WARL: Wrench-Augmented Reinforcement Learning for Task-Agnostic Learning in Legged Robots

Keita Yoneda[1], Kento Kawaharazuka[1,2], Kei Okada[1]

***Abstract*— While reinforcement learning for legged robots has achieved high motor performance, it has been constrained by the limited exploration capability of actions confined to the joint space. To address this issue, this study proposes a new method, Wrench-Augmented Reinforcement Learning (WARL), which introduces a wrenche (force and torque) into the action space. The proposed method combines wrench-guided exploration with a success rate-based curriculum mechanism to expand exploration capabilities in the early stages of learning, with the ultimate goal of acquiring behaviors based solely on joint control. Experiments using a quadruped robot demonstrated that WARL can learn robustly across diverse terrains and motor tasks without requiring terrain-specific reward adjustments or complex curriculum designs. Furthermore, an ablation study verified the effectiveness of the Switching Curriculum, which gradually eliminates the wrench. On the other hand, we also show that introducing a wrench can encourage behaviors that do not sufficiently exploit the robot's physical embodiment. These findings suggest that while wrench-based exploration enhancement is effective for improving learning efficiency, designing it in a way that is consistent with the robot's physical structure is a critical future challenge.**

I. Introduction

Legged robots have garnered attention as highly mobile platforms capable of traversing diverse environments. In recent years, reinforcement learning has become a standard method for achieving dynamic motions such as walking on uneven terrain and jumping [1]–[3]. However, conventional frameworks often involve numerous heuristics in the learning environment and reward design for specific tasks, posing a challenge due to the high design cost for new tasks.

For example, in tasks that involve traversing uneven terrain with large steps, a standard approach is to use a heuristic curriculum design that gradually increases the height of the steps as learning progresses [4]. Even when not using curriculum-based changes in terrain, task-specific, meticulous reward design, such as reward shaping that incorporates fine-grained behaviors for learning jumping motions, is often necessary [5]. While these methods have been successful in learning specific tasks, they lack the versatility to be applied to a wide range of tasks.

In this study, we hypothesize that a primary reason for the need for task-specific design is the difficulty of exploration, which stems from the complex relationship between the action space and the task space. Here, the task space refers to the space that characterizes the motion of the legged robot, which, for locomotion tasks, includes the robot's center of mass position and orientation. In conventional reinforcement learning for legged robots, the action space is typically the joint space, such as joint angles or joint torques [6], [7]. In general, the mapping from the joint space to the task space is complex because it is strongly influenced by discrete and nonlinear factors such as contact states and joint configurations. This complexity can lead to slow learning speeds and convergence to local optima. Previous research has attempted to address this problem through modifications in environment and reward design [8]–[10].

However, a more direct solution is to reduce the mismatch between the action and task spaces. For example, in robot arm control, it has been reported that using the end-effector position as the action improves exploration efficiency [11]–[13]. For legged robots, methods have been proposed that use the foot-end space as the action [14], hierarchical methods that combine model-based control with high-level actions such as foot-end position and gait phase [15], [16], and methods that introduce prior structures such as periodic trajectories [17]. Although these methods contribute to improving learning efficiency, there are limits to the improvement

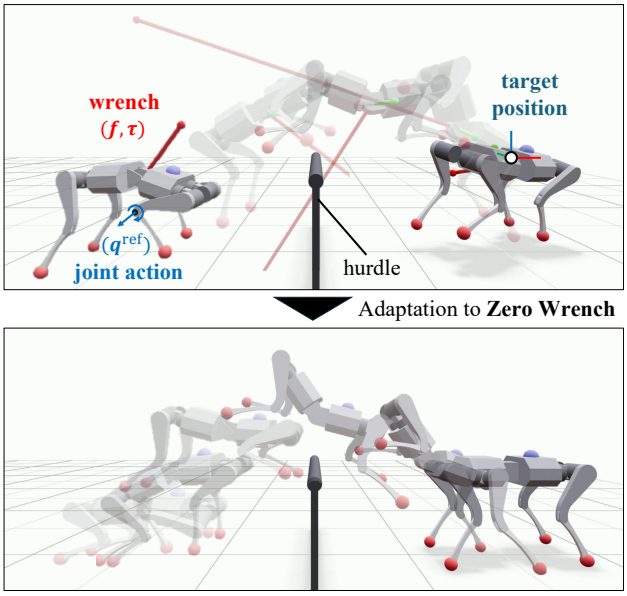

Fig. 1. Overview of Wrench-Augmented Reinforcement Learning (WARL). By introducing a wrench into the action space, exploration is promoted, and a policy that does not use the wrench is ultimately acquired through a curriculum learning process.

[1] The authors are with the Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan. [yoneda, kawaharazuka, k-okada]@jsk.imi.i.u-tokyo.ac.jp

[2] The author is with the AI Center, Graduate School of Information Science and Technology, The University of Tokyo, Japan.

in exploration performance because nonlinearity still exists between the foot-end space and the robot's body position and orientation space. Furthermore, design constraints for real-world deployment often limit the choice of action parameterization.

One of the essential features of legged robot locomotion is the change in the robot's position and orientation. By incorporating a wrench ($w = (f, \tau) \in \mathbb{R}^6$), which directly acts on these quantities, into the action space, the relationship to the task space can be simplified, thereby improving exploration. Previous studies have introduced wrenches into motor learning, including EFGCL, which applies external forces to facilitate backflip, barrel roll, and jumping motions [18], ZEST, which uses model-based wrench generation for reference trajectory tracking [19], and A2CF, which outputs wrenches to accelerate learning [20]. However, EFGCL relies on manually designed external forces requiring heuristic intuition, ZEST does not use wrenches for exploration, and A2CF employs them mainly as auxiliary assistance rather than as exploratory actions. Moreover, evaluations have been limited to tasks solvable without wrench assistance, leaving the effectiveness of wrench-guided exploration insufficiently validated.

To address these limitations, we propose a method that introduces a wrench as a primary exploratory action and gradually removes its influence through a curriculum learning scheme. During early training, behavior is generated mainly through the wrench, whose contribution is progressively reduced as learning proceeds. By directly applying policy-generated wrenches to the robot's torso, the proposed method promotes efficient exploration in locomotion tasks while ultimately learning an equivalent joint-only control policy.

The main contributions of this study are as follows:

1) We show that by introducing a wrench into the action space in the motor learning of legged robots, exploration can be made more efficient, and versatile learning that does not heavily depend on task-specific environment or reward design becomes possible.
2) We propose a curriculum learning method for acquiring a policy that can be applied to real robots by gradually removing the wrench introduced for exploration guidance.
3) We analyze that there is a trade-off between the improvement in spatial exploration ability by this method and the difficulty in acquiring dexterous motions that utilize the environment, such as footholds. Resolving this trade-off is a future challenge.

II. Background

A. Difficulty of Exploration in Reinforcement Learning for Legged Robots

Reinforcement learning is a framework for optimizing a policy through interaction with an environment, and is typically formulated as a Markov Decision Process (MDP). An MDP consists of a state s, an action a, a reward r, and a transition distribution, and the agent learns a policy $\pi(a|s)$ that maximizes the expected cumulative reward. In motor learning for legged robots, the state includes the robot's posture, velocity, and contact information, and the action is output as joint torques or joint angles.

Proximal Policy Optimization (PPO) [21] is widely used as an algorithm for reinforcement learning in legged robots. PPO is an algorithm that ensures learning stability by constraining the policy update amount, and it is also easy to combine with curriculum learning. Therefore, this study also adopts PPO as the baseline algorithm. In PPO, which assumes a stochastic policy, exploration is performed by sampling actions based on a Gaussian distribution:

$$a_t \sim \pi_\theta(a_t|s_t) = \mathcal{N}(\mu_\theta(s_t), \Sigma_\theta(s_t)) \tag{1}$$

Since this randomness-based exploration is performed in the action space, its design greatly affects exploration efficiency. Conventional joint-space actions are typically of the form where the target joint angles are output by a scale transformation as follows, and then converted to joint torques by PD control.

$$\boldsymbol{q}^{\text{ref}} = \boldsymbol{q}^{\text{range}} \odot \text{clip}(\lambda^{\text{joint}} \boldsymbol{a}^{\text{joint}}, -1, 1) + \boldsymbol{q}^{\text{offset}} \tag{2}$$

Here, $\boldsymbol{a}$ is the action output by the policy, which is converted to the target joint angle $\boldsymbol{q}^{\text{ref}}$ using a scaling factor λ^{joint}, an action range $\boldsymbol{q}^{\text{range}}$, and an offset $\boldsymbol{q}^{\text{offset}}$.

B. Difference in Spatial Exploration Capability due to Action Space Design

We investigate the effect of action space design on spatial exploration capability by comparing a conventional joint-space action with an action space that includes a wrench. Using the quadruped robot KLEIYN [10] in Isaac Gym, we compare the travel distance when Gaussian noise is directly injected into the action outputs before scaling. The joint angle action is scaled to a joint angle in the form of Eq. 2 and then converted to a joint torque by PD control. The wrench action is calculated directly by scale transformation and clipping as follows.

$$\boldsymbol{w} = \boldsymbol{w}^{\max} \cdot \text{clip}(\lambda^{\text{wrench}} \boldsymbol{a}^{\text{wrench}}, -1, 1) \tag{3}$$

$$\boldsymbol{w}^{\max} = (\boldsymbol{f}^{\max} \in \mathbb{R}^3, \boldsymbol{\tau}^{\max} \in \mathbb{R}^3) \in \mathbb{R}^6 \tag{4}$$

Here, $\boldsymbol{a}^{\text{wrench}}$ is the wrench action, which is converted to the final wrench $\boldsymbol{w}$ using the scaling factor λ^{wrench} and the maximum wrench value $\boldsymbol{w}^{\max}$. Here, we set four types of wrench scales $(f^{\max}[N], \tau^{\max}[Nm]) = \{(0,0), (100,10), (500,100), (1000,100)\}$ and compare the exploration range under each condition. The action scales were set to $\lambda^{\text{joint}} = 0.2, \lambda^{\text{wrench}} = 0.5$, and the joint angle and wrench actions $(\boldsymbol{a}^{\text{joint}}, \boldsymbol{a}^{\text{wrench}})$ were independently sampled according to a standard normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Figure 2 shows the trajectories when this random action was applied to 100 robots for 10 seconds.

In the case of only joint angles with the wrench disabled ($0\ N, 0\ Nm$) and when the wrench scale was small ($100\ N, 10\ Nm$), the robots could only move within a

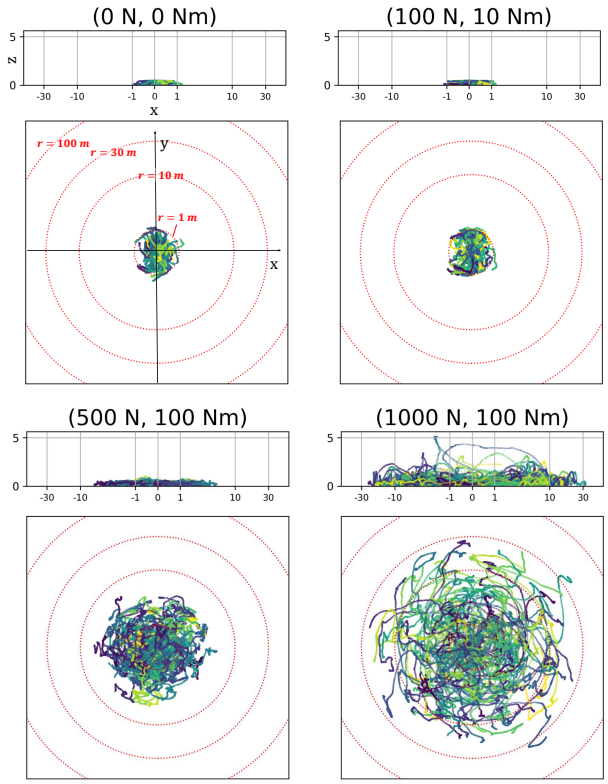

Fig. 2. Comparison of exploration behavior with different wrench scales. The larger the wrench scale, the more direct exploration in the robot's position space becomes possible, and the spatial exploration capability is dramatically improved.

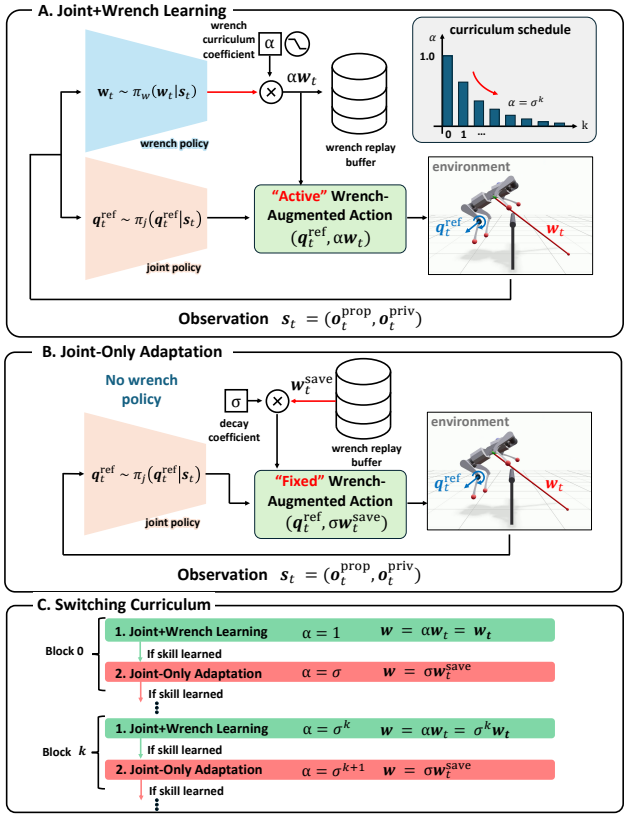

Fig. 3. Learning flow of WARL. Joint+Wrench learning promotes exploration and stores successful wrench sequences. Joint-only adaptation reduces wrench dependence using a decayed coefficient α.

range of about 1m in 10 seconds. In contrast, with a large wrench scale (500 N, 100 Nm), trajectories of more than 10m in 10 seconds were observed, and with an even larger scale (1000 N, 100 Nm), vertical displacement also increased substantially.

These results indicate that by including a wrench in the action space, it becomes possible to explore directly in the robot's position space with only Gaussian noise, and the spatial exploration capability is dramatically improved.

C. Gradual Attenuation of Wrench using Curriculum Learning

Although introducing a wrench improves exploration efficiency, it is not possible to directly apply a wrench to the robot's body in reality. Therefore, previous studies [19], [20] have proposed curriculum learning that gradually attenuates the wrench output based on indicators such as success rate and the magnitude of the auxiliary wrench.

In this study, we follow the framework of previous studies that gradually attenuate the wrench and propose a curriculum learning method that uses only the success rate as an indicator.

III. Method

A. Overview of WARL

We propose Wrench-Augmented Reinforcement Learning (WARL), a framework that exploits wrench-based exploration while ultimately acquiring a joint-only control policy. As illustrated in Fig. 3, WARL alternates between two phases: (i) a Joint+Wrench learning phase, where both a wrench policy π_w and a joint policy π_j are optimized, and (ii) a Joint-only adaptation phase, where the motion is reproduced by π_j under attenuated previously saved wrenches. This Switching Curriculum enables efficient spatial exploration in early learning and gradually removes the dependency on the wrench through a decaying curriculum coefficient α. To make the RL problem tractable, we adopt a Teacher-Student framework [17], where privileged observations are used during training and removed through a student distillation process.

B. Action Space Design

The wrench policy π_w outputs a 7-dimensional vector:

$$(\boldsymbol{a}^{\text{force}} \in \mathbb{R}^3,\ \boldsymbol{a}^{\text{torque}} \in \mathbb{R}^3,\ a^{\text{scale}} \in \mathbb{R}) = \pi_w(\boldsymbol{s}). \tag{5}$$

The wrench components are computed as

$$\lambda^{\text{scale}} = \text{sigmoid}(a^{\text{scale}}), \tag{6}$$

$$\boldsymbol{f} = f^{\max} \cdot \text{clip}(\lambda^{\text{wrench}} \boldsymbol{a}^{\text{force}}, -1, 1) \cdot \lambda^{\text{scale}} \cdot \alpha, \tag{7}$$

$$\boldsymbol{\tau} = \tau^{\max} \cdot \text{clip}(\lambda^{\text{wrench}} \boldsymbol{a}^{\text{torque}}, -1, 1) \cdot \lambda^{\text{scale}} \cdot \alpha. \tag{8}$$

Here, λ^{wrench} is a fixed scaling factor and α is the curriculum coefficient. The resulting wrench is transformed to the global frame using the torso orientation $R \in \text{SO}(3)$ and applied to the torso link.

Algorithm 1 Wrench-Augmented Reinforcement Learning (WARL)

Initialize π_w, π_j, $\alpha = 1$
while $\alpha > \epsilon$ **do**
  Joint+Wrench Learning
  while $\zeta < \gamma$ or $|\Delta\zeta| > \delta$ **do**
    $\boldsymbol{q}_t^{\text{ref}} = \pi_j(\boldsymbol{s}_t)$
    $\boldsymbol{w}_t = \pi_w(\boldsymbol{s}_t)$
    $\boldsymbol{s}_{t+1}, r_t = \text{Step}(\boldsymbol{q}_t^{\text{ref}}, \alpha\boldsymbol{w}_t)$
    Update π_j, π_w
    Store wrench sequences from successful episodes
  end while
  Freeze π_w and set $\alpha \leftarrow \sigma\alpha$
  Joint-only Adaptation
  while $\zeta < \gamma$ or $|\Delta\zeta| > \delta$ **do**
    $\boldsymbol{q}_t^{\text{ref}} = \pi_j(\boldsymbol{s}_t)$
    $\boldsymbol{w}_t = \mathcal{W}_{\text{saved}}(t)$
    $\boldsymbol{s}_{t+1}, r_t = \text{Step}(\boldsymbol{q}_t^{\text{ref}}, \sigma\boldsymbol{w}_t)$
    Update π_j
  end while
end while
return π_j

C. Switching Curriculum

Optimizing joint control and wrench-guided exploration simultaneously can cause training instability. To address this, we propose a Switching Curriculum that alternates between the following two phases.

Joint+Wrench Phase: Both π_w and π_j are optimized under the scaled wrench $\alpha\boldsymbol{w}_t$. Successful wrench sequences are stored for subsequent adaptation. This phase primarily enhances spatial exploration.

Joint-only Adaptation Phase: The joint policy π_j is trained using attenuated saved wrenches $\sigma\mathcal{W}_{\text{saved}}$ to encourage reproduction of the motion with reduced wrench assistance.

Each phase terminates when the success rate ζ exceeds γ and its change $|\Delta\zeta|$ falls below δ. After adaptation, α is updated as $\alpha \leftarrow \sigma\alpha$. This process continues until $\alpha < \epsilon$, yielding a joint-only policy.

D. Saving Wrench Sequences

For each successful episode i, we define the 6D wrench sequence $\mathcal{W}^i$, its norm $|\mathcal{W}^i|$, and the sum of two sequences $\mathcal{W}^A + \mathcal{W}^B$ as

$$\mathcal{W}^i = \{\boldsymbol{w}_t^i\}_{t=0}^{T_i} = \{(\boldsymbol{f}_t^i, \boldsymbol{\tau}_t^i)\}_{t=0}^{T_i}. \tag{9}$$

$$|\mathcal{W}^i| = \sum_{t=0}^{T_i} \sqrt{\|\boldsymbol{f}_t^i\|^2 + \kappa\|\boldsymbol{\tau}_t^i\|^2}. \tag{10}$$

$$\mathcal{W}^A + \mathcal{W}^B = \{\boldsymbol{w}_t^A + \boldsymbol{w}_t^B\}_{t=0}^{T}. \tag{11}$$

κ is a weighting factor to balance force and torque magnitudes. When the buffer accumulates N^{succ} sequences, the median sequence is selected and the saved wrench is updated:

$$\mathcal{W}_{\text{saved}}^j = \eta\mathcal{W}_{\text{saved}}^{j-1} + (1-\eta)\mathcal{W}^{i^*}. \tag{12}$$

This median-based weighted update suppresses extreme cases and stabilizes adaptation.

E. Teacher-Student Learning

Learning legged locomotion using only proprioceptive inputs can be unstable due to the high-dimensional state space and sparse rewards. To address this, we adopt a teacher-student framework [17] that stabilizes training by leveraging privileged observations.

IV. Learning Setup

A. Robot Models

To verify the versatility of our method, we use two quadruped robot models with different sizes and joint configurations: KLEIYN (total length 750 mm, weight 18 kg, max joint torque 24 Nm, max joint speed 480 rpm) and Unitree Go2 (total length 700 mm, weight 15 kg, max joint torque 24 Nm, max joint speed 230 rpm). We select Isaac Gym [22] as our simulation environment. Specifically, we apply wrenches to the robot's torso link using the `apply_rigid_body_force_tensor` method, and control its joints via torque commands using the `set_dof_actuation_force_tensor` method.

B. Task Settings

To verify the effect of exploration guidance by wrench, we set up two types of tasks: goal-reaching tasks and dynamic tasks.

a) Goal-reaching tasks: These consist of four tasks: *HurdleJump*, *LedgeJump*, *GapLeap*, and *IslandTraverse*, all of which require the robot to reach a specified target position $\boldsymbol{p}^*$ and target orientation $\boldsymbol{q}^*$. Specifically, in *HurdleJump*, a 0.5 m obstacle blocks the path to a goal 2 m away. In *LedgeJump*, the goal is located on a 0.5 m high ledge 2 m away. In *GapLeap*, the robot must clear a 1 m wide gap to reach a destination 2 m away. In *IslandTraverse*, the path to a destination 2.4 m away includes a 1.4 m gap containing two 0.2 m square footholds. The objective of this learning is to acquire motions to overcome these diverse terrains without relying on reward tuning or environmental curricula. Therefore, a common reward is used for all tasks, and the terrain is also fixed with a fixed height of steps and size of gaps for learning.

To also evaluate the ability to remain stationary, with a probability of 10% the target position is set to the robot's initial position, effectively requiring the robot to stay in place.

The success rate was calculated based on the criterion that an episode is considered successful if the distance from the target position $|\boldsymbol{p} - \boldsymbol{p}^*|$ is within 0.2 m and the error from the target orientation $|\boldsymbol{\theta}|$ is within 45°.

TABLE I

NOTATION USED IN OBSERVATION AND REWARD FORMULATION

Symbol	Description
$\psi(x, \lambda)$	Gaussian shaping kernel : $\exp(-x^2/\lambda^2)$
w_{cur}	Curriculum weight at iteration i : $0.01^{0.9999^i}$
N_{contact}	Number of penalized contacts
$\boldsymbol{f}$	External force component of wrench
$\boldsymbol{\tau}$	External torque component of wrench
$\boldsymbol{\omega}$	Base angular velocity
$\boldsymbol{v}$	Base linear velocity
$\boldsymbol{g}$	Projected gravity vector in base frame
g_x, g_y	x, y components of $\boldsymbol{g}$
$\boldsymbol{p}$	Robot base position
$\boldsymbol{p}^*$	Target position
$\Delta\boldsymbol{p} = \boldsymbol{p}^* - \boldsymbol{p}$	Position error
$\boldsymbol{q}$	Joint configuration
$\boldsymbol{q}_{\text{def}}$	Default joint configuration
$\dot{\boldsymbol{q}}$	Joint velocity
$\boldsymbol{e}_{\text{dof}}$	$\boldsymbol{q} - \boldsymbol{q}_{\text{def}}$
$\boldsymbol{\theta}$	Orientation error (axis-angle form)
θ	Target rotation angle (pitch/roll)
z	Base height
z^*	Standing height of the robot
ϕ	Phase variable : $t^3/(1+t^3)$
$\boldsymbol{h}$	Measured terrain heights $\in \mathbb{R}^{187}$
$\boldsymbol{\omega}_{xz}$	Angular velocity with zero y component
$\boldsymbol{\omega}_{yz}$	Angular velocity with zero x component

TABLE II

REWARD TERMS FOR GOAL-REACHING TASKS

Term	Formulation	Weight
Target position	$\left(\psi(\lvert\boldsymbol{p}-\boldsymbol{p}^*\rvert, 0.5) - \frac{\lvert\boldsymbol{p}-\boldsymbol{p}^*\rvert^2}{5^2}\right)$	1.0
Stand pose	$\left(\lvert\boldsymbol{\theta}\rvert/\pi\right)^2 \psi(\lvert\boldsymbol{p}-\boldsymbol{p}^*\rvert, 0.1)$	−1.0
Stand DOF	$\psi(\lvert\boldsymbol{e}_{\text{dof}}\rvert, 1)\, \psi(\lvert\boldsymbol{p}-\boldsymbol{p}^*\rvert, 0.1)\, \psi(\lvert\boldsymbol{\theta}\rvert, \pi/2)$	1.0
Angular velocity	$\lvert\boldsymbol{\omega}\rvert^2\, w_{\text{cur}}$	−0.03
Collision	$N_{\text{contact}}\, w_{\text{cur}}$	−5.0
No wrench	$(\lvert\boldsymbol{f}\rvert^2 + 100\lvert\boldsymbol{\tau}\rvert^2)\, w_{\text{cur}}$	-1×10^{-5}

b) Dynamic tasks: We target two dynamic tasks, *Backflip* and *Barrel-Roll*, to evaluate the effectiveness of wrench-guided exploration for motions involving drastic attitude changes. The goal is to perform a backflip and a barrel roll, respectively. Similarly, in these tasks, with a probability of 10% the target position is set to the robot's start position so that the robot is required to remain at its initial location.

In the Dynamic task, the success rate was calculated based on the criterion that an episode is considered successful if the rotation angle error around the target axis $|\theta|$ is within 22.5° and the difference between the robot's height z and the target height z^* is within 0.1 m.

C. Reward and Observation design

We define the observations and rewards for the Goal-reaching tasks and Dynamic tasks below. The definitions of each variable are shown in Table I, and the respective reward functions are shown in Table II and Table III.

a) Observation design: For each task, we define a proprioceptive observation $\boldsymbol{o}^{\text{prop}}$ and a privileged observation $\boldsymbol{o}^{\text{priv}}$. The proprioceptive observation includes information that can be obtained from a real robot, such as the robot's posture, velocity, and contact information. The privileged observation includes information that can only be obtained in a simulation environment, such as the robot's position error and terrain information. Here, ϕ is a phase variable introduced for learning stabilization, and is defined as $\phi = t^3/(1+t^3)$, where t is the time step in the episode.

TABLE III

REWARD TERMS FOR DYNAMIC TASKS

Term	Formulation	Weight
Target angle	$\psi(2\pi - \theta, \pi/8)$	1.0
Stand height	$\psi(2\pi - \theta, \pi/8)\, \psi(z - z^*, 0.1)$	1.0
Stand DOF	$\psi(\lvert\boldsymbol{e}_{\text{dof}}\rvert, 0.5)\, \psi(2\pi - \theta, \pi/8)\, \psi(\lvert\boldsymbol{\theta}\rvert, \pi/8)$	1.0
Angular velocity	$r_\omega = \begin{cases} \lvert\boldsymbol{\omega}_{xz}\rvert^2 w_{\text{cur}}, & \text{Backflip} \\ \lvert\boldsymbol{\omega}_{yz}\rvert^2 w_{\text{cur}}, & \text{Barrel roll} \end{cases}$	−0.03
No tilt	$r_{\text{tilt}} = \begin{cases} \lvert g_y\rvert, & \text{Backflip} \\ \lvert g_x\rvert, & \text{Barrel roll} \end{cases}$	−0.3
Collision	$N_{\text{contact}}\, w_{\text{cur}}$	−5.0
No wrench	$(\lvert\boldsymbol{f}\rvert^2 + 100\lvert\boldsymbol{\tau}\rvert^2)\, w_{\text{cur}}$	-1×10^{-5}

Goal-reaching task

$$\boldsymbol{o}^{\text{prop}} = \left[\boldsymbol{\omega},\ \boldsymbol{g},\ \boldsymbol{e}_{\text{dof}},\ \dot{\boldsymbol{q}},\ \phi\right], \boldsymbol{o}^{\text{priv}} = \left[\Delta\boldsymbol{p},\ \boldsymbol{\theta},\ \boldsymbol{h},\ \boldsymbol{v}\right]$$

Dynamic task

$$\boldsymbol{o}^{\text{prop}} = \left[\boldsymbol{\omega},\ \boldsymbol{g},\ \boldsymbol{e}_{\text{dof}},\ \dot{\boldsymbol{q}},\ \phi\right], \boldsymbol{o}^{\text{priv}} = \left[\theta,\ (z - z^*),\ \boldsymbol{v}\right]$$

D. Learning Parameters

The scaling factor for the wrench action λ^{wrench} was set to 0.5. Based on the results of the exploration capability experiment in simulation Fig. 2, the parameters that characterize the exploration capability of WARL, $(f^{\max}, \tau^{\max})$, were set to $(1000N, 100Nm)$. The parameters of the WARL curriculum were set as follows.

$$\gamma = 0.6, \quad \delta = 2e-6, \quad \sigma = 0.8, \quad \epsilon = 0.01$$

$$N^{\text{succ}} = 100, \quad \eta = 0.9, \quad \kappa = 100.0$$

Furthermore, the standing height z^* of each robot was set to the base height at the default joint angle $\boldsymbol{q}_{\text{def}}$, which was 0.43m for KLEIYN and 0.3m for Unitree Go2.

We used PPO for learning, and trained 4096 robots in parallel for 5000 iterations. Each policy was updated at 50Hz, and one iteration consisted of 48 steps (one episode is 4 seconds). The learning time was about 5 hours on a single RTX 3090 Ti.

V. LEARNING EXPERIMENTS AND RESULTS

A. Learning Results in Simulation

Using the proposed method WARL, we trained two robots, KLEIYN and Unitree Go2, on the six tasks defined in Sec. IV (four Goal-reaching tasks and two Dynamic tasks). The learning curves and the motions learned by KLEIYN are shown in Fig. 4. Each graph shows the mean and variance of five trials. No robot-specific tuning was performed other than changing the standing height z^*.

Learning was successful for both robots on a variety of goal-reaching and dynamic motion tasks. The learning

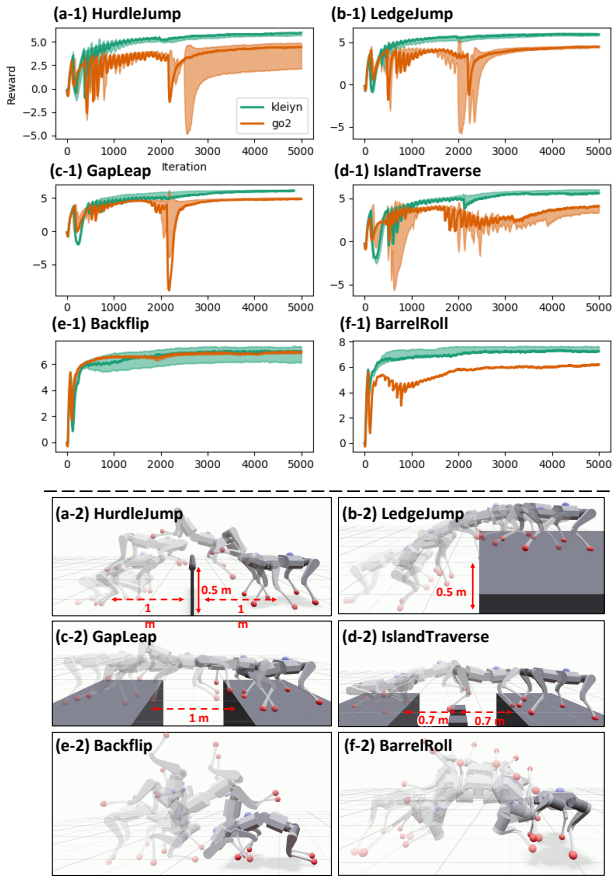

Fig. 4. [a-f]-1: Transition of rewards when learning six tasks with KLEIYN and Go2, respectively. [a-f]-2: Snapshots of the learned motions in each task.

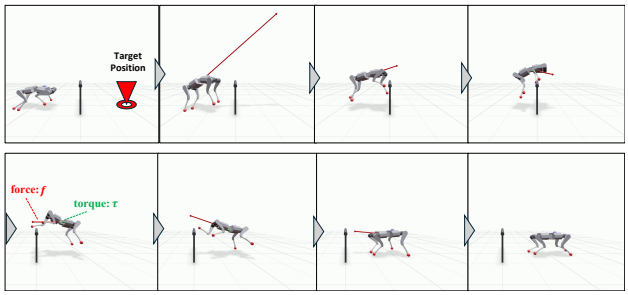

Fig. 5. Learned wrench and joint actions in *HurdleJump* at iteration 500. Wrench assistance is applied to clear the obstacle, while joint control is coordinated to lift the feet.

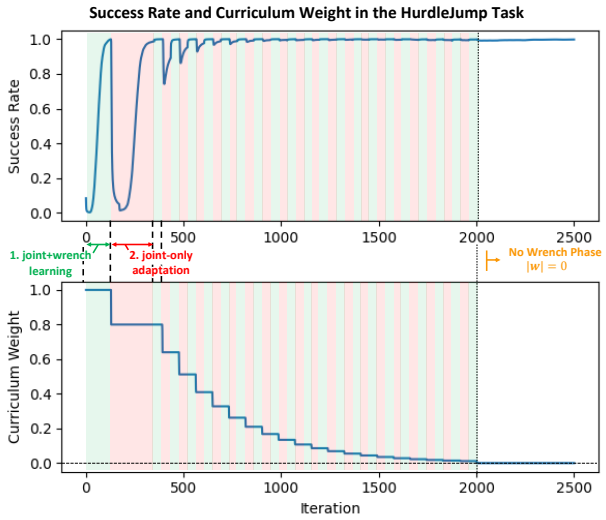

Fig. 6. Transition of success rate and curriculum weight up to 2500 iterations in the HurdleJump task.

curves show an overall trend of reward improvement, with temporary drops caused by curriculum-driven wrench decay followed by recovery through adaptive learning. Learning also progressed using the same parameters across different robots.

In the IslandTraverse task, the robot acquired a motion that directly jumped over the 1.4 m gap without using the intermediate footholds. This likely occurred because wrench-based exploration dominated the learning process, and motions utilizing footholds were not discovered.

The learned wrench and joint motion at iteration 500 in the HurdleJump task are shown in Fig. 5. The results indicate that the hurdle-crossing motion is largely generated by the learned wrench, while coordinated leg motions are also acquired to lift the feet and clear the hurdle.

To illustrate how the curriculum weight is updated based on the success rate, the transition of the success rate and curriculum weight during learning of the HurdleJump task is shown in Fig. 6.

The success rate temporarily decreases when switching from the Joint+Wrench learning phase to the Joint-only adaptive learning phase, but recovers through subsequent adaptive learning and gradually improves.

It takes about 2000 iterations to complete the curriculum, whereas the success rate reaches nearly 1.0 within about 100 iterations in the first Joint+Wrench phase. This suggests that the target motion itself can be acquired quickly using the wrench.

Most of the training time is spent on adaptive learning to remove wrench dependence and reproduce the motion using only joint control. This indicates that while the rough structure of the motion can be obtained quickly, improving the efficiency of adaptive learning is important for reducing total training time.

B. Ablation Study

To verify the effectiveness of each component of the proposed method, we conduct an ablation study under four conditions: (1) WARL, (2) w/o Switch, (3) w/o Wrench Reward, and (4) Baseline. (2) w/o Switch: The adaptation phase is omitted, and learning is always performed in the exploration phase. (3) w/o Wrench Reward: The penalty term for the wrench amount in the exploration phase is removed from the reward function. (4) Baseline: Training is performed using only PPO and joint control, without including the wrench in the action space. All other training conditions, such as the reward function, are identical to the proposed method.

We learn the HurdleJump task under these conditions and show the transition of the success rate in Fig. 7. Since condition (2) succeeded in only 3 out of 10 trials, both successful and unsuccessful learning curves are shown.

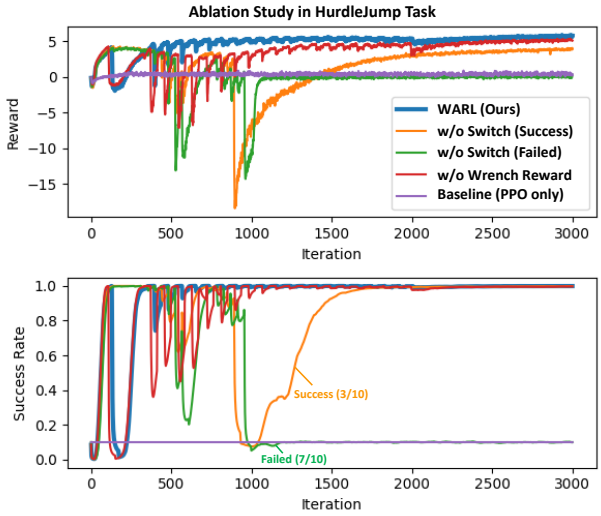

Fig. 7. Results of the ablation study of the proposed method on the HurdleJump task. Learning was performed under four conditions: (1) WARL, (2) w/o Switch, (3) w/o Wrench Reward, and (4) Baseline, and the transition of the success rate is shown.

(1) WARL showed the most stable improvement in success rate overall. In (2) w/o Switch, since the Joint-only adaptation phase was omitted, the learned policy depended on the wrench and the success rate tended to decrease in the final stage of the curriculum. Learning success after the curriculum was inconsistent, indicating that the Switching Curriculum stabilizes learning.

In (3) w/o Wrench Reward, learning eventually succeeded, but the decrease in success rate at the phase switch was larger than in (1) WARL. Without a penalty on wrench usage, policies with stronger wrench dependence tend to be learned during the Joint+Wrench phase.

In (4) Baseline, since the wrench was not included in the action space, exploration guidance was not obtained and the success rate remained low throughout learning. This indicates that the task is difficult to explore using only joint control and that wrench-based exploration guidance is highly effective.

These comparisons indicate that the Switching Curriculum plays a central role in removing wrench dependence, while the Wrench Reward helps suppress excessive reliance on the wrench.

C. Execution of Policy for Real Robot in MuJoCo Simulator

The policy $\pi_j(\boldsymbol{a}|\boldsymbol{s})$ acquired by WARL includes privileged observation and cannot be directly applied to a real robot. Therefore, we train a policy $\pi_s(\boldsymbol{a}|\boldsymbol{o}^{\text{prop}})$ that takes only Proprioceptive Observation as input using Teacherâ€“Student learning.

Snapshots of the HurdleJump, LedgeJump, and GapLeap motions executed on the MuJoCo [23] simulator are shown in Fig. 8.

The learned policy successfully reproduces similar motions on the MuJoCo simulator.

VI. Discussion

a) Advantages and Challenges of Wrench-Based Exploration: We experimentally investigated how introducing a wrench into the action space affects learning characteristics. The advantages and challenges are summarized as follows.

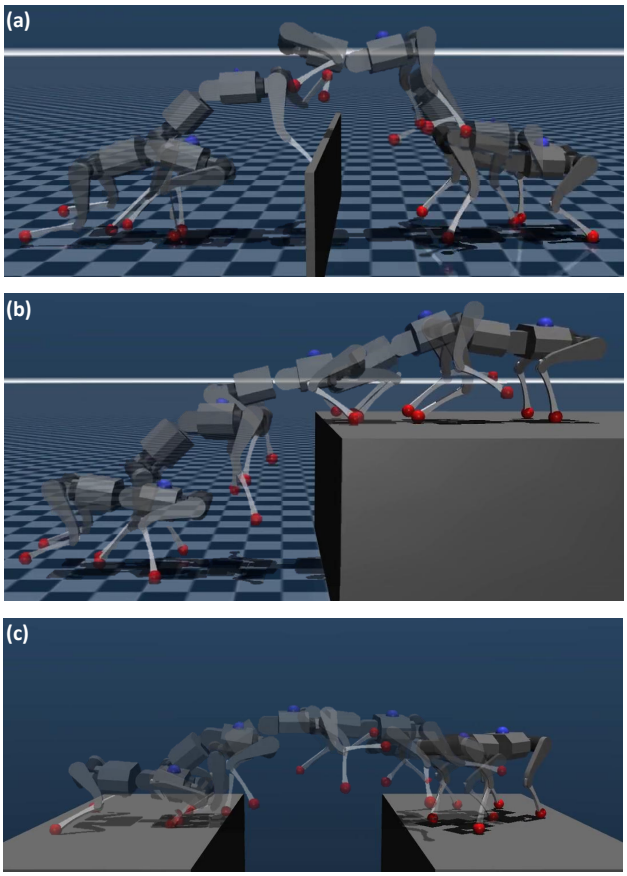

Fig. 8. Snapshots of the learned policy executed on the MuJoCo simulator. (a) HurdleJump, (b) LedgeJump, (c) GapLeap

The first advantage is accelerated learning. By including a wrench in the action space, the agent obtains higher exploration capability compared to joint-space control alone. This expands the range of exploration that leads to the discovery of target motions and shortens the time required to acquire them.

The second advantage is the reduction of heuristic learning design. When learning only with joint actions, the limited exploration capability often requires artificial adjustments such as detailed reward tuning and step-by-step curricula. Introducing a wrench broadens the exploration space, allowing learning to progress with less reliance on such auxiliary designs.

b) Improvement of the Learning Framework: For simplicity, the wrench policy and joint policy were trained with the same reward function, although their roles are inherently different. The primary role of the wrench policy is to facilitate learning of the joint policy.

Using the same reward for both policies may cause the system to favor a wrench policy that does not sufficiently consider embodiment. To address this issue, the reward for the wrench policy should instead evaluate how effectively it promotes the autonomous achievement of the target motion by the joint policy. One possible approach is to introduce a structure in which joint policy learning forms an inner

loop while its progress is optimized externally, similar to hierarchical reinforcement learning.

There is also room for improvement on the joint-policy side. The current framework adapts to a gradually decaying wrench, but another approach is to use wrench-generated motions as an imitation target. For example, trajectories obtained from wrench-driven motions could be used as teacher signals, enabling the joint policy to acquire motions that are more consistent with the robot's embodiment.

c) Future Development Potential: The idea of incorporating a wrench into the action space is not limited to a specific robot morphology. Although this study focused on a quadruped robot, the approach may also be effective for robots with different locomotion forms, such as bipedal or multi-legged systems, particularly in tasks requiring dynamic control.

Furthermore, the wrench policy is relatively weakly tied to a specific body structure. This abstraction suggests the possibility of transferring motion knowledge between different embodiments. By using the wrench as a body-independent representation, it may be possible to construct a framework for sharing and transferring motor knowledge across robots.

From this perspective, the proposed approach contributes not only to improving learning efficiency but also to expanding the design principles of reinforcement-learning-based robot control.

VII. Conclusion

In this study, we proposed a new method, WARL, which introduces a wrench into the action space to overcome the constraints on exploration capability in reinforcement learning for legged robots. The proposed method integrates wrench-guided exploration and success rate-based curriculum learning.

The experimental results demonstrated that by using WARL, it is possible to learn motions uniformly under various conditions without requiring individual design for each terrain or task. In addition, an ablation study verified that the Switching Curriculum, which gradually removes the wrench, is effective for the autonomous acquisition of a joint policy.

On the other hand, it also became clear that the introduction of a wrench tends to lead to the learning of motions that do not fully utilize the robot's specific embodiment. This issue is attributed to the fact that the wrench is an abstract input that is relatively independent of the body structure.

In the future, further refinement of the learning framework is necessary, such as a method for generating wrenches that explicitly considers embodiment and a transition from exploration using wrenches to the acquisition of a joint control policy.

References

[1] 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*, vol. 4, no. 26, 2019.

[2] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," *arXiv preprint arXiv:2309.05665*, 2023.

[3] H. Kim, H. Oh, J. Park, Y. Kim, D. Youm, M. Jung, M. Lee, and J. Hwangbo, "High-speed control and navigation for quadrupedal robots on complex and discrete terrain," *Science Robotics*, vol. 10, no. 102, p. eads6192, 2025.

[4] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proceedings of the 2022 Conference on Robot Learning*, 2022, pp. 91–100.

[5] V. Atanassov, J. Ding, J. Kober, I. Havoutis, and C. Della Santina, "Curriculum-based reinforcement learning for quadrupedal jumping: A reference-free design," *IEEE Robotics & Automation Magazine*, 2024.

[6] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572–587, 2024.

[7] S. Ha, J. Lee, M. van de Panne, Z. Xie, W. Yu, and M. Khadiv, "Learning-based legged locomotion: State of the art and future perspectives," *The International Journal of Robotics Research*, vol. 44, no. 8, pp. 1396–1427, 2025.

[8] D. Vogel, R. Baines, J. Church, J. Lotzer, K. Werner, and M. Hutter, "Robust ladder climbing with a quadrupedal robot," *arXiv preprint arXiv:2409.17731*, 2024.

[9] G. Bellegarda, C. Nguyen, and Q. Nguyen, "Robust quadruped jumping via deep reinforcement learning," *Robotics and Autonomous Systems*, vol. 182, p. 104799, 2024.

[10] K. Yoneda, K. Kawaharazuka, T. Suzuki, T. Hattori, and K. Okada, "Kleiyn: A quadruped robot with an active waist for both locomotion and wall climbing," *arXiv preprint arXiv:2507.06562*, 2025.

[11] R. Martín-Martín, M. A. Lee, R. Gardner, S. Savarese, J. Bohg, and A. Garg, "Variable impedance control in end-effector space: An action space for reinforcement learning in contact-rich tasks," in *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2019, pp. 1010–1017.

[12] K. Jiang, Z. Fu, J. Guo, W. Zhang, and H. Chen, "Learning whole-body loco-manipulation for omni-directional task space pose tracking with a wheeled-quadrupedal-manipulator," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1481–1488, 2024.

[13] E. Aljalbout, F. Frank, M. Karl, and P. van der Smagt, "On the role of the action space in robot manipulation learning and sim-to-real transfer," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5895–5902, 2024.

[14] H. Duan, J. Dao, K. Green, T. Apgar, A. Fern, and J. Hurst, "Learning task space actions for bipedal locomotion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1276–1282.

[15] G. Bellegarda and A. Ijspeert, "Cpg-rl: Learning central pattern generators for quadruped locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 547–12 554, 2022.

[16] D. Kang, M.-G. Kim, T.-G. Song, H. Kim, S. Ha, and H.-W. Park, "Dynamic policy learning for legged robot with simplified model pretraining and model homotopy transfer," *arXiv preprint arXiv:2512.24698*, 2025.

[17] 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*, vol. 7, no. 62, 2022.

[18] K. Yoneda, K. Kawaharazuka, and K. Okada, "Efgcl: Learning dynamic motion through spotting-inspired external force-guided curriculum learning," *IEEE Robotics and Automation Letters*, vol. 11, no. 5, pp. 5907–5913, 2026.

[19] J. P. Sleiman, H. Li, A. Adu-Bredu, R. Deits, A. Kumar, K. Bergamin, M. Bhardwaj, S. Biddlestone, N. Burger, M. A. Estrada *et al.*, "Zest: Zero-shot embodied skill transfer for athletic robot control," *arXiv preprint arXiv:2602.00401*, 2026.

[20] Z. Cao, Y. Zhang, B. Nie, H. Lin, H. Li, and Y. Gao, "Learning motion skills with adaptive assistive curriculum force in humanoid robots," *arXiv preprint arXiv:2506.23125*, 2025.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[22] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox, "Gpu-accelerated robotic simulation for distributed reinforcement learning," 2018.

[23] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.