Reinforcement Learning for Humanoid Locomotion

Qiao Sun

***Abstract*— This report investigates the use of reinforcement learning (RL) to develop bipedal gaits in humanoid robots, emphasizing stability and adaptability on diverse terrains. It examines three methods: 'Sim-to-Real Learning of Bipedal Gaits,' 'Rapid Motor Adaptation for Bipedal Robots,' and 'Learning Humanoid Locomotion with Transformers.' These approaches utilize Proximal Policy Optimization (PPO) in simulations, applying dynamics randomization to counter real-world disturbances. The aim is to detail these methods, their implementation, and their impact on bipedal robot mobility, contributing to robotics research.**

I. INTRODUCTION

The quest for bipedal robots capable of moving with human-like agility and precision across different terrains has evolved significantly since the first humanoid robot in the 1970s [1]. Advances in control systems [2]–[5] and optimization techniques [6]–[9] have improved robot dynamics. Reinforcement Learning (RL) has been pivotal since the 1990s in enhancing locomotion for both quadrupeds [10]–[13] and bipedal models.

This report reviews three RL strategies enhancing bipedal movement: 'Sim-to-Real Learning of Bipedal Gaits' [14], [15], aimed at generating versatile gaits; 'Rapid Motor Adaptation for Bipedal Robots' [16], which enhances terrain adaptability; and 'Learning Humanoid Locomotion with Transformers' [17], [18], focusing on an end-to-end learning model. These methods demonstrate progress towards adaptable and intelligent systems for real-world applications. The review outlines these advancements and their relevance to future robotics.

II. METHODS

This section delves into several reinforcement learning (RL) methods aimed at improving humanoid robot mobility. It highlights significant contributions in this area.

A. *Sim-to-Real Learning for Bipedal Gaits*

1) Approach Overview: Traditional RL approaches for bipedal gait learning face significant challenges, including limited adaptability and inefficient tuning processes. Reference-based methods [19]–[23] can achieve certain gait characteristics but fail to provide the needed flexibility for developing robust gaits. On the other hand, reference-free methods, exemplified by OpenAI Gym [6], often lead to tedious tuning, rarely attaining optimal results.

To overcome these issues, they introduce a sim-to-real RL framework [14] for the autonomous learning of common bipedal gaits. This framework eliminates the reliance on reference motions, opting instead for simple, periodic cost functions based on kinetic and dynamic criteria to tailor gait-specific rewards. This approach facilitates easy parameter adjustment for various gaits, ensuring rewards align with desired gait behaviors.

A key innovation is the use of periodic reward functions that resonate with the cyclic nature of bipedal movements. By targeting the dynamics of swing and stance phases with a probabilistic method for applying cost functions at specific intervals, the framework enhances gait learning. This method dispenses with the need for external references or benchmarks, representing a significant step forward in robotics by enabling more natural bipedal locomotion, in Figure 1.

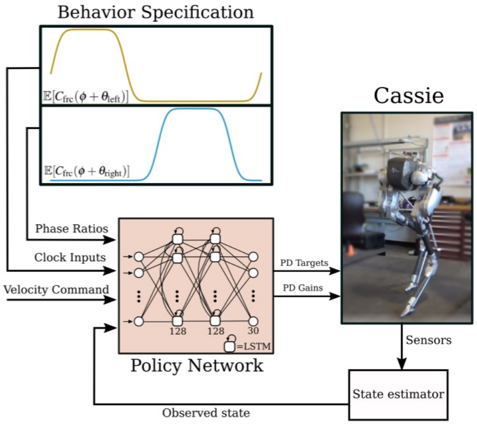


Fig. 1. Illustration of the reward design framework enabling the learning of locomotion behaviors such as standing, walking, and running in robots. This framework incorporates gait parameters into an LSTM policy to generate PD joint position targets and gains, optimizing robotic movements.

2) Problem Formulation: The study addresses the challenge of sim-to-real transfer in RL for bipedal gait learning. It models the problem as a discrete-time Markov Decision Process (MDP) with continuous states and actions (S and A), a transition function $T(s, a, s')$, and a dynamic reward function $R(s, t)$. The goal is to develop a control policy $\pi(a|s)$ that maximizes the expected T-horizon discounted return, $J(\pi)$:

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(S_t, t)\right] \tag{1}$$

where γ is the discount factor, and S_t represents the state at time t under policy π and transition dynamics T.

3) Learning Bipedal Gaits with Periodic Reward Composition: At the core of the method is the periodic reward function, $R(s, \phi)$, which varies with the cycle time Φ, cycling from 0 to 1. This allows for cyclic application of costs:

$$R(s, \phi) = \beta + \sum_i R_i(s, \phi) \tag{2}$$

Each component $R_i(s, \phi)$ includes a phase coefficient c_i, a phase indicator $I_i(\phi)$, and a measurement of phase-specific reward $q_i(s)$. This setup facilitates phase-focused reward adjustments, aiming to minimize foot forces during swing phases for better foot lifting and reduce velocities during stance phases for improved stability.

4) Describing Bipedal Gaits: The framework categorizes bipedal gaits by modifying rewards for swing and stance phases, crucial for bipedal movement. These phases may occur alternately or simultaneously for the left and right feet. For instance, walking involves alternating these phases between legs for continuous motion, whereas hopping engages both feet simultaneously, as depicted in 2.

The expected total reward from bipedal actions is computed by combining phase-specific rewards for each foot, factoring in phase offsets (θ_{left} and θ_{right}):

$$\begin{aligned} \mathbb{E}[R_{\text{bipedal}}(s, \phi)] = {} & \mathbb{E}[C_{\text{frc}}(\phi + \theta_{\text{left}})] \cdot q_{\text{left frc}}(s) \\ & + \mathbb{E}[C_{\text{frc}}(\phi + \theta_{\text{right}})] \cdot q_{\text{right frc}}(s) \\ & + \mathbb{E}[C_{\text{spd}}(\phi + \theta_{\text{left}})] \cdot q_{\text{left spd}}(s) \\ & + \mathbb{E}[C_{\text{spd}}(\phi + \theta_{\text{right}})] \cdot q_{\text{right spd}}(s) \end{aligned} \tag{3}$$

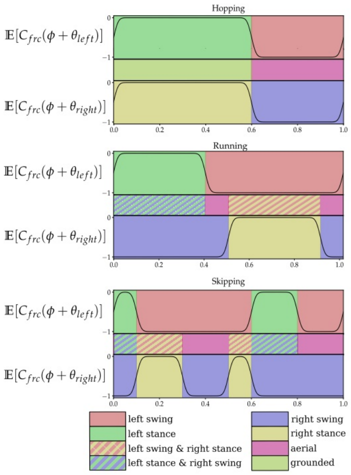


Fig. 2. Expected force phase coefficient for various bipedal gaits. Hopping displays negligible phase shift ($|\theta_{\text{left}} - \theta_{\text{right}}| \approx 0$) between feet. Walking and running show a moderate phase shift ($|\theta_{\text{left}} - \theta_{\text{right}}| \approx 0.5$), indicative of a galloping transition. Skipping requires four phases and mirrors the moderate shift ($|\theta_{\text{left}} - \theta_{\text{right}}| \approx 0.5$). Shaded areas represent hybrid phases combining behaviors of both feet.

5) Model Architecture: they use a Long Short-Term Memory (LSTM) network with two 128-unit layers, chosen for sequence handling abilities essential for bipedal locomotion. This network tracks the robot's state, inputs, and gait phases, outputting 10 joint positions and their PD gains at 40Hz, matching the robot's PD controllers at 2000Hz for dynamic posture adjustments.

The state space encompasses cycle time ϕ, cycle offsets (θ_{left} and θ_{right}), and gait timing, with sine wave clocks and a phase duration ratio vector aiding in phase management and leg coordination.

6) Policy Learning and Optimization: they apply Proximal Policy Optimization (PPO) for training, with a batch size of 32 trajectories, a learning rate of 0.0001, and a replay buffer of 50,000 samples. Training spans 4 epochs per iteration over 150,000,000 samples, typically taking 24 to 36 hours using the cassie-mujoco-sim based on MuJoCo.

7) Reward Function Design: The reward function is formulated as:

$$\mathbb{E}[R(s, \phi)] = \mathbb{E}[R_{\text{bipedal}}(s, \phi)] + R_{\text{smooth}}(s) + R_{\text{cmd}}(s) + \beta \tag{4}$$

where

$$\begin{aligned} R_{\text{cmd}}(s) &= (-1) \cdot q_{i_x}(s) \qquad (5) \\ &+ (-1) \cdot q_{i_y}(s) \qquad (6) \\ &+ (-1) \cdot q_{\text{orientation}}(s) \qquad (7) \\ R_{\text{smooth}}(s) &= (-1) \cdot q_{\text{action diff}}(s) \qquad (8) \\ &+ (-1) \cdot q_{\text{torque}}(s) \qquad (9) \\ &+ (-1) \cdot q_{\text{pelvis acc}}(s) \qquad (10) \end{aligned}$$

$\mathbb{E}[R_{\text{bipedal}}(s, \phi)]$ quantifies the expected reward for bipedal gait patterns, emphasizing the cyclic aspects of movement.

$R_{\text{smooth}}(s)$ imposes penalties on sudden movements and high joint torques to encourage smoother, energy-efficient motions.

$R_{\text{cmd}}(s)$ rewards the robot's compliance with specified velocities and orientations, ensuring precise execution of commands.

with penalties for abrupt movements and rewards for adhering to velocities and orientations, balanced by beta to encourage desirable actions and deter undesired ones.

8) Stair Climbing Extension: they extend the RL framework to stair climbing by incorporating terrain randomization, including varied stair dimensions. This enhances proprioceptive control for autonomous stair navigation in ascent and descent, demonstrated in 3.

B. Adapting Rapid Motor Adaptation for Bipedal Robots

1) Approach Overview: The Adapted Rapid Motor Adaptation (A-RMA) algorithm [16], an extension of the RMA framework developed for quadrupeds, facilitates bipedal robot support. It starts with a base policy [26], [27] informed by gait references, streamlining the initial learning phase. A-RMA [28] enhances policy refinement through Proximal Policy Optimization (PPO), keeping the adaptation module constant, thus improving bipedal adaptability in varied environments, as illustrated in 4.

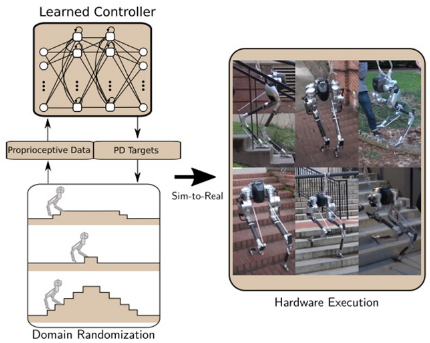


Fig. 3. Methodology for training stair navigation policies without visual cues, improving proprioceptive adaptability to ground height changes for versatile application.

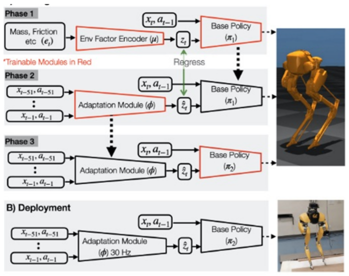


Fig. 4. Illustration of A-RMA's training and deployment phases, following RMA's framework [25] with an additional phase for fine-tuning the base policy using PPO, maintaining the adaptation module unchanged. This step precisely addresses external estimation inaccuracies.

2) Bootstrapping the Base Policy: Initialization of the bipedal robots' base policy utilizes Hybrid Zero Dynamics (HZD)-based gaits, offering a variety of walking patterns. The policy π, from state x_t and vector $z \in \mathbb{R}^8$, predicts action a_t for joint target positions, translated into torque via PD control. Vector z_t, a condensed environmental representation by function μ, contains essential adaptation data:

$$
\begin{aligned}
z_t &= \mu(e_t), \\
a_t &= \pi_1(x_t, a_{t-1}, z_t).
\end{aligned}
$$

This process is enhanced by implementing μ and π_1 as Multi-Layer Perceptrons (MLPs), jointly trained through model-free reinforcement learning to optimize the policy's expected return:

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right],$$

where τ represents the agent's trajectory under policy π_1.

3) Fine-tuning with PPO and Fixed Adaptation Module: Incorporation of the adaptation module ϕ aids in base policy refinement against external estimate inaccuracies. Proximal Policy Optimization (PPO) is applied for fine-tuning the base policy π, improving responsiveness to external variances without altering the adaptation module, thus bolstering the policy's real-world adaptability.

4) Reward Function Design: The reward function is crafted to optimize motor precision, pelvis imitation, efficiency, and reaction forces, including:

1) Motor imitation: $\exp\left(-\rho_1 \left\|q_m^r - \hat{q}_m\right\|_2^2\right)$
2) Pelvis position imitation: $\exp\left(-\rho_2 \left\|q_p - \hat{q}_p\right\|_2^2\right)$
3) Pelvis velocity imitation: $\exp\left(-\rho_3 \left\|\dot{q}_p^r - \dot{\hat{q}}_p\right\|_2^2\right)$
4) Pelvis Rot-imitation: $\exp\left(-\rho_4\left(1 - \cos\left(q_T^r - \hat{q}_T^r\right)\right)\right)$
5) Rotational velocity imitation: $\exp\left(-\rho_5 \left\|\dot{q}_T^r - \dot{\hat{q}}_r\right\|_2^2\right)$
6) Torque penalty: $\exp\left(-\rho_6 \left\|u\right\|_2^2\right)$
7) Ground reaction force penalty: $\exp\left(-\rho_7 \left\|F\right\|_2^2\right)$

C. Learning Humanoid Locomotion with Transformers

1) Approach Overview: This method applies Transformers for dynamic humanoid locomotion, targeting direct real-world application [29]–[37]. Successful in various domains, Transformers now address robotics challenges, such as behavior cloning and depth integration, through online reinforcement learning, bypassing the need for offline datasets.

Simulation training teaches the robot to refine movements and adapt to different terrains by processing observation and action sequences. Initially, it optimizes a state policy in a fully observable simulation, then transitions to an observation policy using KL divergence, improving environmental adaptability, as shown in 5.

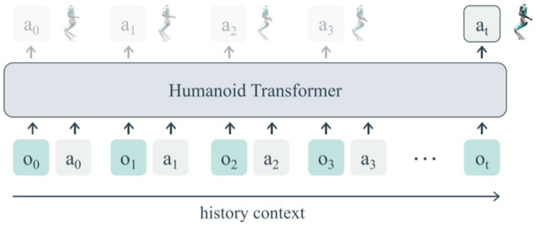


Fig. 5. Exploring blind bipedal locomotion limits, this study introduces a training pipeline for generating policies enabling real-world stair ascent and descent without visual input. These policies develop proprioceptive reflexes for substantial ground height disturbances, achieving robust performance across diverse environments.

2) Problem Formulation: Humanoid locomotion is framed within an MDP, expanded into a Partially Observable MDP for real-world complexity.

MDP to POMDP Transition:

- State Space(S): Describes potential states a robot might encounter, including various environmental configurations.
- Action Space(A): Enumerates actions available to the robot, spanning a comprehensive range of movement options for navigation.
- Transition Function$P(s_{t+1}|s_t, a_t)$:
  Determines the likelihood of moving from one state to another after an action is taken, reflecting environmental variability.

- Reward Function: $R(s_{t+1}|s_t, a_t)$:
  Assigns a scalar value to each transition, assessing the action's impact.
- Observation Space(O): Captures the robot's partial insights into the current state, considering sensor limitations and environmental factors.
- Observation Function$Z(o_t|s_t)$:
  Manages uncertainties in observations, crucial for operation in partially observable settings.

3) Model Architecture:

Transformer Model Implementation: Transformers [39], adapted for robotics, process observation-action sequences in POMDPs, using self-attention for action prediction accuracy and direct applicability to locomotion without offline data.

Self-Attention Mechanism: The self-attention mechanism in Transformers prioritizes relevant input features, identifying key movement strategies. It incorporates sinusoidal positional encodings to maintain temporal context, enhancing future action predictions.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

4) Policy Learning and Optimization:

State Policy Development: The initial step is to develop a state policy (π_s) in a simulated environment, focusing on state-driven actions. This process is enhanced by adjusting reward functions for improved gait and interaction with the environment.

Observation Policy Refinement: Refinement to an observation policy is achieved by applying Kullback-Leibler (KL) divergence, transitioning from a state-dependent to a partially observable decision-making framework. This process mirrors the teacher-student learning model [11], [12], adapting the policy for complexities and uncertainties encountered outside simulated conditions.

Joint Optimization Approach: A joint optimization strategy addresses the discrepancies between state and observation spaces, aligning decision-making processes. It merges reinforcement learning (RL) loss with state-policy supervision into a comprehensive objective function that harmonizes these aspects, ensuring an integrated approach to policy development.

$$L(\pi_o) = L_{RL}(\pi_o) + \lambda D_{KL}(\pi_o||\pi_s) \tag{12}$$

balances RL loss and the divergence between the observation and state policies, with the weighting factor.

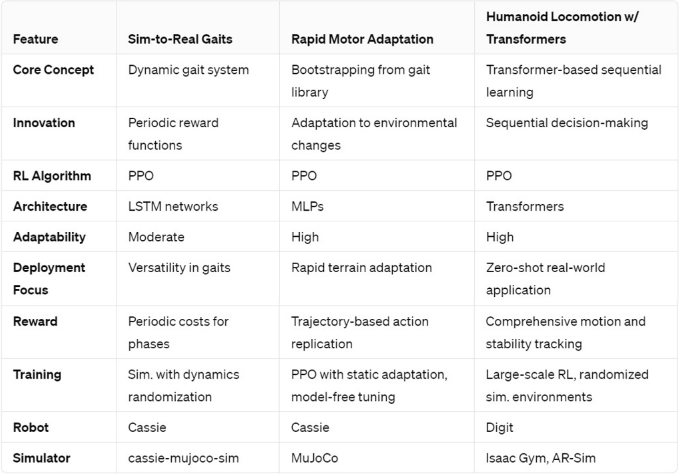
5) Reward Function Design: The reward function for robot locomotion covers dynamics and environmental interactions, focusing on velocity, stability, and efficient movement. It includes metrics for velocity tracking, stability, ground interaction, and joint efficiency, aiming for smooth and realistic motion.

$$r_v := \exp\left(-\frac{1}{2\sigma_{xy}}\left\|v_{xy} - v^*_{xy}\right\|^2\right) \tag{13}$$

$$r_{bm} := v_z^2 + 0.5 * \|\omega_{xy}\|^2 \tag{14}$$

$$r_f := \sum_{i \in \text{foot}} \begin{cases} \|F(i) - f_{\max}\|^2 & \text{if } \|F(i)\|^2 > f_{\max} \\ 0 & \text{if } \|F(i)\|^2 \leq f_{\max} \end{cases} \tag{15}$$

III. DISCUSSION



Feature	Sim-to-Real Gaits	Rapid Motor Adaptation	Humanoid Locomotion w/ Transformers
Core Concept	Dynamic gait system	Bootstrapping from gait library	Transformer-based sequential learning
Innovation	Periodic reward functions	Adaptation to environmental changes	Sequential decision-making
RL Algorithm	PPO	PPO	PPO
Architecture	LSTM networks	MLPs	Transformers
Adaptability	Moderate	High	High
Deployment Focus	Versatility in gaits	Rapid terrain adaptation	Zero-shot real-world application
Reward	Periodic costs for phases	Trajectory-based action replication	Comprehensive motion and stability tracking
Training	Sim. with dynamics randomization	PPO with static adaptation, model-free tuning	Large-scale RL, randomized sim. environments
Robot	Cassie	Cassie	Digit
Simulator	cassie-mujoco-sim	MuJoCo	Isaac Gym, AR-Sim

Sim-to-real Gaits Using LSTM networks, this method allows humanoid robots to learn gait patterns without predefined trajectories, reflecting the cyclic nature of bipedal locomotion. It produces hopping, walking, and running behaviors by adjusting ratios and cycle offsets, showing the method's flexibility.

In further developments, training included randomized stair scenarios with the same reward function. Analysis showed that stair-trained policies, while increasing foot clearance and leg angle retraction speed, also led to higher energy use, indicating a balance between training strategy adaptability and energy efficiency.

Outdoor experiments demonstrated the practical effectiveness of these policies across varied terrains. Stair-trained LSTM policies were particularly effective in stair navigation, achieving high success in ascending and descending, even without visual input. These policies adeptly managed tasks like hopping onto sidewalks and traversing uneven surfaces, demonstrating smooth gait transitions and ground slope adaptation, emphasizing effective force management.

Rapid Motor Adaptation employs a gait library to facilitate adaptive learning, ensuring immediate operational capability and adaptability to various terrains. This approach provides stability and enables environmental adaptation.

In field tests with a Cassie bipedal robot, the A-RMA policy adapted to diverse conditions without post-simulation fine-tuning. It effectively managed tasks such as following dynamic commands on standard terrain and towing heavy loads, demonstrating stability and directional control under varying forces.

On slippery surfaces, A-RMA maintained stability and accurately followed commands, despite significant slips and contact changes. This highlights the policy's capability to adjust to rapid surface condition changes.

In uneven terrain tests, including soft foams and wooden planks, A-RMA navigated without imbalance, adapting to variations in surface softness and contact, showing resilience to untrained conditions.

Humanoid Locomotion with Transformers leverages Transformer models for improved movement prediction by analyzing locomotion dynamics. This technique demonstrated stability and adaptability in real-world tests, managing varying loads while maintaining stable gait, indicating balance and weight adaptability.

The controller navigated diverse terrains, from varying friction levels to obstacle-laden paths, without external sensory input, showing durability and adaptability in complex environments.

Notably, it exhibited dynamic adaptation in new challenges, like climbing untrained steps and recovering from simulated motor malfunctions. The robot's ability to adjust leg movements for step climbing and recover from malfunctions underscores its dynamic learning and adjustment capabilities based on historical data.

Commonalities:
All three methods employ Proximal Policy Optimization (PPO) as their core learning algorithm, based on PPO's capability to efficiently manage the high-dimensional action spaces common in humanoid robotics, ensuring stable progress in learning.

To address sim-to-real discrepancies, each approach incorporates some form of Dynamics Randomization. This technique diversifies the simulation dynamics by randomizing physical parameters like damping, mass, and friction at the start of each training episode. Such randomization prevents overfitting to the simulated environment, facilitating a smoother transition to real-world applications. This strategy acknowledges the inevitable modeling errors in simulations and ensures robustness against them by exposing the learning algorithms to a wider range of dynamic conditions.

None of the methods rely on exteroceptive sensors, such as cameras, for terrain estimation. This design choice stems from the recognition that robots need to operate reliably in diverse and unpredictable real-world conditions, where reliance on such sensors could introduce vulnerability to occlusions, lighting variations, or other environmental factors. Instead, these approaches emphasize proprioception, aiming for a level of robustness that allows for effective navigation across a broad spectrum of human environments without detailed external sensory input.

Difference:
Adaptation Strategies to Diverse Terrains:

The Sim-to-Real Gaits method excels in pre-simulated, predictable environments by directly tuning to specific terrains, offering precise adaptability within known parameters. Its strength lies in customizing locomotion for anticipated scenarios, ensuring optimal performance in familiar settings but struggling with unexpected terrain changes.

Rapid Motor Adaptation (RMA) utilizes a gait library for flexibility, enabling swift adaptation to unpredictable terrains by leveraging a wide array of pre-learned movements. This approach combines the reliability of predefined gaits with on-the-fly learning's adaptability, making it suitable for environments with both known and novel terrain types.

Humanoid Locomotion with Transformers adopts an end-to-end learning strategy from historical data, dynamically adapting to new terrains. By analyzing complex patterns over time, this method provides nuanced adjustments to unforeseen environments, highlighting its capability for forward-looking adaptability without prior explicit programming for each potential scenario.

In summary, these methods illustrate different approaches to navigating the complex landscape of robotic locomotion. Sim-to-Real Gaits focuses on precision for known terrains, RMA balances pre-defined reliability with adaptive learning for mixed environments, and Transformers offer a forward-thinking solution for unpredictable terrain changes. Each strategy reflects a unique balance between leveraging known gait patterns and adapting dynamically to the challenges of real-world mobility.

IV. CONCLUSIONS

This report synthesizes findings from key studies on ”Reinforcement Learning for Humanoid Locomotion,” focusing on three innovative strategies: Sim-to-Real Learning, Rapid Motor Adaptation, and Learning with Transformers. It outlines progress in increasing the adaptability and stability of bipedal robots, with each method addressing distinct aspects of navigation in real-world conditions. Sim-to-real learning ensures the effective application of simulation-trained behaviors, Rapid Motor Adaptation emphasizes quick adaptability to changing terrains, and Learning with Transformers offers sophisticated decision-making capabilities for dynamic locomotion. This analysis contributes to the scholarly discussion and sets a foundation for future investigations into optimizing humanoid robot movement.

V. FUTURE PERSPECTIVES

A. *Integration of Multi-modal Sensory Data*

Future research should explore incorporating visual and tactile feedback into reinforcement learning algorithms, aiming to enhance robots' environmental perception and navigation abilities. This approach could bring robots closer to human sensory processing, potentially increasing their autonomy and adaptability in complex environments.

B. Parallel Reinforcement Learning Across Multiple Robots

Investigating parallel reinforcement learning among multiple robots could accelerate learning and develop more robust locomotion strategies. This collective experience and data sharing could improve learning algorithm efficiency and behavior generalization, enhancing humanoid robots' scalability and real-world applicability.

REFERENCES

[1] Ichiro Kato. Development of about 1. Biomechanism, 1973.

[2] Steve Collins, Andy Ruina, Russ Tedrake, and Martijn Wisse. Efficient bipedal robots based on passive-dynamic walkers. Science, 2005.

[3] Shuuji Kajita, Fumio Kanehiro, Kenji Kaneko, Kazuhito Yokoi, and Hirohisa Hirukawa. The 3d linear inverted pendulum mode: A simple modeling for a biped walking pattern generation. In IROS, 2001.

[4] Marc H Raibert. Legged robots that balance. MIT Press, 1986.

[5] Eric R Westervelt, Jessy W Grizzle, and Daniel E Koditschek. Hybrid zero dynamics of planar biped walkers. IEEE transactions on automatic control, 2003.

[6] Jared Di Carlo, Patrick M Wensing, Benjamin Katz, Gerardo Bledt, and Sangbae Kim. Dynamic locomotion in the mit cheetah 3 through convex model-predictive control. In IEEE/RSJ international conference on intel- ligent robots and systems (IROS), 2018.

[7] Scott Kuindersma. Recent progress on atlas, the world's most dynamic humanoid robot, 2020. URL https://youtu. be/EGABAx52GKI.

[8] Scott Kuindersma, Robin Deits, Maurice Fallon, Andr´es Valenzuela, Hongkai Dai, Frank Permenter, Twan Koolen, Pat Marion, and Russ Tedrake. Optimization- based locomotion planning, estimation, and control de- sign for the atlas humanoid robot. Autonomous robots, 2016.

[9] Yuval Tassa, Tom Erez, and Emanuel Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In IROS, 2012.

[10] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. Science Robotics, 2019.

[11] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. RSS, 2021.

[12] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. Science robotics, 2020.

[13] Gabriel B Margolis and Pulkit Agrawal. Walk these ways: Tuning robot control for generalization with mul- tiplicity of behavior. CoRL, 2022.

[14] Jonah Siekmann, Yesh Godse, Alan Fern, and Jonathan Hurst. Sim-to-real learning of all common bipedal gaits via periodic reward composition. In ICRA, 2021.

[15] Jonah Siekmann, Kevin Green, John Warila, Alan Fern, and Jonathan Hurst. Blind bipedal stair traversal via sim- to-real reinforcement learning. RSS, 2021.

[16] Ashish Kumar, Zhongyu Li, Jun Zeng, Deepak Pathak, Koushil Sreenath, and Jitendra Malik. Adapting rapid motor adaptation for bipedal robots. In IROS, 2022.

[17] Radosavovic I, Xiao T, Zhang B, Darrell T, Malik J, Sreenath K. Learning Humanoid Locomotion with Transformers. arXiv preprint arXiv:2303.03381. 2023 Mar 6.

[18] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-World Humanoid Locomotion with Reinforcement Learning." arXiv, Dec. 14, 2023.

[19] H. Zhang, S. Starke, T. Komura, and J. Saito, "Mode-adaptive neural networks for quadruped motion control," ACM Transactions on Graphics, vol. 37, no. 4, pp. 1–11, 2018.

[20] S. Starke, Y. Zhao, T. Komura, and K. Zaman, "Local motion phases for learning multi-contact character movements," ACM Transactions on Graphics, vol. 39, no. 4, 2020.

[21] Z. Xie, P. Clary, J. Dao, P. Morais, J. Hurst, and M. van de Panne, "Learning locomotion skills for cassie: Iterative design and sim-to-real," L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., ser. Proceedings of Machine Learning Research, vol. 100, PMLR, 2020.

[22] J. Siekmann, S. Valluri, J. Dao, L. Bermillo, H. Duan, A. Fern, and J. Hurst, "Learning Memory-Based Control for Human-Scale Bipedal Locomotion," 2020.

[23] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "DeepMimic," ACM Transactions on Graphics, vol. 37, no. 4, pp. 1–14, 2018.

[24] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," arXiv preprint arXiv:1606.01540, 2016.

[25] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "RMA: Rapid Motor Adaptation for Legged Robots," in Robotics: Science and Systems, 2021.

[26] J. W. Grizzle, C. Chevallereau, A. Ames, and R. Sinnet, "3d bipedal robotic walking: Models, feedback control, and open problems," in IFAC Symposium on Nonlinear Control Systems, 2010

[27] X. Da, O. Harib, R. Hartley, B. Griffin, and J. W. Grizzle, "From 2d design of underactuated bipedal gaits to 3d implementation: Walking with speed tracking," IEEE Access, vol. 4, pp. 3469–3478, 2016.

[28] Z. Fu, A. Kumar, J. Malik, and D. Pathak, "Minimizing energy consumption leads to the emergence of gaits in legged robots," Conference on Robot Learning, 2021.

[29] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Sub- biah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee- lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. NeurIPS, 2020.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2018.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[32]] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transform- ers. In ECCV, 2020.

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[34] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real- world robot learning with masked visual pre-training. arXiv:2210.03109, 2022.

[35] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real- world robot learning with masked visual pre-training. arXiv:2210.03109, 2022.

[36] Ruihan Yang, Minghao Zhang, Nicklas Hansen, Huazhe Xu, and Xiaolong Wang. Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers. arXiv:2107.03996, 2021.

[37] Anthony Brohan, Noah Brown, Justice Carbajal, Yev- gen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv:2212.06817, 2022.

[38] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic ma- nipulation. arXiv:2209.05451, 2022.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017.