Real-Time Hybrid Detection of Adversarial Attacks on Learning-Based Locomotion Controllers for Quadrupedal Robots

Upinder Kaur * kauru@purdue.edu

Ali A. Noghabi † a.noghabi2002@gmail.com

Abstract

Quadrupedal robots with learning-based locomotion controllers are highly adaptable but vulnerable to adversarial attacks that induce small perturbations in sensor data or control signals, potentially leading to catastrophic failures. This paper introduces a hybrid detection framework, combining a Kalman Filter and a Physics-Informed Neural Network (PINN), to detect adversarial attacks in real-time. Using a comprehensive dataset collected under various conditions, including normal and adversarial scenarios, we demonstrate significant improvements in detection accuracy and robustness.

1 Introduction

Quadrupedal robots are increasingly used in dynamic applications requiring adaptability and robust movement across diverse terrains. Learning-based controllers empower these robots with such capabilities; however, they are susceptible to adversarial attacks—small perturbations that exploit vulnerabilities in decision-making processes. Previous studies explored these vulnerabilities, but robust real-time detection mechanisms remain limited. We present a novel hybrid detection framework that combines a Kalman Filter with a Physics-Informed Neural Network (PINN) to enhance resilience against adversarial attacks. The source code for this work is available on GitHub.

2 Methodology

2.1 Dataset Collection

We developed a diverse dataset of quadrupedal robot trajectories using simulators like Isaac Gym, PyBullet, and MuJoCo. The dataset includes normal conditions as well as controlled perturbations designed to simulate realistic disturbances.

2.1.1 Normal and Adversarial Conditions

Normal Conditions involve standard terrain and sensor accuracy, providing a baseline for the robot's stable behavior.

Adversarial Conditions simulate different disturbances, including noise injection to mimic sensor and actuator noise, external perturbations that represent impacts or environmental interferences, varied terrain textures with changes in ground textures and slopes, and intentional adversarial attacks targeting control signals.

2.2 Adversarial Perturbation Agent

Our reinforcement learning (RL) agent is tailored to create a range of adversarial scenarios by varying perturbations across different time frames—short-term and long-term—to effectively test the robot's resilience. This design challenges the agent to explore a range of destabilizing conditions, ensuring that it generates perturbations with diverse impacts on the robot's trajectory and stability. The reward function R_t at time t is:

$$\begin{aligned} R_t = w_{\text{short}} & \sum_{k=t}^{t+N_{\text{short}}} \left(w_{\text{pos\_short}}\|\Delta\mathbf{p}_k\| + w_{\text{orient\_short}}\|\Delta\theta_k\|\right) \\ & + w_{\text{long}} \sum_{k=t}^{t+N_{\text{long}}} \left(w_{\text{pos\_long}}\|\Delta\mathbf{p}_k\| + w_{\text{orient\_long}}\|\Delta\theta_k\|\right) \\ & + w_{\text{torque}} \sum_j \text{ReLU}\left(\frac{|\tau_j|}{\tau_{\text{lim},j}} - 1\right) - w_{\text{reg}}\|\mathbf{a}_t\| + w_{\text{fail}}\mathbb{I}_{\text{fail}}. \end{aligned} \tag{1}$$

In this formulation:

w_{short} and w_{long}: Weights for short- and long-term planning horizons.

$\Delta\mathbf{p}_k$ and $\Delta\theta_k$: Position and orientation deviations at step k, scaled by corresponding weights

$w_{\text{pos\_short}}, w_{\text{orient\_short}}, w_{\text{pos\_long}}, w_{\text{orient\_long}}$.

Torque term: Includes a ReLU penalty for exceeding joint torque limits, encouraging realistic perturbations.

Failure term $w_{\text{fail}}\mathbb{I}_{\text{fail}}$: Rewards the agent for achieving destabilization, reinforcing the failure objective.

Our **Adversarial Perturbation Agent** is inspired by the work of Dobrovolskiy et al.(1)

2.3 Hybrid Detection Framework

The proposed hybrid detection framework leverages both model-based and data-driven approaches to achieve real-time detection of adversarial attacks on quadrupedal robots. The framework consists of two main components: a Kalman Filter (KF) for state estimation and a Physics-Informed Neural Network (PINN) for enforcing physically plausible predictions.

2.3.1 Kalman Filter (KF) for State Estimation

The Kalman Filter (KF) estimates the robot's state (e.g., position, velocity, and orientation) by smoothing noisy sensor data and providing an optimal estimate of the robot's true state. The Kalman Filter operates continuously during locomotion, enabling real-time monitoring across tasks and terrains.

2.3.2 Physics-Informed Neural Network (PINN)

The Physics-Informed Neural Network (PINN) incorporates physical laws and constraints into its learning process to ensure that the robot's predicted behaviors remain physically plausible. the PINN Embeds equations of motion and physical constraints (e.g., joint torque limits, friction coefficients) into the learning process. This ensures that predictions adhere to real-world physics, reducing false positives and improving detection accuracy.

2.3.3 Real-Time Detection Process

The hybrid framework continuously monitors the robot's behavior to detect adversarial attacks using the following steps:

1. **State Prediction:** The Kalman Filter predicts the robot's state based on the current model, while the PINN predicts the next state based on physical laws.

2. **Deviation Analysis:** Compares predicted states (from KF and PINN) with actual sensor data. Significant discrepancies signal potential anomalies.

3. **Anomaly Classification:** Contextual data (terrain type, speed, trajectory) is used to classify anomalies as adversarial attacks or benign deviations.

3 Experimental Setup

To assess the performance of our hybrid detection framework, we tested it across various terrains and locomotion tasks. The tasks include walking in a straight line, random walking, and goal-reaching, under both normal and adversarial conditions. We used a comprehensive dataset collected through simulators like Isaac Gym, PyBullet, and MuJoCo.

3.1 UKF Prediction Visualization

To visualize the effectiveness of the Unscented Kalman Filter (UKF) in state estimation, we present a plot showing the predicted vs. actual states of the robot during a typical locomotion task.

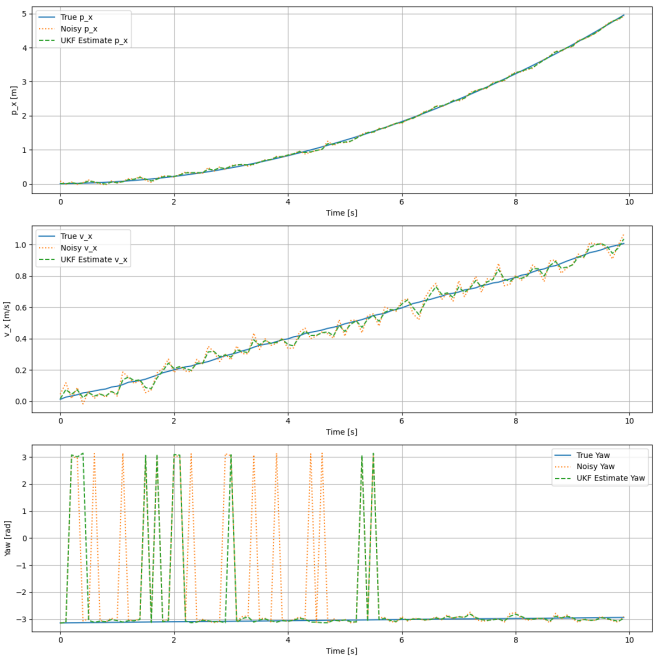

Figure 1: Prediction of the Unscented Kalman Filter (UKF) in real-time during robot locomotion.

4 Results

The results show that the framework significantly enhances detection accuracy and robustness against adversarial attacks, with minimal latency, making it suitable for real-time applications.

References

[1] F. Shi, C. Zhang, T. Miki, J. Lee, M. Hutter, and S. Coros, "Rethinking robustness assessment: Adversarial attacks on learning-based quadrupedal locomotion controllers," *arXiv preprint arXiv:2405.12424v2.*