Proceedings of Machine Learning Research vol 278:1–11, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Survey on Path Planning Based on Deep Reinforcement Learning

Lihan Xu and **Wenzhi Zhang*** ROBOTZWZ@163.COM
Inner Mongolia University of Technology electromechanical department, Hohhot 010000, China
**Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

In recent years, deep reinforcement learning (DRL) has demonstrated significant potential in the field of path planning and control, offering breakthrough solutions for path planning in dynamic and complex environments. DRL has been widely applied in UAV obstacle avoidance, autonomous vehicle path optimization, multi-robot coordination, and complex terrain navigation, demonstrating ad-vantages such as superior path quality, improved smoothness, and enhanced safety. This paper provides a systematic review of recent advances and applications of DRL core techniques. Value-based methods (e.g. DQN) significantly improve decision-making efficiency through optimized reward design and network architectures. Policy gradient algorithms (such as PPO, DDPG, and TD3) achieve high-precision control in continuous action spaces. The Actor-Critic framework, combined with double Q-networks and delayed update mechanisms (e.g. TD3), further expands the application scenarios. Future research should focus on enhancing cross-scenario generalization capabilities and improving deployment efficiency at the industrial level, thereby promoting the practical application of DRL in autonomous driving and industrial robotics.

Keywords: Mobile robots; Path planning algorithms; Reinforcement learning; Deep reinforcement learning.

1. Introduction

With the rapid development of automation and artificial intelligence (AI), mobile robots have been widely deployed in fields such as autonomous driving and logistics warehousing. As a core challenge in robot navigation and autonomous driving, path planning requires the generation of efficient and safe trajectories in complex environments. Traditional methods, such as A* and RRT, rely on deterministic models and struggle to handle dynamic environments and high-dimensional state spaces (Xu et al., 2023; Niu et al., 2022; Huang et al., 2023). Deep reinforcement learning (DRL) integrates the perception capability of deep learning with the decision-making mechanism of reinforcement learning, offering a new paradigm for solving complex path planning problems (Wang et al., 2024; Liu et al., 2019a,b; Feng et al., 2021; Sun et al., 2021).

2. Path Planning Algorithms

In this review, a comprehensive literature search was conducted across several major academic databases, including Web of Science, IEEE Xplore, ScienceDirect, CNKI, and Google Scholar. The search primarily focused on publications from 2018 to 2023, while also including classic papers on fundamental algorithms. Keywords such as “deep reinforcement learning”, “path planning”, “dynamic obstacle avoidance”, “UAV navigation”, and “multi-objective DRL” were used. Articles were selected based on their relevance, citation impact, and experimental validation of DRL algorithms in path planning tasks. Priority was given to studies providing empirical results or comparative analyses that highlight the advantages and limitations of specific DRL methods.

The core task of path planning is to determine the optimal motion trajectory for mobile agents (such as robots and autonomous vehicles) from a starting point to a target point within a given environment. The key objectives include ensuring safety, efficiency, and dynamic adaptability. Existing path planning methods can be categorized into four main classes: traditional search algorithms, optimization-based methods, intelligent algorithms, and deep reinforcement learning (DRL). These approaches exhibit significant differences in performance.

Traditional search algorithms (e.g., A*, Dijkstra, and RRT variants) are known for their high computational efficiency. Among them, the A* algorithm (Abdelfetah et al., 2024) guarantees global optimality, while RRT variants perform well in high-dimensional space exploration (Li et al., 2024). However, these methods rely heavily on precise environment modeling, exhibit poor dynamic adaptability, and struggle to handle multi-objective optimization problems effectively (Zhou et al., 2022; František et al., 2014).

Optimization-based methods (e.g. model predictive control, MPC) excel at multi-objective joint optimization tasks, including path length, comfort, and energy consumption. These methods are particularly suitable for real-time adjustments in dynamic environments. However, their high computational complexity, sensitivity to model accuracy, and limited real-time performance in high-dimensional scenarios remain major challenges (Siboo et al., 2023).

Intelligent algorithms (e.g., genetic algorithms and ant colony optimization) do not require prior knowledge of the environment and possess global search capabilities, making them suitable for unstructured environments. However, they suffer from slow convergence, limited capability in handling high-dimensional state spaces, and weak robustness due to reliance on empirical parameter tuning (Song et al., 2019; Masehian and Sedighizadeh, 2010; Guo et al., 2020).

A comparative summary of these algorithms is presented in Table 1.

Table 1: Comparison of Path Planning Algorithms

Algorithm Category	Representative Methods	Advantages	Limitations
Traditional Search Algorithms	A*, Dijkstra, RRT, RRT*	High computational efficiency; guaranteed global optimality (A*)	Dependence on precise environment modeling; poor dynamic adaptability
Optimization-based Methods	MPC	Capable of multi-objective optimization; suitable for dynamic environments	High computational complexity; limited real-time performance
Intelligent Algorithms	ACO, GA	No prior environmental knowledge required; strong global search capability	Slow convergence; limited capability in handling high-dimensional state spaces

3. Deep Reinforcement Learning for Path Planning

3.1. Deep Reinforcement Learning

Traditional Reinforcement Learning (RL), such as Q-learning (Křetínský and Meggendorfer, 2019; Sutton, 1988; Zhao et al., 2016), suffers from the curse of dimensionality in high-dimensional state spaces (Abdelfetah et al., 2024). Deep Reinforcement Learning (DRL) integrates deep neural networks (DNNs) (He et al., 2015) to efficiently approximate value functions (e.g., state value V(s) and action value Q(s,a)) or policy functions (e.g., deterministic policies in DDPG and stochastic policies in PPO), thereby overcoming the limitations of traditional methods (An et al., 2023; Sutton and Barto, 1998), as illustrated in Figure 1.

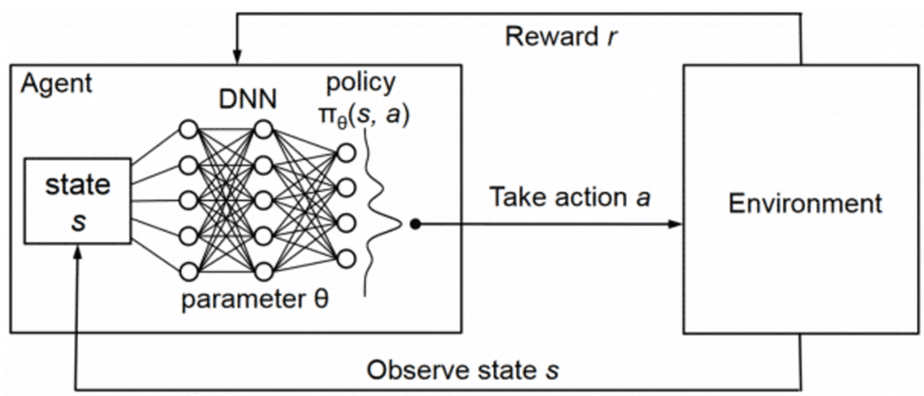

Figure 1: Basic workflow of DRL

DRL is theoretically grounded in the Markov Decision Process (MDP) framework, where environmental interactions are modeled by the tuple (S, A, P, R). The objective is to maximize the cumulative re-ward, as defined in Eq.1, which guides the agent's decision-making and learning process.

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{1}$$

In DRL, deep neural networks are responsible for extracting and processing high-dimensional state features. The experience replay mechanism improves sample efficiency by reusing historical data, while the target network stabilizes Q-value estimation and reduces oscillations during training (Zhang et al., 2021).

3.2. DRL Path Planning Algorithm

Value-Based Methods. Value-based methods are typically applied in discrete state and action spaces. These methods first evaluate the value function and then optimize the policy accordingly. Depending on the input variables, value functions are classified into the state value function $V(s)$ and the state-action value function $Q(s, a)$. The state value function, as defined in Eq.2, represents the expected cumulative return when the agent is in state s. A higher value indicates a more favorable state. The state-action value function, as shown in Eq.3, estimates the cumulative return obtained when the agent takes action a in state s.

$$V(s) = E[\sum_t \gamma^t r_t | s] \tag{2}$$

$$Q(s,a) = E[\sum_t \gamma^t r_t | s, a] \tag{3}$$

The Deep Q-Network (DQN) represents a significant breakthrough in DRL by integrating deep neural networks with Q-learning, thereby overcoming the curse of dimensionality faced by traditional reinforcement learning in high-dimensional state spaces. The core innovations of DQN include Experience Replay and a Target Network, which enhance training efficiency by reusing past experiences and stabilizing the learning targets (Zhang et al., 2021; Zhu et al., 2018).

In the context of path planning, notable improvements to DQN include Dueling DQN and Hierarchical DQN. Dueling DQN separates the state value function and the advantage function, thereby improving adaptability in complex environments (Masehian and Sedighizadeh, 2010). Hierarchical DQN reduces learning complexity by decomposing tasks, such as separating global path planning from local obstacle avoidance (Yin et al., 2020). Additionally, DQN has been integrated with traditional path planning algorithms, such as the Probabilistic Roadmap (PRM), to develop the PRM+DRL algorithm, which enhances generalization capability in dynamic environments (Wu et al., 2023).

Policy Gradient Algorithms. Policy Gradient (PG) algorithms directly optimize the parameters of the policy function to maximize the agent's expected cumulative reward, as defined in Eq.4. Unlike value-based methods, PG algorithms do not require explicit estimation of the state-action value function $Q(s, a)$; instead, the learned policy model directly determines the agent's actions.

$$maxE_{\tau\ \pi\theta}[R(\tau)] \tag{4}$$

Based on the policy characteristics, PG algorithms can be categorized into stochastic and deterministic approaches. Stochastic policies inherently promote exploration and eliminate the reliance on tradition-al ε-greedy strategies, making them more prevalent in practical applications. In contrast, deterministic policy gradients directly output specific actions a for a given state s, as shown in Eq.5.

$$a_t = \mu(s_t) \tag{5}$$

The Proximal Policy Optimization (PPO) algorithm improves training stability by constraining policy updates through a clipping mechanism and enhances sample efficiency by leveraging the advantage function to evaluate action performance (Iskandar and Kovács, 2024). PPO has demonstrated excellent performance in dynamic obstacle avoidance and generalization across diverse scenarios (Azar et al., 2021).

Actor-Critic Based Methods. The reinforcement learning method that combines both value func-tions and policy functions is referred to as the Actor-Critic method. This method introduces two neural networks: the Actor network, which outputs the actions, and the Critic network, which evaluates the value of these actions, simultaneously learning the optimal policy and estimating the optimal value function.

The advantage of the Actor-Critic method lies in its ability to simultaneously optimize both the val-ue function and policy function, thus enhancing learning efficiency and decision-making performance. Moreover, the Actor-Critic method effectively addresses continuous action space problems and adapts well to noisy environments. Compared to traditional policy gradient methods, it offers superior efficiency and performance.

DDPG: The Actor network outputs continuous actions, while the Critic network evaluates the value of these actions, utilizing a target network and experience replay to stabilize the training process , mak-ing it suitable for continuous action spaces. DDPG has demonstrated excellent performance in drone path planning and autonomous driving, effectively handling fine control problems within continuous action spaces (Siboo et al., 2023).

TD3: An improved version of DDPG. By incorporating a double Q-network and delayed policy up-date mechanisms, TD3 effectively mitigates the overestimation of action values (Tariq et al., 2024). TD3 uses two Critic networks to evaluate Q-values, selecting the smaller value as the target to reduce overestimation, and delays updates to the Actor network to ensure the stability of the Critic network. In dynamic envi-ronments, TD3 outperforms DDPG and PPO, particularly in dynamic obstacle handling and real-time path re-planning tasks.

A comparison of these algorithms is summarized in Table 2.

Table 2: Comparison of DRL algorithms

Method	Core Mechanism	Application Scenario	Advantages	Limitations
DQN	Experience replay and target network for stable Q-value estimation	Grid map navigation	Guaranteed theoretical convergence and computational efficiency	Requires discretization of continuous actions; suffers from dimensionality explosion
Dueling DQN	Separation of state-value function V(s) and advantage function A(s,a)	Complex urban navigation	Enhanced capability for environmental feature extraction	Increased network complexity
PPO	Clipped surrogate objective constraining policy updates	Dynamic obstacle avoidance and multi-objective optimization	High training stability and adaptability to high-dimensional states	Sensitive to hyperparameters; limited real-time performance
DDPG	Deterministic policy gradient with target network	Continuous UAV attitude control	High-precision continuous output	Low exploration efficiency; prone to local optima
TD3	Twin delayed Q-networks, delayed policy updates, and target policy smoothing	Real-time replanning in dynamic obstacle environments	Mitigates overestimation and improves stability	Delayed policy updates may slow convergence

4. Innovations and Applications of DRL-Based Path Planning Algorithms

In dynamic environments, classic reinforcement learning algorithms such as DDQN and SARSA have been applied to real-time path planning and dynamic obstacle avoidance tasks, significantly improving the autonomous flying capabilities and environmental adaptability of UAVs (Yao et al., 2024). Furthermore, Deep Q-Network (DQN), as an end-to-end reinforcement learning framework, can directly learn strategies from high-dimensional sensory inputs. It has achieved near-human-level performance in tasks such as Atari 2600 games, establishing the foundation of value-based methods for path planning in complex environments (Mnih et al., 2015).

To address the overestimation bias in value-based methods, researchers have proposed an improved strategy based on Double Q-Learning, which limits overestimation by using the minimum value between two critics, thereby enhancing the algorithm's performance in OpenAI Gym tasks (Fujimoto et al., 2018). Additionally, Dueling DQN decouples the state-value function and the advantage function, improving policy evaluation efficiency, particularly in scenarios where action values are similar (Wang et al., 2016).

In terms of policy optimization, the Proximal Policy Optimization (PPO) algorithm has been intro-duced, which combines the stability of TRPO with simpler implementation. PPO demonstrates a good balance between sample efficiency and performance in robot control and Atari games, becoming one of the mainstream policy gradient methods (Schulman et al., 2017).

At the same time, the Prioritized Experience Replay (PER) mechanism has been improved by inte-grating immediate rewards, TD errors, and actor loss functions to calculate experience priorities. It adaptively adjusts for active samples, reducing collision frequency, accelerating training speed, and enhancing path planning performance (Cheng et al., 2023). Tang et al. (2023) combined PER with D3QN to propose a UAV path planning method, introducing a collision threat assessment model and designing an action space that effectively enhances the algorithm's safety and generalization ability.

With the development of sensor technology, researchers have combined Deep Reinforcement Learning (DRL) with LiDAR (Light Detection and Ranging) to propose the TD3-DWA hybrid algorithm. This method integrates the traditional Dynamic Window Approach (DWA) with Double Delayed Deep Deterministic Policy Gradient (TD3) and optimizes the sampling interval parameters, effectively avoiding both static and dynamic obstacles. It significantly enhances the reliability and safety of robot path planning (Liu et al., 2024).

In the context of autonomous driving applications, researchers have designed a collision prediction model based on Gaussian processes and vehicle dynamics, integrating it into a reinforcement learning framework. This model enables explicit risk perception and post-event interpretability. Experimental results show that this approach outperforms traditional models (such as the intelligent driver model) in terms of safety and speed, with the average collision rate reduced by 15% (Candela et al., 2023).

Furthermore, a combination of the Double Bilinear Delayed Deep Deterministic Policy Gradient (TD3) and Probabilistic Roadmap (PRM) methods has been applied to indoor mobile robot path plan-ning, effectively improving the model's generalization ability and development efficiency (Gao et al., 2020).

To address the local obstacle avoidance and path planning issues in unfamiliar environments, re-searchers proposed a UAV autonomous local path planning algorithm based on the TD3 strategy. This method was validated in the Gazebo simulation environment, achieving a path planning success rate of 93% in obstacle-free environments and 92% in environments with obstacles, demonstrating excel-lent autonomous decision-making ability and environmental adaptability (Zhao et al., 2024).

5. Technical Challenges And Development Trends

5.1. Technical Challenges

Although DRL shows great potential in path planning, several challenges remain:

Low Sample Efficiency and High Training Costs. DRL heavily relies on large-scale interaction data, resulting in high computational demands and prolonged training cycles, which limit its application in data-scarce or resource-constrained scenarios (Mnih et al., 2015; Fujimoto et al., 2018).

Limited Generalization in Dynamic Environments. Most existing algorithms perform poorly when facing randomly moving obstacles or sudden hazards, reducing adaptability in complex and dynamic environments (Yao et al., 2024; Tang et al., 2023; Liu et al., 2024).

Safety and Real-Time Performance Constraints. The high computational complexity of DRL models makes it difficult to meet real-time requirements in tasks such as autonomous driving . Additionally, the lack of comprehensive safety validation limits their application in safety-critical scenarios (Liu et al., 2024; Candela et al., 2023).

Challenges in Multi-Objective Optimization. Current reward-based approaches often rely on empirically set weights to balance objectives such as path length, energy consumption, and comfort. This limits the ability to achieve globally optimal solutions in complex tasks (Schulman et al., 2017; Zhao et al., 2024).

5.2. Development Trends

Despite the challenges faced by DRL in path planning, its future development directions are clear.

Enhancing Generalization Ability. Improving cross-scenario adaptability is crucial. Techniques such as meta-learning (e.g., MAML) and domain randomization help models extract general strategies, enabling better transferability to new environments and improving real-world robustness.

Improving Deployment Efficiency. Integrating multi-modal perception (e.g., LiDAR, vision, and graph neural networks) enhances environmental understanding. Meanwhile, model optimization tech-niques like neural architecture search (NAS) and pruning can effectively reduce computational load, supporting real-time deployment on embedded platforms.

Digital Twin Integration. Digital twin systems enable virtual-real closed-loop learning, which not only reduces energy consumption but also improves task efficiency, offering promising applications in smart manufacturing and autonomous driving.

Emphasis on Safety and Multi-Objective Optimization. Future research will focus on incorporating safety constraints and improving interpretability. Advanced optimization methods, such as Pareto-based approaches and risk-sensitive reinforcement learning, are expected to enhance solution quality and reliability.

6. Conclusion

Deep reinforcement learning (DRL) has demonstrated its powerful potential in path planning, especially in fields such as autonomous driving, robot navigation, and drone path planning. Through algorithm optimization, multi-objective trade-offs, and adaptation to dynamic environments, DRL offers new solutions to address the limitations of traditional path planning methods. However, issues such as sample efficiency, adaptation to dynamic environments, and safety remain key challenges for future research. Future studies should further integrate traditional planning methods, hardware acceleration technologies, and interdisciplinary approaches (e.g., game theory, cognitive science) to achieve more efficient, reliable, and universal path planning systems.

Acknowledgments

Thanks for the support of the following projects: 2021 Inner Mongolia Autonomous Region Science and Technology Plan project - Development and demonstration application of large pressure vessel automatic inspection robot. Project Fund Number: Inner Mongolia Autonomous Region Science and Technology Program (2021GG0260).

References

Hentout Abdelfetah, Maoudj Abderraouf, and Kouider Ahmed. Shortest path planning and efficient fuzzy logic control of mobile robots in indoor static and dynamic environments. *Romanian Journal of Information Science and Technology*, 27(94):21–36, 2024. ISSN 1453-8245. doi: 10.59277/ROMJIST.2024.94.02.

Tianyi An, Ning Li, and Chao Wang. Research on lightweight algorithms for deep reinforcement learning. *Computer Science and Application*, 13(04):779–788, 2023.

Ahmad Taher Azar, Anis Koubaa, Nada Ali Mohamed, Habiba A. Ibrahim, Zahra Fathy Ibrahim, Muhammad Kazim, Adel Ammar, Bilel Benjdira, Alaa M. Khamis, Ibrahim A. Hameed, and Gabriella Casalino. Drone deep reinforcement learning: A review. *Electronics*, 10(9), 2021. doi: 10.3390/electronics10090999.

Eduardo Candela, Olivier Doustaly, Leandro Parada, Felix Feng, Yiannis Demiris, and Panagiotis Angeloudis. Risk-aware controller for autonomous vehicles using model-based collision prediction and reinforcement learning. *Artificial Intelligence*, 320:103923, 2023. doi: https://doi.org/10.1016/j.artint.2023.103923.

Nuo Cheng, Peng Wang, Guangyuan Zhang, Cui Ni, and Erkin Nematov. Prioritized experience replay in path planning via multi-dimensional transition priority fusion. *Frontiers in Neurorobotics*, Volume 17 - 2023, 2023. doi: 10.3389/fnbot.2023.1281166.

Shuo Feng, Hong Shu, and Buqing Xie. Path planning in 3d environments based on improved deep reinforcement learning. *Computer Applications and Software*, 38(01):250–255, 2021. ISSN 1000-386X.

Duchoň František, Babinec Andrej, Kajan Martin, Beňo Peter, Florek Martin, Fico Tomáš, and Jurišica Ladislav. Path planning with modified a star algorithm for a mobile robot. *Procedia Engineering*, 96:59–69, 2014. ISSN 1877-7058. doi: https://doi.org/10.1016/j.proeng.2014.12.098. Modelling of Mechanical and Mechatronic Systems.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 10–15 Jul 2018.

Junli Gao, Weijie Ye, Jing Guo, and Zhongjuan Li. Deep reinforcement learning for indoor mobile robot path planning. *Sensors*, 20(19), 2020. doi: 10.3390/s20195493.

Xinghai Guo, Mingjun Ji, Ziwei Zhao, Dusu Wen, and Weidan Zhang. Global path planning and multi-objective path control for unmanned surface vehicle based on modified particle swarm optimization (pso) algorithm. *Ocean Engineering*, 216:107693, 2020. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2020.107693.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. Deep reinforcement learning with an unbounded action space. *CoRR*, abs/1511.04636, 2015. URL http://arxiv.org/abs/1511.04636.

Yuzhou Huang, Lisong Wang, and Xiaolin Qin. A deep reinforcement learning-based dual-layer path planning method for unmanned vehicles. *Computer Science*, 50(01):194–204, 2023. ISSN 1002-137X.

Alaa Iskandar and Béla Kovács. Investigating the impact of curriculum learning on reinforcement learning for improved navigational capabilities in mobile robots. *Inteligencia Artificial*, 27(73): 163–176, Mar. 2024. doi: 10.4114/intartif.vol27iss73pp163-176.

Jan Křetínský and Tobias Meggendorfer. Of cores: A partial-exploration framework for markov decision processes. In *International Conference on Concurrency Theory*, 2019.

Xiaojuan Li, Tao Chen, Ruichun Han, and Jianxuan Liu. Path planning of robotic arms based on improved rrt algorithm in uncertain picking environments. *Transactions of the Chinese Society for Agricultural Machinery*, 45(04):193–198+337, 2024. ISSN 2095-5553. doi: 10.13733/j.jcam.issn.2095-5553.2024.04.028.

Hao Liu, Yi Shen, Chang Zhou, Yuelin Zou, Zijun Gao, and Qi Wang. Td3 based collision free motion planning for robot navigation. In *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 247–250, 2024. doi: 10.1109/CISCE62493.2024.10653233.

Jianwei Liu, Feng Gao, and Xionglin Luo. A review of deep reinforcement learning based on value functions and policy gradients. *Journal of Computer Science*, 42(06):1406–1438, 2019a. ISSN 0254-4164.

Kun Liu, Tingting Zhang, and Lai Chai. Path optimization of intelligent agents based on reinforcement learning algorithms. In *In Proceedings of the 7th China Command and Control Conference*, pages 452–457, 2019b.

Ellips Masehian and Davoud Sedighizadeh. A multi-objective pso-based algorithm for robot path planning. In *2010 IEEE International Conference on Industrial Technology*, pages 465–470, 2010. doi: 10.1109/ICIT.2010.5472755.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Yifeng Niu, Tianqing Liu, Jie Li, and Shengde Jia. A review of cooperative maneuvering flight motion planning methods for drones in dense environments. *Journal of National University of Defense Technology*, 44(04):1–12, 2022. ISSN 1001-2486.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Sanjna Siboo, Anushka Bhattacharyya, Rashmi Naveen Raj, and S. H. Ashwin. An empirical study of ddpg and ppo-based reinforcement learning algorithms for autonomous driving. *IEEE Access*, 11:125094–125108, 2023. doi: 10.1109/ACCESS.2023.3330665.

Xiaoru Song, Yiyue Ren, Song Gao, and Chaobo Chen. A review of path planning for mobile robots. *Computer Measurement & Control*, 27(04):1–5+17, 2019. ISSN 1671-4598. doi: 10.16526/j.cnki.11-4762/tp.2019.04.001.

Huihui Sun, Weijie Zhang, Runxiang Yu, and Yujie Zhang. Motion planning for mobile robots—focusing on deep reinforcement learning: A systematic review. *IEEE Access*, 9:69061–69081, 2021. doi: 10.1109/ACCESS.2021.3076530.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988. doi: 10.1007/BF00115009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 1998.

J. Tang, J. Yang, K. Li, et al. Uav path planning based on prioritized experience replay d3qn. In *Proceedings of the 3rd Unmanned Systems Summit Forum*, pages 203–211, 2023.

Zain Ul Abideen Tariq, Emna Baccour, Aiman Erbad, and Mounir Hamdi. Reinforcement learning for resilient aerial-irs assisted wireless communications networks in the presence of multiple jammers. *IEEE Open Journal of the Communications Society*, 5:15–37, 2024. doi: 10.1109/OJCOMS.2023.3334489.

Zhihao Wang, Weiqiang Yan, and Mingjun Yang. Research on path planning algorithm based on deep reinforcement learning. In *2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pages 268–273, 2024. doi: 10.1109/ICCASIT62299.2024.10827883.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1995–2003. PMLR, 20–22 Jun 2016.

Qu Wu, Yi Zhang, Kun Guo, and Xi Wang. Research on path planning method based on dpes-duelingdqn. *Computer Applications and Software,*, 40(06):147–153+233, 2023.

Hongxin Xu, Zhizhou Wu, and Yunyi Liang. A review of path planning methods for autonomous vehicles based on reinforcement learning. *Application Research of Computers*, 40(11):3211–3217, 11 2023.

Jiangyi Yao, Xiongwei Li, Yang Zhang, Kaiyan Chen, Danyang Zhang, and Jingyu Ji. Deep reinforcement learning path planning algorithm based on sarsa. In *Proceedings of 2024 12th China Conference on Command and Control*, pages 46–56, Singapore, 2024. Springer Nature Singapore.

Changsheng Yin, Ruopeng Yang, Wei Zhu, Xiaofei Zou, and Feng Li. A review of multi-agent hierarchical reinforcement learning. *CAAI Transactions on Intelligent Systems*, 15(04):646–655, 2020.

Rongxia Zhang, Changxu Wu, Tongchao Sun, and Zengshun Zhao. Research progress of deep reinforcement learning and its application in path planning. *Journal of Computer Engineering & Applications*, 57(19):44–56, 2021.

Dongbin Zhao, Kun Shao, Yuanheng Zhu, Dong Li, Yaran Chen, Haitao Wang, Derong Liu, Tong Zhou, and Chenghong Wang. A survey on deep reinforcement learning: A discussion on the development of computer go. *Control Theory & Applications*, 33(06):701–717, 2016.

Feiyu Zhao, Dayan Li, Zhengxu Wang, Jianlin Mao, and Niya Wang. Autonomous localized path planning algorithm for uavs based on td3 strategy. *Scientific Reports*, 14(1):763, 2024. doi: 10.1038/s41598-024-51349-4.

Chengmin Zhou, Bingding Huang, and Pasi Fränti. A review of motion planning algorithms for intelligent robots. *Journal of Intelligent Manufacturing*, 33(2):387–424, 2022. ISSN 1572-8145. doi: 10.1007/s10845-021-01867-z.

Fei Zhu, Wen Wu, Quan Liu, and Yuchen Fu. A deep q-network method based on upper confidence bound experience sampling. *Journal of Computer Research and Development*, 55(08):1694–1705, 2018.