

MASQ: Multi-Agent Reinforcement Learning for Single Quadruped Robot Locomotion

Qi Liu[†], Jingxiang Guo[†], Sixu Lin, Shuaikang Ma, Jinxuan Zhu, Yanjie Li^{*}

Abstract—This paper proposes a novel method to improve locomotion for a single quadruped robot using multi-agent deep reinforcement learning (MARL). Many existing methods use single-agent reinforcement learning for an individual robot or MARL for the cooperative task in multi-robot systems. Unlike existing methods, this paper proposes using MARL for the locomotion of a single quadruped robot. We propose a learning structure called Multi-Agent Reinforcement Learning for Single Quadruped Robot Locomotion (MASQ), considering each leg as an agent to explore the action space of the quadruped robot, sharing a global critic, and learning cooperatively. Experimental results show that MASQ not only speeds up learning convergence but also enhances robustness in real-world settings, suggesting that applying MASQ to single robots such as quadrupeds could surpass traditional single-robot reinforcement learning approaches. Our study provides insightful guidance on integrating MARL with single-robot locomotion.

I. INTRODUCTION

Reinforcement learning (RL) has made remarkable progress in various robot control learning [1], such as quadruped robot [2], [3], biped robot [4], [5], and unmanned aerial vehicle [6]. This paper focuses on the quadruped robot control learning.

In the process of applying deep RL to a single robot, it is prevalent to use single-agent algorithms [7]–[10]. However, single-agent algorithms may have limitations in managing coordination in specific problems. Many existing methods use single-agent RL algorithms for individual robot learning or multi-agent deep reinforcement learning (MARL) for multi-robot systems in cooperative tasks [11]–[13]. Cooperative MARL algorithms have been widely demonstrated in game-AI [14], [15] to have advantages in multi-agent cooperation.

Unlike existing approaches, this paper proposes using MARL for the locomotion of a single quadruped robot to enhance cooperation between its four legs, thereby enabling it to navigate complex terrains and perform intricate tasks. By proposing cooperative MARL, where each leg acts as an agent, the quadruped robot can better coordinate its movements. This collaborative learning structure, termed Multi-

This work was supported by the National Natural Science Foundation of China [61977019, U1813206] and Shenzhen Fundamental Research [JCYJ20220818102415033, JSBG20201103093802006, KJZD20230923114222045]. (Corresponding author: Yanjie Li, autolyj@hit.edu.cn)

The authors are with the Guangdong Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics and the School of Mechanical Engineering and Automation, the Harbin Institute of Technology Shenzhen, 518055, China.

* Corresponding author.

[†] Equal contribution.

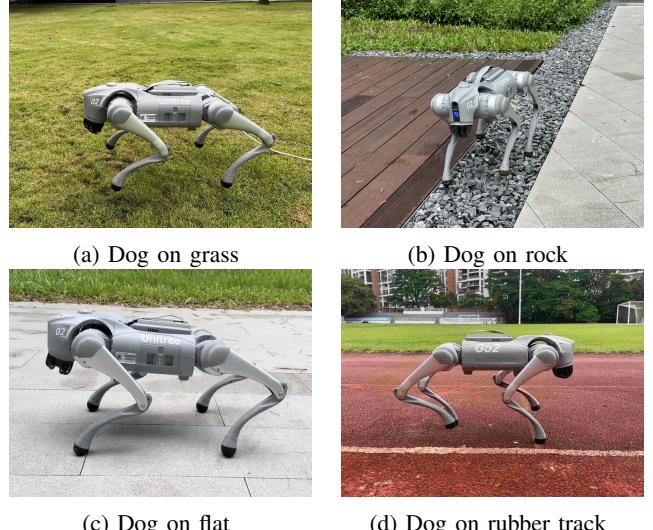


Fig. 1: Quadruped robot on various terrains

Agent Reinforcement Learning for Single Quadruped Robot Locomotion (MASQ), allows the robot to share experiences. Fig. 1 presents the deployment of the MASQ algorithm on a quadruped robot tested across different terrains, including grass, rocks, flat surfaces, and rubber tracks, demonstrating its performance in diverse environments. Fig. 2 shows a sim-to-real comparison of trot gaits in a quadruped robot, highlighting the consistency between simulated and real-world gait patterns.

The main contributions of this paper can be summarized as follows:

- This paper proposes MASQ, a method that treats each leg of a quadruped robot as an individual agent. The locomotion learning is modeled as a cooperative multi-agent reinforcement learning (MARL) problem and solved using a MARL algorithm.
- Experimental results show that the proposed method enhances performance in executing gaits, improves training efficiency and robustness, and achieves better final performance, demonstrating the value and impact of the approach.

II. RELATED WORK

A. Deep RL for Single Robot Control

Recent advances in deep RL for quadruped robots are driven by simulation technologies such as Isaac Gym [16],

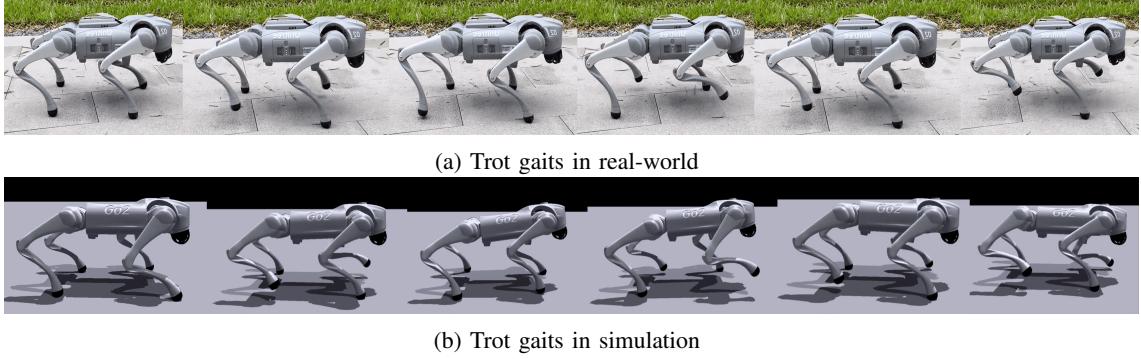


Fig. 2: Sim-to-Real comparison of trot gaits

[17]. Hardware advancements have shown robust performance on various tasks through Sim2Real transfer with zero shot [18]–[21]. Current research focuses on task-specific reward composition and training paradigms to bridge the Sim2Real gap [22]–[24], often using Proximal Policy Optimization (PPO) [25] with an emphasis on task and reward design rather than novel RL algorithms. Although some studies have explored learning algorithms to improve efficiency [26]–[31], most efforts remain centered on efficiency rather than modifying algorithms to control the characteristics of the object. Additionally, research has delved into task-specific strategies like rapid motor adaptation (RMA) [22], hierarchical RL for multi-skill tasks [32], and symmetry-based data enhancement [33], [34], yet challenges remain in using robot symmetry through single-agent methods, indicating that algorithmic research on this aspect is still underdeveloped.

B. MARL for Multi-robot Control

In multi-agent settings, algorithms like Multi-Agent Proximal Policy Optimization (MAPPO) [35], Temporally Extended Multi-Agent Reinforcement Learning (TEMP) [36] have demonstrated strong capabilities in addressing multi-agent robotic challenges, such as drone fleet control [37], [38] and autonomous vehicle fleets [39], [40]. Methods like Reinforced Inter-Agent Learning (RIAL) and Differentiable Inter-Agent Learning (DIAL) [41] further enhance collaborative performance by developing communication protocols among agents. This paper proposes modeling single-quadruped robot locomotion as a cooperative MARL problem, where each leg is treated as an independent agent, contrasting with previous approaches that treat the robot as a monolithic entity [42]–[45] or cooperative groups of multi-robots [12], [46], [47].

III. PRELIMINARIES

This paper considers a finite-horizon Markov decision process (MDP) [48], defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, T)$. \mathcal{S} denotes the state space, $\mathcal{A} := \{a_0, a_1, \dots, a_{|\mathcal{A}|-1}\}$ represents a finite action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ represents the state transition distribution, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ denotes a reward function, $\gamma \in [0, 1]$ denotes a discount factor, and T is a time horizon. At each time step t , the policy π selects an

action $a_t \in \mathcal{A}$. After entering the next state by sampling from $\mathcal{P}(s_{t+1} | s_t, a_t)$, the agent receives an immediate reward $r(s_t, a_t)$. The agent continues to perform actions until it enters a terminal state or t reaches the time horizon T . RL aims to learn a policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ for decision-making problems by maximizing discounted rewards. For any policy π , the state-action value function (Q function) is defined as

$$Q^\pi(s, a) = \mathbb{E}^\pi \left[\sum_{t=0}^T \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right] \quad (1)$$

Proximal Policy Optimization (PPO) [25] enhances the stability and performance of policy gradient methods by limiting policy updates to prevent destabilizing deviations. The core PPO update rule optimizes a clipped surrogate function:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (2)$$

where

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \quad (3)$$

and

$$\hat{A}_t = Q^\pi(s_t, a_t) - V_\psi(s_t) \quad (4)$$

where ψ denotes the parameters of value function (V_ψ) network, ϵ denotes a coefficient. Policy parameters θ are updated as:

$$\theta \leftarrow \theta + \alpha \nabla_\theta L^{CLIP}(\theta) \quad (5)$$

PPO's constrained updates stabilize training and improve performance, making it practical for complex single-agent RL tasks.

IV. MULTI-AGENT REINFORCEMENT LEARNING FOR SINGLE QUADRUPED ROBOT LOCOMOTION

In this paper, we use the collaborative potential of multiple agents to improve the learning process for a single robot's locomotion, resulting in faster training convergence, robustness and better final performance in real-world environments. Specifically, each leg of the quadruped robot is treated as a separate agent within the multi-agent structure, with individual observations and a shared global critic, significantly improving the cooperation among the robot's limbs for more effective locomotion.

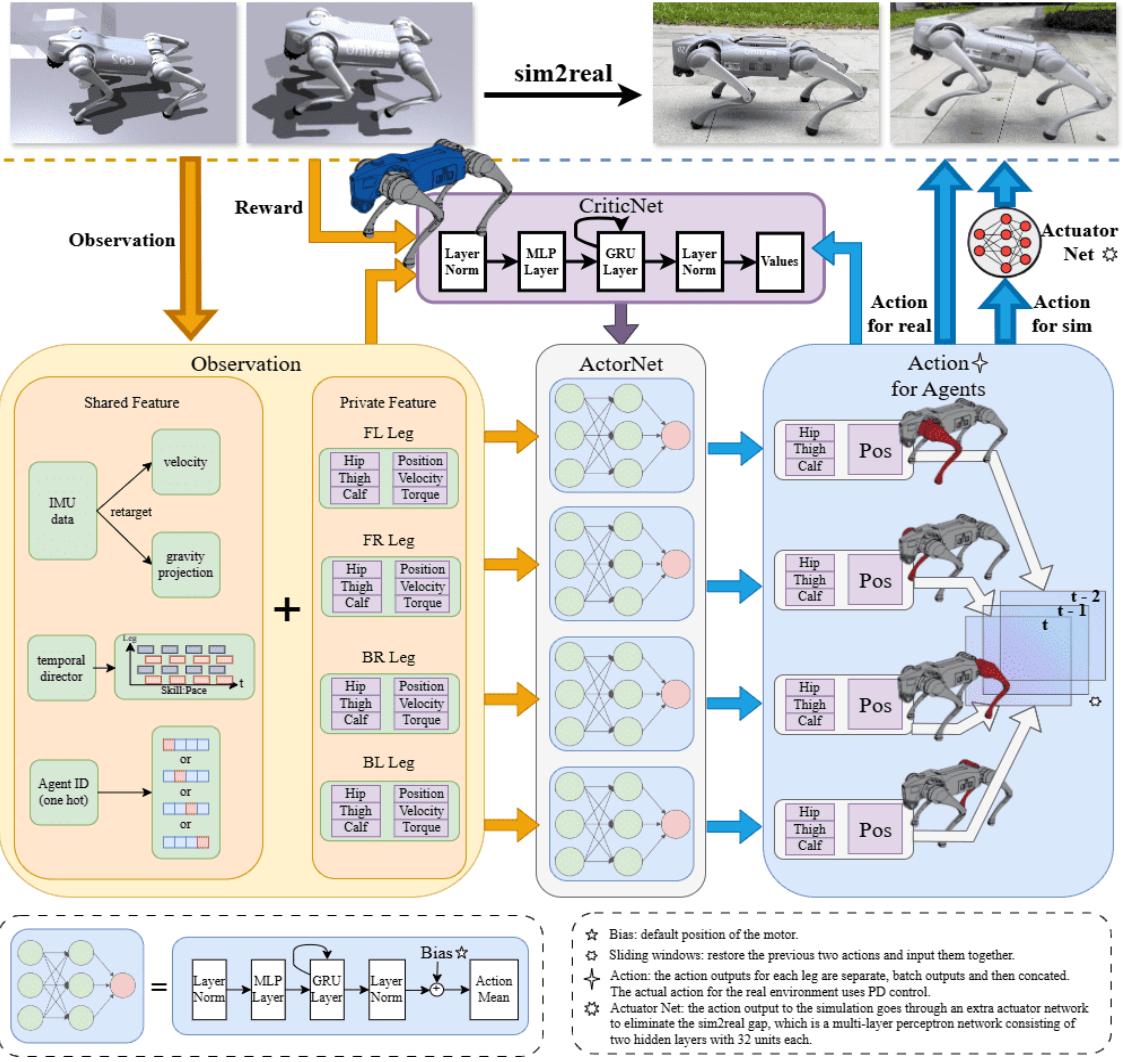


Fig. 3: The framework of MASQ

A. MASQ Modeling

This paper models a single quadruped robot locomotion as a cooperative multi-agent problem, which is described as a partially observable decentralized Markov decision process (decPOMDP) [49]. The decPOMDP is defined by the tuple $G = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathcal{Z}, \mathcal{O}, N, \gamma, T)$. \mathcal{S} is the state space, \mathcal{A} is the action space, \mathcal{P} is the state transition distribution, r is the reward function, \mathcal{Z} is the observation space, \mathcal{O} is the observation function, N is the number of agents, γ is the discount factor, and T is the time horizon. At each time step t , each agent $n \in \{1, \dots, N\}$ selects an action $a_t^n \in \mathcal{A}$, resulting in a joint action $\mathbf{a}_t = \{a_t^1, a_t^2, \dots, a_t^N\}$. The environment transitions to a new state s_{t+1} according to $\mathcal{P}(s_{t+1}|s_t, \mathbf{a}_t)$ and provides a shared reward $r(s_t, \mathbf{a}_t)$. Each agent receives an observation z_t^n from $\mathcal{O}(s_t, n)$ and maintains an observation-action history τ_t^n . MARL aims to learn policies $\{\pi^n\}_{n=1}^N$ that maximize expected cumulative rewards:

$$J(\pi) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r(s_t, \mathbf{a}_t) \right] \quad (6)$$

This paper proposes Multi-Agent Reinforcement Learning for Single Quadruped Robot Locomotion (MASQ), which applies MARL principles to treat different parts of a single quadruped robot as independent agents, trained collaboratively using shared rewards. Specifically, this paper uses MAPPO [35] to solve the modeled multi-agent problem. MAPPO optimizes the following objective function in a multi-agent context:

$$L_{MAPPO}^{CLIP}(\theta) = \sum_{i=1}^n \mathbb{E}_t \left[\min \left(r_t^i(\theta_i) \hat{A}_t, \text{clip}(r_t^i(\theta_i), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (7)$$

where

$$r_t^i(\theta_i) = \frac{\pi_{\theta_i}(a_t^i | o_t^i)}{\pi_{\theta_{i,\text{old}}}(a_t^i | o_t^i)} \quad (8)$$

where i denote the i -th agent in MARL. Each agent updates its policy parameters as follows:

$$\theta_i \leftarrow \theta_i + \alpha \nabla_{\theta_i} L_i^{CLIP}(\theta_i) \quad (9)$$

This paper uses centralized training with decentralized execution (CTDE) [50] to handle multi-agent learning challenges. This approach maintains stability in a constantly changing environment. In implementations, separate neural networks learn a policy (π_θ) and a value function ($V_\psi(s)$) similar to the single-agent version. The value function helps reduce training variability and can take in additional global information. This approach leads to better final performance, faster learning speed, and improved robustness in deployment.

B. MASQ Settings

State Space and Observation: The shared actor network takes a concatenated observation from four agents, each with 35 dimensions, including motor positions $q_t \in \mathbb{R}^3$, motor speeds $\dot{q}_t \in \mathbb{R}^3$, previous actions $a_{t-1} \in \mathbb{R}^3$ and $a_t \in \mathbb{R}^3$, a gait sequencing director $d_t \in \mathbb{R}^1$, projected gravity $g_t \in \mathbb{R}^3$, command values $v_t^{cmd} \in \mathbb{R}^{15}$, body speeds $v_b \in \mathbb{R}^3$, and a one-hot encoding $e_t \in \mathbb{R}^1$. The total input for the actor network is $o_t^{\text{actor}} \in \mathbb{R}^{140}$ (35x4). The architecture of the actor network consists of a normalization layer, followed by an MLP layer, a GRU layer, and another normalization layer, with the final output being the joint angle commands for each leg. Additionally, a bias is applied to the output for better precision in control. On the other hand, the critic network uses global observations and has a 73-dimensional input, including motor positions $q_t \in \mathbb{R}^{12}$, motor speeds $\dot{q}_t \in \mathbb{R}^{12}$, previous actions $a_{t-1} \in \mathbb{R}^{12}$ and $a_t \in \mathbb{R}^{12}$, gait directors $d_t \in \mathbb{R}^4$, projected gravity $g_t \in \mathbb{R}^3$, command values $v_t^{cmd} \in \mathbb{R}^{15}$, and body speeds $v_b \in \mathbb{R}^3$. The output of the critic network consists of continuous V values $V_t \in \mathbb{R}^4$, which are used to calculate the advantage function.

Action Space: The output of the actor-network consists of continuous actions $a_t \in \mathbb{R}^{12}$, and the system then uses these to calculate the torques for the 12 motors. Details can be found in Section IV-D.

Reward Function: The reward functions in Table I are designed to optimize the robot's performance by encouraging desired behaviors and penalizing undesired ones. Key rewards include: *tracking linear velocity*, which uses an exponential decay function to minimize velocity error; *linear velocity Z* and *angular velocity XY*, both penalizing unwanted motions to ensure stability; *torques*, *DOF velocity*, and *DOF acceleration*, which promote energy efficiency and smoother movements; and *collision*, which penalizes excessive contact forces. Additional rewards focus on action smoothness, accurate jumping, minimizing foot slip and impact velocities, and enhancing locomotion stability using Raibert's heuristic [51] and foot velocity tracking.

C. Multi-agent Actor and Global Critic Networks

In the simulation environment of Isaac Gym [16], the robot receives observations and rewards to facilitate its learning. The learning process involves dividing the observations into shared and private features, and both the actor and critic networks use the rewards to train the policy.

TABLE I: Reward function

REWARD SETTINGS, CORRESPONDING EQUATIONS, AND THEIR SCALES		
Reward Term	Equation	Scale
Tracking Linear Velocity	$\exp\left(-\frac{\ \mathbf{v}_{cmd} - \mathbf{v}_b\ ^2}{\sigma_v}\right)$	1.0
Tracking Angular Velocity	$\exp\left(-\frac{(\omega_{cmd,z} - \omega_{b,z})^2}{\sigma_{yaw}}\right)$	0.5
Linear Velocity Z	$\ \mathbf{v}_{b,z}\ ^2$	-2×10^{-2}
Angular Velocity XY	$\sum_i \ \omega_i\ ^2$	-1×10^{-3}
Angular Velocity Torques	$\sum_i \ \tau_i\ ^2$	-1×10^{-5}
DOF Velocity	$\sum_i \ \mathbf{v}_{d,i}\ ^2$	-1×10^{-4}
DOF Acceleration	$\sum_i \left(\frac{\mathbf{v}_{d,i,t} - \mathbf{v}_{d,i,t-1}}{\Delta t}\right)^2$	-2.5×10^{-7}
Collision	$\sum_i (1.0 \cdot (\ \mathbf{f}_i\ > 0.1))$	-5.0
Action Rate	$\sum_i \ \mathbf{a}_t - \mathbf{a}_{t-1}\ ^2$	-1×10^{-2}
Jump	$-(h_b - h_j)^2$	10.0
Feet Slip	$\sum_i (c_s \cdot \ \mathbf{v}_f\ ^2)$	-4×10^{-2}
Action Smoothness 1	$\sum_i (\mathbf{a}_t - \mathbf{a}_{t-1})^2$	-0.1
Action Smoothness 2	$\sum_i (\mathbf{a}_t - 2\mathbf{a}_{t-1} + \mathbf{a}_{t-2})^2$	-0.1
Feet Impact Velocity	$\sum_i (c_s \cdot \ \mathbf{v}_{f,p}\ ^2)$	-0.0
Raibert Heuristic	$\sum_i \ \mathbf{e}_r\ ^2$	-10.0
Tracking Contacts Shaped Velocity	$\sum_i \left(c_d \cdot \left(1 - \exp\left(-\frac{\ \mathbf{v}_f\ ^2}{\sigma_{gv}}\right)\right)\right)$	4.0

The actor-network consists of a multi-layer perception (MLP) base and an activation layer that produces actions and their associated log probabilities. Similarly, the critic network uses an MLP base, ending in an output layer that predicts value functions. The actor-network outputs actions based on observations, while the critic network assesses the value of these actions to guide the learning process. The actions produced for the agents are processed by an actuator network [52] to simulate real-world conditions, enhancing deployment effectiveness in natural environments. After training, the trained actor-network is deployed onto the robotic dog to perform actions directly. Fig. 3 illustrates the entire learning process.

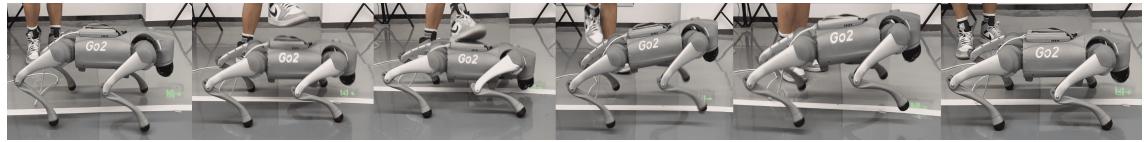
In the context of a quadruped robot, we consider each leg as an individual agent. All four agents share a standard actor-network. Using a shared-parameter network instead of four separate actor networks helps reduce computational load and better fits the nature of a quadruped robot. Unlike typical multi-agent environments, such as StarCraft [14], where each soldier is an independent agent, the quadruped robot is a single entity with four legs symmetrically positioned around the body's center. Therefore, a shared-parameter actor-network is more suitable for this scenario.

We express the policy for each leg as follows:

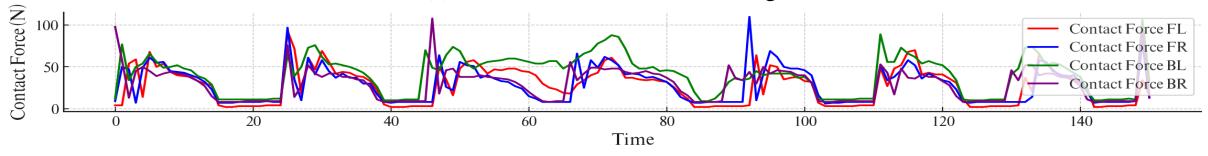
$$\pi_{\theta_i}(a_{i,t} | s_{i,t}) = \pi_\theta(a_{i,t} | s_{i,t}) \quad (10)$$

where $i = 1, 2, 3, 4$ corresponds to the four legs.

The quadruped robot has four legs, each with three motors: hip, thigh, and calf joint motors. Each motor's position, velocity, and torque are observable, so we use these details as the independent observations for each agent. Additionally, to enhance the coordination among the agents, we augment each independent observation with shared observations, including speed and gravity projections calculated from inertial measurement unit (IMU) data, temporal director, and agent identifier (ID). The temporal director $T_i(t)$ guides the gait sequence of each leg under different movement postures, while the agent ID is necessary for the shared-parameter



(a) Robustness in outside disturbing



(b) Contact force in different legs while encountering force

Fig. 4: Robustness test

network. This setup ensures the independence of each agent while improving their cooperative capabilities. The temporal director helps to synchronize the movements of different legs, ensuring smooth gait patterns. It can be defined as:

$$T_i(t) = \sin(2\pi(kt + \Delta_i)) \quad (11)$$

where

- k is the scaling factor of scaling factor of gait cycle.
- Δ_i is the phase offset for the i -th leg, which determines its relative timing within the gait cycle to ensure coordinated movement.

The Global Critic uses a centralized value function approach to consider global information, which fits into the CTDE. It relies on a global value function to coordinate individual agents, such as single-agent algorithms like PPO. The Critic network takes in global observations, which consist of the separate observations of the four agents and shared global information.

During the training in the simulation environment, we used a multi-gait curriculum learning [53]. This curriculum comprises four gaits: pace, trot, bound, and pronk. In the simulation, the quadruped robot learns these four gaits simultaneously, and the progress is updated based on evaluating whether the gaits' rewards meet specific thresholds. This method allows the robot to learn different gaits effectively and is also helpful in testing the generalization capabilities of our proposed approach across various tasks in experimental settings.

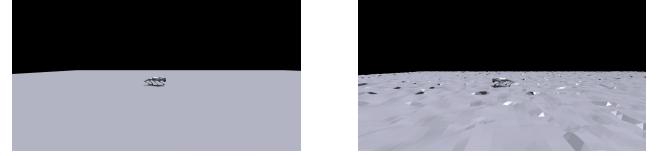
D. Sim-to-real

To bridge the gap between simulation and real-world performance, we used domain randomization and an Actuator Network [52]. Domain randomization involves randomizing various parameters to train a robust policy under different conditions [30], [54], [55]. These parameters include robot body mass, motor strength, joint position calibration, ground friction, restitution, orientation, and magnitude of gravity. We also independently randomize friction coefficients like ground friction. Gravity is randomized every 8 seconds with a gravity impulse duration of 0.99 seconds. Time steps are randomized every 6 seconds, with the overall randomization occurring every 4 seconds. These measures enhance the

model's robustness and adaptability. The training process for the Actuator Network captures the non-ideal relationship between PD error and the torque realized [52], thereby improving the model's performance in real-world applications.

V. EXPERIMENTS

We conducted experiments using the Unitree Go2 quadruped robot with various experiments. Section V-A proposes the experimental setup. Section V-B shows the experimental results of the simulation. Section V-C shows the real-world experiments and comparisons.



(a) Flat terrain

(b) Uneven terrain

Fig. 5: Various terrains

A. Experiments Setup

In the simulation environment, we designed two types of terrain: flat and uneven, shown in Fig. 5. Uneven terrain is a composite of pyramid-sloping terrain and random uniform terrain. Fifteen commands, such as linear velocity, angular velocity and height of the base, sampled within a specified range, guide the gait-learning process. The robot learns to adapt and develop various skills by following these commands.

B. Simulation Experiments

The MASQ was compared with the PPO student-teacher baseline in simulation environments, including two types: flat and uneven terrain. The quadruped robot was trained using the curriculum learning method, with hyperparameters such as the learning rate tuned for optimal performance in both scenarios. Reward curves in Fig. 6 show that the proposed method achieves faster convergence and better final performance compared to the PPO baseline. Experimental results are presented as the mean of rewards over five tests for a fair comparison.

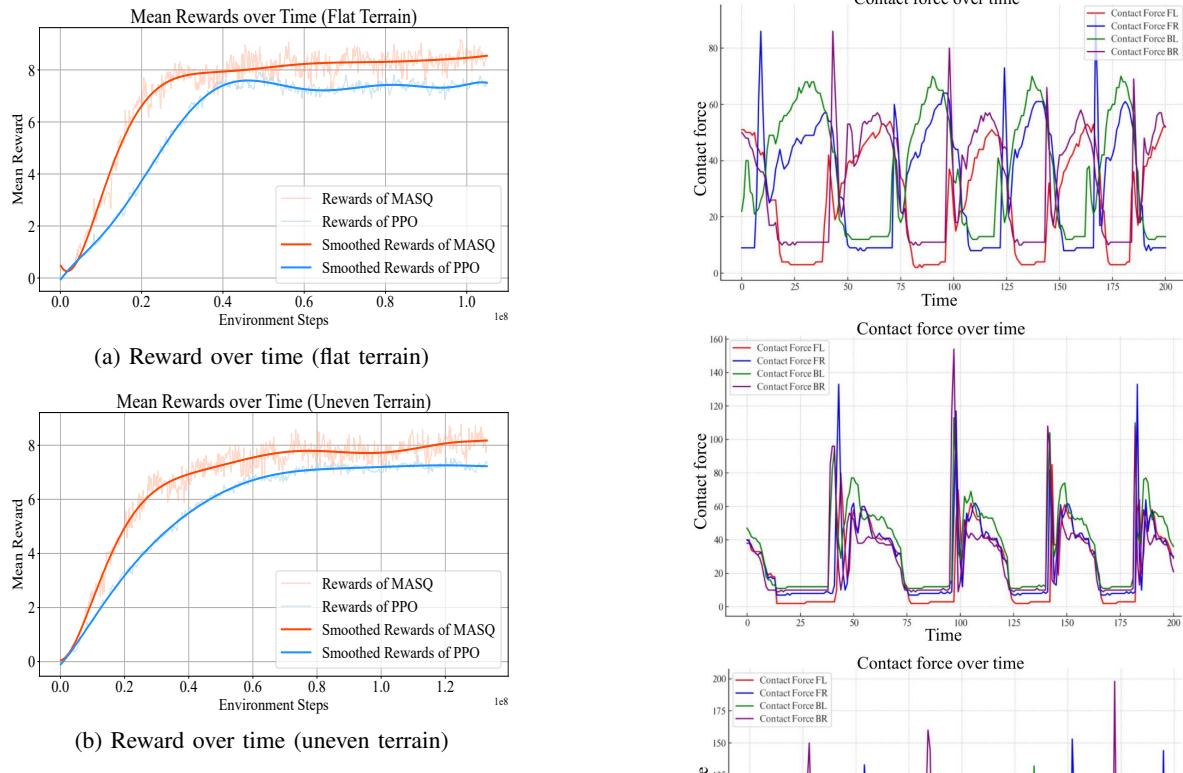


Fig. 6: Return on various terrains

C. Real-word Experiments

This paper deploys MASQ on a quadruped robot and tests it on various terrains, including flat ground, grass, and sand in Fig. 1, where it performs excellently. In addition, we conducted a series of robustness tests for heavy impacts and side kicks. The robot can quickly return to its normal state after being disturbed.

1) **Gaits test:** Fig. 7 illustrates the periodic relationship of quadruped force feedback for each of the four legs under the training of four different skills. They reflect the impact of incorporating the temporal director in our observations on the learning and switching of other skills.

2) **Robustness test:** To validate the robustness of the gaits trained with our method, we conduct external disturbance tests on the robot. We test the robot in a bounding gait with a cycle period of 20ms. During robustness tests, the robot performs continuous jumps in this bounding gait. The impacts of human steps on the robot are applied to propose disturbances during its jump cycle. Under normal conditions, the robot landed simultaneously on all four legs. We monitored this process by recording the force sensor values in the robot's feet, thereby documenting the transition from a normal state through the induced disturbance and back to the normal state. As shown in Fig. 4, the robot returns to its normal state after experiencing a disturbance within just one gait cycle.

VI. CONCLUSION AND FUTURE WORK

This paper proposed a multi-agent-based motion control system for quadruped robots, utilizing a shared-parameter

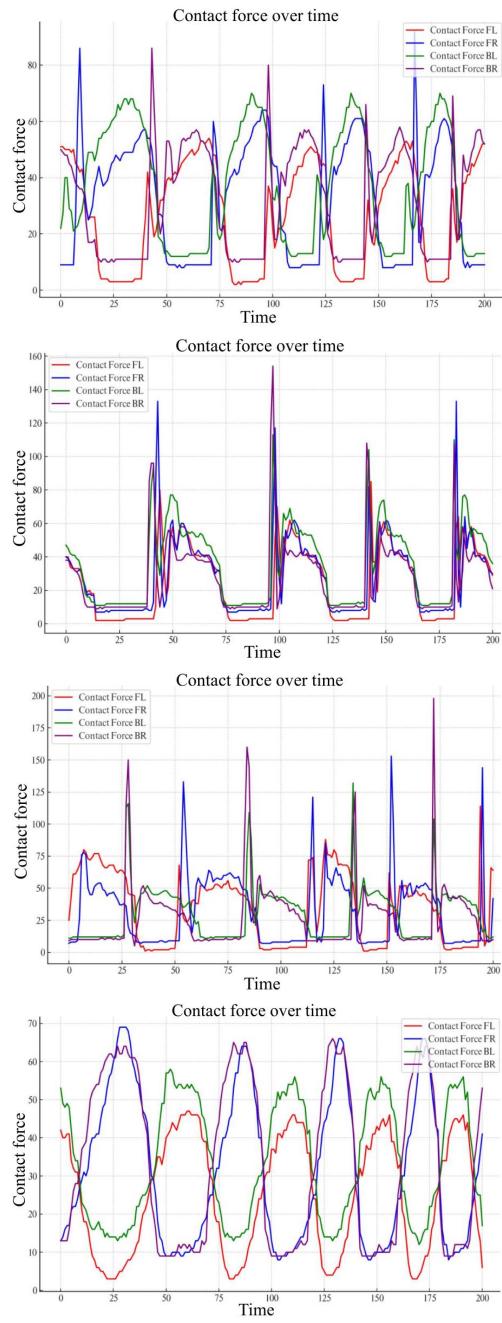


Fig. 7: Contact force of four legs in different gaits

actor network and a centralized critic network within the CTDE framework. The proposed approach, implemented in Isaac Gym, demonstrated substantial improvements in training speed, robustness, and final performance, benefiting from curriculum learning and domain randomization. These advances enabled efficient limb coordination and smoother sim-to-real transitions. Experimental results confirmed the effectiveness of the method in enhancing both performance and efficiency in motion control for symmetric robots. Future work will extend the approach to other symmetric robots and explore its application in more complex dynamic environments.

REFERENCES

- [1] Y. Hou, H. Sun, J. Ma, and F. Wu, "Improving offline reinforcement learning with inaccurate simulators," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5162–5168.
- [2] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*. PMLR, 2023, pp. 22–31.
- [3] I. M. Aswin Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5078–5084.
- [4] H. Benbrahim and J. A. Franklin, "Biped dynamic walking using reinforcement learning," *Robotics and Autonomous systems*, vol. 22, no. 3-4, pp. 283–302, 1997.
- [5] H. Duan, B. Pandit, M. S. Gaddé, B. Van Marum, J. Dao, C. Kim, and A. Fern, "Learning vision-based bipedal locomotion for challenging terrain," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 56–62.
- [6] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2315–2322, 2018.
- [7] B. Jia and D. Manocha, "Sim-to-real robotic sketching using behavior cloning and reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 18 272–18 278.
- [8] A. Lobbezoo, Y. Qian, and H.-J. Kwon, "Reinforcement learning for pick and place operations in robotics: A survey," *Robotics*, vol. 10, no. 3, p. 105, 2021.
- [9] H. Jiang, T. Chen, J. Cao, J. Bi, G. Lu, G. Zhang, X. Rong, and Y. Li, "Sim-to-real: Quadruped robot control with deep reinforcement learning and parallel training," in *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2022, pp. 489–494.
- [10] S. Lyu, H. Zhao, and D. Wang, "A composite control strategy for quadruped robot by integrating reinforcement learning and model-based control," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 751–758.
- [11] L. Canese, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Re, and S. Spanò, "Multi-agent reinforcement learning: A review of challenges and applications," *Applied Sciences*, vol. 11, no. 11, p. 4948, 2021.
- [12] J. Orr and A. Dutta, "Multi-agent deep reinforcement learning for multi-robot applications: A survey," *Sensors*, vol. 23, no. 7, p. 3625, 2023.
- [13] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 5150–5163.
- [14] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.
- [15] D. Ye, G. Chen, P. Zhao, F. Qiu, B. Yuan, W. Zhang, S. Chen, M. Sun, X. Li, S. Li, J. Liang, Z. Lian, B. Shi, L. Wang, T. Shi, Q. Fu, W. Yang, and L. Huang, "Supervised learning achieves human-level performance in moba games: A case study of honor of kings," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 3, pp. 908–918, 2022.
- [16] 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)*, 2021.
- [17] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 91–100.
- [18] 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, Jan. 2019.
- [19] 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.
- [20] 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, p. eabk2822, 2022.
- [21] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, and I. Havoutis, "Real-time trajectory adaptation for quadrupedal locomotion using deep reinforcement learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5973–5979.
- [22] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *Robotics: Science and Systems XVII*, 2021.
- [23] A. Kumar, Z. Li, J. Zeng, D. Pathak, K. Sreenath, and J. Malik, "Adapting rapid motor adaptation for bipedal robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 1161–1168.
- [24] S. Choi, G. Ji, J. Park, H. Kim, J. Mun, J. H. Lee, and J. Hwangbo, "Learning quadrupedal locomotion on deformable terrain," *Science Robotics*, vol. 8, no. 74, p. eade2256, 2023.
- [25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [26] Y. Kim, H. Oh, J. Lee, J. Choi, G. Ji, M. Jung, D. Youm, and J. Hwangbo, "Not only rewards but also constraints: Applications on legged robot locomotion," *IEEE Transactions on Robotics*, 2024.
- [27] L. Ye, J. Li, Y. Cheng, X. Wang, B. Liang, and Y. Peng, "From knowing to doing: Learning diverse motor skills through instruction learning," *arXiv preprint arXiv:2309.09167*, 2023.
- [28] B. L. Semage, T. G. Karimpanal, S. Rana, and S. Venkatesh, "Zero-shot sim2real adaptation across environments," *arXiv preprint arXiv:2302.04013*, 2023.
- [29] X. Gu, Y.-J. Wang, and J. Chen, "Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer," *arXiv preprint arXiv:2404.05695*, 2024.
- [30] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2018.
- [31] B. Wang, Z. Liu, Q. Li, and A. Prorok, "Mobile robot path planning in dynamic environments through globally guided reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6932–6939, 2020.
- [32] X. Zhou, X. Wen, Z. Wang, Y. Gao, H. Li, Q. Wang, T. Yang, H. Lu, Y. Cao, C. Xu, and F. Gao, "Swarm of micro flying robots in the wild," *Science Robotics*, vol. 7, 2022.
- [33] C. Zhang, N. Rudin, D. Hoeller, and M. Hutter, "Learning agile locomotion on risky terrains," *arXiv preprint arXiv:2311.10484*, 2023.
- [34] M. Mittal, N. Rudin, V. Klemm, A. Allshire, and M. Hutter, "Symmetry considerations for learning task symmetric robot policies," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.
- [35] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 611–24 624.
- [36] M. C. Machado, A. Barreto, D. Precup, and M. Bowling, "Temporal abstraction in reinforcement learning with the successor representation," *Journal of Machine Learning Research*, vol. 24, no. 80, pp. 1–69, 2023.
- [37] Y. Alon and H. Zhou, "Multi-agent reinforcement learning for unmanned aerial vehicle coordination by multi-critic policy gradient optimization," *IEEE Internet of Things Journal*, 2020.
- [38] T. Jacob, D. Duran, T. Pfeiffer, M. Vignati, and M. Johnson, "Multi-agent reinforcement learning for unmanned aerial vehicle capture-the-flag game behavior," in *Intelligent Systems Conference*. Springer, 2023, pp. 174–186.
- [39] S. Sainz-Palacios, "Deep reinforcement learning for shared autonomous vehicles (sav) fleet management," *arXiv preprint arXiv:2201.05720*, 2022.
- [40] C. Schmidt, D. Gammelli, F. C. Pereira, and F. Rodrigues, "Learning to control autonomous fleets from observation via offline reinforcement learning," in *2024 European Control Conference (ECC)*. IEEE, 2024, pp. 1399–1406.
- [41] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

- [42] T. R. X, “Lifelike agility and play on quadrupedal robots using reinforcement learning and generative pre-trained models,” *Nature Machine Intelligence*, 2023.
- [43] Y. Wang, R. Sagawa, and Y. Yoshiyasu, “Learning advanced locomotion for quadrupedal robots: A distributed multi-agent reinforcement learning framework with riemannian motion policies,” *Robotics*, vol. 13, no. 6, p. 86, 2024.
- [44] X. Han and M. Zhao, “Learning quadrupedal high-speed running on uneven terrain,” *Biomimetics*, vol. 9, no. 1, p. 37, 2024.
- [45] A. names not provided, “Deep reinforcement learning for real-world quadrupedal locomotion: a comprehensive review,” *OAEPublish*, 2023.
- [46] X. Lan, Y. Qiao, and B. Lee, “Towards pick and place multi robot coordination using multi-agent deep reinforcement learning,” in *2021 7th International Conference on Automation, Robotics and Applications (ICARA)*. IEEE, 2021, pp. 85–89.
- [47] Y. Liu, Z. Cao, H. Xiong, J. Du, H. Cao, and L. Zhang, “Dynamic obstacle avoidance for cable-driven parallel robots with mobile bases via sim-to-real reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1683–1690, 2023.
- [48] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [49] S. C. Ong, S. W. Png, D. Hsu, and W. S. Lee, “Pomdps for robotic tasks with mixed observability,” in *Robotics: Science and Systems*, vol. 5, 2009, p. 4.
- [50] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. Torr, P. Kohli, and S. Whiteson, “Stabilising experience replay for deep multi-agent reinforcement learning,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1146–1155.
- [51] M. H. Raibert and E. R. Tello, “Legged robots that balance,” *IEEE Expert*, vol. 1, no. 4, pp. 89–89, 1986.
- [52] 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, p. eaau5872, 2019.
- [53] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML ’09. New York, NY, USA: Association for Computing Machinery, 2009, p. 41–48.
- [54] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, “Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 619–12 629.
- [55] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.