A Survey on Reinforcement Learning-Based Multi-Agent Path Planning

Jiawei Hu

Artificial Intelligence, Dongguan City University, Dongguan, China

Abstract. With the advancement of artificial intelligence (AI) and Internet of Things (IoT) technologies, industrial systems have witnessed a growing demand for intelligent management. Multi-agent systems (MAS), being applied across various domains, not only significantly enhance the autonomy and efficiency of intelligent systems but also demonstrate irreplaceable potential in hazardous and complex environments where mobile robots may even prove indispensable. Deep reinforcement learning (DRL) has emerged as one of the most effective solutions for multi-agent path finding (MAPF) problems, having successfully achieved global path planning in known environments. However, its performance remains unstable in map-free scenarios, a persistent challenge that has become a focal point in recent MAPF research. This paper synthesizes contemporary studies to provide an overview of current research progress in MAPF. We systematically categorize existing algorithms into centralized and decentralized frameworks, tracing their evolutionary trajectories. Through established evaluation metrics, the paper analyze and compares mainstream algorithms while summarizing their developmental processes. Furthermore, this work examines critical challenges in DRL-based MAPF implementations and proposes potential directions for future research, which may focus on enhancing environmental adaptability, optimizing multi-agent coordination mechanisms, and improving real-time decision-making capabilities in dynamic scenarios. The discussion aims to provide theoretical references for advancing intelligent path planning in complex industrial applications.

1 Introduction

With the advancement of artificial intelligence and Internet of Things technologies,industrial systems have demonstrated increasing demands for intelligent management. The application of multi-agent reinforcement learning technology has been widely applied in promoting the development of unmanned aerial vehicle (UAV) delivery systems,intelligent transportation systems,and related industries, effectively enhancing industrial efficiency and competitiveness while enabling mobile robots to exhibit progressively higher levels of intelligence [1]. Multi-Agent Path Finding (MAPF) refers to the process of planning collision-free paths for multiple agents with distinct starting positions to reach their respective target locations by utilizing acquired environmental

hujiawei202235010408@dgcu.edu.cn

information,with the objective of identifying optimal paths within the environment [2]. A critical challenge in multi-agent systems lies in balancing the trade-off between individual agent path optimization and computational efficiency in group path planning. However, within MAPF scenarios, algorithmic approaches prioritize the generation of computationally tractable and conflict-free trajectories over pursuing strictly optimal path solutions [3]. Multi-agent reinforcement learning focuses on training agent teams to collaboratively accomplish mission objectives that prove challenging for individual agents,where resolving path planning issues remains a key research focus in the field of cooperative multi-agent reinforcement learning.

First, MAPF algorithms can be broadly categorized into two classes: search-based traditional algorithms and learning-based intelligent algorithms.These two categories of algorithms address different environmental requirements [4]. Classical path planning algorithms demonstrate rapid and high-quality performance in static or small-scale environments due to their observable and computationally feasible characteristics. However, as the number of agents and environmental complexity increase, replanning becomes highly time-consuming [4]. Researchers continuously enhance their performance in uncertain environments through optimization or integration with other algorithms. Nevertheless, in many practical application scenarios, such as unknown disaster rescue sites or complex indoor environments, obtaining accurate map information is often challenging or even impossible [5]. In mapless environments, effective cooperation among agents becomes crucial for successful path planning.Therefore, investigating path planning algorithms under mapless conditions holds significant practical implications.

With the evolution of neural networks, distributable deep reinforcement learning methods have been extensively employed to address Multi-Agent Path Finding (MAPF) problems [6].Researchers model MAPF as a discrete, partially observable Markov decision process, enabling agents to acquire sensor-derived observational data as states based on their individual decision-making models. Through continuous interaction with the environment, agents select optimal actions to reach their targets [7]. This learning-based approach requires no prior knowledge and can be effectively scaled to accommodate arbitrary numbers of agents, demonstrating favorable scalability [8].

With advancements in deep learning and reinforcement learning, methods based on reinforcement learning and imitation learning have been increasingly employed to address Multi-Agent Path Finding (MAPF) problems [9]. However, learning-based algorithms face challenges in achieving inter-agent collaboration due to their reliance on localized information for decision-making, inevitably leading to deadlock situations [10]. Nevertheless, deep reinforcement learning remains a highly promising approach for solving MAPF challenges. Compared to traditional centralized planning methods, learning-based approaches aim to develop a universal decentralized planning strategy, enabling agents to make decisions based on their locally observable information. This distributed paradigm of "centralized training with decentralized execution" facilitates better scalability to larger-scale tasks without increasing computational complexity [11]. Learning-based algorithms also incorporate hybrid models to adapt to diverse application scenarios [9]. Communication learning has recently gained significant traction in multi-agent systems, with ongoing explorations into integrating communication mechanisms or combining them with reinforcement learning methods for MAPF problems [12]. Current communication approaches, however,may introduce excessive redundant information that could interfere with agents'decision-making processes in reinforcement learning frameworks. Among these learning-based solutions, Deep Reinforcement Learning (DRL) has emerged as one of the most actively researched domains for developing safe and robust MAPF solutions. DRL-based methods have demonstrated encouraging progress across various fields, attributed to their strong decentralized execution capabilities, high adaptability to novel

environments and uncertainties, and suitability for partially observable environments [13]. This paper synthesizes recent relevant research to present a comprehensive review of current developments in Multi-Agent Path Finding (MAPF). It systematically outlines mainstream algorithmic designs and their evolutionary trajectories in MAPF research, accompanied by a critical summary of current mainstream algorithms through the lens of evaluation metrics. Finally, the paper discusses potential future directions for deep reinforcement learning-based multi-agent path planning algorithms, proposing insights into their development trajectory.

2 Related Technologies and Developments

2.1 Multi-Agent Path Finding (MAPF) problem

The Multi-Agent Path Finding (MAPF) problem entails planning and coordinating collision-free paths for a group of agents to navigate from their respective initial positions to designated target locations [2]. This research domain demonstrates significant application potential across diverse fields including outdoor hazardous environments, intelligent warehousing systems, urban transportation networks, gaming applications, and aerospace operations [4]. A critical constraint requires that agents must be capable of simultaneously executing their planned trajectories without mutual interference.

The resolution of MAPF problems typically necessitates the integration of sophisticated path planning algorithms with effective conflict resolution strategies to satisfy multi-objective optimization requirements. Substantial progress in single-agent path finding (SAPF) research has established a robust theoretical foundation for MAPF technologies [14]. Conceptually, MAPF extends the SAPF framework to address the coordination challenges of multiple autonomous agents operating within shared environments. From a theoretical perspective, SAPF can be regarded as a special case within the broader MAPF paradigm [15]. Contemporary MAPF algorithms are primarily categorized into two architectural paradigms: centralized and decentralized approaches.

The classical MAPF framework employs discrete-time models characterized by synchronized action execution and scenario-based conflict constraints, providing formal mathematical foundations for algorithmic development [16]. However, these simplified assumptions often create discrepancies between theoretical models and practical implementations.Learning-based MAPF algorithms encounter particular challenges in partially observable environments where individual agents face operational uncertainties and potential deadlock situations [10]. Incorporating comprehensive constraint handling mechanisms becomes imperative to prevent deadlocks and ensure system safety. Although current methodologies can generate collision-free optimal paths, computational complexity escalates exponentially with increasing agent populations - a critical limitation driving ongoing research efforts. Recent advancements focus on enhancing algorithmic adaptability to dynamic environments and improving computational efficiency for large-scale multi-agent systems. Researchers continue to refine MAPF solutions through innovative approaches that balance path optimality, computational tractability, and real-world applicability.

2.2 Development of Deep Reinforcement Learning Algorithms

Extensive research on single-agent path planning has yielded substantial theoretical advancements, establishing a robust foundation for subsequent multi-agent planning algorithms [14]. The inception of Deep Reinforcement Learning (DRL) can be traced to

DeepMind, which pioneered the integration of Deep Learning (DL) and Reinforcement Learning (RL) to develop a novel DRL framework. A landmark application of DRL was demonstrated by AlphaGo,which defeated human champions in 2016 [17]. Subsequently, OpenAI introduced the multi-agent DRL (MADRL)-based OpenAI-Five algorithm, achieving unprecedented success in 5v5 Dota 2 matches against professional players [18], thereby showcasing MADRL's superior capability in handling high-dimensional, multi-agent interactive environments.

DRL synergizes the reward-driven environmental interaction mechanism of RL with the high-dimensional feature extraction capabilities of DL. By leveraging deep neural networks to directly process complex environmental data from multiple agents, DRL enables the derivation of optimal path planning strategies. Among learning-based solutions, DRL-based approaches have emerged as a prominent research direction for developing safe and robust Multi-Agent Path Finding (MAPF) solutions [19]. These methods exhibit significant advantages, including decentralized execution, enhanced adaptability to novel environments and uncertainties, and improved applicability in partially observable scenarios [20]. With the maturation of DL techniques, the integration of Deep Q-Networks (DQN) with RL has become a focal point, laying the groundwork for the rise of Multi-Agent Deep Reinforcement Learning (MADRL) and shifting research toward more realistic and complex environments [20].

In early explorations of DRL for MAPF, researchers combined RL principles with multi-agent path planning in simplified grid environments, enabling agents to learn collision-avoidance strategies through trial-and-error interactions. Subsequent studies introduced iterative improvements to enhance algorithmic performance. For instance, the PRIMAL algorithm incorporated Asynchronous Advantage Actor-Critic (A3C) frameworks with behavior cloning for supervised RL training, achieving improved planning efficiency [21]. However, its reliance on centralized planners limited scalability in large-scale multi-agent systems and dynamic scenarios. To address these limitations, decentralized approaches such as Distributed Heuristic Coordination (DHC) were proposed. DHC enhances flexibility by integrating heuristic inputs of all potential shortest paths and employs multi-head attention mechanisms to model inter-agent interactions [22]. Such decentralized methods leverage local observations for decision-making, demonstrating superior robustness to environmental disruptions.

Despite these advancements, learning-based algorithms face inherent challenges. Early RL-based MAPF solutions predominantly utilized Independent Q-Learning (IQL), which inherently induces environmental non-stationarity, constrained decision-making capabilities, and insufficient inter-agent coordination. Solely relying on local observations often leads to deadlock scenarios due to the lack of collaborative mechanisms. Recent developments in multi-agent communication learning have prompted the integration of communication protocols into MAPF frameworks, enabling agents to exchange partial observations for distributed decision-making [23]. However, indiscriminate communication may introduce redundant information, increasing bandwidth overhead and degrading system efficiency. To mitigate this, the Decentralized Causal Communication (DCC) algorithm selectively transmits critical information among neighboring agents, effectively reducing communication redundancy while maintaining decision quality [24].

3 Reinforcement Learning-Based Algorithms and Models for MAPF

3.1 DDPG and Extended

Deep Deterministic Policy Gradient(DDPG) is a deep reinforcement learning algorithm designed for continuous action spaces, integrating the core principles of the Deterministic Policy Gradient (DPG) and Deep Q-Network (DQN) [25]. Building upon the DPG framework proposed by Silver et al., DDPG optimizes action selection through deterministic policies to mitigate the variance issues inherent in stochastic policies, while leveraging DQN's experience replay and target network mechanisms to address training instability in deep neural networks [26]. However, the original DPG was limited to low-dimensional states and linear function approximations, restricting its applicability to complex tasks. DDPG innovatively combines these mechanisms with DPG, establishing its theoretical foundation by directly optimizing action selection via deterministic policies. A notable advantage of DDPG lies in its architectural simplicity, requiring only a basic actor-critic framework and learning algorithm with minimal complexity, thereby facilitating implementation and scalability to intricate problems and large-scale networks.

To extend DDPG to multi-agent environments, researchers developed the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm. In multi-agent scenarios, each agent's behavior is influenced by the policies of others, significantly increasing environmental dynamics and complexity [27]. MADDPG adopts a dual-network structure (critic and policy networks) incorporating target network techniques. Each agent is equipped with an independent actor-critic architecture, in which the critic network accesses global information during training, while the actor network relies solely on local observations for decentralized execution. This centralized training with decentralized execution (CTDE) paradigm effectively addresses cooperation and competition challenges in complex multi-agent environments, substantially broadening DDPG's applicability [27]. By integrating centralized learning with distributed decision-making, MADDPG enhances agents' capability to learn strategies in mixed cooperative-competitive scenarios, establishing a versatile framework for multi-agent reinforcement learning. Despite challenges in dimensionality and computational efficiency, the algorithm's innovation and empirical validation have provided critical insights for subsequent research.

Further improvements to the DDPG framework include Twin Delayed DDPG (TD3), which addresses overestimation bias and policy update instability in DDPG. TD3 introduces dual Q-networks to reduce value overestimation and implements delayed policy network updates to enhance stability. In continuous control tasks, TD3 demonstrates superior performance compared to DDPG, achieving more stable and efficient policy learning, thereby advancing the capabilities of DDPG-based algorithms in complex environments [28].

3.2 PPO Algorithm and Extended

Proximal Policy Optimization (PPO) represents a pivotal advancement in reinforcement learning (RL), specifically designed to address training instability in policy gradient methods. Its predecessor, Trust Region Policy Optimization (TRPO), introduced by Schulman et al. in 2015, aimed to ensure monotonic policy improvement by constraining policy update steps via Kullback-Leibler (KL) divergence bounds [29]. While TRPO stabilizes training through KL divergence constraints, its computational complexity and implementation challenges limited practical adoption. Traditional policy gradient methods, including TRPO, faced inherent limitations such as unstable training dynamics and low sample efficiency. PPO emerged as a streamlined and efficient alternative, retaining the core benefits of TRPO while overcoming its computational barriers.

The PPO algorithm, first proposed by Schulman et al., introduces a clipping mechanism to constrain policy update ratios (i.e., probability ratios) within a predefined range, effectively preventing excessively large policy deviations [30]. Additionally, PPO incorporates adaptive KL penalty mechanisms and truncated importance sampling to regulate update magnitudes. By alternating between data sampling via environmental interactions and optimizing a surrogate objective function using stochastic gradient ascent, PPO addresses the instability issues plaguing conventional policy gradient methods like TRPO [30]. Key advantages of PPO include implementation simplicity, computational efficiency, and robust performance across diverse tasks such as robotic control and Atari gaming benchmarks [31]. Empirical studies demonstrate PPO's superiority in continuous control tasks (e.g., MuJoCo environments), outperforming traditional algorithms like TRPO and Advantage Actor-Critic (A2C) in terms of stability and sample efficiency. The integration of clipping mechanisms with first-order optimization enables PPO to achieve state-of-the-art performance in sample efficiency, training stability, and task generalization.

Notably, Engstrom et al. (2020) revealed that code-level optimizations in PPO implementations — including value function clipping, reward scaling, orthogonal initialization, and layer normalization—play a critical role in its performance enhancements [32]. While prior studies attributed PPO's superiority to its ratio-clipping mechanism, this work demonstrated that code-level adjustments exert a more significant impact on final rewards than algorithmic core mechanisms. These findings prompted researchers to reassess methodologies for quantifying the contributions of individual algorithmic components to overall performance.

3.3 SAC Algorithm and Extended

The Soft Actor-Critic (SAC) algorithm represents a paradigm in deep reinforcement learning (DRL) under the maximum entropy framework, deriving its theoretical foundation from maximum entropy reinforcement learning. This framework extends traditional RL objectives by incorporating entropy regularization to encourage policy stochasticity, thereby balancing exploration-exploitation trade-offs and mitigating premature policy convergence in continuous action spaces [33]. SAC evolved from earlier methodologies such as Soft Q-Learning (SQL) and hybrid Actor-Critic architectures integrating policy gradients with value functions [34].

The algorithm unifies the conventional goal of maximizing expected returns with entropy maximization, promoting action diversity while pursuing high rewards. This dual objective transforms policy optimization into a joint maximization of expected rewards and policy entropy, enhancing sample efficiency and robustness. SAC employs an Actor-Critic architecture augmented with double Q-networks to mitigate Q-value overestimation bias and target networks for stable parameter updates, effectively addressing local optima and training instability [34]. Empirical evaluations demonstrate SAC's superior performance over DDPG across all MuJoCo benchmarks. As an off-policy algorithm leveraging experience replay buffers, SAC achieves higher sample reuse rates and asymptotically matches or surpasses PPO in most tasks without manual temperature parameter tuning. Its exceptional performance in high-dimensional complex tasks underscores its significance in continuous control domains. SAC's development reflects the evolutionary trajectory of RL from deterministic to stochastic policies and from single-objective optimization to multi-objective equilibrium, with its core innovation—the integration of entropy regularization and double Q-learning—establishing a framework for efficient exploration and stable learning in intricate environments.

Recent advancements propose the CTSAC algorithm, which integrates Transformer networks with SAC to enhance environmental reasoning capabilities and address traditional SAC's limitations in long-term dependency modeling and training efficiency [35]. CTSAC incorporates Transformer layers into SAC's Actor-Critic architecture, leveraging self-attention mechanisms to capture long-range dependencies in historical states and environmental contexts. This innovation enables improved avoidance of local optima and enhances the foresight of path planning in complex scenarios. Additionally, a curriculum learning strategy based on progressive task difficulty escalation is introduced, employing periodic revisitation of prior tasks to alleviate catastrophic forgetting.

4 Conclusion

In recent years, with the development and maturation of MAPF technology, it has been widely applied in real life. Researchers have continuously improved algorithms to address the limitations of mainstream methods, such as insufficient scalability and deadlock issues, optimizing MAPF algorithms to minimize redundant search efforts and balance solution efficiency with quality. Meanwhile, enhancing environmental adaptability in increasingly complex and diverse application scenarios remains a challenge for algorithmic optimization. With the rapid advancement of artificial intelligence, MAPF algorithms based on multi-agent deep reinforcement learning (DRL) have broader application prospects. Efficient algorithms can improve the response speed and processing capability of path planning, enabling effective collaboration among multiple agents. Through systematic analysis of recent DRL-based MAPF algorithms, this paper identifies potential future directions:

(1) DRL remains one of the most promising approaches for developing safe and robust MAPF solutions. Future work should optimize learning algorithms to handle large-scale agents, reduce computational complexity and resource consumption, and improve scalability while balancing efficiency and solution quality.

(2) Transitioning from centralized to distributed planning is critical, emphasizing independent path planning for individual agents alongside global consistency through communication and coordination mechanisms. Traditional methods excel in small-scale static environments, while learning-based approaches show higher adaptability and efficiency in large-scale dynamic scenarios.

(3) Improving adaptability to dynamic environments remains a priority. While DRL has succeeded in global path planning for known maps, its performance in unknown environments without prior maps varies significantly. Future research may integrate frameworks like communication systems to enhance adaptability in dynamic settings.

References

1. cloud-based multi-hop multi-robot wireless networks', IETE Tech. Rev., vol. 37, no. 1, pp. 98 – 107, 2020
2. Liu, Z. F., Cao, L., Lai, J., et al.: 'Overview of multi-agent path finding.' Computer Engineering and Applications, 2022, 58(20): 43-62
3. Stern, R., Sturtevant, N. R., Felner, A., et al.: 'Multi-agent pathfinding: Definitions, variants, and bench-marks.' In: Twelfth Annual Symposium on Combinatorial Search. 2019
4. Wu, W. J., Wang, T. D., Sun, Y., et al.: 'Survey of multi-agent path finding technology.' Journal of Beijing University of Technology, 2024, 50(10): 1263-1272

5. Oliveira, I. R. L., & Brandão, A. S.: 'Deep reinforcement learning for mapless robot navigation systems.' 2023 Latin American Robotics Symposium (LARS), 2023 Brazilian Symposium on Robotics (SBR), and 2023 Workshop on Robotics in Education (WRE) (pp. 331-336). Salvador, Brazil. 2023
6. Shi, D. X., Peng, Y. X., Yang, H. H., et al.: 'DQN based Multi-agent Motion Planning Method with Deep Reinforcement Learning.' Computer Science, 2024, 51(2):268-277.
7. Sartoretti, G., Kerr, J., Shi, Y., et al.: 'Primal: Pathfinding via reinforcement and imitation multi-agent learning.' IEEE Robotics and Automation Letters, 2019, 4(3):2378-85
8. Zhang, Y. L., Li, Z. W., Xu, J. H., Jiang, Y. C., Cui, Y.: 'Multi - Agent Path Planning Based on Congestion Awareness and Caching Communication.' Computer Science. 2024
9. Riviere, B., Hönig, W., Yue, Y., et al.: 'Glas: Global-to-local safe autonomy synthesis for multi-robot motion planning with end-to-end learning.' IEEE Robotics and Automation Letters, 2020, 5(3): 4249-4256.
10. Ye, Z. H.: 'Warehouse Robot Path Planning Based on Multi - Agent Reinforcement Learning.' Harbin Institute of Technology, 2023
11. Chung, J., Fayyad, J., Younes, Y. A., et al.: 'Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding.' Artificial Intelligence Review, 2024, 57(2):41.
12. Li, Q., Gama, F., Ribeiro, A., et al.: 'Graph neural networks for decentralized multirobot path planning' // 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, USA, 2020: 11785-11792.
13. Laurent, F., Schneider, M., Scheller, C., et al.: 'Flatland competition 2020: Mapf and marl for efficient train coordination on a grid world.' In: NeurIPS 2020 Competition and Demonstration Track, PMLR, pp 275 – 301. 2021
14. Liu, Z., Cao, L., Lai, J., Chen, X., Chen, Y.: 'A review of multi-intelligent body path planning.' Comput. Eng. Appl. 2022, 58, 43 – 62
15. Yang, L., Li, P., Qian, S., Quan, H., Miao, J., Liu, M., Hu, Y., & Memetimin, E.: 'Path Planning Technique for Mobile Robots: A Review.' Machines, 11(10), 980. 2023 https://doi.org/10.3390/machines11100980
16. Reijnen, Y., Zhang, W., Nuijten, C., and Goldak-Altgassen, M.: 'Combining Deep Reinforcement Learning with Search Heuristics for Solving Multi-Agent Path Finding in Segment-based Layouts', 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 2020, pp. 2647-2654.
17. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D., 'Mastering the game of go with deep neural networks and tree search,' Nature, vol. 529, no. 7587, pp. 484 – 489, Jan. 2016
18. Berner, C., et al.: 'Dota 2 with large scale deep reinforcement learning,' arXiv:1912.06680. Available: https://arxiv.org/abs/1912.06680. 2019
19. Manchella, K., Umrawal, A. K. and Aggarwal, V., 'FlexPool: A distributed model-free deep reinforcement learning algorithm for joint passengers and goods transportation,' IEEETrans. Intell. Transp. Syst., vol. 22, no. 4, pp. 2035 – 2047, Apr. 2021.

20. Honari, H., Khodaygan, S.: 'Deep reinforcement learning-based framework for constrained any-objective optimization.' J Ambient Intell Humaniz Comput pp 1 – 17. 2023
21. Sartoretti, G. et al.: 'PRIMAL: Pathfinding via Reinforcement and Imitation Multi-Agent Learning,' in IEEE Robotics and Automation Letters, vol. 4, no. 3, pp. 2378-2385, 2019
22. Ma, Z., Luo, Y., and Ma, H.: 'Distributed Heuristic Multi-Agent Path Finding with Communication,' 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 8699-8705
23. Seoung, K. L. & Nakju, D.: 'Local path planning scheme for car-like robots' shortest turning motion using geometric analysis. Intelligent Service Robotics, (prepublish), 1-39. 2025
24. Das, S., Biswas, A., Saxena, A.: 'DCC: A Cascade-Based Approach to Detect Communities in Social Networks.' Springer, Singapore. 2024. https://doi.org/10.1007/978-981-99-6690-5_28
25. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., & Wierstra, D.: 'Continuous control with deep reinforcement learning.' arXiv preprint arXiv:1509.02971. 2015
26. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. & Riedmiller, M.: 'Deterministic Policy Gradient Algorithms.' Proceedings of the 31st International Conference on Machine Learning, in Proceedings of Machine Learning Research. 32(1):387-395. 2014
27. Lowe, R., et al.: 'Multi-Agent Actor-Critic for Mixed Cooperative Competitive Environments.' Advances in Neural Information Processing Systems 30: 6379-6390. 2017
28. Fujimoto, S., Hoof, H. & Meger, D.: 'Addressing Function Approximation Error in Actor-Critic Methods.' Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research. 80:1587-1596. 2018
29. Schulman, J., Levine, S., Abbeel, P., Jordan, M. & Moritz, P.: 'Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, in Proceedings of Machine Learning Research. 37:1889-1897. 2015
30. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.: 'Proximal Policy Optimization Algorithms.' arXiv preprint arXiv:1707.06347. 2017
31. Andrychowicz, M., Baker, B., Chociej, M., Józefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., & Zaremba, W.: 'Learning dexterous in - hand manipulation. Journal of Robotics', 39(1), 1 - 13. 2020
32. Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., & Mądry, A.: 'Implementation matters in deep policy gradients: A case study on PPO and TRPO.' arXiv preprint arXiv:2005.12729. 2020
33. Ziebart, B. D., Maas, A., Bagnell, J. A., & Dey, A. K.: 'Maximum Entropy Inverse Reinforcement Learning.' Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 1433 - 1438. 2008
34. Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S.: 'Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.' Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research. 80:1861-1870. 2018

35. Yang, C., Bi, S., Xu, Y., & Zhang, X.: 'CTSAC: Curriculum-Based Transformer Soft Actor-Critic for Goal-Oriented Robot Exploration.' arXiv preprint arXiv:2503.14254. 2025