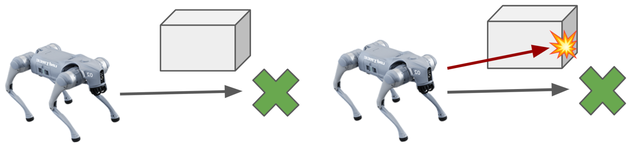
Sensor Adversarial Robustness Training for Robot Dogs under Perception Attacks

Noah Manuel, Josiah Norton, Yufeng Yang, Dacheng Shen and Mengyu Liu

MOTIVATION

- Quadrupedal robots increasingly rely on external sensors for navigation and obstacle avoidance.
- In reach-avoid tasks, the robot dog must use sensor observations to infer free space, avoid obstacles, and reach the target.
- However, small adversarial perturbations on sensor observations may cause the policy to misperceive obstacles and generate unsafe commands.
- This raises a key question: **How to train a more robust policy that enables a robotic dog to safely complete its tasks, even when its sensors are under attack?**
- To answer this question, we evaluate FGSM-based perception attacks on Sensors' observations and investigate whether adversarial robustness training can produce a policy that still completes reach-avoid tasks when its sensor inputs are attacked.



TASK AND POLICY SETUP

- We evaluate the policy in a Sensor-based reach-avoid task, where the robot must reach a target while avoiding obstacles.

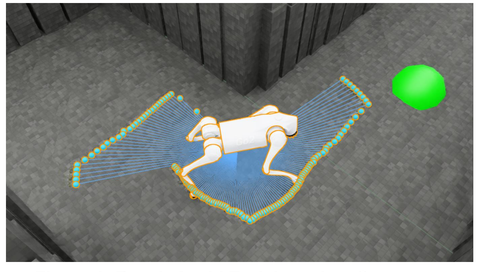


Figure 1: Reach-Avoid Task based on Sensors

- The policy takes sensor data, proprioception, and goal information as input, and outputs high-level velocity commands: *Vx, Vy*, and *Wz*.
- Attacking sensor observations can mislead the policy and cause unsafe actions.

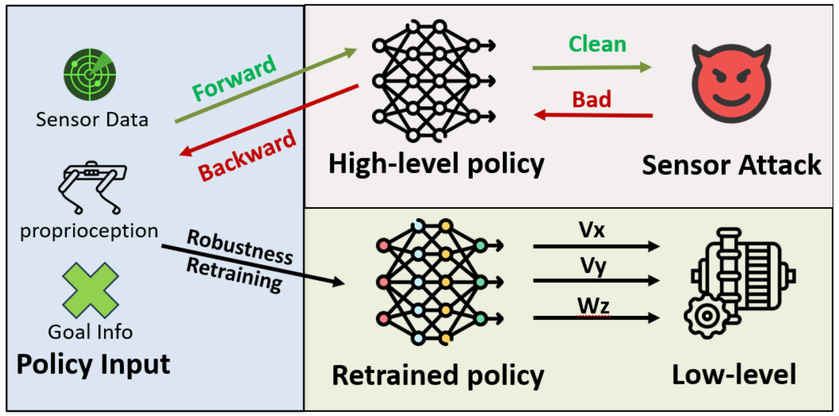


Figure 2: Overview of Sensor Attack and Adversarial Retraining Framework for Robot dogs. The high-level policy receives sensor data, proprioception, and goal information, and outputs velocity commands to the low-level controller. FGSM perturbs the sensor observation before policy inference, while mixed adversarial retraining exposes the policy to both clean and attacked observations.

ATTACK DESIGN: FROM RANDOM NOISE TO FGSM

- **Random Noise Sanity Check:** We first inject random noise into Sensor observations. The policy still achieves 100% success rate under this random perturbation.
  This suggests that the clean policy is naturally robust to unstructured noise, and simple noise is insufficient to expose its vulnerability.
- **FGSM Sensor Attack**: Fast Gradient Sign Method, FGSM, is a one-step gradient-based adversarial attack proposed by Goodfellow et al. [1]. It generates adversarial examples by adding a small perturbation in the direction of the input gradient, which can efficiently push the model output toward an undesired behavior.
  In this work, we adapt FGSM to attack the normalized Sensor observation before policy inference. Let x denote the normalized Sensor observation and π(x) denote the high-level policy output. The adversarial observation is generated as:

$$\delta = -\epsilon \cdot \mathrm{sign}\left(\nabla_x \mathcal{L}(\pi'(x) a_{\mathrm{bad}})\right),$$

  where ϵ controls the attack strength and a_{bad} is an unsafe target action used to guide the perturbation. The perturbation is applied to the sensor observation.
  This attack is designed to make the policy misinterpret free space and obstacle locations, leading to unsafe commands possible collisions.

RETRAIN DESIGN: MIXED ADVERSARIAL RETRAINING

- At each environment step, the policy uses either a clean observation or an adversarial perturbed sensor observation. Clean steps maintain the original task performance. Adversarial steps generate sensor perturbations using the gradient of an attack loss toward an unsafe target action. The perturbation is applied only to the sensor dimensions, while other input data remain unchanged. The attacked observation is used for policy inference and rollout collection.
- PPO is then updated normally using a rollout buffer containing both clean and adversarial transitions.
- **Clean Step:** obs_clean → policy → action → env step → rollout buffer
- **Adversarial step:** obs_clean → policy → attack loss → sensor gradient → obs_adv → policy → action
- **PPO update:** mixed rollout buffer → normal PPO update

EVALUATION

- Experimental Task:
  Reach-avoid navigation requires the robot to reach a target while avoiding obstacles.
- Evaluation Setting:
  We compare the original policy and the retrained policy under clean observations, random noise, and FGSM sensor attacks.
- Evaluation Metrics:
  Success and Fail rate of 100 tasks.

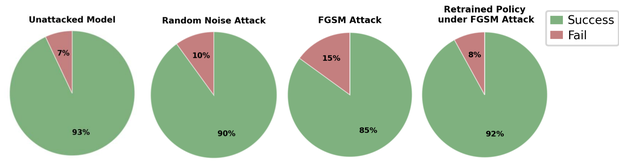


Figure 3: Policy performance under clean observations, random sensor noise, FGSM sensor attacks, and FGSM-attacked observations using the retrained policy.

- Under FGSM-attacked observations, the retrained policy achieves a 92% success rate, which is lower than the clean setting but higher than both the random-noise setting and the original policy under FGSM attack. This shows that mixed retraining improves robustness against adversarial sensor perturbations while maintaining strong task performance.

REFERENCES

[1] Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR 2015.

ACKNOWLEGEMENTS

This work is supported by funding for the VICEROY Northwest Institute for Cybersecurity Education and Research (CySER) provided by The Office of the Undersecretary of Defense for Research and Engineering, in collaboration with the Air Force Research Laboratory and Griffiss Institute.

WASHINGTON STATE UNIVERSITY