Multivariate Confidence Calibration for Object Detection

Fabian Küppers[1], Jan Kronenberger[1], Amirhossein Shantia[2], Anselm Haselhoff[1]

{fabian.kueppers, jan.kronenberger, anselm.haselhoff}@hs-ruhrwest.de, ashantia@visteon.com

[1]Ruhr West University of Applied Sciences
Bottrop, Germany

[2]Visteon Electronics GmbH
Kerpen, Germany

Abstract

Unbiased confidence estimates of neural networks are crucial especially for safety-critical applications. Many methods have been developed to calibrate biased confidence estimates. Though there is a variety of methods for classification, the field of object detection has not been addressed yet. Therefore, we present a novel framework to measure and calibrate biased (or miscalibrated) confidence estimates of object detection methods [1]. The main difference to related work in the field of classifier calibration is that we also use additional information of the regression output of an object detector for calibration. Our approach allows, for the first time, to obtain calibrated confidence estimates with respect to image location and box scale. In addition, we propose a new measure to evaluate miscalibration of object detectors. Finally, we show that our developed methods outperform state-of-the-art calibration models for the task of object detection and provides reliable confidence estimates across different locations and scales.

1. Introduction

Modern object detection algorithms are widely used in several domains like autonomous driving or medical diagnosis. Most of these object detectors, based on neural networks, provide a score for a certain class and a proposal of the object location and scale. In previous work it has been shown that modern architectures for classification tasks tend to be overconfident with their predictions compared to the observed accuracy [3] and thus are called *miscalibrated*. Calibration in the scope of predictors means that the observed accuracy of the predictor matches its mean confidence. Miscalibration can be observed for modern object detection models as well. However, a well-calibrated object detector is crucial

---

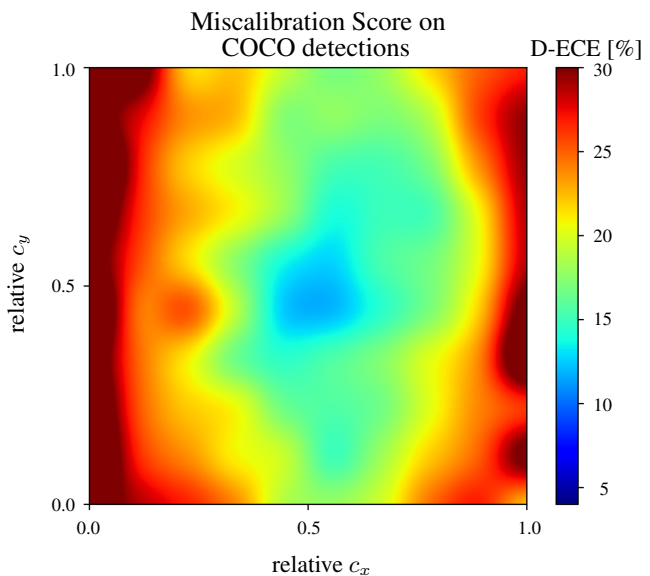[1]The framework is publicly available at: https://github.com/fabiankueppers/calibration-framework

Figure 1: Miscalibration evaluation on the COCO dataset for a R-FCN ResNet-101 [1]. The miscalibration score highly depends on the object location (the center coordinates c_x and c_y) and increases as the prediction gets close to the image boundaries. The evaluation is based on the detection expected calibration error (D-ECE), a variant of the expected calibration error (ECE) [14].

especially in safety-critical applications like pedestrian detection for autonomous driving applications.

For image classification, there are methods to calibrate the predictions of a network as a post-processing step (logistic calibration/Platt scaling [19], temperature scaling [3], etc.). These methods can easily be applied to object detection tasks. However, these well-known methods only take the confidence scores of the detections into account. As illustrated in Fig. 1, the calibration error highly depends on the object location and increases as the prediction gets close to the image boundaries. A similar behavior can be observed with respect to the object width and height. These
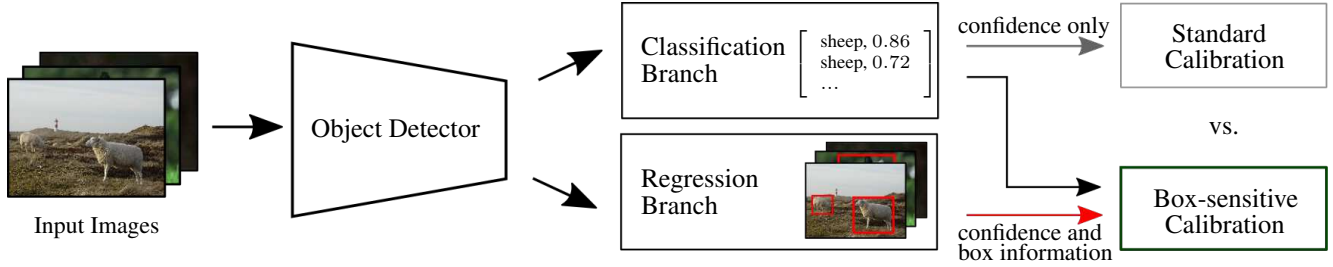
Figure 2: Extended box-sensitive or multivariate calibration proposed in this paper (bottom) vs. standard confidence calibration (top). An object detector predicts a class with a certain confidence and a box proposal containing information about location and scale. Standard calibration methods only take the confidence into account. Our approach also includes the regression branch of an object detector in order to improve the calibration results.

location-dependencies are not captured by conventional calibration methods but might be very important especially in safety-critical domains. For example, in road traffic it often occurs that a pedestrian suddenly appears at the roadside to cross the road and is therefore located at the image boundaries. If we observe a miscalibration especially at the boundaries, we cannot trust in the confidence score.

Motivated by the empirical observation that the calibration error depends on bounding box properties, our goal is to develop box-sensitive calibration methods for detectors. We propose a simple and effective framework to measure and optimize the calibration properties of object detectors. The approach treats the detector as a black-box and is therefore applied after the regular training process. Our framework is agnostic to the used detection method as well as to the underlying calibration algorithm. In other words, we do not modify the learning objective of a detector and thus our work is orthogonal to ongoing research in regularization methods for calibration [15, 22, 18]. We show how to extend common calibration algorithms, *e.g.* logistic calibration, and how to measure miscalibration location and/or scale dependent. We use pretrained state-of-the-art object detection methods (SSD [12], Faster R-CNN [20] and R-FCN [1]) to evaluate our considerations about the developed calibration algorithms.

2. Related Work

Common methods for confidence calibration in the domain of classification are histogram binning [23], isotonic regression [24], Bayesian binning into quantiles (BBQ) [14], ensemble of near isotonic regression (ENIR) [13], logistic calibration/Platt scaling [19], temperature scaling [3] and beta calibration [9]. These methods are applied during inference as a post-processing step after classification. A straight forward transfer of these methods to object detectors is possible, but might not utilize the full potential - dependencies between the estimated confidences and the regression output might have an influence.

In contrast to post-processing calibration methods, [22] has recently proposed a method to obtain well-calibrated classifiers during model training. The authors have shown that the uncertainty obtained by Monte-Carlo dropout is correlated with the miscalibration and hence they use the uncertainty to control an additional regularization term. This approach is related to the ongoing research in the domain of model uncertainty. A common way to obtain uncertainties are Bayesian neural networks (BNNs) [2, 6]. An additional approach to obtain uncertainties for one-stage object detectors is presented by [7]. However, it has been shown that even those uncertainty estimates are biased and thus need calibration [2, 8, 7]. In [8], a method is proposed to perform regression calibration on the confidence interval obtained by a BNN. In contrast, [18] use a regression calibration framework to obtain calibrated object locations. These approaches are orthogonal to ours and could be used in tandem with ours. For example, the framework proposed in this paper can either be used to integrate a dependence to the regression output or to further optimize the calibration in a post-processing step.

Similarly, IoU-Net [5] and GIoU [21] are a valuable extension to our proposed framework to demonstrate further orthogonal calibration directions. In contrast to our approach, IoU-Net focuses on fine-tuning the regression output to improve the IoU between predicted and ground-truth position. They introduce an additional location confidence (regression) in conjunction with the confidence (categorical). They even support the need for our calibration method because they show that the 'regular' confidence is not a direct measure for misalignment. We agree that an improved localization might have an impact on the precision and thus on the confidence calibration as well, but the necessity of confidence calibration remains.

An approach to obtain well-calibrated confidence estimates for object detection is proposed in [15]. They perform a temperature scaling [3] by adding an auxiliary output to the network that scales the remaining logits. Additionally, the authors propose a variation of the expected

calibration error (ECE) [14] as the preferred miscalibration metric for object detection. However, neither the calibration nor the proposed variation of the ECE are explicitly designed to capture properties of the regression output.

3. Methods for Confidence Calibration

In this section we introduce our calibration methodology for object detectors. The main idea is to integrate the classification and regression branch of a detector to perform a box-sensitive calibration (Fig. 2). First, we discuss some background about object detectors that is relevant for the calibration task. Afterwards, we introduce the calibration framework which is an extension to the multivariate domain to handle the regression branch. A simple artificial calibration example is provided to demonstrate the advantage of the proposed method. Finally, we show how to integrate standard methods, e.g. beta calibration, into the calibration framework.

3.1. Object Detectors

Modern object detectors typically take an image as the input x and deliver a prediction in form of a class label y, a confidence score p and a bounding box $r = (c_x, c_y, w, h)$, where (c_x, c_y) is the position of the box center and the scale is determined by the width and height (w, h). We interpret these quantities as random variables that follow a joint ground-truth distribution $\pi(x, y, r) = \pi(y, r|x)\pi(x)$. The detector can be interpreted as function $h(x) = (\hat{y}, \hat{p}, \hat{r})$, that takes the input images and generates predictions for the class, confidence and location. The training process for fitting the model parameters is based on a probabilistic model $\hat{\pi}(y, r|x)$, where the cross-entropy between the training data $\pi_{data}(x, y, r)$ (empirical distribution of $\pi(x, y, r)$ given by the dataset $\mathcal{D}$) and the model $\hat{\pi}(y, r|x)$ is minimized. The evaluation of a detector differs from classification in the sense that both, the class label and the regression output for the location, must match the ground-truth. For the bounding box prediction $\hat{r}$ it is nearly impossible to match the ground-truth perfectly, therefore, a certain overlap score (e.g. Intersection of Union (IoU)) is required to assign detections to ground-truth annotations. In other words, the class prediction and a predefined threshold, e.g. IoU > 0.5, is used to classify whether a prediction is a true or false positive.

The confidence score of a regular classifier is calibrated with respect to the accuracy. For black-box detectors, e.g. without any information about anchors, rejected proposals, etc., it is impossible to determine the accuracy. Instead, the common metrics precision and recall are used to evaluate the performance of an object detector. Precision is the fraction of correct detections among all detections while recall denotes the proportion of actual objects that are correctly identified as such among all ground-truth objects.

Common methods like temperature scaling perform a logistic fit between the estimated confidence and the corresponding ground-truth label. If we want to calibrate with respect to the recall, we also need to integrate ground-truth objects into the calibration that are not identified as such and thus have no confidence score. Therefore, a calibration with respect to the recall is not possible in a black-box setting. In this work we thus use precision as a surrogate for accuracy. That is to say, given 100 detections, each predicted with 0.9 confidence, 90 out of the 100 detections should be correctly classified. This is our measure to define calibration for black-box classifiers. Thus, an object detection model is perfectly calibrated if

$$\underbrace{\mathbb{P}(\mathrm{m} = 1|\hat{\mathrm{p}} = p, \hat{\mathrm{y}} = y, \hat{\mathrm{r}} = r)}_{\text{precision given } p, y, r} = \underbrace{p}_{\text{confidence}} \quad (1)$$
$$\forall p \in [0, 1], y \in \mathcal{Y}, r \in [0, 1]^J$$

is fulfilled, where $\mathrm{m} = 1$ denotes a correctly classified prediction that matches a ground-truth object and $\mathrm{m} = 0$ denotes a mismatch. Thus, $\mathbb{P}(\mathrm{m} = 1)$ is the shorthand notation for approximating $\mathbb{P}(\hat{\mathrm{y}} = \mathrm{y}, \hat{\mathrm{r}} = \mathrm{r})$ with a certain IoU threshold. We use bounding box information relative to the image size $r \in [0, 1]^J$, where J is the dimension of the used box encoding. Similar to [3], we assume that a perfect calibration is not achievable by any known detection method.

3.2. Multivariate Calibration Framework

Our goal is to develop calibration methods for detectors that can be applied after the training independently of the underlying detector architecture (black-box). For simple classification tasks, a confidence map g is applied on top of a miscalibrated scoring classifier $\hat{p} = h(x)$ to deliver a calibrated confidence score $\hat{q} = g(h(x))$. The key ingredient that can be used to improve the calibration of a detector is that we have additional information about the estimated bounding box $\hat{r}$. Therefore, the calibration map is not only a function of the confidence score, but also of $\hat{r}$. To define a general calibration map for binary problems, we use the logistic function and the combined input $s = (\hat{p}, \hat{r})$ of size K by

$$g(s) = \frac{1}{1 + e^{-z(s)}}. \quad (2)$$

Similar to the work presented in [9] regarding beta-calibration, we interpret the logit z as the logarithm of the posterior odds

$$z(s) = \log \frac{f(\mathrm{m} = 1|s)}{f(\mathrm{m} = 0|s)} \approx \log \frac{f(s|\mathrm{m} = 1)}{f(s|\mathrm{m} = 0)} = \ell r(s). \quad (3)$$

For simplicity, we assume a uniform prior $f(\mathrm{m})$ and use the approximation which in turn is the log-likelihood ratio

$\ell r(s)$. In contrast to classifier calibration, we can use multivariate probability density functions to model $f(s|m)$ and therefore include the bounding box information. Even though it is possible to evaluate the distribution parameters directly, we use a discriminative approach to obtain a better calibration mapping [16]. Thus, the calibration process can then be posed as an optimization of a logistic regression problem. Nevertheless, using different models $f(s|m)$ results in an extremely varying expressiveness of the calibration mapping, as we explain in Fig. 3.

3.3. Conditional Independent Calibration

The log-likelihood ratio in Eq. 3, that is used to get a calibrated confidence estimate, is based on two multivariate probability density functions. If we assume conditional independence of all random variables s_k contained in s, we can simplify the joint density and the log-likelihood ratio can be expressed as

$$\ell r(s) = \sum_{k=1}^{K} \log \frac{f(s_k|m=1)}{f(s_k|m=0)}. \qquad (4)$$

This assumptions allows multiple univariate density functions to be combined in an easy fashion. However, keeping track of dependencies and correlations between the variables is not possible. In short, dependencies between confidence, location and scale parameters are neglected.

Independent Logistic Calibration: The well-known logistic calibration or Platt scaling [19] used for classification calibration is one of the simplest calibration methods. The logits of a neural network are scaled and shifted to obtain a calibrated confidence estimate. Logistic calibration is obtained using univariate normal distributions for $g(s)$ [9]. If we assume multiple univariate normal distributions for each box quantity with equal variance, we can simply extend this method to object detection by extending the scale factor to a scale vector $w \in \mathbb{R}^K$. However, instead of using the uncalibrated confidence estimate $\hat{p}$, we use the logit of the network as part of s to be conform with the original formulation in [19] and [9]. If we assume equal variances for both normal distributions, the log-likelihood ratio of the independent logistic calibration for object detection is given by

$$\ell r_{\text{LC}}(s) = s^T w + c, \qquad (5)$$

with the bias $c \in \mathbb{R}$. This method needs $K+1$ parameters to compute a calibration mapping.

Independent Beta Calibration: In standard beta calibration [9] the confidence score is assumed to follow a beta distribution in order to get a better fit in the interval $[0, 1]$. Similar to Platt scaling, the beta calibration

method can also be extended to the object detection task by introducing beta distributions for each box quantity s_k. Then the log-likelihood ratio can be expressed as

$$\ell r_{\text{BC}}(s) = \sum_{k=1}^{K} \log \left(\frac{\text{Beta}(s_k|\alpha_k^+, \beta_k^+)}{\text{Beta}(s_k|\alpha_k^-, \beta_k^-)} \right), \qquad (6)$$

with the beta distribution $\text{Beta}(s_k|\alpha, \beta)$. The shape parameters for $m = 1$ and $m = 0$ are denoted by α_k^+, β_k^+ and α_k^-, β_k^- respectively. We can reparametrize this expression by using

$$\ell r_{\text{BC}}(s) = c + \sum_{k=1}^{K} a_k \log(s_k) - b_k \log(1 - s_k), \qquad (7)$$

where the distribution parameters are summarized by $a_k = \alpha_k^+ - \alpha_k^-$, $b_k = \beta_k^- - \beta_k^+$ and $c = \sum_k \log B(\alpha_k^-, \beta_k^-) - \log B(\alpha_k^+, \beta_k^+)$. The beta function B is a normalization constant. The reparametrization allows for standard optimization methods to determine $g(s)$ (cp. [9]). The monotony for the confidence dimension is kept by restricting the parameters $a_1, b_1 > 0$ as proposed in [9]. In summary, this model needs $2K + 1$ parameters.

3.4. Conditional Dependent Calibration

As demonstrated in Fig. 3, it can be beneficial to also capture correlations between the confidence and the bounding box quantities. If we keep the original formulation from Eq. 3, we can use multivariate probability density functions to define a calibration map while keeping track of possible relationships.

Dependent Logistic Calibration: Instead of using an independent normal distribution for each quantity separately, we can also insert two multivariate normal distributions for the conditional distributions $f(s|m)$ given in Eq. 3 and obtain

$$\ell r_{\text{LC}}(s) = \frac{1}{2} \left[(s_-^T \Sigma_-^{-1} s^-) - (s_+^T \Sigma_+^{-1} s^+) \right] + c, \qquad (8)$$

with $\mu^+, \mu^- \in \mathbb{R}^K$ and $\Sigma^+, \Sigma^- \in \mathbb{R}^{K \times K}$ as the mean vectors and covariance matrices for $m = 1$ and $m = 0$, respectively. Furthermore, we define $c = \log \frac{|\Sigma^-|}{|\Sigma^+|}$ as well as $s^+ = s - \mu^+$ and $s^- = s - \mu^-$. To preserve symmetric and positive definite covariance matrices during optimization, we estimate the entries of two matrices $V_+^{-1}, V_-^{-1} \in \mathbb{R}^{K \times K}$ and assume $\Sigma^{-1} = (V^T V)^{-1} = V^{-1}(V^{-1})^T$. This method needs $2(K^2 + K) + 1$ parameters to be optimized.

Dependent Beta Calibration: The extension of the independent beta calibration from section 3.3 is not that straight forward to enable a tracking of possible correlations
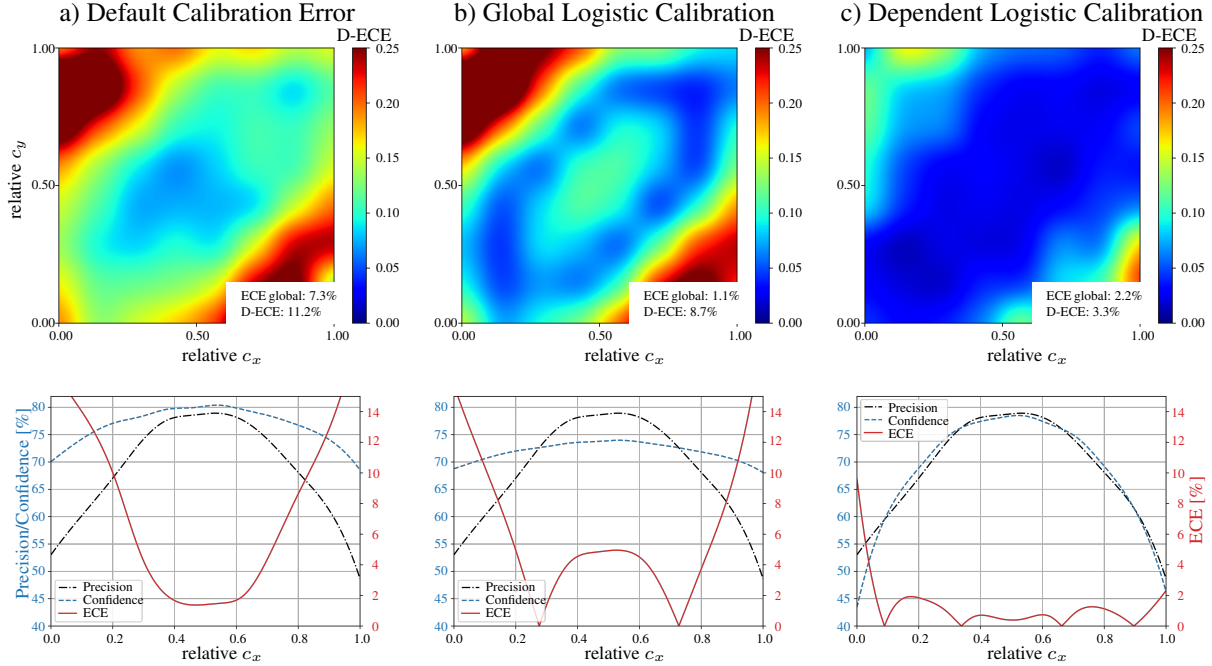
Figure 3: a) Consider a miscalibrated detector where the observed precision and the confidence score decrease closer to the image boundaries. Furthermore, assume a strong positive correlation with regard to c_x and c_y (top). Due to different slopes, precision and confidence exhibit a strong deviation at the boundaries and thus we have a high expected calibration error (ECE); especially in that border area (bottom). b) If we now apply a standard calibration map that only depends on the confidence score, we basically shift or in this case decrease the confidence curve (bottom). In other words, we optimize the average value of the ECE over all image locations, but don't influence the shape of the confidence curve. c) The proposed box-sensitive or multivariate calibration has access to location information and is capable of depicting dependencies between the confidence score, c_x and c_y. In this way, it is possible to adapt the confidence score across all locations differently.

between the variables. The natural multivariate extension of a beta distribution is a Dirichlet distribution. However, the random variables represented by a Dirichlet are defined over a K simplex which is not suitable for calibration. Therefore, we propose to use a generalized beta distribution for calibration established by [10] and defined as

$$f(s|\mathbf{m}) = \frac{1}{B(\alpha_0, ..., \alpha_K)} \frac{\prod_{k=1}^K \left[\lambda_k^{\alpha_k}(s_k^*)^{\alpha_k-1}\left(\frac{s_k^*}{s_k}\right)^2\right]}{\left[1 + \sum_{k=1}^K \lambda_k s_k^*\right]^{\sum_{k=0}^K \alpha_k}},$$

(9)

with the shape parameters α_k, β_k restricted to $\alpha_k, \beta_k > 0 \quad \forall k \in \{0, \ldots, K\}$. Using $\lambda_k = \frac{\beta_k}{\beta_0}$ and $s_k^* = \frac{s_k}{1-s_k}$, the resulting log-likelihood ratio $\ell r_{\mathrm{BC}}(s)$ is given by

$$\sum_{k=1}^K \left[\alpha_k^+ \log(\lambda_k^+) - \alpha_k^- \log(\lambda_k^-) + (\alpha_k^+ - \alpha_k^-) \log(s_k^*)\right] +$$

(10)

$$\sum_{k=0}^K \left[\alpha_k^- \log\left(\sum_{j=1}^K \lambda_j^- s_j^*\right) - \alpha_k^+ \log\left(\sum_{j=1}^K \lambda_j^+ s_j^*\right)\right] + c,$$

where $c = \log B(\alpha_0^-, ..., \alpha_k^-) - \log B(\alpha_0^+, ..., \alpha_k^+)$. Positive correlations are captured by the distribution, whereas negative correlations are neglected [17]. With this formulation, we need to optimize $4(K+1)+1$ parameters.

Multidimensional Histogram Binning: Histogram binning is one of the simplest methods to calibrate confidence estimates. The confidence space is divided into N equally sized bins to group all predictions of a classifier. This allows calculating the gap between average precision and confidence in each bin. Afterwards, a confidence map $g(s)$ is obtained by assigning a new confidence score to each prediction that equals the observed precision in the according bin. Instead of integrating the histogram binning in our framework (Eq. 2), we directly use a multidimensional binning for each dimension of s. This allows a fine-grained calculation of the average precision with respect to different box information like image location, width, and height to obtain new confidence scores. The number of parameters however grows exponentially with the number of dimensions $N_{\mathrm{total}} = \prod_{k=1}^K N_k$ and thus N_k needs to be chosen carefully for each dimension.

4. Measuring Miscalibration of Detectors

For classification problems, the expected calibration error (ECE) is used to measure the miscalibration of a model [14, 3]. In the classification domain, this score describes the expected deviation of the observed accuracy from the estimated confidence

$$\mathbb{E}_{\hat{p}}\Big[\big|\mathbb{P}(\hat{y} = y|\hat{p} = p) - p\big|\Big]. \qquad (11)$$

Since $\hat{p}$ is a continuous random variable, the ECE is approximated by binning the confidence space of $\hat{p}$ into N equally spaced bins where $I(n)$ is the set of all samples in a single bin. The ECE is thus approximated by

$$\text{ECE} = \sum_{n=1}^{N} \frac{|I(n)|}{|\mathcal{D}|} \cdot \Big|\text{acc}(n) - \text{conf}(n)\Big|, \qquad (12)$$

where $|\mathcal{D}|$ is the total amount of samples. The average accuracy and confidence in each bin is denoted by $\text{acc}(n)$ and $\text{conf}(n)$, respectively. Intuitively, we could use this formulation to measure the miscalibration of detectors as well. The example evaluation of different methods in Fig. 3 (top) however shows the weakness of the standard ECE. Comparing the different heatmaps for each method reveals a significant improvement of the confidence calibration when we use additional box information instead of a global calibration. However, this fact is reflected very poorly by the standard ECE score. In this case, the global calibration (ECE = 1.1%) outperforms the dependent logistic calibration (ECE = 2.2%) in terms of the ECE (cp. Fig. 3). The ECE uses the confidence of each sample independent of the box properties to apply binning and to calculate an average precision. But for object detection tasks, the average precision may vary in different areas of the image. However, our proposed methods for location-dependent calibration adjust the confidence differently for different image regions. This cannot be captured by the standard ECE - thus, a location-dependent calibration may not lead to an improvement of the standard ECE score.

For this reason, we also integrate the box information into the ECE calculation. Therefore, we define the detection ECE (D-ECE) as the expected deviation of the observed precision with respect to the given box properties. Using Eq. 1 we get

$$\mathbb{E}_{\hat{p},\hat{r}}\Big[\big|\mathbb{P}(m = 1|\hat{p} = p, \hat{y} = y, \hat{r} = r) - p\big|\Big]. \qquad (13)$$

Thus, the D-ECE for a given class y depends on the IoU and on the number of additional box information. If we calculate the expectation with respect to a subset of the variables, e.g. $\hat{p}$, calibration heatmaps can be generated (cp. Fig. 4).

Similar to Eq. 12, we calculate the multidimensional D-ECE by partitioning the confidence space as well as the box property space in each dimension k into N_k equally spaced bins. If we iterate over all bins in all dimensions, the total amount of bins is given by $N_{\text{total}} = \prod_{k=1}^{K} N_k$ and the D-ECE is thus given by

$$\text{D-ECE}_K = \sum_{n=1}^{N_{\text{total}}} \frac{|I(n)|}{|\mathcal{D}|} \cdot \Big|\text{prec}(n) - \text{conf}(n)\Big|, \qquad (14)$$

where $\text{prec}(n)$ and $\text{conf}(n)$ denote the average precision and confidence in each bin, respectively.

On the one hand, a higher number of bins leads to a more accurate evaluation of the D-ECE score, especially if the data density is a very steep or quickly changing function. On the other hand, the number of samples per bin decreases as we increase the number of bins. Thus, the estimation of the average precision and confidence in each bin is less accurate. Furthermore, due to the non-linearity and bin migration of samples, the D-ECE scores with different number of dimensions and/or bins cannot be compared directly to each other. Therefore, we suggest to evaluate calibration methods for object detection with a D-ECE of at least the same dimensionality as was used for calibration.

Consider the following example: Method A uses only the confidences for calibration without any additional box information ($K = 1$) while method B also takes the c_x position of the bounding boxes into account ($K = 2$). Method C calibrates with respect to all three quantities, confidence $\hat{p}$ and c_x, c_y position ($K = 3$). To compare these models, we need the D-ECE$_K$ ($K = 3$ with c_x, c_y position) because C uses the confidence and 2 additional box quantities for calibration. A comparison with the standard ECE is not reasonable. Note that calibration methods with other box information (e.g. width or height) should not be compared to B or C.

5. Experiments

5.1. Evaluation Protocol

We apply our extended calibration methods to several pretrained object detectors available in [4]. All networks use a non-max suppression with a probability threshold of 0.3 and an IoU threshold of 0.6. We use the COCO validation dataset 2017 [11] with images licensed for commercial use only to demonstrate the effectiveness of our approach. Unfortunately, we cannot use the according COCO test set as there are no annotations publicly available. Thus, we are using a random split of the validation set and use 70% for building the calibration scheme and the remaining images for testing. The random split is repeated 20 times to obtain an average calibration result for each method.

In general, one should use all influential factors (c_x, c_y, w, h) for measuring the D-ECE metric and for box-

		$(\hat{p})$	$(\hat{p},c_x,c_y)$	$(\hat{p},w,h)$	full
baseline		19.235	18.868	18.265	15.255
HB	conf.	**0.890**	5.121	5.181	5.790
	dep.	-	**4.557**	**3.739**	5.727
LC	conf.	1.282	5.244	5.321	5.752
	indep.	-	5.283	4.300	5.011
	dep.	-	5.310	4.118	**4.935**
BC	conf.	1.321	5.250	5.283	5.751
	indep.	-	5.296	4.557	5.097
	dep.	-	5.290	4.078	5.105

(a) F. R-CNN ResNet-50 [20], IoU@0.6.

		$(\hat{p})$	$(\hat{p},c_x,c_y)$	$(\hat{p},w,h)$	full
baseline		11.675	11.623	10.490	8.781
HB	conf.	**1.178**	5.392	4.332	4.930
	dep.	-	**4.991**	**3.360**	4.975
LC	conf.	1.232	5.404	4.437	4.891
	indep.	-	5.412	3.387	4.298
	dep.	-	5.460	3.419	**4.181**
BC	conf.	1.299	5.431	4.450	4.917
	indep.	-	5.460	3.646	4.237
	dep.	-	5.412	3.422	4.369

(b) R-FCN ResNet-101 [1], IoU@0.6.

		$(\hat{p})$	$(\hat{p},c_x,c_y)$	$(\hat{p},w,h)$	full
baseline		6.648	8.286	11.400	10.239
HB	conf.	**1.278**	5.766	10.723	7.429
	dep.	-	**5.615**	6.544	7.716
LC	conf.	1.306	6.085	10.707	7.575
	indep.	-	6.103	8.356	5.763
	dep.	-	6.017	5.638	5.111
BC	conf.	1.297	6.016	10.721	7.533
	indep.	-	6.036	5.265	5.090
	dep.	-	6.041	**4.715**	**4.916**

(c) SSD Inception v2 [12], IoU@0.6.

		$(\hat{p})$	$(\hat{p},c_x,c_y)$	$(\hat{p},w,h)$	full
baseline		32.458	31.805	30.990	26.468
HB	conf.	**0.910**	4.987	5.678	5.821
	dep.	-	**4.392**	**4.532**	6.340
LC	conf.	1.667	5.321	5.603	5.290
	indep.	-	5.279	5.419	4.883
	dep.	-	5.196	4.776	**4.755**
BC	conf.	1.266	5.367	5.370	5.201
	indep.	-	5.214	5.485	5.047
	dep.	-	5.327	4.733	4.920

(d) F. R-CNN ResNet-50 [20], IoU@0.75.

		$(\hat{p})$	$(\hat{p},c_x,c_y)$	$(\hat{p},w,h)$	full
baseline		26.087	25.207	23.862	20.394
HB	conf.	**1.014**	5.279	4.583	5.343
	dep.	-	**5.005**	4.171	5.967
LC	conf.	1.679	5.769	4.231	4.849
	indep.	-	5.768	4.099	4.740
	dep.	-	5.609	3.976	**4.734**
BC	conf.	1.485	5.769	4.188	4.785
	indep.	-	5.604	4.144	4.778
	dep.	-	5.756	**3.872**	4.772

(e) R-FCN ResNet-101 [1], IoU@0.75.

		$(\hat{p})$	$(\hat{p},c_x,c_y)$	$(\hat{p},w,h)$	full
baseline		19.668	18.761	19.470	18.067
HB	conf.	**1.262**	5.654	10.772	7.581
	dep.	-	**5.383**	5.846	7.718
LC	conf.	1.688	5.888	10.599	7.308
	indep.	-	5.872	7.700	4.958
	dep.	-	5.865	5.541	**4.130**
BC	conf.	1.796	6.002	10.649	7.366
	indep.	-	5.923	4.839	4.140
	dep.	-	5.893	**4.713**	4.184

(f) SSD Inception v2 [12], IoU@0.75.

Table 1: Calibration scores for different networks. The baseline is the default miscalibration of the network. Each column shows the D-ECE score for a specific dimensionality, that is determined by the additional bounding box information (global or confidence only, $(\hat{p},c_x,c_y)$, $(\hat{p},w,h)$ or full). Each row contains the evaluation of a calibration method (e.g. histogram binning (HB), logistic (LC) or beta (BC) calibration). Each method is subdivided into the categories 1) confidence only (conf.), 2) box information assumed to be independent (indep.) or 3) box information assumed to be dependent (dep.). The best results are mainly achieved by multidimensional histogram binning for confidence only calibration, whereas multivariate logistic and beta calibration outperform the histogram binning in the box-sensitive case.

sensitive confidence calibration. However, for anchor-free methods [25] or semantic segmentation only the position is available. Furthermore, for safety critical applications the calibration goal depends on the assurance case which could be independent of the scale. Thus, we evaluated several setups to demonstrate our generic approach. For the one-dimensional calibration case (confidence only), we use 15 bins for the histogram binning and 20 bins for the computation of the ECE. For $K = 3$ ($(\hat{p},c_x,c_y)$ or $(\hat{p},w,h)$), we use $N_k = 5$ for histogram-based calibration and $N_k = 8$ for the D-ECE. Finally, in the complete box-sensitive calibration case $K = 5$, the histogram calibration is performed with $N_k = 3$ bins in each dimension while using a binning of $N_k = 5$ for the D-ECE. The robustness of the D-ECE calculation is increased by neglecting bins with less than 8 samples.

5.2. Results

A baseline is obtained by evaluating each detector without calibration, but with D-ECE measures of different dimensionality. The results are presented in the first row of Table 1. In the first experiment, we apply global calibration methods without any additional box information. The results for histogram binning, logistic and beta calibration are presented in the first column of Table 1. In contrast to the observations for classification presented in [3], the simple histogram binning outperforms the logistic and the beta calibration in all cases. In the next two experiments (second and third column), additional bounding box quantities like position or scale are evaluated separately. Finally, all bounding box quantities are used for calibration (fourth column). As proposed in the previous section, the computation of the D-ECE is always chosen according to the dimensionality. For the evaluation, we applied the same IoU score of 0.6 as used for the non-max suppression of the networks. In addition, we also measured the miscalibration for a higher IoU score of 0.75. The results for different network architectures and different IoU scores are shown in Table 1 with the best D-ECE scores highlighted. Fig. 4 illustrates the improved calibration in terms of a heatmap. The effect of box-sensitive confidence calibration is also qualitatively demonstrated in Fig. 5.

The results reveal that the simple histogram binning achieves the lowest calibration error for the global calibration with confidence only and fair results for calibration with respect to either $(\hat{p},c_x,c_y)$ or $(\hat{p},w,h)$. For the confidence calibration with all box information available, the multivariate logistic or the beta calibration outperform the histogram binning, especially in the dependent case. Thus, it is preferable to keep track
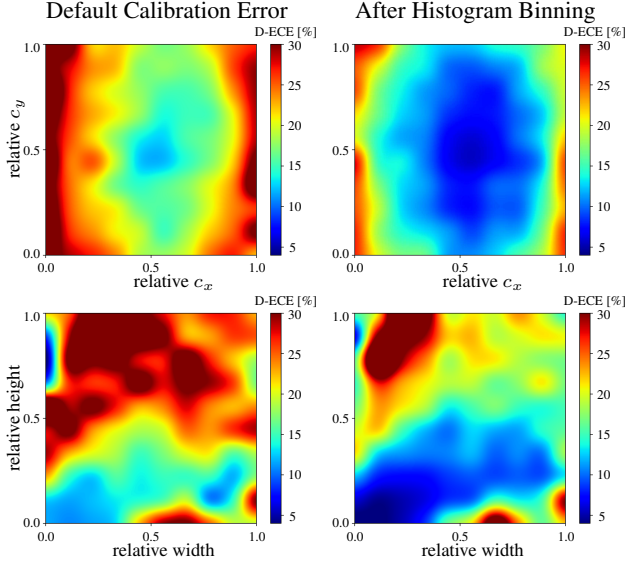
Figure 4: Calibration heatmap of R-FCN [1, 4] detections on COCO with IoU @ 0.6. The left column shows the baseline miscalibration measured over the relative c_x and c_y coordinates (top row) and the relative width/height (bottom row) respectively. The right column shows the D-ECE scores after calibration with histogram binning that is also able to capture those additional box information. For the corner case, low relative width and high relative height, there are only a few data samples available. Thus, the D-ECE calculation is likely to be inaccurate in that area.

of correlations between the box quantities in order to improve the D-ECE score, especially for high dimensional calibration mappings. In general, the histogram binning has offered a high representational power for detection calibration but in the case of high dimensionality it is outperformed by the other methods. It should be clear that an increasing number of bins makes the histogram binning prone to overfitting and, in addition, the amount of bins grows exponentially with the number of dimensions. A representative histogram-based calibration mapping requires a dataset that compensates for the exponential bin growth. Therefore, we conclude that either the logistic calibration or beta calibration should be preferred for high dimensional calibration mappings with only a small or medium amount of data available, whereas the histogram binning might be used for low-dimensional cases or on large datasets.

6. Conclusion

In this paper we examined how to extend common calibration methods to the object detection task. We claim that the calibration depends on the regression output of the detector, which is conform with our empirical observations
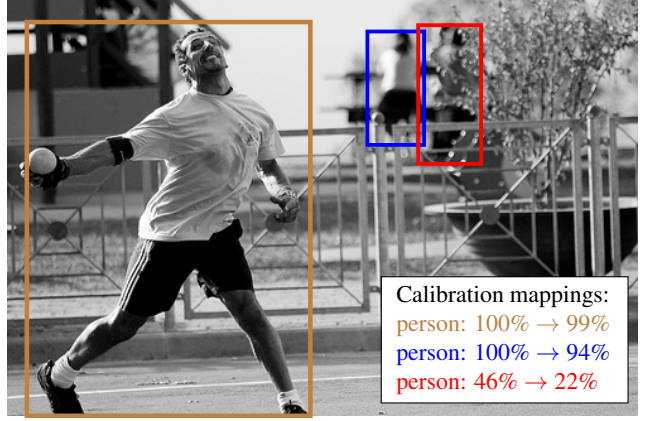

Figure 5: Qualitative R-FCN calibration results for predictions on COCO with equal confidence but with different positions and scales. Using the dependent logistic calibration, the confidence of the brown detection is slightly adapted, whereas the remaining detections undergo major changes to indicate the uncertainty for predictions with those confidences, positions, and scales.

on the COCO dataset. Therefore, we propose a novel framework to measure miscalibration and to calibrate object detection models. To reveal and overcome the limitations of traditional methods, a location and scale dependent treatment of measuring miscalibration is needed. The same holds for the calibration methodology. Using the proposed framework, we are able to outperform state-of-the-art calibration methods by including additional bounding box information from the detector. The proposed framework is also useful to measure the miscalibration with respect to the box regression output. We note that this new measurement provides insights into the properties of a detector that are not revealed by any traditional evaluation metric. We have shown that multidimensional histogram binning is a good choice for low dimensional calibration mappings. The multivariate logistic or beta calibration achieve superior results for high dimensional mappings using considerably fewer parameters.

Acknowledgments

References

[1] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 379–387, 2016.

[2] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, August 2017.

[4] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

[5] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–799, 2018.

[6] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems (NIPS)*, pages 5574–5584, 2017.

[7] Florian Kraus and Klaus Dietmayer. Uncertainty estimation in one-stage object detection. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 53–60. IEEE, 2019.

[8] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *International Conference on Machine Learning (ICML)*, pages 2801–2809, 2018.

[9] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631, 2017.

[10] David L Libby and Melvin R Novick. Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational Statistics*, 7(4):271–294, 1982.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot MultiBox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016.

[13] Mahdo Naeini and Gregory Cooper. Binary Classifier Calibration Using an Ensemble of Near Isotonic Regression Models. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 360–369, December 2016.

[14] Mahdo Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2901–2907, 2015.

[15] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection. In *Workshop on Machine Learning for Intelligent Transportation Systems (NIPS)*, 2018.

[16] Andrew Y. Ng and Michael I. Jordan. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[17] Ingram Olkin and Thomas A Trikalinos. Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54–60, 2015.

[18] Buu Phan, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Taylor Denouden, and Sachin Vernekar. Calibrating Uncertainties in Object Localization Task. In *Third workshop on Bayesian Deep Learning (NIPS)*, 2018.

[19] John Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[21] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[22] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for Single-Shot Confidence Calibration in Deep Neural Networks Through Stochastic Inferences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Bianca Zadrozny and Charles Elkan. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 609–616, 2001.

[24] Bianca Zadrozny and Charles Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 694–699, 2002.

[25] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 840–849, 2019.