Robust and Safe Locomotion under Actuator Failures using Deep Reinforcement Learning

Alessandra La Monica
Politecnico di Turin, Italy

Keven Abad
Politecnico di Turin, Italy

***Abstract*—This work investigates robustness and safety in deep reinforcement learning for robotic locomotion under actuator-level failures. Instead of focusing solely on environmental uncertainties, we explicitly model motor degradations such as reduced torque and partial joint impairment. Policies are trained using Soft Actor-Critic on two benchmark environments, Hopper and Ant, representing underactuated and redundant morphologies respectively. By combining actuator-level domain randomization with safety-aware reward shaping, we analyze whether exposure to degraded actuation during training enables policies to maintain stable behavior under unseen failure conditions. Experimental results highlight fundamental differences in how robustness emerges across morphologies, revealing trade-offs between performance, stability, and failure tolerance.**

***Index Terms*—Deep Reinforcement Learning, Robust Policy Learning, Domain Randomization, Goal-Conditioned Reinforcement Learning, Safety-Aware RL, Simulated Robotics**

I. Introduction

A. Motivation and Challenges

Reinforcement Learning (RL) has demonstrated strong potential for controlling legged robots in simulation, enabling the emergence of complex locomotion behaviors without relying on explicit dynamic models or handcrafted control laws [1], [2]. However, policies trained under nominal conditions often exhibit poor robustness when exposed to uncertainty, hardware degradation, or safety-critical constraints. Even small variations in actuator behavior or contact dynamics can lead to significant performance degradation, limiting the applicability of RL-based controllers in realistic robotic scenarios.

In particular, most standard locomotion benchmarks implicitly assume ideal actuation and stable contact conditions. When these assumptions are violated, for example due to partial actuator failures or changes in ground properties, learned policies frequently fail in an abrupt and unpredictable manner. This sensitivity highlights the need for training methodologies that explicitly account for degraded and uncertain operating conditions.

B. Robustness through Domain Randomization

Domain Randomization (DR) has emerged as an effective approach to improve robustness and generalization by exposing the policy to a distribution of environment dynamics during training rather than a single nominal model [3], [4]. While DR is commonly applied to global physical parameters such as link masses or ground friction, actuator-level degradation represents a particularly realistic and safety-critical source of failure that remains comparatively underexplored.

In real robotic systems, actuators may experience partial loss of torque, asymmetric degradation, or permanent joint impairments. Modeling such failures explicitly during training can encourage the learning of locomotion strategies that are less dependent on nominal motor performance. In this work, robustness is addressed by introducing actuator-level Domain Randomization, where the effective strength of motors is randomized during training and fixed deterministic faults are applied during evaluation. This setup enables a principled assessment of resilience to actuator failures under controlled Sim-to-Sim transfer conditions.

C. Safety-Aware Goal-Conditioned Locomotion

Beyond robustness, safety represents a fundamental requirement for deploying learning-based controllers in real-world environments. Policies optimized solely for speed or cumulative reward often exploit aggressive and unstable behaviors, which may be unacceptable in scenarios involving human interaction or fragile hardware [5].

To address this issue, this project adopts a goal-conditioned locomotion formulation, where the robot is required to reach spatial targets rather than simply maximizing forward velocity [6]. Safety-aware behavior is encouraged through reward shaping and task design, including penalties on excessive control effort, torso instability, and falls. Success is defined through robust goal-reaching criteria that require sustained stability rather than instantaneous achievement.

To further improve generalization, terrain variability is introduced via Domain Randomization of ground friction. By exposing the agent to varying contact conditions during training, the learned policy is encouraged to adopt stable locomotion strategies that are less sensitive to precise ground properties.

D. Contributions and Scope

This project investigates robustness and safety in legged locomotion using Reinforcement Learning on two benchmark environments: Hopper and Ant, representing underactuated and redundant morphologies, respectively [7]. All experiments are based on the Soft Actor-Critic (SAC) algorithm [2], chosen for its stability in continuous control tasks and its effective handling of exploration through entropy maximization.

The main contributions of this work are:

Identify applicable funding agency here. If none, delete this.

- the explicit modeling of actuator degradation through actuator-level Domain Randomization;
- the evaluation of policy resilience under deterministic actuator fault scenarios;
- the formulation of safety-aware, goal-conditioned locomotion with shaped rewards and stability constraints;
- the integration of terrain variability through friction Domain Randomization.

All robustness and safety mechanisms are implemented through modular environment wrappers, without modifying the underlying learning algorithm. This design enables controlled and interpretable comparisons across experimental settings. The experimental results analyze the trade-off between locomotion performance, robustness to actuator and terrain variations, and safety-related behavior across different robot morphologies.

II. Related Work

A. Reinforcement Learning for Legged Locomotion

Reinforcement Learning is widely used for learning locomotion policies in continuous-control simulated environments such as MuJoCo Ant and Hopper, which capture key challenges of legged robots, including nonlinear dynamics, intermittent contacts, and high-dimensional continuous state and action spaces. Policies are typically trained end-to-end from proprioceptive observations to torque commands, with episode termination triggered by falls or severe instability. Among commonly used algorithms, Soft Actor-Critic (SAC) is often preferred for its stability in continuous domains and its entropy-regularized objective, which promotes effective exploration and improved robustness. For this reason, SAC is adopted as the reference learning algorithm for both Ant and Hopper in this project, enabling consistent comparisons across different randomization and reward-shaping settings.

B. Robust Reinforcement Learning and Actuator Domain Randomization

A well-known limitation of simulation-trained policies is their sensitivity to variations in system dynamics. Domain Randomization (DR) addresses this issue by training the agent over a distribution of environment parameters rather than a single nominal model, encouraging policies that are robust to unobserved uncertainties. Beyond mass and contact parameter randomization, recent work highlights the importance of actuator-level randomization, which is particularly relevant for modeling hardware degradation and partial failures. In this project, actuator robustness plays a central role: in Hopper, actuator degradation is modeled through action scaling, while in Ant it is implemented directly at the dynamics level via MuJoCo actuator gain randomization, including both global and per-joint variations. Robustness is further evaluated through deterministic fault tests, in which torque reductions are applied to specific actuators at controlled times. This clearly separates stochastic randomization used during training from structured fault conditions used during evaluation.

C. Goal-Conditioned Locomotion and Safety-Aware Reward Shaping

An important extension of locomotion tasks is goal-conditioned Reinforcement Learning, where the robot must reach specific target positions instead of merely maximizing forward velocity. This formulation enables more interpretable performance metrics, such as success rate and time-to-goal, and is implemented in this project by augmenting observations with goal-relative information and shaping rewards based on goal progress. At the same time, increasing attention is being paid to safety-aware Reinforcement Learning, since policies optimized solely for return may exhibit aggressive or unstable behaviors. A practical approach is to incorporate safety considerations directly into the reward function, for example by penalizing excessive torso inclination or high vertical impact velocity, without resorting to fully constrained RL formulations. This allows explicit analysis of the trade-off between performance and safety. Finally, ground-friction domain randomization is commonly used to improve generalization across different surface conditions. In this work, friction randomization is combined with actuator randomization and safety-aware reward shaping, resulting in a multi-domain randomization setup that trains agents to reach goals safely under varying dynamics.

III. Methodology

A. Learning Algorithm

All experiments are based on the Soft Actor-Critic (SAC) algorithm, which serves as the common learning backbone for both the Ant and Hopper environments. SAC is an off-policy actor–critic method designed for continuous action spaces and is well suited for locomotion tasks due to its stability and entropy-regularized objective. In the implemented pipeline, SAC is used without algorithmic modifications: the policy, value functions, and optimization procedure remain identical across all experiments. As a result, any observed differences in performance, robustness, or safety behavior can be attributed exclusively to changes in environment dynamics, reward shaping, or randomization strategies rather than to differences in the learning algorithm itself.

B. Environment Structure and Training Pipeline

The experiments are conducted in MuJoCo-based locomotion environments, specifically Ant and Hopper, both of which model articulated legged robots with continuous state and action spaces. The base environments provide standard proprioceptive observations (joint positions, joint velocities, and body state) and continuous actuator commands mapped internally to joint torques. The implementation follows a modular environment design. Core physical dynamics are handled by the base environment (or a custom extension in the Ant case), while task logic, reward shaping, and evaluation protocols are implemented via environment wrappers. This separation ensures that changes in robustness mechanisms or task formulation do not interfere with the underlying dynamics. Training is performed using vectorized environments

and multiple random seeds to reduce variance. Observation normalization is applied consistently across training and evaluation, while reward normalization is disabled to preserve interpretability of shaped returns. Periodic evaluations are executed on dedicated evaluation environments that share normalization statistics with the training environment, ensuring fair and reproducible comparisons.

C. Robustness via Actuator Domain Randomization

Robustness to hardware degradation is addressed through actuator-level Domain Randomization. Actuator degradation is modeled as a multiplicative reduction in actuation strength, corresponding to a loss of available torque. In the Hopper environment, actuator randomization is implemented via a wrapper that scales selected action components before they are applied to the simulator. New actuator strength coefficients are sampled at every episode reset, exposing the agent to varying actuation capabilities across episodes while keeping actuation conditions fixed within each episode. In the Ant environment, actuator randomization is implemented directly at the dynamics level by modifying MuJoCo actuator gain parameters. Two formulations are supported: independent per-actuator weakening and global symmetric scaling applied to all actuators. In both cases, actuator gains are randomized episodically and restored to nominal values at reset to avoid cumulative effects. During training, actuator degradation is stochastic and sampled from predefined ranges. During evaluation, actuator faults are instead applied deterministically, using fixed fault specifications that define which actuator is affected, the remaining strength, and the fault onset time. This strict separation between training-time randomization and evaluation-time faults enables a clear assessment of fault tolerance and generalization.

D. Goal-Conditioned and Safety-Aware Locomotion

In addition to pure locomotion, the project implements goal-conditioned locomotion. Instead of maximizing forward velocity, the agent is required to reach a spatial goal sampled at the beginning of each episode. Goal information (relative position and distance) is appended to the observation space through a dedicated wrapper. The reward function is shaped to encourage progress toward the goal, penalize residual distance, and provide a success bonus when the goal is reached within a specified tolerance. To ensure robustness of success detection, success may be required to persist for multiple consecutive steps before being confirmed. Safety considerations are incorporated through reward shaping and environment logic, rather than through explicit constrained RL formulations. Safety-related quantities such as torso pitch and vertical velocity are monitored at each timestep. When predefined thresholds are exceeded, penalties proportional to the violation magnitude are applied to the reward. Importantly, safety violations do not necessarily terminate the episode, allowing the agent to experience unsafe states and learn recovery behaviors. This design encourages stable and recoverable locomotion without imposing hard constraints.

E. Terrain Domain Randomization with Curriculum

To improve robustness to environmental variability, the methodology includes ground friction domain randomization. Ground contact friction coefficients are scaled multiplicatively within a predefined range, affecting slip conditions and ground reaction forces. In the Ant experiments, friction randomization is combined with a curriculum strategy, where the randomization range is progressively increased during training. The agent is initially trained under near-nominal friction conditions and is gradually exposed to more variable terrain dynamics. This curriculum stabilizes early learning while ultimately promoting robustness to a wider range of contact conditions. Friction randomization is applied exclusively at the dynamics level and does not alter the task definition, reward structure, or observation space. As a result, performance differences can be attributed directly to changes in terrain dynamics rather than to task-related confounding factors.

F. Unified Perspective

Taken together, the methodology implements a unified robustness-oriented training framework that combines actuator domain randomization, goal-conditioned task formulation, safety-aware reward shaping, and terrain variability. Importantly, these mechanisms are applied in a modular and controlled manner: actuator and friction randomization affect only the transition dynamics, while goal conditioning and safety shaping define the task objective. This design enables controlled ablation and evaluation of different robustness dimensions. Actuator randomization targets resilience to hardware degradation, friction randomization addresses environmental uncertainty, and safety-aware goal conditioning ensures meaningful and stable behavior. By keeping the learning algorithm fixed and varying only environment dynamics and task structure, the methodology provides a clear and interpretable assessment of how robustness and safety emerge from experience in Reinforcement Learning–based locomotion.

IV. Experimental Evaluation

This section presents the experimental evaluation of the proposed approaches in the Ant and Hopper environments. The experiments assess robustness to actuator degradation, goal-conditioned locomotion performance, and safety and stability under domain shift. All results are obtained using the evaluation pipeline described in the Methodology section and are consistent with the implementations reported in the accompanying project documentation.

A. Experimental Scenarios and Compared Methods

Two main experimental settings are considered. The first focuses on robustness to actuator degradation, comparing agents trained with and without actuator domain randomization. In this setting, a baseline agent trained under nominal actuation is evaluated against a resilient agent trained with randomized actuator strength. The second setting investigates goal-conditioned locomotion under environmental variability, comparing three agents: a baseline goal-conditioned agent, an

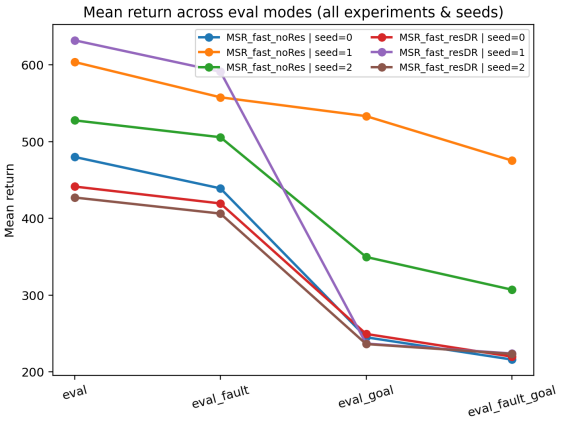

Fig. 1. Mean episode return across evaluation modes and random seeds.

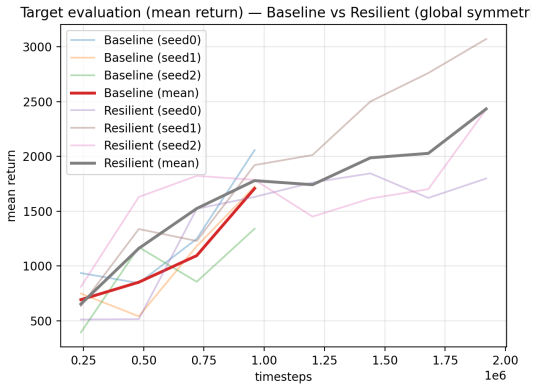

Fig. 2. Target-domain mean return under nominal actuation.

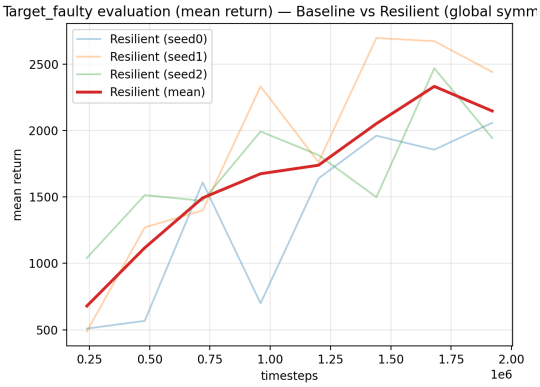

Fig. 3. Target-domain mean return under actuator degradation.

improved baseline variant, and a goal-conditioned agent augmented with safety-aware reward shaping and terrain friction randomization. All agents are trained using the same learning algorithm and comparable training budgets. Performance is evaluated using identical protocols and multiple random seeds to ensure a fair and reproducible comparison.

B. Evaluation Protocol and Metrics

- **eval**: nominal dynamics without faults;
- **eval_fault**: deterministic actuator degradation applied at evaluation time;
- **eval_goal**: goal-conditioned task without actuator faults;
- **eval_fault_goal**: goal-conditioned task with actuator degradation.

The following metrics are reported:

- **Mean episode return**, measuring overall task performance;
- **Success rate**, defined as the fraction of episodes in which the goal is reached;
- **Mean time-to-goal**, computed only over successful episodes;
- **Safety violation rate**, measuring the frequency of violations of torso pitch or vertical speed limits.

While mean return is informative for locomotion tasks, it is inherently dependent on the reward definition and task structure. For goal-conditioned experiments, success rate and time-to-goal provide more interpretable indicators of task performance.

Figure 1 reports the mean return across evaluation modes for all experiments and seeds, providing a global overview of performance degradation under increasing task difficulty.

C. Robustness to Actuator Degradation

Robustness to actuator degradation is evaluated by introducing deterministic torque reductions during evaluation while keeping training conditions unchanged. Under nominal evaluation conditions, baseline and resilient agents achieve comparable performance, indicating that actuator randomization does not hinder learning in the absence of faults. When actuator degradation is introduced, the baseline agent exhibits a sharp performance drop, reflected in a significant reduction in mean return. In contrast, the resilient agent maintains substantially higher performance across faulted scenarios. This behavior is consistent across random seeds and fault configurations. These results indicate that actuator domain randomization during training improves generalization to unseen actuation conditions. Rather than relying on nominal actuator effectiveness, the resilient agent learns control strategies that distribute effort across multiple joints, making the policy less sensitive to localized loss of actuation. The learning dynamics and the effect of faults are illustrated in Figures 2 and 3, which report target-domain performance with and without actuator degradation.

TABLE I
ANT ENVIRONMENT: FINAL PERFORMANCE UNDER NOMINAL AND FAULTY ACTUATION.

Method	Target (final)	Target-faulty (final)	Rel. drop
Baseline	1706.85 ± 293.17	–	–
Resilient (global sym)	2432.86 ± 519.47	2148.46 ± 212.67	0.086 ± 0.164

D. Goal-Conditioned Locomotion under Domain Shift

In the goal-conditioned setting, agents are evaluated based on their ability to reliably reach spatial targets under variable

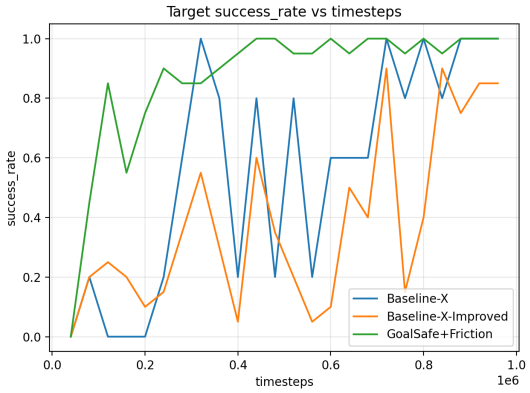

Fig. 4. Target success rate during goal-conditioned training.

Fig. 5. Target mean return during goal-conditioned training (secondary metric).

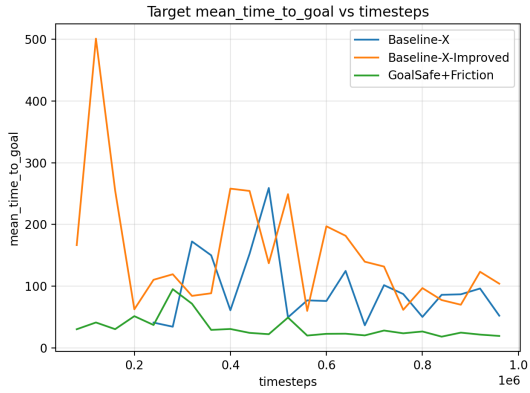

Fig. 6. Mean time-to-goal on the target domain.

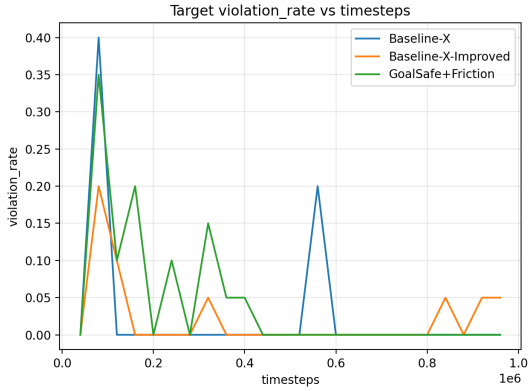

Fig. 7. Safety violation rate during goal-conditioned training.

dynamics. The goal-conditioned agent with safety-aware reward shaping and friction randomization consistently outperforms baseline variants in terms of learning stability and final performance. Baseline agents display high variance in success rate and require substantially more training steps to achieve consistent goal-reaching behavior. In contrast, the safety-aware agent rapidly converges to high success rates and maintains stable performance throughout training. Mean return trends are reported for completeness, but should be interpreted with caution due to their dependence on shaped rewards. Success rate curves, shown in Figure 4, provide a clearer picture of task completion capability and highlight the benefits of combining goal conditioning with environmental randomization.

E. Efficiency and Safety Analysis

Beyond success rate, behavior quality is assessed in terms of efficiency and safety. The safety-aware agent consistently reaches targets in fewer timesteps, as reflected by lower mean time-to-goal values compared to baseline agents (Figure 6). At the same time, safety violation rates decrease steadily during training and converge to near-zero values. Baseline agents, instead, continue to exhibit occasional violations even at later training stages, indicating more aggressive or unstable behaviors. These results show that incorporating safety-related penalties into the reward function does not merely discourage unsafe behavior, but actively guides the learning process toward smoother and more stable locomotion strategies, without compromising task success.

To further quantify the observed differences, Table II reports the final evaluation metrics for goal-conditioned navigation on the Ant environment. The GoalSafe+Friction agent achieves a 100% success rate on both source and target domains, while baseline variants exhibit lower success rates and higher variance across domains.

In terms of efficiency, the safety-aware agent reaches the target in approximately 19 timesteps on average, compared to values above 50 and up to 104 timesteps for baseline configurations. Importantly, this improvement in efficiency does not come at the cost of safety: violation rates converge to zero, and no residual instability is observed at evaluation time.

These results confirm that safety-aware reward shaping and terrain friction randomization jointly promote fast, stable, and reliable goal-reaching behavior under domain shift, outperforming baseline goal-conditioned agents across all considered metrics.

TABLE II
FINAL EVALUATION METRICS FOR GOAL-CONDITIONED NAVIGATION ON THE ANT ENVIRONMENT.

method	domain	timesteps	success_rate	mean_traj_to_goal	violation_rate	mean_violations	mean_final_dist	mean_return	mean_qtn
Baseline-X	source	960,048	0.800	69.000	0.000	0.000	0.526	2.442	255.200
Baseline-X	target	960,048	1.000	52.200	0.000	0.000	0.469	1.655	52.200
Baseline-X Improved	source	960,048	0.800	124.250	0.100	0.150	0.643	1.289	307.000
Baseline-X Improved	target	960,048	0.850	104.059	0.050	0.100	0.564	1.645	273.350
GoalSafe+Friction	source	960,048	1.000	19.850	0.000	0.000	0.188	15.668	20.300
GoalSafe+Friction	target	960,048	1.000	19.350	0.000	0.000	0.232	15.870	26.400

F. Unified Discussion Across Tasks and Domains

Across the different experimental settings, a consistent pattern emerges: robustness and safety arise from structured exposure to variability during training rather than from explicit fault modeling or hard constraints at evaluation time. Actuator domain randomization improves resilience to hardware degradation by encouraging policies that do not rely on precise actuation strength. Terrain friction randomization and safety-aware reward shaping promote stable and efficient goal-directed behavior under environmental uncertainty. By keeping the learning algorithm fixed and modifying only environment dynamics and reward structure, the experiments demonstrate that robust and stable locomotion behaviors can emerge naturally from experience. This unified evaluation highlights the effectiveness of domain randomization and reward shaping as practical tools for improving the reliability of reinforcement learning–based locomotion controllers.

V. DISCUSSION

The results demonstrate that robustness and safety in reinforcement learning–based locomotion can be effectively improved by exposing agents to structured variability during training. Across both Ant and Hopper environments, actuator-level domain randomization leads to higher resilience under degraded actuation while also improving generalization under nominal target dynamics, indicating a regularization effect rather than excessive conservatism. In goal-conditioned tasks, safety-aware reward shaping combined with terrain randomization yields faster convergence, higher success rates, and reduced safety violations compared to baseline agents. Task-level metrics such as success rate and time-to-goal prove more informative than raw return in this setting, showing that increased reliability does not come at the expense of efficiency. Overall, the consistent trends observed across tasks and robot morphologies suggest that robust and stable locomotion behaviors can emerge naturally by keeping the learning algorithm fixed and modifying only environment dynamics and reward structure, without relying on explicit fault models or constrained optimization frameworks.

VI. CONCLUSION

This work investigated robustness and safety in reinforcement learning–based locomotion through a unified experimental framework evaluated on the Ant and Hopper environments. By keeping the learning algorithm fixed and modifying only environment dynamics and reward structure, the study isolated the effects of actuator domain randomization, terrain variability, and safety-aware task formulation. The results show that exposing agents to structured variability during training leads to policies that generalize better under domain shift and degrade more gracefully in the presence of actuator degradation. In particular, actuator-level domain randomization improves fault tolerance without sacrificing nominal performance, while goal-conditioned locomotion with safety-aware reward shaping yields higher success rates, faster task completion, and significantly reduced safety violations. Across both environments, robustness and safety emerge as properties learned through experience rather than enforced through explicit constraints or external controllers. These findings support the use of domain randomization and reward shaping as practical and modular tools for improving the reliability of reinforcement learning controllers in legged locomotion tasks.

VII. FUTURE WORK

Several directions can be explored to extend the results presented in this work. First, the current experiments consider either actuator degradation or terrain variability in isolation. Future work could investigate the combined effect of multiple simultaneous sources of uncertainty, such as concurrent actuator faults and terrain randomization, to further stress-test policy robustness. Second, while safety is enforced through reward shaping and termination logic, future extensions could incorporate explicit safety constraints or constrained reinforcement learning formulations to provide stronger theoretical guarantees. Third, the current evaluation is limited to simulation-based Sim-to-Sim transfer. Extending the framework to Sim-to-Real transfer would represent a natural and impactful next step, particularly for assessing the practical relevance of actuator-level domain randomization. Finally, the framework could be extended to more complex robotic platforms and tasks, such as dynamic obstacle avoidance or multi-agent scenarios, to further evaluate scalability and generalization. The modular structure of the environment wrappers makes such extensions straightforward and highlights the flexibility of the proposed methodology.

REFERENCES

[1] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel, "Trust region policy optimization," *International Conference on Machine Learning (ICML)*, 2015.

[2] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *International Conference on Machine Learning (ICML)*, 2018.

[3] 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," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[4] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[5] G. Dalal, D. Gilboa, S. Mannor, and G. Chechik, "Safe exploration in continuous action spaces," *arXiv preprint arXiv:1801.08757*, 2018.

[6] T. Schaul, D. Horgan, K. Gregor, and D. Silver, "Universal value function approximators," *International Conference on Machine Learning (ICML)*, 2015.

[7] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.