Curiosity-Driven Learning of Joint Locomotion and Manipulation Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract: Learning complex locomotion and manipulation tasks presents significant challenges, often requiring extensive engineering of, e.g., reward functions or curricula to provide meaningful feedback to the Reinforcement Learning (RL) algorithm. This paper proposes an intrinsically motivated RL approach to reduce task-specific engineering. The desired task is encoded in a single sparse reward, i.e., a reward of "+1" is given if the task is achieved. Intrinsic motivation enables learning by guiding exploration toward the sparse reward signal. Specifically, we adapt the idea of Random Network Distillation (RND) to the robotics domain to learn holistic motion control policies involving simultaneous locomotion and manipulation. We investigate opening doors as an exemplary task for robotic applications. A second task involving package manipulation from a table to a bin highlights the generalization capabilities of the presented approach. Finally, the resulting RL policies are executed in numerous real-world experiments on a wheeled-legged robot in biped mode. We experienced no failure in our experiments, which consisted of opening push doors (16 times in a row) and manipulating packages (6 times in a row).

Keywords: Curiosity, Reinforcement Learning, Wheeled-Legged Robots

1 Introduction

Recent advancements in Reinforcement Learning (RL) have propelled legged and wheeled-legged robots beyond the confines of the lab, empowering them to fulfill a broad range of practical applications [1, 2, 3]. Nevertheless, tasks in environments designed for human interaction remain challenging as they frequently require coordination between locomotion and manipulation. Research often aims to reduce task complexity by manually dividing the problem into multiple stages [4]. However, this approach lacks the exploration of holistic solutions that consider the interplay between different stages. Solving these problems end-to-end with minimal task-specific engineering remains an open challenge. This work proposes an RL approach that solves the task as a whole with little need for task-specific engineering. We adopt a sparse reward setting, rewarding "+1" when the desired task is achieved. Although simple in its formulation, the discovery of sparse rewards is in general challenging and requires strategies for the exploration of the environment. In our work,

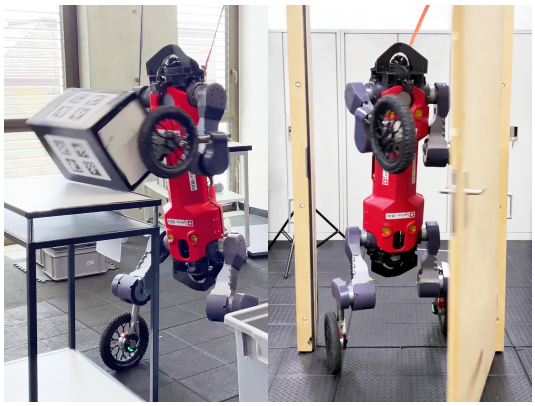

Figure 1: A wheeled-legged robot picking up a package and opening a door while balancing on two wheels. Motions can be seen in the supplementary video.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

intrinsic motivation is employed to guide the agent. For the first time, the concept of Random Network Distillation (RND) [5], which represents intrinsic motivation, has been successfully applied and validated in a real-world robotic task.

A dynamic wheeled-legged robot in biped mode [6] serves as the experimental platform. We validate the proposed method in simulation and in real-world experiments by solving tasks that are common in human environments, such as door opening and package manipulation, as shown in Fig. 1. Thereby, we are able to show the task-independent properties of the chosen curiosity-driven exploration approach and its robustness on hardware. Succinctly, our main contributions can be summarized as follows:

1. A curiosity-driven sparse reward RL approach for learning end-to-end locomotion and manipulation tasks without laborious, task-specific engineering (Section 4)
2. Successful real-world validation of RL policies on a wheeled-legged robot in biped mode, demonstrating a high level of repeatability (Section 5.2)
3. The notion of curiosity states as guiding mechanisms (Section 3.2)
4. An analysis of emerging learning behaviors in sparse reward settings (Section 5.1)

2 Related Work

The increase in computational power has made RL a viable choice for control [7]. Especially in control problems that require contact-rich interactions, RL controllers achieve state-of-the-art performance [8, 9]. However, the reward function commonly includes a variety of dense reward terms that need to be tuned extensively to result in the desired behavior. In contrast, learning in sparse reward settings, while often being easier to define, can be inefficient or not possible. Because the reward only exists in a small fraction of the state space, random exploration induced by, e.g., epsilon-greedy policies is not sufficient to pick up the reward signal in reasonable time [10].

2.1 Mechanisms to Guide Exploration

Guided exploration toward the reward can improve learning performance over random exploration. We group existing approaches into the following three main categories:

- **Expert Demonstrations**, usually in the form of state or state-action trajectories, can be an effective tool to teach a desired skill [11, 12, 6, 13]. However, demonstrations are predominantly hand-crafted in a time-consuming process to achieve a single task. Furthermore, the agent is biased towards the provided, potentially suboptimal, trajectories.
- **Curriculum Learning** [14] involves gradually increasing task difficulty during training. In most RL approaches, this is achieved by altering either the environment [15], reward function [16], or initial or terminal state distribution [17, 18]. However, a manual definition of a curriculum is necessary. While some work exists on automated curriculum discovery [17, 19], generation and efficient scheduling of intermediate tasks are subject to ongoing research with no general solution.
- **Intrinsic Motivation**, more specifically curiosity, is an essential driver for human learning [20]. The ability to learn without external rewards allows agents to acquire skills that can be valuable for future tasks or facilitate the learning of more intricate abilities. Sparked by novel stimuli, curiosity rewards exploring the unknown for the pure sake of knowledge gain. Integrating this mechanism into our learning algorithms holds great promise, particularly due to its task-agnostic nature, which distinguishes it from expert demonstrations or curriculum learning.

2.2 Mechanisms for Intrinsic Motivation

Past research has explored three main approaches to modeling intrinsic motivation:

- **Rewarding Surprise:** In an initial endeavor to incorporate curiosity into an RL algorithm, [21] introduced a reward based on the Euclidean distance between the prediction of a model learning the environmental forward dynamics and the observed transition. Intuitively, the model is expected to perform poorly in unfamiliar regions with limited data, where it is "surprised" by unforeseen state transitions [10]. Consequently, the agent receives a high reward, encouraging subsequent visits to the novel region. As the dynamics model improves over time, the reward diminishes. The concept of measuring novelty using prediction error from a dynamics model raises a significant concern: it tends to favor stochastic transitions, as they are difficult or even impossible to predict accurately. For instance, a TV screen displaying white noise would be unpredictable, leading to a high reward and an agent that might prefer watching TV indefinitely [22]. To address this challenge, [23] adopts a different approach by predicting a state embedding instead of the complete world state.
- **Rewarding Learning Progress:** Another line of work focuses on estimating the expected learning progress within specific regions of the state space [24, 25, 26]. To apply such methods, the state and action spaces are discretized, enabling their implementation in a tabular RL setting. However, measuring learning progress in high dimensional continuous state spaces remains computationally infeasible [23].
- **Rewarding State Novelty:** Count-based exploration methods reward state novelty directly by keeping track of the number of visits for each state and prioritizing less visited states [27]. To tackle exploration in continuous state spaces, [28] proposes pseudo-counts, a generalization of count-based exploration methods. Burda et al. [5] propose RND, a method based on predicting information about the current state to measure state novelty. They train a prediction model alongside the RL agent in a supervised fashion, improving its accuracy for visited states during the learning process. The prediction error can then be utilized as an intrinsic reward signal, where familiar states yield more accurate predictions compared to less visited or unvisited states. While rewarding surprise might suffer from stochastic transitions and rewarding learning progress involves complex measures, estimating state novelty is straightforward and intuitive, even in continuous state spaces. Given our focus on loco-manipulation[1] tasks characterized by continuous state spaces and potentially complex task dynamics, we consider the third option as the most suitable.

2.3 Loco-Manipulation Tasks

Existing approaches to solving complex loco-manipulation tasks include splitting the problem into a learning-based controller for locomotion and a model-based controller for manipulation [29], or using human demonstrations and curricula [30].

In this study, the primary task under examination is opening doors. A quadrupedal robot by Boston Dynamics equipped with an arm has demonstrated impressive performance in the task of opening doors [31]. However, limited information is available regarding the specific approach employed, only that it is model-based. To break down complexity, a common approach is to divide the task into sub-tasks that can be solved in sequence [32, 4]. This requires task-specific modeling and tailoring of the control scheme, which can be identified as a common shortcoming of the studied approaches. In the following, we aim for a more holistic approach.

3 Curiosity Formulation

Structured exploration is a crucial factor for successful learning in sparse reward settings. RND [5] offers an intuitive and computationally efficient approach to intrinsic motivation, while the proposed curiosity state focuses exploration.

[1] Moving while manipulating objects.

3.1 Random Network Distillation

The RND module consists of two function approximators. A randomly initialized target network $\boldsymbol{f}$ encodes states $\boldsymbol{s} \in \mathcal{S}$ into an unknown embedding $\boldsymbol{f}(\boldsymbol{s})$. The target stays fixed during the whole training process. A predictor network $\hat{\boldsymbol{f}}$ estimates the target's embedding, given the same input $\boldsymbol{s}$ as the target. The predictor network is trained alongside the RL agent on the visited states in a supervised fashion with a Mean Squared Error (MSE) loss. The prediction error, i.e., the difference between the outputs of both networks serves as the intrinsic reward signal defined by

$$r_{\text{intrinsic}} = \left\| \boldsymbol{f}(\boldsymbol{s}) - \hat{\boldsymbol{f}}(\boldsymbol{s}) \right\|_2 . \tag{1}$$

Intuitively, familiar state regions yield small prediction errors as the predictor is already trained on similar states. Not yet visited state regions lead to large errors and therefore large intrinsic rewards. As the agent visits an unfamiliar region repeatedly, the prediction error decreases.

3.2 Curiosity State

While [5] uses the agent's observations $\boldsymbol{o}$ instead of the state $\boldsymbol{s}$, we propose a more flexible formulation where a *curiosity state* $\boldsymbol{s}_c = \boldsymbol{\phi}(\boldsymbol{s})$ is passed to the RND module to stay independent of the observations. The mapping $\boldsymbol{\phi}$ can be freely chosen, as long as the state $\boldsymbol{s}$ can be implicitly inferred from the environment's feedback. This way, curiosity can be focused on the desired quantities. While this formulation allows for arbitrary mappings $\boldsymbol{\phi}$, it suffices to consider $\boldsymbol{\phi}$ to select a subset of the full state without further modification. A sketch of the implemented RND module is shown in Fig. 2.

4 Task-Specific RL Formulation

We illustrate an effective sparse reward formulation on the exemplary task of door opening, detailing the chosen rewards, observations, and the introduced curiosity state. Subsequently, we adapt the formulation to the task of package manipulation to demonstrate the straightforward generalization of the proposed approach to different tasks.

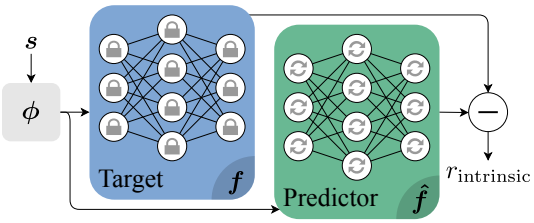

Figure 2: Random Network Distillation (RND) module. A mapping $\boldsymbol{\phi}$ encodes the full system state $\boldsymbol{s}$ into the curiosity state $\boldsymbol{s}_c$. The target network with fixed weights (blue) embeds the curiosity state $\boldsymbol{s}_c$ into an unknown representation. The predictor network (green) is updated during training and attempts to match the target's output. The difference between outputs serves as the intrinsic reward signal.

4.1 Rewards

The chosen reward function consists of the three distinct reward terms

$$r = r_{\text{intrinsic}} + r_{\text{task}} + r_{\text{shaping}}. \tag{2}$$

The first term $r_{\text{intrinsic}}$ is defined by equation 1 and motivates the agent to explore the relevant part of the state space. In this work, we use Multi Layer Perceptrons (MLPs) with 1 and 2 hidden layers with 5 neurons and a one-dimensional output for the target and predictor network, respectively. The second term r_{task} is the only task-specific reward. For the task of door opening, it is defined intuitively by repeatedly giving a reward of +1 while the task is achieved, i.e., when the door hinge angle q_{hinge} is in a desired interval, and no reward otherwise:

$$r_{\text{task}} = \begin{cases} 1, & \text{if } 1.5 < q_{\text{hinge}} < 2.1 \\ 0, & \text{otherwise} \end{cases} . \tag{3}$$

The inclusion of the third term r_{shaping} serves as an incentive for the robot to maintain a standing posture and adopt more refined, less forceful strategies. This emphasis on smoother and less aggressive policies is crucial for achieving effective sim-to-real deployment within the realm of RL control [15, 33]. For a comprehensive breakdown of the rewards and their respective weights, please refer to the supplementary material.

4.2 Observations

In terms of observations, our approach focuses on minimizing task dependency and avoids the use of complex exteroceptive information. The only observation that pertains specifically to the door is the relative position ${}_{\mathcal{C}}\boldsymbol{r}_{CH}$ of the door handle origin in the robot's camera frame. Additionally, to enable the robot to determine the degree to which the door is open, we provide the initial door handle position ${}_{\mathcal{C}}\boldsymbol{r}_{CH_{\text{init}}}$ as input to the policy.

As for standard observations that are not task-specific and have been covered in previous works [6, 3], we provide a list of these observations in the supplementary material. We note that the observations are subject to empirical normalization, to prevent bad scaling of input variables.

4.3 Curiosity Implementation

An advantage of the proposed curiosity state $\boldsymbol{s}_c$ is its independence from the robot's observations. Since the RND module is only needed for training, it is possible to include quantities that are easy to attain in simulation but hard to estimate in reality and thus unfit as observations. To focus the agent's curiosity on the door, door hinge and handle angles $\boldsymbol{q}_{\text{door}}$ as well as their angular velocities $\dot{\boldsymbol{q}}_{\text{door}}$ are included in the curiosity state. Adding the distance between the robot and the door handle d_{CH} induces faster interaction with the door. To avoid an intrinsic reward signal for moving too far away from the door, the distance is clipped according to

$$d_{CH} = \min(\|\boldsymbol{r}_{CH}\|_2, 2), \tag{4}$$

and the curiosity state is therefore defined as

$$\boldsymbol{s}_c = \begin{bmatrix} \boldsymbol{q}_{\text{door}}^\top & \dot{\boldsymbol{q}}_{\text{door}}^\top & d_{CH} \end{bmatrix}^\top. \tag{5}$$

4.4 Generalization to Package Manipulation

We show the task independence and generalization capability of the proposed approach by subjecting it to a second task requiring different locomotion and manipulation skills. We choose the exemplary task of package manipulation, i.e. grabbing, moving, and dropping a package, as it involves a freely moving object, as opposed to the fixed-base articulated door. To encode the task, we define a similar sparse task reward r_{task}, given by

$$r_{\text{task}} = \begin{cases} 1, & \text{if } {}_{\mathcal{I}}\boldsymbol{r}_{\text{package}} \in \mathcal{S}_{\text{bin}} \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where ${}_{\mathcal{I}}\boldsymbol{r}_{\text{package}}$ is the position of the package in the inertial frame $\mathcal{I}$ and $\mathcal{S}_{\text{bin}}$ includes the space in and above the bin. We include the space above the bin to reduce reward sparsity as the agent still learns to drop the package. It is however not necessary to include this space for successful learning. Instead of observing the door handle position, the agent now observes the relative package, table, and bin positions ${}_{\mathcal{C}}\boldsymbol{r}_{CP}$, ${}_{\mathcal{C}}\boldsymbol{r}_{CT}$, and ${}_{\mathcal{C}}\boldsymbol{r}_{CB}$, respectively. As before, relative positions are observed in the camera frame. The last necessary change concerns the curiosity state. We define it as

$$\boldsymbol{s}_c = \begin{bmatrix} {}_{\mathcal{I}}\dot{\boldsymbol{r}}_{\text{package}}^\top & d_{PB} & d_{CP} \end{bmatrix}^\top, \tag{7}$$

where ${}_{\mathcal{I}}\dot{\boldsymbol{r}}_{\text{package}}$ is the linear package velocity in the inertial frame. The distances between package and bin d_{PB} and camera and package d_{CP} are again clipped as in equation 4.

5 Experimental Validation and Discussion

To validate our approach, we conduct experiments in simulation and the real world. For implementation details please refer to the supplementary material. All experiments are conducted with the wheeled-legged quadrupedal robot in Fig.1, a dynamic robot that can perform hybrid motions between walking and driving. Recently, [6] discovered a bipedal locomotion mode through RL, further increasing the robot's versatility. In bipedal mode, the front legs can serve as arms to manipulate objects.

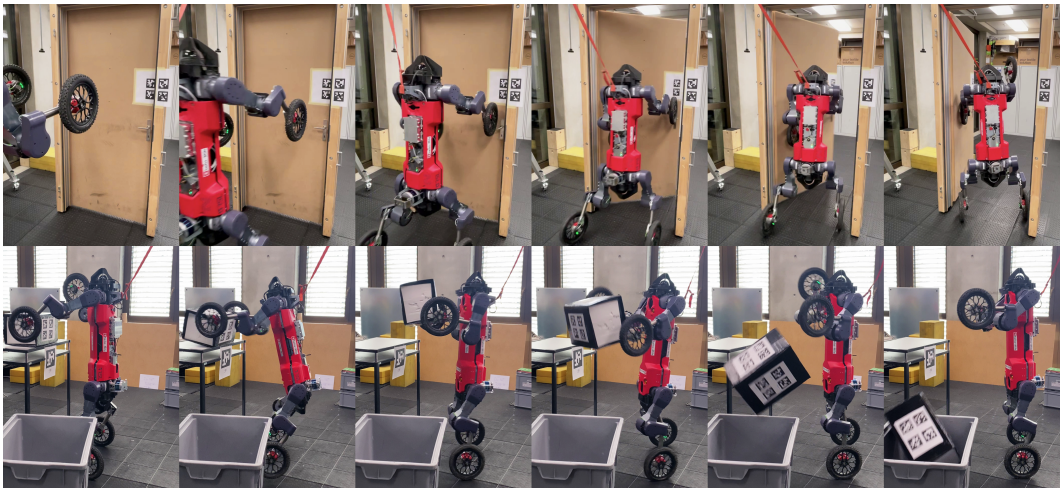

Figure 3: Snapshot sequences (left to right) of the push door and package manipulation experiments. Opening the door takes 2.5 seconds, grasping and dropping the package 1.5 seconds.

5.1 Simulation Results

The main task we investigate is door opening. To enable pull door experiments, we equip the robot with basic hooks attached to the front wheels. Qualitative examples of the door opening and package grabbing motions can be seen in the supplementary video. In the following, we discuss the relevant findings of learning with curiosity.

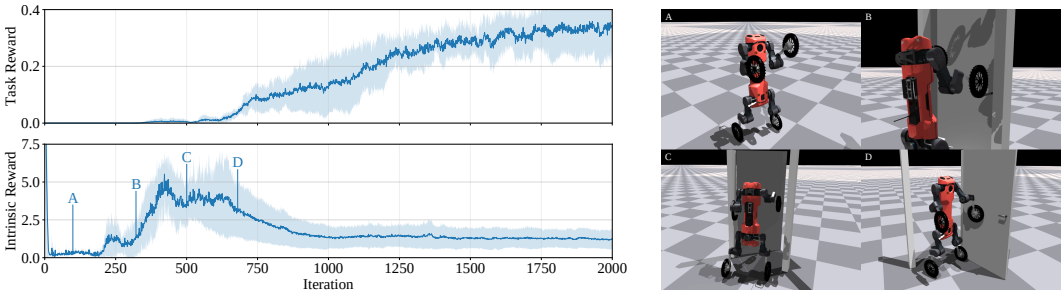

Figure 4: Task specific reward r_{task} and intrinsic reward $r_{\text{intrinsic}}$ over the course of learning to open a push door. Out of 6 training runs with different randomization seeds, 4 pick up the sparse reward signal and constitute the shown mean and standard deviation. After learning how to stand (A), the robot starts to play with the door handle and the intrinsic reward increases (B). Next, the robot opens the door slightly while still manipulating the door handle (C). The agent then discovers the sparse door opening reward and starts to optimize towards it (D).

- **Training Evaluation:** The need for guided exploration can be seen in a training run with extrinsic reward signals only. Indeed, the investigated skills are not learned. Instead, the learning process ends in a local optimum and the agent learns to stand without moving. The added intrinsic reward signal guides the agent towards states that involve manipulating the door or package. An exemplary learning process is shown in Fig. 4 for a push door. Interestingly, the skill is not discovered in every training run. We attribute this phenomenon to shaping rewards that hinder skill discovery. Shaping rewards increase the threshold of the sparse task reward signal needed to optimize the skill. If the threshold is not reached by exploring the door through curiosity, the skill is not learned. The learning process for package manipulation evolves similarly. First, the robot discovers the package and moves it on the table to increase the intrinsic reward signal. After

learning to grab the package with both wheels, the robot starts to wiggle the package and moves it closer to the bin. Once the sparse task reward is found, the intrinsic reward decreases as the agent optimizes its behavior to achieve the desired task.

- **RND Evaluation:** Experiments show that the network architecture of the RND module can be kept minimal. Small networks are efficient in terms of computational resources and suffice to describe a complex function over the state space. Too large networks hinder successful learning and intrinsic reward convergence. While a predictor of the same size as the target should be able to fully approximate the target, choosing a larger predictor leads to more consistent intrinsic reward curves over different training runs.
  The weight of the intrinsic reward does not require extensive tuning as learning is successful for a large interval. Choosing a weight that is too small can slow down training, whereas selecting a weight that is too large can keep the agent curious and leads to policies that continue to play with the door. Normalizing the reward empirically as in [5] does not improve training and prevents the reward from converging to values close to zero. Again, convergence to small values is necessary to model the loss of interest as the agent gets more familiar with the environment. Normalization scales the reward with its standard deviation and thus leads to larger rewards at the end of training. This work omits normalized intrinsic rewards for the given reasons.
  In contrast, normalization of the curiosity state is crucial for a proper decay of the intrinsic reward over the course of training. The curiosity state includes quantities that can reach high magnitudes like velocities. Directly using those as input complicates optimization due to, e.g., differences in scales for different input variables and vanishing gradients. The curiosity state is therefore subjected to empirical normalization, i.e. it is normalized with a running estimate of its mean and variance. While normalization improves the convergence properties of the intrinsic reward, it violates the theoretical foundation of state visitation-based curiosity. During training, the same state keeps changing its normalized representation due to varying normalization parameters. Different states could have almost identical representations for different points in time and thus also almost identical target embeddings. Even though this distorts the measure for state familiarity, curiosity state normalization does not seem to have a negative influence on exploration and is highly recommended to improve convergence to a small reward signal.

- **Emerging Behaviors:** A key finding of this paper is the sensitivity of the learning process to small changes in the training environment. Changing the seed used to initialize networks or set randomized quantities is enough to alter the resulting policy significantly. RL approaches usually rely on dense task rewards that heuristically steer the agent towards a feasible trajectory. The sparse task reward setting does not bias the agent towards any trajectory. Randomness in the exploration can thus lead to the discovery of different minima and trajectories, especially in contact-rich scenarios. As a result, different behaviors emerge as shown in Fig. 5 where the robot is holding the door open in a variety of poses. For the task of package manipulation, differences are more subtle. During task execution, policies mainly differ in the leg configuration and the stepping pattern of the robot. The emergence of different behaviors can be beneficial to find solutions that would have been overlooked in a heuristically influenced setting. However, fine-tuning a policy's behavior is not straightforward.

5.2 Real-World Results

We demonstrate the effectiveness of the proposed approach by opening a push door over 16 times in a row without a single failure in a lab environment. Learned policies are able to let the robot stand and navigate toward the door. The robot reaches for the door handle with its right wheel and attempts to press it while pushing against the door. As soon as the door is unlocked, the robot swings the door open and holds it open while standing still, as shown in the top row of Fig. 3. Sim-to-real transfer benefits from modeling the robot's Field of View (FOV) in training, explained in the supplementary material. Since the real robot lacks the added hooks for pulling doors, we leave pull door experiments to future work.

For the task of package manipulation (shown in the bottom row of Fig. 3), the robot robustly grabs the package and drops it into the bin 6 times in a row. In one instance out of all tests, the robot lost its grip on the package. However, the robot quickly tried to re-grab the package and successfully delivered it in the end.

Policies exhibit highly dynamic behavior. If the task requires more gentle motion, a task reward that does not favor quick task completion would be more appropriate.

5.3 Limitations

Depending on the desired task, the intrinsic-motivation approach might come with some drawbacks. Firstly, the exploration process is governed by randomness. On the one hand, this propels the discovery of behaviors that might have been overlooked with dense approaches. On the other hand, it complicates the process of enforcing a specific behavior. Although dense shaping rewards can be added to enforce the desired behavior, they might inhibit exploration of the chosen task, leading to unsuccessful training.

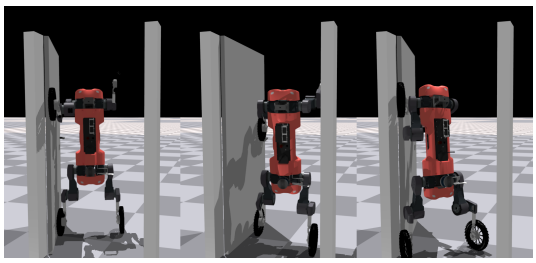

Figure 5: A wheeled-legged robot holding the door open in a variety of poses.

Secondly, since the intrinsic reward signal is not vanishing completely, the agent stays curious about its environment even after finding the task reward. In our case, this did not cause unwanted behaviors. If it does for other applications of RND, the weight of the reward could be scheduled.

Lastly, as the predictor network is trained with supervised learning, overfitting to specific regions of the state space could occur. Although not observed in this work, this might be exploited by the RL agent in a back-and-forth fashion. Subsequent switching to different state space regions would gain intrinsic reward repeatedly. Common methods to avoid overfitting such as regularization or dropouts could be employed in that case.

6 Conclusions and Future Work

We show that intrinsic motivation for exploration proves successful in simulation, yielding motion policies for complex tasks that involve locomotion and manipulation. Our RL method proves to generalize over multiple tasks requiring different sets of skills. We present results for two example tasks: door opening and package manipulation. We note that different behaviors emerge for small changes in the training environment. This phenomenon is explained by the absence of dense task rewards that bias the agent towards specific trajectories and is inherent to the sparse reward setting.

The introduced notion of a curiosity state guides exploration toward the reward in an efficient manner and allows learning of various tasks with basic task-dependent observations. We note that curiosity state normalization is crucial for proper reward convergence during the course of training, even though it distorts the measure of state familiarity.

To validate the proposed method, trained motion policies are executed on a wheeled-legged robot in biped mode. Experiments show that the robot is able to successfully and robustly open a push door in a lab environment, over 16 times in a row without failure, as well as manipulate packages through a grabbing, moving, and dropping motion over 6 times in a row without failure. We conclude that the investigated approach proves valuable for the robotics control domain as it enables the learning of highly complex skills with a minimal amount of task-specific engineering.

Future research could involve further investigation into the chosen curiosity formulation. Other potential continuations of this work could consist of implementing the wheel-attached hooks to open pull doors or investigating setups including multiple tasks.

References

[1] J. Lee, J. Hwangbo, and M. Hutter. Robust recovery controller for a quadrupedal robot using deep reinforcement learning. *arXiv e-prints*, 2019.

[2] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):2822, 2022.

[3] J. Lee, M. Bjelonic, and M. Hutter. Control of wheeled-legged quadrupeds using deep reinforcement learning. In *Robotics in Natural Settings: CLAWAR 2022*, pages 119–127. Springer, 2022.

[4] H. Ito, K. Yamamoto, H. Mori, and T. Ogata. Efficient multitask learning with an embodied predictive model for door opening and entry with whole-body control. *Science Robotics*, 7 (65):8177, 2022.

[5] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. In *Seventh International Conference on Learning Representations*, pages 1–17, 2019.

[6] E. Vollenweider, M. Bjelonic, V. Klemm, N. Rudin, J. Lee, and M. Hutter. Advanced skills through multiple adversarial motion priors in reinforcement learning. *arXiv preprint arXiv:2203.14912*, 2022.

[7] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

[8] N. Rudin, D. Hoeller, M. Bjelonic, and M. Hutter. Advanced skills by learning locomotion and local navigation end-to-end. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2497–2503. IEEE, 2022.

[9] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, M. Wulfmeier, J. Humplik, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *arXiv e-prints*, 2023.

[10] J. Achiam and S. Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

[11] M. Kalakrishnan, L. Righetti, P. Pastor, and S. Schaal. Learning force control policies for compliant manipulation. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 4639–4644. IEEE, 2011.

[12] J. Ramírez, W. Yu, and A. Perrusquía. Model-free reinforcement learning from expert demonstrations: a survey. *Artificial Intelligence Review*, pages 1–29, 2022.

[13] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4): 1–20, 2021.

[14] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[15] N. Rudin, D. Hoeller, P. Reist, and M. Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning*, pages 91–100. PMLR, 2022.

[16] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. Wiele, V. Mnih, N. Heess, and J. T. Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pages 4344–4353. PMLR, 2018.

[17] C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pages 482–495. PMLR, 2017.

[18] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba. Hindsight experience replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5055–5065, 2017.

[19] C. Florensa, D. Held, X. Geng, and P. Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528. PMLR, 2018.

[20] W. McDougall. *An introduction to social psychology*. Methuen, 1908.

[21] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[22] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Mental Develop.*, 2(3):230–247, 2010.

[23] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[24] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.*, 11(2):265–286, 2007.

[25] A. Baranes and P.-Y. Oudeyer. Robust intrinsically motivated exploration and active learning. In *2009 IEEE 8th International Conference on Development and Learning*, pages 1–6. IEEE, 2009.

[26] M. Frank, J. Leitner, M. Stollenga, A. Förster, and J. Schmidhuber. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics*, 7:25, 2014.

[27] G. Ostrovski, M. G. Bellemare, A. Oord, and R. Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[28] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1479–1487, 2016.

[29] Y. Ma, F. Farshidian, T. Miki, J. Lee, and M. Hutter. Combining learning-based locomotion policy with model-based manipulation for legged mobile manipulators. *IEEE Robot. Autom. Lett.*, 7(2):2377–2384, 2022.

[30] J. Merel, S. Tunyasuvunakool, A. Ahuja, Y. Tassa, L. Hasenclever, V. Pham, T. Erez, G. Wayne, and N. Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)*, 39(4):39, 2020.

[31] Boston Dynamics. Testing robustness, 2018. URL https://youtu.be/aFuA50H9uek.

[32] M. Stuede, K. Nuelle, S. Tappe, and T. Ortmaier. Door opening and traversal with an industrial cartesian impedance controlled mobile robot. In *Proc. IEEE Int. Conf. Robot. Autom.*, pages 966–972. IEEE, 2019.

[33] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):5986, 2020.