek to obtain various guarantees. One such guarantee related to object localization, that will be addressed in this paper, may read: ensure that at least a significant portion (i.e. a user-specified fraction) of the objects recognized in visual images satisfy this property: their true bounding boxes are fully covered[6] by the boxes predicted by a given object detection model. This type of guarantee may be helpful, for example, to build reliable models for tumor discovery, obstacle detection or trajectory estimation.

Uncertainty Quantification for Object Detection Several techniques such as Deep Ensembles methods [18, 25] or Monte Carlo-Dropout methods [2, 9, 26, 27] have been developed to provide epistemic uncertainty quantification. Other methods such as Direct Modeling add additional layers on top of the object detector to achieve such estimations [19]. More recent works introduce *probabilistic object detectors* which distinguish between the aleatoric and epistemic uncertainties and estimate their variances separately. Based on the work of [15], the authors extend Bayesian neural networks to object detectors [14,16]. A complete survey on uncertainty estimation for object detectors can be found in [11].

Many of the cited methods have been applied to uncertainty quantification in object detection tasks with various success. Nevertheless, to the best of our knowledge, none of these works provide statistical guarantees about the estimated uncertainties, e.g., that the relevant objects are correctly classified with high probability, or that they are correctly localized in the image most of the time, or both.

Main Contributions and Outline of the Paper In this work, we consider a relatively recent family of statistical methods called *Conformal Prediction*, which are post-processing methods to compute guaranteed "error margins" for various learning tasks. Our main contribution is the first application of such ideas to a practical object detection use-case, namely, pedestrian *localization* (i.e. correct prediction of the minimum area bounding box encompassing objects classified as pedestrians). This may be further used to increase the reliability of, e.g., collision avoidance or assisted braking functions. To that end, the paper is organized as follows:

- After presenting the main ingredients of conformal prediction in Section 2 we describe the experimental setting in Section 3.
- In Sections 4 and 5 we show several ways to apply conformal prediction methods for object localization, with various statistical guarantees.
- In Section 6 we emphasize subtle pitfalls that a user may fall into, to help interpret conformal prediction guarantees when applied to object detection.

[6] All the coordinates of the true box will be found inside the rectangle defined by the predicted bounding box of the object.

Fig. 1. Conformalization example (box-wise, risk level $\alpha = 0.1$) on a BDD100k image with **Ground Truth**, **Inference** and **Conformalized** boxes.

2 Background: Conformal Prediction

Consider a supervised learning task (e.g. classification or regression), where we want to predict an unknown label y (e.g. a class or a real number) given an observed input x (e.g. an image). Typical ML models such as deep neural networks output predictions $\hat{f}(x)$ with little or no hint as to whether $\hat{f}(x)$ is close to the unknown label y. To that end, *Conformal Prediction* [1, 20, 33] is a family of post-processing methods that are useful to compute guaranteed "error margins", under some assumptions on the data (see Theorem 1 below for an example). The overall process from learning to inference typically unfolds as follows.[7]

1. **Data collection:** Two different datasets are collected: a *training set* and a *calibration set*, which will be used to learn and evaluate a ML model. (See below for independence and distribution requirements on the data.)
2. **Training step:** a machine learning model $\hat{f}$ is learned on the *training set*. The underlying model can be of virtually any kind (a deep neural network, a random forest, etc).
3. **Conformalization step:** the learned model $\hat{f}$ is evaluated on the *calibration set*. This step consists in measuring the errors of $\hat{f}$ on the calibration set, and in reporting a quantile q_α of these errors for some pre-specified risk level $\alpha \in (0, 1)$. More precisely, given a *non-conformity score* $s\big(\hat{y}, y\big)$ to assess the "distance" between a prediction $\hat{y} = \hat{f}(x)$ and a ground truth y, we compute the errors of $\hat{f}$ on all data points (x_i, y_i) of the calibration set[8]:

$$R^i = s\big(\hat{y}_i, y_i\big) \;, \quad i = 1, \ldots, n_c \;, \tag{1}$$

[7] More complex variants exist. The typical process outlined here is more precisely known as *split conformal prediction*.

[8] The errors are sometimes called "residuals" (hence the R^i notation).

where n_c is the size of the calibration set. (For example, in regression, we can take the absolute difference $s\big(\hat{y}, y\big) = |y - \hat{y}|$.) Then, the quantile q_α is defined as the $\lceil (1-\alpha)(n_c+1) \rceil$-th largest value among the observed errors R^i.

4. **Inference step:** Given a new input x, instead of outputting a simple prediction $\hat{f}(x)$, we output a prediction set $C^\alpha(x)$, with the goal of containing the unknown ground truth y. It is defined as the set of all labels y' that are "close enough" to the prediction $\hat{y} = \hat{f}(x)$ of the ML model:

$$C^\alpha(x) = \Big\{ y' : \; s\big(\hat{y}, y'\big) \leq q_\alpha \Big\} \,, \tag{2}$$

where q_α is the error quantile reported at the end of the conformalization step, and serves as an "error margin". For example, in the regression example mentioned above, the prediction set is given by $C^\alpha(x) = [\hat{y} - q_\alpha ; \hat{y} + q_\alpha]$. Other non-conformity scores lead to other prediction sets, as shown later.

Dataset Requirements. In order to be able to prove that a prediction set $C^\alpha(x)$ contains the unknown label y "most of the time", the datasets must satisfy some requirements. Sufficient requirements are that:[9]

(i) data from all 3 datasets (training, calibration, inference) are independent;
(ii) data distributions at calibration and inference steps are identical.

Requirement (i) is useful to avoid overfitting issues. Requirement (ii) is useful to make sure that errors measured during the conformalization step are representative of errors at inference time. Interestingly though, training data can be distributed differently, which can be useful when computational resources or data for training are rather scarce, while an ML model carefully trained for a close distribution is already available. (Of course, a model that was pre-trained for a very different distribution will perform poorly at the conformalization step, and thus the error margin q_α will be large.)

Under the above dataset requirements, the conformal prediction process 1-4 outlined above satisfies the following probabilistic guarantee.

Theorem 1 (see, e.g., [1, 20, 33]). *Assume the training, calibration, and inference datasets satisfy Requirements (i) and (ii) above. Then, on average over the choice of the calibration set and the new data point* (X, Y)*,*

$$P\big(Y \in C^\alpha(X)\big) \geq 1 - \alpha \,.$$

We say in this case that the method has a *coverage* of $1 - \alpha$. The above guarantee means that, for a fraction $1-\alpha$ of all possible calibration sets in Step 3 and possible data points (x, y) in Step 4, the prediction set $C^\alpha(x)$ contains the true label y. In other words, if we repeated the overall conformal prediction process 1-4 many times independently, it would err a fraction at most α of the time. Details about dangers of interpretation are given in Section 6.

[9] Mathematically speaking, it is in fact sufficient that the calibration data and the data at inference time are exchangeable, conditionally on the training data.

3 Experimental Setting and Goals

In the following sections we describe how conformal prediction methods can be applied to post-process (i.e. shrink or enlarge) the prediction boxes provided by a pedestrian detector. The goal is that the new boxes, called *conformalized boxes*, fully cover the true bounding boxes of the objects of interest, "most of the time". For example, in Figure 1 we would like to cover all gold boxes with green boxes, which are obtained by adjusting the predicted boxes in cyan. The precise interpretation and limitations of the "most of the time" statement will be detailed shortly, when we apply conformalization procedures on different levels respectively: per coordinate, per bounding box or per image.

Both the level at which conformalization is conducted, and the design of non-conformity scores are *engineering choices*, as they depend on the actual usage of the model in real-world applications. For the pedestrian detection case considered here we may be interested in providing guarantees related to individual objects. In this case, a box-wise conformalization seems more appropriate. However, in other cases we may be interested in image-wise conformalization, as to ensure that a majority of images satisfy a desired property (e.g. all or most of the objects of interest in the image are "well" localized).

In all our experimental settings we consider the YOLOv3 object detector [30] pretrained on the COCO training dataset [23] (i.e. Step 2 described in Section 2 is fixed). As stated in Section 2 we can conformalize on a calibration dataset with a distribution that is different from that of the training set. Therefore we conduct all our experiments on the BDD100k dataset [34] by considering its training set as our calibration set (denoted by $\mathcal{D}^{calib}_{\text{BDD}}$ thereafter) and its validation set as our test set (denoted by $\mathcal{D}^{test}_{\text{BDD}}$). Since we focus on pedestrian detection, the original $70k+10k$ images of training and validation sets reduce to 22213 and 3220 images with at least one person/pedestrian (we include riders), containing 91349 and 13262 annotated persons respectively.

In the following sections we propose non-conformity scores for each level of analysis (coordinate, box or image) and discuss practical implications of these choices. Finally, we emphasize some statistical aspects that are essential to any correct interpretation of the obtained guarantees.

4 Coordinate-Wise and Box-Wise Conformalization

As mentioned above, our goal is to post-process the boxes predicted by an object detector such that they cover the true bounding boxes of the objects of interest. In this section, we compute error margins at box level, by treating boxes as individual data points for the conformal prediction process of Section 2. We compute error margins for each coordinate $x_{\min}, x_{\max}, y_{\min}, y_{\max}$ separately (Sec. 4.2), and then show how to correct them to obtain guarantees at box level (Sec. 4.3).

4.1 Preliminary Assignment

On a given image, in order to compare predicted boxes with true boxes at box level (that is, compare box A with box B), we need to assign predicted boxes to

true boxes. A preliminary pre-processing *assignment stage* is thus necessary. This step is performed with the Hungarian matching algorithm [17] based on the IoU metric. By applying it, we retain exclusively the true positive bounding boxes for calibration. Therefore, at inference time our probabilistic guarantee will only ensure that true positives are correctly covering the ground truth, while false negatives might still exist. In the following experiments, unless otherwise stated, a detection box must have a confidence score higher than 0.5 and an IoU with a ground truth object above 0.5 to assign the predicted box to a true box. Based on this assignment stage, the BDD100k calibration and test sets reduce to 42824 and 6138 assigned persons respectively.

4.2 Coordinate-Wise Conformalization

We explain in details how to instantiate Steps 3 and 4 of Section 2.

Conformalization step. Assume that we have assigned predicted boxes to true boxes as in Section 4.1. In order to compare the i-th predicted box with the i-th true box, we compare each of the four predicted coordinates $\hat{x}^i_{\min}, \hat{x}^i_{\max}, \hat{y}^i_{\min}, \hat{y}^i_{\max}$ with the four true coordinates $x^i_{\min}, x^i_{\max}, y^i_{\min}, y^i_{\max}$, by counting errors positively when the truth lies outside the prediction (e.g. $\hat{x}^i_{\min} > x^i_{\min}$ or $\hat{x}^i_{\max} < x^i_{\max}$), and negatively otherwise. This choice is less conservative than considering absolute error values and leads to the following four errors (cf. Eq. (1)): for $i = 1, \dots, n_c$,

$$\begin{array}{ll} R^i_{x_{\min}} = \hat{x}^i_{\min} - x^i_{\min} & R^i_{y_{\min}} = \hat{y}^i_{\min} - y^i_{\min} \\ R^i_{x_{\max}} = x^i_{\max} - \hat{x}^i_{\max} & R^i_{y_{\max}} = y^i_{\max} - \hat{y}^i_{\max} \end{array} \tag{3}$$

Note that n_c is given by the total number of predicted objects assigned to a true object, which is larger than the number of images in the calibration set.

Then, following Step 3 of Section 2, we compute a quantile q_α for each of the four errors above, defined as the $\lceil (1-\alpha)(n_c+1) \rceil$-th largest value among the observed errors R^i. These four quantiles will serve as error margins for each coordinate.

As an illustration, the errors $R^i_{y_{\max}}$ on $\mathcal{D}^{calib}_{\text{BDD}}$ are represented on the histogram on the left side of Fig. 2 (in red: the quantile q_α for $\alpha = 0.1$, i.e. specified coverage of 0.9). The right side shows the evolution of the quantile q_α w.r.t. the parameter α. High guarantees ($\alpha < 0.05$) imply large margins, whereas low guarantees only require small modifications of the predicted coordinates.

Inference step. We now instantiate Eq. (2) of Section 2 to compute a prediction set C^α for each coordinate $x_{\min}, x_{\max}, y_{\min}, y_{\max}$, given four predicted coordinates $\hat{x}_{\min}, \hat{x}_{\max}, \hat{y}_{\min}, \hat{y}_{\max}$ as inputs. These prediction sets are intervals:

$$\begin{array}{ll} C^\alpha_{x_{\min}} = [\hat{x}_{\min} - q^{x_{\min}}_\alpha, +\infty) & C^\alpha_{y_{\min}} = [\hat{y}_{\min} - q^{y_{\min}}_\alpha, +\infty) \\ C^\alpha_{x_{\max}} = (-\infty, \hat{x}_{\max} + q^{x_{\max}}_\alpha] & C^\alpha_{y_{\max}} = (-\infty, \hat{y}_{\max} + q^{y_{\max}}_\alpha] \end{array} \tag{4}$$

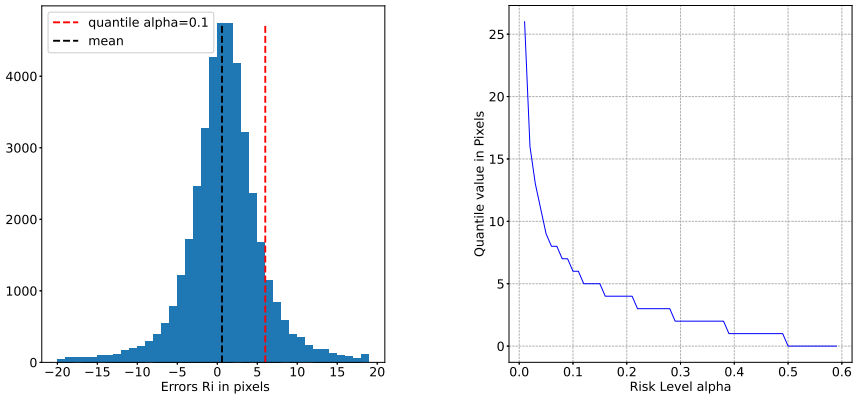

Fig. 2. Left: histogram of the errors R^i for the coordinate $y_{\max}$, and the corresponding quantile q_α for $\alpha = 0.1$. Right: Evolution of the quantile value q_α with risk level α.

In Table 1, the first four lines (coordinate-wise) give the evaluation of the observed coverage on the $\mathcal{D}^{test}_{\text{BDD}}$ set, i.e., for each coordinate, the proportion of (true positive) boxes for which the true coordinate lies within the corresponding prediction set. We can see that Thm. 1 is verified whatever the specified coverage.

Table 1. Evaluation of observed coverage on $\mathcal{D}^{test}_{\text{BDD}}$ using the quantile evaluated for three specified coverage values (in red when the specified coverage is not reached).

Method	specified coverage (1-α)	0.7	0.9	0.95
		Observed coverage		
Coordinate-Wise §4.2	$x_{\min}$	0.76	0.91	0.96
	$x_{\max}$	0.78	0.91	0.96
	$y_{\min}$	0.70	0.92	0.95
	$y_{\max}$	0.71	0.91	0.95
Box-Wise §4.3	w/o Bonferroni	0.35	0.73	0.86
Box-Wise §4.3	with Bonferroni	0.79	0.92	0.96

4.3 Bonferroni Correction for Box-Level Guarantees

In this section we seek the following guarantee: at inference time, among all true bounding (pedestrian) boxes that are detected, a fraction $1 - \alpha$ of them are correctly covered by conformalized boxes.[10] We explained in Eq. (4) how to compute error margins to locate unknown coordinates $x_{\min}, x_{\max}, y_{\min}, y_{\max}$ of a box, given predictions $\hat{x}_{\min}, \hat{x}_{\max}, \hat{y}_{\min}, \hat{y}_{\max}$. It might be tempting to define the

[10] The $1 - \alpha$ guarantee only holds on average over all calibration sets, see Section 6.

conformalized box as the largest (worst-case) box whose coordinates are within the intervals C^α, i.e., the box with coordinates

$$\hat{x}_{\min} - q_\alpha^{x_{\min}} \qquad \hat{x}_{\max} + q_\alpha^{x_{\max}} \qquad \hat{y}_{\min} - q_\alpha^{y_{\min}} \qquad \hat{y}_{\max} + q_\alpha^{y_{\max}} \ .$$

However, looking at the experimental results of Table 1 (line w/o Bonferroni), we remark that for, e.g., a specified coverage of 95%, only 86% of true boxes are covered by conformalized boxes. The lower-than-expected coverage is a direct consequence of considering each coordinate of the bounding boxes independently, since errors on each coordinate can happen on different boxes.

One simple solution is to apply a so-called *Bonferroni correction* (e.g., [5]), a statistical adjustment to account for this $4-way$ dependency. This is straightforward, as it amounts to perform the exact same procedure as described in the previous section, but replacing q_α with $q_{\alpha/4}$. With this slight change the new coverage on the test set becomes within the expected levels (see Table 1, line with Bonferroni). Note that since $q_{\alpha/4} \geq q_\alpha$, the conformalized boxes will be larger. For example, for the $y_{\max}$ coordinate and for a specified coverage of 0.9 ($\alpha = 0.1$), we obtain $q_\alpha = 6$ pixels and $q_{\alpha/4} = 14$ pixels.

5 Image-Wise Conformalization

The method of the previous section aimed at guaranteeing that, at inference time, among all true (pedestrian) bounding boxes that are detected, a fraction $1-\alpha$ of them are correctly covered by conformalized boxes.[10] This guarantee has two limitations: (i) false negatives (undetected pedestrians) are not taken into account, without any control on their occurrence rate; (ii) the fraction of detected true boxes that are covered on a given image might be much different from $1-\alpha$ (the $1-\alpha$ coverage is an average over all boxes across the test set). While the box-wise approach can still be useful for pedestrian detection (e.g., for tracking true positives), for medical applications such as cancer cell detection, (i) and (ii) imply that we might miss too many cancer cells on too many images.

Next we study another non-conformity score to pursue a guarantee at image level rather than at object level. We aim at the following guarantee: on average over the choice of the calibration set, a fraction $1-\alpha$ of images at inference will be such that a fraction $1-\beta$ of true boxes in the image will be correctly covered by conformalized boxes.

To that end, we consider a non-conformity score $s_\beta(\hat{B}, B)$ that is a close variant of the *partial (or quantile) directed Hausdorff distance* [31]. It compares two sets of boxes on a given image: the set $\hat{B}$ of all predicted boxes, and the set B of all true boxes. Our score $s_\beta(\hat{B}, B)$ is defined as the smallest margin $r \geq 0$ that it suffices to add to *all* predicted boxes in $\hat{B}$ (on *all* four coordinates) so that a fraction at least $1-\beta$ of true boxes in B are correctly covered by the union of enlarged predicted boxes. We then follow the whole process of Section 2. In particular, the quantile q_α computed at conformalization is the margin that will be added to all predicted boxes at inference.

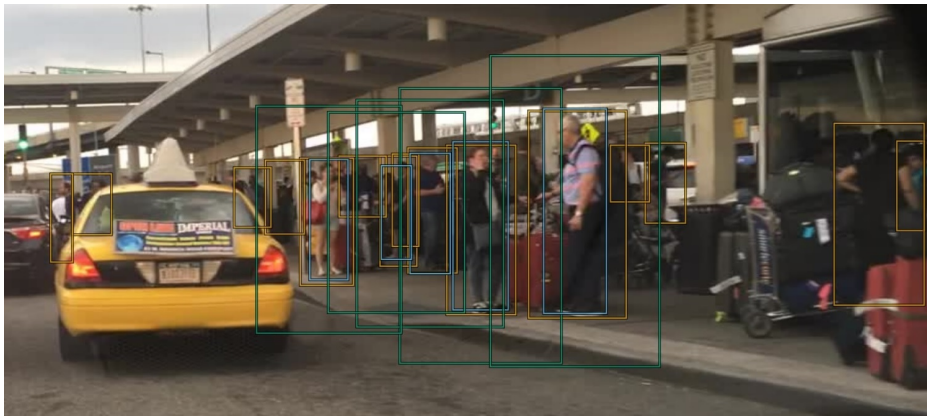

Fig. 3. Conformalization example (image-wise, risk level $\alpha = 0.1$, $\beta = 0.25$) on a BDD100k image with **Ground Truth**, **Inference** and **Conformalized** boxes.

A BDD100k example is shown on Figure 3, for $\alpha = 0.1$ and $\beta = 0.25$ (the aim is to correctly locate 75% of all pedestrians on each of 90% of all possible images). Interestingly, though only 5 pedestrians were detected (in cyan), 9 pedestrians (in gold) are located within conformalized boxes (in green). This positive effect is due to the presence of multiple pedestrians nearby. Though this effect is not frequent in our use-case (pedestrians are often more isolated) and prevents the system from tracking pedestrians individually, it seems more useful for applications where objects of interest are often nearby, such as cancer cell detection.

6 Statistical Pitfalls

While very useful in practice, the probabilistic guarantees behind conformal prediction such as Theorem 1 should be interpreted with care. They rely on assumptions and have some limitations, which we outline below.

A guarantee "on average" over the box/image domain. As explained after Theorem 1, the inequality $P(Y \in C^\alpha(X)) \geq 1 - \alpha$ is a guarantee *on average* over calibration sets and test data points. In our use-case, this implies that the $1 - \alpha$ coverage is correct on average over all boxes/images at inference time, but might be incorrect in some subsets of the box/image domain. For example, for our box-wise conformalization experiment on the BDD100k dataset, when restricting the test set $\mathcal{D}^{test}_{\text{BDD}}$ to pedestrians that are close to the camera (height larger than 150 pixels), the coverage is smaller than 64% for a specified coverage of 70% ($\alpha = 0.3$), and smaller than 81% for a specified coverage of 90% ($\alpha = 0.1$).[11] Of course, a simple solution here is to apply conformal prediction for close pedestrians only (predicted height larger than 150 pixels). However, this

[11] These coverage values include statistical error margins at level 95%.

solution cannot work in settings where a very large number of subdomains need to be distinguished (since calibration sets need to be large enough) or where these subdomains are not known a priori. In any case, due to the statistical nature of conformal prediction methods, one must keep in mind that there are some boxes or images on which these methods will fail at inference.

A guarantee "on average" over calibration sets. Similarly, while the $1-\alpha$ coverage is correct on average over all possible calibration sets, its value might be different for the single calibration set used in practice. The way the coverage varies from one calibration set to another was described in details in [1]. Next we illustrate this variability on BDD100k with box-wise conformalization (as in Section 4.3) and $\alpha = 0.1$, by re-sampling various calibration sets and reporting the associated test coverage values. The histogram on Figure 4 shows a large variability of coverage values, which means that different calibration sets lead to different coverage values at inference. Fortunately here, most values are above the specified coverage of 0.9 (since the margins are a little conservative due to the Bonferroni correction), but the tail probability on the left of 0.9 implies that the user has still a chance to use a calibration set that would result in a lower-than-expected coverage at inference. Recent works have proposed variants of conformal prediction (with more conservative margins) to deal with this variability (e.g., [4, 10]).

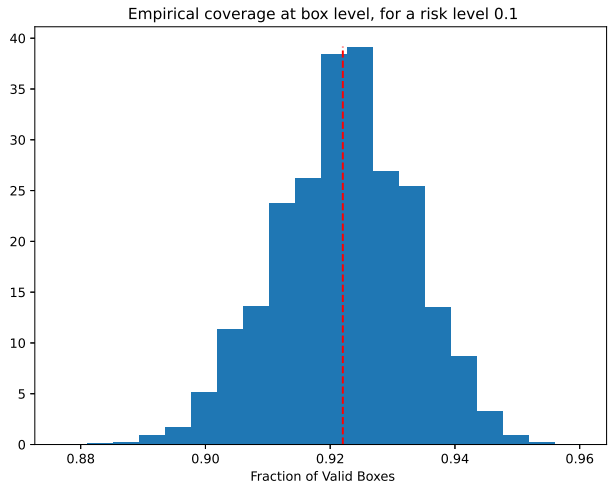

Fig. 4. Empirical coverage distribution measured on the same test set when sampling different calibration sets and applying box-wise conformalization as in Section 4.3.

Datasets requirements: independence assumption. As recalled in Section 2, several properties on the datasets are required for the probabilistic guarantee of Theorem 1 to be valid. This is not at all surprising due to the statistical

nature of conformal prediction. The independence requirement (i) between all data involved (training, calibration, test) is to avoid overfitting issues. In particular, dependencies between training and calibration data could lead to measure artificially small errors R^i at conformalization, while dependencies within the calibration set may reduce statistical power in an uncontrolled manner.[12] In practice, the independence assumption seems difficult to guarantee or test, but data collection practices should aim at it. For the box-wise procedure of Section 4.3, it is not clear a priori that errors R^i are independent, because of possible intra-image dependencies between boxes. Dependencies seem a little easier to prevent in the image-wise setting of Section 5 (by, e.g., discarding neighboring images in a sequence). For our use-case however, our box-wise and image-wise test coverage values, which reach specified coverage, are really encouraging.

Distribution shift between calibration and test. We now investigate the importance of the dataset requirement (ii) of Section 2. To that end, we use our model calibrated on $\mathcal{D}^{calib}_{\text{BDD}}$ for the box-wise criterion, as in Section 4.3, but we test its coverage on a new dataset, namely the Cityscape validation set [8]. This might be a reasonable thing to do, as both datasets are quite similar (they include urban scenes for autonomous driving scenarios). Yet the obtained coverage on this new dataset is 0.790 for $\alpha = 0.1$, i.e. significantly lower than expected. On the opposite, the coverage on $\mathcal{D}^{test}_{\text{BDD}}$ is 0.924, which is consistent with the expected guarantee. See Figure 5 for results on a range of α values, the black dotted line showing the expected coverage. In our haste to apply our methods on new data, we misinterpreted the offered guarantee: even if conceptually similar, the two datasets (the one used for calibration and test) probably do not share the same intrinsic data distribution. This invalidates the data-related assumptions described in Section 2, and consequently the offered guarantee. While this is obviously a contrived example, in many practical situations it is quite easy to mislead oneself into believing that the data processed at inference time follows exactly the same distribution as the one used for the calibration of the model. Recent works in conformal methods such as [3, 32] attempt to address such challenges related to *distribution shift*. In our future work, we aim to apply these results on our own experiments.

7 Discussions and Future Works

In this paper we present a practical guideline of how conformal prediction methods can be applied to object localization to provide uncertainty estimations with probabilistic guarantees. We illustrate the essential aspects on a pedestrian localization use case. We propose several variants of conformal prediction methods that provide guarantees at various levels (coordinate, bounding box or image) and with different non-conformity scores. Finally, we highlight several statistical aspects that one must take into account, and discuss the interpretation of the obtained guarantees.

[12] In fact, conformal guarantees work slightly beyond independence—under a so-called "exchangeability" assumption [1, 20, 33].

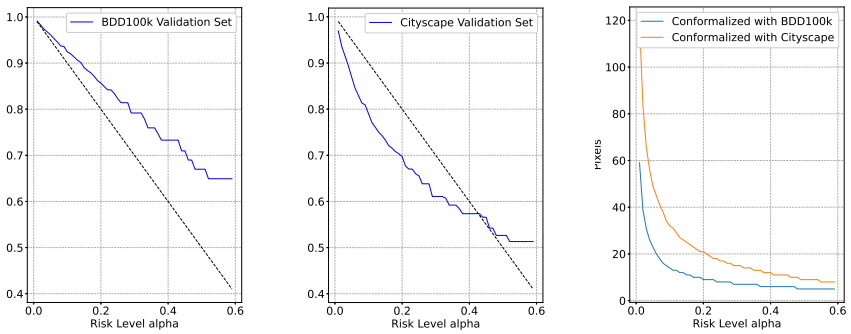

Fig. 5. Box-wise empirical coverage oberved on (left) BDD100k validation set vs. (middle) Cityscape validation set (expected coverage is black dotted line). Right: quantile curves for $Y_{\max}$ with BDD100k or Cityscape training sets considered for calibration.

An engineering choice. Conformal prediction methods rely on the choice of a non-conformity score that can be designed according to the targetted task and a priori knowledge. For example, if we wish to differentiate the level of uncertainties w.r.t. the size of the object we can normalize the scores of Eq. 3 by the width and height of the predicted box, i.e., $\tilde{R}^i_{x_{\min}} = R^i_{x_{\min}}/\Delta^i_x, \tilde{R}^i_{x_{\max}} = R^i_{x_{\max}}/\Delta^i_x$, where $\Delta^i_x = |\hat{x}^i_{\max} - \hat{x}^i_{\min}|$ (and similarly for the y coordinates). Such scores would lead to margins that scale multiplicatively with the size of the predicted object. Though this is undesirable for our pedestrian use-case (since pedestrians that are close to the camera correspond to the largest objects), this might be useful in applications where we are ready to pay larger margins for larger objects. Likewise, we can extend proposed scores as to compute simultaneously an inner and outer conformal bounding box covering most of the true boxes (by considering errors in absolute value). Note that the choice of the non-conformity score can also help to analyze the data quality in test or calibration sets: high scores can be indicative of 'extreme cases', such as suspicious inputs or anomalous annotations.

Future work. This paper focused on the localization task. Next we will address the classification problem, and the interplay between the two—that is, the global detection problem. At a system level, it would also be important to build a link between such statistical guarantees and safety-related risks (e.g., fault rate). This task is however difficult, as this necessarily relies on additional assumptions and uncertainties [6]. All these aspects, together with the numerous potential statistical pitfalls, open up very challenging research questions.

Acknowledgements This work has benefited from the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-P3IA-0004. The authors gratefully acknowledge the support of the DEEL project.[13]

[13] https://www.deel.ai/

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification (2021), arXiv:2107.07511
2. Azevedo, T., de Jong, R., Maji, P.: Stochastic-yolo: Efficient probabilistic object detection under dataset shifts (2020), arXiv:2009.02967
3. Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability (2022), arXiv:2202.13415
4. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.I.: Distribution-free, risk-controlling prediction sets. Journal of the ACM **68**(6) (2021)
5. Bickel, P.J., Doksum, K.A.: Mathematical Statistics: Basic Ideas and Selected Topics, vol. 1. Chapman and Hall/CRC (2015)
6. Bonnin, H., Jenn, E., Alecu, L., Fel, T., Gardes, L., Gerchinovitz, S., Ponsolle, L., Mamalet, F., Mussot, V., Cappi, C., Delmas, K., Lefevre, B.: Can we reconcile safety objectives with machine learning performances? In: Proc. of the 11th Edition of European Congress of Embedded Real Time Systems (ERTS) (2022)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision – ECCV 2020. Springer International Publishing (2020)
8. 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: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. Deepshikha, K., Yelleni, S.H., Srijith, P.K., Mohan, C.K.: Monte carlo dropblock for modelling uncertainty in object detection (2021), arXiv:2108.03614
10. Ducoffe, M., Gerchinovitz, S., Sen Gupta, J.: A high-probability safety guarantee for shifted neural network surrogates. In: Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020). pp. 74–82 (2020)
11. 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)
12. Girshick, R.B.: Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
13. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014)
14. Harakeh, A., Smart, M., Waslander, S.L.: BayesOD: A Bayesian approach for uncertainty estimation in deep object detectors (2019), arXiv:1903.03838
15. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems (2017)
16. Kraus, F., Dietmayer, K.: Uncertainty estimation in one-stage object detection. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (2019)
17. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2) (1955)
18. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (2017)
19. Le, M.T., Diehl, F., Brunner, T., Knol, A.: Uncertainty estimation for deep neural object detectors in safety-critical applications. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (2018)

20. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018)
21. Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
22. Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis Machine Intelligence **42**(02) (2020)
23. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Computer Vision – ECCV 2014. Springer International Publishing (2014)
24. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: Ssd: Single shot multibox detector (2016)
25. Lyu, Z., Gutierrez, N., Rajguru, A., Beksi, W.J.: Probabilistic object detection via deep ensembles. In: European Conference on Computer Vision. Springer (2020)
26. Miller, D., Dayoub, F., Milford, M., Sunderhauf, N.: Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In: 2019 International Conference on Robotics and Automation (ICRA) (2019)
27. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: 2018 International Conference on Robotics and Automation (ICRA) (2018)
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
29. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
30. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement (2018), arXiv:1804.02767
31. Rucklidge, W.J.: Efficiently locating objects using the hausdorff distance. International Journal of Computer Vision **24**, 251–270 (1997)
32. Tibshirani, R.J., Barber, R.F., Candes, E.J., Ramdas, A.: Conformal prediction under covariate shift. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (2019)
33. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer-Verlag, Berlin, Heidelberg (2005)
34. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: BDD100K: a diverse driving dataset for heterogeneous multitask learning (2018), arXiv:1805.04687