2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

When the Adversary Knows You Better: Adversarial Training for Learning-Based Legged Robots

Qinchao Xu[1], Satoshi Yagi[1], Satoshi Yamamori[1,2], Jun Morimoto[1,2]

***Abstract*— Deep reinforcement learning has emerged as the dominant paradigm for training legged robots to locomote, however, when deployed in unstructured, dynamically varying real-world environments, the safety of neural network based controllers remains insufficiently guaranteed. Prior studies have demonstrated that sequential adversarial attacks, formulated via reinforcement learning, can effectively expose latent vulnerabilities in controllers and thus serve as a valuable complement to Domain Randomization techniques. These methods, however, are inherently constrained by the assumption that both the adversary and the locomotion policy share identical state space inputs. In contrast, our approach overcomes this limitation by incorporating privileged information into the adversarial network's observation input, thereby more than doubling the attack success rate. Furthermore, we mitigate the controller's tendency toward overly conservative behavior under attacks by introducing stochastic termination criteria. We validate the proposed method in real-world deployments, showing that it not only significantly enhances robustness but also preserves original task performance.**

I. Introduction

Deep reinforcement learning (DRL) empowers the automatic synthesis of locomotion policies via large-scale, simulation-based trial and error [1], dramatically reducing the need for laborious hand-engineering of control heuristics [2], [3]. When applied to quadrupedal robots, DRL enables the emergence of agile and adaptive behaviors that exploit body dynamics to cope with diverse terrains. By leveraging techniques such as Domain Randomization [4] and curriculum learning [1], [5], DRL agents can acquire complex [6]–[8], non-intuitive gaits that generalize across a wide array of environments. Nonetheless, these neural network controllers remain largely opaque [9] and can fail unpredictably when exposed to out-of-distribution scenarios, sensor corruptions, or subtle perturbations, posing significant safety risks in real-world deployments [10], [11]. Systematically probing and reinforcing policies against adversarial disturbances is therefore crucial to uncover latent vulnerabilities and bolster controller reliability under unforeseen conditions.

Adversarial attacks have been widely applied in image recognition [12], [13], speech recognition [14], and autonomous driving [15], typical attacks inject imperceptible perturbations into neural-network inputs to compromise their performance. In the robotics domain, Shi et al. [10] diversified these approaches by proposing multi-modal attacks tailored to locomotion controllers. Their multi-modal attack approach extends adversarial attacks into the perturbation space, command space, and observation space, corresponding to continuous perturbation forces, malicious commands, and sensor noise, respectively. Although their framework offers valuable insights into enhancing the safety and robustness of locomotion controllers, it still faces several limitations: (1) By relying on hand-crafted auxiliary rewards to encourage exploration, the adversary agent becomes vulnerable

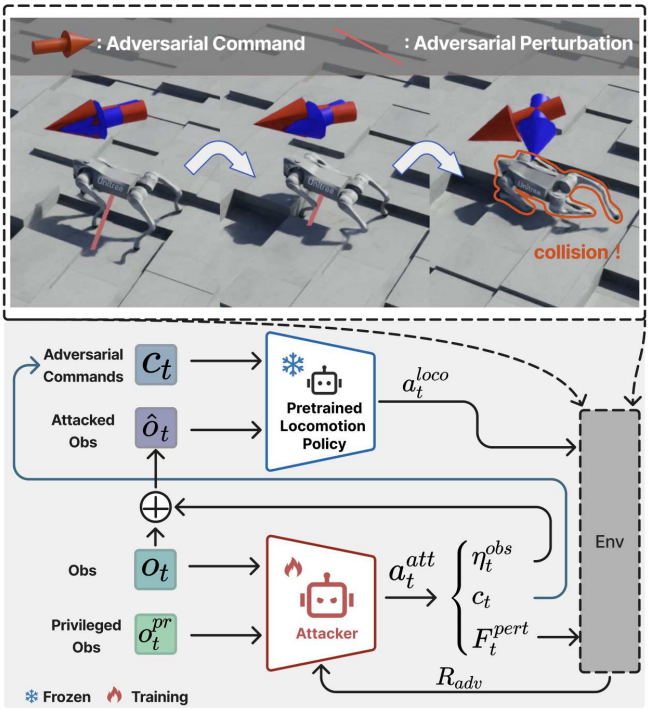


Fig. 1: Proposed framework for training adversarial attack policies. For details, see Sec. III-B. The locomotion policy and adversarial policy actions are denoted by a_t^{loco} and a_t^{att}, respectively. Observation noise η_t^{obs}, adversarial command c_t, and perturbation force F_t^{pert} are applied to the system, while the adversarial reward R_{adv} is computed as defined in Eq. 1. The snapshots illustrate a successful attack sequence: red arrows and lines depict adversarial commands and perturbations, with their lengths and directions indicating magnitudes and orientations of malicious inputs, while blue arrows represent the robot's body linear velocity.

*This work was supported by JST Moonshot R&D, Grant Number: JPMJMS223B-3, and JSPS KAKENHI Grant Number: 22H04998 and 23K24925.

[1]Learning Machines Group, Graduate School of Informatics, Kyoto University, Kyoto, Japan.
xu.qinchao@lm.sys.i.kyoto-u.ac.jp, {yagi, morimoto}@i.kyoto-u.ac.jp

[2]Dept. of Brain Robot Interface, Computational Neuroscience Labs, ATR, Kyoto, Japan. yamamori@atr.jp

to poorly designed incentives, causing it to converge on suboptimal local optima; (2) The adversary policy in their framework need not be transferred to real-world deployment, yet their approach constrains the adversary's ability to probe for vulnerabilities; (3) The proposed finetuning method under attacks still induces overly conservative behavior.

To address these shortcomings, we build upon previous work with three key enhancements. First, we integrate a more general form of curiosity-driven Random Network Distillation (RND) [16] into the framework to enhance exploration in complex tasks. Second, diverging from typical black-box attacks constrained by limited information [17]–[20], adversaries in robotic domains can exploit privileged state information to construct an informational advantage. Finally, during the controller's finetuning phase, we replace naive episode termination with a stochastic termination criterion [21], which enriches the informational content of the feedback available to the locomotion controller. To validate the effectiveness of our method and its enhancements, we conducted experiments both in simulation and on a real quadruped robot.

The contributions of this work can be summarized as follows:

1) A curiosity-driven reward design for adversarial agent training that eliminates the need for task-specific auxiliary rewards.
2) We pioneer the integration of privileged information into adversarial training, significantly enhancing the attack policy's capacity to uncover and exploit controller vulnerabilities.
3) A stochastic termination mechanism for the controller finetuning phase that prevents overly conservative behaviors, yielding gains in the finetuned controller's overall performance.

The remainder of this paper is organized as follows. Section II introduces related studies. Section III introduces our training pipeline. Section IV presents the experimental setups and the results. In Section V, we discuss the limitations and future work. Finally, Section VI concludes this paper.

II. Related Work

A. Adversarial Attacks via Privileged Information

Adversarial attacks remain the predominant paradigm for probing and enhancing model robustness [22]–[26]. They are broadly classified into black-box and white-box settings, distinguished by the adversary's level of access to the target model's internals: in the black-box scenario, perturbations are crafted solely from input–output queries [17]–[20], whereas the white-box scenario grants full access to architecture [27], [28], weights, and gradients, enabling direct optimization of adversarial examples [29]. Methodologically, our approach predominantly falls within the black-box paradigm, circumventing the need for intrusive access to model internal. However, we deliberately construct the adversary's informational advantage through extra privileged information, thereby bridging the efficacy gap between black-box and white-box attack scenarios.

Our perspective aligns with a broader trend in RL, where privileged information is increasingly exploited to improve decision-making. Advances in simulation fidelity [30], [31] have spurred the development of training frameworks enhanced by privileged information for improved information acquisition. For instance, privileged information integration via teacher-student architectures [6], [32], [33] has shown superior performance in various legged motion control tasks. These works directly demonstrate that privileged information provides valuable guidance to agents.

B. Robustifying Policy with Adversarial Training

Contemporary research has established adversarial training as a principled approach for hardening control policies against disturbances, with documented successes in manipulation tasks [34] and competitive multi-agent scenarios [35], [36]. Moreover, Tang et al. [37] pioneered the integration of adversarial training into locomotion control, achieving agile motion behaviors. These studies collectively demonstrate that competitive training enhances performance across a wide range of tasks.

Research on using adversarial training to improve robot robustness has been conducted previously. Takuto et al. [20] proposed adversarial torque perturbations for robot joints; their approach remained confined to the controller's action space. [38] incorporated an H_∞ constraint to generate perturbations based on the current state of the quadruped robot. However, their method overlooks instability scenarios arising from the cumulative effect of perturbations. In contrast, RL is inherently well-suited for handling long-term, sequential decision making problems [10], [19], [20].

III. Method

Our overall training pipeline is as follows: train a locomotion policy, freeze the locomotion policy and train an adversarial attack policy, freeze the adversarial policy and resume locomotion training under attacks. All reinforcement learning agents are trained using the Proximal Policy Optimization (PPO) algorithm [39] with an actor–critic network. Stage II is trained for 8,000 iterations, with convergence of the adversarial policy used as the stopping criterion. Stage III requires only 4,000 iterations during resume training. The simulation includes slope, rough, block, and stair terrains, organized via a curriculum to progressively increase difficulty.

A. Locomotion Controller

Securing a sufficiently robust locomotion controller is essential for subsequent stages; If simple attacks can compromise the controller, efforts to enhance its ability to withstand unexpected disturbances become ineffective. Building on Domain Randomization (cf. Table I), we enhanced controller robustness prior to adversarial training by adjusting the command-update sampling interval to (1, 10) seconds and introducing two additional reward terms—feet stumble and feet slide (cf. Table II). The locomotion policy's observation space, shown in Table III as $o_t \in \mathbb{R}^{235}$. Its action space is a 12-dimensional vector a_t, corresponding to the robot's desired joint angles.

TABLE I: Domain Randomization

Randomization Term	Range	Unit
Friction	$[0.3, 1.2] \times$ nominal value	-
Restitution	$[0.0, 0.2]$	-
Payload mass	$[-3, 3]$	kg
External force	$[-10, 10]$	N
External torque	$[-10, 10]$	N·m
Joint position offset	$[-0.2, 0.2]$	rad
Joint velocity offset	$[-2.0, 2.0]$	$\mathrm{m \cdot s^{-1}}$
Joint K_p	$[0.8, 1.3] \times$ nominal value	-
Joint K_d	$[0.8, 1.3] \times$ nominal value	-
Push by velocity	$[-0.65, 0.65]$	$\mathrm{m \cdot s^{-1}}$

B. Privileged Information-Based Adversarial Attacks

a) Reward: Adversarial attacks on robotic controllers constitute a classic sparse-reward task: the adversary only receives a reward upon inducing a final failure state in the robot. In this work, the combined reward R_{adv} is defined as:

$$R_{\mathrm{adv}} = \lambda_{\mathrm{term}} \cdot \mathbb{1}_{\mathrm{term}} + \lambda_{\mathrm{int}} \cdot r_{\mathrm{int}}, \tag{1}$$

where λ is the reward weight, the term $\mathbb{1}_{\mathrm{term}}$ corresponds to episode termination in adversarial training, the adversary agent receives a positive reward via $\mathbb{1}_{\mathrm{term}}$ only when the robot satisfies the termination condition. r_{int} is the intrinsic reward, calculated by RND method.

We instantiate two Multi Layer Perceptrons (MLPs) of identical architecture, $f(o)$ and $\hat{f}(o)$. $f(o)$ is the target network, which is randomly initialized and kept fixed, while $\hat{f}(o)$ is the predictor, updated by the mean squared error (MSE) as:

TABLE II: Reward Component and its weight.

Note: $\exp(\cdot)$ and $\mathbb{1}$ denote the exponential function and the indicator function, $\|\cdot\|$ is Euclidean norm. Superscripts $(\cdot)^{\mathrm{cmd}}$ and $(\cdot)^{\mathrm{des}}$ indicate commanded and desired values, respectively. Subscript $(\cdot)_{xyz}$, $(\cdot)_{yaw}$ and $(\cdot)_f$ represent the robot's body coordinate frame (xyz), the yaw rotation and the foot. Here, v, ω, g, t, θ, τ, h and F denote, respectively, the linear velocity, yaw rate, gravity vector, time for the foot in the air, joint position, joint torque, base height relative to the ground and leg net contact force, respectively. In particular, $\mathbb{1}_{\mathrm{first}}$ and $\mathbb{1}_{\mathrm{contact}}$ indicate, respectively, whether the contact is the foot's first touchdown and whether the foot is currently in contact with the ground.

Reward Component	Equation	Weight
Lin. velocity tracking	$\exp\big(-(v_{xy}^{\mathrm{cmd}} - v_{xy})^2\big)$	2.5
Ang. velocity tracking	$\exp\big(-(\omega_{yaw}^{\mathrm{cmd}} - \omega_{yaw})^2\big)$	0.75
Alive	$\mathbb{1}_{\mathrm{alive}}$	0.2
Base flat	$(1 - g_z)^2$	0.1
Feet air time	$(t^{\mathrm{air}} - t^{\mathrm{des}}) \cdot \mathbb{1}_{\mathrm{first}}$	1.0
Linear velocity (z)	v_z^2	-2.0
Angular velocity (xy)	ω_{xy}^2	-0.05
Joint accelerations	$\|\ddot{\theta}\|^2$	-5×10^{-7}
Joint velocities	$\|\dot{\theta}\|^2$	-5×10^{-3}
Joint torques	$\|\tau\|^2$	-2.5×10^{-5}
Base height	$(h^{\mathrm{des}} - h)^2$	-5.0
Action rate	$(a_t - a_{t-1})^2$	-0.05
Feet stumble	$\mathbb{1}_{\max(\|F_{xy}\|) > 3\,F_z}$	-0.5
Feet slide	$\big\|v_{f,xy}\big\| \cdot \mathbb{1}_{\mathrm{contact}}$	-0.1

$$r_{\mathrm{int}} = \Big\| f(o_c) - \hat{f}(o_c) \Big\|_2, \tag{2}$$

the input $o_c \in \mathbb{R}^{72}$ comprises the standard observation $o'_t \in \mathbb{R}^{48}$ and privileged observation $o_t^{\mathrm{pr}} \in \mathbb{R}^{24}$ (cf. Table III), but with all height scan dots removed. Motivated by the RND encoding strategy in [40], we consider height scan dots irrelevant to adversarial exploration.

b) Observation and Action: during adversarial training, creating an informational advantage over the victim policy. The privileged observation $o_t^{\mathrm{pr}} \in \mathbb{R}^{24}$ comprises enhanced foot-state features and joint net contact force measurements, enabling the adversary to identify the robot's vulnerable states such as foot liftoff and joint–terrain collisions.

The action space of the adversarial policy, termed the "adversarial space", follows the configuration in [10], comprising the perturbation space, the command space, and the observation space. Considering practical relevance, we bound each adversarial size as follows: base perturbation forces within [-25, 25] N; command within [-0.5, 0.5] m/s; and observation-noise injections with total magnitude constrained to [-0.1, 0.1]. The per-step update magnitude for any adversarial attacks is capped at 0.05 $\times$ its maximum allowable bound. Moreover, adversarially generated malicious commands supplant the locomotion task's randomly sampled commands, with updates occurring at a fixed interval of 1 seconds. Notably, our observation space attacks diverge from prior work [10], which constrained disturbances solely to IMU noise. We contend that joint-encoder noise [41]—stemming from simulation-to-reality gaps, thermal drift, encoder wear, and similar factors—constitutes a significant vulnerability requiring targeted adversarial exploration. The observation-space attacks are defined by the $o_t^{\mathrm{att}} \in \mathbb{R}^{220}$ in the Table III, excluding the velocity commands and last action.

c) Terminations: In this section, we discuss the formulation of termination criteria within the adversarial training framework. For the adversary, the termination criteria serve as the goal in a grid-world RL task, marking the successful completion of an attack. For quadrupedal robots, preventing falls—and the resulting damage to critical sensors or cameras—is paramount. Thus, we determine the criteria as two

TABLE III: Observation Space Compositions

Observation Type		Input	Dims
Observation o_t	Attack o_t^{att}	base linear velocity	3
		base angular velocity	3
		project gravity vector	3
		joint position	12
		joint velocity	12
		height scan dots	187
	Unattacked o_t^{unatt}	velocity commands	3
		last action	12
Privileged Observation o_t^{pr}		feet airtime	4
		thigh and shank contact	8
		feet contact forces	12

parts:

$$t_{\text{orientation}} = \begin{cases} 1, & \text{if } |\arccos\left(-\mathbf{g}_{\text{proj}} \cdot \hat{\mathbf{z}}\right)| > \theta_{\text{lim}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

$$t_{\text{contact}} = \begin{cases} 1, & \text{if } \max_{j \in \{\text{hip, head}\}} \|f_j\|_2 > F_{\text{th}}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

here, $\mathbf{g}_{\text{proj}}$ denotes the gravity projection vector, $\hat{\mathbf{z}}$ the unit vector along the z-axis, and θ_{lim} the prescribed angular threshold. In the definition of t_{contact}, contacts occurring at the hip joints and at the robot's head are undesirable, f_j is the contact force vector and F_{th} is the predefined force threshold.

C. Locomotion Controller Finetuning

In this section, we revisit controller training by freezing the adversary policy and resuming the original locomotion training from the previous checkpoint. In RL locomotion training, it is typical to employ naive termination mechanisms [6], [42] with hand-crafted conditions. However, naive termination is ill-suited for the finetuning phase: adversarial attacks can trivially trigger termination, causing premature episode truncation and severely limiting informative learning signals for the locomotion policy. Conversely, removing termination altogether provides no explicit constraint on undesirable behaviors. This motivates a principled middle ground between hard termination and no termination.

Accordingly, in this training phase, we replace the naive termination with a stochastic termination criterion. This method is adopted from [21], and we refer readers to that work for further details. Specifically, we adopt the same return computation as follows:

$$\max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \left(\prod_{t'=0}^{t} \gamma(1 - \delta(s_{t'}, a_{t'})) \right) r(s_t, a_t) \right], \tag{5}$$

where $r(s_t, a_t)$ follows the same reward composition as Table II. We reformulate Eq. 3 and Eq. 4 into the corresponding constraint function, based on which the termination probability function $\delta(s_{t'}, a_{t'})$ is computed as:

$$\delta = \max_{i \in I} p_i^{\max} \text{clip}\left(\frac{c_i^+}{c_i^{\max}}, 0, 1 \right), \tag{6}$$

Here, c_i^+ is the violation of constraint function i. For the critical hyperparameter $p_i^{\max}$, unlike [21], where events such as falling are treated as hard constraints with $p_i^{\max} = 1$, this setting is unsuitable in adversarial training scenarios where constraint violations occur frequently. Instead, we adopt a curriculum strategy in which $p_i^{\max}$ gradually increases from 0.05 to 0.4 over training iterations.

IV. Experiments

In this section, we empirically evaluate our adversarial training framework in both simulation and the real robot, focusing on three aspects: the contribution of each component, the performance of the controller after finetuning, and the transferability of robustness from simulation to real-world deployment.

A. Ablation Study

To evaluate the contribution of each component in our adversarial training, we conducted the ablation study shown in Fig. 2. In this ablation experiment only, we attenuate the adversary's capabilities by restricting its action space to perturbation-force attacks bounded within [-25, 25] N; The rationale for this design is that multi-modal attacks induce controller failures so readily that they mask the gains in attack efficacy; accordingly, we deliberately weaken the adversary's capabilities and conduct this simplified experiment. In Fig. 2, the horizontal axis denotes training iterations, and the vertical axis is positively correlated with the attack success rate.

Through comparison, we found that privileged information is the most effective way to improve attack performance, doubling the success rate of attacks. This will also create an advantage when facing sufficiently robust advanced controllers in the future. As for RND, our motivation for integrating RND was to encourage exploration, as we observed that adversary tend to converge to monotonous and easily predictable behaviors. In the training curves shown in Fig. 2 for the model (purple line) augmented with privileged information and RND, although convergence speed is somewhat reduced, the exploration-driven gains in attack success rate persist over the long term.

Building on this experiment, we further conduct an ablation study to investigate which types of privileged information contribute most to the adversary's ability to identify vulnerabilities. As shown in Table IV, feet airtime contributes the most to the adversary's probing capability among all privileged information components. Although feet contact forces may implicitly indicate foot lift-off through the vertical contact force, the adversary agent lacks access to historical information, making it infeasible to infer airtime from instantaneous contact forces alone. Regarding feet airtime, we further hypothesize that granting the adversary access to richer temporal history would further increase attack success rates.

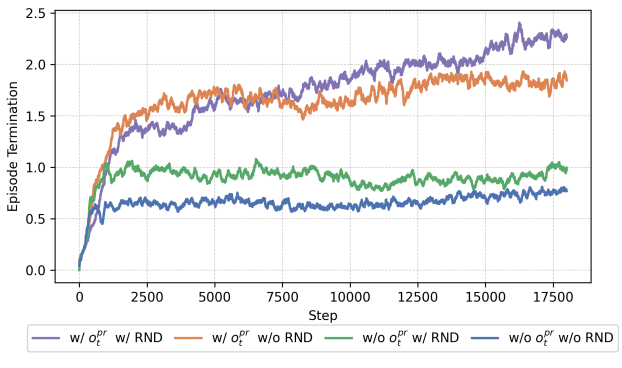


Fig. 2: Contribution of each component in our adversarial training. The horizontal axis denotes training steps, and the vertical axis plots the episodic count aggregated over all termination terms.

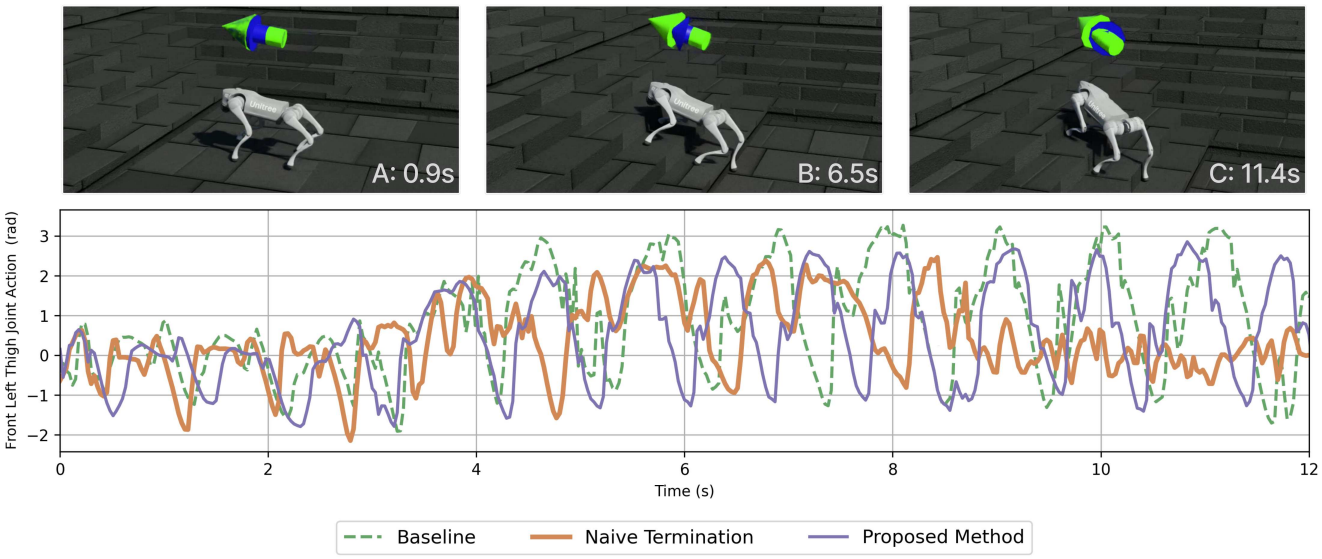


Fig. 3: Staircase-climbing performance different policies. Figures A–C illustrate the overly conservative behavior exhibited by the naive termination approach. A: flat-ground locomotion; B: first two stairs; C: hesitation on the stairs. All tests were conducted from identical initial positions under a constant 0.5 m/s linear velocity command, indicated by the green arrow.

B. Locomotion Task Performance

For comparative evaluation, we consider the following methods:

- Baseline: The original victim policy that we trained by Sec. III-A.
- Naive Termination: Refers to the locomotion policy finetuned using the method proposed by [10], only 5% of parallel agents are under adversarial attacks, and with a naive termination mechanism.
- Proposed Method: The finetuning method we proposed in Sec. III-C.

TABLE IV: Comprehensive ablation results across different privileged observations term.

Method	Episode Termination Metric
Ours w/o feet airtime	1.125 ± 0.014
Ours w/o thigh and shank contact	1.458 ± 0.031
Ours w/o feet contact forces	2.250 ± 0.113
Ours w/ total Privileged Info. o^{pt}	2.417 ± 0.066

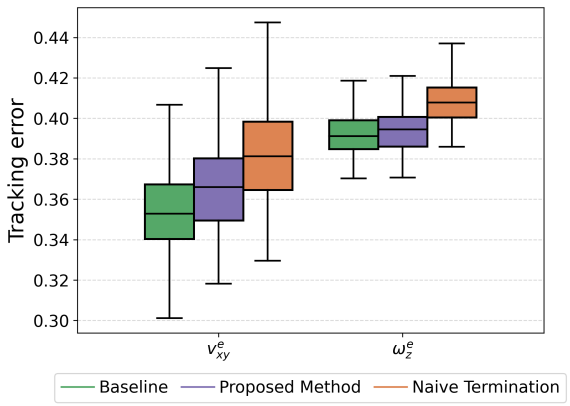


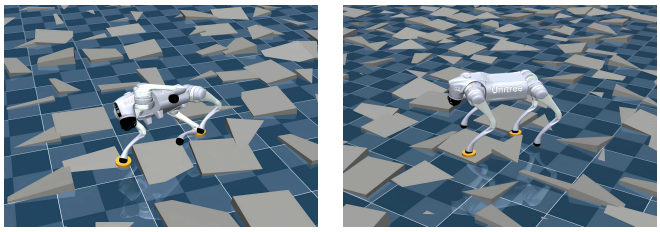
Fig. 4: Command tracking error. In the horizontal axis, v^e_{xy} (measured in m/s) denotes the tracking error between commanded and actual linear velocities in the robot's XY body frame, and w^e_{xy} (measured in rad/s) represents the yaw rotational velocity error.

While finetuning with Shi et al.'s [10] method effectively prevents overly conservative behavior on most terrains, we observe that on staircases, the policy inevitably exhibits hesitation. The reason is that, compared to other terrains, stairs make it very easy for attacks to succeed. As shown in Fig. 3, under normal circumstances, the locomotion policy equipped with exteroception (height scan dots) can easily climb stairs, but the policy finetuned using the naive termination method begins to hesitate halfway through. Moreover, across multiple trials, the proposed stochastic termination approach produced joint action magnitudes slightly below the baseline's, yet still completed the staircase-climbing task.

We concurrently evaluate all policies on the original command tracking locomotion task to assess fundamental performance. We evaluate each policy with 240 parallel agents in simulation to ensure fair average performance. The robot receives velocity commands drawn uniformly from [–1, 1] m/s, resampled at random intervals between [1, 10] s. As shown in the boxplot in Fig. 4, locomotion performance is analyzed using the mean tracking error. The baseline policy without adversarial training achieves the highest performance. Although our method outperforms previous approaches, it still falls short of this baseline. In conjunction with Fig. 3, we further posit that our training method yields policies that are sufficiently cautious without being overly conservative.

TABLE V: Average Minimum Push Velocity across Different Policies

Policy	Min. push (m/s)
Baseline	0.88
Shi [10]	1.09
Ours w/o Privileged Info. o^{pt}	1.13
Ours w/o RND	1.21
Ours w/o Stochastic Termination	1.23
Ours	**1.25**



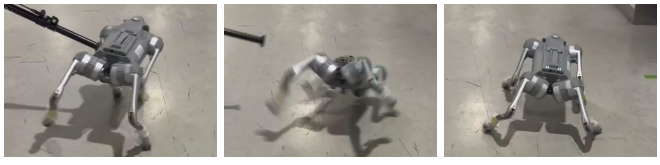
(a) (b)

Fig. 5: Sim2Sim performance on the unseen terrain. Under malicious command attacks, the baseline locomotion controller (a) performs substantially worse than the controller (b) produced by our adversarial training pipeline.

C. Robustness Analysis

The primary purpose of adversarial attacks is to enhance a specific aspect of the original policy's robustness, which in this work is primarily its physical robustness. In the IsaacLab simulation, we deployed 240 parallel agents, randomly distributed across various terrains, to evaluate the robustness of the policies described in Sec. IV-B. To simulate instantaneous perturbations, we apply random velocity impulses in the base's XY body frame every 3 seconds. We record the smallest v^{push} that causes the robot to be terminated. As shown in Table V, our method outperforms both the baseline and Shi's state-of-the-art approach [10], the latter of which indicates that it does not make use of privileged information and relies on a simple termination method. Moreover, through the additional ablation experiments, we demonstrate that privileged information yields the largest robustness gain in adversarial training, thanks to the adversary's more comprehensive vulnerability probing enabled by its informational advantage.

Shi [10]



Ours

Fig. 6: Challenging indoor perturbation experiments were conducted, the robot's feet were wrapped with smooth plastic film to simulate low-friction terrain. Through comparison, it was found that the policy trained with our method exhibited greater robustness improvements not only in simulation but also on the real robot.

Fig. 7: Field trials were conducted across diverse challenging terrains, including dense grass, slippery slopes, and gravel paths, where the controller trained with our method maintained its ability to traverse in difficult terrains.

D. Sim2Sim and Sim2Real

To facilitate policy transfer, we trained a blind locomotion policy using an asymmetric actor–critic: the critic receives the full observation o_t (cf. Table III), while the actor is deprived of base linear velocity and height scan dots inputs. We also performed malicious command injections in a more realistic MuJoCo simulation [31], targeting both the baseline locomotion policy and the policy trained with our proposed adversarial training pipeline, as shown in Fig 5. On the terrain not seen during training, we issued malicious commands via keyboard, namely sudden changes in command direction and oversized command vectors. In 15 trials, the baseline policy fell 9 times, whereas our finetuned policy failed only twice. It must be acknowledged that many training regimes [20], [38] did not account for malicious commands, representing a significant vulnerability in their robustness.

The adversarially trained controller was directly deployed on the Go2 robot's onboard Jetson Nano, running the policy at 50 Hz with the same k_p and k_d parameters as in the simulation environment. To evaluate whether the finetuned controller retains the same performance in the real world as in simulation, we first conducted robustness tests, as shown

in Fig. 6. Subsequently, we demonstrated the controller's locomotion capability and traversability across various outdoor terrains Fig. 7.

V. Limitation and Future Work

The attack-finetuning process can be regarded as one round of adversarial training. While we do not further explore alternating training procedures for jointly optimizing the locomotion controller and the adversary, this is not the primary focus of our work. Similar training procedures have already been established in paradigms such as GANs [43], we more expect our method to serve as a foundation for more comprehensive training frameworks in the future.

Future work will focus on developing constrained alternating training procedures. While progressively intensifying the adversary during training is a plausible strategy, we argue that a constrained method is necessary to limit attack magnitude and thus provide the controller sufficient breathing room to recover under strong attacks.

VI. Conclusion

In this paper, we equip the adversary with privileged information to bridge the efficacy gap between black-box and white-box attacks, substantially improving its ability to uncover controller vulnerabilities. We also advocate a more general adversarial training paradigm, which incorporates exploration-oriented rewards for the adversary and replaces the naive termination mechanism with stochastic termination in the finetuning phase. Empirical evaluation in simulation and on a real robot shows that our approach both improves vulnerability discovery and yields more robust, better-performing controllers compared with prior methods.

References

[1] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *5th Annual Conference on Robot Learning*, 2021.

[2] D. Kim, J. Di Carlo, B. Katz, G. Bledt, and S. Kim, "Highly dynamic quadruped locomotion via whole-body impulse control and model predictive control," *arXiv preprint arXiv:1909.06586*, 2019.

[3] S. Kajita, F. Kanehiro, K. Kaneko, K. Fujiwara, K. Harada, K. Yokoi et al., "Biped walking pattern generation by using preview control of zero-moment point," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 2, 2003, pp. 1620–1626 vol.2.

[4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[5] Z. Xu, X. Han, H. Shen, H. Jin, and K. Shimada, "Navrl: Learning safe flight in dynamic environments," *IEEE Robotics and Automation Letters*, 2025.

[6] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 11 443–11 450.

[7] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "Anymal parkour: Learning agile navigation for quadrupedal robots," *Science Robotics*, vol. 9, no. 88, p. eadi7566, 2024. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adi7566

[8] Z. Zhuang, Z. Fu, J. Wang, C. G. Atkeson, S. Schwertfeger, C. Finn et al., "Robot parkour learning," in *7th Annual Conference on Robot Learning*, 2023.

[9] S. Kuutti, S. Fallah, and R. Bowden, "Training adversarial agents to exploit weaknesses in deep control policies," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 108–114.

[10] F. Shi, C. Zhang, T. Miki, J. Lee, M. Hutter, and S. Coros, "Rethinking robustness assessment: Adversarial attacks on learning-based quadrupedal locomotion controllers," 2024. [Online]. Available: https://arxiv.org/abs/2405.12424

[11] X. Wang, S. Nair, and M. Althoff, "Falsification-based robust adversarial reinforcement learning," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 205–212.

[12] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *2019 IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4724–4732.

[13] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey et al., "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, p. 107332, 2021.

[14] X. Yuan, J. Zhang, K. Chen, C. Wei, R. Li, Z. Ma et al., "Adversarial attack and defense for commercial black-box chinese-english speech recognition systems," *ACM Trans. Priv. Secur.*, vol. 28, no. 1, Dec. 2024. [Online]. Available: https://doi.org/10.1145/3701725

[15] M. Zhou, W. Zhou, J. Huang, J. Yang, M. Du, and Q. Li, "Stealthy and effective physical adversarial attacks in autonomous driving," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6795–6809, 2024.

[16] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1lJJnR5Ym

[17] X. Y. Lee, Y. Esfandiari, K. L. Tan, and S. Sarkar, "Query-based targeted action-space adversarial policies on deep reinforcement learning agents," in *Proc. 12th international conference on cyber-physical systems*, 2021, pp. 87–97.

[18] Y. Ren, H. Zhang, X. Cao, C. Yang, J. Zhang, and H. Li, "Promoting or hindering: Stealthy black-box attacks against drl-based traffic signal control," *IEEE Internet of Things Journal*, vol. 11, no. 4, pp. 5816–5825, 2024.

[19] Y. Ren, H. Zhang, L. Du, Z. Zhang, J. Zhang, and H. Li, "Stealthy black-box attack with dynamic threshold against marl-based traffic signal control system," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 10, pp. 12 021–12 031, 2024.

[20] T. Otomo, H. Kera, and K. Kawamoto, "Adversarial joint attacks on legged robots," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Oct. 2022, p. 676–681.

[21] E. Chane-Sane, P.-A. Leziart, T. Flayols, O. Stasse, P. Souères, and N. Mansard, "Cat: Constraints as terminations for legged locomotion reinforcement learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 303–13 310.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[23] M. Lechner, A. Amini, D. Rus, and T. A. Henzinger, "Revisiting the adversarial robustness-accuracy tradeoff in robot learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1595–1602, 2023.

[24] M. Lechner, R. Hasani, R. Grosu, D. Rus, and T. A. Henzinger, "Adversarial training is not ready for robot learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4140–4147.

[25] L. Schott, J. Delas, H. Hajri, E. Gherbi, R. Yaich, N. Boulahia-Cuppens et al., "Robust Deep Reinforcement Learning Through Adversarial Attacks and Training : A Survey," Mar. 2024, 57 pages, 16 figues, 2 tables. [Online]. Available: https://hal.science/hal-04521876

[26] N. Carlini and D. Wagner, " Towards Evaluating the Robustness of Neural Networks ," in *2017 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2017, pp. 39–57.

[27] X. Li, Y. Li, Z. Feng, Z. Wang, and Q. Pan, "Ats-o2a: A state-based adversarial attack strategy on deep reinforcement learning," *Computers & Security*, vol. 129, p. 103259, 2023.

[28] Y. Wang, J. Liu, X. Chang, R. J. Rodríguez, and J. Wang, "Di-aa: An interpretable white-box attack for fooling deep neural networks,"

Information Sciences, vol. 610, pp. 14–32, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025522008520
[29] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in *27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, p. 3905–3911.
[30] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller et al., "Orbit: A unified simulation framework for interactive robot learning environments," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3740–3747, 2023.
[31] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.
[32] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.
[33] H. Wang, H. Luo, W. Zhang, and H. Chen, "Cts: Concurrent teacher-student reinforcement learning for legged locomotion," *IEEE Robotics and Automation Letters*, 2024.
[34] J. Duan, Q. Wang, L. Pinto, C.-C. Jay Kuo, and S. Nikolaidis, "Robot learning via human adversarial games," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1056–1063.
[35] J. Morimoto and K. Doya, "Robust reinforcement learning," *Neural Computation*, vol. 17, no. 2, pp. 335–359, 02 2005. [Online]. Available: https://doi.org/10.1162/0899766053011528
[36] J. Tu, T. Wang, J. Wang, S. Manivasagam, M. Ren, and R. Urtasun, "Adversarial attacks on multi-agent communication," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 7768–7777.
[37] Y. Tang, J. Tan, and T. Harada, "Learning agile locomotion via adversarial training," in *2020 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. IEEE, 2020, pp. 6098–6105.
[38] J. Long, W. Yu, Q. Li, Z. Wang, D. Lin, and J. Pang, "Learning h-infinity locomotion control," in *8th Annual Conference on Robot Learning*, 2024.
[39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[40] C. Schwarke, V. Klemm, M. van der Boon, M. Bjelonic, and M. Hutter, "Curiosity-driven learning of joint locomotion and manipulation tasks," in *7th Annual Conference on Robot Learning*, 2023.
[41] H. Zhang, W. Qin, Y. Gao, Q. Li, Z. Chen, and J. Zhao, "Disturbance elimination for the modular joint torque sensor of a collaborative robot," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 2405134, 2020.
[42] J. Long, Z. Wang, Q. Li, L. Cao, J. Gao, and J. Pang, "Hybrid internal model: Learning agile legged locomotion with simulated robot response," in *The Twelfth International Conference on Learning Representations*, 2024.
[43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.