



OPEN ACCESS

*CORRESPONDENCE

Yunhong Ma,
✉ mayunhong@nwpu.edu.cn
Weina Fang,
✉ fangweina@mail.nwpu.edu.cn

RECEIVED 29 June 2026
REVISED 07 August 2026
ACCEPTED 28 August 2026
PUBLISHED 09 September 2026

CITATION

Ma Y, Fang W, Wang Y and Kang X (2026)
Research on multi-UAV collaborative
reconnaissance based on MADDPG-
VFNBV algorithm.
Aerosp. Res. Commun. 4:17249.
doi: 10.3389/arc.2026.17249

COPYRIGHT

© 2026 Ma, Fang, Wang and Kang. This is
an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Research on multi-UAV collaborative reconnaissance based on MADDPG-VFNBV algorithm

Yunhong Ma*, Weina Fang*, Yiran Wang and Xinzhe Kang

Northwestern Polytechnical University, Xi'an, China

To address uneven target allocation, repeated exploration, long path lengths, and decision instability caused by dynamic obstacles in multi-UAV cooperative reconnaissance in unknown 3D environments, this paper proposes a MADDPG-VFNBV algorithm that integrates Multi-Agent Deep Deterministic Policy Gradient with the Voronoi-Frontier-Next-Best-View framework. A global 3D voxel map is constructed through multi-UAV map fusion, and exploration targets are generated and assigned using Frontier detection, MeanShift clustering, and dynamic Voronoi partitioning. During NBV decision-making, the MADDPG policy adaptively balances information gain, path cost, coordination cost, and obstacle risk. Experiments show that the proposed algorithm improves reconnaissance efficiency, adaptive decision-making, and cooperative exploration performance in dynamic 3D environments.

KEYWORDS

Frontier exploration, multi-agent deep deterministic policy gradient (MADDPG), multi-UAV, next-best-view(NBV), reinforcement learning

Introduction

Recent advances in UAV platforms, three-dimensional perception, embedded computing, and autonomous control have enabled UAVs to perform mapping, search, inspection, rescue, and reconnaissance in unknown or hazardous environments. With high maneuverability, broad coverage, and low terrain dependence, UAVs are well suited for rapid information acquisition without prior maps, making autonomous exploration path planning a key capability. Unknown-environment exploration aims to maximize information gain under constraints on time, energy, and sensing range. Frontier-based methods guide UAVs toward boundaries between known and unknown space, while systems such as FUEL, FAEP, and RACER improve Frontier management, viewpoint selection, path generation, and trajectory replanning for efficient three-dimensional exploration [1–4]. However, these methods generally rely on manually designed heuristic functions whose weights are difficult to adapt to different exploration stages, environmental structures, and dynamic risks.

Multi-UAV exploration improves efficiency and robustness through parallel perception but also introduces task conflicts, redundant exploration, communication overhead, path intersections, and congestion. Existing studies address these challenges through cooperative grouping, decentralized coordination, spatial partitioning, and distributed planning. Ma et al. proposed a consensus-based grouping algorithm to reduce task-allocation conflicts and resource competition [5]. RACER improves decentralized multi-UAV exploration through spatial decomposition and coverage-path optimization [3], while Dong et al. employed dynamic topological graph-based Voronoi partitioning to allocate tasks with reduced communication burden [6]. Ribeiro and Basiri further combined Frontier and NBV

strategies in a distributed three-dimensional exploration framework [7]. These studies show that effective multi-UAV exploration depends not only on individual path planning but also on coordinated task allocation, information sharing, and decision-making.

Deep reinforcement learning offers a promising approach to autonomous decision-making in unknown and dynamic environments by learning policies that optimize long-term returns. Multi-agent reinforcement learning extends this capability to cooperative tasks through centralized training and decentralized execution [8, 9]. Recent studies have applied these methods to UAV navigation, coverage, and exploration. Westheider et al. developed adaptive multi-UAV path planning for cooperative information gathering [10], while Tan et al. introduced macro-actions for decentralized exploration under communication constraints [11]. Xue and Chen addressed partial observability and environmental non-stationarity in multi-UAV navigation [12], and Chen et al. combined Transformers with reinforcement learning to improve scalability with varying team sizes [13]. MARVEL further integrated graph attention networks for cooperative exploration using limited-field-of-view sensors [14]. These studies demonstrate the potential of reinforcement learning for cooperative decision-making, environmental adaptation, and complex policy learning in multi-UAV systems. Recent studies have also extended learning-based UAV optimization to communication, resource management, and trajectory design. Li et al. developed a distributed many-agent reinforcement learning framework for energy-efficient UAV-driven MEC, jointly considering UAV trajectories, task offloading, and communication/computing resources [15]. Li and Aghvami proposed a hybrid D3QN-TD3 approach for radio resource management in cellular-connected UAV networks [16], while Li et al. introduced a DRL method with quantum-inspired experience replay for UAV trajectory optimization [17]. Moreover, Li et al. employed MADDPG to jointly optimize multi-UAV trajectories and resource allocation in MEC networks [18]. Related studies have also addressed multi-UAV task allocation and coalition formation using contract-net and coalition-based methods [19–20]. Learning- and data-driven approaches have further been applied to multi-UAV path planning and maneuvering target tracking [21–22]. These studies demonstrate the potential of intelligent methods for cooperative decision-making and planning in complex UAV systems.

As summarized in Table 1, existing learning-based UAV studies mainly address resource management and trajectory optimization, while conventional exploration methods rely on predefined viewpoint-evaluation criteria. To overcome these limitations, MADDPG-VFNBV integrates 3D voxel mapping, frontier-based viewpoint generation, dynamic Voronoi allocation, and MADDPG-based adaptive NBV 80 coefficients for cooperative exploration and path planning.

Problem modeling and algorithm design

Multi-UAV cooperative reconnaissance problem model

In an unknown three-dimensional environment, multi-UAV cooperative reconnaissance aims to maximize exploration

coverage while ensuring flight safety and coordination and minimizing redundant exploration and path costs. Consider a swarm of N UAVs, where the position of the i -th UAV is (x_i, y_i, z_i) . Each UAV acquires local environmental information through onboard sensors and updates the global map via map sharing.

This paper adopts a three-dimensional voxel map to discretely represent the environment. Let the environment map at time t be denoted as shown in Equation 1:

$$M_t = \{E_t, O_t, X_t\} \quad (1)$$

where E_t , O_t and X_t represent the free, occupied, and unknown states, respectively. During the exploration of an unknown environment, a Frontier point is defined as a boundary point between a known free region and an unknown region, as expressed in Equation 2:

$$F_t = \{f_1, f_2, \dots, f_n\} \quad (2)$$

Based on the above problem model, this paper further constructs the Voronoi-Frontier-NBV algorithm to realize candidate viewpoint generation, cooperative region partitioning, and next-best-view selection.

Voronoi-Frontier-NBV algorithm

The algorithm detects and clusters Frontier points to generate candidate exploration targets, partitions them among UAVs using dynamic Voronoi regions, and selects the next-best viewpoints through an evaluation function. A path planner then generates the corresponding UAV trajectories.

To reduce the number of candidate points and improve decision-making efficiency, this paper clusters the Frontier point set to obtain several cluster centers which are obtained using the MeanShift algorithm. Starting from the initial seed points, each center is iteratively updated toward the mean position of the Frontier points within the specified bandwidth until convergence. The converged locations are then merged to form the final cluster centers g_k , as given in Equation 3:

$$g_k = \frac{\sum_{f_j \in F_t} K_h(f_j - g_k) f_j}{\sum_{f_j \in F_t} K_h(f_j - g_k)} \quad (3)$$

where $K_h(\cdot)$ is a Gaussian kernel function with bandwidth h . The reachable frontier point closest to each cluster center is selected as a candidate viewpoint, as defined in Equation 4, and these viewpoints form the candidate viewpoint set given in Equation 5:

$$c_k = \arg \min_{f_j \in F_k} \|f_j - g_k\|_2 \quad (4)$$

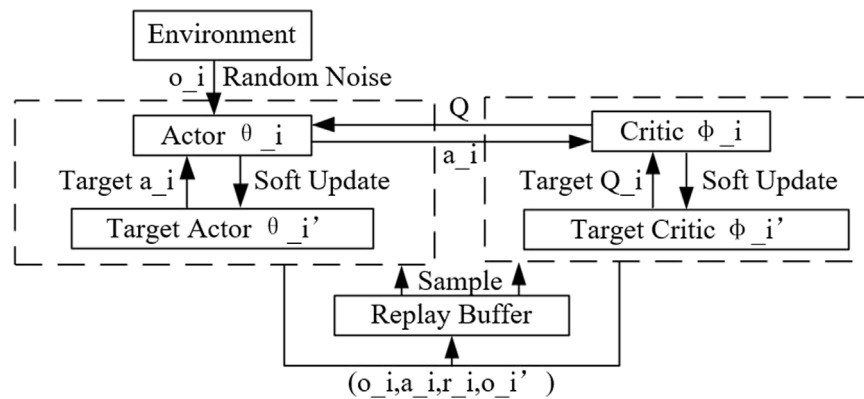
$$C_t = \{c_1, c_2, \dots, c_k\} \quad (5)$$

In the multi-UAV scenario, to reduce repeated exploration caused by different UAVs selecting identical or nearby targets, the concept of dynamic Voronoi partitioning is adopted to allocate candidate points. The arrival cost from the i -th UAV to the position q is defined in Equation 6:

$$R_{di}(q) = \alpha_d d_i(q) + \alpha_h \Delta h_i(q) + \alpha_p \rho_i(q) \quad (6)$$

TABLE 1 Comparison with representative related studies.

Study	Main task	Learning/coordination method	Key feature
Zhou et al. [1]	3D exploration	Frontier + hierarchical planning	Efficient frontier exploration
Zhou et al. [3]	Multi-UAV exploration	Decentralized spatial decomposition	Cooperative exploration without learning
Ribeiro and Basiri [7]	Multi-UAV exploration	Frontier + NBV	Distributed viewpoint planning
Westheider et al. [10]	Information gathering	DRL	Learning-based cooperative path planning
Chiun et al. [14]	Multi-robot exploration	MARL	Cooperative exploration with limited FoV
Li et al. [15]	UAV-assisted MEC	Distributed MADRL	Joint trajectory and resource optimization
Li and Aghvami [16]	Radio resource management	D3QN + TD3	Adaptive communication-resource allocation
Li et al. [17]	UAV path planning	DRL-QiER	Learning-based trajectory optimization
Li et al. [18]	Multi-UAV MEC	MADDPG + PSO	Joint trajectory and resource allocation
MADDPG-VFNBV	3D cooperative reconnaissance	Voronoi + Frontier-NBV + MADDPG	Adaptive NBV weighting for information, path, coordination, and risk

FIGURE 1
The framework diagram of the MADDPG algorithm.

where $d_i(q)$ denotes the length of the shortest feasible path, $\Delta h_i(q)$ denotes the altitude variation cost, and $\rho_i(q)$ denotes the local risk density. The dynamic Voronoi region is defined in Equation 7:

$$V_i(t) = \{q \in F_i \cup C_i | R_{di}(q) \leq R_{dj}(x), \forall j \neq i\} \quad (7)$$

Each UAV preferentially performs NBV scoring within its own candidate region and selects the candidate viewpoint with the highest score as its flight target. The NBV scoring function is given in Equation 8, and the corresponding coefficient vector is defined in Equation 9:

$$S_i(c, t) = \omega_{ig}^i(t)I(c) - \omega_{dist}^i(t)D_i(c) - \omega_{coop}^i(t)P_i(c) - \omega_{obs}^i(t)R_{obs}(c) \quad (8)$$

$$\omega_i(t) = [\omega_{ig}^i, \omega_{dist}^i, \omega_{coop}^i, \omega_{obs}^i] \quad (9)$$

where $I(c)$ denotes information gain, $D_i(c)$ denotes path distance, $P_i(c)$ denotes coordination cost, and $R_{obs}(c)$ denotes obstacle risk cost.

MADDPG algorithm

Multi-UAV target selection depends on UAV states, path costs, target distribution, unknown regions, and obstacle risks, constituting a multi-agent cooperative decision problem. Independent learning may cause non-stationarity, while the continuous NBV weights required in this study are unsuitable for discrete-action methods. Therefore, MADDPG is adopted to learn adaptive NBV coefficients through centralized training with joint states and actions and decentralized execution based on local observations.

Each UAV in the system corresponds to an agent. The local observation of the i -th UAV at time t is denoted as o_i . The Actor network outputs an action according to this observation, as given in Equation 10:

$$a_i = \mu_i(o_i; \theta_i) \quad (10)$$

where μ_i represents the Actor network of the i -th agent and θ_i denotes its network parameters. During training, each Critic

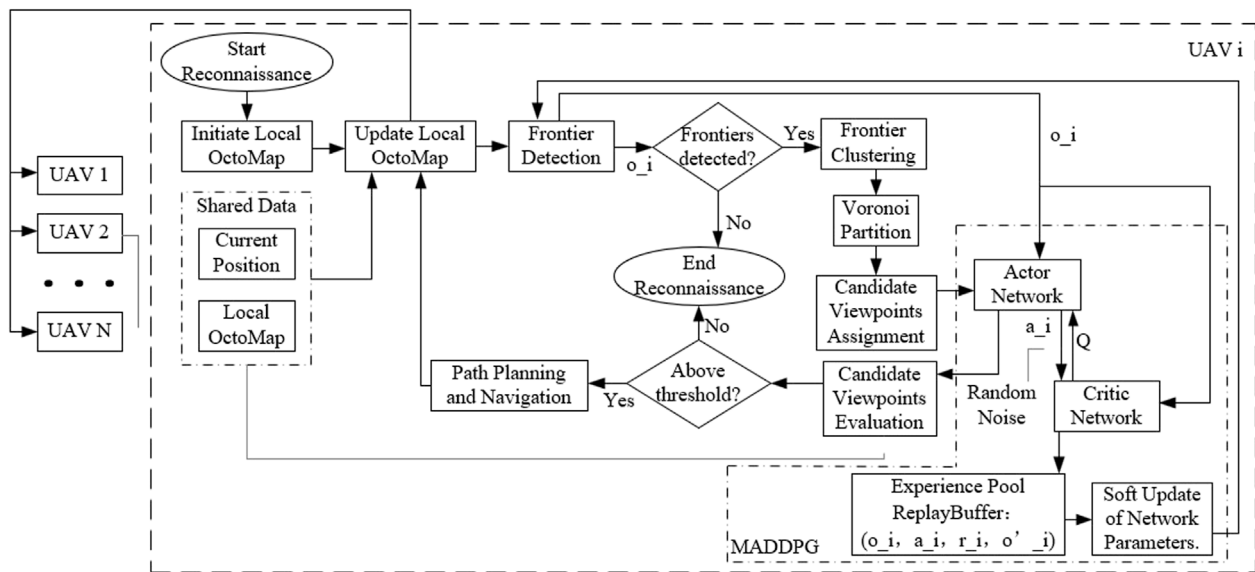


FIGURE 2
Schematic of the MADDPG-VFNBV algorithm.

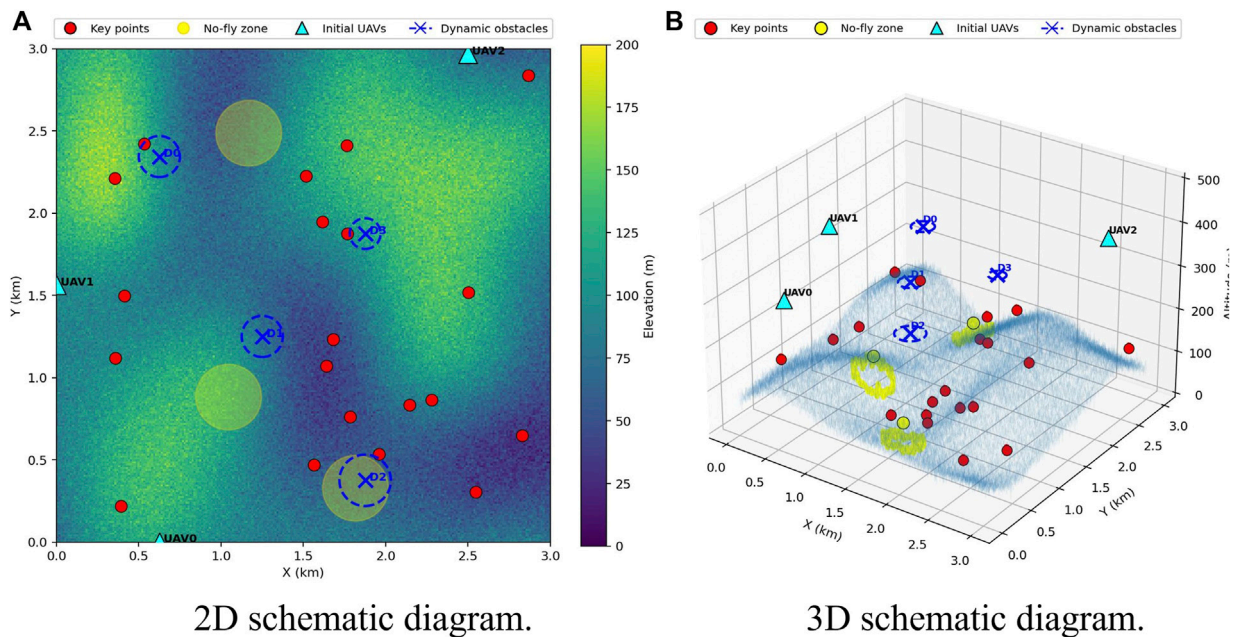


FIGURE 3
Schematic diagrams of the initial experimental environment. (A) 2D schematic diagram. (B) 3D schematic diagram.

estimates the action value Q_i from the joint observations and actions of all UAVs. This centralized information allows it to account for interactions among agents, improving the stability of multi-UAV cooperative policy learning.

For the sample (o, a, r, o') in the experience replay buffer, the target Q-value of the i -th agent can be expressed as shown in Equation 11:

$$y_i = r_i + \gamma Q'_i(o'_i, \dots, o'_n, a'_i, \dots, a'_n) \quad (11)$$

where γ denotes the discount factor, Q'_i denotes the target Critic network, and a'_i denotes the action output by the target Actor network at the next state. The Critic network is updated by minimizing the mean squared error between the current Q-value and the target Q-value, as defined in Equation 12:

$$L(\phi_i) = \min \frac{1}{B} \sum (Q_i - y_i)^2 \quad (12)$$

where B denotes the batch size, and ϕ_i represents the parameters of the i -th Critic network. The objective of the Actor network is to select actions that can obtain higher Q-values. Therefore, the update objective of the Actor network can be formulated as maximizing the Critic's evaluation of the current action, as expressed in Equation 13:

$$J(\theta_i) = \max \frac{1}{B} \sum Q_i \quad (13)$$

To improve training stability, MADDPG introduces target networks and a soft update mechanism. The parameters of the target Actor and Critic networks are slowly updated according to Equations 14, 15, respectively:

$$\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i \quad (14)$$

$$\phi'_i \leftarrow \tau \phi_i + (1 - \tau) \phi'_i \quad (15)$$

where τ denotes the soft update coefficient. This mechanism can prevent drastic fluctuations in the target values, thereby improving the stability of the training process. The framework diagram of the MADDPG algorithm is shown in Figure 1.

MADDPG-VFNBV algorithm

Based on the above Voronoi-Frontier-NBV algorithm and the MADDPG algorithm, this paper introduces MADDPG into the evaluation process of NBV candidate viewpoints and constructs a MADDPG-VFNBV multi-UAV cooperative reconnaissance model.

State space

Let the local observation state of the i -th UAV in the system at time t be denoted as shown in Equation 16:

$$o_i(t) = [\bar{x}_i, \bar{y}_i, \bar{z}_i, Cov_t, \bar{N}_f, \bar{N}_c, \bar{N}_{c_i}, \rho_i^u, \rho_i^d, \rho_i^d] \quad (16)$$

where $\bar{x}_i$, $\bar{y}_i$, and $\bar{z}_i$ denote the normalized 3D coordinates of the i -th UAV; Cov_t is the global map coverage; $\bar{N}_f$ and $\bar{N}_c$ are the normalized numbers of Frontier points and global candidate viewpoints, respectively; $\bar{N}_{c_i}$ is the normalized number of viewpoints assigned to the i -th UAV; ρ_i^u is the local unknown-voxel ratio; and ρ_i^d and ρ_i^d represent global and local obstacle information, respectively.

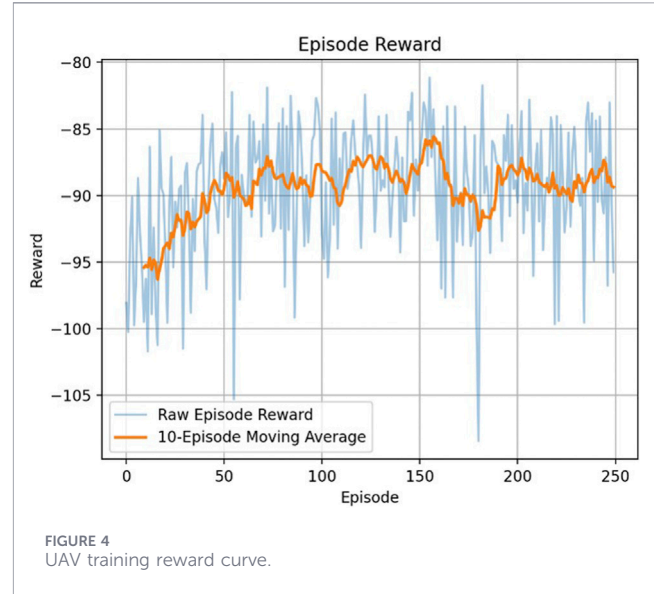
Action space definition

The action of the i -th UAV at time t is defined in Equation 17:

$$a_i(t) = \omega_i(t) = [\omega_{ig}^i, \omega_{dist}^i, \omega_{coop}^i, \omega_{obs}^i] \quad (17)$$

During the actual training process, the action output by the Actor network is normalized using the Softmax function, which ensures that each action component satisfies the normalization and non-negativity constraints are given in Equations 18, 19, respectively:

$$\omega_{ig}^i + \omega_{dist}^i + \omega_{coop}^i + \omega_{obs}^i = 1 \quad (18)$$



$$\omega_{ig}^i, \omega_{dist}^i, \omega_{coop}^i, \omega_{obs}^i > 0 \quad (19)$$

Reward function design

The reward function of the i -th UAV at time t is defined in Equation 20, where the coverage increment and the coefficient-sum constraint are given in Equations 21, 22, respectively:

$$r_i(t) = \lambda_C \Delta C_t - \lambda_m L_i(t) - \lambda_r \rho_i^{rep}(t) - \lambda_d R_i^{obs}(t) - \lambda_s I_i^{stuck}(t) \quad (20)$$

$$\Delta C_t = C_t - C_{t-1} \quad (21)$$

$$\lambda_C + \lambda_m + \lambda_r + \lambda_d + \lambda_s = 1 \quad (22)$$

where ΔC_t denotes the increment in map coverage, $L_i(t)$ denotes the travel distance, $\rho_i^{rep}(t)$ denotes the repeated exploration cost, $R_i^{obs}(t)$ denotes the obstacle risk, and $I_i^{stuck}(t)$ denotes the stagnation cost.

Viewpoints covering more unknown areas receive higher rewards, encouraging information-rich exploration. Travel-distance and repeated-exploration penalties reduce energy consumption and redundant coverage, while obstacle-risk and stagnation costs promote flight safety and prevent prolonged hovering or repeated target selection.

MADDPG-VFNBV algorithm procedure

The main procedure of the MADDPG-VFNBV algorithm is as follows:

1. Initialize the three-dimensional voxel map, UAV states, Actor network, Critic network, and experience replay buffer;
2. Each UAV updates its local map according to sensor observations, and the local maps are fused to obtain the global map;
3. Detect Frontier points in the map and generate the candidate viewpoint set through clustering;
4. Assign a subset of candidate viewpoints to each UAV according to the idea of dynamic Voronoi partitioning;

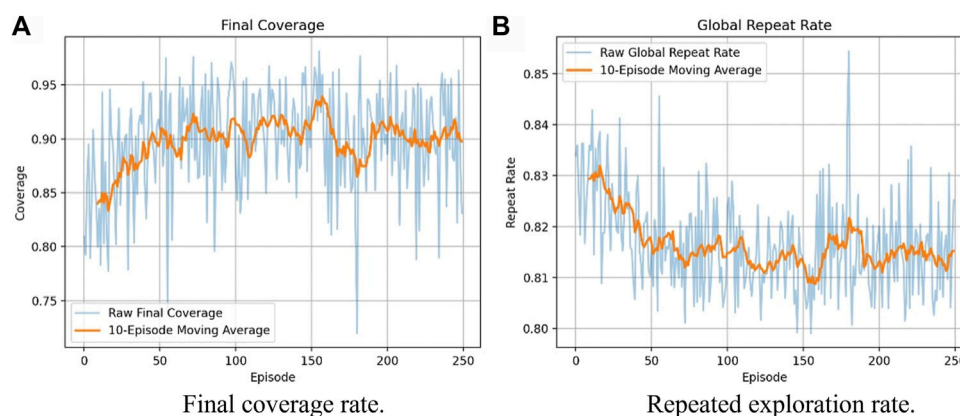


FIGURE 5

Training curves of the final coverage rate and repeated exploration rate. (A) Final coverage rate. (B) Repeated exploration rate.

5. Each UAV inputs its current state into the Actor network to output the NBV scoring coefficients;
6. Select the next best viewpoint for each UAV according to the NBV scoring function;
7. Use the A* algorithm for path planning and control the UAVs to move;
8. Calculate the reward according to the coverage increment, path cost, repeated exploration rate, and obstacle risk;
9. Store the experience samples in the experience replay buffer and update the Actor-Critic networks;
10. Repeat the above process until the maximum number of steps is reached or the map coverage satisfies the task requirements.

The schematic diagram of the MADDPG-VFNBV algorithm is shown in Figure 2.

Results

Simulation environment and initial map

The simulations are conducted in an unknown three-dimensional terrain environment covering a $3 \text{ km} \times 3 \text{ km}$ area with undulating terrain, key information points, no-fly zones, dynamic obstacles, and predefined UAV initial positions. Figure 3 presents two-dimensional and three-dimensional views of the initial environment. Red dots denote key information points, yellow circles indicate no-fly zones, cyan triangles show UAV initial positions, and blue dashed circles and crosses represent dynamic obstacles and their influence areas. The top view illustrates the horizontal distribution of task elements, while the three-dimensional view shows terrain elevation and UAV initial altitudes. No-fly zones and dynamic obstacles affect exploration path planning and candidate-viewpoint selection.

Three UAVs cooperatively explore an unknown area using onboard sensors while avoiding static impassable regions and dynamic obstacles. The objective is to maximize map coverage

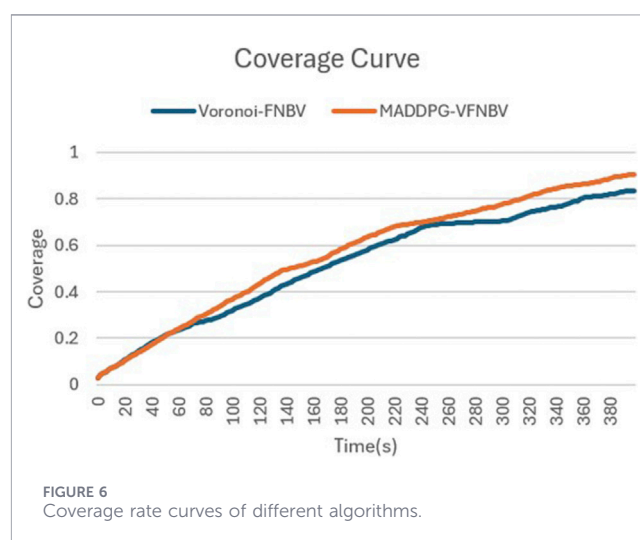


FIGURE 6

Coverage rate curves of different algorithms.

within a specified time while minimizing path redundancy and repeated exploration.

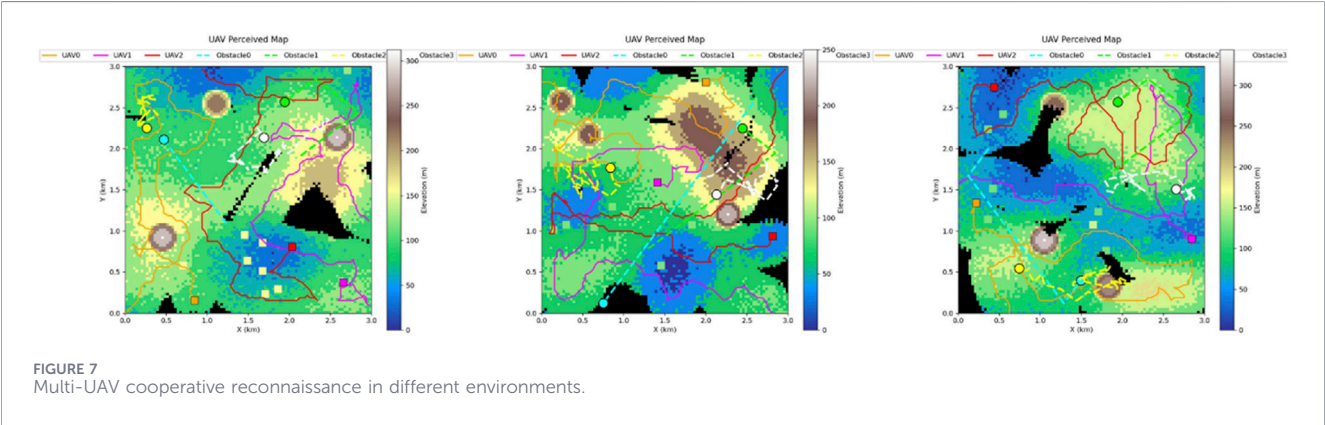
Analysis of the training process

The Voronoi-Frontier-NBV method uses fixed coefficients without training, where as MADDPG-VFNBV requires model training. The Actor network consists of two hidden layers with 128 neurons each, while the Critic network contains two hidden layers with 256 neurons each. Both networks are optimized with learning rates of 1×10^{-4} and 1×10^{-3} for the Actor and Critic, respectively. The batch size is 128, the discount factor is 0.95, and the soft-update coefficient is 0.007. Gaussian noise is adopted for exploration during training. At the end of each episode, the cumulative reward, final coverage, and repeated exploration rate are recorded, with the episode rewards shown in Figure 4.

Figure 4 shows noticeable fluctuations in episode rewards due to changes in candidate-viewpoint distribution, dynamic obstacles, and environmental noise. However, the 10-episode moving average

TABLE 2 Comparison of exploration performance of different algorithms.

Algorithm	Final coverage rate	Total path length (m)	Longest Single-UAV path (m)	Repeated exploration rate
Voronoi-frontier-NBV	0.835728	27,532.049	9,321.432	0.813026
MADDPG-VFNBV	0.905165	26,543.136	9,114.667	0.789434



generally increases and gradually converges, indicating that the Actor network learns an increasingly effective strategy for adjusting the NBV coefficients.

Figure 5 shows that, as training progresses, the final coverage rate increases and stabilizes at approximately 0.90, while the repeated exploration rate decreases to a low level. This indicates that the trained reinforcement learning module improves viewpoint selection and reduces competition among UAVs.

The training curves show that MADDPG-VFNBV gradually converges, with increasing episode rewards and coverage rates and decreasing repeated exploration, indicating that the reinforcement learning module effectively learns to adjust the NBV coefficients.

Comparative experimental setup

To evaluate the improvement of MADDPG-VFNBV over Voronoi-Frontier-NBV, both methods are tested in the same three-dimensional environment under identical conditions, including UAV number, sensor range, flight time, and path planning method.

The experimental results are shown in Figure 6. The MADDPG-VFNBV algorithm achieves higher coverage during most of the simulation, particularly in the middle and later stages. According to Table 2, compared with Voronoi-Frontier-NBV under the same flight time, it increases final coverage by approximately 6.94%, reduces total path length by 3.59% and the longest single-UAV path by 2.22%, and lowers repeated exploration by 2.90%. These results demonstrate that reinforcement learning improves the adaptability of the baseline method to different exploration stages and environmental changes, thereby enhancing multi-UAV cooperative exploration performance.

Adaptability verification of the MADDPG-VFNBV algorithm in different environments

To further verify the adaptability of the MADDPG-VFNBV algorithm under different environmental conditions, this paper conducts tests in multiple randomly generated three-dimensional terrain environments. The different test environments vary in terms of UAV initial positions, terrain elevation distributions, no-fly zone locations, and dynamic obstacle distributions.

In Figure 7, square markers and solid lines denote UAVs and their trajectories, while circles and dashed lines represent dynamic obstacles and their trajectories. The UAVs avoid no-fly zones and dynamic obstacles while expanding toward unexplored regions. When key points and unknown areas are dispersed, they explore in different directions to reduce regional overlap. These results demonstrate the algorithm’s adaptability to varying terrain and obstacle distributions.

Discussion

The results show that MADDPG-VFNBV improves coverage while reducing path length and repeated exploration compared with the fixed-coefficient Voronoi-Frontier-NBV method. This indicates that adaptive NBV weighting can better balance exploration efficiency, coordination, and obstacle risk.

Unlike conventional methods using predefined evaluation coefficients, MADDPG-VFNBV dynamically adjusts the NBV coefficients according to the exploration state. Tests in different environments also demonstrate its adaptability to varying terrain and obstacle distributions.

The current study is limited to simulation with a small UAV team. Further work will include multi-seed statistical evaluation, ablation studies, computational-complexity analysis, larger UAV teams, and comparisons with other multi-agent reinforcement learning methods.

Overall, this paper demonstrates that integrating MADDPG with the Voronoi-Frontier-NBV framework provides an effective approach for multi-UAV cooperative reconnaissance in unknown three-dimensional environments. By adaptively adjusting the NBV evaluation coefficients according to the exploration state, the proposed method improves exploration coverage while reducing path length and repeated exploration. These results highlight the potential of adaptive reinforcement-learning-based viewpoint selection for improving cooperative exploration efficiency.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YM provided supervision and guidance for the research and reviewed the manuscript. WF was responsible for the methodology, algorithm development, and writing of the original draft. YW and

XK assisted with result checking, manuscript revision, and proofreading. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

References

- Zhou B, Zhang Y, Chen X, Shen S. FUEL: fast UAV exploration using incremental frontier structure and hierarchical planning. *IEEE Robotics Automation Lett* (2021) 6(2): 779–86. doi: 10.1109/LRA.2021.3051563
- Cao C, Zhu H, Yang F, Xia Y, Choset H, Oh J, et al. Autonomous exploration development environment and the planning algorithms. In: *IEEE International Conference on Robotics and Automation (ICRA)* (2022). p. 8921–8. doi: 10.1109/ICRA46639.2022.9812330
- Zhou B, Xu H, Shen S. RACER: Rapid collaborative exploration with a decentralized Multi-UAV system. *IEEE Trans Robotics* (2023) 39(3):1816–35. doi: 10.1109/TRO.2023.3236945
- Zhao Y, Yan L, Xie H, Dai J, Wei P. Autonomous exploration method for fast unknown environment mapping by using UAV equipped with limited FOV sensor. *IEEE Trans Ind Electronics* (2024) 71(5):4933–43. doi: 10.1109/TIE.2023.3285921
- Ma Y, Li X, Wang H, Kang Y, Cao T, Zhang Y, et al. Multi-UAVs collaborative task allocation based on improved consensus-based grouping algorithm. In: *IEEE 18th International Conference on Control and Automation (ICCA)* (2024). p. 850–5. doi: 10.1109/ICCA62789.2024.10591950
- Dong Q, Xi H, Zhang S, Bi Q, Li T, Wang Z, et al. Fast and communication-efficient Multi-UAV exploration via voronoi partition on dynamic topological graph. In: *2024 International Conference on Intelligent Robots and Systems* (2024). p. 14063–70. doi: 10.1109/IROS58592.2024.10801613
- Ribeiro A, Basiri M. Efficient 3D exploration with distributed Multi-UAV teams: integrating frontier-based and next-best-view planning. *Drones* (2024) 8(11):630. doi: 10.3390/drones8110630
- Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. *Artif Intelligence Rev* (2022) 55(2):895–943. doi: 10.1007/s10462-021-09996-w
- Canese L, Cardarilli GC, Di Nunzio L, Fazzolari R, Giardino D, Re M, et al. Multi-agent reinforcement learning: a review of challenges and applications. *Appl Sci* (2021) 11(11):4948. doi: 10.3390/app11114948
- Westheider J, Rueckin J, Popovic M. Multi-UAV adaptive path planning using deep reinforcement learning. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2023). p. 649–56. doi: 10.1109/IROS55552.2023.10342516
- Tan AH, Bejarano FP, Zhu Y, Ren R, Nejat G. Deep reinforcement learning for decentralized multi-robot exploration with macro actions. *IEEE Robotics Automation Lett* (2023) 8(1):272–9. doi: 10.1109/LRA.2022.3224667
- Xue Y, Chen W. Multi-agent deep reinforcement learning for UAVs navigation in unknown complex environment. *IEEE Trans Intell Vehicles* (2024) 9(1):2290–303. doi: 10.1109/TIV.2023.3298292
- Chen D, Qi Q, Fu Q, Wang J, Liao J, Han Z. Transformer-based reinforcement learning for scalable Multi-UAV area coverage. *IEEE Trans Intell Transportation Syst* (2024) 25(8):10062–77. doi: 10.1109/TITS.2024.3358010
- Chiun J, Zhang S, Wang Y, Cao Y, Sartoretti G. MARVEL: multi-agent reinforcement learning for constrained field-of-view multi-robot exploration in large-scale environments. In: *International Conference on Robotics and Automation-ICRA-Annual* (2025). p. 11392–8. doi: 10.1109/ICRA55743.2025.11127700
- Li Y, Madhukumar AS, Zheng Hui Ernest T, Zheng G, Saad W, Aghvami AH. Energy-efficient UAV-driven multi-access edge computing: a distributed many-agent perspective. *IEEE Trans Commun* (2025) 73(9):8405–20. doi: 10.1109/TCOMM.2025.3552746
- Li Y, Aghvami AH. Radio resource management for cellular-connected UAV: a learning approach. *IEEE Trans Commun* (2023) 71(5):2784–800. doi: 10.1109/TCOMM.2023.3262826
- Li Y, Aghvami AH, Dong D. Path planning for cellular-connected UAV: a DRL solution with quantum-inspired experience replay. *IEEE Trans Wireless Commun* (2022) 21(10):7897–912. doi: 10.1109/TWC.2022.3162749
- Li Z, Xu C, Zhang Z, Wu R. Deep reinforcement learning based trajectory design and resource allocation for task-aware multi-UAV enabled MEC networks. *Comput Commun* (2024) 213:88–98. doi: 10.1016/j.comcom.2023.11.006
- Ma Y, Kang Y, Wang H, Fang W, Li Y. Multi-UAV task allocation based on improved contract net algorithm. In: *Trends in Advanced Unmanned Aerial Systems* (Singapore: Springer) (2026), 117–124. doi: 10.1007/978-981-95-7840-5_12
- Ma Y, Bai S, Xu Z, Yang J. The formation method of task coalition for multi-UAV cooperation. In: *16th International Conference on Control & Automation (ICCA)* (IEEE) (2020), 713–718. doi: 10.1109/ICCA51439.2020.9264544
- Ma Y, Fang W, Wang H, Kang Y, Li Y. Research on multi-UAV path planning based on the MATD3-APF algorithm. In: *Trends in Advanced Unmanned Aerial Systems* (Singapore: Springer) (2026), 108–116. doi: 10.1007/978-981-95-7840-5_11
- Ma Y, Cao T, Wang Z, Li X, Wang Y. A data-driven neural network used on maneuvering target tracking. In: *International Conference on Cyber-Physical Social Intelligence (ICCSI)* (IEEE) (2023), 362–367. doi: 10.1109/ICCSI58851.2023.10304033