Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Efficient learning of robust multigait quadruped locomotion for minimizing cost of transport

Zhicheng WANG[1], Xin ZHAO[†‡2], Meng Yee (Michael) CHUAH[3], Zhibin LI[4], Jun WU[1,5], Qiuguo ZHU[1,5]

[1]*Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310000, China*

[2]*Chinese Scholartree Ridge State Key Laboratory, China North Vehicle Research Institute, Beijing 100072, China*

[3]*Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore*

[4]*Department of Computer Science, University College London, London WC1E, United Kingdom*

[5]*State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310000, China*

[†]E-mail: pekingbit@163.com

Received Dec. 16, 2024; Revision accepted May 12, 2025; Crosschecked

Unedited

Abstract: Quadrupedal robots are able to exhibit a range of gaits, each with its own traversability and energy efficiency characteristics. By actively coordinating between gaits in different scenarios, energy-efficient and adaptive locomotion can be achieved. This study investigates the performance of learned energy-efficiency policies for quadrupedal gaits under different commands. We propose a training-synthesizing framework that integrates learned gait-conditioned locomotion policies into an efficient multiskill locomotion policy. The resulting control policies achieved low-cost smooth switching and controllable gaits. Our results of the learned multiskill policy demonstrate seamless gait transitions while maintaining energy optimality across all commands.

Key words: Reinforcement learning; Locomotion; Motor learning; Energy efficiency
https://doi.org/10.1631/FITEE.2401070 **CLC number:**

1 Introduction

Quadruped robots have successfully navigated complex environments using various control approaches, but their adaptability and efficiency still fall short compared to biological animals. In addition to differences in body structures, another significant reason is that animals can easily adopt the most suitable gait pattern and frequency and switch between gaits rapidly, smoothly, and robustly. Different gaits enable animals to effectively handle diverse terrain conditions at different speeds while maintaining good energy efficiency (Hildebrand, 1965; Hoyt and Taylor, 1981). As quadruped robots continue to be deployed in complex and unstructured environments, they will inevitably encounter new challenges that require emulating the strategies used by their natural counterparts.

Controllable gaits and active switching capabilities offer significant advantages for quadruped robot control (Haynes and Rizzi, 2006; Hsiao-Wecksler et al., 2010). By incorporating higher-level decision inputs, robots can activate different gaits in real-time (Xi et al., 2016). Moreover, these capabilities enable robots to go beyond locomotion and perform tasks such as dancing and leaping using specially designed gait controllers (Margolis and Agrawal, 2022).

However, the integration of multiple gaits to improve adaptability has been largely unexplored in existing research. Most frameworks adopt a single fixed gait type, or they neglect controllable gaits altogether and let the policy itself decide. Although this approach prioritizes simplicity and ease of tuning, it sacrifices optimality and controllability. Due to the manual fine-tuning typically required for

‡ Corresponding author
ORCID: Zhicheng WANG, https://orcid.org/0000-0002-2657-7591
© Zhejiang University Press 2025

multigait integration, many researchers opt not to explore this avenue.

Methods that emphasize gait-conditioned locomotion controllers and gait transitions, such as those utilizing central pattern generators, often rely on heuristics-based reference trajectories and constructed oscillators (Iscen et al., 2018; Shao et al., 2022; Tan et al., 2023). However, these approaches often introduce additional parameters that require time-consuming manual tuning, which imposes limitations on their performance.

As shown in Fig. 1, in this study, we investigate learned gait-conditioned locomotion policies and present a solution that addresses the considerations of energy efficiency, robustness, and robot hardware safety. Our goal is to develop an efficient multiskill locomotion policy that seamlessly integrates these different gait policies while maintaining a balanced approach to the factors mentioned above. By achieving smooth transitions between the various policies, our framework aims to optimize the overall performance and effectiveness in quadruped locomotion tasks.

The main contributions of this work are as follows: (1) development of a unified framework capable of training diverse and robust gait-conditioned

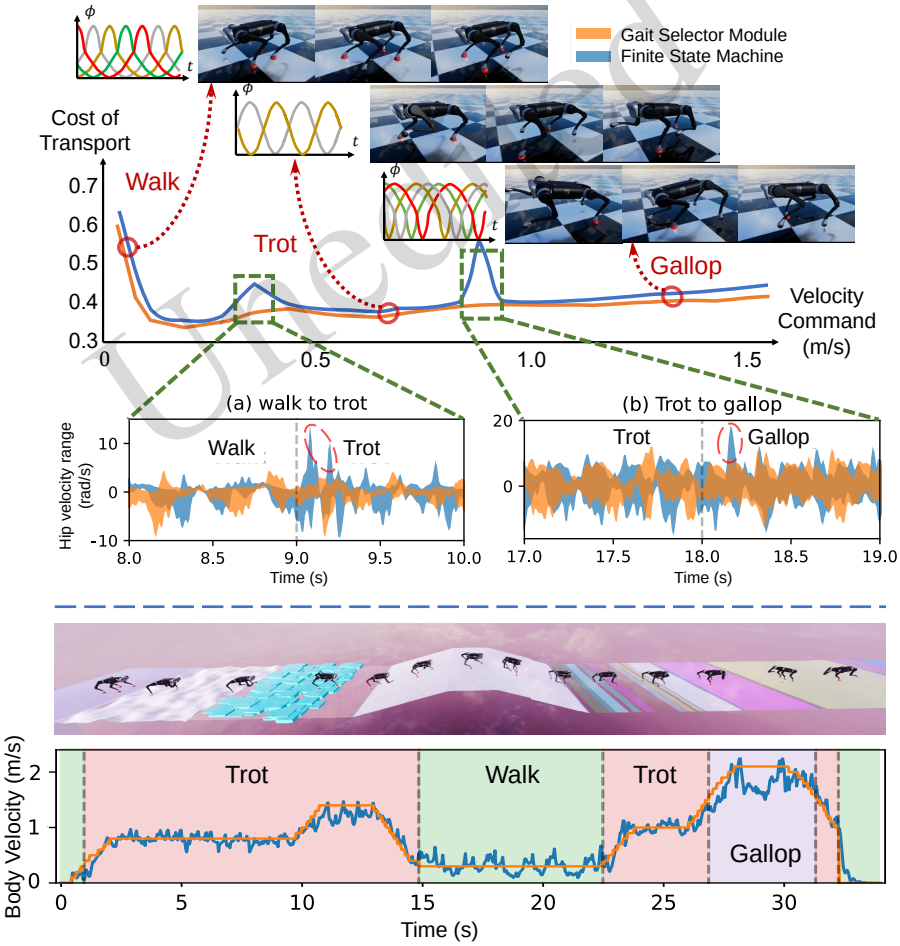


Fig. 1: Efficient gait-conditioned locomotion learning framework. Top: Single skill learning, analysis, and integration. Single-gait policies are trained to follow corresponding phase command signals and velocity commands. Through the use of reinforcement learning, a gait selector module can be trained. The performance of the proposed policy (orange plots) demonstrated a smoother transition than the state machine (blue plots) as shown in joint velocity plots (a) and (b). The vertical dotted lines indicate the gait transitions, and the undesired large spikes caused by improper switching schemes are marked out by dashed red circles. Bottom: The proposed multigait policy is robust enough to traverse complex environments. The snapshot and plot demonstrate the proposed policy deployed on a simulated robot traversing different terrains with gait transitions.

Unedited

locomotion policies; (2) proposal of a safe and cost-effective architecture for integrating trained gait policies to form an adaptive multimodal locomotion skill with smooth transition; (3) testing of the proposed policies in simulation environments and an analysis of their performance, which demonstrates their optimality.

2 Related work

2.1 Privileged learning-based locomotion

The base states of a robot, especially body velocity, are vital for achieving robust locomotion. However, estimators based on proprioceptive observations can be very noisy and inaccurate. To overcome this challenge, learning-based approaches that leverage privileged information during training have emerged as a powerful tool (Chen et al., 2019). By utilizing precise states available in simulation, these approaches enable more effective learning.

Since Lee et al. (2020) and Miki et al. (2022) made the first successful implementation and Kumar et al. (2021) and Fu et al. (2021) achieved reference-free sim-to-real transfer, privileged training has been widely adopted in the field of learning-based locomotion research (Agarwal et al., 2022; Kumar et al., 2022; Margolis and Agrawal, 2022; Margolis et al., 2022; Loquercio et al., 2023). Motivated by these advancements, our work aims to combine privilege training with gait regulation to develop an efficient and robust multigait policy.

2.2 Gait-conditioned locomotion

The study of legged locomotion has been a long-standing area of research. Gait-conditioned policy can be obtained through various approaches, such as learning gait library residuals (Iscen et al., 2018; Siekmann et al., 2020; Kumar et al., 2021; Xie et al., 2021), imitation learning (Peng et al., 2020; Tsounis et al., 2020; Jin et al., 2022) or reward design. One common approach for reward formulation is penalizing contact state errors to encourage tracking a predefined contact reference (Tan et al., 2018; Acero et al., 2022; Shao et al., 2022). Bio-inspired heuristic policies can be used to generate smoother transitions between different gaits (Iscen et al., 2018; Shao et al., 2022). Siekmann et al. (2021a) introduced phase-dependent rewards that penalize contact force and foot velocity to ensure proper swing-stance cycles, leading to the development of gait-conditioned policies for legged robots (Siekmann et al., 2020, 2021b; Shao et al., 2022). Fu et al. (2021) demonstrated that indirect energy rewards can also lead to the emergence of gaits, although they may not strictly follow explicit gait patterns.

However, the energy efficiency of these learned gaits has not been extensively explored in existing works. In this study, we focus on investigating the energy efficiency of different learned gaits and identifying the optimal gait to perform under various velocity commands. We aim to provide insights into the energy characteristics of these gaits and determine the most suitable gait for different locomotion requirements.

2.3 Motor skill integration

Skill synthesis plays a critical role in achieving multigait locomotion. One common approach is to design finite state machines (FSMs) and activate the desired controller based on specific conditions (Lee et al., 2019). However, this method can result in brittle behaviors during switching. Alternative synthesizing methods, including integrating motion reference datasets into neural policies using techniques such as multiplicative compositional policies (Peng et al., 2019), and latent adversarial methods (Luo et al., 2020; Peng et al., 2021, 2022) have also been proposed. Other approaches involve fusing skills into a single policy, such as policy distillation (Hinton et al., 2015; Fu et al., 2021), mixture of experts (Jacobs et al., 1991), or expert policy mixture (Zhang et al., 2018; Yang et al., 2020).

Our study uses a trained gait selector module that uses velocity commands and terrain information to decide gait parameters. We have observed that parameter-level switching is sufficient to provide robust and adaptive integrating behavior, without having to distill multiple policies with divergent behavior patterns.

3 Methodology

3.1 Control architecture

In our framework, the control system includes a group of learning-based single-gait policies, a trained gait selector module, and a gait phase generator, as

Unedited

depicted in Fig. 2, the details of which are explained in Section 3.2.

To enable efficient training, we refer to the architecture proposed in (Nahrendra et al., 2023) when training the single-gait policy. A single-gait policy with gait type j, denoted as σ_j, consists of a variational autoencoder (μ_j, η_j) to predict the proprioceptive observation, compress the proprioceptive observation history into latent encodings z, and explicitly estimate the robot's linear velocity $\hat{v}$; a backbone policy ρ conditions joint-level position targets on the latest proprioceptive states, estimations, and encodings.

The input of a single-gait policy is composed of the linear velocity command $\boldsymbol{v}_t^*$, gait phase vector $\mathbf{g}(t)$, an h-step proprioceptive states history $\boldsymbol{s}_{t:t-h+1}$, and the previous action $\boldsymbol{a}_{t-1}$. The proprioceptive states $\boldsymbol{s}$ refer to the available sensor states of real-world robots, including gravity vector $\boldsymbol{k}$, body angular velocity $\boldsymbol{\omega}$, joint space position $\boldsymbol{q}$, and velocity $\dot{\boldsymbol{q}}$. The action space of a single-gait policy consists of a 12-dimensional joint position command. The single-gait policies run at 50Hz. A low-stiffness 1kHz proportional-derivative (PD) controller is used to track these joint position commands while maintaining compliance.

To achieve automatic gait switching, a gait selector module ψ is trained with reinforcement learning (RL). During training, the single-gait policies are frozen. The gait selector module computes gait parameters $\boldsymbol{\theta}_t$ and a one-hot vector $\mathbf{w}$. Its input is composed of the height map around the robot $\boldsymbol{H}$, linear velocity command $\boldsymbol{v}_t^*$, the proprioceptive history $\boldsymbol{s}_{t:t-h+1}$, and the previous action $\boldsymbol{\theta}_{t-1}$. The gait selector module runs at 5Hz.

The formulation of the control system can be described by the following equations.

$$[\mathbf{w}, \boldsymbol{\theta}] = \psi(\boldsymbol{v}^*, \boldsymbol{H}, \boldsymbol{s}_{t:t-h}, \boldsymbol{\theta}_{prev}) \tag{1}$$

$$[z, \hat{v}] = \mu_j(\boldsymbol{s}_{t-1:t-h}) \tag{2}$$

$$\hat{\boldsymbol{s}}_{t+1} = \eta_j(\boldsymbol{s}_{t-1:t-h}) \tag{3}$$

$$\boldsymbol{a}_{j,t} = \rho_j(\boldsymbol{z}, \hat{\boldsymbol{v}}, \boldsymbol{v}^*, \mathbf{g}_\theta(t), \boldsymbol{s}_t, \boldsymbol{a}_{j,t-1}) \tag{4}$$

$$a_t = \sum_{j=1}^{N} w_j \cdot \boldsymbol{a}_{j,t} \tag{5}$$

3.2 Gait formulation

The choice of gait in legged locomotion involves a trade-off between energy consumption and traversability, as different gaits prioritize different performance aspects. In general, a longer flight phase within a given period allows the policy to achieve higher velocity but at the cost of increased energy consumption.

To achieve controllable gait patterns within a unified framework, we design a phase generator to generate the gait phase vector as an input component of the policies. Inspired by central-patern-generator-based works (Ijspeert, 2008; Iscen et al., 2018), we adopt periodic signals based on sine waves to specify gait phase commands. Fig. 3 illustrates the contact sequences based on gait phase signals for typical gaits. By modulating the period and offset of every leg, the robot can learn to perform any desired gait. This approach provides flexibility for designing gaits for other performance considerations, as energy consumption may not be the only consideration in challenging environmental conditions.

Unedited

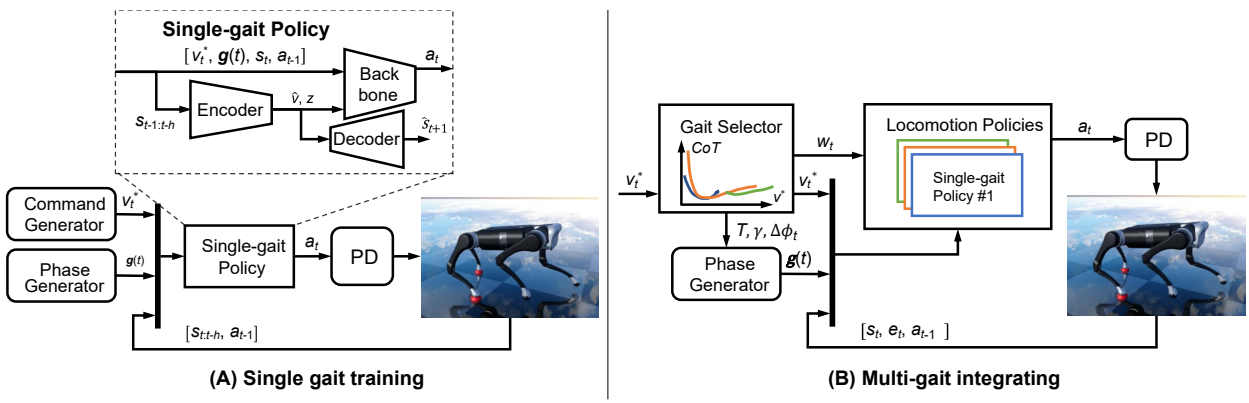


Fig. 2: Main control architecture. (A) Training single-gait policy; (B) formulation of multiskill locomotion policy. The notations are explained in Section 3.1.

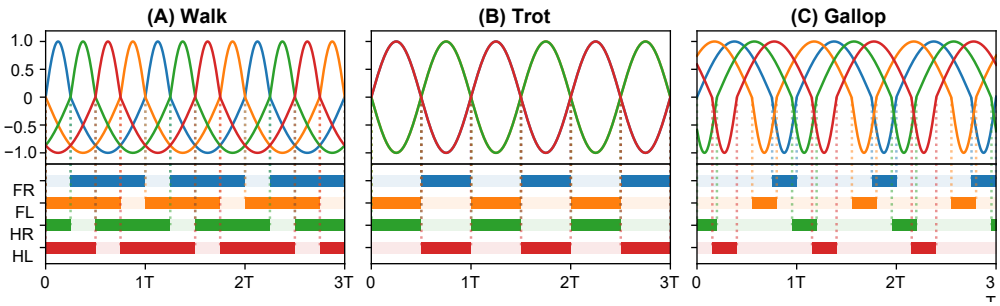


Fig. 3: Gait phase signals and foot contact sequence of different gaits. The horizontal axis represents time, the vertical axis of upper plots refers to the value of gait signal axis, and the vertical axis of lower plots refers to the contact states of robot's end effectors. (A) Quasi-static walk; (B) trot; and (C) gallop.

The phase generator adopted in our implementation can be formulated as Eqs. (6)-(9):

$$\mathbf{g}_\theta(t) = [\boldsymbol{g}_0, \cdots, \boldsymbol{g}_3]^T \tag{6}$$

$$\boldsymbol{\theta} = [\Delta\phi_0, \cdots, \Delta\phi_3, \gamma_0, \cdots, \gamma_3, T_0, \cdots, T_3] \tag{7}$$

$$\boldsymbol{g}_i = [\sin\phi_i, \cos\phi_i]^T, i = 0, 1, 2, 3 \tag{8}$$

$$\phi_i = \begin{cases} \frac{\pi}{\gamma_i T_i} t + \Delta\phi_i & 0 \le t < \gamma_i \\ \frac{\pi t}{(1-\gamma_i)T_i} + \frac{1-2\gamma_i}{1-\gamma_i}\pi + \Delta\phi_i & \gamma_i \le t < T_i \end{cases} \tag{9}$$

where $\mathbf{g}_\theta(t)$ is the periodic phase command generated by the generator, which is a function of time in a period t and gait parameter $\boldsymbol{\theta}$, T_i is the period of the current gait, γ_i is the duty cycle of the specified gait, representing the proportion of flight phase within a period, $\Delta\phi_i$ is the phase offset of the corresponding gait, ϕ_i is the linear phase indicator for foot i, and $\boldsymbol{g}_i$ is the trigonometric phase signal for foot i. Two state channels per foot are designed to guarantee that every time point can only correspond to one phase state within a given period. Note that π here is the mathematical constant instead of the policy.

To demonstrate the multigait capability of our proposed framework, we have selected quasistatic walking, dynamic trotting, and galloping. These gaits possess diverse features that can address most basic locomotion demands.

3.3 Policy training

3.3.1 RL approach

The state of the robot at a specific time is constrained by the previous state and the action taken by the robot therefore, both the gait-conditioned locomotion control and the gait decision can be described by a Markov decision process, which is suitable for RL. The action performed by the robot using its policy influences the probability distribution of the state transition, and the result of the transition leads to a corresponding reward value, which indicates how successful the state is. Hence, optimizing the performance of the controller is equivalent to maximizing the total reward over an infinite horizon. In our implementation, the single-gait policy and the gait selector module are trained with proximal policy optimization (PPO) (Schulman et al., 2017). Parameters related to PPO and policy architecture can be found in the Appendix.

3.3.2 Single-gait reward design

Our reward terms, which focus more on periodic gait behavior and energy efficiency, can be categorized into three distinct classes: task-oriented, gait-conditioned, and efficiency-related rewards. The gait-related reward design is based on (Siekmann et al., 2021a). We use the cost of transport (CoT) as a representation of energy consumption. The reward function is formulated as follows:

$$R_{\text{velocity}} = F_{\alpha_1,\beta_1}(||\boldsymbol{v}^* - \boldsymbol{v}||) \tag{10}$$

$$R_{\text{orientation}} = F_{\alpha_2,\beta_2}(||\boldsymbol{k}^* - \boldsymbol{k}|| \tag{11}$$

$$R_{\text{height}} = F_{\alpha_3,\beta_3}(|p_z^* - p_z|) \tag{12}$$

$$R_{\text{swing}} = F_{\alpha_4,\beta_4}(\sum_i (C_i||\boldsymbol{F}_i||) \tag{13}$$

$$R_{\text{stance}} = F_{\alpha_5,\beta_5}(\sum_i ((1-C_i)||\boldsymbol{v}_{f,i}||)) \tag{14}$$

$$R_{\text{energy}} = F_{\alpha_6,\beta_6}(\frac{|\dot{\boldsymbol{q}}^T| \cdot |\boldsymbol{\tau}|}{mg||\boldsymbol{v}||}) \tag{15}$$

$$R_{\text{smooth}} = F_{\alpha_7,\beta_7}(||\boldsymbol{\tau} - \boldsymbol{\tau}_{\text{prev}}||) \tag{16}$$

$$R_{\text{jointvel}} = F_{\alpha_8,\beta_8}(||\dot{\boldsymbol{q}}||) \tag{17}$$

Unedited

$$F_{\alpha,\beta}(x) = \alpha e^{-\beta x^2} \tag{18}$$

$$C_i = \begin{cases} 1 & \phi_i > 0 \\ 0 & \phi_i \leq 0 \end{cases} \tag{19}$$

where $i \in \{FR, FL, HR, HL\}$ is the foot index, C_i is the flight phase coefficient indicating whether the corresponding foot i should be in swing phase, $\boldsymbol{F}_i$ is the contact force on foot i, $\boldsymbol{v}_{f,i}$ is the world-frame velocity of foot i, τ is the torque command applied on joints, and $F_{\alpha,\beta}(x)$ is a Gaussian kernel applied to all terms to ensure reward survival. The coefficient α represents the importance coefficient related to the weight of this term in the final reward composition, and β is the scaling coefficient, which is related to the physical meaning of the error term. CoT is computed by summing the absolute joint powers to represent total energy consumption.

3.3.3 Gait selector reward design

Aimed at selecting the optimal gait policy and parameters, while preventing the gait selection from oscillating too much, the reward function is defined as follows:

$$R_{\text{velocity}} = F_{\alpha_9,\beta_9}(||\boldsymbol{v}^* - \boldsymbol{v}||) \tag{20}$$

$$R_{\text{survival}} = -\alpha_{10} K_{\text{dones}} \tag{21}$$

$$R_{\text{smooth}} = -\alpha_{11} ||\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{prev}}|| \tag{22}$$

$$R_{\text{jointvel}} = -\alpha_{12} \max ||\dot{\boldsymbol{q}}|| \tag{23}$$

$$R_{\text{energy}} = -\alpha_{13} \frac{|\dot{\boldsymbol{q}}^T| \cdot |\boldsymbol{\tau}|}{mg||\boldsymbol{v}||} \tag{24}$$

where K_{dones} is the failing times between two timesteps, and $\boldsymbol{\theta}_{\text{prev}}$ represents the gait parameters given by the gait selector at the last timestep.

3.3.4 Domain randomization

To facilitate adaptation to a wide range of tasks, we adopt natural constraints and randomization, including terrain friction, motor strength, and payload, for better sim-to-real transfer, with reference to the work of (Kumar et al., 2021). The detailed ranges are given in the Appendix.

3.3.5 Curriculum training

game-inspired curriculum training approach is used to ensure the learning of gait-conditioned policies. The same curriculum terrain is applied to both single-gait training and gait selector training. Detailed parameters can be found in the Appendix.

4 Experimental results

4.1 Experimental setup

In our work, we utilized the Jueying Lite3 and Unitree A1, both small-size 12-degree-of-freedom quadrupedal robots. All processes run on a laptop equipped with an Intel Core i7-11800H and an NVIDIA GeForce RTX 3070. The training environments are implemented in IsaacGym (Makoviychuk et al., 2021), experiments are performed in a realistic simulator, RaiSim (Hwangbo et al., 2018), and the policy architecture is built with PyTorch (Paszke et al., 2019). To ensure reproducibility, the random seeds were set based on the system time upon process initialization.

To evaluate the performance of trained policies and the proposed gait selector module, we set up comparison groups against the proposed method as follows:

- **Single-gait policies**(SGPs): To demonstrate the performance of single-gait policies, we apply two typical gait frequencies to every policy for evaluation. All policies are trained with the same parameter configuration, as listed in the Appendix. To ensure the repeatability of our results, we trained 20 policies for each gait with different random seeds.

- **SGP w/ finite state machine**(FSM): Given a user command v^*, this policy chooses the policy with the best energy efficiency. The gait frequency is fixed to a medium frequency. The FSM switching logic is listed in the Appendix.

Meanwhile, to compare the energy efficiency with that of existing methods, the proposed method is compared with the following methodologies from the literature:

- **Emergent gait (EG) distillation (Fu et al., 2021)**: Instead of using explicit gait signal commands, this method uses energy reward to train policies with different gaits.

- **Model Predictive control (Di Carlo et al., 2018) (MPC)**: This method is a model-based convex MPC controller. It does not explicitly optimize energy terms.

Unedited

4.2 Evaluation metrics

4.2.1 Task performance

To evaluate task performance, we measure the average velocity tracking the root mean square (RMS) errors (RMSEs) under typical velocity commands on flat terrain over a duration of 40s. The velocity commands, chosen based on related works (Di Carlo et al., 2018; Fu et al., 2021), include 0.375m/s, 0.9m/s, and 1.5m/s. The corresponding velocity tracking errors are represented as RMS(Δv_L), RMS(Δv_M), and RMS(Δv_H), respectively.

4.2.2 Energy efficiency

Energy efficiency is assessed using the average CoT, denoted as CoT, under feasible velocity commands. Feasibility is defined as the ability of the policy to operate for >40s without failure on flat ground. Velocity commands are sampled at intervals of 0.02m/s, and each trial is conducted for a duration of 40s. To simulate real-world conditions, a Gaussian noise with a standard deviation of 0.1m/s is added to the constant velocity command in each trial.

4.2.3 Robustness

Robustness is determined by the traversing success rate, denoted as TSR, over 20 trials on terrains of 10m length representing common challenges. Demonstrated by Fig. 4, these terrains include fractal terrain with a height standard deviation of 0.08m, gaps with a maximum length of 0.1m, discrete terrain with a maximum height difference of 0.1m, slopes of 20°, and projectile impacts, where a 0.1kg sphere is shot at 40m/s from a random direction every 0.5s. In our implementation, failure is defined as the inability to cross the terrain within 60s. The velocity commands for each trial are randomly sampled from the feasible range, ensuring diverse testing scenarios.

4.2.4 Performance summary

A summary of performance under different conditions is provided in Table 1, and a comparison of performance relative to other methods is presented in Table 2. The optimal performances are highlighted in bold characters.

4.2.5 Safety

Safety is evaluated based on the ratio between the maximum joint velocity during the first second after switching and that of the next second. A larger maximum joint velocity ratio suggests that the transition behavior is more violent than the normal limit cycle. We collected data from 2000 switchings happening at random times. This test only applies to the policies with explicit gait transition, which includes the proposed method and FSM.

4.3 Gait-conditioned policies

4.3.1 Feasibility region evaluation for single-skill policies

The performance of the single-gait policies is summarized in Table 1 and Fig. 5. These results demonstrate that all single-gait policies are capable of effectively tracking velocity and gait command signals, highlighting the robustness of our training framework in enabling the policies to master locomotion skills. More detailed evaluations are shown

Unedited

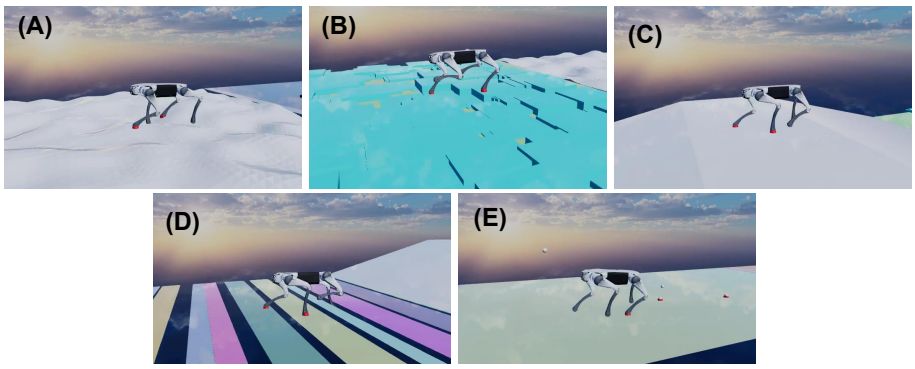
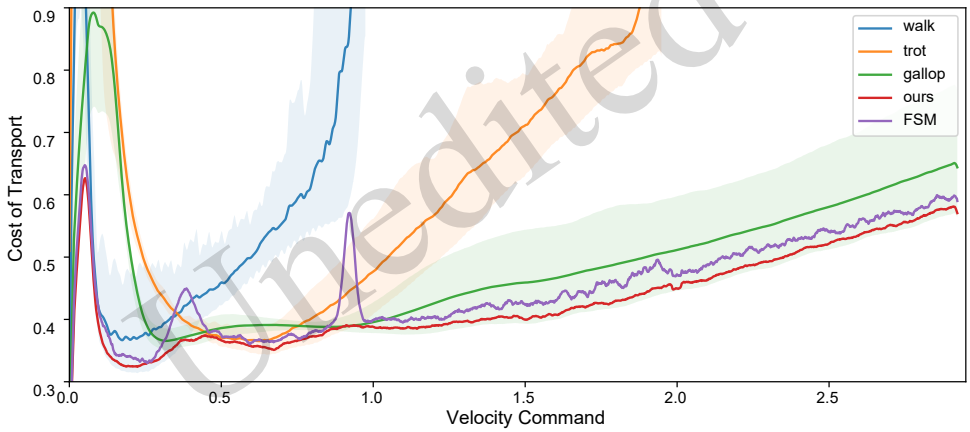


Fig. 4: Snapshots of different regions in the testing course: (A) fractal terrain; (B) random platforms; (C) sloped ground; (D) region with gaps; and (E) projectile impact region.

Table 1: Performance summary under different conditions

	Walk		Trot		Gallop		FSM	Ours
Frequency(Hz)	0.8	1.6	1.2	2.4	2.0	3.2		
RMS(Δv_L)	0.1034	**0.0920**	0.0954	0.0982	0.0966	0.0941	0.1125	0.0929
RMS(Δv_M)	0.3768	0.1430	0.1365	0.0860	0.0621	0.0625	0.0827	**0.0620**
RMS(Δv_H)	0.8916	0.4997	0.3042	0.0880	0.0760	**0.0500**	0.0650	0.0522
CoT	0.7241	0.4733	0.6886	0.6041	0.5675	0.4514	0.4531	**0.4306**
$\text{TSR}_{\text{fractal}}$	0.90	0.95	**1.00**	**1.00**	0.80	0.85	**1.00**	**1.00**
$\text{TSR}_{\text{slope}}$	0.85	0.85	**1.00**	**1.00**	0.65	0.70	**1.00**	**1.00**
$\text{TSR}_{\text{discrete}}$	0.80	0.95	**1.00**	0.95	0.70	0.60	0.75	**1.00**
TSR_{gaps}	0.80	0.80	0.90	**1.00**	0.80	0.80	**1.00**	0.95
$\text{TSR}_{\text{projectile}}$	0.95	**1.00**	**1.00**	**1.00**	1.00	**1.00**	**1.00**	**1.00**



Unedited

Fig. 5: Cost of transport comparison between multigait policies and single-gait member policies.

Table 2: Performance comparison with other methods

v^*(m/s)	**Ours**	**EG**	**MPC**
0.375	**0.3551**	0.8059	2.2390
0.9	**0.3803**	0.3936	0.7506
1.5	**0.3937**	0.5841	0.5871

in the Appendix.

Each single-gait policy exhibits trade-offs across different metrics. Walking policies excel in energy efficiency under low-velocity commands and demonstrate superior robustness but are unable to track high-velocity commands. Conversely, galloping policies achieve high-speed locomotion but exhibit reduced robustness and increased energy consumption. These trade-offs underline the inherent limitations of single-gait strategies and the potential benefits of combining them in multigait policies.

To ensure that the policies are applied under appropriate circumstances, we need to determine the feasible working velocity command range of each policy. We consider the policy to be reliable when failure rate per second, mean velocity command tracking error, and mean contact state error are all below manually defined thresholds. These results are also affected by gait frequency and command velocities. The feasibility evaluation calculation and figures are listed in the Appendix.

Figs. 6-8 show the performances of learned policies under different conditions. The feasible command range of all policies under different frequencies can be summarized from them. Generally, we can conclude that higher gait frequencies allow the policies to achieve a wider velocity range, and trotting

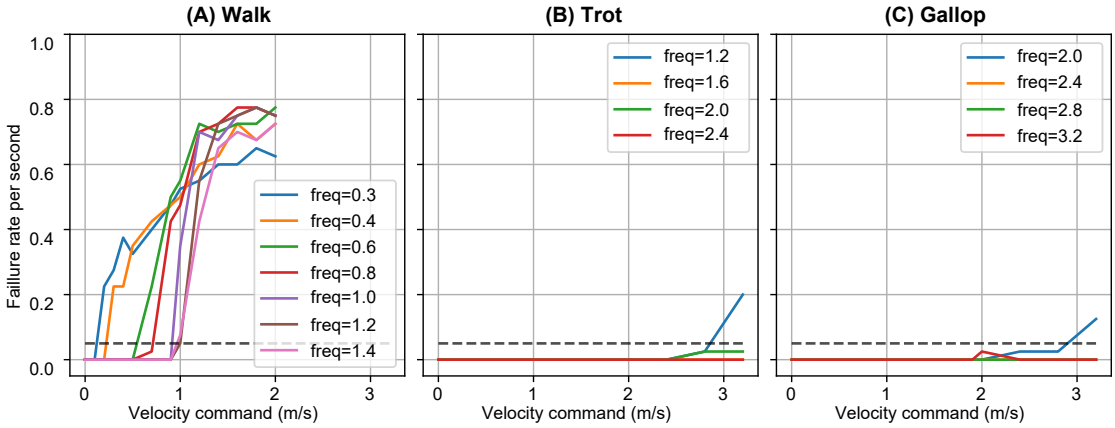
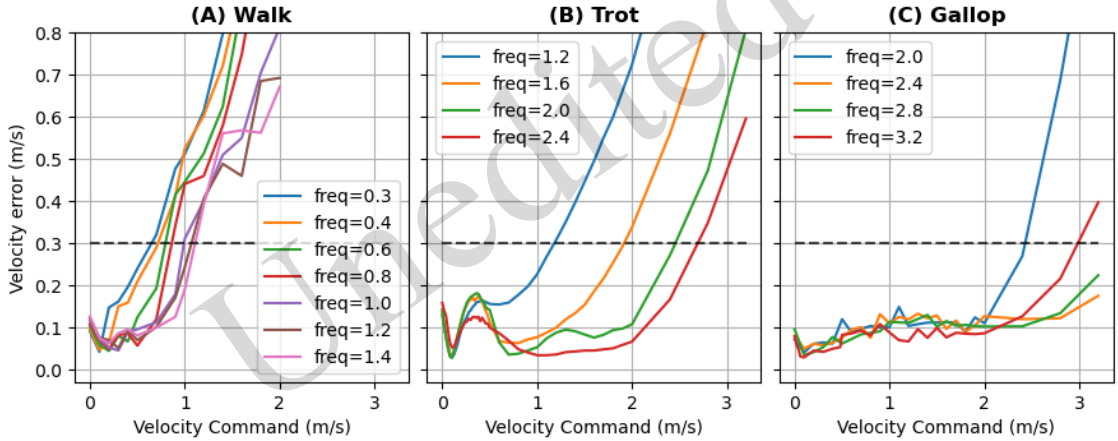


Fig. 6: Failure rates of trained single-gait policies under various velocity commands. Failure rate per second is the total failure counts in a time period divided by the length of the episode. To ensure high robustness for real robot transfer, we set the failure threshold at 0.05 (represented by the dotted line).



Unedited

Fig. 7: Velocity tracking performance over a range of velocity commands. Velocity command tracking error is the RMSE between world-frame velocities and commands in one test episode. According to our experience of using a real quadruped robot, we set the threshold to 0.3m/s (represented by the dotted line).

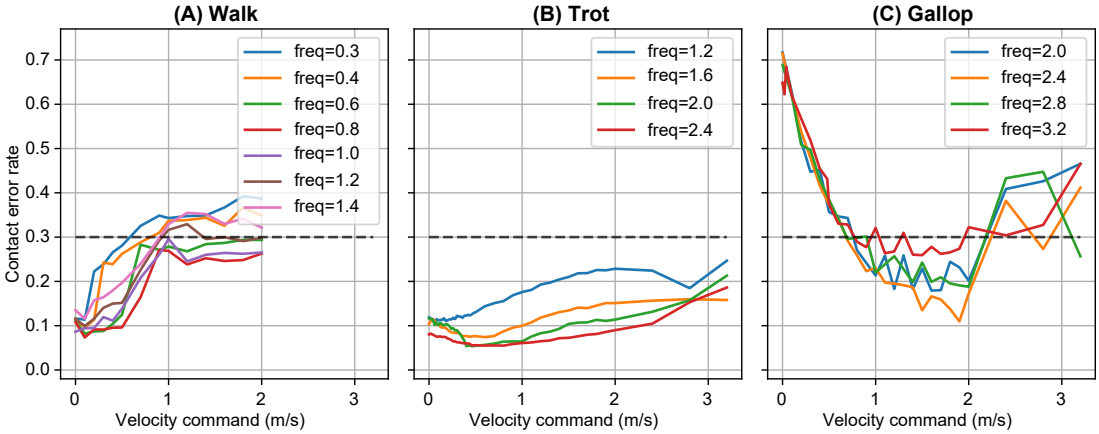


Fig. 8: Average contact state error over a range of velocity commands. Contact state error is the average absolute error between binary contact states and expected contact states in one test episode. To make sure that the desired gait is performed while leaving a flexible margin for flight phases, the threshold is set to 0.36.

gait has the widest feasible range compared with other gait policies. Specifically, in Fig. 8, we disqualify low-speed galloping because it becomes grounded and inconsistent with the command phase signal under lower velocity targets.

4.3.2 Gait transition

The performance of the multigait policies is presented in Table 1. Both the proposed method and the FSM group effectively leverage the advantages of all single-gait policies. The velocity tracking accuracy of both groups is comparable to the best-performing single-gait policies, with the proposed method slightly outperforming the FSM group. Additionally, the proposed policy demonstrates superior adaptability to challenging terrains compared to the FSM group. Notably, the gait selector tends to favor the robust trotting policy when traversing complex terrain conditions, such as discrete regions and slopes.

A detailed plot of energy efficiency is shown in Fig. 5. The proposed policy consistently achieves the highest energy efficiency across all velocity commands. While the FSM group is also efficient for most velocity commands, two notable spikes in energy consumption occur around gait-switching points. These spikes are attributed to oscillations between policies caused by noisy command velocities, which result in discrete fluctuations in phase vectors. In contrast, the gait selector activates gait transitions at more optimal timing, minimizing such oscillations and ensuring smoother transitions.

Regarding safety, the joint velocity ratio distribution is illustrated in Fig. 9. The FSM group exhibits higher peak joint velocities during gait transitions, indicating more pronounced fluctuations. In comparison, the gait selector produces smoother transitions with fewer fluctuations, aligning with the energy efficiency findings. A video demonstrating the gait transition behavior is included in the Supplementary materials.

5 Conclusions

This work presents a learning framework for training and organizing robust gait-conditioned policies, revealing the relationship between velocity command and energy efficiency across different gaits. Based on these findings, an adaptive multimodal locomotion controller is developed, and it outper-

Unedited

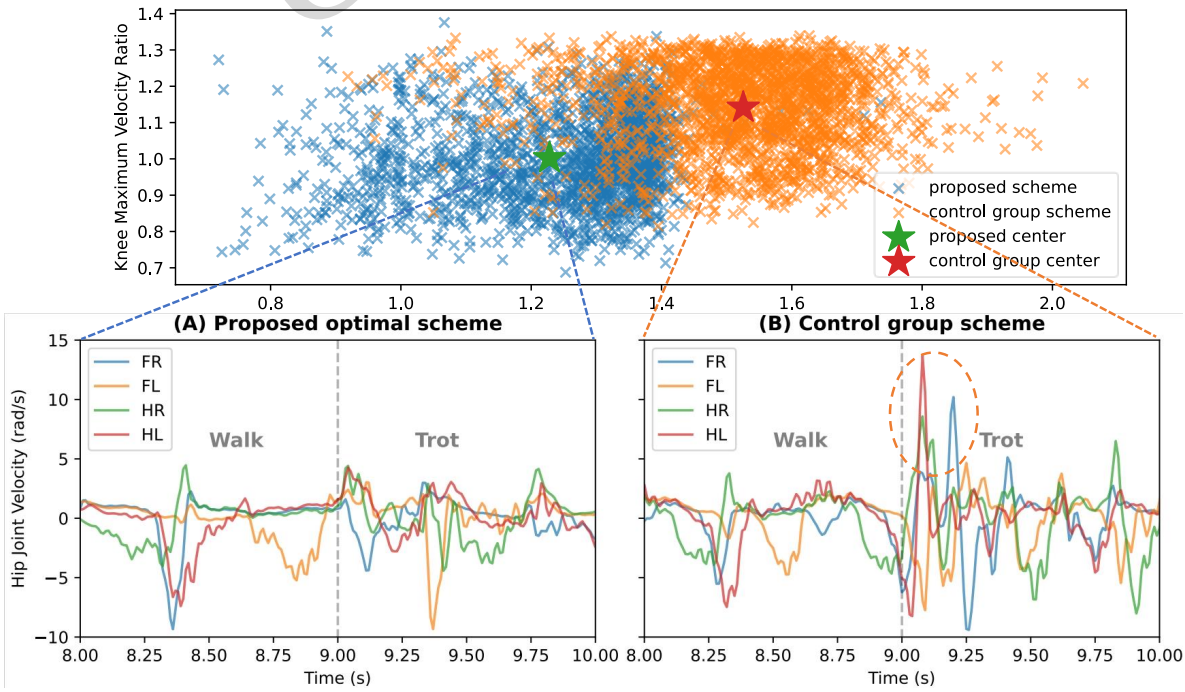


Fig. 9: Joint velocity comparison. Top: Joint velocity ratio distribution. The FSM cluster has more intense hip joint movements. Bottom: Typical joint velocity plots during the time of switching. The undesired joint velocity spikes are marked out by a dotted circle.

forms single-gait policies in terms of velocity tracking accuracy, CoT, and robustness.

5.1 Limitation on constraints

In the presented work, the single-gait policies switch discretely, and the constraints are expressed in rewards when training the gait selector. Future work can focus on developing safer and more sophisticated switching strategies that take into account strict constraints.

5.2 Future work on elastic bodies

The energy consumption analysis reveals that while the proposed policy exhibits advantages relative to existing works, it still has lower energy efficiency compared to animals. This difference can be attributed to the fact that animals have elastic body structures, such as tendons, that allow them to store and release energy. Future work could explore incorporating elastic structures into a robot to achieve better energy performance.

5.3 Future work on motor skills

The framework presented allows for the use of other motor skills by higher-level controllers. Future research could focus on developing advanced motor skills that integrate perception information to enhance the capabilities and adaptability of the robot.

Contributors

Meng Yee (Michael) CHUAH conceptualized and designed the research. Zhicheng WANG coded the robot learning and controlling system and processed the data. Zhicheng WANG drafted the manuscript. Meng Yee (Michael) CHUAH helped organize the manuscript. Zhicheng WANG and Meng Yee (Michael) CHUAH revised and finalized the paper. Funding Acquisition is done by Qiuguo ZHU, Jun WU and Xin ZHAO. Qiuguo ZHU and Zhibin LI were in charge of supervision and project administration. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Acero F, Yuan K, Li ZB, 2022. Learning perceptual locomotion on uneven terrains using sparse visual observations. *IEEE Robot Autom Lett*, 7(4):8611-8618. https://doi.org/10.1109/LRA.2022.3188108

Agarwal A, Kumar A, Malik J, et al., 2022. Legged locomotion in challenging terrains using egocentric vision. Proc 6th Annual Conf on Robot Learning, p.403-415.

Chen D, Zhou B, Koltun V, et al., 2019. Learning by cheating. Proc 3rd Annual Conf on Robot Learning, p.66-75.

Di Carlo J, Wensing PM, Katz B, et al., 2018. Dynamic locomotion in the mit cheetah 3 through convex model-predictive control. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.1-9. https://doi.org/10.1109/IROS.2018.8594448

Fu ZP, Kumar A, Malik J, et al., 2021. Minimizing energy consumption leads to the emergence of gaits in legged robots. Proc 5th Annual Conf on Robot Learning, p.928-937. https://doi.org/10.1109/ROBOT.2006.1641859

Haynes GC, Rizzi AA, 2006. Gaits and gait transitions for legged robots. Proc IEEE Int Conf on Robotics and Automation, p.1117-1122. https://doi.org/10.1109/ROBOT.2006.1641859

Hildebrand M, 1965. Symmetrical gaits of horses. *Science*, 150(3697):701-708. https://doi.org/10.1126/science.150.3697.701

Hinton G, Vinyals O, Dean J, 2015. Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531

Hoyt DF, Taylor CR, 1981. Gait and the energetics of locomotion in horses. *Nature*, 292(5820):239-240. https://doi.org/10.1038/292239a0

Hsiao-Wecksler ET, Polk JD, Rosengren KS, et al., 2010. A review of new analytic techniques for quantifying symmetry in locomotion. *Symmetry*, 2(2):1135-1155. https://doi.org/10.3390/sym2021135

Hwangbo J, Lee J, Hutter M, 2018. Per-contact iteration method for solving contact dynamics. *IEEE Robot Autom Lett*, 3(2):895-902. https://doi.org/10.1109/LRA.2018.2792536

Ijspeert AJ, 2008. Central pattern generators for locomotion control in animals and robots: a review. *Neural Netw*, 21(4):642-653. https://doi.org/10.1016/j.neunet.2008.03.014

Iscen A, Caluwaerts K, Tan J, et al., 2018. Policies modulating trajectory generators. Proc 2nd Annual Conf on Robot Learning, p.916-926.

Jacobs RA, Jordan MI, Nowlan SJ, et al., 1991. Adaptive mixtures of local experts. *Neural Comput*, 3(1):79-87. https://doi.org/10.1162/neco.1991.3.1.79

Jin YB, Liu XW, Shao YC, et al., 2022. High-speed quadrupedal locomotion by imitation-relaxation reinforcement learning. *Nat Mach Intell*, 4(12):1198-1208. https://doi.org/10.1038/s42256-022-00576-3

Kumar A, Fu ZP, Pathak D, et al., 2021. RMA: rapid motor adaptation for legged robots. Proc 17th Robotics: Science and Systems, p.1-9. https://doi.org/10.15607/rss.2021.xvii.011

Unedited

Kumar A, Li Z, Zeng J, et al., 2022. Adapting rapid motor adaptation for bipedal robots. Proc IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.1161-1168. https://doi.org/10.1109/IROS47612.2022.9981091

Lee J, Hwangbo J, Hutter M, 2019. Robust recovery controller for a quadrupedal robot using deep reinforcement learning. https://doi.org/10.48550/arXiv.1901.07517

Lee J, Hwangbo J, Wellhausen L, et al., 2020. Learning quadrupedal locomotion over challenging terrain. *Sci robot*, 5(47):eabc5986. https://doi.org/10.1126/scirobotics.abc5986

Loquercio A, Kumar A, Malik J, 2023. Learning visual locomotion with cross-modal supervision. Proc Int Conf on Robotics and Automation, p.7295-7302. https://doi.org/10.1109/ICRA48891.2023.10160760

Luo YS, Soeseno JH, Chen TPC, et al., 2020. Carl: controllable agent with reinforcement learning for quadruped locomotion. *ACM Trans Graph*, 39(4):38. https://doi.org/10.1145/3386569.3392433

Makoviychuk V, Wawrzyniak L, Guo YR, et al., 2021. Isaac gym: high performance gpu based physics simulation for robot learning. Proc 35th Conf on Neural Information Processing Systems, p.1-12.

Margolis GB, Agrawal P, 2022. Walk these ways: tuning robot control for generalization with multiplicity of behavior. Proc 6th Annual Conf on Robot Learning, p.1-9.

Margolis GB, Yang G, Paigwar K, et al., 2022. Rapid locomotion via reinforcement learning. Proc 18th Robotics: Science and Systems, p.1-9.

Miki T, Lee J, Hwangbo J, et al., 2022. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Sci Robot*, 7(62):eabk2822. https://doi.org/10.1126/scirobotics.abk2822

Nahrendra IMA, Yu B, Myung H, 2023. Dreamwaq: learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning. Proc IEEE Int Conf on Robotics and Automation, p.5078-5084. https://doi.org/10.1109/ICRA48891.2023.10161144

Paszke A, Gross S, Massa F, et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Proc 33rd Int Conf on Neural Information Processing Systems, p.8024-8035.

Peng XB, Chang M, Zhang G, et al., 2019. Mcp: Learning composable hierarchical control with multiplicative compositional policies. Proc 33rd Int Conf on Neural Information Processing Systems, article 331.

Peng XB, Coumans E, Zhang TN, et al., 2020. Learning agile robotic locomotion skills by imitating animals. Proc 16th Robotics: Science and Systems, p.1-9.

Peng XB, Ma Z, Abbeel P, et al., 2021. Amp: adversarial motion priors for stylized physics-based character control. *ACM Trans Graph*, 40(4):144. https://doi.org/10.1145/3450626.3459670

Peng XB, Guo YR, Halper L, et al., 2022. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Trans Graph*, 41(4):94. https://doi.org/10.1145/3528223.3530110

Schulman J, Wolski F, Dhariwal P, et al., 2017. Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347

Shao YC, Jin YB, Liu XW, et al., 2022. Learning free gait transition for quadruped robots via phase-guided controller. *IEEE Robot Autom Lett*, 7(2):1230-1237. https://doi.org/10.1109/LRA.2021.3136645

Siekmann J, Valluri S, Dao J, et al., 2020. Learning memory-based control for human-scale bipedal locomotion. Proc 16th Robotics: Science and Systems, p.1-8. https://doi.org/10.15607/rss.2020.xvi.031

Siekmann J, Godse Y, Fern A, et al., 2021a. Sim-to-real learning of all common bipedal gaits via periodic reward composition. Proc IEEE Int Conf on Robotics and Automation, p.7309-7315. https://doi.org/10.1109/ICRA48506.2021.9561814

Siekmann J, Green K, Warila J, et al., 2021b. Blind bipedal stair traversal via sim-to-real reinforcement learning. Proc 17th Robotics: Science and Systems, p.1-9. https://arxiv.org/abs/10.15607/rss.2021.xvii.061

Tan DCH, Zhang J, Chuah M, et al., 2023. Perceptive locomotion with controllable pace and natural gait transitions over uneven terrains. https://arxiv.org/abs/2301.10894

Tan J, Zhang TN, Coumans E, et al., 2018. Sim-to-real: Learning agile locomotion for quadruped robots. Proc 14th Robotics: Science and Systems, p.1-9. https://doi.org/10.15607/RSS.2018.XIV.010

Tsounis V, Alge M, Lee J, et al., 2020. Deepgait: Planning and control of quadrupedal gaits using deep reinforcement learning. *IEEE Robot Autom Lett*, 5(2):3699-3706. https://doi.org/10.1109/LRA.2020.2979660

Xi WT, Yesilevskiy Y, Remy CD, 2016. Selecting gaits for economical locomotion of legged robots. *Int J Robot Res*, 35(9):1140-1154. https://doi.org/10.1177/0278364915612572

Xie ZM, Da XY, Van de Panne M, et al., 2021. Dynamics randomization revisited: A case study for quadrupedal locomotion. Proc IEEE Int Conf on Robotics and Automation, p.4955-4961. https://doi.org/10.1109/ICRA48506.2021.9560837

Yang CY, Yuan K, Zhu QG, et al., 2020. Multi-expert learning of adaptive legged locomotion. *Sci Robot*, 5(49):eabb2174. https://doi.org/10.1126/scirobotics.abb2174

Zhang H, Starke S, Komura T, et al., 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Trans Graph*, 37(4):145. https://doi.org/10.1145/3197517.3201366

Unedited