ITM Web of Conferences **78**, 01021 (2025) https://doi.org/10.1051/itmconf/20257801021
CSEIT 2025

Deep Reinforcement Learning in Continuous Control: Advances and Challenges

Buze Ke[1], *Tianqi Qiu*[2], *Yikai Shen*[3*]

[1]School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
[2]Multiple media and web development, Fuzhou University, Fuzhou, China
[3]International College, Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. With the widespread application of Deep Reinforcement Learning (DRL) in continuous control, its learning and decision-making capabilities in high-dimensional state spaces have brought new opportunities for intelligent control systems. This paper provides a systematic review of the research progress of DRL in continuous control tasks, covering representative algorithms from classic Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3) to the soft policy method Soft Actor-Critic (SAC), and further focuses on key breakthroughs in algorithm structure and sample efficiency in recent years. In the latest research, one class of methods effectively improves the accuracy and stability of policy generation by aligning the behavior of diffusion models with the Q function; another class of methods employs Euclidean data augmentation techniques to enhance data diversity and generalization performance during the training process; additional studies introduce successor features and concurrent policy combination mechanisms, significantly improving transfer efficiency and adaptability in multi-task learning. These innovative algorithms show great potential in alleviating training instability, enhancing sample utilization, and improving policy generalization capabilities. Looking ahead, DRL research in continuous control can further explore efficient exploration mechanisms, model prediction integration, multi-agent collaboration, and real-world deployment, laying the foundation for building reliable and adaptive intelligent control systems.

1 Introduction

In recent years, the rapid advancement of artificial intelligence technology has led to significant interest in DRL, particularly for its efficacy in addressing complex tasks within the domain of Continuous Control. Continuous control problems are applied in various areas of daily life, including robotic manipulation, autonomous driving, robotic arm operation, and drone navigation. In this situation, the main challenge is developing an efficient and stable learning strategy which enables agents to perform complex tasks with high-dimensional continuous action spaces. In contrast to conventional reinforcement learning, DRL combines deep neural networks with policy optimization methods, enabling

* 2022214972@stu.cqupt.edu.cn

the automatic extraction of high-dimensional features from unprocessed state inputs and the direct formulation of effective control strategies. This integration markedly improves the learning capabilities and generalization performance of complex systems. Consequently, a systematic review of recent advancements in this field is essential for both academic inquiry and practical engineering.

In existing research, early work such as DDPG first introduced the actor-critic architecture to continuous control problems, successfully achieving end-to-end policy learning. Subsequently, TD3 alleviated the issue of policy overestimation through mechanisms like delayed updates and target smoothing, further enhancing training stability. SAC introduced an entropy term to encourage policy exploration, balancing learning efficiency and policy robustness. These methods form the technical foundation of mainstream algorithms in current continuous control tasks.

Building on this, recent research has further advanced the diversification and deep optimization of methods. On one hand, some studies have proposed using diffusion models to construct behavior policies, and then improving sample efficiency and policy generation quality through alignment with the Q function, addressing the instability issues of traditional policy updates. On the other hand, geometric structure enhancement methods have also gained attention, where researchers have introduced data transformation techniques in Euclidean space to achieve broader coverage of state distributions and improve model generalization capabilities. Additionally, in the field of multi-task reinforcement learning, concurrent combination methods based on successor features have emerged, modeling the composability of policies to enable transfer learning and rapid adaptation across tasks. These works not only theoretically expand the modeling paradigm of continuous control but also achieve excellent performance in practical evaluations.

Nevertheless, there are still several urgent issues to be addressed in current DRL methods for continuous control, such as poor training stability, low sample efficiency, weak task generalization ability, and difficulties in deployment in real systems. Therefore, systematically organizing the research context in this direction, summarizing its mainstream methods and representative achievements, and evaluating the contributions and shortcomings of existing work in both theory and practice will provide valuable references for future research.

This article provides a review of the current research status and development trends of DRL in continuous control. It first introduces the mainstream methods adopted in this field in recent years, including representative algorithms based on deterministic policies, soft policies, and model-assisted strategies, and analyzes their advantages and limitations in terms of performance, sample efficiency, and applicable scenarios. Next, the article summarizes commonly used experimental platforms and datasets in current research, such as MuJoCo, Roboschool, and DeepMind Control Suite, and introduces common evaluation metrics such as cumulative rewards, learning curves, and policy convergence time to help standardize performance comparison. Subsequently, the article summarizes the policy levels and actual performance achieved by existing methods in various standard tasks, showcasing the key progress made by DRL in continuous control. On this basis, it further discusses the key challenges still faced today, such as low sample efficiency, insufficient generalization ability, weak policy robustness, and deployment limitations, and analyzes the proposed solutions in related research, including the introduction of model predictive mechanisms, enhancing the flexibility of policy structures, and strengthening task transfer and multi-task adaptability. Through a systematic organization and structured analysis of the above content, this article aims to present readers with a clear technical context and development trends in this field, helping subsequent researchers quickly grasp research directions and providing theoretical support and methodological insights for further optimization and application of DRL in continuous control.

2 Overview of Mainstream Approaches in Recent Years

2.1 Model-Free DRL for Continuous Control

Table 1 Comparison of Technical Advantages and Disadvantages of Different Models

Aspect	**Model-Free DRL**	**PopSAN Hybrid Framework**	**Implicit World Model (TD-MPC2)**
Core Principle	Model-free DRL (PPO/SAC) with dense reward shaping and policy fine-tuning	Integration of DRL (DDPG/TD3/SAC/PPO) with population-coded SNNs on neuromorphic hardware	Implicit world modeling with joint reward/value prediction, task embedding for cross-task transfer, and latent MPC planning
Technical Advantages	1. High generalization (geometry/configurations) 2. 100% zero-shot transfer success 3. Sample-efficient fine-tuning	1. 140× energy efficiency 2. Real-time performance 3. Noise robustness (event-driven computation)	1. Unified model for 104 tasks 2. Strong cross-task transfer 3. Heterogeneous action space support
Technical Disadvantages	1. Requires high-fidelity simulators 2. Prolonged training (million-scale steps) 3. Complex reward engineering	1. SNN training complexity (gradient approximation) 2. Precision constraints from spike rates 3. Neuromorphic hardware dependency	1. Limited visual input resolution 2. Discrete action incompatibility 3. Manual task embedding initialization

Asad Ali Shahid et al. introduced a model-free DRL framework [1], utilizing Proximal Policy Optimization (PPO) and SAC algorithms. This approach trains continuous control policies in simulated environments through meticulously designed dense reward functions, achieving a 100% zero-shot transfer success rate on real-world Franka Emika Panda robots. A key innovation lies in its policy fine-tuning mechanisms, which enable rapid adaptation to novel tasks while preserving generalization across diverse object geometries and initial configurations. These features make the method highly suitable for industrial robotic grasping applications.

2.2 Hybrid Learning Framework with Spiking Neural Networks

Guangzhi Tang et al. proposed a hybrid learning architecture that integrates population-coded Spiking Neural Networks (PopSAN) with DRL optimization [2]. Implemented on neuromorphic hardware (Intel Loihi), this framework demonstrates 140× higher energy efficiency compared to conventional platforms like the Jetson TX2. The population coding mechanism enhances input-output representation capabilities, enabling effective handling of high-dimensional continuous control tasks while retaining the event-driven computational advantages inherent to spiking neural networks.

2.3 Implicit World Model-Based Predictive Control

Nicklas Hansen et al. developed the TD-MPC2 algorithm [3], a model predictive control method that leverages implicit world models for large-scale multi-task continuous control. This unified framework supports training across 104 diverse tasks spanning domains such as DMControl and ManiSkill2 by implicitly modeling task commonalities. Technical innovations include task embedding for semantic similarity learning, latent space planning via MPC/MPPI optimizers, and parameter-efficient cross-task knowledge transfer. The framework also maintains compatibility with heterogeneous action spaces, broadening its applicability. Table 1 presents a comparison of the technical advantages and disadvantages of different models.

3 Common datasets and evaluation criteria

3.1 Common Datasets

In the field of reinforcement learning for robots, multiple datasets and simulation environments are widely used for algorithm validation and performance evaluation [4]. With its high-fidelity physics simulation capabilities, the MuJoCo physics engine has become the platform of choice for complex robot dynamics models (e.g., 7-degree-of-freedom robotic arm control) and dynamic balancing tasks (bipedal walking, quadrupedal running), supporting custom environmental parameters (friction coefficient, target position) to test the robustness of the algorithm. OpenAI Gym integrates classic control tasks (such as CartPole and Pendulum) and complex robot tasks (such as HalfCheetah) through standardized interfaces, which is convenient for quickly comparing the performance of different algorithms (PPO, SAC) in continuous action space control and sparse reward scenarios [5].

For multi-task learning scenarios, DMControl and Meta-World cover diverse tasks from humanoid robot movement to fine operations (drawer opening, stacking building blocks) [6], while ManiSkill2 and MyoSuite focus on fine object manipulation and muscle-driven control, respectively, providing a unified benchmark for transfer learning and generalization ability evaluation of world models (such as TD-MPC). IsaacGym relies on IsaacGym for large-scale parallel training, and its GPU-accelerated capabilities based on the NVIDIA PhysX engine support collaborative training of thousands of robot instances (such as dynamic obstacle avoidance and multi-robot handling), significantly improving the efficiency of distributed training.

In the field of offline reinforcement learning, the D4RL dataset integrates expert strategy and random strategy data (such as the Franka robotic arm kitchen task), and verifies the robustness of CQL and other algorithms on out-of-distribution data through standardized evaluation protocols. In addition, DeepMind Control Suite combines low-dimensional state and high-dimensional pixel observation (such as Cheetah Run) to support domain randomization technology, which is often used to verify the effect of data augmentation methods on the generalization ability of algorithms.

3.2 Evaluation Criteria

The algorithm performance evaluation should comprehensively consider the quality, efficiency and robustness of the task completion. The task success rate (e.g., the success rate of Franka manipulator grasping) and the cumulative reward value (e.g., the normalized reward in D4RL) are the core indicators, the former directly reflects the effectiveness of task execution, and the latter measures the ability of the algorithm to optimize the goal through the sum of rewards, especially in the sparse reward scenario (e.g., maze navigation), which can compare the exploration efficiency of different algorithms. In order to make fair comparisons across tasks, normalized scores (e.g., TD-MPC2 vs. DreamerV3) normalize the rewards to a uniform interval to facilitate quantification of the multi-task transfer learning effect.

Sample efficiency focuses on the amount of data required by an algorithm to achieve a specific performance, such as the compression effect of data augmentation technology on the number of training steps in DeepMind Control Suite, or the data utilization of offline reinforcement learning on the D4RL dataset. The robustness of the strategy evaluates the stability of the algorithm through perturbation tests (e.g., friction coefficient changes, observed noise), such as the success rate of muscle-driven robots in MyoSuite under parameter perturbations, which is crucial for simulation-to-real migration.

Computational efficiency becomes a key metric in real-time control and large-scale training scenarios, including training throughput in the IsaacGym environment (e.g., simulated steps per second) and inference latency on neuromorphic processors (e.g., real-time decision-making speed in PopSAN). Together, these standards constitute a comprehensive evaluation system for algorithms from the laboratory to the actual implementation.

4 Discussion and analysis of the problems and possible solutions in this direction

4.1 Existing Problems

4.1.1 Low sample efficiency

Most mainstream DRL methods in current continuous control are model-free algorithms, such as DDPG, TD3, and SAC, which rely on a large number of interactions with the environment to learn policies. Although this approach is flexible, the training process typically requires millions of interaction samples. In practical applications such as robotic control and autonomous driving, the data collection process is often characterized by a significant demand for time and resources, alongside potential issues related to equipment deterioration and safety risks [7]. These considerations substantially constrain the applicability of the algorithm under discussion. Consequently, enhancing sample efficiency and minimizing reliance on real-world environments has emerged as a critical challenge that necessitates attention within this domain.

4.1.2 Limitations of Computing Resources and Energy Consumption

To adapt the limitations of resource-constrained edge devices and embedded systems, it is necessary to develop more efficient DRL algorithms. For example, using lightweight neural network architectures, such as Depthwise Separable Convolutions and Model Pruning techniques, can efficiently reduce energy consumption and computational complexity. Moreover, hardware acceleration (likes specialized AI chips) can be used to stimulate the training and inference processes, thereby decreasing reliance on GPUs and improving the Deployability of algorithms on low-power devices. The progression of green and low-carbon computing has rendered the development of DRL methodologies that function with minimal energy consumption a significant goal.

4.1.3. Enhancing Multi-tasking and Generalization Ability

To address the limitations of existing methods that cannot transfer when facing new tasks or changes in the environment, transfer learning and meta-learning can be employed to improve multi-task processing capabilities. These methods allow for knowledge sharing across different tasks, thereby avoiding the need to train from scratch. For instance, meta-learning algorithms can learn how to quickly adapt to new tasks and transfer existing experiences to new tasks, significantly improving the efficiency of multi-task learning. Additionally, task embedding and shared representation learning can share underlying feature representations across multiple tasks, enhancing the model's generalization ability to adapt to different tasks and environmental changes.

4.1.4. Strategy Robustness and Safety

In environments characterized by dynamism or significant uncertainty, reinforcement learning methodologies frequently demonstrate limited robustness [8]. This vulnerability renders them prone to failure in response to minor alterations in state distribution, potentially resulting in unstable or hazardous behaviors.This issue is particularly severe in scenarios involving real physical devices or human-machine interactions. Furthermore, existing algorithms generally lack effective response mechanisms for extreme states or sudden events, resulting in the inability of the strategy to make safe decisions in the face of abnormal situations. Therefore, enhancing the stability and safety assurance capabilities of strategies is a necessary prerequisite for the practical application of DRL in critical real-world control systems.

4.2 Possible Solutions

4.2.1 Improving Sample Efficiency

In order to accommodate the limitations of resource-constrained edge devices and embedded systems, it is imperative to develop more efficient DRL algorithms. For instance, the implementation of lightweight neural network architectures, including Depthwise Separable Convolutions and Model Pruning techniques, can substantially decrease both computational complexity and energy consumption. Furthermore, the utilization of hardware acceleration, such as specialized artificial intelligence chips, can enhance the optimization of training and inference processes, thereby diminishing dependence on GPUs and facilitating the deployment of algorithms on low-power devices. As the field of green low-carbon computing progresses, the design of DRL methodologies that function with minimal energy consumption has emerged as a critical objective.

4.2.2 Limitations of Computing Resources and Energy Consumption

To adapt to resource-constrained edge devices and embedded systems, it is necessary to design more efficient DRL algorithms. For example, using lightweight neural network architectures, such as Depthwise Separable Convolutions and Model Pruning techniques, can significantly reduce computational complexity and energy consumption. Additionally, hardware acceleration (such as specialized AI chips) can be used to optimize the training and inference processes, thereby reducing reliance on GPUs and improving the deployability of algorithms on low-power devices. With the advancement of green and low-carbon computing, designing DRL methods that can operate under low energy consumption has also become a key objective.

4.2.3 Enhancing Multi-tasking and Generalization Ability

To address the limitations of existing methods that cannot transfer when facing new tasks or changes in the environment, transfer learning and meta-learning can be employed to improve multi-task processing capabilities. These methods allow for knowledge sharing across different tasks, thereby avoiding the need to train from scratch. For instance, meta-learning algorithms can learn how to quickly adapt to new tasks and transfer existing experiences to new tasks, significantly improving the efficiency of multi-task learning. Additionally, task embedding and shared representation learning can share underlying feature representations across multiple tasks, enhancing the model's generalization ability to adapt to different tasks and environmental changes.

4.2.4 Enhancing Strategy Robustness and Safety

To improve the robustness and safety of reinforcement learning strategies in dynamically changing and uncertain environments, safety components can be embedded in the reward function to ensure that the agent considers safety constraints while pursuing long-term rewards, thereby avoiding dangerous behaviors. Additionally, by employing uncertainty modeling methods, the agent's ability to respond to unexpected situations can be improved by estimating its behavior in uncertain environments. Furthermore, utilizing a safe reinforcement learning framework can introduce safety control mechanisms in the strategy learning process, enhancing the strategy's responsiveness to extreme states or abnormal situations.

4.3 Future Research Directions

4.3.1 Development of General-Purpose Agents

To achieve end-to-end learning from high-level instructions to low-level control, future research can explore the integration of multimodal models encompassing language, vision, and action [9]. By enabling cross-modal knowledge fusion, agents can comprehend complex instructions and execute tasks in real-world environments. For instance, agents may perceive the environment through visual inputs, interpret task requirements via natural language instructions, and perform actions through motor control. This paradigm enhances agents' autonomy and adaptability in dynamic and complex scenarios, significantly advancing the development and practical deployment of general-purpose intelligent agents.

4.3.2 Unsupervised Pretraining

Leveraging large-scale unlabeled data for pretraining facilitates the construction of world models that can learn the dynamics of environments without reliance on extensive manual annotations. This approach not only reduces the dependence on labeled datasets but also exploits the richness of environmental interactions to improve generalization. Through unsupervised learning, agents can autonomously discover latent patterns from interactions with their surroundings, enabling informed decision-making and control. This significantly lowers the cost of human intervention and promotes more scalable and efficient learning systems.

4.3.3 Real-Time Adaptive Control

Developing mechanisms for online and adaptive policy updates allows agents to dynamically adjust their control strategies in response to real-time data and environmental changes. This is particularly critical in highly dynamic and unpredictable domains such as autonomous driving and robotic control. Continuous strategy refinement ensures more accurate and efficient task execution. Real-time adaptive control enhances the flexibility and responsiveness of decision-making policies, enabling agents to better operate under real-world uncertainties and complexities [10].

4.3.4 Co-Optimization of Hardware and Algorithms

Jointly advancing hardware and algorithmic design is key to deploying continuous control systems in real-world scenarios. For example, leveraging dedicated hardware accelerators such as neuromorphic chips or FPGAs can support the efficient execution of reinforcement learning and deep learning algorithms on resource-constrained devices. Concurrently, algorithmic optimizations can be tailored to maximize hardware efficiency and performance. This co-optimization framework reduces computational overhead, accelerates inference speed, and supports scalable and high-performance intelligent systems, thereby promoting the widespread adoption of intelligent agents in practical applications.

5 Conclusion

The application of deep reinforcement learning (DRL) in the field of continuous control has made significant progress in recent years, providing intelligent solutions for complex tasks such as robot control, autonomous driving, and drone navigation. This paper systematically reviews the research status of DRL in continuous control, and comprehensively analyzes the mainstream methods, experimental platforms, evaluation criteria, existing problems and future directions, aiming to provide a clear picture of the technical context and development trend for academia and industry.

At the method level, DRL realizes end-to-end control from high-dimensional state space to continuous action space by combining deep neural network and policy optimization technology. Early algorithms such as DDPG, TD3 and SAC have laid the basic framework for model-independent DRL, and improved the stability and exploration efficiency of the strategy through deterministic policy optimization, deferred update and entropy regularization. In recent years, researchers have proposed further innovative improvements: the behavior strategy alignment technique based on the diffusion model improves the quality of strategy generation by optimizing the Q function; The data transformation method based on geometric structure enhancement expands the coverage of state distribution and enhances the generalization ability of the model. The implicit world model (e.g., TD-MPC2) and multi-task transfer framework (e.g., policy combination based on successor features) significantly improve the adaptability and scalability of the algorithm in multi-task scenarios through cross-task knowledge sharing and latent space planning.

In terms of experimental verification, simulation platforms such as MuJoCo, DMControl, and ManiSkill2 provide a standardized test environment for continuous control research, and quantify the algorithm performance through core indicators such as task success rate, cumulative reward, and sample efficiency. The results show that the current mainstream methods have approached or surpassed the level of human experts in complex control tasks (such as humanoid robot motion and multi-joint robotic arm operation), but at the same time, they also expose bottlenecks such as low sample efficiency, high computing resource consumption, and insufficient strategy robustness. For example, model-independent DRLs typically require millions of interactive samples to be trained, while hybrid frameworks based on neuromorphic computing can improve energy efficiency but are limited by hardware dependency. The generalization ability of multi-task strategy in dynamic environment still needs to be improved.

In order to solve the above challenges, the researchers propose a multi-dimensional optimization path: at the algorithm level, by introducing model predictive control (MPC), unsupervised pre-training and meta-learning mechanisms, the dependence on real interaction data can be reduced and cross-task migration can be accelerated; At the engineering deployment level, lightweight network architecture and hardware co-design (such as neuromorphic chips) can reduce computing overhead and promote the deployment of edge devices. In terms of security, the reward function design and uncertainty modeling techniques that integrate security constraints can enhance the fault tolerance of the strategy in a dynamic environment. In addition, the construction of multimodal general-purpose agents that integrate language, vision and movement, and the exploration of real-time adaptive control mechanisms will become the focus of future research.

Looking forward to the future, the development of DRL in continuous control needs to take into account the balance between algorithm innovation and implementation practice. On the one hand, it is necessary to break through the theoretical bottleneck and develop more efficient exploration mechanisms, interpretable policy representations, and robust world models. On the other hand, it is necessary to solve the domain adaptation problem from simulation to real objects to promote the reliable deployment of algorithms in real physical systems. With the deepening of multidisciplinary integration (such as neuroscience heuristic computing and low-carbon hardware design), DRL is expected to give rise to safer, adaptive, and sustainable intelligent control systems in the fields of industrial automation and smart cities, and finally realize the leapfrog development from laboratory to industrial application.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Asad, A. S., Dario, P., Francesco, B.: ‘Continuous control actions learning and adaptation for robotic manipulation through reinforcement learning.’ Autonomous Robots (2022), 46: 483–498
2. Tang, G. Z., Neelesh, K., Raymond, Y., Konstantinos, P. M.: ‘Deep Reinforcement Learning with Population-Coded Spiking Neural Network for Continuous Control.’ Rutgers University. Proceedings of Machine Learning Research. Proceedings of the 2020 Conference on Robot Learning. 2020
3. Nicklas, H., Su, H., Wang, X. L.: ‘TD-MPC2: Scalable, Robust World Models for Continuous Control.’ ICLR 2024. 2024
4. Liu, Y. T., and Aamir, A.: ‘Multi-Task Reinforcement Learning in Continuous Control with Successor Feature-Based Concurrent Composition.’ IEEE 25-28 June 2024
5. Chen, H. Y., Zheng, K. W., Su, H., Zhu. J.: ‘Aligning Diffusion Behaviors with Q-functions for Efficient Continuous Control.’ NeurIPS. 2024
6. Luo, J. Z., Chen, J. Y., Zhang. Q.: ‘Reinforcement Learning with Euclidean Data Augmentation for State-Based Continuous Control.’ Machine Learning (cs.LG); Artificial Intelligence (cs.AI) 16 Oct 2024

7. Fang, Z. Y.: ‘Research on Train Operation Control Strategy Based on Deep Reinforcement Learning’, Dalian Jiaotong University. 2023
8. Zhang, W. G.: ‘Research on indoor mobile robot navigation based on deep reinforcement learning’, Southeast University. 2021
9. Hou, X. H.: ‘Research on Obstacle Avoidance of Field Robots Based on Deep Reinforcement Learning.’ Northeastern University, 2020
10. Pang, L. Z.: ‘Research on Key Technologies of Follow up Robots Based on Deep Learning’ Northeastern University. 2019