2024 International Conference on Advanced Robotics and Mechatronics (ICARM) | 979-8-3503-8572-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICARM62033.2024.10715938

Effects of Prior Knowledge for Stair Climbing of Bipedal Robots Based on Reinforcement Learning

Mengya Su, Yasen Jia, and Yan Huang

***Abstract*— Bipedal robots with two legs are capable of traversing various terrains like level ground and staircases. Applying Reinforcement Learning (RL) based control can realize stable climbing stairs of bipedal robots. Prior knowledge of robots such as surrounding terrain information may improve the performance of climbing stairs. However, the impacts of prior knowledge on locomotion of bipedal robots across various terrains have not been systematically studied. In this work, we analyzed the effects of the amount of prior knowledge about terrain in front of the robot with RL-based control. Simulation results showed that introducing prior knowledge about terrain to bipedal robots can increase the maximum allowable ground height variation, realize smooth transition from level-ground walking to stair climbing, and improve disturbance rejection and energy efficiency. Prior knowledge is a trade-off between environmental information acquisition and computational complexity reduction. We find that there exists an optimal amount of prior knowledge for disturbance rejection and energy efficiency of stair climbing.**

I. Introduction

Bipedal robots with two legs are capable of traversing various terrains. This merit enables bipedal robots work alongside humans in unstructured environment such as uneven terrains. Staircase is a typical and common uneven terrain in human living environment. Therefore, many researches have developed motion control methods for biped robots to climb stairs and realize the transitions from level-ground walking to stair climbing. Bipedal robots equipped with camera or LiDAR sensors can perceive the information of surrounding terrain in front of itself and adjust gait before terrain variation. In this work, terrain information is represented as prior knowledge in robot motion control. Prior knowledge may help robots overcome the ground height variation and climbing stairs.

Model-based motion control methods with or without prior knowledge for stair climbing of bipedal robots have shown successful progress. Fevre et al. developed a controller based on velocity decomposition and applied it to a planar five-link biped robot without prior knowledge. The robot can climb stairs with step height as 5.2% leg length [1]. Caron et al. presented a walking stabilization control method by combining a wrench distribution quadratic program and a whole-body admittance controller [2]. The HRP-4 robot successfully climbed an industrial staircase with $18.5cm$ (24% leg length) step height by using this method. Gutmann et al. demonstrated the small QRIO robot with a stereo vision system can climb a staircase which has step height of $3cm$ (5% robot height) [3]. Okada et al. combined a precise 3D planar surface detection method and a practical 3D footstep planner method to develop a walking control system. The control system enabled the robot HRP2 to climb a $10cm$ height step (14.5% leg length) [4]. It can be seen that model-based control methods have successfully solved some difficult tasks by mathematical models. However, this kind of approach to bipedal locomotion on uneven terrain may require complex control design and may not fully exploit the robots capabilities [5, 6].

Model-free Reinforcement Learning (RL) approaches, have a significant impact on legged robot control. Without complicated modeling and laborious designing, RL methods can generate controller by automatic trial-and-error. Some RL-based methods without prior knowledge have been applied to bipedal robot locomotion on stairs. Siekmann et al. proposed periodic reward functions for RL based on foot forces and velocities, allowing the policy explore more action space[7]. The robot Cassie can ascend stairs blindly using this method [8]. Peng et al. successively proposed two imitation learning frameworks, which could generate stylized character motions with reference datasets [9, 10]. The results demonstrated that level-ground walking can be transferred to stair climbing. Yang et al. combined human bias with Mixture of Experts (MoE) learning architecture, enabling the robot Valkyrie to walk robustly over stairs with step height as $10cm$ (5.6% robot height) in simulation [11].

In addition to the RL-based control methods without prior knowledge mentioned before, RL-based control method with prior knowledge have also been developed and applied to motion control of bipedal robots. Duan et al. designed a vision-based RL framework that enabled the robot Cassie to climb stairs with the step height ranging from $0.05m$ to $0.2m$, and overcome a $0.5m$ (60% leg length) high step up [12]. Marum extended the belief encoder network from the quadruped robot to the bipedal robot, generating stable and robust biped gaits over stairs with noisy exteroception inputs [13]. Radosavovi et al. replaced normal neural network architectures such as Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN) with Transformer,

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

Mengya Su is with the School of Mechantronical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: mengys@bit.edu.cn).

Yasen Jia is with the School of Mechantronical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: jason_1120202397@163.com).

Yan Huang is with the School of Mechantronical Engineering, Beijing Institute of Technology, Beijing 100081, China; Key Laboratory of Biomimetic Robots and Systems, Ministry of Education, Beijing 100081, China; Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing 100081, China (corresponding author, e-mail: yanhuang@bit.edu.cn).

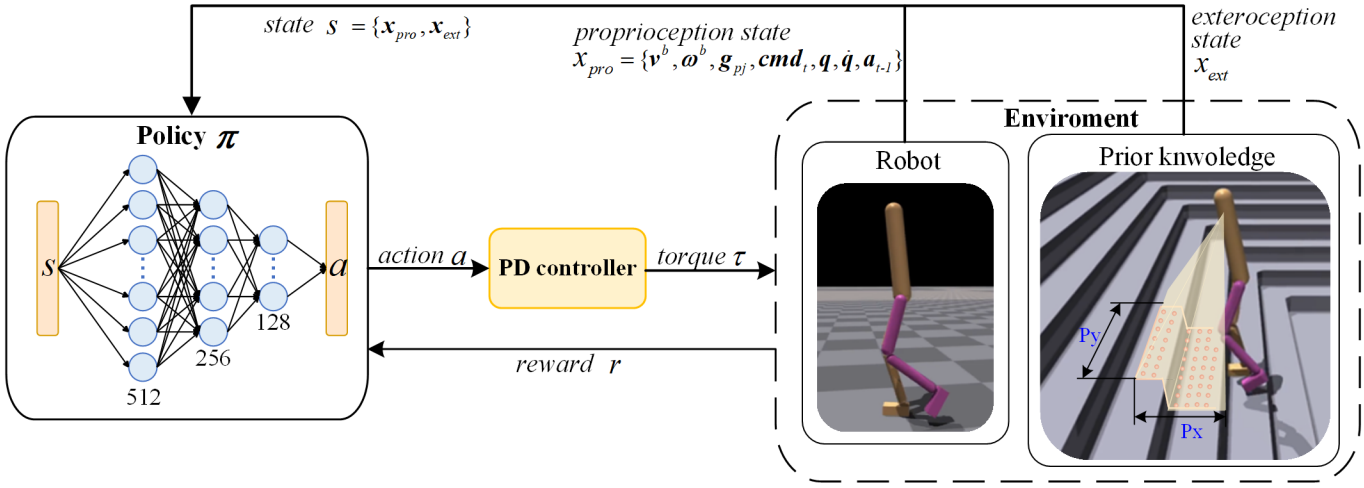

Fig. 1: The control architecture of the proposed method. Policy network receives state and reward from the environment. The prior knowledge consists of the vertical height difference between the base of the robot and the measurement points on ground (shown as red points). The perception distances along x-axis and along y-axis are represented as P_x and P_y, respectively. The policy outputs the action to the PD controller to command the robot.

realizing Digit humanoid robot whole-body and versatile locomotion in real world [14].

These current control methods based on RL with environment prior knowledge have good performance on the transition from walking to climbing stairs. However, to the best of our knowledge, there has been no systematical study on the effects of prior knowledge amount on climbing stairs of bipedal robots based on RL. In this work, we apply a RL-based control method with reward shaping and loss symmetry to a bipedal robot to achieve stair climbing and transition between terrains. The effects of prior knowledge amount on motion performance have also been studied.

The remainder of this paper is organized as follows: Section II describes the bipedal robot control method based on RL. Section III demonstrates simulation results, including the maximum one-step height, kinematics and dynamics analysis, and energy consumption. Section IV concludes the study.

II. Method

A. Overview of Motion Control

The control architecture is illustrated in Fig. 1 including the policy network, PD controller and environment. The bipedal robot model is a seven-link model with 16 Degree of Freedoms (DoFs), 6 for float base, 3 for each hip, 1 for each knee and 1 for each ankle. The coronal plane is in the x-axis direction, and the sagittal plane is in the y-axis direction.

The bipedal robot motion control problem on level ground and stairs in this work can be seen as a discrete-time stochastic process, Markov Decision Process (MDP) which can simulate the agent in environment with Markov property to realize stochastic policy and return. MDP consists of the state s_t, action a_t, reward r_t and policy π_θ. At each time step, the agent perceive the current state, and apply action to environment to change the next state and get a scalar reward. The accumulation of rewards over time is denoted as return $\bar{R}_\theta$. In this work, we use RL to solve the MDP problem, and the aim is to maximize the expected return by optimizing the network parameters θ^*. At each simulation step, the policy network receives the current state s_t and r_t from environment, and then outputs action a_t to actuate the robot. The state consists of the proprioception and exteroception (prior knowledge of environment) information and will be detailed in the following section.

B. State and Action Representation

The state is represented as $\boldsymbol{s}_t = \{\boldsymbol{v}^b, \boldsymbol{\omega}^b, \boldsymbol{g}_{pj}, \boldsymbol{cmd}_t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{ext}\}$, consisting of the base linear velocities $\boldsymbol{v}^b$, base angular velocities $\boldsymbol{\omega}^b$, projected gravity $\boldsymbol{g}_{pj}$, desired velocity commands $\boldsymbol{cmd}_t$, joint positions, joint velocities $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$, last actions $\boldsymbol{a}_{t-1}$ and the vertical height difference between the base of robot and measurement points $\boldsymbol{x}_{ext}$. The density of measurement points is $0.05m$ within the range of perception distance x (P_x) and perception distance y (P_y), as illustrated in Fig. 1. To evaluate the effect of prior knowledge on walking up stairs, we adjust the range of P_x from $0m$ to $1.25m$ while keeping P_y constant at $0.5m$. The action $\boldsymbol{a}_t$ is composed of the joint positions, and is sent to PD controller to output the joint torques to control the robot.

C. Network Architecture

The network follows an Actor-Critic architecture structured by MLP, in which the policy and value networks share parameters. The policy neural network π_θ, as shown in Fig. 1, maps the state s_t to a distribution over action p_θ modeled as a Gaussian distribution $\mathcal{N}(\mu(s), \sigma^2)$, where $\mu(s)$ represents the average value, and σ is the standard deviation. The state s undergoes processing through three hidden layers of size $[512, 256, 128]$ with ReLU activation functions, followed by a linear output layer that produces the action a. The network architecture for the value network is similar, but the output layer consists of only one linear cell.

D. Reward Formulation

We define rewards r_t to encourage the bipedal robot to successfully realize walking up stairs gaits and the transition from level-ground walking to stair climbing while penalizing undesirable behaviors such as falling down, unnatural movement. The formulation of the reward r_t is shown as (1).

$$r_t = r_{tlv} + r_{tla} + r_{tor} + r_{acc} + r_{limit} + r_{no\_fly} + r_a \quad (1)$$

The detailed items are shown as follows.

- Tracking linear or angular velocity reward r_{tlv}, r_{tla}: this term encourages the robot to track the x-axis (forward) and y-axis (left and right) linear and angular velocity commands:

$$\begin{aligned} r_{tlv} &= \exp\left(-\left\|\hat{v}_{xy} - v_{xy}\right\|^2 / \sigma_{xy}\right) \\ r_{tla} &= \exp\left(-\left\|\hat{\omega}_{yaw} - \omega_{yaw}\right\|^2 / \sigma_{xy}\right) \end{aligned} \quad (2)$$

  where $\hat{v}_{xy}$ and v_{xy} are the desired and real forward velocities of robot base, respectively. $\hat{\omega}_{yaw}$ and ω_{yaw} are the desired and real yaw velocity of the base, and $\sigma_{xy} = 0.25$.
- Torques penalty reward r_{tor}: this term penalizes joint torques, preventing the robot generating labor-intensive actions.

$$r_{tor} = -\left\|\sum_i \tau_i\right\|^2 \quad (3)$$

  where τ_i is the torque of i-th joint.
- Joint acceleration penalty reward r_{acc}: this term penalizes drastic change of joint accelerations which may lead to shakiness.

$$r_{acc} = -\sum_i \left\|\dot{q}_i(t) - \dot{q}_i(t-1)\right\|^2 \quad (4)$$

  where $\dot{q}_i(t)$ is the current i-th joint velocity, and $\dot{q}_i(t-1)$ is the i-th joint velocity at last simulation step.
- Joint position limits reward r_{limit}: this term serves as a soft limit on joint positions, manifested in the form of a reward function that penalizes excessive joint limitations.

$$r_{limit} = \begin{cases} q - q_{max}, & \text{if} \quad q > q_{max} \\ q_{min} - q, & \text{if} \quad q < q_{min} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

- Feet air time reward r_{air}: this term encourages the robot's foot to be in the air at the swing phase, contributing to the generation of a walking gait with foot clearance of swing leg.

$$r_{air} = -\sum_j^2 \left\|(t_{air,j} - 0.5)\right\|^2 \quad (6)$$

  where j is the foot index, and $t_{air,j}$ represents the time of j-th foot in air.
- No fly reward r_{no_fly}: this term prevents both feet from being in the air simultaneously.

$$r_{no\_fly} = \begin{cases} 1, & \text{if} \quad single\ contact \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

  where $single\ contact$ means at least one leg is in contact with ground.
- Action rate reward r_a: this term penalizes the output of the policy network with excessively large differences and thus encourage to generate smooth actions.

$$r_a = -\sum_i \left\|a_i(t) - a_i(t-1)\right\|^2 \quad (8)$$

  where $a_i(t)$ and $a_i(t-1)$ are the policy output of i-th joint at current and last simulation step, respectively.

E. Algorithm and Training Setup

In this work, we employ Actor-Critic networks trained with the Proximal Policy Optimization (PPO) algorithm [15]. At each time step, joint position commands are sampled from the policy controller to generate rollouts. The robot executes the corresponding action until the termination condition is triggered or the maximum episode number is reached.

Inspired by Heess et al. [16], we incorporate terrain and speed command curriculum into the training setup. During training stage, the terrain difficulty varies along the forward direction of the robot except flat plane. We enlarge the speed command sampling range once the reward of tracking linear velocity has reached 80% of its maximum. The initial range of the forward speed is $[-0.5, 0.5] m/s$, and the upper bound will plus 0.5 and the lower bound will subtract 0.5 on the basis of the initial range once the condition is reached.

The human gait pattern can produce symmetric biped locomotion temporally and spatially, and the walking gait is symmetric in left-right stride, swing time and joint forces [17]. To encourage policy to produce symmetric biped locomotion, we apply the symmetry loss [18] in our system. So the final optimization problem can be described as:

$$\pi_{\theta^*} = \underset{\theta}{\text{argmin}} \quad L_{PPO}(\theta) + wL_{sym}(\theta) \quad (9)$$

where π_{θ^*} is the policy, and θ^* is the parameter of the policy network. $L_{PPO}(\theta)$ is the loss computed by the PPO algorithm. $L_{sym}(\theta)$ is the symmetry loss computed by bilateral symmetry. Here, w represents the symmetry loss coefficient, representing the weight of the symmetry loss in the total loss.

III. Results

A. Experiment Setup

We verified the proposed method in the IsaacGym simulator [19], which can train agents in massive environments parallelly [20]. In the training stage, we chose 512 parallel environments and iterated $30,000$ times for a bipedal robot to learn walking on level-ground, climbing stairs and transition between the two gaits. Specifically, the width of stair is $0.5m$, and the height of the stair varies from $0.05m$ to $0.3m$ based on the curriculum learning. The frequency of the control policy is 50Hz. In this study, we changed the perception distance of the robot in the forward direction P_x (shown in Fig. 1) to adjust the amount of prior knowledge. In the simulation, we analyzed the effects of P_x on the maximum allowable ground height variation of a single step for the

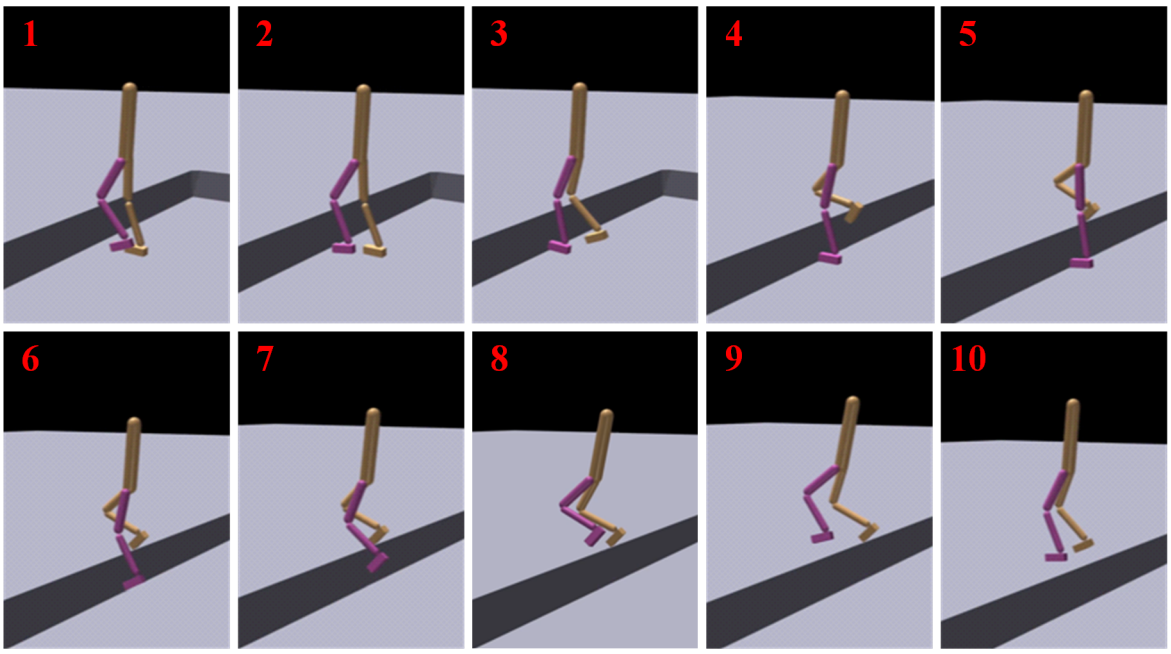

Fig. 2: The bipedal robot climbs a single stair. The perception distance along x-axis is $0.75m$ and along y-axis is $0.5m$, and the bipedal robot is climbing a single stair with step height as $0.42m$.

robot, the kinematics and dynamics of the bipedal robot transitioning from level-ground walking to stair climbing, and the energy efficiency of climbing stairs.

B. Climbing a Single Stair

In this sub-section, the robot climbed one single stair using the trained policy. We analyzed the maximum stair height the robot can overcome with perception distance P_x varying from $0m$ to $1.25m$. The results indicated that prior knowledge can significantly improve the maximum allowable stair height. The maximum stair height is $0.16m$ without prior knowledge, while exceeds $0.3m$ with prior knowledge. There existed an optimal P_x with a value of $0.75m$ (shown in Fig. 3). When P_x reaches its optimal value, the maximum allowable stair height is $0.42m$ (39% leg length). This trend can be explained as the robot may lack sufficient terrain information when the perception distance is too small, while the input dimensions and the computational load of the policy increase significantly when the perception distance is too large, making it difficult to learn gaits with good performance. This relation represented that selection of prior knowledge is a trade-off between environmental information acquisition and computational complexity reduction.

C. Kinematics and Dynamics Analysis

In this sub-section, we analyzed the correlation between the prior knowledge and the transition from level-ground walking to stair climbing. Joint angles and joint torques during transition are shown in Fig. 4. The curves of joint angles and torques are similar to those of human walking in [21], which indicates that the proposed policy can generate human-like gaits.

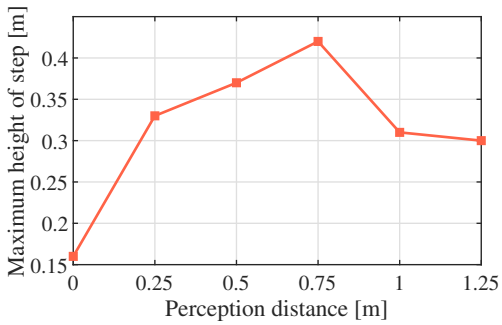

Fig. 3: The maximum stair height that the bipedal robot with varying perception distance can climb. The step width is $0.5m$.

From curves of joint angles, one can find that the joint angles of the robot with prior knowledge exhibit less variation and smoother transitions because the robot perceives terrain information and adjust gaits before terrain changes. The robot adjusts its gait earlier with a larger P_x. With the maximum P_x of $1.25m$, the robot exhibits a gait very close to stair climbing before walking up the stairs. Large perception distance can enable the robot to change its behaviors in advance of the terrain variation.

The joint torques of the robot with or without prior knowledge have similar trend and show minor difference. The amplitude of sagittal hip torques with prior knowledge is slightly smaller than without prior knowledge because joint angle curves with prior knowledge are smoother.

D. Energy Efficiency

Energy consumption is considered as a significant factor in bipedal locomotion. Cost Of Transport (COT) is a common and dimensionless measure of energy consumption. The

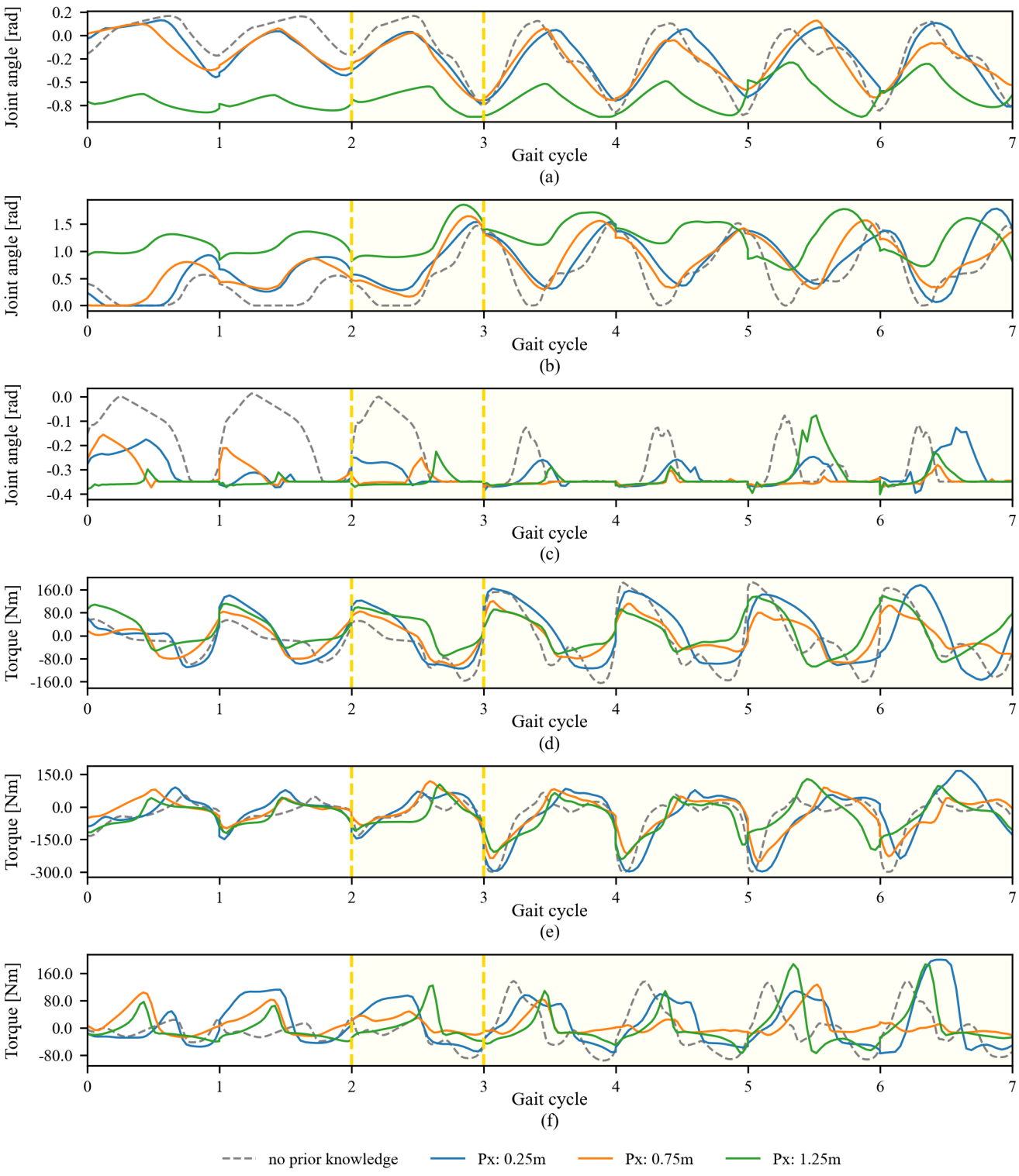

Fig. 4: The joint angles and torques of robot in transition from level-ground walking to stair climbing. (a)-(c): the hip joint angle, knee joint angle and ankle joint ankle in sagittal plane. (e)-(f): the hip joint torque, knee joint torque and ankle joint ankle in sagittal plane. A gait cycle is defined as the period from heel-strike of leg a to the next heel-strike of leg a (The leg firstly climbing on the stair is denoted as leg a, $a = left\ or\ right$). The first two gait cycles with white background are in level-ground walking phase and other gait cycles with yellow background are in stair climbing phase. Specifically, at the end of the second gait cycle, both legs are on the level ground; in the third gait cycle, the robot transitions from level ground to the stairs as leg a climbs on the first stair; at the end of the third gait cycle, leg a is on the first stair while the other leg is on the level ground.

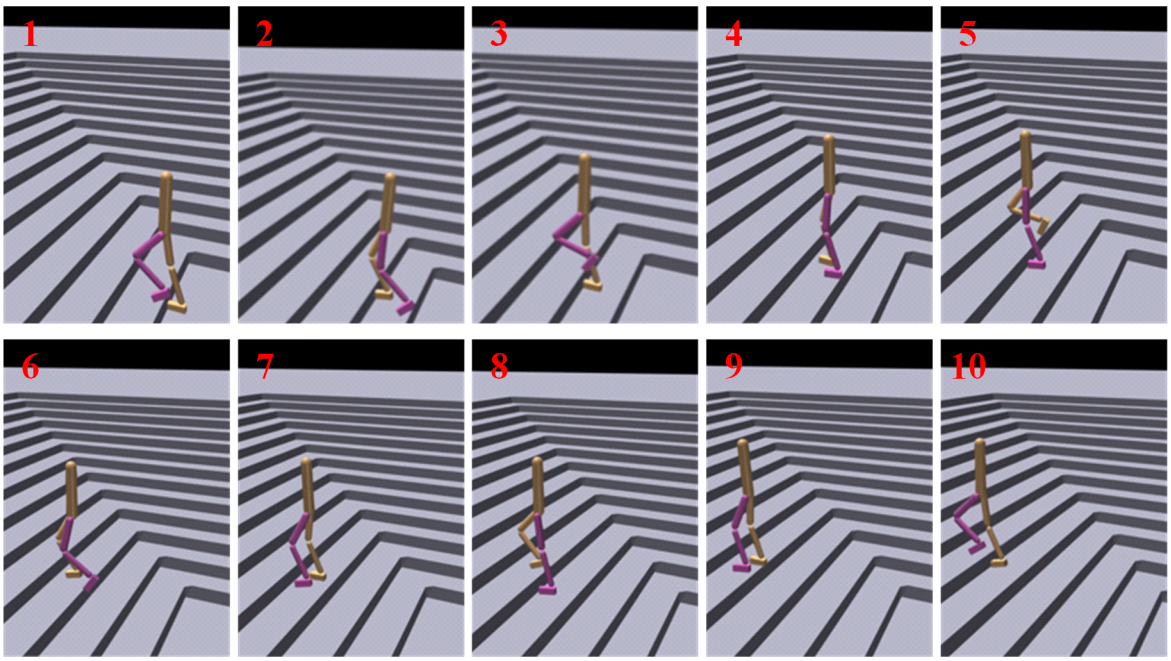

Fig. 5: The bipedal robot climbs stairs. The perception distance along x-axis is $0.75m$ and along y-axis is $0.5m$, and the bipedal robot is climbing stairs with step height as $0.2m$ and step width as $0.5m$.

formulation of COT is demonstrated as (10):

$$\text{COT} = \frac{\int_{t_0}^{t_1} \left(\sum_i^n T_i \dot{q}_i\right) dt}{mg\left(x_1 - x_0\right)} \tag{10}$$

where T_i is the ith joint torque, and $\dot{q}_i$ is the ith joint velocity. m is the mass of the biped robot. $(x_1 - x_0)$ is the distance traveled during the period from t_0 to t_1.

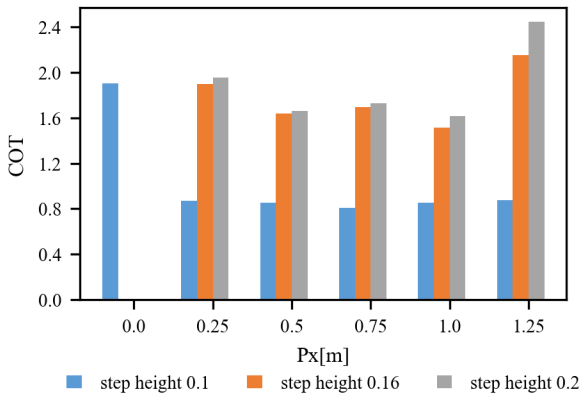

Fig. 6: COT of bipedal robot in climbing stairs with different perception distance P_x. The robot possessing prior knowledge failed to climb stairs with step height as $0.16m$ and $0.2m$.

In order to analyze the correlation between the perception distance and energy efficiency, we compute the COT of stair climbing with various P_x as shown in Fig. 6. The robot without prior knowledge failed to climb stairs with height as $0.16m$ and $0.2m$ due to the lack of disturbance rejection ability. The energy efficiency without prior knowledge of climbing stairs with step height as $0.1m$ is much lower than that with prior knowledge. For climbing stairs with prior knowledge, there exists an optimal value of P_x for energy efficiency. The optimal value lies within the range from $0.5m$ to $1.0m$.

From both perspectives of overcoming maximum ground height variation and energy efficiency, the optimal perception distance of prior knowledge is about $0.75m$, which means that a proper P_x can maximize the ability of disturbance rejection and minimize the energy consumption of stair climbing.

Previous study shows that human gaze remains within 2-4 steps ahead during stair ascent [22]. The step length of walking generated by our method is between $0.2m$ to $0.3m$. Thus, the optimal perception distance in this study approximates 2-4 step length. Consequently, the optimal amount of prior knowledge obtained by our analysis is close to the range of human gaze when approaching stairs, indicating that our method captures certain mechanism of human locomotion.

There is certain difference between the proposed method and human motion control mechanism. The amount of prior knowledge in humans, which means the distance can be perceived or seen ahead when a person walks or climbs the stairs, would not significantly affect the locomotive performance but requires more mental attention to be processed. In comparison, excessive perception information would impact the performance (e.g. energy efficiency) of robots for the reason that it may expend the exploration space during training process, making it difficult to find the optimal solution.

Nonetheless, the similarity between the results of our study and previous study can still be manifested. Within the optimal value of perception distance, maximum allowable stair height and energy efficiency are both improved as the amount of prior knowledge increases. For the perception distance

above the optimal value, despite the difference between the proposed method and human motion control, the effect of the amount of prior knowledge on locomotive performance is similar. To further elaborate, too much prior knowledge is unnecessary, causing heavier workload of perception and decision, which explains the existence of the optimal value to some extent.

IV. Conclusions

In this work, we employed an RL-based control method with reward shaping and loss symmetry to achieve stair climbing and transition between terrains of a bipedal robot, and systematically analyzed the effects of prior knowledge amount of the robot on stair climbing and walking-stair climbing transition. Simulation results showed that robot with prior knowledge can climb on a single step with higher height, generate smoother transition from level-ground walking to stair climbing. Moreover, analysis of COT indicated that the robot with prior knowledge had better disturbance rejection ability and energy efficiency than that without prior knowledge. The selection of prior knowledge is a trade-off between environmental information acquisition and computational complexity reduction. There exists an optimal amount of prior knowledge for disturbance rejection and energy efficiency of stair climbing. Moreover, the optimal amount of prior knowledge obtained by our analysis is close to the range of human gaze when approaching stairs.

References

[1] M. Fevre, B. Goodwine, and J. P. Schmiedeler, "Terrain-blind walking of planar underactuated bipeds via velocity decomposition-enhanced control," *Int J Rob Res*, vol. 38, no. 10-11, pp. 1307–1323, 2019.

[2] S. Caron, A. Kheddar, and O. Tempier, "Stair climbing stabilization of the hrp-4 humanoid robot using whole-body admittance control," in *Proc IEEE Int Conf Rob Autom*, 2019, pp. 277–283.

[3] J.-S. Gutmann, M. Fukuchi, and M. Fujita, "Stair climbing for humanoid robots using stereo vision," in *IEEE Int Conf Intell Rob Syst*, vol. 2, 2004, pp. 1407–1413.

[4] K. Okada, T. Ogura, A. Haneda, and M. Inaba, "Autonomous 3d walking system for a humanoid robot based on visual step recognition and 3d foot step planner," in *Proc IEEE Int Conf Rob Autom*, 2005, pp. 623–628.

[5] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Sci Robot*, vol. 5, no. 47, p. eabc5986, 2020.

[6] Z. Xie, G. Berseth, P. Clary, J. Hurst, and M. van de Panne, "Feedback control for cassie with deep reinforcement learning," in *IEEE Int Conf Intell Rob Syst*, 2018, pp. 1241–1246.

[7] J. Siekmann, Y. Godse, A. Fern, and J. Hurst, "Sim-to-real learning of all common bipedal gaits via periodic reward composition," in *Proc IEEE Int Conf Rob Autom*, 2021, pp. 7309–7315.

[8] J. Siekmann, K. Green, J. Warila, A. Fern, and J. Hurst, "Blind bipedal stair traversal via sim-to-real reinforcement learning," in *Robot. Sci. Syst.*, 2021.

[9] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Trans Graph*, vol. 37, no. 4, pp. 1–14, 2018.

[10] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Trans Graph*, vol. 40, no. 4, pp. 1–20, 2021.

[11] C. Yang, K. Yuan, S. Heng, T. Komura, and Z. Li, "Learning natural locomotion behaviors for humanoid robots using human bias," *IEEE Robot Autom Lett*, vol. 5, no. 2, pp. 2610–2617, 2020.

[12] H. Duan, B. Pandit, M. S. Gadde, B. J. van Marum, J. Dao, C. Kim, and A. Fern, "Learning vision-based bipedal locomotion for challenging terrain," *2023, arXiv:2309.14594*.

[13] B. v. Marum, "Learning perceptive bipedal locomotion over irregular terrain," Ph.D. dissertation, 2023.

[14] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Learning humanoid locomotion with transformers," *2023, arXiv:2303.03381*.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *2017, arXiv:1707.06347*.

[16] N. Heess, D. Tb, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami *et al.*, "Emergence of locomotion behaviours in rich environments," *2017, arXiv:1707.02286*.

[17] I. Handi and K. B. Reed, "Perception of gait patterns that deviate from normal and symmetric biped locomotion," *Front Psychol*, vol. 6, 2015.

[18] W. Yu, G. Turk, and C. K. Liu, "Learning symmetric and low-energy locomotion," *ACM Trans Graph*, vol. 37, no. 4, pp. 1–12, 2018.

[19] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance GPU based physics simulation for robot learning," in *2021, arXiv:2108.10470*.

[20] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proc. Mach. Learn. Res.*, 2022, pp. 91–100.

[21] D. Neumann, *Kinesiology of the Musculoskeletal System*. Elsevier Health Sciences, 2016.

[22] V. Miyasike-daSilva, F. Allard, and W. E. McIlroy, "Where do we look when we walk on stairs? gaze behaviour on stairs, transitions, and handrails," *Exp Brain Res*, vol. 209, pp. 73–83, 2011.