Uncertainty-Aware Online Extrinsic Calibration: A Conformal Prediction Approach

Mathieu Cocheteux, Julien Moreau, Franck Davoine

▶ **To cite this version:**

Mathieu Cocheteux, Julien Moreau, Franck Davoine. Uncertainty-Aware Online Extrinsic Calibration: A Conformal Prediction Approach. Winter Conference on Applications of Computer Vision (WACV), IEEE; CVF, Feb 2025, Tucson, United States. hal-04881740

HAL Id: hal-04881740
https://hal.science/hal-04881740v1

Submitted on 12 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Uncertainty-Aware Online Extrinsic Calibration: A Conformal Prediction Approach

Mathieu Cocheteux[1], Julien Moreau[1], Franck Davoine[2]
[1]Université de technologie de Compiègne, CNRS, Heudiasyc, France
[2]CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, France
{mathieu.cocheteux, julien.moreau}@hds.utc.fr, franck.davoine@cnrs.fr

Abstract

Accurate sensor calibration is crucial for autonomous systems, yet its uncertainty quantification remains underexplored. We present the first approach to integrate uncertainty awareness into online extrinsic calibration, combining Monte Carlo Dropout with Conformal Prediction to generate prediction intervals with a guaranteed level of coverage. Our method proposes a framework to enhance existing calibration models with uncertainty quantification, compatible with various network architectures. Validated on KITTI (RGB Camera-LiDAR) and DSEC (Event Camera-LiDAR) datasets, we demonstrate effectiveness across different visual sensor types, measuring performance with adapted metrics to evaluate the efficiency and reliability of the intervals. By providing calibration parameters with quantifiable confidence measures, we offer insights into the reliability of calibration estimates, which can greatly improve the robustness of sensor fusion in dynamic environments and usefully serve the Computer Vision community.

1. Introduction

In the dynamic field of autonomous systems, extrinsic calibration, which determines the spatial relationships between sensors, is essential for effective data fusion from multiple sensors. The calibration quality can thus directly impact subsequent tasks such as object detection or segmentation. In real-world scenarios, even slight calibration errors can significantly impact the safety and performance of autonomous vehicles and robots [5].

Traditional calibration methods, relying on manual procedures or controlled environments, are increasingly inadequate for the demands of modern autonomous systems. These approaches, while precise in controlled settings, do not allow for on-the-fly calibration, and thus to maintain a correct calibration in operation. This limitation has created a pressing need for robust techniques capable of real-time online calibration in natural environments.

Recent years have seen significant advancements in the state of the art for online calibration, particularly with the integration of deep learning-based methods. Models such as those proposed in [6, 7, 18, 26, 33, 37, 44] have demonstrated remarkable improvements in calibration efficiency and accuracy. However, the quantification of uncertainty in the calibration process is yet to be studied. Quantifying the reliability of calibration estimates is essential in ensuring a consistent calibration quality.

In this context, we focus on model uncertainty (epistemic uncertainty), which reflects the confidence of the model in its predictions. Unlike data uncertainty (aleatoric uncertainty), which is inherent and irreducible, model uncertainty can be mitigated with more data or improved models [20]. This focus is crucial for calibration tasks, where the reliability of the model's predictions directly impacts system performance and safety.

To address this challenge, we propose an approach that integrates Monte Carlo Dropout (MCD) [11] with Conformal Prediction (CP) [42]. Our method generates prediction intervals with statistically guaranteed coverage—the probability that the true outcome falls within the predicted interval—enabling robust quantification of calibration parameter uncertainty in dynamic environments. To the best of our knowledge, this work pioneers the study of uncertainty in online extrinsic calibration and introduces the first CP-based framework for providing statistically guaranteed intervals in this context. Our key contributions are:

- A Conformal Prediction [42] framework tailored for online extrinsic calibration, providing reliable and statistically sound uncertainty estimates.
- A deep learning-based approach that seamlessly integrates Monte Carlo Dropout [11] for model uncertainty estimation with Conformal Prediction [42], enabling uncertainty quantification in dynamic environments.

- A comprehensive validation on real-world benchmark datasets with different sensor modalities (KITTI [14] for RGB-LiDAR and DSEC [13] for Event Camera-LiDAR calibration), demonstrating the generalization and effectiveness of uncertainty-aware calibration across different sensor types.

2. Related Work

We would like to draw the attention of the reader on the two different uses of the word *calibration* in this work. *Sensor calibration*, or *extrinsic calibration*, refers to the spatial transformation between sensors. Conversely, *uncertainty calibration* refers to a process part of the CP method and described in Section 3.2.1.

2.1. Extrinsic Calibration for Multi-Sensor Systems

Extrinsic calibration, the process of estimating spatial relationships between heterogeneous sensors, is crucial for accurate data fusion in autonomous driving and robotics [30]. Traditional methods often relied on manual procedures or controlled environments [45], which proved impractical for dynamic, real-world scenarios. This limitation has driven research towards automated, robust, and online calibration techniques.

Early work in automated calibration saw significant advancements. [31] introduced an automatic extrinsic calibration method for LiDAR-camera systems using mutual information maximization. [15] proposed a single-shot approach for camera and range sensor calibration, leveraging checkerboards for feature extraction.

Lead by the seminal work of Schneider *et al*. [35], recent years have witnessed significant advancements with the apparition of deep learning-based calibration techniques [6, 7, 18, 26, 33, 37, 44], demonstrating the potential of end-to-end learning in handling complex spatial relationships between sensors in uncontrolled environments. While these methods have improved calibration accuracy, they do not explore uncertainty quantification. Our work addresses this gap by proposing a framework to provide uncertainty estimates and safe intervals for existing calibration models. We demonstrate our approach using slightly modified versions of state-of-the-art lightweight models UniCal [6] and MuLi-Ev [7] as case studies. This framework aims to enhance the reliability and interpretability of extrinsic calibration in dynamic, real-world scenarios.

2.2. Uncertainty Quantification in Computer Vision

Uncertainty quantification has become indispensable in computer vision, particularly for safety-critical applications [1, 20]. Modern deep learning models contend with both aleatoric (data-inherent) and epistemic (model-related) uncertainties [20]. Recent advancements have yielded diverse techniques to address these challenges.

Bayesian Neural Networks (BNNs) provide a principled approach by approximating posterior distributions over network weights [27]. To mitigate their computational complexity, methods like Monte Carlo Dropout have emerged, interpreting dropout as a Bayesian approximation during inference [11]. This enables efficient uncertainty estimation in large-scale vision applications with minimal architectural modifications.

Ensemble methods, such as Deep Ensembles [23], aggregate predictions from multiple independently trained models, effectively capturing model uncertainty while enhancing predictive performance. In medical imaging, the Probabilistic U-Net [22] combines U-Net architecture with a conditional variational autoencoder to generate multiple plausible segmentations, addressing inherent ambiguities.

Prior Networks [29] explicitly model distributional uncertainty, crucial for distinguishing various uncertainty types, including out-of-distribution samples.

These methods have demonstrated efficacy across various computer vision tasks, including object detection [17], semantic segmentation [19], and depth estimation [20]. In autonomous driving, uncertainty estimates help identify low-confidence situations, potentially triggering safety interventions [10]. However, challenges remain, including computational overhead and distributional assumptions, presenting opportunities for future research in efficient and scalable uncertainty quantification for real-world vision systems.

2.3. Conformal Prediction (CP)

CP has emerged as a powerful framework for uncertainty quantification, offering distribution-free guarantees for prediction intervals [2, 42]. It has a low computational cost at runtime, and works with minimal assumptions (mostly exchangeability). Unlike traditional methods that rely on specific distributional assumptions, CP provides valid prediction sets under the minimal assumption of exchangeability, making it widely applicable across various domains [36].

The core principle of CP lies in its ability to construct prediction intervals that contain the true outcome with a user-specified probability, regardless of the underlying data distribution [40, 43]. This is achieved through a nonconformity measure, which quantifies the dissimilarity between a new example and a set of previously observed examples [40]. The resulting prediction intervals adapt to the complexity of the data, offering tighter bounds in regions of high confidence and wider intervals where uncertainty is greater [3].

Several variants of CP have been developed to enhance its efficiency and applicability. Inductive Conformal Prediction [32] simplifies the original framework by keeping out an uncertainty calibration set on which are computed nonconformity scores, reducing computational complexity for

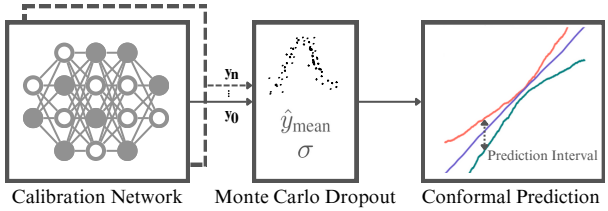


Figure 1. Overview of the uncertainty-aware calibration pipeline. (Left) The deep learning Calibration Network estimates parameters from sensor data. (Center) MCD is applied to generate multiple predictions, producing a mean estimate $\hat{y}_{\text{mean}}$ (the prediction) and a standard deviation σ (measuring the uncertainty). (Right) The CP method is beforehand calibrated on a separate uncertainty calibration set, then can be used to estimate intervals with a $1-\alpha$ coverage.

large datasets. Split Conformal Prediction [24] further refines this approach, providing a simple yet powerful method for constructing prediction intervals in regression tasks.

Recent advancements have focused on integrating CP with modern machine learning techniques. Conformalized Quantile Regression [34] combines CP with quantile regression, yielding more efficient prediction intervals. The Jackknife+ method [4] offers a computationally efficient alternative that produces asymptotically valid prediction intervals under weaker assumptions.

The distribution-free nature of CP has led to its successful application in various fields, including computer vision [3], medical diagnosis [25], and time series forecasting [39]. In computer vision, CP has been used to provide uncertainty estimates for image-to-image regression tasks [3], demonstrating its potential for enhancing the reliability of deep learning models.

To date, the application of CP in the field of sensor calibration remains unexplored. Our work aims to integrate CP with deep learning models to develop a novel framework for online extrinsic calibration with reliable uncertainty estimates.

3. Method

This section introduces our uncertainty-aware online extrinsic calibration approach, which is built to be used on top of a deep learning extrinsic calibration model. In this work, we conduct experiments using models based on [6,7]. Figure 1 illustrates our proposed calibration pipeline. We focus on quantifying model uncertainty (epistemic) rather than data uncertainty (aleatoric) for two key reasons: (1) model uncertainty is reducible through improved modeling and additional data, critical for enhancing calibration reliability, and (2) in our context, it is mostly linked to small variations in time synchronization, with an impact much smaller than that of the error due to the model quality. Our approach generates prediction intervals for calibration parameters with coverage guarantees, ensuring reliable calibration in dynamic environments.

In real-world application, the extrinsic calibration parameters would then be updated in real-time using the mean prediction from MCD, and the uncertainty estimate and width interval from CP would be used to adjust the robustness of the system. If the uncertainty exceeds a certain threshold, the system could trigger a recalibration process or adjust the confidence in sensor fusion tasks.

The following subsections detail our uncertainty estimation techniques and the application of CP.

3.1. Dropout as a Bayesian Approximation

Bayesian Neural Networks [27] are good at estimating model uncertainty, but are often too computationally heavy for real-time applications. We thus employ a lighter method, Monte Carlo Dropout (MCD), leveraging its interpretation as an approximate Bayesian inference method [11]. By applying dropout during both training and inference, we sample from the approximate posterior distribution of the network's weights, enabling uncertainty estimation in our calibration predictions.

Given a neural network $f(x;\theta)$ with weights θ, MCD generates N stochastic forward passes, each with a different dropout mask:

$$\hat{y}_i = f(x;\theta_i), \quad i = 1,\dots,N \tag{1}$$

where θ_i is the randomly masked weights. The mean prediction and model uncertainty are then estimated as:

$$\hat{y}_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i, \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \hat{y}_{\text{mean}})^2} \tag{2}$$

To optimize MCD for real-time online calibration, where we process single data points sequentially, we could implement a parallel execution strategy, which consists in replicating the input N times and treat it as a N-size batch, then perform N forward passes simultaneously with different dropout masks $\{\theta_i\}$. This approach would reduce computational overhead, making it suitable for real-time applications without compromising uncertainty estimation quality.

3.2. Building Intervals with Conformal Prediction

To complement MCD and provide a guarantee on the prediction intervals, we integrate CP. CP is a distribution-free technique which, given a desired maximal error rate α defining a coverage level $1-\alpha$, produces a prediction interval that is guaranteed to contain the true calibration parameters with at least a $1-\alpha$ probability.

Our method is particularly inspired by the implementation in [3], which provides a framework for generating CP intervals from a scalar uncertainty measure.

3.2.1 Nonconformity Measure and Calibration

Our CP method integrates uncertainty estimation with rigorous statistical guarantees [2]. We define a nonconformity measure that quantifies the discrepancy between predictions and ground truth, accounting for model uncertainty [36].

Given an uncertainty calibration set $\{(y_k^{\text{true}}, \hat{y}_k, \hat{\sigma}_k)\}_{k=1}^{m}$, where m is the number of samples, $\hat{y}_k$ is the predicted calibration parameter, y_k^{true} is the ground truth, and $\hat{\sigma}_k$ is the estimated uncertainty from MCD, we compute the nonconformity score s_k for each sample:

$$s_k = \frac{|\hat{y}_k - y_k^{\text{true}}|}{\hat{\sigma}_k} \tag{3}$$

This score, inspired by the Mahalanobis distance [28], normalizes the prediction error by the estimated uncertainty, allowing for adaptive confidence intervals that account for varying levels of uncertainty across different predictions [34].

The uncertainty calibration phase involves computing these scores for the entire calibration set, resulting in $\{s_1, s_2, \ldots, s_m\}$. This set forms the basis for determining the appropriate quantile used in constructing prediction intervals, ensuring that our method maintains the desired coverage level across diverse scenarios [24, 42].

3.2.2 Quantile Determination

To ensure that the prediction intervals have the desired coverage level $1 - \alpha$, we compute a quantile $Q_{1-\alpha}$ from the nonconformity scores obtained in the calibration phase. Specifically, the quantile $Q_{1-\alpha}$ is determined by sorting the nonconformity scores in ascending order $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(m)}$ and selecting the $(m+1)(1-\alpha)$-th score:

$$Q_{1-\alpha} = s_{(\lceil (m+1)\cdot(1-\alpha) \rceil)} \tag{4}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. This selection ensures that the proportion of nonconformity scores less than or equal to $Q_{1-\alpha}$ is at least $1 - \alpha$, providing the desired coverage.

3.2.3 Prediction Interval Computation

Finally, for a new test input, we use the quantile $Q_{1-\alpha}$ to compute the prediction interval for the extrinsic calibration parameter. Given a new prediction $\hat{y}$ and its associated uncertainty $\hat{\sigma}$, the prediction interval is calculated as:

$$\text{Prediction Interval} = [\hat{y} - Q_{1-\alpha} \cdot \hat{\sigma}, \hat{y} + Q_{1-\alpha} \cdot \hat{\sigma}] \tag{5}$$

This interval provides a guarantee that the true calibration parameter will fall within the prediction interval with at least probability $1 - \alpha$.

By following these steps, our method leverages both the predictive power of deep learning models and the robustness of CP to provide uncertainty-aware extrinsic calibration intervals that are reliable in dynamic environments.

3.3. Implementation Details

Our calibration uncertainty framework is demonstrated on existing deep learning-based calibration models [6, 7], which serve as the backbone for extrinsic calibration. The implementation of MCD requires minimal modifications to the original architecture, and CP can be treated as a post-processing step. Training and inference are conducted using PyTorch, and the models are evaluated on real-world datasets. The models were trained from scratch using NVIDIA V100 GPUs. To facilitate the implementation of our CP method, we utilize the Fortuna [8] framework. A separate uncertainty calibration set is used to establish the nonconformity thresholds, ensuring that the prediction intervals are valid across diverse scenarios. In line with what is done in [6, 7], we introduce artificial decalibrations on the input during the training and testing. As we are mostly interested in the range of small decalibrations most often encountered in real-world scenarios, and requiring the best accuracy, with $\pm 1^\circ$ and $\pm 10cm$ on all axes.

4. Experiments

4.1. Datasets

We evaluate our uncertainty-aware online extrinsic calibration method on two datasets: KITTI [14], which provides synchronized RGB images and 64-channel LiDAR data, and DSEC [13], which offers event camera data and 16-channel LiDAR. These datasets cover diverse sensor modalities and conditions, enabling a robust assessment of our approach.

For KITTI, we split the data into training (60%), validation (15%), calibration (15%), and testing (10%) sets. For DSEC, we use 70% for training, 15% for validation, and 15% for calibration, with a separate test set. The calibration subsets are used to compute nonconformity scores for the conformal prediction method, ensuring well-calibrated uncertainty intervals.

4.2. Evaluation Metrics

To rigorously evaluate the performance of our method, we employ a set of well-established metrics specifically tailored for CP intervals. These metrics assess both the reliability and efficiency of the prediction intervals generated by our approach.

4.2.1 Prediction Interval Coverage Probability (PICP)

PICP is a widely used metric [21, 38], which quantifies the proportion of true calibration parameters that fall within the predicted intervals. PICP is crucial for assessing the reliability of the prediction intervals, ensuring that the true values are captured within the intervals at the desired coverage level. For our method, we calculate PICP as follows:

$$\text{PICP} = \frac{1}{m}\sum_{k=1}^{m} \mathbb{I}\left(y_k \in [\hat{y}_k^{\text{lower}}, \hat{y}_k^{\text{upper}}]\right) \tag{6}$$

where y_k represents the true calibration parameter, $\hat{y}_k^{\text{lower}}$ and $\hat{y}_k^{\text{upper}}$ denote the lower and upper bounds of the predicted interval, and $\mathbb{I}(\cdot)$ is the indicator function. A PICP close to the desired coverage level (*e.g.* 90%) indicates that the prediction intervals are reliable.

4.2.2 Mean Prediction Interval Width (MPIW)

MPIW [9] measures the average width of the prediction intervals, providing insight into the trade-off between interval width and coverage. It is defined as:

$$\text{MPIW} = \frac{1}{m}\sum_{k=1}^{m}(\hat{y}_k^{\text{upper}} - \hat{y}_k^{\text{lower}}) \tag{7}$$

While narrower intervals are generally preferred, they must still maintain the desired coverage as indicated by the PICP. MPIW helps quantify this balance, with lower MPIW values being more desirable provided that the PICP is maintained close to the target coverage level.

4.2.3 Interval Score (IS)

IS [16] provides a balanced evaluation by penalizing both the width of the prediction intervals and any instances where the intervals fail to cover the true value. IS is particularly useful for assessing the efficiency of the intervals, as it encourages intervals that are both narrow and reliable. The Interval Score for our method is computed as:

$$\begin{aligned}\text{IS} = \frac{1}{m}\sum_{k=1}^{m}\Big[&(\hat{y}_k^{\text{upper}} - \hat{y}_k^{\text{lower}}) \\ &+ \frac{2}{\alpha}\cdot(\hat{y}_k^{\text{lower}} - y_k)\cdot\mathbb{I}(y_k < \hat{y}_k^{\text{lower}}) \\ &+ \frac{2}{\alpha}\cdot(y_k - \hat{y}_k^{\text{upper}})\cdot\mathbb{I}(y_k > \hat{y}_k^{\text{upper}})\Big]\end{aligned} \tag{8}$$

where α is the significance level (*e.g.* 0.1 for a $1-\alpha = 90\%$ coverage interval). A lower IS indicates better overall performance, reflecting more efficient and reliable intervals.

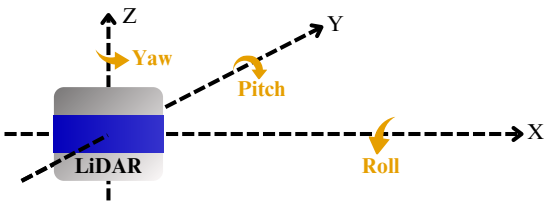


Figure 2. Axes of rotation and translation of the spatial transformation in the LiDAR frame.

4.3. Results

Throughout this subsection, results are analyzed for each axis in translation and in rotation. These axes are represented in Figure 2. As we pioneer the introduction of uncertainty estimation and interval prediction for online extrinsic calibration, we will mostly assess the quality of our results in the absolute, and show that they provide added value by providing uncertainty estimates of which we will demonstrate the robustness.

4.3.1 Interval Quality Analysis

To evaluate our method, we analyze the prediction interval quality across the KITTI and DSEC datasets. Table 1 shows results for three key metrics: Prediction Interval Coverage Probability (PICP), Mean Prediction Interval Width (MPIW), and Interval Score (IS) at different target coverage levels.

Our method demonstrates consistent and reliable performance, with PICP values closely matching target coverage levels across both datasets. This robustness holds despite the datasets' diverse sensor modalities and environmental conditions, highlighting the method's effectiveness in maintaining desired uncertainty bounds.

Analysis of IS and MPIW metrics reveals important insights. For translational estimates, the Y-axis consistently achieves the lowest MPIW and IS values, indicating precise lateral translation estimates crucial for accurate lane-level localization in autonomous driving.

Exceptional precision is observed in Roll and Yaw estimates, with the KITTI dataset showing 99% coverage MPIWs of 0.26° and 0.29°, and corresponding IS values of 0.26° and 0.30°. These narrow intervals suggest high confidence and accuracy in rotational estimates.

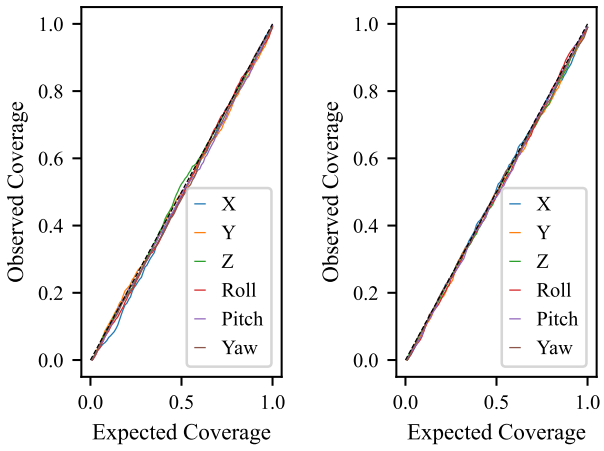
However, Pitch estimation shows greater uncertainty compared to other rotational parameters, with a 99% coverage MPIW of 0.66° (IS: 0.67°) in KITTI and 0.88° (IS: 0.89°) in DSEC, indicating inherent challenges likely due to sensor limitations (see Section 4.4.3).

Overall, these results demonstrate the method's efficacy in providing well-calibrated uncertainty estimates, balancing reliability and precision.

Dataset	Metric	X (cm) Target Coverage			Y (cm) Target Coverage			Z (cm) Target Coverage			Roll (°) Target Coverage			Pitch (°) Target Coverage			Yaw (°) Target Coverage		
		90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%
KITTI [6, 14]	PICP (%)	89.4	95.3	98.8	88.7	94.3	98.9	89.2	94.0	98.6	89.3	93.7	99.0	90.5	94.5	99.3	88.8	94.9	98.6
	MPIW ↓	2.81	3.62	5.17	1.64	2.14	3.42	2.43	3.10	4.65	0.14	0.17	0.26	0.31	0.39	0.66	0.15	0.19	0.29
	IS ↓	3.56	3.96	5.27	2.17	2.46	3.55	3.35	3.63	4.86	0.17	0.19	0.26	0.37	0.42	0.67	0.19	0.21	0.30
DSEC [7, 13]	PICP (%)	88.2	94.5	99.0	90.1	94.6	98.8	90.2	95.5	98.9	92.0	95.3	98.9	90.7	95.8	99.2	90.4	94.2	98.8
	MPIW ↓	1.45	1.95	3.09	1.20	1.54	2.32	2.09	2.79	4.06	0.09	0.12	0.20	0.36	0.49	0.88	0.13	0.17	0.27
	IS ↓	1.83	2.12	3.13	1.45	1.66	2.34	2.09	3.01	4.11	0.12	0.13	0.20	0.47	0.55	0.89	0.16	0.19	0.27

Table 1. Evaluation metrics for our uncertainty-aware online extrinsic calibration method on KITTI [14] and DSEC [13] datasets. We report Prediction Interval Coverage Probability (PICP), Mean Prediction Interval Width (MPIW), and Interval Score (IS) for different target coverage levels across six degrees of freedom. Lower values of IS and MPIW indicate more precise and tighter prediction intervals.

Figure 3. LiDAR point cloud projections onto an image frame from KITTI [14] under varying Pitch calibration. (Top) Ground truth Pitch minus 0.25°. (Middle): Ground truth Pitch. (Bottom) Ground truth Pitch plus 0.25°. This ±0.25° range represents an extreme case of our 90% confidence interval for Pitch. The visual differences in projections are minimal, demonstrating the high precision and practical robustness of our calibration method in challenging scenarios.



(a) KITTI [14] (b) DSEC [13]

Figure 4. Calibration curves for extrinsic parameters on the KITTI [14] and DSEC [13] datasets, showing observed versus expected coverage for each degree of freedom. Better seen on screen.

As seen in Figure 3, even on the Pitch axis, and in outlier cases where the predicted 90% interval is among the widest predicted by our network, at 0.5° (*i.e.* ±0.25°), the difference between lower and higher bounds of the interval remains visually barely visible, as desired.

4.3.2 Quality of the Uncertainty Calibration

To assess the calibration quality of our uncertainty estimates, we present calibration curves for both datasets in Figures 4a and 4b. Both figures show the alignment of observed and expected coverage probabilities with the diagonal, indicating well-calibrated uncertainties across different coverage levels. For KITTI (Figure Figure 4a), translational parameters (X, Y, Z) demonstrate excellent calibration, with slight deviations in some rotational parameters. The DSEC results (Figure Figure 4b) exhibit a similar trend, confirming consistency with the PICP values in Table 1.

4.3.3 Interval Visualization

To provide a more intuitive understanding of our method's performance, we present ordered interval plots for the KITTI dataset in Figure 5. Those for DSEC can be found in the supplementary materials. These plots offer a visual representation of the prediction intervals and their relationship to the ground truth values for each degree of freedom.

We observe that the prediction intervals consistently envelop the ground truth values. This visual confirmation aligns with the PICP values reported in Table 1 close to the target coverage. The intervals appear to widen at the extremes of the value range, indicating increased uncertainty

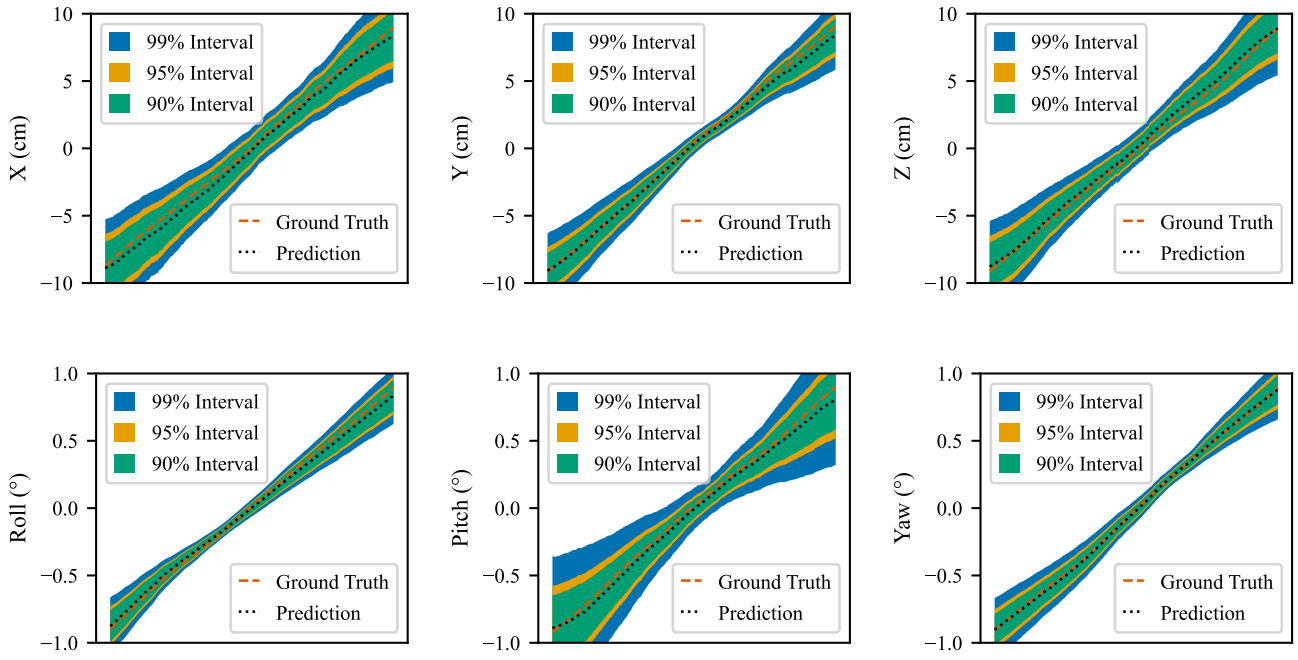


Figure 5. Ordered prediction interval plots for the six degrees of freedom (in translation and rotation) on the KITTI [14] dataset. The overlaid shaded regions represent the intervals corresponding to expected coverage levels of 90%, 95%, and 99%. The ground truth values should fall within the respective intervals for at least the specified proportion of samples. The X-axis represents the ordered test samples sorted by the ground truth values for each degree of freedom, while the Y-axis indicates the deviation from the ground truth. All curves are smoothed using a moving average to enhance readability.

in these regions. This increased uncertainty is especially visible on the Pitch interval curves, which are much wider. This behavior demonstrates our method's ability to adapt its uncertainty estimates based on the difficulty of the calibration task in different scenarios.

4.4. Discussion and Ablation Studies

The results discussed above demonstrate the effectiveness of our method across different datasets and sensor modalities. It is important to mention that these results are reached without deteriorating the original accuracy of the calibration model. For example, we achieved on KITTI an average rotation error of $0.04°$ and a translation error of $0.46\,\text{cm}$, comparable to [6] (rotation errors between $0.03°$ and $0.04°$, and translation errors from $0.33\,\text{cm}$ to $0.89\,\text{cm}$). To reflect on these results and their robustness, we conducted ablations studies and discuss our method's strengths and weaknesses below.

4.4.1 Impact of Monte Carlo Dropout Parameters

We conducted an ablation study to investigate the effect of varying the number of forward passes in MCD on uncertainty estimation. Figure 6 illustrates the relationship between the number of forward passes and IS for both translation and rotation errors.

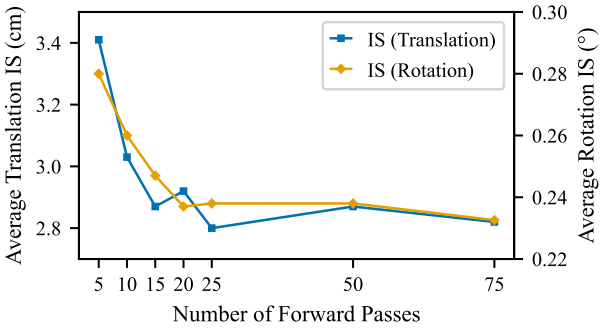


Figure 6. Impact of the number of Monte Carlo Dropout forward passes on the Interval Score (IS) averaged for translation parameters and rotation parameters. Experiment realized on KITTI [14].

The results reveal a clear trend: as the number of forward passes increases, IS decreases, indicating more precise uncertainty estimates. This trend is particularly pronounced in the range of 5 to 25 forward passes for both translation and rotation errors. Beyond 25 passes, the rate of improvement

Target (%)	PICP (%)					
	X	Y	Z	Roll	Pitch	Yaw
90	77.3	90.3	76.9	79.69	85.0	80.6
95	82.3	92.0	81.1	83.85	88.9	85.8

Table 2. Results of the ablation study on KITTI [14] showing the observed coverage (PICP) when using only MCD with a normal distribution assumption, instead of our proposed MCD+CP method.

diminishes significantly, suggesting a point of diminishing returns. For an optimal balance between computational efficiency and uncertainty estimation quality, we selected 25.

As demonstrated in [41], the optimal dropout rate for estimating epistemic uncertainty depends on the network architecture and its size. Following the grid search approach outlined in [11, 12], we tested rates between 0.05 and 0.5. We found that the smallest rate yielding consistent results was optimal, as increasing the rate degraded performance, though with a lesser impact than the number of forward passes. Consequently, a dropout rate of 0.25 was applied to the backbone layers, while the best results were achieved with 0.05 on KITTI and 0.10 on DSEC for the head.

4.4.2 Necessity of Applying Conformal Prediction

To evaluate the effectiveness of our CP approach, we performed an ablation study on KITTI using MCD [11] alone, with a normal distribution assumption for interval estimation. In this setup, MCD was applied during inference by performing multiple forward passes with dropout enabled. The mean and standard deviation of these predictions were then calculated to estimate quantiles under the normal distribution. This method differs from our primary approach, where σ is directly incorporated into the CP framework.

As shown in Table 2, the MCD with a normal distribution assumption consistently underestimates interval widths, resulting in lower-than-expected coverage rates at both 90% and 95% target levels. This underperformance likely arises from the inadequacy of the normal distribution in capturing true uncertainty and the inherent limitations of MCD in providing calibrated uncertainty estimates for this task. These findings highlight the superiority of our CP-based method for generating more accurate and reliable prediction intervals in extrinsic calibration.

4.4.3 Impact of LiDAR Vertical Resolution

Our experiments reveal a significant pattern in the uncertainty estimates, particularly for the Pitch angle. We observe consistently wider prediction intervals for Pitch compared to Roll and Yaw, which we attribute to the inherent limitations of LiDAR vertical resolution. This sparse vertical sampling introduces challenges in precisely aligning LiDAR and camera data along the vertical axis.

This phenomenon is especially pronounced in the DSEC dataset, which utilizes a 16-channel LiDAR, offering substantially lower vertical resolution than the 64-channel LiDAR employed in KITTI. As evidenced in Table 1, the MPIW for Pitch (up to 0.88°) in DSEC significantly exceeds those for Roll (0.20°) and Yaw (0.27°). This underscores the importance of comprehensive uncertainty quantification in extrinsic calibration, especially when dealing with sensors that have inherent resolution limitations.

4.4.4 Generalization and Practical Implications

While we observe slight variations in performance between the KITTI and DSEC datasets, these differences are relatively minor. Overall, our method demonstrates strong generalization capabilities.

The consistency of the uncertainty estimates demonstrated in Section 4.3.2 and the tightness of the intervals shown in Sections 4.3.1 and 4.3.3 are noteworthy. These results suggest that our method provides reliable and actionable uncertainty estimates and prediction intervals for autonomous driving systems. Its integration in calibration systems could potentially improve safety and calibration-related decision-making in real-world deployments. The method's ability to adapt to sensor limitations, such as poor vertical resolution in LiDARs, makes it particularly valuable for robust calibration in diverse autonomous driving scenarios.

5. Conclusion

We introduced a novel approach to online extrinsic calibration that incorporates rigorous uncertainty quantification through a combination of Monte Carlo Dropout (MCD) and Conformal Prediction (CP). MCD captures model uncertainty by providing a probabilistic measure of calibration parameters, while CP offers prediction intervals with formal guarantees, ensuring coverage of the true calibration parameters with a user-specified probability $1 - \alpha$.

Our method enhances existing calibration models by adding the capacity to evaluate the uncertainty of its estimates and provide statistically-guaranteed intervals. The results show that our approach maintains calibration accuracy while offering insights into the reliability of these estimates, which is vital for robust sensor fusion in autonomous systems.

Future work could explore the integration of this framework in real-world systems and optimize the use of uncertainty measures to ensure consistent calibration quality.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, Dec. 2021. 2

[2] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, Mar. 2023. 2, 4

[3] Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning (ICML)*, pages 717–730. PMLR, June 2022. 2, 3

[4] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, Feb. 2021. 3

[5] Mokhtar Bouain, Denis Berdjag, Nizar Fakhfakh, and Rabie Ben Atitallah. An extrinsic sensor calibration framework for sensor-fusion based autonomous vehicle perception. In *14th International Conference on Informatics in Control, Automation and Robotics*, volume 1, pages 505–512, 2017. 1

[6] Mathieu Cocheteux, Aaron Low, and Marius Bruehlmeier. UniCal: a single-branch transformer-based model for camera-to-LiDAR calibration and validation. *arXiv preprint arXiv:2304.09715*, 2023. 1, 2, 3, 4, 6, 7

[7] Mathieu Cocheteux, Julien Moreau, and Franck Davoine. MULi-Ev: Maintaining unperturbed LiDAR-event calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4579–4586, June 2024. 1, 2, 3, 4, 6

[8] Gianluca Detommaso, Alberto Gasparin, Michele Donini, Matthias Seeger, Andrew Gordon Wilson, and Cedric Archambeau. Fortuna: A library for uncertainty quantification in deep learning. *arXiv preprint arXiv:2302.04019*, 2023. 4

[9] Nicolas Dewolf, Bernard De Baets, and Willem Waegeman. Valid prediction intervals for regression problems. *Artificial Intelligence Review*, 56(1):577–613, 2023. 5

[10] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for LiDAR 3D vehicle detection. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 3266–3273, 2018. 2

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059. PMLR, June 2016. 1, 2, 3, 8

[12] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 8

[13] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, July 2021. 2, 4, 6

[14] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, Sept. 2013. 2, 4, 6, 7, 8

[15] Andreas Geiger, Frank Moosmann, Ömer Car, and Bernhard Schuster. Automatic camera and range sensor calibration using a single shot. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3936–3943, May 2012. 2

[16] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. 5

[17] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2888–2897, 2019. 2

[18] Xin Jing, Xiaqing Ding, Rong Xiong, Huanjun Deng, and Yue Wang. DXQ-Net: Differentiable LiDAR-camera extrinsic calibration using quality-aware flow. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6235–6241, Oct. 2022. 1, 2

[19] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 2

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1, 2

[21] Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F. Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 22(9):1341–1356, 2011. 5

[22] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 2

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017. 2

[24] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 3, 4

[25] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12008–12016, 2022. 3

[26] Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. LCCNet: LiDAR and camera self-calibration using cost volume network. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2888–2895, 2021. 1, 2

[27] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 2, 3

[28] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018. 4

[29] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018. 2

[30] Faraz M Mirzaei, Dimitrios G Kottas, and Stergios I Roumeliotis. 3D LiDAR–camera intrinsic and extrinsic calibration: Identifiability and analytical least-squares-based initialization. *The International Journal of Robotics Research*, 31(4):452–467, Apr. 2012. 2

[31] Gaurav Pandey, John R McBride, and Ryan M Eustice. Automatic extrinsic calibration of vision and LiDAR by maximizing mutual information. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2942–2949, 2015. 2

[32] Harris Papadopoulos, Kyriacos Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive conformal prediction. *European Conference on Machine Learning (ECML)*, pages 345–356, 2002. 2

[33] Jiakai Park, Qixing Zhou, and Vladlen Koltun. CalibNet: Self-supervised extrinsic calibration using 3D spatial transformer networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1110–1117, 2018. 1, 2

[34] Yaniv Romano, Evan Patterson, and Emmanuel J Candès. Conformalized quantile regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 3, 4

[35] Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. RegNet: Multimodal sensor registration using deep neural networks. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1803–1810, July 2017. 2

[36] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008. 2, 4

[37] Jieying Shi, Ziheng Zhu, Jianhua Zhang, Ruyu Liu, Zhenhua Wang, Shengyong Chen, and Honghai Liu. Calibrcnn: Calibrating camera and LiDAR by recurrent convolutional neural network and geometric constraints. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1, 2

[38] Laurens Sluijterman, Eric Cator, and Tom Heskes. How to evaluate uncertainty estimates in machine learning for regression? *Neural Networks*, 173:106203, 2024. 5

[39] Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:6216–6228, 2021. 3

[40] Paolo Toccaceli. Introduction to conformal predictors. *Pattern Recognition*, 124:108507, 2021. 2

[41] Francesco Verdoja and Ville Kyrki. Notes on the behavior of mc dropout. In *International Conference on Machine Learning (ICML) Workshops*, 2021. 8

[42] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005. 1, 2, 4

[43] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. In *Conformal and Probabilistic Prediction and Applications*, pages 82–102. PMLR, 2017. 2

[44] Shan Wu, Amnir Hadachi, Damien Vivet, and Yadu Prabhakar. NetCalib: A novel approach for LiDAR-camera autocalibration based on deep learning. In *International Conference on Pattern Recognition (ICPR)*, pages 6648–6655, Jan. 2021. 1, 2

[45] Zhao Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2301–2306, 2014. 2