A Multi-UAV Cooperative Navigation Method Based on Policy Decomposition Structure

Li Tan[a,*], Haixia Zhao[a], Jiaqin Chai[a], Xinshi Zhang[a], Grigore Stamatescu[b]

[a] *School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100080, China.*
[b] *Department of Automation and Industrial Informatics, National University of Science and Technology POLITEHNICA of Bucharest, 060042 Bucharest, Romania.*

Cooperative navigation of multiple unmanned aerial vehicles (UAVs) in disaster search-and-rescue scenarios is challenging due to dense obstacles, partial observability, and strong inter-agent coupling, which often result in path conflicts, collision risks, and limited policy generalization. To address these challenges, this paper proposes a Multi-Agent Deep Deterministic Policy Gradient framework with a Graph-Attention-based Staged Actor (GS-MADDPG). Under a centralized training and decentralized execution paradigm, GNNs are employed to model local interaction relationships among UAVs, enabling effective information aggregation and cooperative decision-making under partial observability. Furthermore, the Actor network is decomposed into perception, goal-guidance, and feature fusion subnetworks, allowing hierarchical decoupling and coordinated integration of local obstacle avoidance behaviors and global navigation objectives. The simulation results conducted in the complex three-dimensional urban environment show that compared with traditional methods, GS-MADDPG has improved the navigation success rate, robustness and generalization performance. When the density of environmental obstacles is 50% and the number of unmanned aerial vehicles increases from 2 to 10, the navigation success rate of GS-MADDPG is approximately 40% higher than that of the benchmark algorithm; even in cases with higher obstacle density, GS-MADDPG still achieves a relatively high success rate. This verifies its effectiveness in multi-UAV cooperative navigation for search and rescue tasks.

Keywords: Multi-UAV Systems; Deep Reinforcement Learning; Collaborative Navigation; Path Planning.

1. Introduction

In recent years, frequent disasters have placed higher demands on the intelligence of search and rescue systems [1]. Multi-Unmanned aerial vehicle(UAV) systems, with their high maneuverability, rapid deployment, and collaborative advantages, have become essential tools for search and rescue operations [2]. Coordinated navigation of multiple UAVs to target areas can significantly enhance search and rescue efficiency and success rates. However, obstacles in complex environments are irregularly distributed and densely packed, requiring collision avoidance in path planning. Without a global perspective and with limited communication, coordinating the behavior of multiple UAVs to resolve path conflicts and navigate obstacles poses significant challenges [3, 4].

To address these issues, existing research employs heuristic search algorithms, random sampling methods, artificial potential field approaches, and rule-based distributed control strategies for path planning. While these methods perform reasonably well in structurally simple environments, they struggle to effectively handle environmental uncertainties when confronted with complex obstacles, multi-UAV dynamic collaboration, and limited local perception. Furthermore, they lack good scalability and adaptability, limiting the robustness and intelligence of the system [4–6].

To enhance the generalization and adaptability of navigation strategies, Multi-agent reinforcement learning (MARL) has been extensively applied in multi-UAV cooperative control in recent years [7, 8]. Among these, Multi-agent deep deterministic policy gradient (MADDPG) stands as a representative algorithm[9]. It supports integrating multi-agent information during the training phase, enabling centralized training with distributed execution, and possesses strong collaborative modeling capabilities. However, in scenarios with dense obstacles and complex tasks, relying solely on MADDPG's state representation may still result in low learning efficiency and inadequate obstacle avoidance performance.

During navigation, a drone's perception of its surrounding environment and the status of its teammates is

*Corresponding author.E-mail: tanli@th.btbu.edu.cn.

critical. To enhance collaborative obstacle avoidance capabilities, this paper introduces graph neural networks (GNNs) to model local relationships among drones, facilitating information sharing and collaborative decision-making. This approach improves the robustness and efficiency of navigation strategies [10, 11]. The main contributions of this paper include:

(1) A staged policy architecture based on MADDPG is proposed to achieve collaborative decision-making and information fusion among multiple UAVs under local communication constraints.

(2) By decomposing the Actor into a GNN perception subnetwork, a target subnetwork, and a fusion network, this method achieves hierarchical decoupling between local obstacle avoidance and global navigation, along with dynamic feature fusion.

(3) Experiments demonstrate that the proposed method exhibits outstanding robustness and generalization capabilities in complex, partially observed environments.

The rest of this paper is organized as follows. In Section 2, related work is reviewed. The proposed model is introduced in Section 3. The performance evaluation and results are presented in Section 4, and finally, this paper is concluded in Section 5.

2. Related Work

Multi-UAV cooperative navigation aims to generate safe, continuous, and coordinated flight trajectories in complex three-dimensional environments. The problem is inherently challenging due to high-dimensional continuous state spaces, partial observability of the environment, and strongly coupled interactions among multiple UAVs[12].

Model-based path planning methods are widely employed for collision-free trajectory generation among multiple UAVs, encompassing paradigms such as search, sampling, potential fields, and optimization control. Li et al. [13] modified A*'s bidirectional search by alternately expanding forward and backward nodes while sharing information to accelerate global path planning; Tao et al. [14] introduced an artificial potential field (APF) into RRT*'s variable probability sampling process to generate new nodes biased toward target regions, accelerating autonomous vehicle path planning; Fan et al. [15] concurrently embedded bidirectional search and target-biased sampling into APF-RRT*, enabling UAV trajectory planning to converge faster under obstacle potential field guidance. At the local obstacle avoidance level, artificial potential field methods construct real-time avoidance behavior through gravity-repulsion mechanisms. Liu et al. [16] embedded artificial potential fields into multi-stage optimization to sequentially guide multiple UAVs through three-dimensional urban environments. Zhang et al. [17] combined an improved artificial potential field with optimal consistency control to coordinate obstacle avoidance among multiple UAVs. Model predictive control methods generate feasible trajectories by explicitly modeling dynamics and safety constraints. Du et al. [18] employed model predictive control with online rolling optimization of obstacle avoidance constraints, enabling real-time safe obstacle circumvention during formation flight. Chu et al. [19] embedded PID feedback into model predictive control to correct reference trajectories online, achieving real-time obstacle avoidance and tracking. However, these approaches generally rely on global environmental information or precise model descriptions, leading to rapidly increasing computational complexity in dynamically uncertain environments or large-scale UAV scenarios [20].

Reinforcement learning, which learns continuous control policies through interactive trial-and-error, has been applied to unmanned aerial vehicle navigation problems. Single-agent continuous control methods demonstrate good adaptability in isolated scenarios; however, in multi-UAV systems, independent policies struggle to explicitly model coupling relationships between agents. Multi-agent reinforcement learning achieves effective characterization of cooperative behaviors through joint modeling or centralized training mechanisms [21]. Such cooperative behaviors often involve not only spatial coordination but also temporal consistency among agents, such as simultaneous arrival or synchronized task execution, which further increases the control complexity [22].

In continuous action spaces, MADDPG has been widely applied to multi-UAV cooperative tasks. Liu et al. [23] employed a dual-timescale hierarchical MADDPG to enable simultaneous search and real-time obstacle avoidance for multiple UAVs; Cao and Chen [24] utilized enhanced deep reinforcement learning to coordinate observation sharing among multiple UAVs, synchronously learning obstacle avoidance and navigation strategies under a unified reward. Li et al. [25] guided UAVs to complete cooperative encirclement through staged rewards. Zhang et al. [26] achieved formation stability under changing communication topologies by utilizing shared parameters. Existing MADDPG methods predominantly employ a single Actor network to directly process obstacle avoidance perception and target information, which is prone to feature coupling and training instability in complex 3D environments [27].

To address the aforementioned challenges, this paper proposes a phased Actor network architecture. It incorporates a path-avoidance subnetwork based on graph attention, a goal-oriented subnetwork, and a state enhancement fusion mechanism to enhance the collaborative navigation performance of multiple UAVs in complex 3D environments.

3. System Model

In this paper, the cooperative obstacle avoidance and target navigation problem for multi-UAV systems is formulated as a partially observable multi-agent Markov decision process (POMDP), represented by the tuple $(\mathcal{O}, \mathcal{A}, T, R, \gamma)$, where $\mathcal{O}$ denotes the local observation space, $\mathcal{A}$ is the joint action space, T is the state transition function, R is the global

reward function, and $\gamma \in [0,1]$ is the discount factor. Due to limited onboard sensing and constrained local communication, each UAV can only access partial environmental information, and the system therefore operates under partial observability. The global state S_t consists of the positions, linear velocities, angular velocities of all UAVs, as well as the spatial distribution of obstacles in the environment. Accordingly, at time t, UAV i receives a local observation o_t^i defined as

$$o_t^i = \left[v_t^i,\ \omega_t^i,\ g_{dir}^i,\ d_{goal}^i,\ ray_t^i,\ neigh_t^i\right], \tag{1}$$

where $v_t^i \in \mathbb{R}^3$ and $\omega_t^i \in \mathbb{R}^3$ denote the linear and angular velocities of UAV i, jointly describing its kinematic state. The goal-related information includes the unit direction vector $g_{dir}^i \in \mathbb{R}^3$ pointing toward the target and the normalized distance d_{goal}^i. The obstacle perception vector $ray_t^i \in \mathbb{R}^n$ is obtained from a 3D radar along predefined directions, while $neigh_t^i \in \mathbb{R}^{7k}$ encodes the relative positions and velocities of the k nearest neighboring UAVs in the body coordinate frame, together with an existence indicator. The total observation dimension is therefore $3+3+3+1+n+7k$. To facilitate structured policy learning, the observation vector is partitioned into an obstacle-related state s_o and a goal-related state s_g, where $s_g \in \mathbb{R}^4$ contains the goal direction and distance, and s_o includes the remaining obstacle and neighbor interaction features. The observation structure is illustrated in Fig. 1.

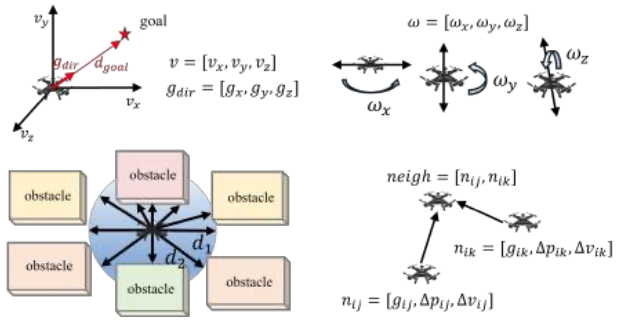


Fig. 1. Structure of the local observation vector for UAV i, including motion state, goal-related information, obstacle perception features, and neighboring UAV interaction features.

The action space is defined as $\mathcal{A} = [-1,1]^3$, representing normalized velocity commands along the three body-frame axes of each UAV. The output of the Actor network a_t^i is constrained by a hyperbolic tangent function and linearly scaled to generate the target velocity

$$\mathbf{v}_{tar}^i = action_scale \cdot a_t^i, \tag{2}$$

where $action_scale$ is a predefined constant. At the control level, a force-based velocity tracking mechanism is employed, allowing the UAV's actual velocity to smoothly track the commanded target velocity. The low-level dynamics of the multi-UAV system are simulated using the PyBullet physics engine and can be expressed as

$$S_{t+1} \sim P(S_t, A_t), \tag{3}$$

where the transition process is nonlinear and stochastic due to aerodynamic damping, collision handling, and elastic contact recovery.

The reward function is designed to jointly encourage navigation efficiency, collision avoidance, and cooperative behavior. For UAV i, the instantaneous reward is defined as

$$r_t^i = r_{time} + r_{goal} + r_{collision} + r_{obstacle} + r_{close} + r_{far} + r_{success}, \tag{4}$$

where r_{time} is a constant negative penalty applied at each time step to discourage prolonged trajectories. The goal progress reward is defined as

$$r_{goal} = d_{t-1}^i - d_t^i, \tag{5}$$

with d_t^i denoting the distance between UAV i and the target. Collision events incur a penalty

$$r_{collision} = I_c^i, \tag{6}$$

where $I_c^i = 1$ indicates a collision with an obstacle or another UAV. To enforce obstacle safety margins, an obstacle proximity penalty is introduced as

$$r_{obstacle} = \max(0, d_{safe} - d_{min}), \tag{7}$$

where d_{min} is the minimum radar-detected distance. Inter-UAV safety is further encouraged through

$$r_{close} = \sum_{j \in \mathcal{N}_i} \max(0, d_{min_UAV} - d_{ij}), \tag{8}$$

where d_{ij} denotes the distance between UAV i and its neighbor j. To prevent excessive dispersion, a neighborhood deviation penalty is defined as

$$r_{far} = \max\left(0, \left\|p^i - p_{centroid}^i\right\| - d_{max_UAV}\right), \tag{9}$$

where $p_{centroid}^i$ is the target-weighted centroid of the local neighborhood. When all UAVs reach the target region, a global task completion reward

$$r_{success} = R_{succ} \tag{10}$$

is assigned.

Based on the above formulation, a Multi-Agent Deep Deterministic Policy Gradient framework with a Graph-Attention-based Staged Actor (GS-MADDPG) is proposed under a centralized training and distributed execution paradigm. As shown in Fig. 2, the architecture consists of a global Critic network and a shared Actor network with a stage-wise structure.

The Critic network takes the joint observations and joint actions of all agents as input to evaluate the global Q-value, while the Actor performs decentralized decision-making with shared parameters. The staged Actor comprises an obstacle avoidance subnetwork based on a Graph Attention Network (GAT) and a target-oriented subnetwork implemented as a Multi-Layer Perceptron. In the obstacle avoidance subnetwork, each UAV is modeled

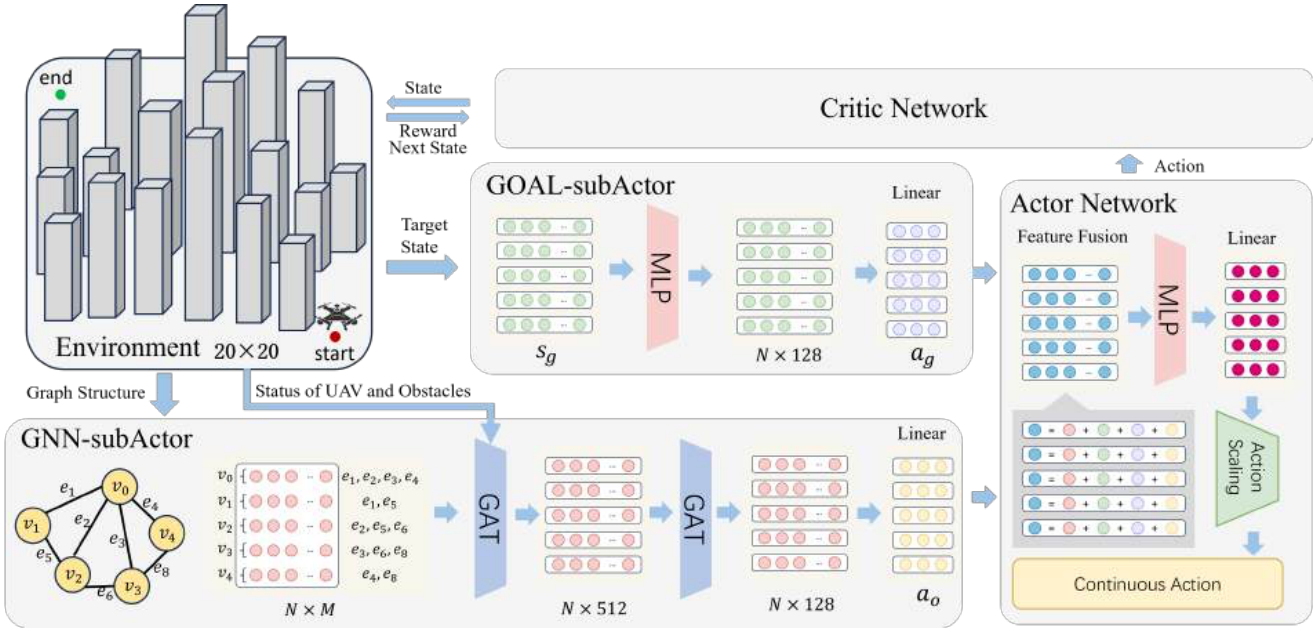


Fig. 2. GS-MADDPG structural framework.

as a graph node with features s_o^i, and edges are constructed based on nearest-neighbor relationships. A multi-head graph attention mechanism is applied to aggregate neighboring features, followed by head averaging to produce a latent representation, which is mapped to an obstacle-avoidance action proposal $a_o^i \in \mathbb{R}^3$. In parallel, the target subnetwork processes s_g^i and outputs a target-oriented action proposal $a_g^i \in \mathbb{R}^3$. The outputs of both subnetworks, together with the original observation features, are concatenated as

$$[s_o^i,\, s_g^i,\, a_o^i,\, a_g^i],$$

and fed into a decision aggregation network. After two fully connected layers with layer normalization and ReLU activation, the final action

$$a_{\text{final}}^i \in \mathbb{R}^3$$

is produced and constrained by a tanh function.

During training, all UAVs share the same Actor and Critic parameters. The Critic is optimized by minimizing the temporal-difference loss

$$L_c = \left(Q(s_t, a_t) - y_t\right)^2, \tag{11}$$

where

$$y_t = r_t + \gamma Q_{target}(s_{t+1}, a_{t+1}). \tag{12}$$

The Actor is optimized by maximizing the expected Q-value estimated by the Critic,

$$L_a = -\mathbb{E}_{s_t \sim \mathcal{D}}\left[Q(s_t, \pi(s_t))\right], \tag{13}$$

where $\mathcal{D}$ denotes the replay buffer. A soft target update mechanism $\theta_{\text{target}} \leftarrow \tau\theta + (1-\tau)\theta_{\text{target}}$ is adopted to stabilize training. Experience replay is used to reduce sample correlation, and a cosine annealing strategy is applied to schedule the learning rate. Owing to the staged Actor design and graph-attention-based interaction modeling, the proposed method achieves robust cooperative navigation and effective obstacle avoidance under partial observability and communication constraints.

4. Experiments

To evaluate the robustness and cooperative navigation performance of the proposed GS-MADDPG in complex urban environments, a structured urban scenario is constructed on a PyBullet-based 3D simulation platform. The simulation considers cooperative obstacle-avoidance navigation tasks for a UAV swarm under partial observability. Each UAV is modeled as a dynamically constrained agent, with its state comprising velocity, attitude, and local perception, while actions are executed through control forces in three-dimensional space. The training process follows a centralized training and decentralized execution paradigm, employing a global critic to assess joint states and actions and shared actor networks for distributed decision-making.

The experimental scenario is discretized into a 20×20 urban grid, with each grid cell representing a 10-meter-square physical area. The starting zone and target zone are positioned diagonally opposite each other. Buildings are randomly generated at a default density of 0.5, with dimensions occupying approximately 50%–80% of a grid cell.

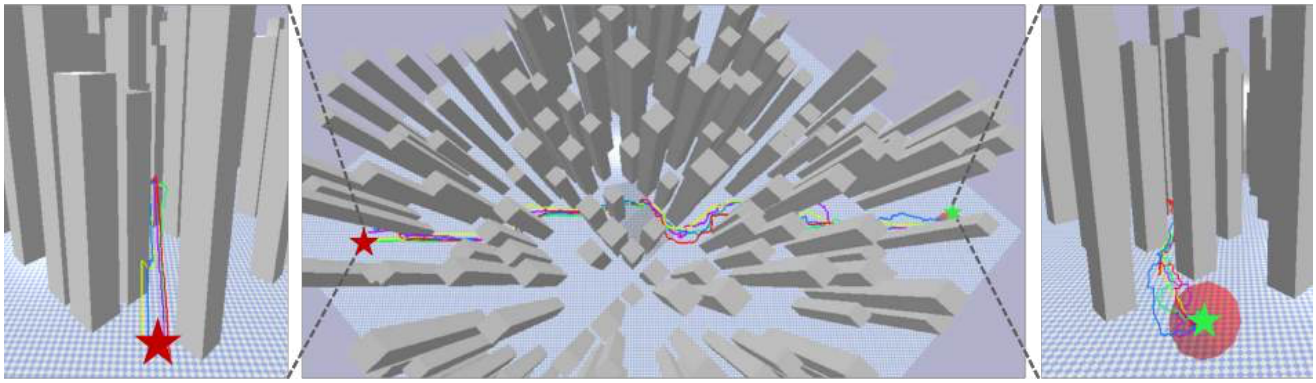


Fig. 3. Trajectory visualization of five UAVs in a simulation environment with 50% obstacle density. Different colored paths denote individual UAV trajectories from the common start point to the target region, while transparent red circles indicate the target arrival zones.

Each drone is equipped with 16 equidistantly distributed laser radar beams, each with a maximum detection range of 20 meters, used to perceive surrounding obstacles and the relative distribution of neighboring nodes. The maximum flight altitude is set to 25 meters, with a maximum of 900 steps per round. Key experimental parameters are shown in Table 1.

During training, all UAVs share a common set of policy parameters, which are optimized using experience replay and gradient-based updates. The replay buffer size is set to 5×10^4, and a cosine annealing schedule is applied to the learning rate to improve training stability and convergence. After training, the learned policies are evaluated over 10 independent simulation episodes.

Table 1. Parameter settings.

Hyperparameters	Values
Max episode	500
Memory size	512
Soft update factor τ	0.005
Discount factor γ	0.99
Learning rate	0.0003

The performance of the proposed method is evaluated using three metrics: success rate, average completion time, and path efficiency. The success rate is defined as the proportion of UAVs that successfully reach the target region within the maximum allowed number of steps. The average completion time measures the mean number of time steps required for successful UAVs to complete the task. The path efficiency is computed as the ratio between the straight-line distance from the starting position to the target and the actual trajectory length traveled by each UAV.

Fig. 3 presents the simulation results under the conditions of 5 UAVs and 50% obstacle density. All UAVs start from the red origin point and gradually navigate toward the target area while avoiding surrounding obstacles under the proposed collaborative strategy.

During multi-UAV navigation, it is generally difficult for all UAVs to arrive at the target simultaneously. Therefore, a target region is defined around each target point using transparent red circular areas. Once a UAV enters this region, it is regarded as having reached the target, which also helps avoid collisions caused by multiple UAVs converging to the same point. The trial is considered successful only when all five UAVs enter their corresponding target regions.

The results demonstrate that the proposed method can effectively balance local obstacle avoidance and global goal-reaching behaviors in complex environments.

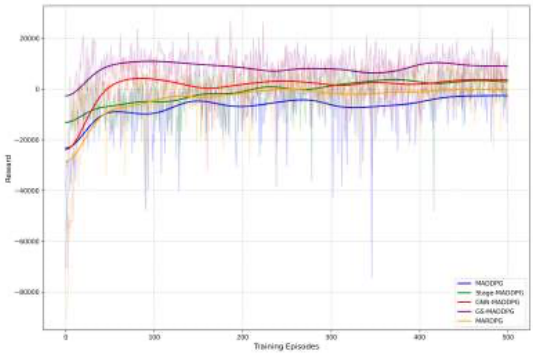


Fig. 4. Training curves of average rewards for four network architectures: MADDPG, MADDPG-GNN, MADDPG-STAGE, and the proposed GS-MADDPG.

To evaluate the impact of different structural designs on policy performance, four network architectures are compared: the standard MADDPG, MADDPG with a graph neural Actor (MADDPG-GNN), MADDPG with a staged

Actor structure (MADDPG-STAGE), and the proposed GS-MADDPG. Fig. 4 shows the training curves of average rewards for all compared models.

The results indicate that the proposed method achieves consistently higher reward values than the standard MADDPG from the early stages of training and maintains this advantage throughout the learning process, demonstrating faster convergence and superior overall performance.

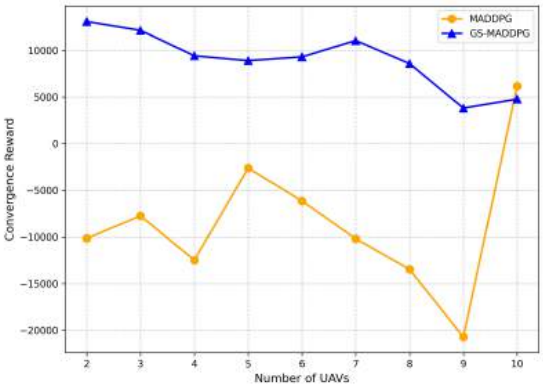


Fig. 5. Comparison of converged average reward values between MADDPG and the proposed method under different fleet sizes (2–10 UAVs).

To evaluate convergence behavior under different fleet sizes, the converged reward values and convergence times of MADDPG and the proposed method are compared for 2 to 10 UAVs. Fig. 5 presents the converged rewards at different scales, while Fig. 6 shows the corresponding convergence times.

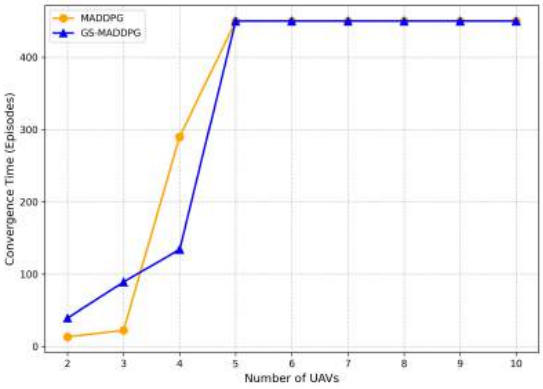


Fig. 6. Comparison of convergence times between MADDPG and the proposed method under different fleet sizes (2–10 UAVs).

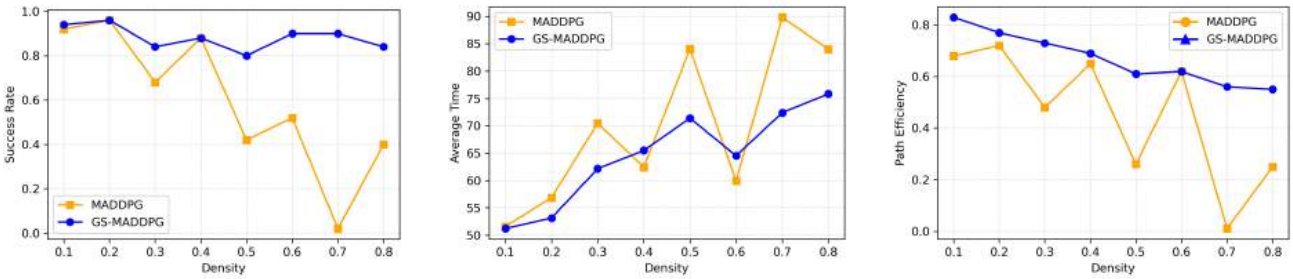
The results demonstrate that our method achieves significantly higher converged reward values than MADDPG across all scale conditions, indicating superior adaptability and stability in complex multi-agent scenarios. Regarding convergence time, the difference between the two methods is negligible, suggesting that our approach enhances performance without substantially increasing training overhead.

In order to further test the robustness of the method proposed in this paper under different levels of environmental complexity, this paper compared the model success rate, average time, and path efficiency under different obstacle densities. As shown in Table 2, the success rate of the method proposed in this paper did not show a significant difference from the baseline algorithm in the low-density environment. However, when the environmental density increased, our algorithm outperformed the MADDPG algorithm.

Table 2. Experimental results of different densities.

Density	Method	Success rate (%)	Average time (s)	Path efficiency
0.1	MADDPG	92	51.59	0.68
	GS-MADDPG	94	51.21	0.83
0.2	MADDPG	96	56.88	0.72
	GS-MADDPG	96	53.14	0.77
0.3	MADDPG	68	70.45	0.48
	GS-MADDPG	84	62.16	0.73
0.4	MADDPG	88	62.45	0.65
	GS-MADDPG	88	65.49	0.69
0.5	MADDPG	42	84.04	0.26
	GS-MADDPG	80	71.39	0.61
0.6	MADDPG	52	59.93	0.62
	GS-MADDPG	90	64.52	0.62
0.7	MADDPG	2	89.83	0.01
	GS-MADDPG	90	72.39	0.56
0.8	MADDPG	40	83.98	0.25
	GS-MADDPG	84	75.84	0.55

According to Table 2, we can draw the graph Fig. 7 showing the changes in experimental results when the density of obstacles changes. It indicates that the method proposed in this paper can maintain relatively stable navigation performance under different obstacle densities. Although in some cases the performance of the benchmark algorithm is comparable or even better than that of the method in this chapter, its volatility is relatively high, suggesting that its navigation process is random and unreliable. The accuracy rate of this method is much higher



(a)Success rate. (b)Average time. (c)Path efficiency.

Fig. 7. Performance comparison under different densities.

than that of the benchmark model, and it still has high path efficiency in complex environments. This shows that this model has a strong resistance to complex environmental interference in the collaborative multi-UAV navigation process.

Table 3. Experimental results of different UAV numbers.

Number	Method	Success rate (%)	Average time (s)	Path efficiency
2	MADDPG	93	75.84	0.82
	GS-MADDPG	96	77.54	0.86
3	MADDPG	89	74.40	0.50
	GS-MADDPG	92	73.53	0.80
4	MADDPG	71	78.11	0.48
	GS-MADDPG	84	71.80	0.70
5	MADDPG	42	84.04	0.26
	GS-MADDPG	80	71.39	0.61
6	MADDPG	58	88.39	0.14
	GS-MADDPG	77	67.55	0.65
7	MADDPG	64	86.25	0.51
	GS-MADDPG	74	78.22	0.62
8	MADDPG	21	84.90	0.26
	GS-MADDPG	89	80.57	0.57
9	MADDPG	53	81.48	0.44
	GS-MADDPG	84	79.12	0.52
10	MADDPG	33	88.65	0.22
	GS-MADDPG	75	80.88	0.42

To examine the system's scalability in terms of the number of UAV, additional experiments were conducted with the number of unmanned aircraft ranging from 2 to 10. The results are presented in Table 3. Overall, our method outperforms the baseline model in all indicators, with particularly outstanding performance in terms of success rate and path efficiency.

To observe the changes, we represented the data in the above table graphically, and the results are shown in Fig. 8.

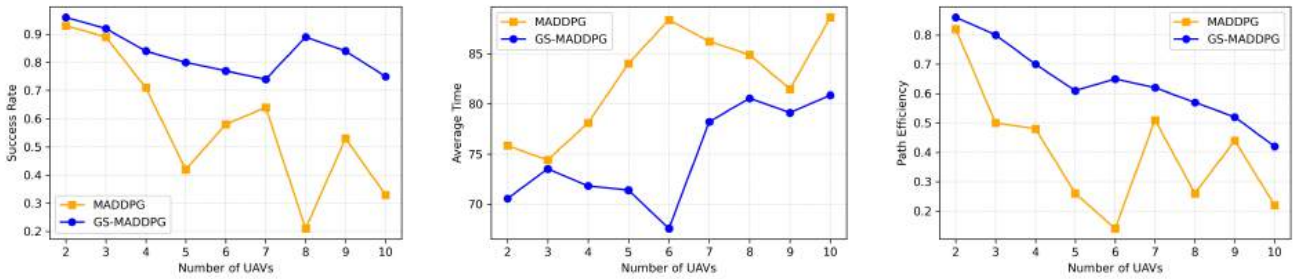
From the experimental results, it can be observed that the average task completion time exhibits a non-monotonic relationship with the number of UAVs, achieving its optimal value when six UAVs are deployed. This behavior suggests that insufficient UAVs limit environmental perception, while an excessive number increases coordination and avoidance complexity, both of which negatively affect navigation efficiency. A moderate number of UAVs therefore provides a more effective balance between perception and cooperation.

Ablation experiments were conducted to examine the contributions of the GNN module and the staged (STAGE) structure. As shown in Table 4, introducing the GNN module leads to a noticeable increase in success rate, while its influence on path efficiency remains relatively limited. This improvement can be attributed to the GNN's ability to encode inter-UAV topological relationships and aggregate neighboring information, which enhances cooperative perception and collision avoidance.

Table 4. Ablation experiment.

Method	Success rate	Average time	Path efficiency
MADDPG	0.42	84.04	0.26
+GNN	0.67	78.63	0.31
+STAGE	0.52	79.91	0.58
GS-MADDPG	0.80	71.39	0.61

In contrast, incorporating the STAGE structure results in a more pronounced improvement in path efficiency, with a comparatively smaller effect on success rate. This suggests that the target-oriented subnetwork primarily promotes more efficient target-approach behaviors, enabling UAVs to generate shorter and more rational paths under



(a)Success rate. (b)Average time. (c)Path efficiency.

Fig. 8. Performance comparison under different numbers of UAVs.

safe conditions. The combination of these two structures allows the model to simultaneously improve mission completion reliability and navigation efficiency.

Overall, the GNN module primarily enhances cooperative perception and safety awareness, while the staged architecture refines global decision-making and path-planning efficiency. Their integration enables the proposed method to achieve simultaneous improvements in effectiveness and efficiency in complex multi-UAV navigation scenarios.

5. Conclusion

This paper studies cooperative navigation and obstacle avoidance for multi-UAV systems in complex environments with partial observability. A stage-wise multi-UAV reinforcement learning framework integrating graph neural networks is presented. By decomposing the actor network into functionally specialized modules, the proposed method decouples local obstacle avoidance from global goal-directed navigation, while graph attention mechanisms facilitate information aggregation and cooperative decision-making among neighboring UAVs. Simulation results under varying obstacle densities and different numbers of UAVs show improvements in navigation success rate, path efficiency, and robustness compared with baseline methods, demonstrating the effectiveness and scalability of the proposed approach in complex search-and-rescue scenarios.

Acknowledgments

This work is supported by Natural Science Foundation of China(62472010). Chongging Natural Science Foundation(CSTB2024NSCQ-MSX0687). And Ministry of Education Foundation on Humanities and Social Sciences(23YJAZH129).

References

[1] Yang L, Zhang X, Li Z, et al. A LODBO algorithm for multi-UAV search and rescue path planning in disaster areas[J]. Chinese Journal of Aeronautics, 2025, 38(2): 103301.

[2] Tan L, Zhao H. A multi-UAV rapid post-disaster search and rescue method based on deep reinforcement learning[J]. Complex & Intelligent Systems, 2026, 12(1): 41.

[3] Zhang H, Tan L, Liu Y, et al. Energy-efficient multi-UAV collaborative path planning using levy flight and improved gray wolf optimization[C]//2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024: 1-8.

[4] Huang Y, Duan M, Mo L. Multiagent containment control with nonconvex states constraints, nonuniform time delays, and switching directed networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(11): 5021-5028.

[5] Varela G, Caamaño P, Orjales F, et al. Autonomous UAV based search operations using constrained sampling evolutionary algorithms[J]. Neurocomputing, 2014, 132: 54-67.

[6] Yao P, Wang H, Su Z. Real-time path planning of unmanned aerial vehicle for target tracking and obstacle avoidance in complex dynamic environment[J]. Aerospace Science and Technology, 2015, 47: 269-279.

[7] Huang Y, Duan M, Mo L. Multiagent containment control with nonconvex states constraints, nonuniform time delays, and switching directed networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(11): 5021-5028.

[8] Xue Y, Chen W. Multi-agent deep reinforcement learning for UAVs navigation in unknown complex environment[J]. IEEE Transactions on Intelligent Vehicles, 2023, 9(1): 2290-2303.

[9] Zhao D, Huanshi X, Xun Z. Active exploration deep reinforcement learning for continuous action space with forward prediction[J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 6.

[10] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.

[11] Ma Z, Xiong J, Gong H, et al. Adaptive depth

graph neural network-based dynamic task allocation for UAV-UGVs under complex environments[J]. IEEE Transactions on Intelligent Vehicles, 2024, 10(5): 3573-3586.

[12] Zhang H, Li H, Wu X, et al. MTGNet: multi-label mesh quality evaluation using topology-guided graph neural network[J]. Engineering with Computers, 2025, 41(1): 321-333.

[13] Li C, Huang X, Ding J, et al. Global path planning based on a bidirectional alternating search A* algorithm for mobile robots[J]. Computers & Industrial Engineering, 2022, 168: 108123.

[14] Tao F, Ding Z, Wang Z, et al. VPS-APF-RRT*: A variable probability sampling RRT* intelligent algorithm combined with APF for autonomous vehicles path planning[J]. Expert Systems with Applications, 2025: 129965.

[15] Fan J, Chen X, Liang X. UAV trajectory planning based on bi-directional APF-RRT* algorithm with goal-biased[J]. Expert systems with applications, 2023, 213: 119137.

[16] Liu G, Tang Y, Tan Z, et al. A multi-stage optimization strategy for multi-UAV path planning in complex urban 3D environments with obstacle avoidance[J]. Complex & Intelligent Systems, 2026, 12(1): 17.

[17] Zhang P, He Y, Wang Z, et al. Research on multi-UAV obstacle avoidance with optimal consensus control and improved APF[J]. Drones, 2024, 8(6): 248.

[18] Du Z, Zhang H, Wang Z, et al. Model predictive formation tracking-containment control for multi-UAVs with obstacle avoidance[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(6): 3404-3414.

[19] Chu D, Li H, Zhao C, et al. Trajectory tracking of autonomous vehicle based on model predictive control with PID feedback[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(2): 2239-2250.

[20] Kumar P, Pal K, Govil M C. Comprehensive review of path planning techniques for unmanned aerial vehicles (uavs)[J]. ACM Computing Surveys, 2025, 58(3): 1-44.

[21] Wang R, Deng Z. Outdoor Multi-UAV Collaborative Task Planning System Based on Multi-Agent and Improved PPO Algorithm[C]//Proceedings of the 2025 5th International Conference on Automation Control, Algorithm and Intelligent Bionics. 2025: 193-200.

[22] Yao P, Zhao Z. Improved Glasius bio-inspired neural network for target search by multi-agents[J]. Information Sciences, 2021, 568: 40-53.

[23] Liu J, Wei S, Li B, et al. Dual-timescale hierarchical MADDPG for Multi-UAV cooperative search[J]. Journal of King Saud University Computer and Information Sciences, 2025, 37(6): 142.

[24] Zhengyang C A O, Gang C. Enhanced deep reinforcement learning for integrated navigation in multi-UAV systems[J]. Chinese Journal of Aeronautics, 2025, 38(8): 103497.

[25] Li B, Wang J, Song C, et al. Multi-UAV roundup strategy method based on deep reinforcement learning CEL-MADDPG algorithm[J]. Expert Systems with Applications, 2024, 245: 123018.

[26] Zhang Y, Wu Z, Ma Y, et al. Research on autonomous formation of multi-UAV based on MADDPG algorithm[C]//2022 IEEE 17th International Conference on Control & Automation (ICCA). IEEE, 2022: 249-254.

[27] Chen Z, Xu H, Cheng C. Deep reinforcement learning for multi-UAV MEC system trajectory optimization: Minimizing energy consumption[C]//Proceedings of the 2024 2nd International Conference on Internet of Things and Cloud Computing Technology. 2024: 24-30.



Li Tan is currently pursuing the M.S. degree at Beijing Technology and Business University. Her research interests include multi-UAV cooperative technologies, artificial intelligence, and deep reinforcement learning.



Haixia Zhao is currently pursuing the M.S. degree at Beijing Technology and Business University. Her research interests include multi-UAV cooperative technologies, artificial intelligence, and deep reinforcement learning.



Jiaqin Chai is currently pursuing the M.S. degree at Beijing Technology and Business University. Her research interests include path planning technologies, artificial intelligence, and optimize algorithm.



Xinshi Zhang is currently pursuing the M.S. degree at Beijing Technology and Business University. His research interests include path planning technologies, artificial intelligence, and optimize algorithm.



Grigore Stamatescu received the Ph.D. degree from the University POLITEHNICA of Bucharest, in 2012. His research results have been published in over 150 articles. From 2015 to 2016, he was a Fulbright Visiting Scholar with the University California, Merced, on distributed sensing and advanced control in smart buildings, and a 2019 JESH Scholar of Austrian Academy of Sciences at the Graz University of Technology on data-driven modelling of manufacturing systems. He is currently a Professor (Habil. 2019) with the Department of Automation and Industrial Informatics, leading the Artificial Intelligence for SEnsing, Low-Power Networking and ICS SEcurity (AISENSE) Laboratory. His research interests include the Internet of Things and artificial intelligence for industry and energy applications.