Learning Robust Quadrupedal and Bipedal Locomotion on a Quadruped Robot using One Policy

Yasen Jia[1] Shengcong Sang[1] and Yan Huang[1,2,*]

***Abstract*—Locomotion is an essential ability for legged robots to be applied in various conditions. The ability to perform quadrupedal and bipedal locomotion on one robot can bring excellent versatility for legged robots. However, few works have studied learning both quadrupedal and bipedal locomotion using one policy. Using different policies for different locomotion requires switching between polices. The switching process may introduce undesired behaviors during the transition because of distribution shift. In this work, we propose a framework for a quadruped robot to learn quadrupedal and bipedal locomotion in a unified policy via reinforcement learning (RL). Through task-dependent reward designs based on binary task indicators, a policy can learn both kinds of locomotion even they are mutually conflicting. Results show that the quadruped robot can perform robust bipedal and quadrupedal locomotion on complex terrains. Besides, the robot can perform smooth transition within 1 second between two kinds of locomotion even during motion at speeds as high as 1.5 m/s. Our framework outperforms previous methods in terms of training efficiency and still demonstrates comparable terrain traversal performance on stairs and slopes.**

***Index Terms*—reinforcement learning, quadruped, biped, robot**

I. Introduction

Locomotion control of legged robots is an important issue, in which quadrupedal locomotion and bipedal locomotion are two extensively investigated gaits [1]–[6]. While quadrupedal or bipedal locomotion can be realized separately, the ability to perform both of them and switch between them on one robot can endow the robot with superior mobility on rough terrains and excellent versatility for complex tasks [7]. To obtain this ability, the motion control policy for legged robots is crucial. Existing studies have shown the possibility of bipedal locomotion on quadruped robots [8]–[15], and the ability of performing these two locomotion modes on biped robots [7], [16]. Previous methods can be categorized into two classes: model-based control and RL-based control.

Model-based control approaches often use a mathematical model to describe robot dynamics and formulate an optimal control problem for legged locomotion, whose goal is to find the control signal that can minimize the objective function, while satisfying constraints [6], [18]–[20]. For model-based control methods, achieving bipedal and quadrupedal locomotion on one robot usually requires modification of dynamic models and hand-crafted key poses. Kamioka *et al.* [16] proposed an extended planning algorithm for bipedal and quadrupedal locomotion and intermediate transitions, based on a linear time-variant inverted pendulum model that allows variable height of COG. However, it demanded manual design of key poses, which could be tedious. Kobayashi *et al.* [7] achieved automatic selection of bipedal and quadrupedal locomotion on the Gorilla Robot III through the combination of a gait selection strategy and a speed adjustment strategy. However, the resulted speed of bipedal and quadrupedal locomotion was relatively slow ($<$ 0.2 m/s). Amatucci *et al.* [14] accelerated model predictive control solving through decomposing robot dynamics into smaller subsystems and utilizing the Alternating Direction Method of Multipliers (ADMM) to ensure consensus among subsystems. The reduction of computational time was verified to be up to 75% and this method enabled a quadruped robot to perform bipedal gait. Nevertheless, the result was limited to simulation. Kim *et al.* [15] integrated linear complementarity constraint with contact-implicit differential dynamic programming (DDP) and task-specific costs to enable physically feasible contact exploration. It successfully enabled a quadruped robot to perform front-leg rearing motion, which is similar to bipedal locomotion. However, it did not demonstrate the bipedal motion on uneven terrains or under unexpected forces.

RL-based control methods for legged robots have attracted more attention in recent years [2]–[5], [21]. RL-based control methods usually utilize physics simulators to collect data of robots interacting with the environment. The data is then fed into the RL algorithms to optimize the policy network towards accomplishing tasks specified by reward functions. RL-based control methods have also been used to achieve bipedal locomotion on quadruped robots. Smith *et al.* [9] proposed a method based on transfer learning and enabled a quadruped robot to navigate to a specified position with only hind legs contacting with the ground. However, it leveraged the convenience of indoor localization and was thus limited to indoor environments. Li *et al.* [10] trained a quadruped robot to learn bipedal walking and achieved upper limb imitation through the combination of a motion-conditioned policy and a motion target generator. Nevertheless, their method did not show smooth transitions with bipedal and quadrupedal

This work was supported by the National Natural Science Foundation of China (No. 62073038) and Beijing Institute of Technology Research Funds for High-Level Talents

[1] School of Mechantronical Engineering, Beijing Institute of Technology, Beijing 100081, China jason_1120202397@163.com

[2] Key Laboratory of Biomimetic Robots and Systems, Ministry of Education, Beijing 100081, China yanhuang@bit.edu.cn

[*] Corresponding Author

TABLE I: Comparison with existing methods on learning quadrupedal and bipedal locomotion on a quadruped robot.

Method	Real Robot	w/o Prior Trajectory	Unified Policy	Transition during High-speed Motion	Robust Locomotion on Complex terrains
Smith *et al.* [9]	√	√	×	×	×
Li *et al.* [10]	√	√	×	×	×
Peng *et al.* [11]	×	×	×	×	×
Su *et al.* [12]	√	√	×	×	√
Kim *et al.* [13]	√	√	×	×	√
Huang *et al.* [17]	√	√	√	×	√
Ours	√	√	√	√	√

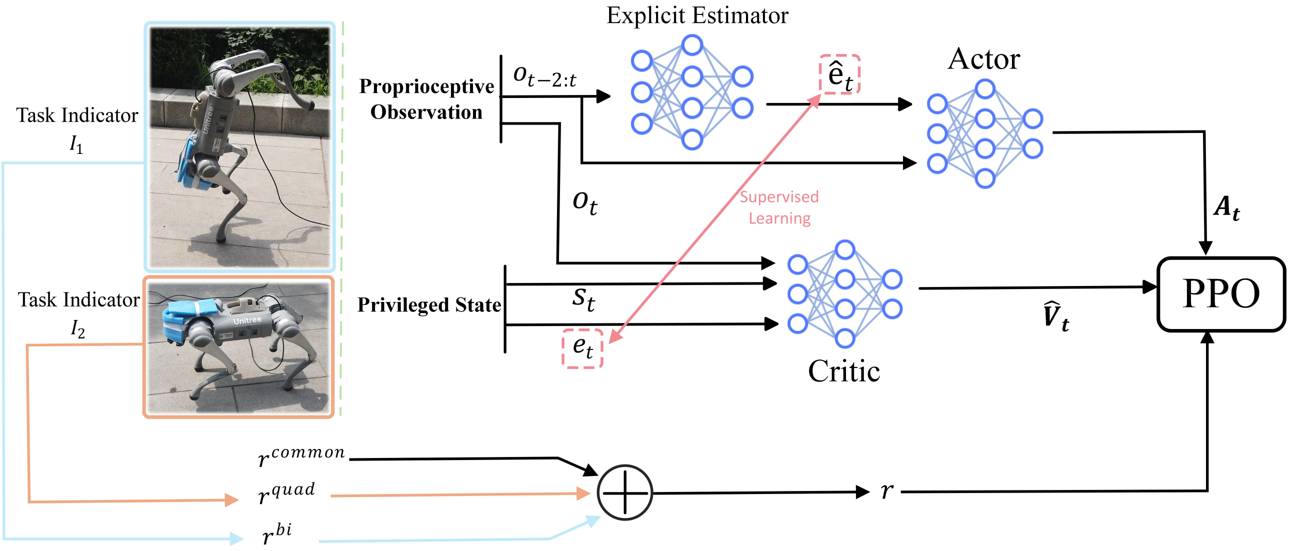

Fig. 1: Overview of our proposed RL training framework. During training, we train an explicit estimator and a control policy concurrently, through supervised learning and reinforcement learning respectively. Total reward r incorporates two kinds of task rewards r^{quad} and r^{bi} with common rewards r^{common}, in which task rewards are constructed through task indicators.

locomotion during agile motion or on complex terrains. Peng *et al.* [11] proposed a framework that incorporated Adversarial Motion Prior (AMP) with a teacher-student policy for a quadruped robot to imitate bipedal locomotion. However, it showed results only in simulation and the reference trajectory for AMP needed to be generated through trajectory optimization, which added more complexity. Su *et al.* [12] proposed two kinds of methods to augment vanilla Policy Proximal Optimization (PPO) algorithm and successfully performed bipedal walking on a quadruped robot. However, its bipedal locomotion in real-world was restricted to level ground and did not demonstrate the possibility of transitions from quadruped to biped while performing agile quadrupedal locomotion. Kim *et al.* [13] proposed barrier-based style rewards to enable multiple legged locomotion (quadruped, tripod, biped) and demonstrated compelling results on a quadruped robot. However, the proposed policy network for different locomotion was separate, which increased the complexity of the whole controller for performing quadrupedal locomotion, bipedal locomotion and the transition between them. Huang *et al.* [17] proposed a Mixture of Experts (MoE) framework to learn bipedal locomotion and quadrupedal locomotion for a quadruped robot. However, it requires training multiple expert networks and falls behind in training efficiency. Qualitative comparison with previous works are summarized in Table I.

Though many works have been done to achieve bipedal or quadrupedal locomotion on one robot, few of existing works have achieved robust bipedal locomotion, quadrupedal locomotion and the transition between them on a real legged robot using one unified policy. Using different policies for different locomotion requires switching between polices. Switching process may introduce undesired behaviors during the transition because of distribution shift [22], such as collisions with surroundings. MoE can alleviate this problem but introduce extra training and on-board computation load.

In this work, we propose an approach for a quadruped robot to learn quadrupedal and bipedal locomotion in one unified policy through reinforcement learning. We utilize binary mask indicators to enable multi-task reinforcement learning for one policy and achieve robust locomotion and smooth transitions between two kinds of locomotion on a quadruped robot. Our framework reduces the training time by 97% compared with [17], and still achieves comparable terrain traversal performance on stairs and slopes.

Our contributions are as follows. First, we propose an efficient learning framework for a quadruped robot to learn quadrupedal and bipedal locomotion in one unified policy based on binary task indicators. Second, we demonstrate quadrupedal locomotion, bipedal locomotion and smooth transitions between them with one unified policy on a real quadruped robot. The policy could perform robust quadrupedal locomotion and bipedal locomotion on uneven terrains like slopes, discrete obstacles and stairs, and could achieve smooth transitions during high-speed motion or on uneven terrains.

II. Task-Dependent Reward Design

A. Overview

To enable one policy to learn multiple tasks, the key is the reward design. We construct binary task indicators to distinguish reward functions belonging to distinct tasks. The final reward is the summation of biped task rewards, quadruped task rewards and common rewards, as shown in Fig. 1.

B. Constructing Binary Task Indicators

A reward function is often designed as a square sum of some physical states of the robot, or designed as an exponential of the square sum. These kinds of designs are common in previous works [23]. In the context of multi-task reinforcement learning, directly designing reward functions like those may cause conflicts between different tasks. We need to design reward functions as task-dependent functions to learn multiple tasks in a unified policy.

For our task of achieving bipedal locomotion, quadrupedal locomotion and the transition between them, we identify bipedal locomotion and quadrupedal locomotion as two tasks. The transition between these two locomotion is an emergent behavior learned by the policy. We specify task one as bipedal locomotion and task two as quadrupedal locomotion. We define task mode variable $x \in \mathbb{Z}^+$ to indicate task selection. Then we can construct task indicators I_i $(i = 1, 2)$, which are defined as follows:

$$I_1 = \begin{cases} 1, & \text{if } x = 1 \\ 0, & \text{if } x = 2 \end{cases} \tag{1}$$

$$I_2 = \begin{cases} 0, & \text{if } x = 1 \\ 1, & \text{if } x = 2 \end{cases} \tag{2}$$

With this definition, we can have $I_1 = 1, I_2 = 0$ when $x = 1$, and $I_1 = 0, I_2 = 1$ when $x = 2$. When $x = i$, it means that task i is activated. When bipedal locomotion is activated ($x = 1$), only biped task rewards and common rewards will function, guiding the policy to learning bipedal locomotion while quadruped task rewards are masked out. The same principle applies to quadrupedal locomotion ($x = 2$).

C. Task-Dependent Reward Design

The task indicator serves as a mask over original reward functions. Task-dependent rewards for task i can be defined as the product of an original reward $r^{original}$ and the task indicator I_i:

$$r^{task} = r^{original} I_i$$

Our rewards r can be classified into three categories: common rewards r^{common}, quadrupedal locomotion rewards r^{quad} and bipedal locomotion rewards r^{bi}. Common rewards r^{common} consist of termination reward, collision reward, angular velocity reward, DoF (Degree of Freedom) velocity reward, DoF acceleration reward, action rate reward, action smoothness reward, torque reward, hip position reward and DoF position limit reward. Common rewards are utilized basically to penalize undesired behaviors and foster better sim-to-real transfer. All reward terms with their expressions and weights are summarized in Table II.

TABLE II: Reward Terms with Expressions and Weights

Term	Expression	Weight
	Common	
Termination	$\mathbb{1}_{termination}$	$-2e^2$
Collision	$\mathbb{1}_{collision}$	-1
Angular velocity	$\|\|\boldsymbol{\omega}_{t,xy}\|\|_2^2$	$-5e^{-2}$
DoF velocity	$\|\|\dot{\boldsymbol{q}}_{\boldsymbol{t}}\|\|_2^2$	$-5e^{-4}$
DoF acceleration	$\|\|\ddot{\boldsymbol{q}}_t\|\|_2^2$	$-2e^{-7}$
Action rate	$\|\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|\|_2^2$	$-1e^{-2}$
Action smoothness	$\|\|\boldsymbol{a}_t - 2\boldsymbol{a}_{t-1} - \boldsymbol{a}_{t-2}\|\|_2^2$	$-1e^{-2}$
Torque	$\|\|\boldsymbol{\tau}_t\|\|_2^2$	$-2e^{-4}$
Hip position	$\|\|\boldsymbol{q}_{t,hip}\|\|_2^2$	-1
DoF position limit	$\mathbb{1}(\boldsymbol{q}_t \notin [\boldsymbol{q}_{min}, \boldsymbol{q}_{max}])$	-1
	Quadruped	
Lin. velocity tracking	$\exp(-4\|\|\boldsymbol{v}_{xy}^* - \boldsymbol{v}_{xy}\|\|_2^2)I_2$	2
Ang. velocity tracking	$\exp(-4\|\|\omega_z^* - \omega_z\|\|_2^2)I_2$	1
Base height	$\exp(-2\|\|h^{*,quad} - h\|\|_2^2)I_2$	1
Base angle	$\exp(-10\|\|\boldsymbol{n}_{b,xy}\|\|_2^2)I_2$	2
Lin. velocity z	$v_z^2 I_2$	-0.5
Periodic gait	$\exp(-2E[R^{quad}(s, \Phi)])I_2$	1.5
Foot clearance	$\exp(-100\ error_{clearance}^{quad})I_2$	1
	Biped	
Lin. velocity tracking	$\exp(-4\|\|\boldsymbol{v}_{xy}^* - \boldsymbol{v}_{zy}\|\|_2^2)\tilde{I}_{up}I_1$	2
Ang. velocity tracking	$\exp(-4\|\|\omega_z^* - \omega_z\|\|_2^2)\tilde{I}_{up}I_1$	1
Base height	$\exp(-2\|\|h^{*,bi} - h\|\|_2^2)I_1$	1
Base angle	$\exp(-2\|\|\boldsymbol{\phi}_{xy}^* - \boldsymbol{\phi}_{xy}\|\|_2^2)I_1$	1
Periodic gait	$\exp(-2E[R^{bi}(s, \Phi)])\tilde{I}_{up}I_1$	2
Foot clearance	$\exp(-100\ error_{clearance}^{bi})I_1$	1
Rear legs position	$\|\|\boldsymbol{q}_{fl} - \boldsymbol{q}_{fr}\|\|_2^2 I_1$	-0.5
Upright contacts	$\mathbb{1}_{front\ feet\ contact}I_1$	1

For quadrupedal locomotion (quad), task-dependent rewards include $r_{tracking\ lin}^{quad}$, $r_{tracking\ ang}^{quad}$, $r_{lin\ vel\ z}^{quad}$, $r_{base\ height}^{quad}$, $r_{base\ angle}^{quad}$, $r_{foot\ clearance}^{quad}$, $r_{periodic\ gait}^{quad}$. The original reward for rewards other than $r_{base\ angle}^{quad}$, $r_{foot\ clearance}^{quad}$ and $r_{periodic\ gait}^{quad}$ are defined the same as [23]. Base angle reward $r_{base\ angle}^{quad}$ is defined as Equation 3 and foot clearance reward $r_{foot\ clearance}^{quad}$ is defined as Equation 4:

$$r_{base\ angle}^{quad} = \exp(-10||\boldsymbol{n}_{b,xy}||_2^2)I_2 \tag{3}$$

where $\boldsymbol{n}_b$ is the normal vector of the surface below the robot base projected into the robot's base frame.

$$r_{foot\ clearance}^{quad} = \exp(-100\ error_{clearance}^{quad})I_2 \tag{4}$$

where clearance error $error^{quad}_{clearance} = \sum_{feet}(p_i - (max(H_{sample,i}) + p_{des}))^2||\boldsymbol{v}_{f,xy,i}||_2$, p_i is the height of the i-th foot, $H_{sample,i}$ is the sampled terrain height around the i-th foot (within 2 cm), p_{des} is the desired foot height (set to 0.08 m), and $\boldsymbol{v}_{f,xy}$ is the velocity of i-th foot along x and y axes. We specify the first leg as front left (fl) leg, the second leg as front right (fr) leg, the third leg as rear left (rl) leg and the fourth leg as rear right (rr) leg.

The reward for periodic gait in quadrupedal locomotion $r^{quad}_{periodic\ gait}$ is constructed respecting the principle in [4], which is defined as follows:

$$\begin{aligned} E[R^{quad}(s, \varPhi)] = \sum_{feet}(&E[C_{frc}(\varPhi + \theta_i)] \cdot q_{i,frc}(s) \\ &+ E[C_{spd}(\varPhi + \theta_i)] \cdot q_{i,spd}(s)) \end{aligned}$$

$$r^{quad}_{periodic\ gait} = \exp(-2E[R^{quad}(s, \varPhi)])I_2$$

where s is the state of the robot, $\varPhi$ is the clock variable ($\varPhi = t/T_{gait}$), θ_i is the phase offset for i-th leg, $q_{i,frc}$ is the norm of contact force of i-th foot, $q_{i,spd}$ is the norm of velocity of i-th foot. We set the gait period T_{gait} to 0.45, $\theta_{fl} = 0.0, \theta_{fr} = 0.5, \theta_{rl} = 0.5, \theta_{rr} = 0.0$, and the swing ratio of the gait is 0.4, which forms a trot gait.

For bipedal locomotion (bi), task-dependent rewards include $r^{bi}_{tracking\ lin}$, $r^{bi}_{tracking\ ang}$, $r^{bi}_{base\ height}$, $r^{bi}_{base\ angle}$, $r^{bi}_{foot\ clearance}$, $r^{bi}_{periodic\ gait}$, $r^{bi}_{rear\ legs}$, $r^{bi}_{upright\ contacts}$. Due to motor capability constraints, it's very hard for a quadruped robot to learn to get up and track given velocity command simultaneously. Thus we introduce an extra upright indicator $\tilde{I}_{up}$ as an assistive technique, instructing the robot to first learn to get up and then learn to perform bipedal gait and to track given velocity command. This upright indicator is constructed as a step function, which is defined as Equation 5. We assume that if the base pitch angle of the robot exceeds 1.4 radians, then the robot is in bipedal pose. Expressions for these reward functions are shown in Table II, where $\boldsymbol{v}^*_{xy}$ is the desired base linear velocity in x and y axes, $\boldsymbol{v}_{zy}$ is the real base linear velocity in z and y axes, ω^*_z is the desired base angular velocity in z axis, ω_z is the real base angular velocity in z axis, $h^{*,bi}$ is the desired base height for bipedal locomotion, h is the real base height, $\boldsymbol{\phi}^*_{xy}$ is the desired base euler angle in x and y axes, $\boldsymbol{\phi}_{xy}$ is the real base euler angle in x and y axes, $\boldsymbol{q}_{fl}$ and $\boldsymbol{q}_{fr}$ are joint positions of front left leg and front right leg.

$$\tilde{I}_{up} = \mathbb{1}(\phi_y - 1.4) \tag{5}$$

Foot clearance reward $r^{bi}_{foot\ clearance}$ and periodic gait reward $r^{bi}_{periodic\ gait}$ have similar forms with quadrupedal counterparts, but only consider front feet, as shown in Equation6 and Equation7. The $error^{bi}_{clearance}$ in Equation6 is defined as $\sum_{front\ feet}(p_i - (max(H_{sample,i}) + p_{des}))^2||\boldsymbol{v}_{f,xy,i}||_2$.

$$r^{bi}_{foot\ clearance} = \exp(-100\ error^{bi}_{clearance})I_1 \tag{6}$$

$$\begin{aligned} E[R^{bi}(s, \varPhi)] = \sum_{front\ feet}(&E[C_{frc}(\varPhi + \theta_i)] \cdot q_{i\ frc}(s) \\ &+ E[C_{spd}(\varPhi + \theta_i)] \cdot q_{i\ spd}(s)) \end{aligned}$$

$$r^{bi}_{periodic\ gait} = \exp(-2E[R^{bi}(s, \varPhi)])I_1 \tag{7}$$

where the gait period T_{gait} is 0.45, $\theta_{fl} = 0.0, \theta_{fr} = 0.5$, and the swing ratio of the gait is 0.4.

III. Implementation Details

Our method is illustrated in Fig. 1. Besides task-dependent reward design, we adopt the concurrent training method in [24] and train a unified estimator-policy set for two kinds of locomotion and the transition between them. Since the estimator and the policy are trained concurrently for two kinds of locomotion, the resulted estimator is capable of estimating explicit states for both quadrupedal and bipedal locomotion.

A. Observation and Action

The observation at time t o_t is defined as

$$o_t = (\boldsymbol{cmd}, \boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{a_{t-1}}, \boldsymbol{m}, \boldsymbol{clock}, \boldsymbol{ratio})$$

where $\boldsymbol{cmd}$ is the velocity command, $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ are the base euler angle and angular velocity, $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$ are the joint positions and velocities, $\boldsymbol{a_{t-1}}$ is the actions of last time step, and $\boldsymbol{m}$ is a one-hot vector [25] with two elements, indicating the activation status of two tasks. Clock input $\boldsymbol{clock}$ is defined similar to [4], as shown in Equation 8 and Equation 9, for quadrupedal locomotion and bipedal locomotion respectively. The phase ratio $\boldsymbol{ratio}$ consists of the swing-phase ratio and the stance-phase ratio of specified gait.

$$\begin{aligned} clock^{quad} = \{&sin(\frac{2\pi(\varPhi + \theta_1)}{L}), sin(\frac{2\pi(\varPhi + \theta_2)}{L}), \\ &sin(\frac{2\pi(\varPhi + \theta_3)}{L}), sin(\frac{2\pi(\varPhi + \theta_4)}{L})\} \end{aligned} \tag{8}$$

$$\begin{aligned} clock^{bi} = \{&sin(\frac{2\pi(\varPhi + \theta_1)}{L}), sin(\frac{2\pi(\varPhi + \theta_2)}{L}), \\ &0, 0\} \end{aligned} \tag{9}$$

The input of the estimator is $o_{t-2:t} = [o_t, o_{t-1}.o_{t-2}]^T$, including observation from current time step and last two time steps. The output of the estimator $\hat{e}_t$ consists of base linear velocity, foot contact probability, normal vector of the surface below the robot base projected into base frame and feet heights. We obtain the label data e_t from the simulation and train the estimator through supervised learning, utilizing Mean Square Error (MSE) loss. The input of the actor is the concatenation of $o_{t-2:t}$ and $\hat{e}_t$. The input of the critic includes proprioceptive observation o_t, estimator label data e_t and domain randomization parameters s_t. The action of the policy a_t is desired joint angles, which is converted into desired torque via a PD controller.

TABLE III: Hyper Parameters for Training

PPO	
Batch size	4096 x 24
Mini-batch size	4096 x 6
Number of epochs	5
Clip range	0.2
Entropy coefficient	0.01
Discount factor	0.99
GAE discount factor	0.95
Desired KL-divergence	0.01
Learning rate	adaptive
Supervised Learning	
Learning rate	$1e^{-4}$
Mini-batch size	768
Number of epochs	5

TABLE IV: Initial State Distribution of Our Training.

State Term	Range	Unit
Hip joint angle	[-0.2, 0.2]	rad
Thigh joint angle	[-1.0, 1.0]	rad
Calf joint angle	[-0.6, 0.6]	rad
Base linear velocity	[-0.5, 0.5]	m/s
Base angular velocity	[-0.5, 0.5]	rad/s
Base pitch angle	[0, 1.3]	rad

B. Environment Setup

We utilize IsaacGym [26] as the simulator, and customize an implementation of training an explicit estimator and a policy concurrently [24] based on legged_gym [23]. PPO [27] is adopted as our training algorithm. The backbone of policy, critic and estimator is MLP (Multi Layer Perceptron). The hidden layers are $[512, 256, 128]$ for the policy network and $[256, 128]$ for the estimator network. Some hyper parameters for training are listed in Table III. We train 4096 parallel environments on complex terrains including slope, stairs and discrete obstacles and it takes 4000 iterations for the policy to obtain satisfying performance, which corresponds to 144 minutes of wall clock time using a Nvidia RTX 3080 for training. We set the max episode length to 18 seconds, corresponding to 900 time steps with a control policy of 50 Hz. Episodes are terminated once the maximum episode length is reached or collision between the robot base and terrains happens. We use Go2 quadruped robot from Unitree Robotics [28] to validate our method and the joint PD controller parameters are set to be $k_p = 20.0$, $k_d = 0.5$. Parallel environments are trained on a terrain curriculum including slopes, stairs and discrete obstacles to enhance the robustness of the policy [23]. During training, the task mode variable x is randomly assigned as 1 or 2 every 6 seconds, thus the policy can undergo mode changes twice for each episode.

We design the initial state distribution delicately to boost the learning of the skill. As shown in Table IV, the randomization range of thigh joint angle and calf joint angle are wide to cover joint angle distribution in quadrupedal and bipedal locomotion. Base pitch angle is randomized in the range of [0,1.3] to foster the learning of bipedal locomotion.

TABLE V: Domain Randomization

Randomization Term	Range	Unit
Friction	[0.2, 1.6]	-
Payload mass	[-1, 1]	kg
Pushing robot	[-1.0, 1.0]	m/s
CoM bias	[-0.01, 0.01]	m
Joint damping	[0.25, 0.35]	N·m·s/rad
Joint friction	[0.005, 0.01]	-
Joint armature	[0.015, 0.025]	kg·m^2
Joint p gains	[0.8, 1.2] × nominal value	N·rad
Joint d gains	[0.8, 1.2] × nominal value	N·rad/s

To bridge the sim-to-real gap, we conduct system identification to better align the motor response between simulation and reality [10]. We suspend the quadruped robot and run a trained policy in the air to collect the motor response data at 200Hz. Then we utilize this deployment data to find joint friction, joint damping and joint armature parameters to match the motor response between simulation and reality best. After this real-to-sim calibration process, the domain randomization range becomes narrower and better matches the robot's motor response. To enhance the robustness of the estimator-policy set, we also apply other domain randomization terms, including friction, payload mass, CoM(center of mass) bias, pushing robots and joint PD gains. The randomized terms with their range are shown in Table V. The pushing robot term is applied every 10 seconds.

IV. Experimental Results

A. Ablation Study

In simulation, we conduct an ablation study to demonstrate the importance of task indicators and upright indicators in our proposed framework. We compare our method with two ablated versions:

- *w/o task indicator*: Task indicators I_i are ablated from rewards, which means the policy is instructed to learn to perform quadrupedal locomotion and bipedal locomotion simultaneously.
- *w/o upright indicator*: Upright indicator $\tilde{I}_{up}$ is ablated from bipedal rewards but task indicators are kept, which means the policy is stimulated to learn to track given velocity commands, perform periodic gait and get up simultaneously.

As shown in Fig. 2, we compared the quadruped tracking linear velocity reward and biped base angle reward across three conditions. We could see that quadruped tracking linear velocity reward of *w/o task indicator* was significantly higher because of the absence of task indicators as masks. However, the biped base angle reward of *w/o task indicator* could not rise in a long term, meaning the policy failed to learn bipedal locomotion, highlighting the importance of task indicators for learning mutually conflicting behaviors in a unified policy. Also, biped base angle reward of *w/o upright indicator* couldn't rise as the learning continued, which meant the presence of upright indicator was essential for learning bipedal and quadrupedal locomotion in one policy.

B. Performance of Bipedal and Quadrupedal Locomotion

We evaluated the performance of the quadruped robot in two locomotion modes across different terrains in the real world,

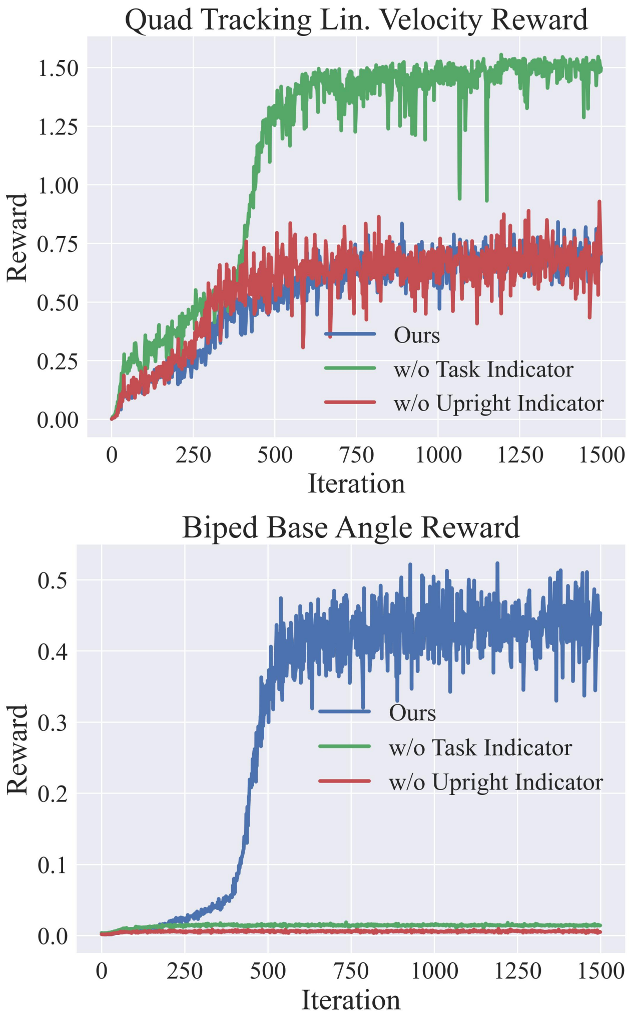

Fig. 2: Results of ablation study. Upper and lower subfigures show tracking linear velocity reward (quadruped) and base angle reward (biped) of three groups respectively. Data was collected under 4096 parallel environments in 1500 iterations.

including stairs, slopes and discrete obstacles, shown in Fig. 3 and the supplemented video. The results demonstrated the effectiveness of our approach in enabling the robot to handle complex real-world environments.

For quadrupedal locomotion, the robot exhibited outstanding performance across various terrains. The robot successfully navigated outdoor stairs with a height of 15 cm, achieving a 100% success rate in all trials, with consistent performance in all directions—highlighting its omnidirectional movement capabilities. Similarly, on indoor slopes with an incline of 30°, the robot maintained a 100% success rate for both ascending and descending, further highlighting the robustness of the quadrupedal gait.

For bipedal locomotion, performance of the same policy was also strong, though with some variations depending on the terrain and orientation. On 12 cm outdoor stairs, the robot achieved a 100% success rate in the backward locomotion for both ascending and descending. In contrast, the forward locomotion yielded a 20% success rate for ascending stairs, while maintaining a 100% success rate for descending, suggesting greater stability in the backward locomotion during ascent. When performing forward bipedal locomotion on stairs, the head of the robot was prone to hitting stair surfaces, which could be improved through specifying higher desired base height. Additionally, the robot demonstrated robust performance on indoor slopes of 30°, achieving a 100% success rate in both ascending and descending, confirming its effectiveness on inclined surfaces in bipedal mode. On discrete obstacles consisting of parts and weight plates, both locomotion styles could carry out robust locomotion to maintain stability. All of the tests on rough terrains were conducted 20 times and the average success rate was reported.

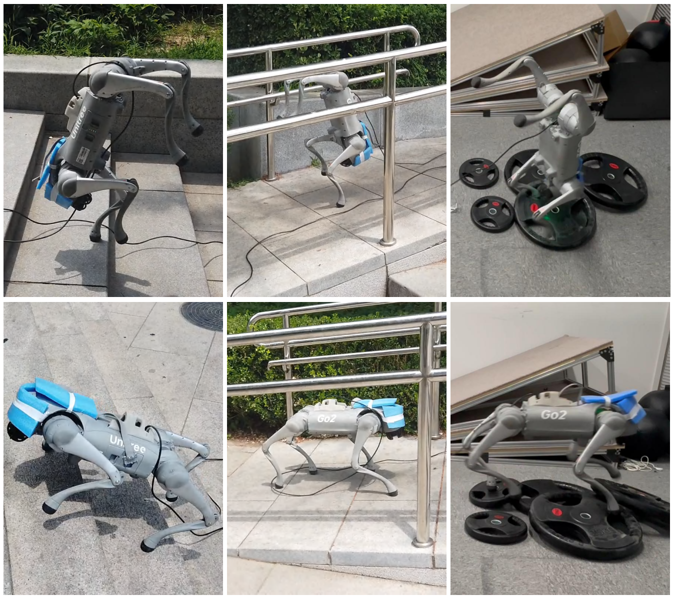

Fig. 3: Our policy deployed on a quadruped robot performs bipedal locomotion and quadrupedal locomotion over rough terrains. Upper row shows bipedal locomotion on stairs, slopes and discrete obstacles respectively. Lower row shows quadrupedal locomotion on stairs, slopes and discrete obstacles respectively.

These results validate the robustness of our method, as the robot consistently performed well across a range of terrains and conditions in both locomotion styles using one unified policy.

C. Transition between Bipedal and Quadrupedal Locomotion

A key feature of our approach is the ability to seamlessly transition between bipedal and quadrupedal locomotion, even under dynamic and challenging conditions. Fig. 4 shows the smooth transition movement. The average transition time is about 1 second. We evaluate the transition performance on both flat ground and sloped terrains.

- On flat ground, the robot accelerated to a speed of 1.5 m/s in quadrupedal mode and then instantly switched to bipedal mode with a 100% success rate. This demon-

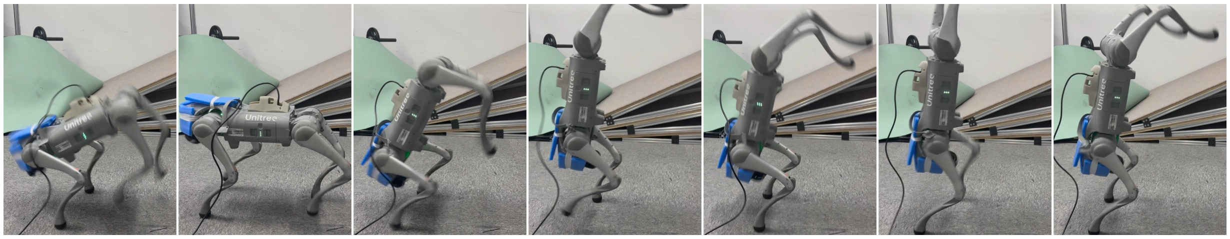

Fig. 4: Transition movement of the robot with our policy. It shows snapshots of the smooth transition from quadrupedal locomotion to bipedal locomotion achieved by the unified policy, which takes 1 second to complete.

TABLE VI: Success rates of two methods on various terrains with two kinds of locomotion. In the bracket, "q" stands for quadrupedal locomotion and "b" stands for bipedal locomotion. In the table, "-" means not being reported by the previous work.

Method	15cm Stairs up (q)	15cm Stairs down (q)	30° Slope (q)	30° Slope (b)	12cm Stairs up (b)	15 cm Stairs down (b)
Ours	1.0	**1.0**	1.0	1.0	**0.6**	**1.0**
MoELoco	1.0	-	1.0	1.0	-	0.95

strates the policy's capability to handle high-speed transitions without compromising stability.

- On a 10° outdoor slope, the robot successfully transitioned from quadrupedal to bipedal mode with a 100% success rate, highlighting its adaptability to complex terrains.

Furthermore, we compared the transition of our unified policy with a controller consisting of two policies, where each policy was trained to perform one kind of locomotion. The two policy controller showed undesired hitting-ground behavior when transitioning from biped to quadruped. On the contrary, our unified policy could execute smooth transition without collision between the base and the ground, as demonstrated in the supplemented video.

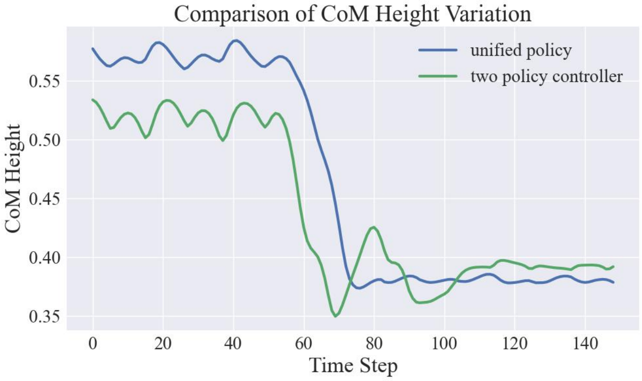

Fig. 5: Comparison of CoM height variation of two groups during the transition from bipedal locomotion to quadrupedal locomotion.

For quantitative analysis, we compared success rates of the transition from bipedal locomotion to quadrupedal locomotion in the simulation. We ran 4096 environments on the plane, which sampled velocity commands from the range of $[-0.5, 0.5]$ m/s. As a result, our unified policy achieved a success rate of 1.0, compared with 0.69 for the two policy controller. Besides, we compared the CoM height variation of two groups in the simulation, which was shown in Fig. 5. We could see that our unified policy achieved smoother transition with smaller CoM height oscillation.

The ability to perform instantaneous and reliable transitions in diverse scenarios is crucial for real-world applications, where the robot may need to quickly adapt its locomotion strategy in response to changing environments or task requirements.

D. Comparison with Previous Works

We compared our method with *MoELoco* [17], which incorporated quadrupedal and bipedal locomotion into one MoE policy. We compared the terrain traversal performance and training efficiency of two methods. Results for terrain traversal performance were summarized in Table VI. We conducted 20 trials for each kind of terrain and reported the average success rate. We could see that the success rate of our method was the same as *MoELoco* in many terrains. Besides, our method could enable bipedal locomotion to walk up 12cm stairs at a success rate of 0.6, which was not seen in *MoELoco*. For the terrain of 12cm stairs up (b), we conducted 10 forward tests and 10 backward tests. The success rates for forward tests and backward tests were 1.0 and 0.2 respectively, and we reported the average success rate of two kinds of tests. Furthermore, our method exceeds *MoELoco* in terms of training efficiency since our method only needs 4000 iterations for 4096 parallel environments to gain the ability but *MoELoco* needs nearly 120000 iterations for 4096 environments. Our method reduces the training time by 97% and still achieves comparable terrain traversal performance on stairs and slopes.

V. Conclusions and Future Works

In this work, we propose an approach for a quadruped robot to learn bipedal locomotion, quadrupedal locomotion and the transition between them through binary indicator guided reward functions. Our method enables a quadruped robot to perform robust quadrupedal locomotion, bipedal locomotion and the smooth transition during agile movement, achieving a stable transition even when running at a speed of 1.5 m/s. Our method is superior in terms of training efficiency and still possesses comparable terrain traversal performance on stairs and slopes compared with previous works.

There are several ways to extend the present study. Incorporating visual information into this framework and triggering different modes via human motion can be an interesting follow-up work. Besides, we can adapt the proposed method for a more capable humanoid robot to perform both quadrupedal and bipedal locomotion, which may bring humanoid robots superior stability on rough terrains.

References

[1] Benjamin Katz, Jared Di Carlo, and Sangbae Kim. Mini cheetah: A platform for pushing the limits of dynamic quadruped control. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6295–6301, 2019.

[2] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

[3] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020.

[4] Jonah Siekmann, Yesh Godse, Alan Fern, and Jonathan Hurst. Sim-to-real learning of all common bipedal gaits via periodic reward composition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7309–7315, 2021.

[5] Jonah Siekmann, Kevin Green, John Warila, Alan Fern, and Jonathan Hurst. Blind Bipedal Stair Traversal via Sim-to-Real Reinforcement Learning. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[6] Patrick M. Wensing, Michael Posa, Yue Hu, Adrien Escande, Nicolas Mansard, and Andrea Del Prete. Optimization-based control for dynamic legged robots. *IEEE Transactions on Robotics*, 40:43–63, 2024.

[7] Taisuke Kobayashi, Tadayoshi Aoyama, Masafumi Sobajima, Kosuke Sekiyama, and Toshio Fukuda. Locomotion selection strategy for multi-locomotion robot based on stability and efficiency. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2616–2621, 2013.

[8] Chen Yu and Andre Rosendo. Multi-modal legged locomotion framework with automated residual reinforcement learning. *IEEE Robotics and Automation Letters*, 7(4):10312–10319, 2022.

[9] Laura Smith, J. Chase Kew, Tianyu Li, Linda Luu, Xue Bin Peng, Sehoon Ha, Jie Tan, and Sergey Levine. Learning and adapting agile locomotion skills by transferring experience, 2023.

[10] Yunfei Li, Jinhan Li, Wei Fu, and Yi Wu. Learning agile bipedal motions on a quadrupedal robot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9735–9742, 2024.

[11] Tianhu Peng, Lingfan Bao, Joseph Humphreys, Andromachi Maria Delfaki, Dimitrios Kanoulas, and Chengxu Zhou. Learning bipedal walking on a quadruped robot via adversarial motion priors, 2024.

[12] Zhi Su, Xiaoyu Huang, Daniel Ordoñez-Apraez, Yunfei Li, Zhongyu Li, Qiayuan Liao, Giulio Turrisi, Massimiliano Pontil, Claudio Semini, Yi Wu, and Koushil Sreenath. Leveraging symmetry in rl-based legged locomotion control. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 6899–6906. IEEE, October 2024.

[13] Gijeong Kim, Yong-Hoon Lee, and Hae-Won Park. A learning framework for diverse legged robot locomotion using barrier-based style rewards, 2025.

[14] Lorenzo Amatucci, Giulio Turrisi, Angelo Bratta, Victor Barasuol, and Claudio Semini. Accelerating model predictive control for legged robots through distributed optimization. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 12734–12741. IEEE, October 2024.

[15] Gijeong Kim, Dongyun Kang, Joon-Ha Kim, Seungwoo Hong, and Hae-Won Park. Contact-implicit model predictive control: Controlling diverse quadruped motions without pre-planned contact modes or trajectories. *The International Journal of Robotics Research*, 44(3):486–510, October 2024.

[16] Takumi Kamioka, Tomoki Watabe, Masao Kanazawa, Hiroyuki Kaneko, and Takahide Yoshiike. Dynamic gait transition between bipedal and quadrupedal locomotion. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2195–2201, 2015.

[17] Runhan Huang, Shaoting Zhu, Yilun Du, and Hang Zhao. Moe-loco: Mixture of experts for multitask locomotion, 2025.

[18] S. Kajita, F. Kanehiro, K. Kaneko, K. Yokoi, and H. Hirukawa. The 3d linear inverted pendulum mode: a simple modeling for a biped walking pattern generation. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)*, volume 1, pages 239–246 vol.1, 2001.

[19] Salman Faraji, Soha Pouya, Christopher G. Atkeson, and Auke Jan Ijspeert. Versatile and robust 3d walking with a simulated humanoid robot (atlas): A model predictive control approach. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1943–1950, 2014.

[20] Kenji Kaneko, Fumio Kanehiro, Mitsuharu Morisawa, Kazuhiko Akachi, Go Miyamori, Atsushi Hayashi, and Noriyuki Kanehira. Humanoid robot hrp-4 - humanoid robotics platform with lightweight and slim body. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4400–4407, 2011.

[21] Suyoung Choi, Gwanghyeon Ji, Jeongsoo Park, Hyeongjun Kim, Juhyeok Mun, Jeong Hyun Lee, and Jemin Hwangbo. Learning quadrupedal locomotion on deformable terrain. *Science Robotics*, 8(74):eade2256, 2023.

[22] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11996–12007. Curran Associates, Inc., 2020.

[23] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 91–100. PMLR, 08–11 Nov 2022.

[24] Gwanghyeon Ji, Juhyeok Mun, Hyeongjun Kim, and Jemin Hwangbo. Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion. *IEEE Robotics and Automation Letters*, 7(2):4630–4637, 2022.

[25] Sarah Harris and David Harris. *Digital Design and Computer Architecture, RISC-V Edition*. Morgan Kaufmann, 2021.

[26] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[28] Unitree go2. https://www.unitree.com/go2. Accessed: Feb. 2025.