2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Leveraging Symmetry in RL-based Legged Locomotion Control

Zhi Su*,[2], Xiaoyu Huang*,[1], Daniel Ordoñez-Apraez[3], Yunfei Li[2], Zhongyu Li[1], Qiayuan Liao[1], Giulio Turrisi[3], Massimiliano Pontil[3], Claudio Semini[3], Yi Wu[2,4], Koushil Sreenath[1]

***Abstract*— Model-free reinforcement learning is a promising approach for autonomously solving challenging robotics control problems, but faces exploration difficulty without information about the robot's morphology. The under-exploration of multiple modalities with symmetric states leads to behaviors that are often unnatural and sub-optimal. This issue becomes particularly pronounced in the context of robotic systems with morphological symmetries, such as legged robots for which the resulting asymmetric and aperiodic behaviors compromise performance, robustness, and transferability to real hardware. To mitigate this challenge, we can leverage symmetry to guide and improve the exploration in policy learning via equivariance / invariance constraints. We investigate the efficacy of two approaches to incorporate symmetry: modifying the network architectures to be strictly equivariant / invariant, and leveraging data augmentation to approximate equivariant / invariant actor-critics. We implement the methods on challenging loco-manipulation and bipedal locomotion tasks and compare with an unconstrained baseline. We find that the strictly equivariant policy consistently outperforms other methods in sample efficiency and task performance in simulation. Additionaly, symmetry-incorporated approaches exhibit better gait quality, higher robustness and can be deployed zero-shot to hardware.**

I. Introduction

The field of robotics has witnessed a surge in the adoption of data-driven reinforcement learning (RL) methods to tackle the control problems of legged locomotion [1], [2], navigation [3], and manipulation [4]. This trend is primarily fueled by the ability of these methods to (i) cope with phenomena that impact the system evolution but are challenging to model analytically, (ii) autonomously acquire control strategies without the need of extensive domain knowledge. However, commonly-used model-free RL methods often treat the robot as a black-box system; by neglecting analytical models of dynamics, they often remain agnostic to the properties of robot's morphology. Furthermore, these methods face exploration difficulties to learn multi-modalities, especially symmetric modalities [5] where the under-exploration of some modes leads to asymmetric behaviors that are often unnatural and sub-optimal. For example, failure to fully capture the two symmetric modalities of bipedal locomotion leads to limping behaviors and reduced control performance, compromising robustness and transferability to real hardware.

Leveraging symmetries in Markov decision processs (MDPs) is a promising direction to alleviate the difficulty in symmetric-modality learning and provides a strong bias that we can leverage to improve the policy's exploration.

*Equal contribution. [1]UC Berkeley, CA, USA. [2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [3]Istituto Italiano di Tecnologia, Italy. [4]Shanghai Qi Zhi Institute, Shanghai, China. Contact: haytham.huang@berkeley.edu

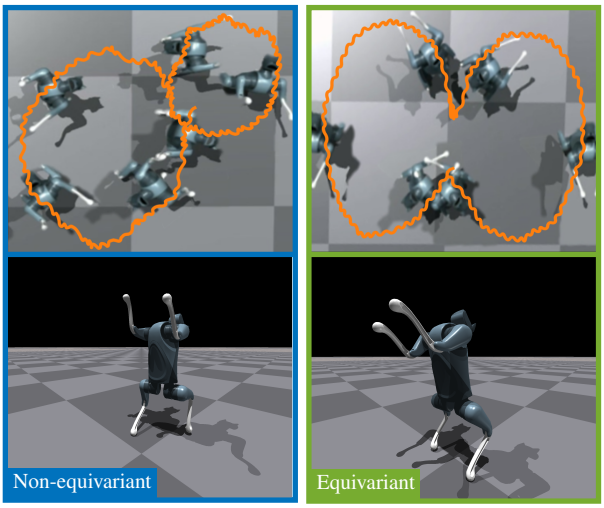

Fig. 1: Comparison between non-equivariant (left) and equivariant (right) control policies, of a quadrupedal robot with a sagittal reflection symmetry, performing a right / left bipedal turning task. Top plots show the trajectories of a commanded turn in opposite directions. Bottom figures visualize the gait pattern the policy learns. While the unconstrained policy learns an asymmetric gait between left and right feet and fails to perform a symmetric turning trajectory, the equivariant policy achieves both motion-level and task-level symmetry. For more experimental results, see `https://youtu.be/Ad1clt4Yi4U`

Specifically, since symmetric MDPs possess equivariant optimal control policies [6], [7], posting equivariance requirements on the RL policy helps it approach such an optimal policy. Since symmetry is ubiquitous in both biological and robotic systems, it is indeed an important but under-explored question of the best way to properly incorporate symmetry into RL algorithms. We note that although equivariance has gained attention in other fields such as computer vision and graphics, there have been limited prior works that apply it on legged robots and demonstrate its efficacy in related tasks.

In this work, we investigate different methods for adding equivariance information in improving the exploration guidance in training model-free RL algorithms. Specifically, we investigate how a loosely equivariant policy trained by data augmentation as carried out in [8] compare to a strictly equivariant policy enforced by its network architecture on legged robot control. We perform extensive experiments in the challenging tasks of loco-manipulation and bipedal locomotion as quadrupeds. We benchmark the two methods against a vanilla RL policy in simulation environments and showcase the optimality of the symmetry-incorporated policies. Our experimental results show that the equivariance-enforced policy consistently outperforms other variants both

in terms of the performance metrics and the acquired gait patterns being more steady and natural.

We demonstrate the sim-to-real capability of both methods for completing a loco-manipulation task and a bipedal locomotion task, and show the enhanced robustness of the symmetry-incorporated policies compared to a vanilla RL policy. We provide a detailed discussion on the robustness of the two variants on real-world hardware, which could provide guidance for future development using symmetry-incorporated RL methods for different tasks.

II. Background

Here, we present a brief introduction to the fundamental concepts and notation required to understand how morphological symmetries can be leveraged in robot control.

A. Symmetry group actions and representations

We will study the robot's morphological symmetry using the principles of group theory, a branch of mathematics that investigates symmetry transformations as abstract mathematical entities, separate from the objects they are linked with. This abstraction allows us to investigate how the robot's reflectional symmetry is imprinted on various vector spaces (such as the robot's state space and the MDP's action space) and functions of interest (including the robot's dynamics, and the MDP's value and reward functions).

Our focus is on the reflection symmetry group, denoted as $\mathbb{G} := \mathbb{C}_2 = \{e, g_s \mid g_s^2 = e\}$. This group consists of the identity transformation, denoted by e, and the reflection transformation, denoted by g_s, such that the transformation is its own inverse $g_s \circ g_s := g_s^2 = e$.

To express the action of this group on a vector space $\mathcal{X} \subseteq \mathbb{R}^n$ we define a group representation on $\mathcal{X}$, denoted $\rho_{\mathcal{X}} : \mathbb{G} \to \mathbb{GL}(\mathcal{X})$. This is a homomorphism (i.e., structure preserving) map between the symmetry transformations and the group of invertible linear maps on $\mathcal{X}$. The group structure is preserved in the sense that we can represent the composition of two symmetries by a matrix-matrix multiplication, $\rho_{\mathcal{X}}(g_a \circ g_b) = \rho_{\mathcal{X}}(g_a)\rho_{\mathcal{X}}(g_b) \mid \forall\ g_a, g_b \in \mathbb{G}$, and symmetry inversion by matrix inversion, $\rho_{\mathcal{X}}(g^{-1}) = \rho_{\mathcal{X}}(g)^{-1} \mid \forall\ g \in \mathbb{G}$. This allows us to represent the action of a symmetry on a point $\boldsymbol{x} \in \mathcal{X}$ as a matrix-vector multiplication, i.e., $g \triangleright \boldsymbol{x} := \rho_{\mathcal{X}}(g)\boldsymbol{x} \in \mathcal{X}$. Here, $(\triangleright) : \mathbb{G} \times \mathcal{X} \to \mathcal{X}$ denotes the action of the group on the vector space.

In this context, we will denote vector spaces that possess such a group action as *symmetric vector spaces*. Next, we will define the symmetric properties of functions of symmetric vector spaces.

B. Equivariant and invariant functions

A function $f : \mathcal{X} \to \mathcal{Y}$ between two symmetric vector spaces typically falls into one of two categories: group invariant or group equivariant. If the output of the function f remains constant regardless of the transformation applied to the input, the function is deemed $\mathbb{G}$-invariant. Conversely, a function is termed $\mathbb{G}$-equivariant if applying a transformation to the input before computing the function yields the same result as computing the function first and then applying the transformation to the output. These conditions can be formally defined as follows for all $g \in \mathbb{G}, \boldsymbol{x} \in \mathcal{X}$:

$$\underbrace{f(\boldsymbol{x}) = f(\rho_{\mathcal{X}}(g)\boldsymbol{x})}_{\mathbb{G}\text{-invariant}} \quad \text{and} \quad \underbrace{\rho_{\mathcal{Y}}(g)f(\boldsymbol{x}) = f(\rho_{\mathcal{X}}(g)\boldsymbol{x})}_{\mathbb{G}\text{-equivariant}}. \tag{1}$$

C. Symmetric Markov Decision Processes

An MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, r, T, p_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition density function, indicating the probability $T(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ of transitioning to state $\boldsymbol{s}'$ given the current state $\boldsymbol{s}$ and action $\boldsymbol{a}$. Lastly, $p_0 : \mathcal{S} \to [0, 1]$ is the probability density of initial states, denoting $p_0(\boldsymbol{s})$ as the probability of starting the Markov process at state $\boldsymbol{s}$.

An MDP is considered to have a symmetry group $\mathbb{G}$ if both state and action spaces are symmetric vector spaces (with the corresponding group representation $\rho_{\mathcal{S}}$ and $\rho_{\mathcal{A}}$), and if it satisfies the following conditions for all $g \in \mathbb{G}, \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$ [6], [7]:

- The transition density is $\mathbb{G}$-invariant

$$T(g \triangleright \boldsymbol{s}'|g \triangleright \boldsymbol{s}, g \triangleright \boldsymbol{a}) = T(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) \tag{2a}$$

- The density of initial states is $\mathbb{G}$-invariant

$$p_0(g \triangleright \boldsymbol{s}) = p_0(\boldsymbol{s}) \tag{2b}$$

- The reward function is $\mathbb{G}$-invariant

$$r(g \triangleright \boldsymbol{s}, g \triangleright \boldsymbol{a}) = r(\boldsymbol{s}, \boldsymbol{a}) \tag{2c}$$

(2a), (2b) describe the necessary constraints on the MDP's dynamics to ensure that at every time t, the state probability density remains $\mathbb{G}$-invariant, $p_t(g \triangleright \boldsymbol{s}) = p_t(\boldsymbol{s}) \mid \forall\ t \geq 0, g \in \mathbb{G}, \boldsymbol{s} \in \mathcal{S}$. This characteristic is indicative of dynamical systems with $\mathbb{G}$-equivariant dynamics, implying that the temporal evolution of a given state $\boldsymbol{s}$, on average, mirrors the evolution of its symmetric states $g \triangleright \boldsymbol{s} \mid \forall\ g \in \mathbb{G}$ (refer to [9]).

The significance of studying such MDPs lies in the known symmetry constraints of the optimal control policy $\pi^* : \mathcal{S} \to \mathcal{A}$ and the optimal value function $V^{\pi^*} : \mathcal{S} \to \mathbb{R}$. Specifically, symmetric MDPs possess $\mathbb{G}$-equivariant optimal control policies [7],

$$g \triangleright \pi^*(\boldsymbol{s}) = \pi^*(g \triangleright \boldsymbol{s}) \mid \forall\ \boldsymbol{s} \in \mathcal{S}, g \in \mathbb{G}, \tag{3}$$

and $\mathbb{G}$-invariant optimal value functions [7],

$$V^{\pi^*}(g \triangleright \boldsymbol{s}) = V^{\pi^*}(\boldsymbol{s}) \mid \forall\ \boldsymbol{s} \in \mathcal{S}, g \in \mathbb{G}. \tag{4}$$

These constraints provide substantial inductive biases that can be leveraged in reinforcement learning algorithms [6], [10], [11]. As we will discuss, the sagittal symmetry inherent in a quadruped robot's morphology allows us to formulate the controlled robot dynamics as a symmetric MDP.

D. Morphological symmetries and symmetric MDPs

In this work we aim to learn loco-manipulaiton control policies for the quadruped robot Xiaomi CyberDog2 [12] and Unitree Go1 [13]. These robots possess a sagittal state symmetry arising from the symmetric mass distribution of the robot's torso, and the replication of the left-right limbs.

This symmetry, referred to as a morphological symmetry, enable us to cast the robot's controlled dynamics as a symmetric MDP, considering that, due to the morphological symmetry, the robot's state vector space, denoted $\mathcal{S} \in \mathbb{R}^n$, and action space, denoted $\mathcal{A} \in \mathbb{R}^{n_a}$, are symmetric vector spaces [9], [14], with the action space being spanned by the robot's $n_a \leq n$ controlled degrees of freedom (DoF). Furthermore, as robots with morphological symmetries feature $\mathbb{G}$-equivariant dynamics [9], the MDP's transition density T is guaranteed to be $\mathbb{G}$-invariant, complying with (2a). Moreover, the required $\mathbb{G}$-invariance of the reward function r (2c) is commonly satisfied as the reward depends only on relative distance/error measurements of the robot state, and other $\mathbb{G}$-invariant terms, such as fall/contact detection.

III. Related Works

Leveraging Symmetry in RL: Previous works on leveraging symmetry in locomotion control learning focus on leveraging morphological symmetries [9] and / or temporal symmetries [15] (i.e., periodicity of locomotion gait).

To encourage periodic gaits, [16], [17] introduce periodic phase signals in the state space $\mathcal{S}$, characterizing the gait cycle and / or each of the limb's gait cycle. Alternatively, [18] defines the MDP's action space $\mathcal{A}$ in terms of the parameters of existing Central Pattern Generators [19] describing periodic motion of the robot's limbs.

While these methods only explore temporal symmetry, in this work we present a method incorporating both temporal and morphological symmetries. Leveraging morphological symmetries to aid in the approximation of the optimal $\mathbb{G}$-equivariant control policy π^* (3) and $\mathbb{G}$-invariant value function V^{π^*} (4) can be achieved either through data-augmentation and soft / hard equivariance / invariance constraints on the models used to approximate these functions.

Inspired by data augmentation commonly used in supervised learning, augmenting RL algorithms with symmetric data involves constructing gradient from both the collected transitions and their symmetric transitions to induce equivariance / invariance for the policy and value function. This simple idea proves to be effective in locomotion tasks [8] and manipulation tasks [20].

Modifying loss functions to regulate equivariance in the policy and invariance in the value function provides additional gradient with symmetry information and poses *soft* symmetry constraints for the RL policy. Prior works show that an effective auxiliary objective for the policy network can be simply adding a regularization loss for equivariant actions [21]. However, this requires delicate hyperparameter tuning. Work in [22] adds a trust region on the regularization loss to improve training stability, and [23] adapts it to Soft Actor Critic algorithm and adds regularization on the variance of the action distribution.

Enforcing $\mathbb{G}$-equivariant policies and $\mathbb{G}$-invariant value functions involves modifying neural network model architecture [11], [24], [25], which introduces *hard* constraints on the symmetry of the learnt policy. In manipulation tasks, prior works [6], [26] have shown that such hard equivariance constraints lead to superior performance among other soft, approximately equivariant policies. However, the applicability of such equivariant RL algorithm and its performance on more dynamic robotics systems, such as legged robots remain underexplored. In this work, we study the effectiveness of data augmentation and hard-constraint RL algorithms on challenging locomotion and loco-manipulation tasks.

Loco-manipulation and Bipedal Locomotion: Recent works have focused on pushing the boundary of agility of legged robots beyond locomotion tasks. One trend is the demonstration of loco-manipulation tasks with legs being the end-effector to interact with objects, such as buttons on the wall [27], [28], soccer balls [29], [30], and move environment objects as simple tools [31]. However, as the task becomes more complex and requires highly dynamic maneuvers, it becomes exponentially more difficult for the policy to learn symmetrical behavior, resulting in suboptimal performance as seen in [30] and [29].

Another agile task commonly demonstrated is to perform bipedal locomotion with quadrupedal robots. Prior works demonstrate bipedal walking with assistive devices to lower the robustness requirements [32], or show single-task bipedal walking without capabilities to effectively turn and walk at various velocities [33]. A recent work shows a more developed controller, but again it shows non-negligible asymmetrical performance in agile motions, especially turning in various directions [34]. For humanoid robots, we again observe asymmetric behaviors on agile skills such as jumping to the side [35] when trained without equivariance constraints. In this work, we aim to investigate how adding symmetry constraints to the RL algorithms may affect the performance on such complex and agile bipedal locomotion tasks.

IV. Incorporating Symmetry in Model-free RL

In this section, we introduce two variations of Proximal Policy Optimization (PPO) [36] which distinctly leverage the robot's morphological symmetry, namely: PPO with data-augmentation in training (**PPOaug**) and PPO with hard equivariance/invariance constraints on neural networks (**PPOeqic**). We will compare these variants with a vanilla **PPO** implementation as a baseline. We opt out symmetric loss function, another method to augment the training process, as it has been consistently outperformed by PPOaug, as prior work [8] shows. To build the symmetry representations ρ_S and ρ_A, we leverage open-sourced repositories MorphoSymm [9] and ESCNN [37].

A. PPO with data-augmentation (PPOaug)

Considering the symmetries of the state and action spaces, and the invariance of the reward (2c), a direct method to

leverage symmetry is data-augmentation on policy and critic learning. Specifically, we perform updates on policy π_θ and critic V_ϕ, parameterized by θ and ϕ, with both the augmented transition tuples $(g \triangleright \boldsymbol{s}', g \triangleright \boldsymbol{a}, g \triangleright \boldsymbol{s}, r(g \triangleright \boldsymbol{s}, g \triangleright \boldsymbol{a}))$ and the online collected transition tuples $(\boldsymbol{s}', \boldsymbol{a}, \boldsymbol{s}, r(\boldsymbol{s}, \boldsymbol{a}))$.

Data augmentation in learning the critic biases the learned value function to be an approximately $\mathbb{G}$-invariant function. This, in turn, guides the optimization of the actor control policy, by constructing an invariant return estimate within the PPO's general advantage estimation. Following (2c), the augmentation process ensures that the values of the symmetric states converge towards identical target returns.

Since PPO is an on-policy algorithm, the fact that the augmented sample does not strictly come from the action distribution of the current policy creates an off-policy scenario. To minimize the effects of off-policy samples, we need to ensure that policy π_θ is approximately symmetric. To achieve this, we seek to minimize the difference between the gradient signals between the original samples and the augmented samples by network initialization and update rules. Specifically, we initialize the policy network with zero mean $\mu = 0$ and a small variance such that for the initialized policy network, $\pi_\theta(g \triangleright \boldsymbol{s}) \approx \pi_\theta(\boldsymbol{s})$ with a large enough standard deviation for samples in the first update. Then, unlike previous work [20] which stores the equivariant transitions in the rollout storage, we opt to perform equivariant augmentation after sampling a mini-batch of original data for each gradient update, such that the loss considers the original and augmented samples equally. Assuming that the policy π_θ before update is approximately equivariant, since the augmented samples provide gradient signals that push the policy towards an equivariant action distribution to a similar extent alone with the original one, we expect the updated policy to be approximately equivariant as well.

B. PPO with hard equivariance / invariance symmetry constraints (PPOeqic)

To enforce $\mathbb{G}$-equivariance constraints on the learned control policy π_θ and $\mathbb{G}$-invariance constraints on the learned value function V_ϕ, we parameterize these functions as equivariant / invariant neural networks. In this work, we design these networks as equivariant / invariant multi-layer-perceptron (EMLP) as a substitute for the unconstraint MLP. Specifically, EMLP network guarantees that $\pi_\theta(g \triangleright \boldsymbol{s}) = g \triangleright \pi_\theta(\boldsymbol{s})$. Regarding the invariant value function, we substitute the critic MLP with an invariant MLP, such that for an invariant NN, $V_\phi(g \triangleright \boldsymbol{s}) = V_\phi(\boldsymbol{s})$. This is done by swapping the last layer of the EMLP with an invariant transformation layer that transforms the regular representation of symmetry group $\mathbb{G}$ into the trivial representation, which is a scalar for the reflection group $\mathbb{C}_2$. We train the equivariant actor and invariant critic networks with vanilla PPO algorithm.

Besides morphological symmetry, we further consider temporal symmetry by assuming gait periodicity, leading to a valid equivariance transformation on the phase signal ψ [15], as in Sec. III. Specifically, for the phase signal ψ that varies from 0 to 1 within each period, the operation $g \triangleright \psi$ yields $(\psi + 0.5) - \lfloor \psi + 0.5 \rfloor$. A potential problem previous works have noted is that equivariant policies might struggle with neutral states [8], [17] where $g \triangleright \boldsymbol{s} = \boldsymbol{s}$ leading to $g \triangleright \boldsymbol{a} = \boldsymbol{a}$, for all $g \in \mathbb{G}$. By incorporating the additional phase ψ into the state of the MDP, we let $g \triangleright \boldsymbol{s} \neq \boldsymbol{s}$ when the robot is in its morphological neutral phase, thus avoiding the problem of neutral states for equivariant policy.

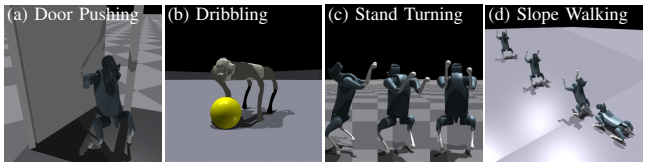

Fig. 2: Challenging tasks to test the efficacy of symmetry-incorporated policies. (a,b) showcase loco-manipulation with task-space symmetry and (c,d) exhibit bipedal locomotion with motion-level symmetry. These tasks push the boundary of agility and dynamic mobile manipulation for quadrupeds.

V. Experiments

We compare the PPO variants described in Sec. IV on four different tasks consisting of loco-manipulation and bipedal locomotion on quadrupedal robots. In loco-manipulation tasks, we investigate how incorporating symmetry influences task-level symmetry, while in bipedal locomotion, we focus on the intrinsic symmetry from the quadruped kinematic and dynamics structure. The observation consists of proprioceptive readings and task-specific observations. For tasks involving bipedal walking, we give dense rewards that encourage up-right orientation of the base and feet stepping periodically, adapted from [34]. All rewards are $\mathbb{G}$-invariant functions with relative measurements of the robot state. We include domain randomization for easing the sim-to-real transfer. All the tasks are trained in the Isaac Gym simulator [38]. Details of the environment setup are described below.

Door Pushing: In this loco-manipulation task (Fig. 2(a)), a robot needs to open a door using its front limbs while standing and adjust for doors that open either left or right. The task-specific observations include its initial relative position and orientation with respect to the door in the robot's frame, and the door's swing direction. The task-specific reward only includes the tracking error in forward velocity, encouraging the policy to learn pushing the door open to advance.

Dribbling: Shown in Fig. 2(b), this task requires a quadrupedal robot to dribble a soccer ball given its desired velocity commands. The ball observation is a relative position with respect to the robot base frame, and the rewards include ball velocity tracking and robot's proximity to the ball. We refer readers to [29] for more detailed settings.

Stand Turning: This agile bipedal locomotion task, adapted from [34], requires a quadrupedal robot to stand up on two feet and follow input commands, consisting of the desired linear velocity and the robot's desired heading, while the tracking errors are used as task rewards.

Slope Walking: This bipedal locomotion task pushes the limit of agility by requiring a quadruped to traverse an inclined flat surface on two feet.Training involves a terrain curriculum with a maximum slope of $11.3°$.

VI. RESULTS

In this section, we discuss the training and evaluation of PPO, PPOaug, and PPOeqic on the aforementioned tasks. We keep the training scheme the same across three methods, and we tune the hyperparameters separately for each task. For quantitative results, we evaluate on 2,000 simulation environments and 10 episodes per environment. We report the mean and variance across three different seeds.

A. Training Performance

For training performance, we focus on both training return (highest return in training) and sample efficiency (number of samples needed to reach the same return) as the metrics.

From Fig. 3, the PPOeqic policy consistently outperforms other PPO variants, showing the effectiveness of equivariance constraints in training. Furthermore, PPOeqic shows consistent improvements in sample efficiency, notably with fewer samples in early stage of training.

This indicates the equivariance constraints provide more efficient guidance in the policy's exploration than other variants. In contrast, PPOaug achieves higher training returns but similar sample efficiency across three out of four tasks compared to vanilla PPO baseline. Unlike PPOeqic, PPOaug needs to learn two aproximately equivariant action distributions, i.e. both $\pi_\theta(g \triangleright \boldsymbol{s})$ *and* $\pi_\theta(\boldsymbol{s})$, which is itself a more complex learning scheme. In comparison, vanilla PPO tends to converge to an inequivariant policy that still receives high training returns, such as leveraging an asymmetric gait.

Notice that these sub-optimal behaviors appeared on the PPO policy, though feasible due to imprecise contact modeling, are overfitted to the simulation environment and lack robustness, which are proved to be infeasible in sim-to-real transfer, as later described in Sec. VI-F.

B. Door Pushing task

We select the door pushing task as the benchmark for task-level symmetry in loco-manipulation, we evaluate the performance using two key metrics: success rate (SR) and the Symmetric Index (SI) [39]. SR quantifies the percentage of episodes where the robot is able to open a door at least 60 degrees wide and traverse it successfully, whereas SI assesses the disparity in success rates between left and right scenarios, calculated as $\text{SI} = \frac{2|X_R - X_L|}{X_R + X_L} \times 100\%$, where X_R is a scalar metric (e.g. the success rate) and X_L is its symmetric version.

1) Success Rates: As shown in Table I, PPOeqic policy consistently outperforms the unconstrained PPO policy with a 4.47% higher mean SR. PPOaug, in comparison, achieves a 6% lower mean SR. However, in addition to mean success rates, we are also interested in the *maximum* success rate each method may achieve across the three seeds. We notice that the difference between mean and maximum SRs for PPO policy is small, indicating that it consistently converges to similar sub-optima. In comparison, both symmetry-incorporated policies achieve a much higher maximum SR, around 88% for PPOeqic and 87% for PPOaug, which is 19% higher than that of PPO policy. This indicates that while there is no guarantee, PPOeqic and PPOaug are indeed more likely to achieve a better performance than PPO, given the optimal policy for the task is equivariant.

Comparing PPOaug and PPOeqic, we notice that while they achieve similar maximum SRs, the mean SR for PPOeqic is 10% higher than the one of PPOaug. This showcases the robustness in training of the hard-constraint equivariant policy compared to the soft-constraint policy.

2) Symmetry Index: Looking at the SI metric, we find that unconstrained PPO policy receives a high SI value, 4 times higher than the PPOaug policy and 8 times higher than the hard-constraint PPOeqic policy. This indicates that although PPO achieves a relatively high SR, it fails to learn a balanced policy between the two symmetric modes (left and right-handed doors). As previously mentioned, vanilla RL faces exploration difficulty in learning multi-modal distributions, such as the two modes in our case. In comparison, both PPOaug and PPOeqic policy achieve low SIs, indicating symmetric behaviors in the *task* space, even though the equivariance constraints are on the *joint* space.

3) Out-of-Distribution Scenarios: The policy's performance is further evaluated on its robustness and generalization to out-of-distribution (OOD) scenarios. Specifically, we randomize the robot's initial roll, pitch, and yaw angles in a wide range of $[-15°, 15°]$ that is unseen during training.

In out-of-distribution scenarios, similar to in-distribution results, PPOeqic outperforms the vanilla PPO policy in both mean and maximum SRs. In addition, PPOeqic again demonstrates more stable training performance than PPOaug policy, achieving a 5.9% higher mean SR across three seeds. This showcases the better generalization of an equivariant policy in response to out-of-distribution scenarios, given the distribution still follows a symmetric MDP setting.

4) Single Mode Training: An intriguing finding is that while PPOeqic has hard equivariance constraint, the policy trained only with one side scenario consistently outperforms the one trained on both sides. This trend is also present in PPOaug, even though the policy is not equivariance-enforced. We attribute this to the fact that during training, the domain randomization of joint properties, such as friction and PD gains, breaks the assumption of $\mathbb{G}$-invariant state transition probability that is necessary for an optimal policy to be equivariant. Consider a case where the joint properties are asymmetrically randomized, if the same environment is trained on only right-handed door, since the direction indicator is a part of the state $\boldsymbol{s}$, for any state transition $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$, we can mirror an equivariant transition $(g \triangleright \boldsymbol{s}, g \triangleright \boldsymbol{a}, g \triangleright \boldsymbol{s}')$ that satisfies equation 2a and does not conflict with the training data. Essentially, we can mirror a symmetric environment not seen in training but exists in the training distribution to form a symmetric MDP together with the original environment. However, if $(g \triangleright \boldsymbol{s}, g \triangleright \boldsymbol{a}, g \triangleright \boldsymbol{s}')$, the mirrored transition with the left-handed door already exist in the training data, equation 2a will not hold because of the asymmetric dynamics. Therefore, in contexts requiring dynamic asymmetry, such as domain randomization, it is beneficial to train on only one mode but deploy on other symmetric modes with equivariant policies. As we show in

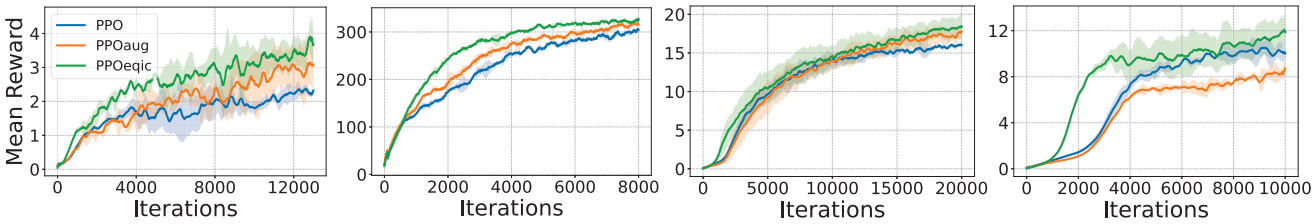

Fig. 3: Comparison of training curves of PPO, PPOaug, and PPOeqic on four tasks from left to right: Door Pushing, Dribbling, Stand Turning, and Slope Walking. Learning curves show mean episodic return and standard deviation for three seeds. PPOeqic consistently demonstrates the highest training returns and sample efficiency in all tasks.

Method		Mean SR (%)	Max SR (%)	RSI	OOD Mean SR (%)	OOD Max SR (%)	OOD RSI
PPO	trained on 1 side	43.40 ± 1.73	44.47	199.96	27.46 ± 2.34	30.04	199.99
	trained on 2 sides	61.18 ± 7.56	69.63	12.25	42.98 ± 2.21	45.51	3.45
PPOaug	trained on 1 side	54.39 ± 32.56	86.98	3.02	38.24 ± 19.17	52.85	1.40
	trained on 2 sides	50.24 ± 36.52	74.98	3.77	36.46 ± 25.98	53.78	**0.17**
PPOeqic	trained on 1 side	**65.65** ± 23.16	**87.96**	**0.98**	**44.15** ± 9.39	50.63	0.88
	trained on 2 sides	59.92 ± 17.39	74.74	1.65	38.56 ± 15.54	**55.95**	0.32

TABLE I: Comparison of success rates (SR) and their symmetry index on door-pushing tasks on training-distribution and out-of-distribution scenarios. Of the three variants, PPOeqic demonstrates both higher success rate and better symmetry index in both cases, indicating a better task-level symmetric policy.

Method	Error (rad)	Error RSI	CoT (Js/m)	CoT RSI
PPO	0.265 ± 0.022	0.0945	2223 ± 102	0.0259
PPOaug	0.259 ± 0.010	0.0212	2378 ± 101	**0.0046**
PPOeqic	**0.254** ± 0.014	**0.0207**	**2026** ± 187	0.0172

TABLE II: Comparison of command tracking error, Cost of Transport and their symmetry index on stand turning tasks for three PPO variants. PPOeqic demonstrates less error and energy consumption, indicating a more optimal policy.

Sec. VI-F, PPOeqic trained with only right-handed doors can be zero-shot deployed on doors that swing on either side.

C. Dribbling task

Consistent with previous loco-manipulation tasks, incorporating symmetry improves soccer dribbling gaits. Qualitatively, PPOeqic and PPOaug keep the ball closer to the robot, while PPO often kicks it away and chases it. Quantitatively, the hard-constraint PPOeqic achieves an average episodic return of 431.86 ± 1.77, slightly outperforming PPO (427.02 ± 2.15) and PPOaug (430.49 ± 3.45), suggesting all methods are near saturation in the simulation environment.

D. Stand Turning task

Besides loco-manipulation, we examine how symmetry helps if the task involves bipedal locomotion only. Qualitatively, the PPO policy repeatedly develops a staggered gait with one foot positioned ahead of the other to maintain balance. This gait results in significant jittering and hinders symmetric behaviors between left and right turns, as shown in Fig. 1. Furthermore, even though the reward function explicitly encourages symmetry, PPO fails to transfer on hardware unless the reward terms are carefully tuned to generate a relatively symmetric gait. Conversely, symmetry-incorporated policies exhibit both symmetric gaits and more importantly symmetric turning trajectories in the task level (Fig. 1).

To measure the optimality of the policies, we gauge the error between the robot's commanded and actual headings. Additionally, we include the Cost of Transport (CoT) to evaluate the energy efficiency of the controller. Here, we calculate CoT as $\frac{\sum_t \sum_{12 \text{ actuators}} [\tau \dot{\theta}]^+}{\sum_t |v|}$, where τ is joint torque, $\dot{\theta}$ is joint angular velocity, and v is the horizontal velocity of the robot base at each simulation timestep. We observe that PPOeqic achieves both lower tracking error and CoT compared to other variants. This indicates the relative optimality induced by the hard equivariance constraint. In comparison, PPOaug achieves lower tracking error and error SI but higher CoT than PPO, indicating that the behavior is similarly symmetric but less optimal compared to PPOeqic.

E. Slope Walking Task

In the challenging task of slope walking, we see significant differences between the three PPO variants. The PPO policy learns a more natural gait compared to other tasks. However, it still shows significant issues with learning a symmetric foot placement, resulting in backward steps and frequent balance loss. This is evident from the plot of the feet positions in the direction of desired motion, shown in Fig. 4(a), where PPO shows undesired behaviors including stepping in the same place, stepping backwards, or slipping on the ground. Additionally, the locomotion gaits are devoid of periodic stepping in some instances. This unregulated gait pattern leads to significant decrement in velocity tracking performance, where the policy walks only half as far as PPOeqic in the same timeframe.

The PPOaug policy improves upon PPO's limitations, exhibiting better alternation between the leading foot and more regulated foot placement. However, as shown in Fig. 4(b) variations in step size and occasional staggering of one leg peSIst, indicating room for improvement.

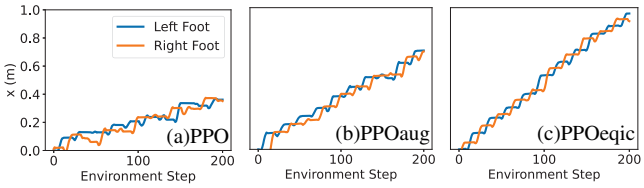

Fig. 4: Plots of the feet positions in the desired walk direction. We observe that vanilla PPO learns an unstable step pattern with backward steps and foot slipping, resulting in 50% slower walking speed. PPOaug improves drastically but asymmetric patterns such as foot dragging still exists. PPOeqic presents the most symmetric interweaving gait pattern and walks at the desired speed.

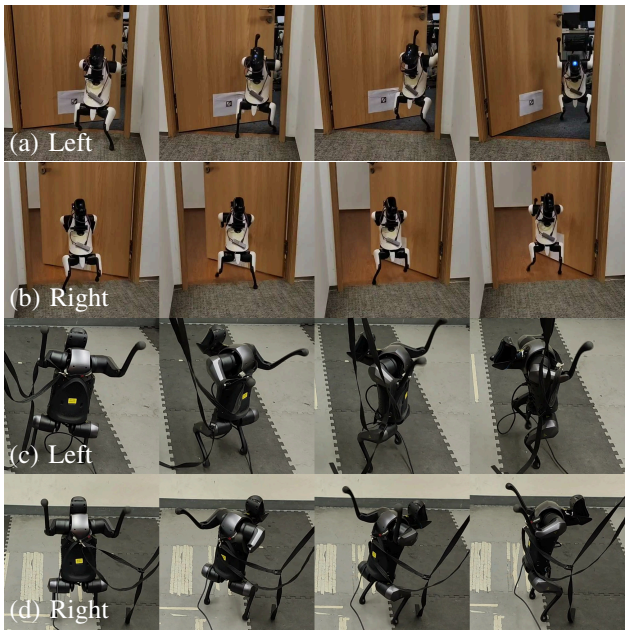

Fig. 5: Snapshots of equivariant policies deployed zero-shot to perform real-world tasks of door pushing (a,b) and stand turning (c,d). For task-level symmetry (e.g. door pushing), PPOeqic trained with only right-handed door can be deployed on both left and right-handed doors with symmetric gait patterns. For motion symmetry, PPOaug is more robust against slightly asymmetric dynamics that exists on actual hardware.

PPOeqic presents the most stable gait. Shown in Fig. 4(c), it maintains consistent foot exchange, regulated contact sequences, and similar step sizes, even on a $11.3°$ incline. This demonstrates PPOeqic's effectiveness in learning challenging tasks, highlighting its optimal performance through the hard equivariance constraints compared to other methods. Quantitatively, it is able to walk at the desired velocity of $0.25\ m/s$, 20% more accurately than PPOaug.

F. Real-world Experiments

In this section, we explore the sim-to-real transfer of these methods through real-world experiments on two selected tasks, thereby evaluating their efficacy beyond simulation.

1) Door Pushing Task: Although vanilla PPO achieves high success rates in simulation, its real-world performance is limited. The asymmetric gaits learned by PPO often cause the robot to fall, likely due to overfitting to simulated contact dynamics that differ from the real world. In contrast, PPOeqic is more robust: even though trained exclusively on right-handed doors, it is able to locomote steadily and open both left and right-handed doors. On the other hand, PPOaug tends to rely on one front leg, showing limited strength in pushing doors. This highlights the importance of equivariance constraints for optimal real-world behavior.

2) Stand Turning Task: Similar to the door pushing task, the vanilla PPO policy is able to locomote without falling over in simulation but fails in real-world contexts. In comparison, both PPOaug and PPOeqic policies demonstrate good zero-shot transferability, executing 90-degree turns to both sides with high consistency. Notably, the PPOaug policy exhibits significantly higher robustness and stability, successfully completing the task in 9 out of 10 trials. This again highlights the substantial enhancement in sim-to-real transfer by incorporating symmetry into the learning process.

3) Discussion: The intricacies of real-world applications often reveal that environments are not perfectly symmetric MDPs. Discrepancies such as uneven robot mass distribution and variance in actuator dynamics introduce inherent asymmetries in the robot's dynamics. Our findings suggest that in loco-manipulation tasks, where task-space symmetry dominates, the robot's intrinsic asymmetries are negligible for the overall MDP, allowing hard symmetry constraints (e.g., PPOeqic) to excel in real-world scenarios. However, in bipedal walking tasks which consist of almost entirely intrinsic symmetry, we observed that PPOeqic is more vulnerable to distribution shifts compared to methods without hard constraints (e.g., PPOaug). By allowing natural adaptation to the robot's asymmetries during training, PPOaug demonstrates enhanced robustness against such imperfect symmetry. This aligns with prior work [8], suggesting that the choice between PPOaug and PPOeqic should be tailored to the specific nuances of the problem.

VII. Conclusion

In this work, we investigate the benefits of leveraging symmetry in model-free RL for legged locomotion. We compare the performance of incorporating data augmentation and hard equivariance constraints across four challenging bipedal locomotion and loco-manipulation tasks against a vanilla PPO baseline. We find that imposing hard symmetry constraints on the learned policy and value functions leads to better performance than other methods. Compared to vanilla PPO, equivariant policies learn notably more steady and symmetric gait patterns, eventually leading to better task-space symmetry. More importantly, equivariant policies trained on single symmetry mode are directly generalizable to other modes. When applied to real-world scenarios, symmetry-incorporated policies demonstrate significantly better robustness than unconstrained policy. Furthermore, PPOaug copes with slight asymmetry in robot's own dynamics, while PPOeqic demonstrates better performance on task-space symmetry.

We hope this work can aspire the exploration of leveraging symmetry constraints on robots with larger symmetry groups than the reflection group concerned in this work. As the number of symmetry modalities increases, the symmetry constraints are expected to play a more crucial role in guiding

the exploration of model-free RL over the increasingly complex state, action spaces. In addition, equivariant policies demonstrate promising potential for even larger performance gains over vanilla PPO, highlighting improvements as large as 26% in some seeds. Future efforts could be on stabilizing training to consolidate this enhancement and develop better symmetry-incorporated RL algorithms.

ACKNOWLEDGEMENTS

X. H., Z. L., Q. L., and K. S. acknowledge financial support from The AI Institute, InnoHK of the Government of the Hong Kong Special Administrative Region via the Hong Kong Centre for Logistics Robotics. G. T., M. P., and C. S. acknowledge financial support from PNRR MUR Project PE000013 "Future Artificial Intelligence Research (hereafter FAIR)", funded by the European Union – NextGenerationEU. The authors thank Prof. Xue Bin Peng for insightful discussions on this work. The authors also thank Xiaomi Inc. for providing CyberDog 2 for experiments.

REFERENCES

[1] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," in *Robotics: Science and Systems*, 2020.

[2] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control," *arXiv preprint arXiv:2401.16889*, 2024.

[3] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *International Conference on Learning Representations*, 2019.

[4] A. Singh, L. Yang, C. Finn, and S. Levine, "End-to-end robotic reinforcement learning without reward engineering," *Robotics: Science and Systems XV*, 2019.

[5] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov, "Efficient exploration via state marginal matching," *arXiv preprint arXiv:1906.05274*, 2019.

[6] D. Wang, M. Jia, X. Zhu, R. Walters, and R. Platt, "On-robot learning with equivariant models," in *6th Annual Conference on Robot Learning*, 2022.

[7] M. Zinkevich and T. Balch, "Symmetry in markov decision processes and its implications for single agent and multi agent learning." Citeseer, 2001.

[8] M. Mittal, N. Rudin, V. Klemm, A. Allshire, and M. Hutter, "Symmetry considerations for learning task symmetric robot policies," *arXiv preprint arXiv:2403.04359*, 2024.

[9] D. Ordoñez-Apraez, G. Turrisi, V. Kostic, M. Martin, A. Agudo, F. Moreno-Noguer, M. Pontil, C. Semini, and C. Mastalli, "Morphological symmetries in robotics," *The International Journal of Robotics Research*, 2024.

[10] H. Huang, D. Wang, R. Walters, and R. Platt, "Equivariant transporter network," in *Proceedings of Robotics: Science and Systems*, 2022.

[11] E. Van der Pol, D. Worrall, H. van Hoof, F. Oliehoek, and M. Welling, "Mdp homomorphic networks: Group symmetries in reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4199–4210, 2020.

[12] Xiaomi, "Cyberdog2," https://www.mi.com/cyberdog2, 2024, accessed: Mar. 2024.

[13] Unitree Robotics, Go1, 2024, https://www.unitree.com/go1, [Online; accessed Mar. 2024].

[14] D. Ordoñez-Apraez, M. Martin, A. Agudo, and F. Moreno-Noguer, "On discrete symmetries of robotics systems: A group-theoretic and data-driven analysis," in *Robotics Science and System (RSS)*, 2023.

[15] J. Ding and Z. Gan, "Breaking symmetries leads to diverse quadrupedal gaits," *IEEE Robotics Autom. Lett.*, 2024.

[16] L. Liu, M. V. D. Panne, and K. Yin, "Guided learning of control graphs for physics-based characters," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, 2016.

[17] F. Abdolhosseini, H. Y. Ling, Z. Xie, X. B. Peng, and M. Van de Panne, "On learning symmetric locomotion," in *Motion, Interaction and Games*, 2019, pp. 1–10.

[18] 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.

[19] G. Bellegarda and A. Ijspeert, "Cpg-rl: Learning central pattern generators for quadruped locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, 2022.

[20] Y. Lin, J. Huang, M. Zimmer, Y. Guan, J. Rojas, and P. Weng, "Invariant transform experience replay: Data augmentation for deep reinforcement learning," *IEEE Robotics and Automation Letters*, 2020.

[21] W. Yu, G. Turk, and C. K. Liu, "Learning symmetric and low-energy locomotion," *ACM Transactions on Graphics (TOG)*, 2018.

[22] M. Kasaei, M. Abreu, N. Lau, A. Pereira, and L. P. Reis, "A cpg-based agile and versatile locomotion framework using proximal symmetry loss," *arXiv preprint arXiv:2103.00928*, 2021.

[23] D. Ordonez-Apraez, A. Agudo, F. Moreno-Noguer, and M. Martin, "An adaptable approach to learn realistic legged locomotion without examples," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4671–4678.

[24] A. K. Mondal, P. Nair, and K. Siddiqi, "Group equivariant deep reinforcement learning," *arXiv preprint arXiv:2007.03437*, 2020.

[25] S. Rezaei-Shoshtari, R. Zhao, P. Panangaden, D. Meger, and D. Precup, "Continuous mdp homomorphisms and homomorphic policy gradient," *Advances in Neural Information Processing Systems*, 2022.

[26] D. Wang, R. Walters, and R. Platt, "SO(2)-equivariant reinforcement learning," in *International Conference on Learning Representations*, 2022.

[27] X. Cheng, A. Kumar, and D. Pathak, "Legs as manipulator: Pushing quadrupedal agility beyond locomotion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[28] E. Vollenweider, M. Bjelonic, V. Klemm, N. Rudin, J. Lee, and M. Hutter, "Advanced skills through multiple adversarial motion priors in reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5120–5126.

[29] Y. Ji, G. B. Margolis, and P. Agrawal, "Dribblebot: Dynamic legged manipulation in the wild," *International Conference on Robotics and Automation*, 2023.

[30] X. Huang, Z. Li, Y. Xiang, Y. Ni, Y. Chi, Y. Li, L. Yang, X. B. Peng, and K. Sreenath, "Creating a dynamic quadrupedal robotic goalkeeper with reinforcement learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[31] M. Xu, W. Yu, P. Huang, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use with large language models," in *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023.

[32] C. Yu and A. Rosendo, "Multi-modal legged locomotion framework with automated residual reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 312–10 319, 2022.

[33] L. M. Smith, J. C. Kew, T. Li, L. Luu, X. B. Peng, S. Ha, J. Tan, and S. Levine, "Learning and adapting agile locomotion skills by transferring experience," in *Robotics: Science and Systems*, 2023.

[34] Y. Li, J. Li, W. Fu, and Y. Wu, "Learning agile bipedal motions on a quadrupedal robot," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[35] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through reinforcement learning," in *Robotics science and systems*, 2023.

[36] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[37] G. Cesa, L. Lang, and M. Weiler, "A program to build e (n)-equivariant steerable cnns," in *International conference on learning representations*, 2021.

[38] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu based physics simulation for robot learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[39] R. Robinson, W. Herzog, and B. M. Nigg, "Use of force platform variables to quantify the effects of chiropractic manipulation on gait symmetry." *Journal of manipulative and physiological therapeutics*, 1987.