Learning Robust Quadrupedal Locomotion Through Terrain Interaction via Deep Reinforcement Learning

Aman Arora, Nalini Ratha
Department of Computer Science and Engineering
University at Buffalo, The State University of New York
Buffalo, USA
{amanaror, nratha}@buffalo.edu

***Abstract*—Quadrupedal robots, with their complex walking patterns, can traverse difficult terrain where wheeled robots falter. Creating a controller for these robots presents challenges due to the necessity to adapt to diverse terrains and the complexity involved in the walking behavior. Recently, controllers learned through Deep Reinforcement Learning have made significant progress by emulating the experiential learning seen in animals. Past methods relying exclusively on proprioceptive feedback often depend on predefined motion patterns or gaits, and their reward mechanisms are often influenced by these motion patterns. Our study presents a learned controller that depends uniquely on proprioception, achieving strong locomotion in tough environments without using predefined motion patterns or gaits, relying on gait-independent rewards.**

I. Introduction

Quadrupedal animals excel in navigating challenging terrains, largely due to the complex walking patterns enabled by their advanced proprioceptive sensing. Although vision is crucial for long-term trajectory planning, effective terrain interaction is vital for achieving stable locomotion and agile movement. Traditional methods for quadrupedal locomotion involve complicated dynamic and kinematic modeling of both robots and environments, requiring significant human effort for model development and parameter optimization. Additionally, the simplifying assumptions needed for such models can restrict the robots' locomotion abilities. [1]–[4]

Recent approaches using reinforcement learning (RL) for locomotion control have shown significant advancements by training in simulations and learning through experience. [5]–[10] Among these, methods incorporating terrain height information in simulations and predicting it via exteroception in real-world settings perform better in trajectory planning and obstacle climbing [7], [10]. However, robust use of proprioception remains crucial, as exteroception using RGB cameras can fail in challenging environments or lighting conditions, and 3D lidar point clouds may misidentify obstacles like tall grass or misinterpret pliable surfaces like snow as passable. In such scenarios, relying on proprioceptive sensing through IMU and joint encoders is essential due to their lightweight and dependable nature. [7]

Additionally, locomotion methods relying solely on proprioception to navigate difficult terrains, like stairs, often use motion priors or predefined gaits [5], [11]. However, these predesigned gaits or motion priors restrict legged robots' adaptability to challenging terrains and limit the agility needed for dynamic maneuvers, especially in smaller robots [12]. Thus, a learning framework capable of achieving robust locomotion through proprioception without motion priors or predefined gaits is necessary. Recent studies suggest that by creating gait-independent rewards based on terrain interactions, it is possible to develop visually appealing gaits, or by minimizing energy use, emergent gaits can be formed [6], [12].

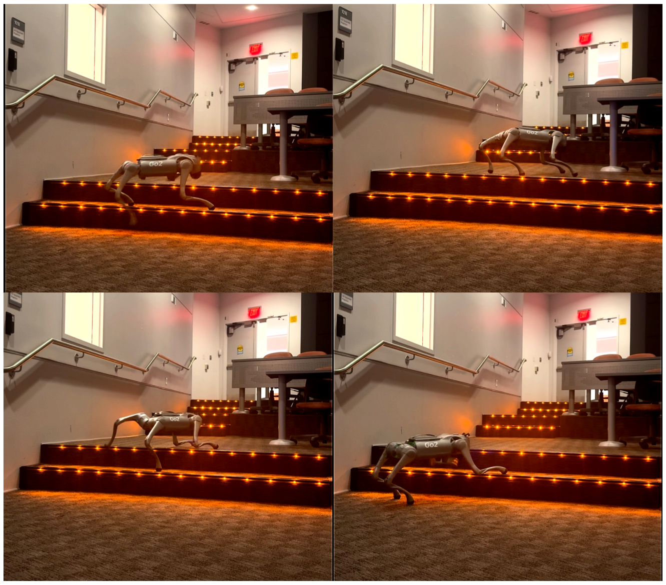

Fig. 1: Top: The Unitree GO2 robot begins ascending stairs approximately 16–17 cm in height. It first interacts with the steps to estimate their height, then adjusts its fooot trajectory accordingly to initiate the climb. Bottom: The same robot performs a descent on the same staircase.

This paper introduces a framework leveraging Deep Reinforcement Learning to develop a robust locomotion policy for quadrupedal robots, utilizing gait-independent rewards for successful transfer to real robots without exteroception. The contributions are:

1) A reinforcement learning framework designed to work from simulation to reality with gait-independent rewards, enabling robust locomotion learning without relying on external perception.
2) An improved method for enhancing Sim-to-Real transfer of learned controller results over existing techniques.

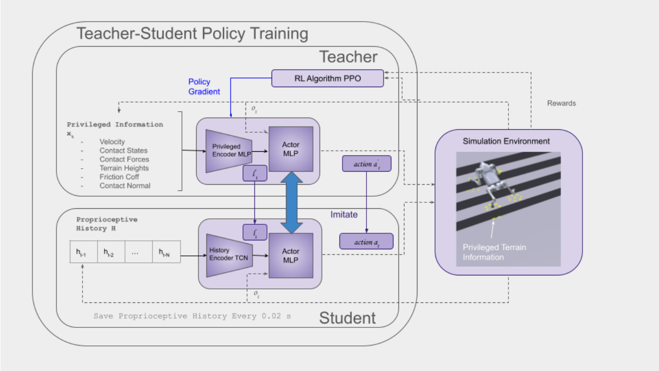

Fig. 2: Overview. By learning a locomotion policy in a simulation through gait independent rewards, the robot can walk through challenging terrains such as stairs and random obstacle and then this policy can be transferred on a real robot using Imitation Learning.

II. Method

A. Preliminaries

The environment in this paper is modeled as an infinite-horizon partially observable Markov decision process (POMDP), represented by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{H}, \mathcal{A}, d_0, p, r, \gamma)$. The continuous complete state, partial observation, and action spaces are given by $\mathbf{s} \in \mathcal{S}$, $\mathbf{h} \in \mathcal{H}$, and $\mathbf{a} \in \mathcal{A}$, respectively. The environment starts with an initial state distribution $d_0(\mathbf{s}_0)$, evolves according to the state transition probability $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, and each transition is associated with a reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$. The discount factor γ is within $[0, 1)$. At time t, a temporal history observation covering the previous T observations is described as $\mathbf{h}_t^T = \begin{bmatrix}\mathbf{h}_t & \mathbf{h}_{t-1} \dots \mathbf{h}_{t-T}\end{bmatrix}^T$. A privileged vector $\mathbf{p}_t$ denotes the subset of the full state information inaccessible to the real robot, expressed as $\mathbf{s}_t = \langle \mathbf{p}_t, \mathbf{h}_t \rangle$. Furthermore, the privileged encoding vector $\mathbf{l}_t$ represents a latent encoding of the privileged vector $\mathbf{p}_t$ and the history encoder vector $\mathbf{e}_t$, which includes a latent encoding of the temporal history observations $\mathbf{h}_t^T$.

B. Teacher Policy

We develop a teacher policy that operates with access to the complete state information within a simulated environment. While this is referred to as a full state, privileged observations are constrained to enable their latent encoding to be effectively approximated by a temporal history of partial observations, specifically proprioceptive history. For example, terrain height measurements are limited to a 10 cm radius around each foot rather than a larger rectangle (1.6 m x 1 m) utilized by others employing exteroception. We employ PPO to train the teacher policy to optimize upon a policy that achieves the maximum expected returns.

1) Actor-Critic Network: The actor or the policy network, $\pi_\theta(\mathbf{a}_t|\mathbf{l}_t, \mathbf{h}_t)$ is a neural network parameterized by θ that infers an action $\mathbf{a}_t$, privileged latent vector, $\mathbf{l}_t$ and partial observation $\mathbf{h}_t$, where $\mathbf{p}_t$ is computed using privileged encoder MLP $\mu(.)$ Formally,

$$\mu(\mathbf{p}_t) = \mathbf{l}_t \tag{1}$$

$$\pi_\theta(\mathbf{l}_t, \mathbf{h}_t) = \mathbf{a}_t \tag{2}$$

The critic network is designed similarly, but it outputs the value of the state $\mathbf{s}_t$, represented as $\mathbf{V}(\mathbf{s}_t)$.

2) Action Space: The action space is a 12×1 vector, $\mathbf{a}_t$, corresponding to the desired joint angle of the robot. To facilitate learning, we train the policy to infer the desired joint angle around the robot's default stand still pose, $\boldsymbol{\theta}_{\text{default}}$. Hence, the robot's desired joint angle is defined as

$$\boldsymbol{\theta}_{\text{target}} = \boldsymbol{\theta}_{\text{default}} + \mathbf{a}_t. \tag{3}$$

The target joint angles are converted to torques by an actuator network that has been specifically trained for the robot intended for policy deployment.

3) Rewards: We define reward functions and their coefficients as shown in the table I. The reward function is designed for following objectives:

1) To follow the given velocity and turning commands.
2) To achieve a good foot clearance for the terrain obstacles surrounding the feet.
3) To achieve a efficient and natural walking behavior.

The linear and angular velocity rewards are related to the first objectives. Instead of making it follow the exact velocity commands we give it a velocity direction in XY plane and compute MSE of the the projected velocity in that direction $\mathbf{v}_{pr}$ with 0.6. This type of velocity reward closely follows [5]. The foot clearance reward pertains to the second objective, following [5], but it is defined in a gait-independent way. A key insight is that foot swing can be determined through terrain interaction, allowing the use of this reward without motion priors or predefined gaits. We define Swing as the time period between two consecutive foot contacts with the terrain, which can be easily retrieved from the simulation environment. This swing definition has been previously utilized for feet airtime reward [6], and we extend it to the feet clearance reward. To our knowledge, this work is the first to define a gait-independent foot clearance reward.

The additional rewards focus on the third objective, with feet airtime and power consumption being pivotal for achieving a natural and efficient gait. The airtime reward encourages a more natural gait appearance, while the power consumption reward promotes symmetry and efficiency in gait. The feet airtime reward closely resembles the approach in [6], and the power consumption reward is akin to [12]. Most of the remaining reward functions are influenced by the rewards in [6].

In Table I, cmd stands for command, and i denotes a foot index. The variable $t_{air,f}$ indicates the time elapsed since the last takeoff, resetting to zero at each touchdown. In the foot slip reward, $C_{f,i}$ is the contact state for each foot. For the foot clearance reward, I_{swing} represents the number of feet in the Swing phase, and the set of collision-free feet is defined as $\mathcal{F}_{clear} = \{i : r_{f,i} > \max(H_{scan,i}), i \in I_{swing}\}$, with $H_{scan,i}$ being the set of scanned heights around the i-th foot.

We define a positive reward sum as $r_{pos} = r_v + r_\omega + \sum_{i=0}^{3} r_{air,i}$ and a negative reward sum as $r_{neg} =$

TABLE I: Reward Functions

Reward	**Expression**
Linear velocity	$r_{lv} = \begin{cases} \exp\big(-4.0\,(v_{pr}-0.6)^2\big), & v_{pr} < 0.6 \\ 1.0, & v_{pr} \geq 0.6 \\ 0.0, & \text{zero command} \end{cases}$
Angular velocity	$r_\omega = k_\omega exp(-4.0(cmd_{\omega_z} - \omega_z)^2)$
Airtime	$r_{\text{air}} = \sum_{f=0}^{4}\big(\mathbf{t}_{\text{air},f} - 0.5\big)$
Foot slip	$r_{slip,i} = k_{slip} C_{f,i} \lvert\lvert V_{f,xy,i} \rvert\rvert^2$
Foot clearance	$r_{fc} = k_{cl} \left(\sum_{i \in I_{\text{swing}}} \frac{\mathbb{1}_{\mathcal{F}_{\text{clear}}}(i)}{\lvert I_{\text{swing}} \rvert}\right) \lvert\lvert V_{xy} \rvert\rvert^{0.5}$
Orientation	$r_{ori} = k_{ori}(angle(\phi_{body,z}, \phi_{world,z}))^2$
Joint torque	$r_\tau = k_\tau \lvert\lvert \tau \rvert\rvert^2$
Joint position	$r_q = k_q \lvert\lvert q_t - q_{nominal} \rvert\rvert^2$
Joint speed	$r_{\dot{q}} = k_{\dot{q}} \lvert\lvert \dot{q}_t \rvert\rvert^2$
Joint acceleration	$r_{\ddot{q}} = k_{\ddot{q}} \lvert\lvert \dot{q}_t - \dot{q}_{t-1} \rvert\rvert^2$
Joint power	$r_P = k_P \left\lvert \boldsymbol{\tau} \cdot \dot{\boldsymbol{\theta}} \right\rvert$
Action smoothness 1	$r_{s1} = k_{s1} \lvert\lvert q_t^{des} - q_{t-1}^{des} \rvert\rvert^2$
Action smoothness 2	$r_{s2} = k_{s2} \lvert\lvert q_t^{des} - 2q_{t-1}^{des} + q_{t-2}^{des} \rvert\rvert^2$
Base motion	$r_{base} = k_{base}(0.8V_z^2 + 0.2\lvert\omega_x\rvert + 0.2\lvert\omega_y\rvert)$

Reward Coefficients					
k_v	3.0	k_{cl}	.45	$k_{\dot{q}}$, $k_{\ddot{q}}$	-6e-4,- 0.02
k_ω	1.5	k_{ori}	-3.0	k_p	-2e-5
k_a	0.15	k_τ	-6e-4	k_{s1}, k_{s2}	-2.5, -1.2
k_{slip}	-0.08	k_q	-0.75	k_{base}	-1.5

$\sum_{i=0}^{3}(r_{slip,i} + r_{cl,i}) + r_{ori} + r_\tau + r_q + r_{\dot{q}} + r_{\ddot{q}} + r_p + r_{s1} + r_{s2} + r_{base}$. The total reward is defined as

$$r_{tot} = r_{pos} \cdot exp(0.2 r_{neg}) \tag{4}$$

This form of a reward function ensures that the resulting reward is always positive and discourages the policy to choose an early termination.

4) Simulation Environment: We employed a game-inspired curriculum [6] to facilitate the progressive learning of locomotion policies across challenging terrains. The terrains included smooth, rough, random obstacles, and stair terrains, with random obstacles being the most prevalent, followed by stair terrains. The terrain features ten levels of inclination ranging from 0 to 22 degrees, with increasing difficulty. We sample terrain height points from a circle with a radius of 10 cm around each foot. We also use foot contact force and contact normal as privileged information to understand the terrain. Additionally, thigh and shank contact states, friction coefficients, and base linear velocity are utilized. A significant function of the simulation environment is defining the swing phase for each foot, for which we use foot contact states.

C. Student Policy

The student's policy network is directly replicated from the teacher policy. The student's history encoder is trained to approximate the privileged latent, enabling the student policy to produce actions that closely align with the teacher's policy. This history encoder is a TCN encoder, which processes partial observation history, specifically proprioceptive history $\mathbf{h}_t^T$, and generates a history encoding $\mathbf{e}_t$. Training from teacher to student employs Imitation Learning conducted online using DAgger. The history encoder is trained with the MSE loss between privileged and history encoding, as well as between student and teacher actions.

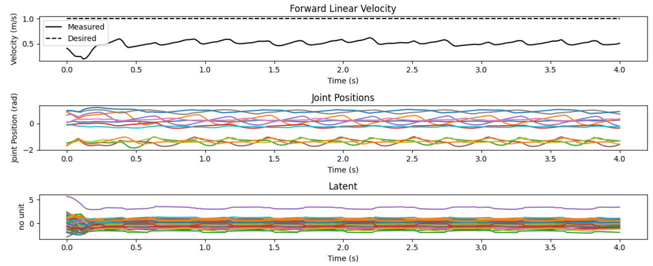

Fig. 3: Actual Velocity, Joint Positions, and Privileged Latent measured from teacher policy deployed in simulation

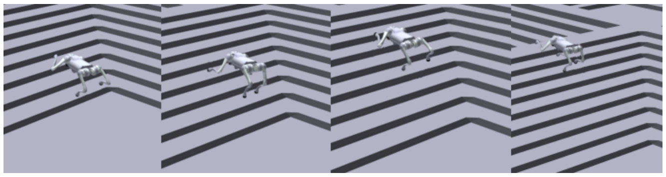

Fig. 4: Teacher Policy deployed in simulation climbing stairs.

$$\mathcal{L} := (\bar{a}_t(l_t, h_t) - a_t(e_t, h_t))^2 + (\bar{l}_t(p_t) - l_t(h_t^T))^2. \tag{5}$$

Quantities indicated with a bar $(\bar{\cdot})$ represent the target values derived from the teacher. We utilize the dataset aggregation method (DAgger) [13]. In particular, training data is produced by executing trajectories using the student policy. For each state encountered, the teacher policy calculates its embedding and action vectors $(\bar{\cdot})$. These outputs from the teacher policy serve as guidance for the respective states.

III. Experiments and Results

We conducted experiments using both the teacher and student policies in a simulation environment and deployed the student policy on a physical robot.

A. Simulation

1) Teacher Policy: In the simulation, the teacher policy exhibits natural walking behavior, where the feet are lifted to a small height on a flat surface. When approaching an obstacle or stair, the feet are raised appropriately for climbing. The teacher policy in the simulation can ascend stairs up to 23 cm in height with an inclination of up to 38 degrees and can navigate random obstacles up to 30 cm high.

2) Student Policy: In simulations, the student policy displays a natural walking behavior, lifting its feet slightly when on flat surfaces. Upon approaching an obstacle or stair, it estimates the height through repeated contact and adjusts its steps accordingly to climb. Better behavior cloning is achieved

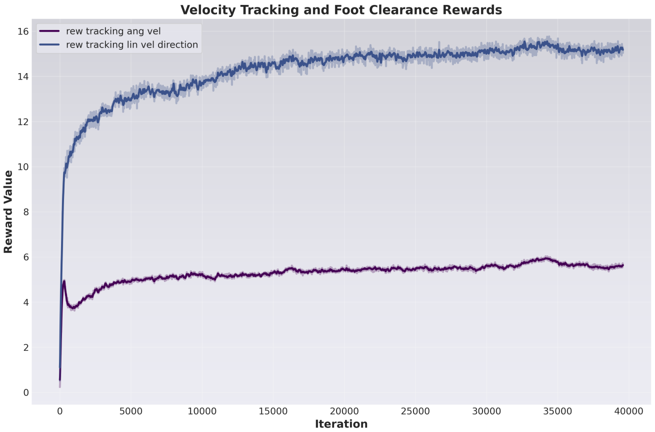

Fig. 5: Velocity and foot clearance rewards acheived in simulation

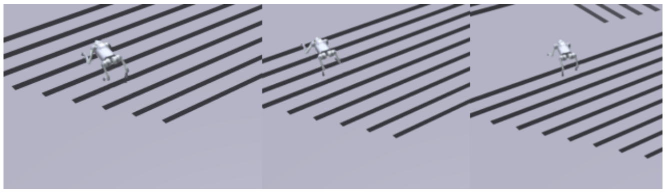

Fig. 6: Student Policy deployed in simulation climbing stairs

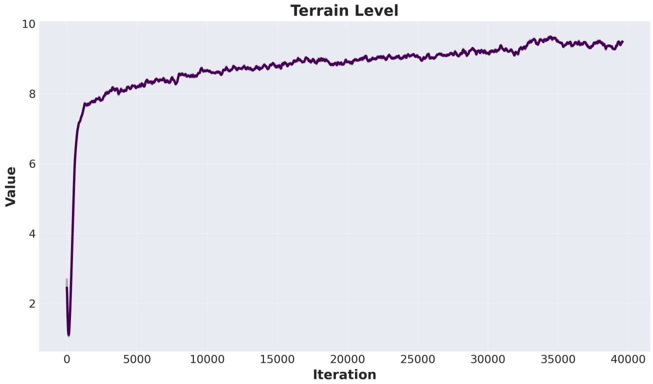

Fig. 7: Mean Terrain Levels achieved during training

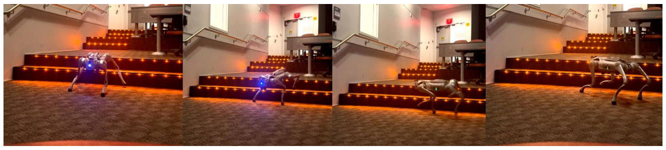

Fig. 8: Policy deployed on real robot demonstrating the emergent behavior

by incorporating an additional linear layer and layer normalization after the TCN encoder, which improves convergence to approximate the privileged latent. In simulations, the student policy successfully climbs stairs up to 18 cm high with an inclination of up to 30 degrees and navigates random obstacles up to 23 cm.

3) Emergent Behavior: The learned policy exhibits interesting emergent behavior when the robot is commanded to climb stairs backward. It climbs the first stair normally, but upon reaching the second stair, it quickly turns its head 180 degrees to face the terrain. This adjustment may help prevent falls by providing better stability during backward climbing. We believe this occurs because the robot's perception is limited to its feet, as it cannot use height points beyond a small circle around them. With no motion priors, the locomotion is guided solely by rewards, allowing the policy to choose behaviors that maximize survival or returns. This specific behavior is absent with discrete obstacles or the first stair, only manifesting once it encounters the second step.

B. Real Robot

The Sim-to-Real transfer is effective due to our comprehensive domain randomization, resulting in the deployed policy on the actual robot demonstrating comparable natural and symmetrical walking behavior.

1) Obstacles and Stairs: Upon encountering an obstacle, the robot demonstrates a similar behavior by assessing the terrain's height and adjusting its gait and foot clearance accordingly. It was tested with a wooden block 10 cm high, which it can climb from any direction with ease. When tested on stairs, the robot can ascend steps between 15 and 18 cm in height. If the steps have an inclination of more than 25 to 30 degrees, it discontinues climbing and maintains balance effectively. In such cases, it can ascend the stairs sideways or at an angle by placing its body on the last step before making the next ascent.

2) Random Push and Friction: The robot is capable of adapting effectively to random pushes and varying friction conditions. We evaluated its performance by pushing it multiple times during walking, demonstrating its resilience to such disturbances. Additionally, it adjusts well to friction changes, as evidenced by its behavior on different surfaces like high-friction carpets and low-friction tiles.

3) Emergent Behavior: The real robot exhibits similar emergent behavior by quickly turning its head towards the stairs when encountering the second stair.

IV. Conclusion

This research demonstrates that effective quadrupedal locomotion relying solely on proprioception can be accomplished without using motion priors by leveraging gait-independent interaction-based rewards. However, a limitation of this approach is that the resulting gaits lack visual appeal and merely focusing on energy minimization is insufficient in challenging training environments. A promising future direction is to test different foot clearance rewards that are known to produce natural-looking gaits. Another area for improvement is incorporating a command curriculum to enhance terrain exploration and refine the robot's control precision.

References

[1] Y. Kim, B. Yu, E. M. Lee, J.-H. Kim, H.-W. Park, and H. Myung, "STEP: State estimator for legged robots using a preintegrated foot velocity factor," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4456–4463, 2022.

[2] M. Bloesch, C. Gehring, P. Fankhauser, M. Hutter, M. A. Hoepflinger, and R. Siegwart, “State estimation for legged robots on unstable and slippery terrain,” in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 6058–6064.

[3] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch *et al.*, “ANYmal – A highly mobile and dynamic quadrupedal robot,” in *Proc. IEEE/RSJ international Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 38–44.

[4] C. Gehring, C. D. Bellicoso, P. Fankhauser, S. Coros, and M. Hutter, “Quadrupedal locomotion using trajectory optimization and hierarchical whole body control,” in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4788–4794.

[5] 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.

[6] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, “Learning to walk in minutes using massively parallel deep reinforcement learning,” in *Proc. Conference on Robot Learning (CoRL)*, 2022, pp. 91–100.

[7] 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.

[8] Z. Fu, A. Kumar, A. Agarwal, H. Qi, J. Malik, and D. Pathak, “Coupling vision and proprioception for navigation of legged robots,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 273–17 283.

[9] W. Yu, D. Jain, A. Escontrela, A. Iscen, P. Xu, E. Coumans, S. Ha, J. Tan, and T. Zhang, “Visual-locomotion: Learning to walk on complex terrains with vision,” in *Proc. Conference on Robot Learning (CoRL)*, 2021, pp. 1291–1302.

[10] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, “Legged locomotion in challenging terrains using egocentric vision,” in *Proceedings PMLR Conference on Robot Learning (CoRL)*, 2022.

[11] J. Siekmann, K. Green, J. Warila, A. Fern, and J. W. Hurst, “Blind bipedal stair traversal via sim-to-real reinforcement learning,” *CoRR*, vol. abs/2105.08328, 2021.

[12] Z. Fu, A. Kumar, J. Malik, and D. Pathak, “Minimizing energy consumption leads to the emergence of gaits in legged robots,” in *Proc. Conference on Robot Learning (CoRL)*, 2021, pp. 928–937.

[13] D. B. S. Ross, G. Gordon, “A reduction of imitation learning and structured prediction to no-regret online learning,” in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 627–635.