

Multi-Agent Deep Reinforcement Learning in UAV Swarm Formation Control: A Review and Future Challenges

Wenchao He

School of Intelligent Engineering, Yanching Institute of Technology, Langfang, Hebei, 065000, China

ABSTRACT

Due to constraints on payload and detection range, a UAV struggles to perform complex dynamic tasks, making swarm coordination a mainstream area of research. Traditional formation control relies on precise dynamic modeling, which limits its adaptability in dynamic environments. Deep Reinforcement Learning (DRL) and Multi-Agent Reinforcement Learning (MARL) possess model-free self-learning capabilities, offering new avenues for intelligent swarm formation. This paper systematically traces the evolution of formation control from traditional model-driven approaches and single-agent DRL to MARL. It compares the applicable scenarios and performance differences among mainstream MARL frameworks (MAPPO, MADDPG, DQN, DDPG, and Dec-MDP) and summarizes progress in integrating MARL with traditional methods such as leader-follower, consistency, and artificial potential fields. It analyzes current core challenges, such as slow convergence, difficulties in scaling, insufficient generalization from simulation to real-world environments, and limitations in on-board deployment, and outlines future trends from the perspectives of engineering implementation, heterogeneous swarms, robustness in complex environments, and algorithm lightweighting. This paper serves as a reference for the algorithmic design and engineering applications of drone swarm formation control.

KEYWORDS

Drone swarms; Formation control; Deep reinforcement learning; Multi-agent deep reinforcement learning; Cooperative control

1 Introduction

1.1 Research Background

Drone technology continues to evolve and improve, and is now widely used in military reconnaissance, emergency search and rescue, agricultural inspections, environmental monitoring, and smart urban management^[1]. Due to limitations in payload capacity, limited detection range, and single-function capabilities, individual drones cannot effectively meet the demands of large-scale detection, supply delivery, and swarm coordination in complex environments. Multi-drone self-organizing formations can effectively expand mission coverage and enhance system fault tolerance and operational reliability through distributed sensing, collective decision-making, and parallel operation mechanisms^[1]; Maintaining task execution within a cluster despite individual node failures, while compensating for performance shortcomings through collaborative task distribution, is a key area of research in the field of unmanned systems. Traditional formation control primarily relies on model-driven methods such as leader-follower control, consistency theory, and model prediction^[2-3]. Relying on the identification of precise dynamic parameters, this approach achieves good control performance only under steady-state conditions. When faced with operating conditions such as wind field disturbances, sudden obstacles, and time-varying communication states, modeling errors can easily lead to a decline in control performance, and the system's environmental adaptability and generalization capabilities may fail to meet the requirements of practical applications. Deep reinforcement learning does not rely on precise system modeling; instead, it iteratively optimizes control strategies through trial and error in interaction with the environment^[4,5]. It can effectively adapt to dynamic and uncertain scenarios, offering new approaches to UAV path planning, formation maintenance, and collaborative obstacle avoidance. Multi-agent reinforcement learning further overcomes the limitations of single-agent learning^[6], has become a core technological trend for intelligent control of swarm formations. However, existing research still faces challenges such as slow training convergence, the curse of dimensionality in large-scale clusters, and insufficient alignment between simulation and real-world systems. To address these issues, this paper systematically reviews technical advancements and identifies core bottlenecks to support future algorithmic improvements and engineering applications.

1.2 Significance of the Study

Cooperative formation flying of drone swarms is a core enabling technology for intelligent unmanned systems. Currently, research in this field remains fragmented, and a comprehensive theoretical framework or unified design framework has yet to be established. This paper traces the evolution of drone formation control technology and reviews the mainstream algorithmic frameworks. It defines the applicable scenarios and technical limitations of multi-agent reinforcement learning in formation tasks and refines the theoretical research framework. Additionally, it summarizes the

strengths and weaknesses of various algorithms and integration approaches, providing guidance for clustered collaborative decision-making, formation stability maintenance, and field deployment in engineering practice. The study holds both theoretical value and practical significance for the optimization of intelligent formation algorithms and the scenario-based implementation of unmanned swarms.

1.3 Research Objectives and Content

1.3.1 Research Objectives

Addressing the issues of high dependency in traditional formation control modeling, poor adaptability to complex environments, as well as low convergence efficiency, dimensional redundancy in large-scale swarms, and insufficient generalization in the transfer between virtual and real environments in multi-agent reinforcement learning, this paper systematically reviews research progress in deep reinforcement learning for unmanned aerial vehicle (UAV) formation control, compares the technical characteristics and application differences of mainstream multi-agent algorithms, and summarizes a fusion architecture integrating reinforcement learning with traditional formation control; It identifies current core technical bottlenecks, forecasts future research priorities and development directions, and provides insights for the optimization and engineering applications of swarm intelligence formation control algorithms.

1.3.2 Research Content

This study begins by systematically reviewing the evolutionary logic of traditional model-driven control, single-agent deep reinforcement learning, and multi-agent deep reinforcement learning. It compares and analyzes the technical characteristics, advantages, and limitations of each paradigm to elucidate the iterative value of intelligent learning methods. Subsequently, the study focuses on mainstream multi-agent reinforcement learning frameworks adapted for formation control, conducting a classification and comparative analysis to clarify the applicable scenarios and performance differences among various algorithms; Building on this foundation, the study summarizes the integrated architectures that combine deep reinforcement learning with classical formation control methods, analyzing the design philosophies, application advantages, and existing shortcomings of various integration schemes. Subsequently, it identifies the key technical bottlenecks in applying multi-agent reinforcement learning to UAV formation control across four dimensions: collaborative decision-making, training optimization, environmental generalization, and engineering transfer. Finally, by addressing practical challenges and technological development needs, the study outlines potential future research directions and trends.

The overall technical approach of this paper is shown in Figure 1.

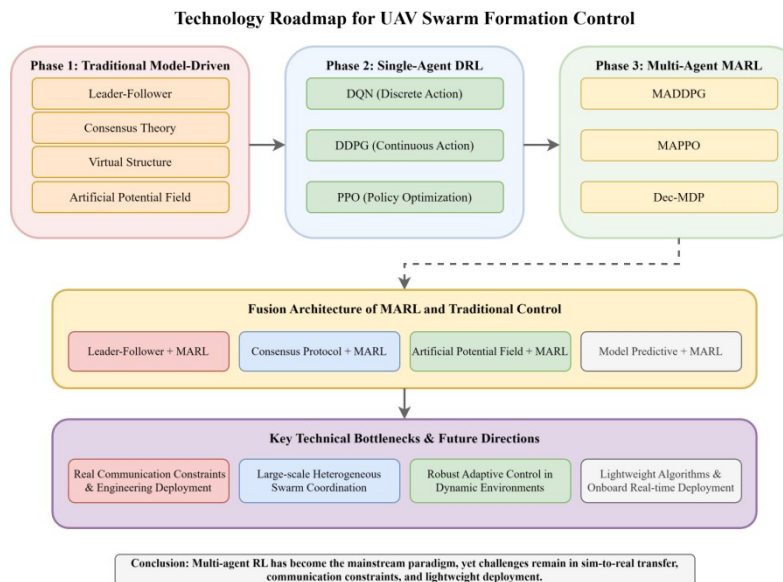


Figure 1 Technology roadmap of UAV swarm formation control with deep reinforcement learning

2 The Developmental Path of Formation Control Methods

In nature, the swarming behavior of organisms such as flocks of birds and schools of fish provides important insights for multi-robot coordination. The three principles of separation, alignment, and cohesion proposed by Reynolds laid the

foundation for swarm control from a behavioral perspective^[7]. Among early classical control methods, the virtual body method proposed by Lewis and Tan treats the entire formation as a rigid virtual body and uses the motion of this virtual rigid body to constrain the formation^[8]. The leader-follower framework developed by Desai et al. achieves formation control through a master-slave hierarchy; it features a simple structure and is easy to implement^[2]. Based on this, traditional model-driven formation control has gradually evolved into two major architectures: centralized and distributed. The centralized architecture relies on a central node for unified decision-making; while its structure is simple, it suffers from high communication overhead and limited robustness and scalability. The distributed architecture achieves global coordination through local information exchange between neighboring machines; with no central node, it is better suited for large-scale clusters. Consistency Theory (Olfati-Saber systematically established a theoretical framework for multi-agent swarm control from the perspectives of graph theory and dynamics^[3]), Model-driven methods, such as model predictive control, rely on precise identification of dynamic parameters and only achieve good control performance under steady-state conditions. When faced with operating conditions such as wind field disturbances, sudden obstacles, and time-varying communication states, modeling errors can easily lead to a decline in control performance, and the system's environmental adaptability and generalization capabilities often fail to meet practical application requirements. This has driven the introduction and development of data-driven reinforcement learning methods. Single-agent deep reinforcement learning, centered on a purely data-driven approach, is often combined with a leader-follower architecture, treating the follower as an independent agent. It relies on algorithms such as DQN, DDPG, and PPO to achieve trajectory tracking and formation maintenance^[4-5,9]. This paradigm does not require precise modeling and can adapt to nonlinear disturbances and dynamic environments, thereby reducing the costs associated with controller design and debugging. However, individual agents optimize only their own behavior and lack constraints for swarm coordination; as a result, system oscillations are likely to occur in complex scenarios, and scalability is insufficient in large-scale swarms. While it can only achieve local optima on a single machine, it lays the groundwork for research into multi-agent collaborative learning. Multi-agent deep reinforcement learning represents a paradigm shift from individual decision-making to collective optimization. It typically employs a centralized training and distributed execution architecture, with MADDPG and MAPPO serving as prime examples^[6,10]. Each drone acts as an independent agent; during the training phase, they integrate global information to optimize joint strategies, while during the deployment phase, they rely solely on local observations to generate control commands. Through a cooperative reward mechanism, they achieve multi-objective optimization for formation maintenance, dynamic obstacle avoidance, and cooperative navigation.

This method is free from the constraints of precise modeling and demonstrates strong adaptability to environmental disturbances, sensor noise, and communication-constrained scenarios; its distributed execution architecture is particularly well-suited for large-scale cluster deployments. However, challenges remain, including environmental non-stationarity, the high dimensionality of the joint action space, and the difficulty of allocating trust among multiple agents. Scaling up the cluster significantly increases training complexity and computational costs, while the stability of policy convergence and the generalization from simulation to real-world scenarios remain major technical bottlenecks. Compared to single-agent algorithms, multi-agent reinforcement learning can achieve global cooperative optimization across the cluster and has become the mainstream research paradigm for intelligent formation control.

3 A Study on the Application of Mainstream MARL Frameworks in Formation Control

3.1 MAPPO Series Frameworks

MAPPO employs a centralized-training-decentralized-execution architecture and is widely used for target allocation, collaborative navigation, and formation control. Its derived algorithms incorporate graph attention networks and hierarchical reward mechanisms to characterize the spatial topology of the swarm and establish a closed-loop perception-decision system, thereby improving convergence speed and disturbance tolerance^[11]. This framework is suitable for mixed static and dynamic targets in environments with multiple obstacles and is easily scalable to large-scale clusters; however, increasing the cluster size substantially increases the computational load, and the online task allocation module incurs additional computational costs^[12].

3.2 MADDPG Framework

MADDPG supports native continuous motion space, making it suitable for maneuvering flight, formation reconfiguration, and coordinated obstacle avoidance^[6]. The improved approach addresses the issues of sparse rewards and egoistic strategies by optimizing the reward mechanism and incorporating attention and LSTM networks, thereby enhancing control accuracy and environmental adaptability^[13-14]. This algorithm performs well on small- and medium-sized clusters, but the network structure is redundant, and there is still room for improvement in convergence speed^[15].

3.3 DQN and Ga-DQN Improved Frameworks

DQN is a classic algorithm for discrete action spaces, commonly used for small-scale path planning and simple obstacle avoidance; however, its exploration is highly stochastic, and it converges slowly in complex environments^[4]. Ga-DQN incorporates an attraction-repulsion mechanism to guide decision-making and employs a multi-dimensional reward function to balance flight efficiency and obstacle avoidance safety^[16]. This framework is lightweight and suitable for low-computing-power onboard platforms; however, due to discrete motion constraints, it is difficult to achieve precise attitude control, making it suitable only for simple, small-scale formations^[16].

3.4 DDPG Framework

DDPG is a deterministic policy algorithm for single agents that can be extended to swarm-based self-organizing formations. Based on the POMDP framework, it utilizes local perception and time-varying communication to achieve stable coordinated swarm motion through composite rewards^[17]. However, the algorithm has limited native capabilities for collaborative coupling modeling and is only suitable for small- to medium-sized swarm formation scenarios^[17].

3.5 Dec-MDP Distributed Decision-Making Framework

Dec-MDP models three-dimensional formation shaping as a distributed decision-making process, enabling global configurations to emerge based solely on the local states of neighboring aircraft, thereby avoiding the curse of dimensionality^[18]. Decentralized architectures are highly resilient to single points of failure and have low computational overhead, making them suitable for large-scale regular geometric formations; however, they lack the end-to-end optimization capabilities of deep learning, resulting in control performance that is inferior to mainstream MARL algorithms in complex scenarios such as dynamic target tracking and dense obstacle avoidance^[18].

Table 1 Comparison of mainstream MARL algorithms for UAV formation control.

Algorithm	ApplicableScenario	Advantage	Disadvantage	Convergence	Complexity
MAPPO	Mixed static/dynamic targets	High scalability, GNN-enhanced	High computation overhead	Medium	High
MADDPG	Continuous action space	Native continuous control	Redundant network, slow convergence	Medium	Medium
DQN/Ga-DQN	Discrete path planning	Lightweight, low computation	Poor fine-grained control	Slow	Low
DDPG	Small/medium swarm	Good for POMDP	Weak coordination modeling	Medium	Medium
Dec-MDP	Large-scale geometric formation	Decentralized, no dimension explosion	Weak in dynamic complex scenarios	Fast	Low

4 Current Status of Research on the Integration of MARL and Traditional Control Methods

Traditional methods such as lead-follow, consensus protocols, and artificial potential fields perform well under ideal conditions, but they are highly dependent on modeling and lack adaptability in dynamic scenarios^[2-3]. Although multi-agent reinforcement learning is model-free and self-learning, it suffers from shortcomings such as slow convergence, sparse rewards, and difficulty in maintaining steady states. The integration of these two approaches allows them to complement each other's strengths and has become an important area of research.

Current mainstream converged architectures fall into four categories:

(1) Lead-Follow Integrated with MARL: Using the traditional lead-follow topology as a framework, this approach employs MARL to optimize the followers' dynamic decision-making and trajectory tracking. Without the need for complex manual rules, it can adapt to maneuvering lead vehicles and complex obstacle environments, effectively enhancing the formation's ability to maintain formation and avoid obstacles;

(2) Consensus protocols combined with MARL: leveraging graph theory and Laplacian matrices to construct communication topology constraints, and using MARL to optimize consensus gains and collaboration weights, thereby enhancing the convergence of cluster collaboration under time-varying topologies and communication constraints;

(3) Integration of Artificial Potential Fields and MARL: This approach employs artificial potential fields to design basic attractive/repulsive constraints, utilizes MARL to overcome the local optima issue inherent in traditional potential fields, and simultaneously optimizes formation spacing and obstacle avoidance safety margins^[15];

(4) Model Prediction and MARL Hierarchical Integration: The upper layer uses model prediction to constrain path planning, while the lower layer employs MARL for real-time online optimization of control variables, thereby balancing

planning rationality with dynamic disturbance rejection capabilities.

While existing fusion research has improved the robustness and adaptability of formation systems, it still has shortcomings: the fusion mechanisms lack a general theoretical framework and are mostly tailored to specific scenarios; the algorithms exhibit high computational complexity after fusion, making them difficult to adapt to onboard edge computing capabilities; and there are significant discrepancies between simulation and actual flight, indicating that model generalization and real-world transferability require further improvement^[13].

Table 2

Fusion Architecture	Traditional Method	MARL Role	Advantage	Limitation
Leader-follower + MARL	Leader-follower	Optimize follower tracking	No manual rules, adaptive to dynamic obstacles	Leader reliance remains
Consensus + MARL	Consensus protocol	Optimize consensus gains	Time-varying topology robustness	Complex graph design
Artificial potential field + MARL	APF	Avoid local optima	Safe separation + collision avoidance	Reward tuning difficult
MPC + MARL	Model predictive control	Online control optimization	Combines planning + disturbance rejection	High complexity, hard to deploy

5 Future Research Directions

5.1 Aligning with Real-World Communication Constraints and Engineering Realization

Most existing studies assume ideal communication conditions and simplified two-dimensional flight scenarios, which do not align with real-world electromagnetic interference, communication delays, and packet loss, nor with three-dimensional flight conditions. Furthermore, the majority of algorithms remain at the simulation verification stage. Future work should move away from these idealized assumptions, develop cooperative control models under time-varying delays and dynamic topologies, advance step-by-step verification through simulation, hardware-in-the-loop testing, and actual flight tests, refine methods for transferring results from virtual to real-world environments, and accelerate the engineering implementation of these algorithms.

5.2 Coordinated Optimization of Large-Scale Heterogeneous Clusters

Current research primarily focuses on small-scale homogeneous UAV formations. In large-scale scenarios, dimensional redundancy and convergence difficulties often arise, and existing approaches lack adaptability to heterogeneous UAV types such as fixed-wing and rotorcraft. Future work could employ strategies such as cluster clustering, task stratification, and topological decoupling to achieve algorithmic dimensionality reduction, thereby alleviating the “curse of dimensionality” and training convergence challenges associated with large-scale UAV swarms.[14] Design a universal formation strategy that is cross-platform and scalable, incorporating mechanisms for automatic formation reconfiguration in the event of single-unit failures or link failures, thereby enhancing the cluster’s error tolerance and resistance to interference.

5.3 Adaptive Control in Complex Dynamic Environments

Existing algorithms do not adequately account for wind disturbances, airflow turbulence, and sudden dynamic obstacles, and their adaptability to complex outdoor and urban environments is limited. By integrating disturbance observations with a reinforcement learning framework, a multi-objective reward mechanism can be established that balances formation maintenance, obstacle avoidance safety, and energy consumption optimization, thereby enhancing the swarm’s ability to make autonomous decisions and operate stably in dynamic environments.

5.4 Algorithm Lightweighting and On-Board Real-Time Deployment

Multi-agent algorithms are complex and computationally intensive, making them ill-suited for the low computational power, low power consumption, and high real-time performance requirements of unmanned aerial vehicles (UAVs). In the future, lightweight techniques such as network pruning, model quantization, and incremental learning could be adopted to reduce computational load and enable real-time deployment at the edge on board the UAV.

6 Conclusion

This review indicates that drone formation control has evolved from traditional model-driven methods to data-driven deep reinforcement learning, and has further developed into the mainstream paradigm of multi-agent cooperative optimization. Through a comparative analysis of algorithms such as MAPPO, MADDPG, DQN, DDPG, and Dec-MDP, it was found that existing methods generally suffer from common issues, including idealized communication assumptions, a disconnect between simulation and real-world generalization, poor adaptability to large-scale heterogeneous fleets, high algorithmic complexity, and difficulties in real-time on-board deployment.

Future research should be grounded in real-world communication and time-varying topological characteristics, focusing on four key directions: large-scale heterogeneous formation optimization, adaptive interference resistance in complex environments, lightweight algorithm design, and simulation-to-real-world transfer validation. The analysis and outlook presented in this paper provide theoretical support and directional guidance for future research on intelligent formation control of UAV swarms.

References

- [1] Zong, Q., Wang, D., Shao, S., et al. (2017) Multi-UAV cooperative formation flight control: Current status and development. *J. Harbin Inst. Technol.*, 49(3): 1–14.
- [2] Desai, J.P., Ostrowski, J.P., Kumar, V. (2001) Modeling and control of formations of nonholonomic mobile robots. *IEEE Trans. Robot. Autom.*, 17(6): 905–908.
- [3] Olfati-Saber, R. (2006) Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Trans. Autom. Control*, 51(3): 401–420.
- [4] Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.
- [5] Lillicrap, T.P., Hunt, J.J., Pritzel, A., et al. (2020) Continuous control with deep reinforcement learning. U.S. Patent 10,776,692.
- [6] Lowe, R., Wu, Y.I., Tamar, A., et al. (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. In: *Advances in Neural Information Processing Systems*. Long Beach, CA. pp. 6379–6390.
- [7] Reynolds, C.W. (1987) Flocks, herds and schools: A distributed behavioral model. In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. New York. pp. 25–34.
- [8] Lewis, M.A., Tan, K.H. (1997) High precision formation control of mobile robots using virtual structures. *Auton. Robots*, 4(4): 387–403.
- [9] Schulman, J., Wolski, F., Dhariwal, P., et al. (2017) Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- [10] Yu, C., Velu, A., Vinitsky, E., et al. (2022) The surprising effectiveness of PPO in cooperative multi-agent games. In: *Advances in Neural Information Processing Systems*, Vol. 35. New Orleans. pp. 24611–24624.
- [11] Sun, S., Jin, B., Yu, H., et al. (2025) Graph neural network enhanced multi-agent reinforcement learning for UAV formation control. In: *2025 3rd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV)*. Singapore. pp. 1–6.
- [12] Sheng, Y., Xie, X., Liu, H., et al. (2026) Dynamic target assignment and cooperative decision-making for UAV swarms based on multi-agent reinforcement learning. *IEEE Internet Things J.*, 13(8): 12345–12360.
- [13] Zhang, Y., Wu, Z., Ma, Y., et al. (2022) Research on autonomous formation of multi-UAV based on MADDPG algorithm. In: *2022 IEEE 17th International Conference on Control & Automation (ICCA)*. Naples. pp. 249–254.
- [14] Wu, J., Li, D., Yu, Y., et al. (2024) An attention mechanism and adaptive accuracy triple-dependent MADDPG formation control method for hybrid UAVs. *IEEE Trans. Intell. Transp. Syst.*, 25(9): 11858–11870.
- [15] Xing, X., Zhou, Z., Li, Y., et al. (2024) Multi-UAV adaptive cooperative formation trajectory planning based on an improved MATD3 algorithm of deep reinforcement learning. *IEEE Trans. Veh. Technol.*, 73(9): 12484–12499.
- [16] Xu, Z., Wang, Q., Kong, F., et al. (2022) Ga-DQN: A gravity-aware DQN based UAV path planning algorithm. In: *2022 IEEE International Conference on Unmanned Systems (ICUS)*. Guangzhou. pp. 1215–1220.
- [17] Salimi, M., Pasquier, P. (2021) Deep reinforcement learning for flocking control of UAVs in complex environments. In: *2021 6th International Conference on Robotics and Automation Engineering (ICRAE)*. Singapore. pp. 344–352.
- [18] Azam, M.A., Mittelman, H.D., Ragi, S. (2021) UAV formation shape control via decentralized Markov decision processes. *Algorithms*, 14(3): 91.