Reinforcement Learning for Quadrupedal Robot Control: Taxonomy, Sim-to-Real, Robustness, and Emerging Trends

Chen Chang

{Zcemcc1@ucl.ac.uk}

Department of Mechanical Engineering, University College London, WC1E 6BT, UK

Abstract. Quadrupedal robots exhibit strong mobility in complex environments where wheeled platforms often perform poorly, but their control remains difficult. In recent years, reinforcement learning (RL) has received growing attention in quadrupedal locomotion, as it supports direct policy optimization without relying entirely on hand-crafted control rules. This paper presents a control-oriented review of RL-based quadrupedal robot control. It first introduces the main physical and algorithmic foundations. Then it examines recent research progress from three complementary aspects: control architecture, observation and information design, and training task formulation. In addition, it analyses several persistent obstacles to practical deployment, namely sim-to-real transfer, reward and objective design, real-time robustness with online adaptation, and perception–actuation integration. The results suggest that future progress will depend not only on stronger learning algorithms, but also on more transparent evaluation, more hardware-aware system design, and more reliable integration of perception, dynamics, and control.

Keywords: Reinforcement learning, Quadrupedal robots, Sim-to-real transfer

1 Introduction

Quadrupedal robots are increasingly used in areas where rubble, bushes and industrial places the wheeled platform is unable to work. Their stability on discontinuous footholds and compliant surfaces allows them to allocate support across numerous contacts and avoids them becoming stuck when faced with terrain uncertainty. The processes of quadrupedal locomotion are extremely nonlinear and also closely coupled with intermittent contacts. The contact occurrences present non-smooth transitions, frictional uncertainty and fast transitions in the feasibility of ground reaction forces. The performance is also multi-objective with trade-offs between the stability, energy efficiency, speed and the hardware safety at actuator limits and delays. Besides, the variability of sensing and environment will establish an ongoing sim-to-real gap. Powerful perception is important in rapid and energy-thrifty locomotion. Lastly, the nature of complex behaviours necessitates uninterrupted transitions between gaits, implying that an inhibition of failure may prove decisive in the development of locomotion strategies [1].

ICMEEA 2026, April 24-26, Singapore, Singapore
Copyright © 2026 EAI
DOI 10.4108/eai.24-4-2026.2364995

Indeed, a significant issue is how one would design and assess a locomotion controller of the real-world quadruped once the legged systems are no longer in quasi-static walking.
Historically, quadrupedal robots have usually been controlled using model-based control pipelines, which typically are designed based on trajectory optimization, model predictive control (MPC) and whole-body control (WBC) [2]. These techniques are mostly interpretable and enable constraints to be written down explicitly, although their usefulness is heavily reliant on the correctness of formulating the model and heavy human configuration. Practically, contact behaviour, structural compliance, and actuator dynamics are hard to model with high fidelity, a handwritten state machine may get more and more complex as the robot is assumed to survive in diverse terrains and cope with behavioural requirements. Another possible framework to implement reinforcement learning (RL) is optimization of closed-loop policies via interaction. This allows controllers to take advantage of rich feedback and acquire nonlinear strategies that are challenging to acquire analytically. Recent advancements in quadruped control with RL have been made in multiple directions, such as multi-skill navigation in hierarchical learning, perception-aware locomotion in improved robustness in unstructured environments [3], and versatile gait adaptation based on animal transition strategies, which can easily switch and recover between terrains without any external sensing [4]. Simultaneously, the advances also demonstrated the unresolved questions, especially on the design of the rewards, safety assurance and evaluation criterion.
It is against this backdrop that the current review looks at RL-based quadrupedal locomotion through the lens of a control-oriented perspective in order to synthesize and critically review the rapidly growing body of literature. Despite its significant potential to produce agile motion, adaptive gait transitions, and robust negotiation of terrains in four-legged robots, significant challenges persist, such as sim-to-real transfer, reward design, evaluation inconsistency, and robust operation in complex settings. In this respect, the theoretical basis, representative control methods, recent research advancement and key unresolved issues concerning RL-based quadrupedal locomotion are discussed in this paper, to explain how the sub-field has evolved, and what may be done to create more robust and deployable quadruped control systems. In the introduction, the paper provides the foundations of the dynamic of quadruped and the RL involved in locomotion, topicalities of the research on the topic in both articles are reviewed, and limitations and up-to-date challenges of the field are discussed before drawing conclusions on the key findings.

2 Preliminaries: quadruped dynamics and RL formulation

2.1 Floating-base dynamics and contact modelling

Quadruped locomotion is commonly modelled as a floating-base, underactuated system in which joint torques act directly on the legs, while trunk motion is generated indirectly through intermittent foot–ground interaction. As illustrated in Fig. 1, this viewpoint provides a common physical basis for analysis, simulation, and learning-based locomotion. A widely used rigid-body formulation is given by [5]:

$$M(q)\ddot{q} + h(q,\dot{q}) = S^{\top}\tau + J_c^{\top}(q)f_c \tag{1}$$

where $q = [q_b^{\top}, q_j^{\top}]^{\top}$stacks the floating-base coordinates q_b and joint coordinates q_j, $M(q)$ is the generalized inertia matrix, $h(q,\dot{q})$ collects gravity, coriolis, and centrifugal terms, τ denotes

the actuated joint torques, S is the actuation selection matrix, and $J_c^{\mathrm{T}}(q)f_c$ represents the contribution of foot–ground contact forces.

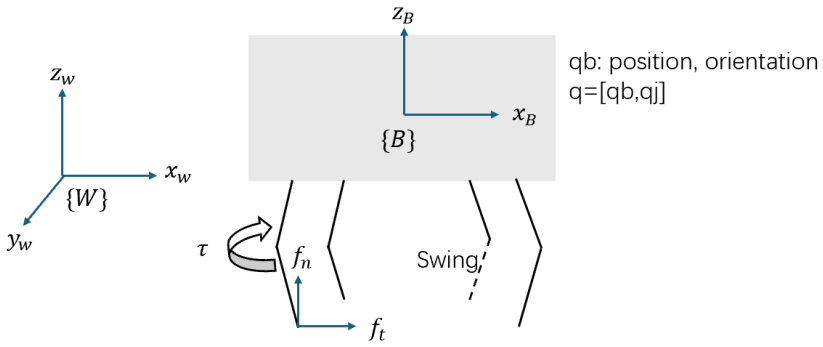

Fig. 1. The basic principles of quadruped robot dynamics. (Picture credit: Original)

This implementation, commonly known as contact-consistent dynamics, is common in locomotion simulators and training systems [2]. Since terrain-dependent friction, compliance, and unmodelled foot deformation make contact-rich models difficult to model accurately, optimization and large-scale training of full-order contact-rich models may become costly.
Reduced-order abstractions are popular in many studies, trading off dominant locomotion physics at reduced computational cost. The centroidal, momentum, and single-rigid-body models are used to describe the acceleration of the robot base as a result of contact forces, whilst the template models, including the spring-loaded inverted pendulum and the linear inverted pendulum, model the important gait mechanics with far lower state dimension [6]. The presence of these abstractions is not that they offer an alternative to full dynamics, but rather that it makes possible rollouts with shorter time, and controller design, reward development, and assessment that have a simpler structure.

2.2 RL formulation for quadruped locomotion

The kinematics of quadrupedal animals provides a universal interface between robots and learning algorithms. In animal kinematics, RL is typically formulated as a sequential decision-making problem, where the policy maps observed results to control actions to maximize long-term motion performance. Fig. 2 shows that at each time step, the policy receives the current observation result, outputs an action, and the environment simultaneously returns a reward and the next observation result. A standard goal is to maximize the expected discounted return [7]

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{2}$$

where $\gamma \in (0,1)$ is the discount factor. This objective is widely used in locomotion learning because it naturally balances immediate stabilization with longer horizon task performance [8]. In practice, however, the effectiveness of this formulation depends less on the generic RL definition itself than on how observations, actions, and rewards are designed for locomotion tasks.

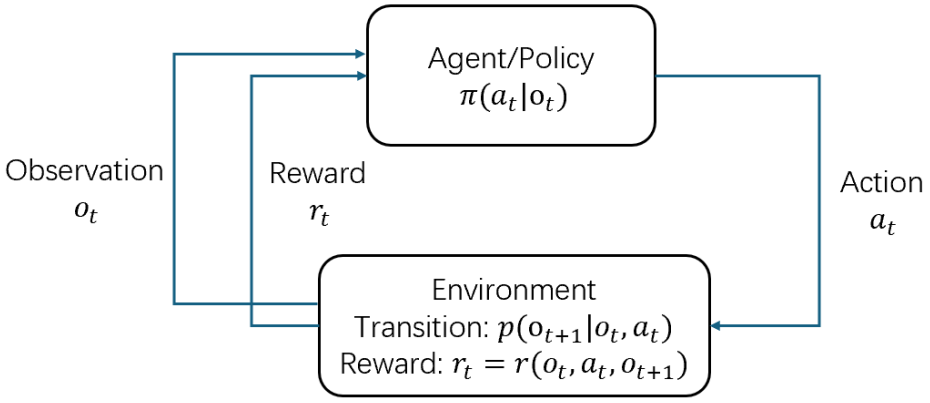

Fig. 2. Schematic illustration of the RL formulation for quadruped locomotion. (Picture credit: Original)

2.2.1 Observation design

In quadruped locomotion, policy inputs are typically built from robot proprioception together with task-level command variables. Common observation components include base orientation and angular velocity from the inertial measurement unit, base linear velocity estimates, joint positions and velocities, foot contact indicators, and commanded motion targets such as desired forward speed or yaw rate [7]. These are popular variables since they can give a succinct description of body posture, motion state and task purpose without the need to measure the full state. In the case where the terrain knowledge is needed, an exterior-sensory depth image or local height map may be incorporated into the policy input [3]. In reality, the perceptive locomotion system is prone to attaching a brief history of observation or to temporally smoothed inputs to minimise the influence of sensing delay, partial observability and momentary perception errors. Consequently, observation design does not simply depend on sensor choice, but also dictates the strength with which the policy would be able to deduce changes in terrain, contacts, and trends of motion using the imperfect measurements. In reality, there are numerous policies applying short observation histories, or time-integrated inputs to enhance resilience to sensing delay, hidden contact state, terrain uncertainty, as well as temporal perception errors.

2.2.2 Action interface

The action interface defines the form that the learned policy transmits commands to the robot and this decision has a high impact on the behaviour and trainability of RL-based locomotion controllers. The various forms of action representations applied to date include joint torques, desired joint positions, desired joint velocities, foot-end references and higher-level gait-related parameters [7]. These interfaces, owing to their varying measures of control granularity, also vary in expressiveness, stability, and deployment. A typical option is to allow the policy to generate desired joint positions or joint position increments and utilise a proportional-derivative controller to follow [8]. The common use of this is that the low-level controller contributes to stabilizing feedback, which in most cases, simplifies policy learning and increases the robustness of hardware experiments. Greater flexibility of control can be provided by more direct torque commands, which are generally harder to teach, as well as more susceptible to

modelling errors. In comparison, action spaces of a higher level can be easier to interpret and can be more transfer-appropriate, although they limit the policy to a smaller control interface. As a result, the action interface should be regarded as a central modelling choice, since it directly affects the trade-off among agility, learning stability, and sim-to-real reliability.

2.2.3 Reward construction

Reward construction specifies which locomotion behaviours are encouraged during training and which are discouraged. Rather than relying on a single performance measure, most quadruped locomotion studies use a weighted combination of terms that jointly encode task achievement, motion quality, and safety [7]. A common structure includes command-tracking terms, posture or stability regularization, energy-related penalties, and smoothness terms that discourage abrupt action changes or large impact events.

Typical tracking rewards encourage the robot to match commanded forward speed, lateral speed, or yaw rate. Stability-related terms often penalize excessive body roll or pitch, large angular velocities, or deviations from a nominal base height. Efficiency terms may penalize large torques, joint power, or unnecessary motion, while smoothness-related shaping can reduce action rate oscillations, impact severity, or foot slip. Many implementations also terminate episodes with a large negative reward when the robot falls or enters unsafe configurations. Because reward design strongly affects the learned gait and can bias comparisons across methods, locomotion papers should report reward components and weights clearly enough to support reproducibility and fair evaluation [6].

3 Research progress of RL-based quadruped control

RL has been widely applied in quadrupedal locomotion, and different application settings place distinct demands on control architecture, observation design, and training objectives.

3.1 Control architecture

Quadruped locomotion methods can first be distinguished by where reinforcement learning is placed within the overall control pipeline. As summarized in Fig. 3, the main architectures can be grouped into three categories, end-to-end policies, hierarchical designs, and hybrid or residual pipelines.

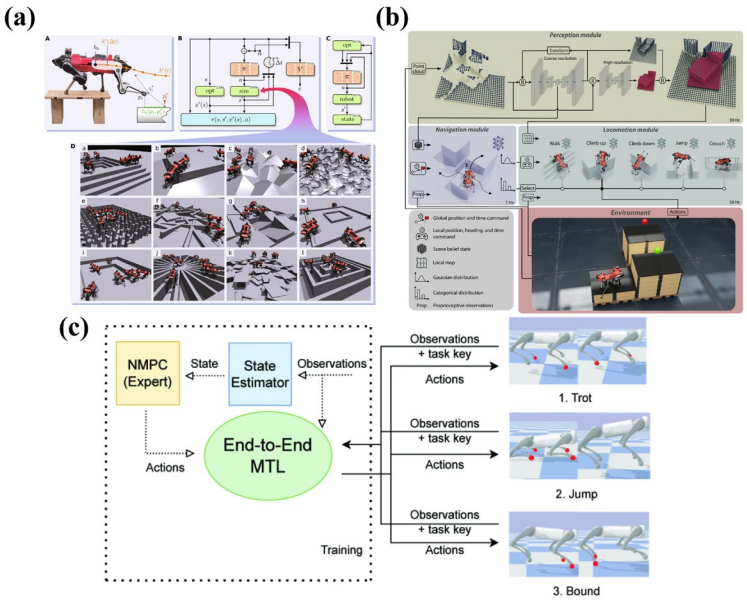

Fig. 3. (a) Hierarchical designs decompose control into higher-level decision modules and lower-level locomotion execution [9]. (b) Hybrid architectures combine learned policies with structured model-based or optimization-based components for improved robustness and precision [10]. (c) End-to-end learned policies directly produce actions from onboard observations and task inputs [11].

3.1.1 End-to-end locomotion policies

In an end-to-end locomotion architecture, a single policy maps onboard observation directly to low-level motor commands, without relying on a hand-designed finite state machine. End-to-end policies can achieve high agility when trained at scale in simulation and can be extended to perception-driven locomotion, where obstacle interaction skills emerge from learned perception to action mappings [12]. Despite these advantages, end-to-end policies are often difficult to interpret and debug. In practice, end-to-end pipelines are most suitable for algorithmic comparisons and for high-agility tasks supported by large-scale simulation training, especially when perception is central to the problem setup.

3.1.2 Hierarchical reinforcement learning: skills, options, and gait schedulers

Hierarchical reinforcement learning (HRL) decomposes quadruped control into a high-level decision module and a low-level motion execution module, which is particularly useful when tasks require gait transitions, skill composition, or long-horizon decision making beyond steady tracking. During execution, the gait selection and transitions are treated as a high-level policy, while rapid motion correction and contact stabilization are handled by the lower-level controller [4]. Compared with a single monolithic policy, this decomposition improves modularity and can reduce the burden of covering widely different regimes within one network. HRL is typically favoured in settings where smooth gait switching and structured behaviour composition are central requirements, such as in terrain courses and parkour-style sequences that demand multiple distinct skills.

3.1.3 Hybrid and residual RL: learning to track or correct structured references

Hybrid control is a widely used direction in deployable quadruped locomotion. In a typical hybrid pipeline, a structured module generates references such as footholds, while reinforcement learning focuses on learning a robust tracking controller around these references. By constraining the policy to track physically plausible references, the search space is reduced and reliability often improves. Meanwhile, residual RL places learning as an additive correction on top of a nominal controller, which helps preserve baseline feasibility while allowing the policy to compensate for unmodeled dynamics and modelling mismatch. Hybrid pipelines often train and transfer more stable than monolithic policies and can be particularly effective when contact feasibility dominates performance [8]. The main drawback is increased system complexity and a dependence on the quality of the reference generator. For these reasons, hybrid and residual RL are most suitable for real-world deployment and for contact critical tasks such as sparse footholds and precision manoeuvres.

3.2 Observation and information design

A second important dimension in quadruped locomotion is the design of observations and information available to the policy. The existing approaches can be broadly grouped into proprioception-only control, perception-fusion control, and privileged-information-based training.

3.2.1 Proprioception-only policies

Proprioception-only policies restrict the observation space to internal sensing, typically including joint positions and velocities from encoders, inertial measurements, an estimated base state, and sometimes simple contact indicators. This setting is widely used because it simplifies the sensing pipeline and therefore tends to reduce sim-to-real risk, since the policy is less exposed to lighting changes, texture variation, and other sources of perception noise. Despite the lack of external sensing, several studies indicate that substantial adaptability is still possible. Bio-inspired gait strategy frameworks show that multi-gait behaviour and state-dependent adjustment can be achieved through structured gait scheduling driven purely by internal signals [4]. Complementary work on emergent gaits without explicit motion references also suggests that, with appropriate rewards and constraints, reinforcement learning can discover diverse gait patterns using proprioceptive inputs alone [13]. However, transfer robustness remains limited on terrains where successful locomotion depends on perceiving discrete footholds or planning foot placement in advance.

3.2.2 Perception

Perception enables anticipation, stepping over gaps, choosing footholds, and planning manoeuvres over obstacles. “In-the-wild” perceptive locomotion highlights how fusing exteroception with proprioception can produce robust behaviour under real terrain variability [3]. Perception is also critical for parkour-style tasks where obstacles are discrete and the required timing and placement precision is high [9]. The trade-off is a larger sim-to-real gap because performance can degrade under sensor noise and latency, as well as lighting and domain shift and imperfect terrain reconstruction.

3.2.3 Privileged learning

A common practical approach is privileged learning, in which simulator-only signals such as terrain height, friction, and contact state are available during training but removed at deployment. A student policy is then distilled to operate with real onboard sensors only. This design improves sample efficiency and helps stabilize learning under partial observability, and it has become a recurring pattern in successful sim-to-real pipelines [7].

3.3 Training tasks

Locomotion methods can also be differentiated by the control objective they prioritize. Beyond architectural and training differences, quadruped RL systems are often evaluated according to the task family they target, including steady command tracking, transitions between behaviours, and recovery under severe disturbances.

3.3.1 Steady locomotion and command tracking

The most frequent family activity is to remain under commanded references of steady locomotion, usually based on linear velocity and the yaw rate, and occasionally of base height or attitude. This environment is commonly adopted as a standard due to its ability to isolate fundamental locomotion skills like balance, foot placement, and energy expenditure, as well as placing the commands and the terrain environment under control and similar across experiments. Simulation at scale indicates that, under conditions of training with a variety of perturbations, and in varied randomized environments, the policies can be trained in an agile and stable manner [3]. In perception-based variants, tracking is further generalized to terrain-aware locomotion through conditioning the policy on the exteroceptive inputs by conditioning the policy such that future contacts and obstacles are predicted as opposed to being reactively responded to [12]. Common assessment is based on monitoring accuracy, base stabilization, step timeliness consistency, power usage, and disturbance evasion.

3.3.2 Gait transition and skill switching

In addition to constant tracking, applications are also focused on gait shifting and switching skills, in which the objective is to transition between behaviours which include walk, trot and pace, or to combine various skills under a common command interface. This family also stresses smoothness and stability in mode changes, as contact time or command target discontinuity can cause loss of balance. Learning based experiments indicate that gait patterns and switching can be developed based on the task goals and limits without a set of motions, which makes conditioning and optimization tasks have a role in determining the resultant locomotion repertoire [13]. Switching is interpreted in other works as a means of pursuing the goals of viability and failure avoidance, which makes not falling a likely dominating force behind transitions under unfavourable circumstances [1]. In environments that have significant obstacles, switching of skill is closely linked to perceptions and extended horizon decision making, especially with respect to parkour-type sequences where individual obstacles necessitate varied interaction patterns [12]. Characteristics of a successful transition usually reported by evaluation include smoothness of transition, delay or switching time, success percentage during the commanded gait and skill sequence, and evidence of the transition period stability.

3.3.3 Recovery and posture regulation

Recovery-focused tasks emphasize robustness under rare but high-consequence events, including push and slip recovery, disturbance rejection during brief contact loss, and self-righting after failure. These tasks test whether a policy has learned stabilizing strategies beyond nominal tracking and often reveal long-tailed failure modes that are not reflected in average case benchmarks. Robustness-oriented studies therefore advocate structured stress tests rather than relying only on random pushes, since targeted perturbations can expose brittle behaviours and overfitting that would otherwise go unnoticed [14]. Recovery objectives are also frequently paired with training approaches that explicitly bias learning toward robustness, such as adversarial disturbances or risk aware objectives, although the defining feature of this task family remains the control goal of rapidly stabilizing and returning to a feasible posture and gait after disruption. Typical metrics include recovery success rate, time to recovery, falls per distance or time under disturbances, and bounds on posture deviation during the stabilization process.

4 Key challenges and technical solutions

RL-based quadruped locomotion has progressed rapidly in simulation, yet deployment on physical robots remains constrained by several recurring technical bottlenecks, including sim-to-real transfer, reward and objective design, real-time robustness with online adaptation, and perception–actuation integration.

4.1 Sim-to-real transfer

Sim-to-real transfer remains one of the central challenges in RL-based quadruped locomotion because policy performance is highly sensitive to mismatch between simulated and physical interaction dynamics. In legged systems, small errors in contact timing, actuator response, state estimation, friction conditions, or terrain properties can accumulate quickly and lead to unstable footholds, degraded tracking, or even complete failure during deployment. To reduce this gap, existing studies typically combine several strategies, including domain randomization, improved system identification, actuator and latency modelling, and noise injection during training. These methods have significantly improved the transferability of learned policies, but they do not fully remove the underlying difficulty that simulation can only approximate real contact-rich dynamics. As a result, the main issue is no longer only whether a policy can be transferred once, but whether it can remain reliable when sensing quality, surface conditions, hardware wear, or external disturbances vary after deployment. Future progress in this area is therefore likely to depend on transfer strategies that balance model fidelity, training diversity, and hardware-tolerant policy design, rather than relying on broader randomization alone.

4.2 Reward shaping and objective design

In quadrupedal locomotion, the task is inherently multi-objective. More specifically, policies must often balance command tracking, stability, energy efficiency, smoothness, foot clearance, and robustness to disturbances at the same time. As a result, poorly designed objectives can lead to reward hacking, overly conservative behaviours, unstable actuation, or policies that perform well only under narrow training conditions. One of the most popular methods is a focus on

command tracking as the primary optimization objective and on the use of unsafe events as explicit termination conditions, which are likely to enhance stability and simplify the interpretation and reproducibility of experimental results [8]. It is also common to make a gradual change in the speed, terrain, or level of perturbation by curriculum training. In more complicated motions, higher-order objectives, such as an imitation-based or reference-guided one, may additionally lighten the load of manual shaping by providing structured motion priors, and then fine-tuning RL to achieve robustness. The reward engineering can, however, be dense, thus enhancing the efficiency of the training but causing confusion as to which of the objective components is the real learnt behaviour. One of the remaining challenges is that the use of reward engineering can enhance the solution of training efficiency and it is even more difficult to determine which of the objective components are actually behind the outcome of behaviour. It is against this reason that making reward terms, termination rules, and curriculum settings clear continues to be relevant to help in reproducibility and fair comparison across studies.

4.3 Real-time robustness and online adaptation

When walking on four legs, the control depends upon fast and repeated contacts and loss can be experienced as well due to a steady decrease in the performance average, as well as infrequent destabilizing incidents. Simultaneously, adaptation should be done with high real-time and onboard computing requirements that restrict the level of complexity of the methods that can be safely implemented on equipment [14]. To this effect, current methods usually involve inclusive robustness-heightening preparation and lightweight online adjustment and neither of the two proposals. One such resolution is to enhance the diversity of training by randomization, disrupting conditions, and a variety of terrain, such that the policies can extrapolate outside a small operating regime. Nevertheless, even though these adaptation techniques may be able to enhance average robustness, they are usually susceptible to sudden failure or truly out-of-distribution instances. In any future work on robust quadruped control, it is expected that tighter coordination of training time robustness with deployment time adaptive behaviour, more standardized stress tests and more failure reporting and evaluation protocol will be more helpful not only in providing meaningful comparisons that can be done between the methods.

4.4 Perception–actuation integration

Exteroceptive perception can be of significant benefit to quadruped locomotion through foothold selection, negotiating obstacles and planning over longer horizons, yet presents significant control and implementation dilemmas. The main challenge is that perceptual data is generally slower, noisier and more vulnerable to domain shift, in comparison to proprioceptive feedback. Due to this, current techniques tend to have perception integrated in such a manner as to maintain the inner control loop in a steady state. Among these designs is to decouple the time scales, such that a fast proprioceptive policy takes care of lower-level stabilization whilst perception provides slower-time scale cues, subgoals, or gait guidance. But enhanced perceptual consciousness does not always result in increased strength of control, especially when there is an uncertainty interaction with quick contact events. It can be anticipated that perception will assume a greater role in quadruped locomotion, yet the architectures of dependable integration will still require architectures that expressly model latency, uncertainty and real-time control constraints [3].

5 Conclusion

In conclusion, this paper presents a control-oriented review of RL-based quadrupedal robot control. It first introduces the main physical and algorithmic foundations. Then it examines recent research progress from three complementary aspects: control architecture, observation and information design, and training task formulation. In addition, it analyses several persistent obstacles to practical deployment, namely sim-to-real transfer, reward and objective design, real-time robustness with online adaptation, and perception–actuation integration. The results suggest that future progress will depend not only on stronger learning algorithms, but also on more transparent evaluation, more hardware-aware system design, and more reliable integration of perception, dynamics, and control.

References

[1] Shafiee, M., Bellegarda, G., Ijspeert, A.: Viability leads to the emergence of gait transitions in learning agile quadrupedal locomotion on challenging terrains. Nat Commun. pp. 3073 (2024).

[2] Di Carlo, J., Wensing, P.M., Katz, B., Bledt, G., Kim, S.: Dynamic locomotion in the MIT Cheetah 3 through convex model-predictive control. Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp. 1-9 (2018).

[3] Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., Hutter, M.: Learning robust perceptive locomotion for quadrupedal robots in the wild. Sci Robot. pp. eabk2822 (2022).

[4] Humphreys, J., Zhou, C.: Learning to adapt through bio-inspired gait strategies for versatile quadruped locomotion. Nat Mach Intell. pp. 1141-1153 (2025).

[5] Cun, C., Yang, Q., Li, Z., Zhou, M.C., Pang, J.: Model predictive optimization and control of quadruped whole-body locomotion. IEEE/CAA J Autom Sin. pp. 2103-2114 (2025).

[6] Sun, B., Mohamed Haris, S., Ramli, R.: A systematic review of deep reinforcement learning for legged robot locomotion. Instruments. pp. 8 (2026).

[7] Ha, S., Lee, J., van de Panne, M., Xie, Z., Yu, W., Khadiv, M.: Learning-based legged locomotion: State of the art and future perspectives. Int J Robot Res. pp. 1396-1427 (2025).

[8] Peng, X.B., van de Panne, M.: 'Learning locomotion skills using DeepRL: Does the choice of action space matter. Proc. ACM SIGGRAPH/Eurographics Symp. Computer Animation, (2017).

[9] Sajja, A., Khorshidi, S., Houben, S., Bennewitz, M.: End-to-end multi-task policy learning from NMPC for quadruped locomotion'. Proc. European Conf. Mobile Robots (ECMR), pp. 1-6 (2025).

[10] Hoeller, D., Rudin, N., Sako, D., Hutter, M.: ANYmal parkour: Learning agile navigation for quadrupedal robots. Sci Robot. pp. eadi7566 (2024).

[11] Jenelten, F., He, J., Farshidian, F., Hutter, M.: DTC: Deep Tracking Control. Sci Robot. pp. eadh5401 (2024).

[12] Cheng, X., Shi, K., Agarwal, A., & Pathak, D. : Extreme Parkour with Legged Robots. arXiv.Org. (2023).

[13] Cooke, L.L., Fisher, C.: Deep reinforcement learning of quadrupedal gaits without motion references: Emergence of pace and other locomotion patterns. Adv Robot. pp. 1447-1462 (2025).

[14] Zhao, H., Song, W., Wang, D., Tong, X., Ding, P., Cheng, X., et al.: MoRE: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp. 11212-11218 (2025).