Perceive With Confidence: Statistical Safety Assurances for Navigation with Learning-Based Perception

Anonymous Author(s)
Affiliation
Address
email

Abstract: Rapid advances in perception have enabled large pre-trained models to be used out of the box for transforming high-dimensional, noisy, and partial observations of the world into rich occupancy representations. However, the reliability of these models and consequently their safe integration onto robots remains unknown when deployed in environments unseen during training. In this work, we address this challenge by rigorously quantifying the uncertainty of pre-trained perception systems for object detection via a novel calibration technique based on conformal prediction. Crucially, this procedure guarantees robustness to distribution shifts in states when perceptual outputs are used in conjunction with a planner. As a result, the calibrated perception system can be used in combination with *any* safe planner to provide an end-to-end statistical assurance on safety in unseen environments. We evaluate the resulting approach, *Perceive with Confidence* (PwC), with experiments in simulation and on hardware where a quadruped robot navigates through previously unseen indoor, static environments. These experiments validate the safety assurances for obstacle avoidance provided by PwC and demonstrate up to 40% improvements in empirical safety compared to baselines.

Keywords: Uncertainty quantification, occupancy prediction, robot navigation

1 Introduction

How can we decide if the outputs of a given perception system are sufficiently reliable for safety-critical robotic tasks such as autonomous navigation? Significant strides in perception over the past few years have enabled large pre-trained models to be used out of the box [1] for tasks such as *occupancy prediction*, which serves as a fundamental building block for navigation. However, current pre-trained models are still not reliable enough for safe integration into many real-world robotic systems. Despite being trained on vast amounts of data, these systems can often fail to generalize to novel environments [2, 3, 4]. In this paper, we ask: *how can we leverage the power of large pre-trained occupancy prediction models while providing safety assurances for robot navigation?*

Consider a legged robot tasked with navigating in a cluttered environment such as a home, office, or warehouse (Figure 1). A typical navigation pipeline for such a system consists of two modules: (i) a perception module that detects obstacles, and (ii) a planner that produces collision-free trajectories assuming accurate perception. However, there are two challenges associated with obtaining reliable outputs from the perception module. First, the environments in which we deploy our robots will be *unseen* during training, and thus require *generalization* to new obstacle geometries, appearances, and other environmental factors. Second, *closed-loop deployment* of the perception system in conjunction with a planner causes a shift in the distribution of *states* (e.g., relative locations to obstacles) that are visited by the robot. Since the robot's planner influences future states, the robot may view obstacles from unfamiliar relative poses (Figure 1) and cause the perception system to fail.

In this paper, we address these challenges by performing rigorous *uncertainty quantification* for the outputs of a pre-trained perception system in order to achieve reliably safe (i.e., collision-free) navigation. We utilize techniques from *conformal prediction* [5] in order to lightly process the outputs of a pre-trained obstacle detection system in a way that provides a *formal assurance* on correctness: with a user-specified probability $1 - \epsilon$, the processed perceptual outputs will correctly

Submitted to the 8th Conference on Robot Learning (CoRL 2024). Do not distribute.

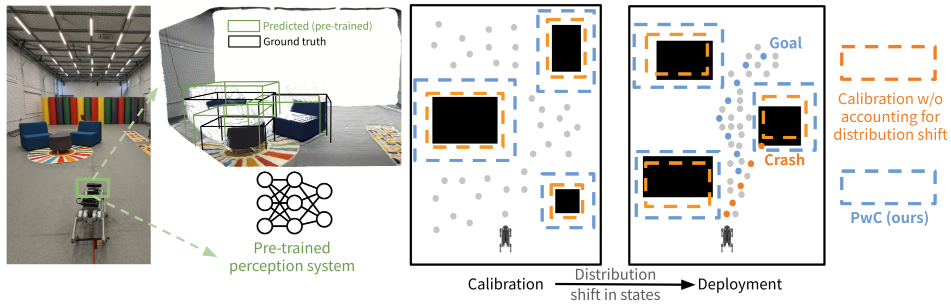


Figure 1: PwC lightly processes the outputs of a pre-trained perception system (green bounding boxes) using conformal prediction in order to ensure a bounded misdetection rate despite *any* distribution shift in states (gray dots). The calibrated perception system (blue boxes) paired with a non-deterministic filter and a safe planner provide an end-to-end statistical assurance on safety in new test environments.

detect obstacles in a *new* environment. To enable this, we assume access to a modest-sized (e.g., $|\cdot| = 400$) dataset of environments that are representative of deployment environments with ground-truth obstacle annotations, and use these for *calibrating* the outputs of the perception system. Crucially, we propose a novel calibration technique that ensures robustness of the perception system to *any closed-loop distribution shift in states*. Hence, the calibrated outputs can be used in conjunction with *any* safe planner to provide an end-to-end statistical assurance on safety in new static environments with a user-specified threshold $1 - \epsilon$. To the best of our knowledge, this is the first work to calibrate a given black-box perception system in a way that ensures robustness to closed-loop distribution shifts in order to provide end-to-end statistical assurances on safe navigation.

Our framework, *Perceive with Confidence* (PwC), is evaluated with experiments in simulation and hardware on the Unitree Go1 quadruped navigating in indoor environments with objects that are unseen during calibration (Figure 1). We validate PwC's ability to provide end-to-end statistical assurances on collision avoidance, while also providing up to 40% increase in safety with only modest reductions in task completion rates compared to baselines that use the pre-trained perception model directly, fine-tune it on the calibration dataset, or utilize conformal prediction for uncertainty quantification but do not account for closed-loop distribution shift.

2 Problem Formulation and Overview

Dynamics and environments. Suppose that the dynamics of the robot are described by $s_{t+1} = f_E(s_t, a_t)$, where $s_t \in \mathcal{S}$ is the robot's state at time-step t, $a_t \in \mathcal{A}$ is the action, and $E \in \mathcal{E}$ is the *environment* that the robot operates in during a given episode. We primarily focus on navigation with static obstacles; in this context, the environment E specifies the locations and geometries of objects. We assume that environments that the robot will be deployed in are drawn from an *unknown* distribution $\mathcal{D}_\mathcal{E}$, e.g., a distribution over possible rooms that the robot may be deployed in. We will make no assumptions on this distribution besides the ability to sample a finite dataset $D = \{E_1, \dots, E_N\}$ of i.i.d. environments from $\mathcal{D}_\mathcal{E}$.

Sensor and perception system. The robot is equipped with a sensor $\sigma : \mathcal{S} \times \mathcal{E} \to \mathcal{O}$ that provides observations $o_t = \sigma(s_t, E)$ (e.g., depth images) based on the robot's state and environment. We assume access to a pre-trained perception model $\phi : \mathcal{O} \to \mathcal{Z}$, which processes raw sensor observations into an occupancy representation of the environment. In this paper, we work with perception models for obstacle detection that output 3D bounding boxes. The representations $(z_0, \dots, z_t)$ up to the current time-step are aggregated into an overall representation $m_t \in \mathcal{M}$ (e.g., a map).

Policy. The representation m_t is used by a planning algorithm in order to produce actions. Denote the resulting end-to-end policy that utilizes a perception model ϕ by $\pi^\phi : \mathcal{O}^{t+1} \to \mathcal{Z}^{t+1} \to \mathcal{M} \to \mathcal{A}$, which maps histories of sensor observations to actions.

Safety and task performance. Let C_E^{safe} be a cost function that captures safety (e.g., obstacle avoidance). Specifically, let $\mathcal{S}_{0,E}$ denote the allowable set of initial conditions in environment E. Then, $C_E^{\text{safe}}(\pi^\phi) \in \{0, 1\}$ assigns a cost of 0 if policy π^ϕ maintains safety from any initial state

$s_0 \in \mathcal{S}_{0,E}$ when deployed over a given time horizon in environment E, and a cost of 1 otherwise. An additional cost function C_E^{task} can be used to capture task performance (e.g., time to reach a goal).

Goal: statistical safety assurance. Our goal is to provide a statistical assurance on safety for the end-to-end policy π^{ϕ}. We propose a procedure that uses a finite dataset D of environments in order to produce a *calibrated* perception system $\bar{\phi} : \mathcal{O} \xrightarrow{\phi} \mathcal{Z} \xrightarrow{\rho} \mathcal{Z}$. Our approach is modular: outputs of the calibrated perception system may be used with *any* safe planner (cf. Section 4) to ensure:

$$C_{\mathcal{D}_{\mathcal{E}}}^{\text{safe}}(\pi^{\bar{\phi}}) := \mathop{\mathbb{E}}_{E \sim \mathcal{D}_{\mathcal{E}}} \left[C_E^{\text{safe}}(\pi^{\bar{\phi}}) \right] \leq \epsilon, \tag{1}$$

for a user-specified safety tolerance ϵ, while also post-processing outputs from ϕ as lightly (i.e., non-conservatively) as possible in order to allow the robot to optimize task performance.

3 Offline: Calibrating the Perception System

In this section, we describe our approach to the uncertainty quantification of a pre-trained perception system. We focus on the challenges highlighted in Section 1: providing statistical assurances on safe generalization to novel environments and ensuring that the offline calibration procedure is robust to shifts in the distribution of states induced by the online implementation of the planner.

3.1 Misdetection Rate and Closed-Loop Distribution Shift

We focus on perception systems that output bounding boxes that predict the locations of objects in the environment. As an example, Figure 1 (left) shows one such real-world environment wherein the union A of the black boxes denotes the ground-truth locations of the chairs. Let B_s denote the union of the green bounding boxes predicted by the perception system ϕ from robot state $s \in \mathcal{S}$. Since the environment in which the robot is deployed may contain partially occluded objects that ϕ was not explicitly trained on, the perception system's outputs may be inaccurate.

Closed-loop distribution shift. In addition to this challenge of generalization, we highlight another challenge that any uncertainty quantification method for perception must tackle. Suppose we fix a policy π^{ϕ} (that uses perception system ϕ) and collect a dataset of observations in different calibration environments from the states that result from applying π^{ϕ}. We can use ground-truth bounding boxes in these environments to produce a calibrated perception system $\bar{\phi}$ with a statistical assurance on correctness for the distribution of observations induced by π^{ϕ}. However, if we now apply the policy $\pi^{\bar{\phi}}$ using the *calibrated* perception system $\bar{\phi}$, the resulting distribution of states will be *different* from the distribution that forms the calibration dataset, thus invalidating the statistical assurance. We refer to this challenge as *closed-loop distribution shift*, which is similar to challenges that arise in offline reinforcement learning [6] and imitation learning [7].

Our key idea for tackling closed-loop distribution shift is to use a *policy-independent* misdetection cost, $\bar{C}_E$, which considers worst-case errors across *all* states in an environment[1], $\bar{C}_E(\phi) := \max_{s \in \mathcal{S}} \ \mathbb{1}_{A \not\subseteq B_s}$. We will present a calibration procedure that allows us to bound this misdetection cost with high probability in a new environment, and thus guarantee the correctness of the calibrated perception system independent of the robot policy using conformal prediction (CP).

3.2 Calibration Procedure

Dataset. We assume access to a dataset of N i.i.d. environments $D = \{E_1, \dots E_N\} \sim \mathcal{D}_{\mathcal{E}}$ (cf. Section 2). In each environment, E_i, we have access to the union A_i of the ground-truth bounding boxes of all the objects in the environment and the unions $B_{s,i}$ of the predicted bounding boxes generated by the pre-trained perception system ϕ from each state $s \in \mathcal{S}$. We construct the calibration dataset either using real-world environments or create simulation environments using real-world data [8, 9, 10] to ensure that the calibration dataset is representative of deployment environments.

[1]It would be infeasible to consider *all* possible states in an environment. In practice, we use a sampling-based motion planner and consider a fixed set of samples for our calibration that could be used by any planner.

Calibration. In each calibration environment E_i, we find the inflation Δ_{q_i} of the bounding box predictions $B_{s,i}$ so as to ensure that all the ground-truth boxes are fully enclosed by the inflated boxes, i.e, $A \subseteq B_{s,i} + \Delta_{q_i}, \forall s \in \mathcal{S}$. Here, $B_{s,i} + \Delta_{q_i}$ refers to the inflation of each bounding box in the union $B_{s,i}$ by $2q_i$ along each dimension. We define the *non-conformity score* for environment E_i to be the minimum required inflation in that environment (background on CP in Appendix A):

$$U_i \;=\; \min_{q_i} \quad q_i \quad \text{s.t} \quad A_i \subseteq B_{s,i} + \Delta_{q_i}, \forall s \in \mathcal{S}. \tag{2}$$

Observe that $U_i \leq 0 \implies A_i \subseteq B_{s,i}, \forall s \in \mathcal{S}$ and a growing U_i signals a worse performance of the pre-trained perception system. We compute the nonconformity scores for all our i.i.d. sampled environments $\{E_1, \ldots, E_N\}$. Hence, the following guarantee holds for the non-conformity score, U_{test}, in a new environment, E_{test}, with probability $1-\delta$ over the sampling of the calibration dataset,

$$\mathbb{P}[U_{\text{test}} \leq \hat{q}_{1-\epsilon} | U_1, \ldots, U_N] \geq \text{Beta}^{-1}_{N+1-v,v}(\delta), \quad v := \lfloor (N+1)\hat{\epsilon} \rfloor, \tag{3}$$

where, $\text{Beta}^{-1}_{N+1-v,v}(\delta)$ is the $\delta-$quantile of the Beta distribution, and we use a modified $\hat{\epsilon}$ for calibration to achieve the desired $1-\epsilon$ coverage, i.e., we compute the associated quantile $\hat{q}_{1-\epsilon}$ as the $\lceil (N+1)(1-\hat{\epsilon}) \rceil^{\text{th}}$ largest value of all the non-conformity scores collected during calibration.[2]

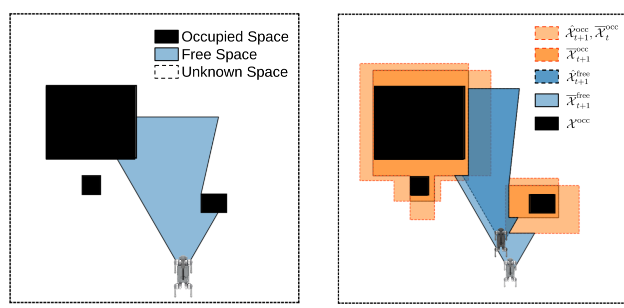
Proposition 1 *Consider the calibrated perception system $\bar{\phi}$ that modifies every bounding box output of the perception system ϕ by scaling the predicted bounding boxes as $\bar{B} = B + \Delta_{\hat{q}_{1-\epsilon}}$. With probability $1-\delta$ over the sampling of the dataset used for calibration, the calibrated perception system, $\bar{\phi}$, is guaranteed to have an ϵ-bounded misdetection rate on* new *test environments:*

$$\mathop{\mathbb{E}}_{E_{test} \sim \mathcal{D}_{\mathcal{E}}} \left[\bar{C}_{E_{test}}(\bar{\phi}) | U_1, \ldots, U_N \right] \leq \epsilon. \tag{4}$$

The above proposition (proof in Appendix B) gives us a formal assurance on the correctness of the perception system *independent of the robot's policy*. As we describe below, the calibrated perception can thus be combined with *any* safe planner to bound the collision rate to ϵ.

4 Online: Perception and Planning

We now focus on the online implementation of the method described in Section 3 to reduce conservatism when used in conjunction with a safe planner. In general, a safe planner takes into account the dynamics of the robot and produces plans in the state space $\mathcal{S}$. We call $\mathcal{X}$ the configuration space of the robot (e.g., x-y location for a point). For any given environment E, we partition $\mathcal{X}$ into three sub-spaces (Figure 2a): the known free space $\mathcal{X}^{\text{free}}$, known occupied space $\mathcal{X}^{\text{occ}}$, and unknown space $\mathcal{X}^{\text{unknown}}$.



(a) A line-of-sight depth sensor along with a bounding box estimator partition the configuration space into three.

(b) The non-deterministic filter takes intersection over the occupied space and takes union over the free space.

Figure 2

Non-deterministic filter. We utilize the assurance obtained from Section 3 to implement a *non-deterministic filter* [11, Ch. 11.2.2] which shrinks the occupied space and grows the known free space over time. Suppose the robot's perceived partition (i.e., map) of the configuration space $\mathcal{X}$ at time t is $m_t := (\overline{\mathcal{X}}_t^{\text{free}}, \overline{\mathcal{X}}_t^{\text{occ}}, \overline{\mathcal{X}}_t^{\text{unknown}})$. At a new time step $t+1$, the robot returns a new set of bounding box predictions, $\hat{\mathcal{X}}_{t+1}^{\text{occ}}$. The filter then updates the new perceived occupied space with $\overline{\mathcal{X}}_{t+1}^{\text{occ}} = \overline{\mathcal{X}}_t^{\text{occ}} \cap \hat{\mathcal{X}}_{t+1}^{\text{occ}}$. We then compute the new estimate of free space $\hat{\mathcal{X}}_{t+1}^{\text{free}}$ based on $\overline{\mathcal{X}}_{t+1}^{\text{occ}}$,

[2] In practice, we choose the calibration threshold $\hat{\epsilon}$ such that the dataset conditional guarantee (3) achieves the desired $(1-\epsilon)-$coverage with probability $1-\delta = 0.99$ over the sampling of the calibration dataset.

considering occlusions and limited field of view. Figure 2b shows the non-deterministic filter applied for one instance. The new perceived free space is updated with $\overline{\mathcal{X}}^{\text{free}}_{t+1} = \overline{\mathcal{X}}^{\text{free}}_{t} \cup \hat{\mathcal{X}}^{\text{free}}_{t+1}$.

The non-deterministic filter pairs effectively with our method in Section 3 for two key reasons: 1) it mitigates the conservatism of our bounding box expansion by intersecting $\overline{\mathcal{X}}^{\text{occ}}_{t}$, rapidly reducing its size even if the initial prediction with CP bounds appears generous; and 2) Prop. 1 ensures that with high probability in a new test environment, $\overline{\mathcal{X}}^{\text{free}}_{t}$ never intersects the true occupied space $\mathcal{X}^{\text{occ}}$. We demonstrate the rapid expansion of known free space in Figure 3 for our simulated setup (Sec. 5).

Safe planning. With our formal assurance on the estimated free space $\overline{\mathcal{X}}^{\text{free}}_{t}$, we can utilize *any* safe planner [12, 13, 14] to ensure end-to-end safety, as long as the planner includes a safety filter that takes into account the robot's dynamics in order to reject potentially unsafe actions with the assumption of known state and static (but unknown) environment [15, Corollary 1.4]. For our simulation and hardware experiments, we use the safe planner proposed in [16], which enforces an inevitable collision set (ICS) constraint [17]. We describe implementation details in Appendix D.

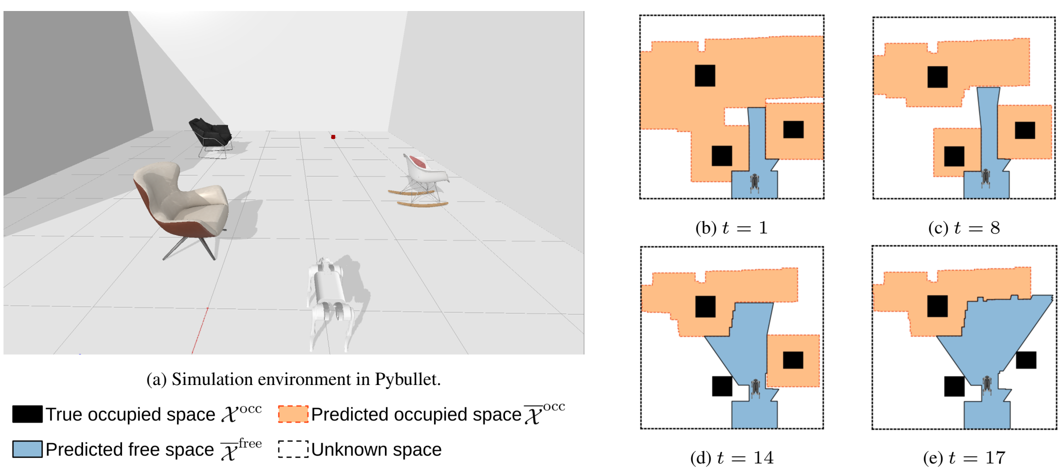
Proposition 2 *For any user-specified safety tolerance ϵ, the* calibrated *perception system $\bar{\phi}$ in Proposition 1 combined with any safe planner that chooses actions based on the outputs of the non-deterministic filter ensures the end-to-end safety for the overall policy $\pi^{\bar{\phi}}$:*

$$C^{safe}_{\mathcal{D}_\mathcal{E}}(\pi^{\bar{\phi}}) := \mathop{\mathbb{E}}_{E\sim\mathcal{D}_\mathcal{E}} \left[C^{safe}_{E}(\pi^{\bar{\phi}}) \right] \leq \epsilon, \tag{5}$$

where $C^{safe}_{E}(\pi^{\bar{\phi}})$ is the cost function for safety from Section 2.

This result (proved in Appendix E) is a direct consequence of the formal assurance on the calibrated perception system that ensures correctness from *any* state in a new test environment (sampled i.i.d. from the same distribution as the calibration environments) with probability $1-\epsilon$ *over environments*.

5 Simulated Experiments: Vision-Based Navigation



(a) Simulation environment in Pybullet. (b) $t = 1$ (c) $t = 8$ (d) $t = 14$ (e) $t = 17$

Figure 3: Simulation and non-deterministic filter updates. **(a)** An example environment in simulation. **(b - d)** The robot begins with large occupied space predictions due to the inflation obtained through offline calibration (Section 3). After a few updates, the predicted occupied space $\overline{\mathcal{X}}^{\text{occ}}$ shrinks significantly.

We evaluate our approach for vision-based navigation in the PyBullet simulator [18] using a diverse set of chairs from the 3D-Front dataset [10]. We use the 3DETR end-to-end transformer model [19] as our pre-trained perception system.

Baselines. We compare our approach (*Perceive with Confidence* — PwC) to three baselines to illustrate its effectiveness in achieving a user-specified safety rate. First, we consider the most common approach of directly using the outputs of the perception system [19] in our planning pipeline. We call this baseline **3DETR**. Next, we consider the common practice of fine-tuning the outputs of the

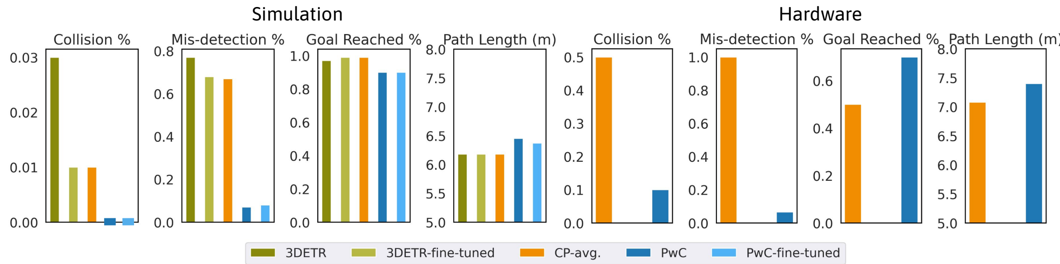


Figure 4: **(Left)** Results for the simulated experiments described in Section 5. Simulations are across 100 new environments with 1 - 5 chairs. **(Right)** Results for the hardware trials described in Section 6. Experiments are across 30 different chair configurations with 4-8 chairs present in each configuration. Here the path length is averaged only for successful trials for both PwC and CP-avg. due to the varying goal locations.

perception system using a small dataset of task-representative environments D_{tune} (cf. Section F.1). We call this perception system **3DETR-fine-tuned**. Lastly, we perform calibration using conformal prediction; however, instead of accounting for the closed-loop distribution shift, we bound the misdetection rate averaged across environments *and* states (similar to [20], which does not utilize conformal prediction, but quantifies expected errors in a perception system for a pre-defined distribution of states). We refer to this baseline as **CP- avg**. We consider two variations of our approach for comparison to the above baselines. First, we refine 3DETR outputs using our calibration procedure described in Section 3. We call this approach **PwC**. Second, the 3DETR outputs are fine-tuned and calibrated using split conformal prediction as described in Appendix F.1; we call this approach **PwC-fine-tuned**. Details regarding calibration and the planner setup are provided in Appendix G.

Results: Misdetection Rate. We first compare our method, PwC, to the baseline CP-avg that is also calibrated using conformal prediction but without accounting for the closed-loop distribution shift. We compare the misdetection rate, i.e., whether obstacles in the scene are classified as free space at any point during a trial. We vary the allowable misdetection bound ϵ for each method, and compute the rate of misdetections in 100 test environments. As seen in Figure 5, our method is guarantees a rate of misdetection lower than the threshold ϵ while CP-avg violates this threshold for every ϵ considered.

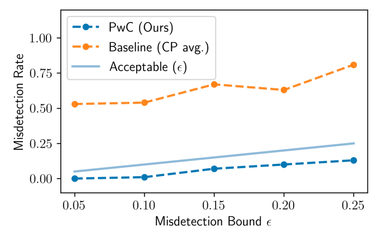


Figure 5: As we relax the confidence threshold by increasing ϵ, the misdetection rate increases but remains bounded for PwC. The baseline method has a misdetection rate much higher than acceptable.

Results: Collision Rate. We compare PwC to the baselines in 100 new environments drawn from the same distribution as calibration environments. Figure 3 illustrates one such test environment and the evolution of the free space in this environment using PwC. Figure 3 shows that though the initial calibrated perception system outputs are inflated, the non-deterministic filter is able to expand the predicted free space in a few time steps and ensure that the robot can navigate without unnecessary conservatism. The results are summarized in Figure 4 and the metrics for success and failure are described in Appendix G. We observe that our proposed approaches, PwC and PwC-fine-tuned, have no collisions in any environments. While the robot reaches the goal in a slightly lower percentage of environments compared to baselines, we emphasize that ours is the only approach that is able to ensure a low, statistically guaranteed misdetection rate across test environments.

To further illustrate the effect of misdetections on safety, we consider a different distribution of environments wherein we randomly place a *single* chair in the straight line path between the initial position of the robot and the goal. For a safety threshold $1 - \epsilon = 0.85$, we compare PwC, CP-avg, and 3DETR. The results are provided in Figure 6 for 100 new test environments, wherein the goal is reached if the robot navigates to within 2 m of the goal. In these environments, the desired safety rate is not met by the baselines while our approach is still statistically guaranteed to be safe.

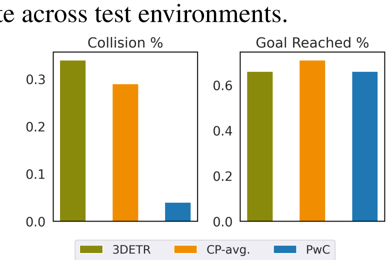
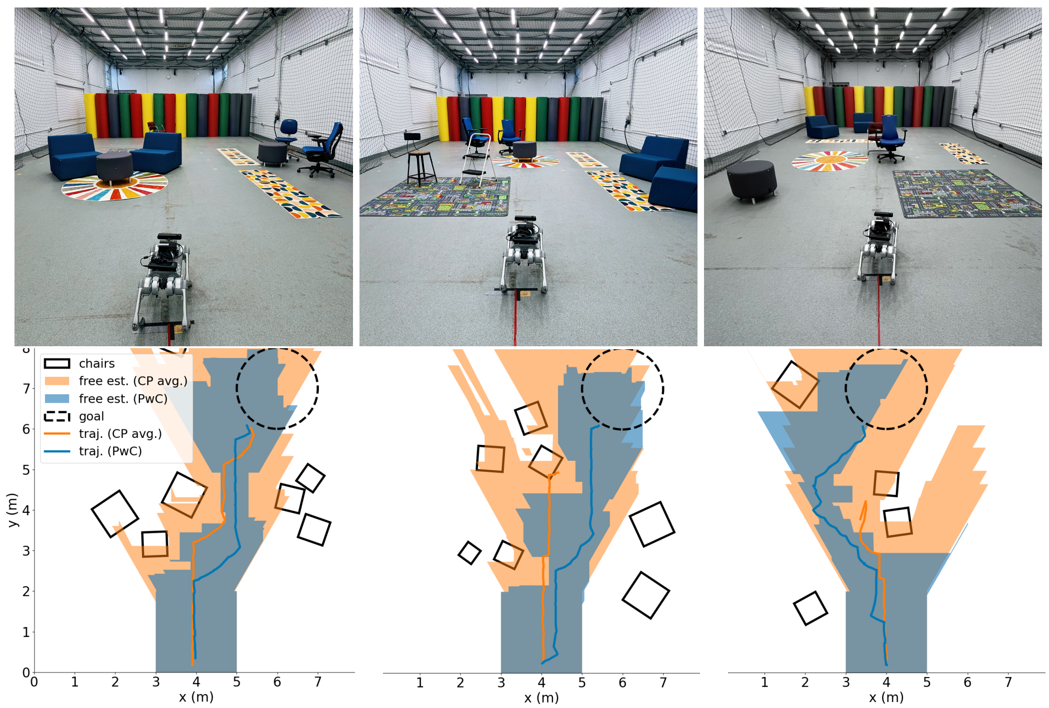


Figure 6: A comparison between the collision rates of different perception systems that use the same safe planner.

We provide additional simulation results that illustrate the effects of 1) closed-loop distribution shifts on safety in Appendix G.2 wherein PwC is robust to an increase in the level of closed-loop distribution shift while the baseline, CP-avg., is not which leads to higher collision rates for CP-avg. and 2) the tradeoff in different partition sizes for fine-tuning using split-CP in Appendix F.1.2.

6 Hardware Validation: Vision-Based Quadruped Navigation

We now validate the end-to-end statistical safety assurance of our approach on a quadrupedal hardware platform. As in our simulation setup in Section 5, the robot is tasked with navigating to a goal location while avoiding different chairs placed in varying configurations across a 8m x 8m room. We utilize the perception system calibrated in simulation with a guaranteed safety rate of $1 - \epsilon = 0.85$, and compare our PwC method against CP-avg. (defined in Section 5) across 30 different physical environments (60 trials total). See Appendix H for more details about the hardware setup.



(a) Environment 1 (b) Environment 4 (c) Environment 14

Figure 7: **(Top)** The physical layouts of the example hardware trails. **(Bottom)** A bird's-eye view of the estimated free spaces (shaded regions), and the trajectories performed by the robot (solid lines) with our method (blue) and the baseline (orange). In all three trials, PwC is able to successfully navigate to the goal through narrow paths (in Environment 1) and occluded areas/goal (in Environment 3). Baseline approach, CP-avg., misdetects free space in all environments leading to collisions in Environments 2 and 3.

Results. For PwC, we used the $\hat{q}_{0.85} = 0.73$m threshold found in simulation to inflate the predicted bounding boxes returned from 3DETR in order to achieve 85% confidence that our robot will remain safe in new environments. We summarize key statistics of PwC compared against CP-avg. ($\hat{q}_{0.85} = 0.02$) across 30 different environments in Figure 4 (right). Importantly, our trials demonstrate that our confidence bound holds on hardware in real environments and without being too conservative. PwC was safe through 90% of the trials and also had comparable path length to the baseline. Meanwhile, the baseline struggled in the real environments by having misdetections in each trial and colliding with a chair in half of the trials. See Figure 7 for trajectories and free space estimations through several environments with narrow spaces, occluded chairs, and occluded goals. The supplementary video contains full example trials.

PwC's low misdetection rate and higher success rate in these trials emphasize the efficacy of the bounding box inflation provided by CP paired with the non-deterministic filter. This pairing, in a principled way, inflates the (potentially poor) bounding box detections to properly capture obstacles but quickly shrinks the occupied space with the filter such that the robot can still navigate effectively.

7 Related Work

Safe planning. Collision avoidance is a crucial goal in autonomous navigation. Safe planning methods typically rely on the assumption that the robot has perfect knowledge of its state and environment [15]. Recent approaches have allowed for occlusion [16, 21, 22, 23] or accounted for losing sight of a previously tracked object [24], but still require either perfect detection of seen objects or bounded sensor noise. Such assumptions are impractical for learning-based perception modules that can fail catastrophically in new environments.

Formal assurances for perception-based control. Proposed methods include control barrier functions (CBFs) [25, 26], verification methods on neural networks (NNs) [27, 28], and other learning-based methods [28, 29, 30, 20, 31, 32, 33, 34]. However, these works either do not guarantee generalization to novel environments [27, 28], or ignore closed-loop distribution shifts [31, 20], or require end-to-end training and a good prior [32, 33, 34], or demand usage/design of specific controllers [25, 26, 29]. Some also make strong assumptions on the perception system [35, 36] that are unrealistic for deployment. In contrast, our method does not need any of the above, and is lightweight and modular, allowing for the use of any downstream safe planners to ensure end-to-end safety.

Conformal prediction. Conformal prediction (CP) [5, 37, 38] is an uncertainty quantification framework particularly suitable for robotics applications [39, 40, 41, 42] where learned modules are deployed in environments drawn form unknown distributions. In this work, we focus on providing uncertainty quantification for the perception system, which usually involves high-dimensional inputs and closed-loop distribution shifts. Prior works [41, 20, 43, 44] either provide guarantees for a single environment, assume known environments, or do not account for closed-loop distribution shifts. To the best of our knowledge, this is the first work to obtain end-to-end safety assurances for the perception and planning system in new environments while being robust to closed-loop distribution shifts and amenable to changes in the planner parameters.

8 Discussion and Conclusions

We presented a modular framework for rigorously quantifying the uncertainty of a pre-trained perception model in order to provide an end-to-end statistical safety assurance for perception-based navigation tasks. Notably, our statistical assurance holds for generalization to new environmental factors (e.g, new obstacle geometries and configurations) and allows for the distribution shift of states that may occur during closed-loop deployment of the perception system with the planner. Our simulation and hardware experiments validated the theoretical safety assurances provided by PwC, while demonstrating significant empirical improvements in safety compared to baseline approaches that do not consider closed-loop distribution shift.

Limitations and future work. One limitation of our work is the assumption of static obstacles. As a future direction, we are interested in quantifying uncertainty in both the state of agents moving in the environment and predictions of their *semantic labels* (e.g., "pedestrian" vs. "bicyclist"), and utilizing game-theoretic planning techniques that account for the uncertainty in the agents' current state and future motion. Additionally, the inflation of bounding boxes we acquire from CP introduces some conservatism. We outline an extension to our approach in Appendix F to address this challenge by utilizing more general occupancy representations beyond bounding boxes, e.g., scene completion networks [45], which produce voxel-wise occupancy confidences. Constructing different non-conformity score functions that incorporate confidences from a pre-trained model could also potentially reduce conservatism. Lastly, we are interested in uncertainty quantification for perception models that support tasks beyond point-to-point navigation, e.g., calibrating the outputs of multi-modal foundation models for language-instructed navigation where we ensure accurate detection. We expect that rigorous uncertainty quantification is a necessary step towards fully leveraging the power of large foundation models [1] while safely integrating them into future robotic systems.

References

[1] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, B. Ichter, D. Driess, J. Wu, C. Lu, and M. Schwager. Foundation models in robotics: Applications, challenges, and the future. *arXiv preprint arXiv:2312.07843*, 2023.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.

[3] A. Wang, M. Islam, M. Xu, Y. Zhang, and H. Ren. SAM meets robotic surgery: An empirical study in robustness perspective. *arXiv preprint arXiv:2304.14674*, 2023.

[4] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, et al. The limits and potentials of deep learning for robotics. *The International journal of robotics research*, 37(4-5):405–420, 2018.

[5] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[6] S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[7] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 15, pages 627–635. PMLR, 2011.

[8] M. A. Uy, Q.-H. Pham, B.-S. Hua, D. T. Nguyen, and S.-K. Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019.

[9] B. Calli, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36:027836491770071, 04 2017. doi:10.1177/0278364917700714.

[10] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, et al. 3D-Front: 3D furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

[11] S. M. LaValle. Planning algorithms. *Cambridge University Press*, 2:3671–3678, 2006.

[12] T. Schouwenaars, É. Féron, and J. How. Safe receding horizon path planning for autonomous vehicles. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, volume 40, pages 295–304. The University; 1998, 2002.

[13] S. Bouraine, T. Fraichard, and O. Azouaoui. Real-time safe path planning for robot navigation in unknown dynamic environments. In *Conference on Computing Systems and Applications*, 2016.

[14] È. Pairet, J. D. Hernández, M. Carreras, Y. Petillot, and M. Lahijanian. Online mapping and motion planning under uncertainty for safe navigation in unknown environments. *IEEE Transactions on Automation Science and Engineering*, 19(4):3356–3378, 2021.

[15] K.-C. Hsu, H. Hu, and J. F. Fisac. The safety filter: A unified view of safety-critical control in autonomous systems. *Annual Review of Control, Robotics, and Autonomous Systems (ARCRAS)*, 2023.

[16] L. Janson, T. Hu, and M. Pavone. Safe motion planning in unknown environments: Optimality benchmarks and tractable policies. *arXiv preprint arXiv:1804.05804*, 2018.

[17] T. Fraichard and H. Asama. Inevitable collision states. A step towards safer robots? In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 388–393, 2003.

[18] E. Coumans and Y. Bai. Pybullet, a Python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2022.

[19] I. Misra, R. Girdhar, and A. Joulin. An end-to-end transformer model for 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021.

[20] D. Sun, B. C. Yang, and S. Mitra. Learning-based perception contracts and applications. *arXiv preprint arXiv:2309.13515*, 2023.

[21] Z. Zhang and J. F. Fisac. Safe occlusion-aware autonomous driving via game-theoretic active perception. *arXiv preprint arXiv:2105.08169*, 2021.

[22] C. Packer, N. Rhinehart, R. T. McAllister, M. A. Wright, X. Wang, J. He, S. Levine, and J. E. Gonzalez. Is anyone there? Learning a planner contingent on perceptual uncertainty. In *Proceedings of the Conference on Robot Learning*, pages 1607–1617. PMLR, 2023.

[23] M. Koschi and M. Althoff. Set-based prediction of traffic participants considering occlusions and traffic rules. *IEEE Transactions on Intelligent Vehicles*, 6(2):249–265, 2020.

[24] F. Laine, C.-Y. Chiu, and C. Tomlin. Eyes-closed safety kernels: Safety for autonomous systems under loss of observability. *arXiv preprint arXiv:2005.07144*, 2020.

[25] S. Dean, A. Taylor, R. Cosner, B. Recht, and A. Ames. Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. In *Proceedings of the Learning for Dynamics and Control (L4DC) Conference*, pages 654–670. PMLR, 2021.

[26] C. Dawson, B. Lowenkamp, D. Goff, and C. Fan. Learning safe, generalizable perception-based hybrid control with certificates. *arXiv preprint arXiv:2201.00932*, 2022.

[27] C. Hsieh, Y. Li, D. Sun, K. Joshi, S. Misailovic, and S. Mitra. Verifying controllers with vision-based perception using safe approximate abstractions. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(11):4205–4216, 2022.

[28] S. M. Katz, A. L. Corso, C. A. Strong, and M. J. Kochenderfer. Verification of image-based neural network controllers using generative models. *Journal of Aerospace Information Systems*, 19(9):574–584, 2022.

[29] S. Ghosh, Y. V. Pant, H. Ravanbakhsh, and S. A. Seshia. Counterexample-guided synthesis of perception models and control. In *Proceedings of the IEEE American Control Conference*, pages 3447–3454. IEEE, 2021.

[30] S. Dean and B. Recht. Certainty equivalent perception-based control. In *Proceedings of the Learning for Dynamics and Control (L4DC) Conference*, pages 399–411. PMLR, 2021.

[31] Y. Liu, N. Mishra, M. Sieb, Y. Shentu, P. Abbeel, and X. Chen. Autoregressive uncertainty modeling for 3d bounding box prediction. In *European Conference on Computer Vision*, pages 673–694. Springer, 2022.

[32] A. Majumdar, A. Farid, and A. Sonar. PAC-Bayes control: Learning policies that provably generalize to novel environments. *The International Journal of Robotics Research*, 40(2-3): 574–593, 2021.

[33] A. Farid, S. Veer, and A. Majumdar. Task-driven out-of-distribution detection with statistical guarantees for robot learning. In *Conference on Robot Learning*, pages 970–980. PMLR, 2022.

[34] A. Farid, D. Snyder, A. Ren, and A. Majumdar. Failure prediction with statistical guarantees for vision-based robot control. In *Proceedings of Robotics: Science and Systems*, 2022.

[35] S. Dean, N. Matni, B. Recht, and V. Ye. Robust guarantees for perception-based control. In *Proceedings of the Learning for Dynamics and Control (L4DC) Conference*, pages 350–360. PMLR, 2020.

[36] G. Chou, N. Ozay, and D. Berenson. Safe output feedback motion planning from images via learned perception modules and contraction theory. In *Algorithmic Foundations of Robotics XV*, pages 349–367. Springer International Publishing, 2023.

[37] V. Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.

[38] A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2022.

[39] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

[40] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 2023.

[41] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Proceedings of the Learning for Dynamics and Control (L4DC) Conference*, pages 300–314. PMLR, 2023.

[42] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, and M. Pavone. Sample-efficient safety assurances using conformal prediction. In *International Workshop on the Algorithmic Foundations of Robotics*, pages 149–169. Springer, 2022.

[43] S. Yang, G. J. Pappas, R. Mangharam, and L. Lindemann. Safe perception-based control under stochastic sensor uncertainty using conformal prediction. *arXiv preprint arXiv:2304.00194*, 2023.

[44] S. Park, O. Bastani, N. Matni, and I. Lee. PAC confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*, 2019.

[45] E. Chatzipantazis, S. Pertigkiozoglou, E. Dobriban, and K. Daniilidis. SE(3)-equivariant attention networks for shape reconstruction in function space. *arXiv preprint arXiv:2204.02394*, 2022.

[46] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal Risk Control, Apr. 2023. URL http://arxiv.org/abs/2208.02814. arXiv:2208.02814 [cs, math, stat].

[47] L. Janson, E. Schmerling, A. Clark, and M. Pavone. Fast marching tree: A fast marching sampling-based method for optimal motion planning in many dimensions. *The International Journal of Robotics Research*, 34(7):883–921, 2015. ISSN 0278-3649.

[48] E. Schmerling, L. Janson, and M. Pavone. Optimal sampling-based motion planning under differential constraints: The driftless case. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2368–2375. IEEE, 2015.

[49] E. Schmerling, L. Janson, and M. Pavone. Optimal sampling-based motion planning under differential constraints: The drift case with linear affine dynamics. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2574–2581. IEEE, 2015.

[50] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019.

[51] K. Neklyudov, D. Molchanov, A. Ashukha, and D. Vetrov. Variance networks: When expectation does not meet your expectations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1GAUs0cKQ.

[52] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[53] T. M. Inc. MATLAB version: 9.13.0 (r2022b), 2022. URL https://www.mathworks.com.