drones

Article

Reinforcement-Learning-Based Multi-UAV Cooperative Search for Moving Targets in 3D Scenarios

Yifei Liu [1], Xiaoshuai Li [1,2,*], Jian Wang [1,2], Feiyu Wei [1] and Junan Yang [1,2]

[1] College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China; liuyifei@nudt.edu.cn (Y.L.)
[2] Anhui Province Key Laboratory of Electronic Restriction, Hefei 230037, China
* Correspondence: xiaoshuai.li@nudt.edu.cn

Abstract: Most existing multi-UAV collaborative search methods only consider scenarios of two-dimensional path planning or static target search. To be close to the practical scenario, this paper proposes a path planning method based on an action-mask-based multi-agent proximal policy optimization (AM-MAPPO) algorithm for multiple UAVs searching for moving targets in three-dimensional (3D) environments. In particular, a multi-UAV high–low altitude collaborative search architecture is introduced that not only takes into account the extensive detection range of high-altitude UAVs but also leverages the benefit of the superior detection quality of low-altitude UAVs. The optimization objective of the search task is to minimize the uncertainty of the search area while maximizing the number of captured moving targets. The path planning problem for moving target search in a 3D environment is formulated and addressed using the AM-MAPPO algorithm. The proposed method incorporates a state representation mechanism based on field-of-view encoding to handle dynamic changes in neural network input dimensions and develops a rule-based target capture mechanism and an action-mask-based collision avoidance mechanism to enhance the AM-MAPPO algorithm's convergence speed. Experimental results demonstrate that the proposed algorithm significantly reduces regional uncertainty and increases the number of captured moving targets compared to other deep reinforcement learning methods. Ablation studies further indicate that the proposed action mask mechanism, target capture mechanism, and collision avoidance mechanism of the AM-MAPPO algorithm can improve the algorithm's effectiveness, target capture capability, and UAVs' safety, respectively.

Keywords: multi-UAV; reinforcement learning; cooperative target search; path planning

Citation: Liu, Y.; Li, X.; Wang, J.; Wei, F.; Yang, J. Reinforcement-Learning-Based Multi-UAV Cooperative Search for Moving Targets in 3D Scenarios. *Drones* **2024**, *8*, 378. https://doi.org/10.3390/drones8080378

Academic Editor: Oleg Yakimenko

Received: 19 June 2024
Revised: 17 July 2024
Accepted: 2 August 2024
Published: 6 August 2024

1. Introduction

In the past few decades, UAV technology has undergone significant advancements, facilitating its extensive application in civil and military fields [1–5]. Multi-UAV systems have garnered considerable attention from both industry and academia due to their superior spatial coverage and efficient cooperative capabilities [6], particularly in target search tasks such as emergency rescue [7,8] and military reconnaissance [9,10]. During the execution of these tasks, the precision and dependability of path planning play crucial roles in ensuring the system's security and stability [11–13].

In the study of the path planning problem for multi-UAV collaborative target search (MCTS), the academic community has primarily proposed two solutions: traditional search methods based on heuristic algorithms [14,15] and intelligent search methods based on reinforcement learning [16,17]. Traditional approaches can yield satisfactory outcomes under specific conditions, but they generally exhibit limited flexibility and adaptability when confronted with complex environments and dynamically evolving tasks. In contrast, reinforcement-learning-based approaches demonstrate significant advantages in terms of intelligence, flexibility, and adaptability. Therefore, the MCTS approach based on reinforcement learning has emerged as a significant area of research [18,19].

The MCTS problem is modeled as a decentralized partially observable Markov decision process (DEC-POMDP) [20], providing a theoretical foundation for employing multi-agent reinforcement learning (MARL) [21]. Employing MARL to address the MCTS problem offers distinct advantages. First, the MARL approach can effectively navigate highly dynamic environmental changes and uncertainties. By continually learning from environmental feedback, this approach can iteratively optimize the decision-making strategy, thereby enhancing the system's adaptability and robustness. Second, the MARL approach supports distributed decision-making, enabling UAVs to achieve autonomous cooperation without central control. This distributed decision-making mechanism significantly enhances the system's expandability and flexibility. Furthermore, the MARL approach can further enhance task execution efficiency by learning to explore potential strategies and cooperation mechanisms in complex tasks [22].

Numerous scholars have applied the MARL algorithms to address the path planning problems of MCTS tasks and have reported significant successes [16,23–25]. However, studies on MARL-based MCTS still face several significant challenges. First, most existing studies primarily focus on static targets, with limited attention to the impact of moving targets. Second, most current research focuses on path planning in two-dimensional (2D) search scenarios, whereas real-world deployment often requires navigating more complex 3D path planning challenges. Modern UAVs are equipped with advanced zoom functionalities that can adjust the field of view and maintain high-quality detection without changing the flight altitude. However, their zoom capabilities are limited, and in certain complex scenarios, relying solely on zoom functionalities may not satisfy all detection requirements. Therefore, studying the 3D path planning of UAVs with different detection capabilities at different altitudes is of great significance for improving search performance. Lastly, traditional MARL algorithms typically exhibit slow convergence rates: a limitation that is especially evident when UAV swarms undertake moving target search tasks in 3D space.

In order to address the challenges above, this paper investigates the path planning problem of multiple UAVs engaged in the search for moving targets within a 3D environment. We propose a path planning strategy utilizing the AM-MAPPO algorithm. This strategy is designed to minimize the uncertainty of the search area and enhance the number of captured moving targets while ensuring effective collision avoidance between UAVs. The study develops a rule-based target capture mechanism and an action-mask-based collision avoidance mechanism to diminish the dimension of the action space of the MARL algorithm, thus improving the convergence speed of the AM-MAPPO algorithm. In addition, to address the issue of dynamic changes in the input dimensions of neural networks caused by the uncertainty of the information observed by UAVs, this study introduces a field-of-view-encoding-based state representation mechanism. The contributions of this paper are summarized as follows.

- Model and problem formulation: A multi-UAV high–low altitude collaborative search architecture is proposed by taking into account the varied detection capabilities of UAVs at different flight altitudes. Under the constraints of UAVs' flight distances, detection abilities, and collision avoidance, an optimization problem for multi-UAV cooperative search for moving targets in a 3D scenario is established to minimize the uncertainty of the search area and maximize the number of captured moving targets.
- Algorithm design: Firstly, a rule-based target capture mechanism is designed to fully utilize the advantages of the extensive field of view of high-altitude UAVs and the high detection quality of low-altitude UAVs. Secondly, an action-mask-based collision avoidance mechanism is designed to eliminate the risk of collisions between UAVs. Thirdly, a field-of-view-encoding-based state representation mechanism is introduced to address dynamic changes in the input dimensions of neural networks. Finally, to address the problem of low learning efficiency caused by the ineffective actions of UAVs in boundary states and the potential for dangerous actions in collision-risk situations, an improved MAPPO algorithm is proposed.

- Experimental verification: To verify the performance of our proposed algorithm, it is compared with other DRL-based methods, including VDN, QMIX, and MAPPO, through simulation. The simulation results demonstrate the proposed algorithm's superiority. In addition, an ablation study is conducted to explore the effectiveness of the action mask mechanism, the target capture mechanism, and the collision avoidance mechanism.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the system model. Section 4 presents the proposed multi-UAV cooperative search method for moving targets. Numerical results are presented in Section 5, and Section 6 concludes this paper.

2. Related Work

The cooperative target search process involving multiple UAVs is characterized by a group of UAVs following predetermined rules to reduce uncertainty and to discover targets within an unknown environment [26]. In theoretical research, this task is often formulated as a multi-objective optimization problem to maximize search efficiency and the probability of target discovery while adhering to constraints such as flight time, distance, and collision avoidance for UAVs [27–30]. This paper aims to deepen the understanding of advancements in this field by focusing on two principal research directions: cooperative target search methods based on heuristic algorithms and those based on reinforcement learning.

2.1. Cooperative Target Search Methods Based on Heuristic Algorithms

In the field of cooperative target search, traditional strategies primarily employ heuristic algorithms such as ant colony, particle swarm optimization, and pigeon-inspired optimization. These algorithms leverage the collective behaviors of biological entities in nature to solve complex optimization problems. Yue et al. [31] proposed a rule-based heuristic multi-ant colony cooperative search algorithm to enhance the capabilities and search efficiency of multiple UAVs engaged in anti-submarine tasks in unknown marine environments. This algorithm extensively considers the cooperation between target submarines and warships, ensuring quick and effective target localization by UAVs in complex adversarial situations. Pérez-Carabaza et al. [32] proposed a search strategy based on the ant colony optimization algorithm to tackle the challenge of multiple UAVs simultaneously searching for multiple targets in uncertain environments. This strategy optimizes UAV flight trajectories to locate all targets in the shortest possible time. Research indicates that this strategy offers significant advantages over existing methods in addressing complex multi-target, shortest-time search problems. While exploring cooperative search and attack missions by multiple UAVs under complex constraints, Duan et al. [33] developed a dynamic discrete pigeon-inspired optimization algorithm. This method provides a solution with superior global optimal search capabilities in discrete environments. Considering factors such as UAV maneuverability, collision avoidance, and communication constraints, Xu et al. [34] established a multi-UAV cooperative path planning model to minimize the planned flight path length for each UAV. They proposed an enhanced particle swarm optimization algorithm that integrates dynamic multi-swarm and comprehensive learning particle swarm approaches, significantly improving the performance of baseline algorithms.

Heuristic algorithms can rapidly provide solutions in specific environments. However, they often exhibit limited flexibility and adaptability, particularly when faced with dynamically changing environments and complex task allocation challenges. This limitation primarily arises from heuristic algorithms' reliance on specific environmental prior knowledge during their design, which restricts their application in unknown or changing environments. During the search process, these algorithms often suffer from inadequate collaboration and a tendency to become trapped in local optima, thereby diminishing the overall search efficiency.

2.2. Cooperative Target Search Methods Based on Reinforcement Learning

Reinforcement learning seeks approximate optimal solutions through continuous trial-and-error learning in unknown environments, demonstrating significant advantages in addressing complex decision-making problems. A principal characteristic of this method is its ability to adapt effectively to dynamic environmental changes without exhaustive prior knowledge of the environment [18]. Against this backdrop, MARL provides crucial theoretical foundations and methodological support for addressing cooperative target search problems involving multiple UAVs. Each UAV functions as an independent agent in multi-UAV systems, accomplishing collaborative tasks without central control. MARL employs local observation and global reward mechanisms to significantly enhance the overall search efficiency of multiple UAVs through collaboration based on individual learning. In addition, the design of MARL effectively addresses the challenges of asymmetric perception information and high decision complexity among multiple UAVs, simultaneously ensuring their adaptability and flexibility to the environment. Therefore, MARL not only conforms to the distributed characteristics of multi-UAV systems but also markedly enhances the efficiency of collaborative search.

In recent years, significant advancements have been achieved in both the application and research of MARL in cooperative search involving multiple UAVs. Considering the potential for detection errors in the sensors and target detection algorithms of UAVs during target search tasks, Luo et al. [24] proposed a centralized deep reinforcement learning (DRL) approach based on deep Q networks aimed at jointly optimizing computational offloading decisions and flight path planning for cooperative target searches involving multiple UAVs. This method utilizes the concept of uncertainty to evaluate the search process, framing the search task as an optimization problem aimed at minimizing uncertainty under energy and time constraints.

However, with the increasing complexity of search tasks, centralized methods that focus computing tasks on central nodes may encounter bottlenecks due to high computational and communication costs. Therefore, the latest research trend focuses on developing distributed multi-UAV collaborative search algorithms to enable autonomous cooperation and mutual coordination among UAVs, with the aim of efficiently conducting search tasks in unknown environments without depending on centralized control systems. Hou et al. [25] proposed a distributed collaborative UAV swarm search strategy to address the challenge of real-time and accurate collection and processing of global information in large-scale search scenarios. This strategy segments complex search scenarios into multiple local regions, thus enabling UAV swarms to conduct searches more efficiently in a distributed manner. Considering the inherent dynamism of UAV swarm search tasks and the necessity of making sequential decisions in constantly changing environments, they introduced a method based on the deep deterministic policy gradient (DDPG) algorithm, modified for multi-agent systems, known as MADDPG. Wang et al. [16] developed a task planning model for multi-agent collaborative search in uncertain environments and proposed an enhanced reinforcement learning algorithm employing a behavioral preference selection strategy. The experimental results demonstrate that this method exhibits significant advantages over other algorithms in terms of task completion rate, target search efficiency, and average search time. Furthermore, the movement trajectory of multi-agent systems proves to be more efficient, highlighting its potential application in multi-objective search tasks. Zhang et al. [21] investigated the multi-UAV collaborative reconnaissance and search (MCRS) problem to ensure sufficient coverage of the mission area and precise positioning of static targets. They introduced a generalized confidence probability mapping model grounded in DS evidence theory for target search and uncertainty evaluation. To address the MCRS problem, a novel deep reinforcement learning algorithm named double criticism deep deterministic policy gradient (DCDDPG) was introduced. The simulation results indicate that DCDDPG surpasses other deep reinforcement learning algorithms in terms of convergence performance and excels at improving search efficiency and reducing uncertainty compared to existing path planning algorithms. However, the research

above primarily concentrates on static target search, limiting its application to moving target environments.

Shen et al. [23] addressed the challenge faced by multiple UAVs with constrained perception ranges and communication capabilities collaborating to locate static targets in environments characterized by dynamic threats. This study's primary goal was to expedite the detection of unknown targets using multiple UAVs, enhance the mission area's coverage, and navigate UAVs away from potential threats. To address the MCTS challenge, a novel MARL approach named DNQMIX was introduced. Simulation outcomes demonstrated that DNQMIX surpassed existing state-of-the-art methods in terms of search rate and area coverage. Given the observed limitations in the generalization and stability of the training process among existing MARL-based multi-agent target search methods when tackling complex tasks, Cao et al. [35] concentrated on exploring target search strategies with enhanced generalization capabilities. They proposed a hierarchical deep Q network incorporating multi-criteria negative feedback (MNF-HDQN) within the deep reinforcement learning framework. MNF-HDQN demonstrates robust generalization capabilities in complex and uncertain target search scenarios, effectively directing multi-agent systems towards efficient target search. However, the studies above primarily concentrate on multi-agent target search and path planning within two-dimensional contexts, whereas real-world applications frequently necessitate more intricate three-dimensional path planning.

2.3. Summary

In summary, reinforcement learning algorithms exhibit superior adaptability to dynamically changing environments compared to heuristic algorithms, bestowing them with significant advantages in dealing with uncertain and complex target search tasks. Within this framework, despite MARL's initial advancements in MCTS, it still encounters several challenges and limitations in the research and application phases. Firstly, the predominant focus of existing research on static target search constrains the models' effectiveness and applicability in more complex moving target environments. Secondly, the paucity of research on path planning for multi-UAV target search in 3D scenes hampers the algorithms' deployment and application in real-world settings.

Compared to static target search scenarios, moving target search scenarios impose higher demands on the real-time performance and response speed of path planning algorithms. In the path planning of static targets, the target's position is fixed, and the path planning algorithm only needs to find the optimal path to reach the target. However, moving targets' positions are constantly changing, and UAVs need to track the target's position in real-time and to continuously adjust their paths. Compared to path planning in 2D scenarios, the computational complexity of 3D path planning is significantly higher. The additional dimension (z-axis) causes the search space to grow exponentially, and path planning algorithms must handle numerous states and actions, dramatically increasing computational complexity and requiring them to manage many variables and constraints.

To address the limitations of existing methods, this paper proposes a novel reinforcement-learning-based path planning method designed to overcome the unique challenges faced by multiple UAVs in the collaborative search for moving targets in 3D environments. This method comprehensively considers various constraints, including the area coverage and target detection capabilities of high-altitude and low-altitude UAVs, flight limitations, and collision avoidance requirements. This innovative work aims to provide novel solutions and theoretical support for addressing the path planning challenges of UAVs in searching for moving targets in 3D environments. Table 1 outlines the differences between this work and existing works based on reinforcement learning algorithms.

Table 1. Comparison of this work with existing works.

Reference	Target Model	Environment Model
[16]	Static target	2D
[21]	Static target	2D
[23]	Static target	2D
[24]	Static target	2D
[25]	Static target	2D
[35]	Static target, randomly moving target, intelligent avoidance target	2D
This work	Moving target	3D

3. System Model

3.1. Environment Model

Figure 1 depicts a scenario in which multiple UAVs collaboratively search for moving targets, including ground targets such as evacuees and emergency vehicles. The target search area A is characterized as a rectangular two-dimensional plane with dimensions $X \times Y$. This area is further divided into $L_X \times L_Y$ small grids, with the coordinates of grid k denoted by (k_x, k_y). The initial positions of moving targets within task area A are unknown, and the targets move randomly at a constant speed. In the initial phase, the probability of target presence p_k in each grid is established to be 0.5, while the uncertainty index χ_k is set to 1. Furthermore, N UAVs initiate their mission by taking off from their starting points and can fly at different altitude layers. Their mission entails exploring the unknown search area, with the aims of reducing uncertainty within the region and endeavoring to capture as many moving targets as possible. It is particularly noteworthy that, due to the specific detection probabilities and false alarm probabilities of the sensors onboard the UAVs, the results of a single scan of the target area is unreliable. Therefore, multiple repeated scans are required to enhance the reliability of the detection results [24]. We denote a set of UAVs as $i \in \mathbb{N} = \{1, 2, \ldots, N\}$, a set of moving targets as $d \in \mathbb{D} = \{1, 2, ..., D\}$, and a set of grids as $k \in \mathbb{K} = \{1, 2, \ldots, K\}$.

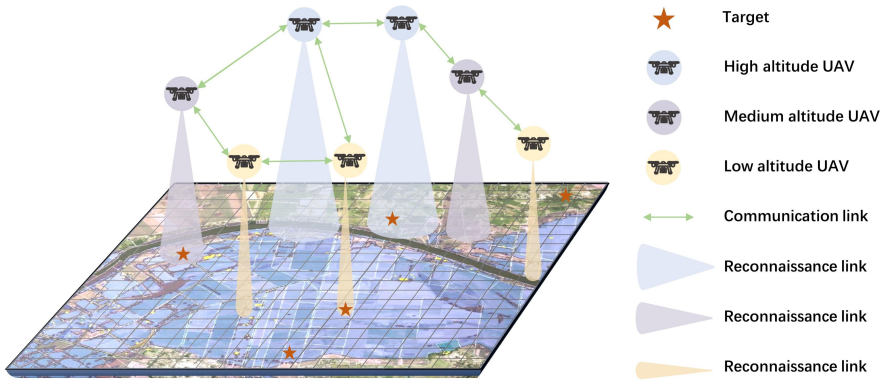

Figure 1. Multi-UAV collaborative target search scenario.

3.2. UAV Model

During each time step, UAV i moves in a 3D space by executing an action a_i^t. The set of available actions $\mathcal{A}$ includes six directions: north (0), east (1), south (2), west (3), up (4), and down (5). Among them, actions 0–3 are available for horizontal movement on a 2D plane, while actions 4 and 5 control vertical movement in a 3D space. As illustrated in Figure 2, upon reaching the boundary of the search area, the movement options for UAVs decrease to a subset of these directions.

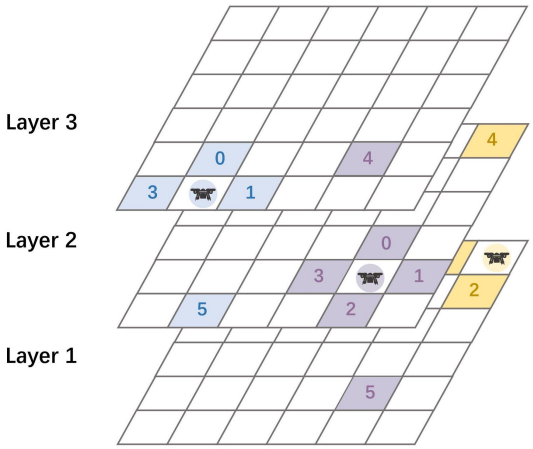

Figure 2. The optional actions for UAVs. The numbers 0, 1, 2, 3, 4 and 5 correspond to the directions north, east, south, west, up and down, respectively.

Given the varying detection capabilities of UAVs across different flight altitudes, this paper uses P_{i,D_z} and P_{i,F_z} to denote the detection probability and false alarm probability, respectively, of UAV i at layer z [36]. Specifically, as the flight altitude of the UAV increases, its detection range broadens. Although this change enhances the coverage area, it also results in a decrease in detection probability and an increase in false alarm probability. Accordingly, UAVs can monitor a broader area when flying at higher altitudes, while the accuracy of target recognition correspondingly diminishes [37].

3.3. Belief Probability Map Update Model

The value range of the target probability distribution map $p^t(k_x, k_y)$ lies between $[0, 1]$, indicating the probability of a target's presence in grid k at time t [38]. Specifically, $p^t(k_x, k_y) = 0$ signifies there is no target present in grid k, while $p^t(k_x, k_y) = 1$ confirms the presence of a target in grid k. For the initialization of the UAV search mission, the probability of a target's presence in all grids is set to 0.5 [25]: that is, $p^0(k_x, k_y) = 0.5$. This initialization is based on an assumption of equal likelihood, which indicates that the possibility of a target's presence in any grid is 50% without any prior information.

During the execution of search missions, UAVs employ onboard sensors to collect detection data. The probability of target presence $p^t(k_x, k_y)$ is dynamically updated based on new observational data [39]. Considering the sensor's detection probability P_{i,D_z} and false alarm probability P_{i,F_z}, the probability of a target's presence $p^{t+1}(k_x, k_y)$ in the grid observed by UAV i at time $t+1$ is updated using Bayes' theorem [40].

$$p_i^{t+1}(k_x, k_y) = \begin{cases} \frac{P_{i,D_z} p_i^t(k_x,k_y)}{P_{i,D_z} p_i^t(k_x,k_y) + P_{i,F_z}(1 - p_i^t(k_x,k_y))}, & b_i^t(k_x, k_y) = 1 \\ \frac{(1-P_{i,D_z}) p^i(k_x,k_y)}{(1-P_{i,D_z}) p_i^t(k_x,k_y) + (1-P_{i,F_z}(1-p_i^t(k_x,k_y))}, & b_i^t(k_x, k_y) = 0 \end{cases} \quad (1)$$

where $b_i^t(k_x, k_y)$ denotes the target detection result of UAV i for grid k at time t, with $b_i^t(k_x, k_y) = 1$ signifying that a target is detected within grid k and, conversely, the absence of a target is implied if the result is otherwise.

The value range of the uncertainty map $\chi^t(k_x, k_y)$ spans from $[0, 1]$ and serves to quantify the degree of knowledge that the UAV possesses regarding the information within grid k at time t. Within this framework, $\chi^t(k_x, k_y) = 0$ indicates that the UAV possesses complete knowledge of the information within grid k, whereas $\chi^t(k_x, k_y) = 1$ signifies that the UAV lacks any knowledge about the information within grid k. The uncertainty of the

grid is quantifiable through the information entropy associated with the probability of a target's presence within a grid, and the expression is given by [41]:

$$\chi_i^t(k_x, k_y) = -p_i^t(k_x, k_y)\log_2 p_i^t(k_x, k_y) - (1 - p_i^t(k_x, k_y))\log_2(1 - p_i^t(k_x, k_y)) \quad (2)$$

3.4. Information Fusion Model

During the process of UAVs collaboratively conducting search missions, each UAV independently maintains a local belief probability map. This strategy allows each UAV to construct a cognitive model of the environment based on its observations and experiences, enabling rapid responses in varied scenarios. However, to achieve swarm intelligence and improve the efficiency of collaborative operations, a mechanism is required to integrate these disparate local cognitions into a global and consistent belief probability map [42]. Drawing on the results of previous research [43], this paper introduces a distributed mechanism for updating local belief maps. The updating rule we employ is the minimum uncertainty rule, which prioritizes the information with the lowest uncertainty for updating the global belief map.

Specifically, each UAV updates its local belief probability map based on the observational data it collects throughout its mission. The UAVs share their local belief information via a predefined communication protocol at each time step. Then, each UAV refreshes its global belief probability map according to the received information, adhering to the minimum uncertainty rule. This process guarantees that each UAV can make decisions based on the latest and most accurate information, thus optimizing the path planning of the entire UAV swarm. The expression is as follows:

$$\chi^t(k_x, k_y) = \min\{\chi_1^t(k_x, k_y), \chi_2^t(k_x, k_y), ..., \chi_N^t(k_x, k_y)\} \quad (3)$$

Further, based on the relationship between the probability of target presence and uncertainty, a principle for updating the probability of target presence can be derived utilizing the minimum uncertainty update principle. The expression is as follows:

$$p^t(k_x, k_y) = p_I^t(k_x, k_y) \quad (4)$$

where $I = \arg\min_i \chi_i^t(k_x, k_y)$, $i = 1, 2, ..., N$.

3.5. Problem Formulation

In the scenario discussed in this paper, the definitions of "collision" and "collision avoidance" are as follows: A collision refers to the event where two or more UAVs make physical contact in space during the collaborative operation of multiple UAVs. Such an event may lead to UAV damage or mission failure, necessitating effective prevention strategies. Collision avoidance refers to using algorithms and strategies to prevent physical contact between multiple UAVs during collaborative operations. This usually involves three aspects: detecting the relative positions of UAVs, predicting potential collision risks, and executing avoidance maneuvers to prevent collisions.

The goal of this paper is to minimize the uncertainty of the search area and maximize the number of captured moving targets by optimizing the path planning strategy of UAVs while ensuring collision avoidance between UAVs. This goal of the cooperative search mission enables UAVs not only to explore unknown areas efficiently but also to locate moving targets precisely while maintaining a safe distance to avert collisions between each other.

The objective function of the problem is formulated as follows:

$$\min \quad f = \sum_{x=1}^{L_X}\sum_{x=1}^{L_Y}\chi^T(k_x, k_y) - \sum_{x=1}^{L_X}\sum_{y=1}^{L_Y} 1_{p^T(k_x,k_y)\geq\xi} \quad (5)$$

subject to:

$$0 \leq u_i^t(x) \leq L_X, \ \forall i \in \mathbb{N} \tag{6}$$

$$0 \leq u_i^t(y) \leq L_Y, \ \forall i \in \mathbb{N} \tag{7}$$

$$u_i^t(z) \in \{1, 2, 3\}, \ \forall i \in \mathbb{N} \tag{8}$$

$$\|\mathbf{u}_i^t - \mathbf{u}_j^t\| > d_{safe}, \ \forall i, j \in \mathbb{N}, \ i \neq j \tag{9}$$

$$P_{i,D_1} > P_{i,D_2} > P_{i,D_3}, \ \forall i \in \mathbb{N} \tag{10}$$

$$P_{i,F_1} < P_{i,F_2} < P_{i,F_3}, \ \forall i \in \mathbb{N} \tag{11}$$

$$V_{i,1} < V_{i,2} < V_{i,3}, \ \forall i \in \mathbb{N} \tag{12}$$

Here, T denotes the maximum execution time of the search task, $1_{p^T(k_x,k_y)\geq\xi}$ represents the number of grids with a target probability exceeding the predetermined threshold ξ, $\mathbf{u}_i^t$ and $\mathbf{u}_j^t$ denote the 3D coordinates of UAV i and UAV j, respectively, at time t, $u_i^t(x)$, $u_i^t(y)$, and $u_i^t(z)$ denote the x, y, and z coordinates, respectively, of UAV i at time t, and d_{safe} indicates the safe distance required to prevent collisions between UAVs. $V_{i,1}$, $V_{i,2}$, and $V_{i,3}$ denote the field-of-view sizes of low-level UAVs, mid-level UAVs, and high-level UAVs, respectively.

The function $f_1 = \sum_{x=1}^{L_X} \sum_{x=1}^{L_Y} \chi^T(k_x, k_y)$ denotes the regional uncertainty upon completion of the search task, and $f_2 = \sum_{x=1}^{L_X} \sum_{y=1}^{L_Y} 1_{p^T(k_x,k_y)\geq\xi}$ denotes the number of targets successfully detected at the conclusion of the search task. Constraints (6) and (7) ensure that the UAVs remain within the designated operational area throughout the search task. Constraint (8) indicates that UAVs can fly at three altitude levels. Constraint (9) mandates that UAVs avoid collisions with one another during task execution. Constraints (10) and (11) specify that UAVs with lower heights have higher detection probabilities and lower false alarm probabilities. Constraint (12) states that the lower the altitude, the smaller the field of view of the UAV.

4. Proposed Multi-UAV Cooperative Search Method for Moving Targets

This paper proposes a multi-UAV cooperative search method to solve the path planning problem for moving target search in a 3D environment. The proposed search method consists of a target capture mechanism, a collision avoidance mechanism, a state representation mechanism, and the AM-MAPPO algorithm. Specifically, the target capture mechanism determines whether to lower the UAVs' altitude to capture the targets based on their current flight altitudes and the obtained positional information of the targets. The collision avoidance mechanism calculates action masks for the subsequent moment based on the UAVs'positions, ensuring the UAVs avoid collisions during flight. The state representation mechanism encodes the existence state of targets, addressing dynamic information changes within the UAVs' fields of view. Then, the observations of all UAVs, the encoded target data within the UAVs' fields of view, and the action masks for collision avoidance are used as inputs for the AM-MAPPO algorithm, which outputs joint actions for all UAVs. Finally, path planning decisions for all UAVs are determined by the outputs of the target capture mechanism and the AM-MAPPO algorithm.

4.1. Rule-Based Target Capture Mechanism

This paper proposes a rule-based target capture mechanism to fully utilize the advantages of a broad field of view for high-altitude UAVs and high detection quality for

low-altitude UAVs. The proposed mechanism segments the search process of multiple UAVs into two stages: broad area coverage search and precise target capture. In the first stage, the UAV conducts broad area coverage tasks at higher flight levels to fully leverage the advantage of having a wide field of view at high altitudes. Once a potential target is detected, the UAV advances to the second stage, lowering the flight altitude to enhance detection probability for precise positioning and capture of the target. This mechanism seeks to balance the relationship between the detection range and the detection accuracy, thereby enhancing the overall efficiency of task execution.

This paper restricts the UAV's subsequent flight actions based on its current altitude and targets' position information. Specifically, the UAV's flight altitude $u_i^t(z)$ is categorized into three levels: high, medium, and low, each with a corresponding field of view $V_{i,z}$ of 9, 5, and 1. When a UAV is flying at a high or medium altitude and a target is suspected within its field of view, it is restricted to lowering its flight altitude, thereby transitioning from a broad area coverage search to precise target capture to enhance detection and capture accuracy. The mathematical model of this process is as follows:

$$\begin{cases} u_i^{t+1}(z) = 2 \quad if\ u_i^t(z) = 3\ and\ g_d^t \in F_{i,3}^t,\ \exists d \in \mathbb{D} \\ u_i^{t+1}(z) = 1 \quad if\ u_i^t(z) = 2\ and\ g_d^t \in F_{i,2}^t,\ \exists d \in \mathbb{D} \end{cases} \tag{13}$$

Here, g_d^t represents the position of target d at time t. $F_{i,2}^t$ and $F_{i,3}^t$ represent the fields of view of a middle-altitude UAV and a high-altitude UAV, respectively.

4.2. Action-Mask-Based Collision Avoidance Mechanism

In the multi-UAV collaborative search task, this paper introduces an efficient collision avoidance mechanism to mitigate the risk of collisions among UAVs. Specifically, an action mask mechanism is initiated when the distance d_{ij} between two UAVs i and j falls below the preset safe distance d_{safe}. This mechanism imposes restrictions on UAVs identified to have potential collision risks. In the subsequent decision-making cycle, these UAVs are constrained to selecting actions that maintain a safe distance from potential collision objects. This entails that upon the detection of a collision risk, the affected UAVs will be directed to move in a specific direction to augment their distance and guarantee the safety of the collaborative search process.

Let the positions of UAVs i and j at time t be $(u_i^t(x), u_i^t(y))$ and $(u_j^t(x), u_j^t(y))$, respectively. The set of permissible actions for the UAV comprises $\mathcal{A} = \{\text{North}, \text{East}, \text{South}, \text{West}, \text{Up}, \text{Down}\}$, and the corresponding action mask is $[1, 1, 1, 1, 1, 1]$, indicating that all actions are permissible. When the action mask is $[0, 1, 1, 1, 1, 1]$, this indicates that all actions except for moving north are permissible. We determine the permissible actions of UAVs at subsequent moments based on their relative positions and derive the corresponding action masks. The specific rules are as follows:

$$\begin{cases} \text{mask}_i^{t+1} = [1,1,0,0,1,1],\ \text{mask}_j^{t+1} = [0,0,1,1,1,1] & if\ u_i^t(x) > u_j^t(x)\ and\ u_i^t(y) > u_j^t(y) \\ \text{mask}_i^{t+1} = [0,1,1,0,1,1],\ \text{mask}_j^{t+1} = [1,0,0,1,1,1] & if\ u_i^t(x) > u_j^t(x)\ and\ u_i^t(y) < u_j^t(y) \\ \text{mask}_i^{t+1} = [1,0,0,1,1,1],\ \text{mask}_j^{t+1} = [0,1,1,0,1,1] & if\ u_i^t(x) < u_j^t(x)\ and\ u_i^t(y) > u_j^t(y) \\ \text{mask}_i^{t+1} = [0,0,1,1,1,1],\ \text{mask}_j^{t+1} = [1,1,0,0,1,1] & if\ u_i^t(x) < u_j^t(x)\ and\ u_i^t(y) < u_j^t(y) \\ \text{mask}_i^{t+1} = [1,1,1,0,1,1],\ \text{mask}_j^{t+1} = [1,0,1,1,1,1] & if\ u_i^t(x) > u_j^t(x)\ and\ u_i^t(y) = u_j^t(y) \\ \text{mask}_i^{t+1} = [1,0,1,1,1,1],\ \text{mask}_j^{t+1} = [1,1,1,0,1,1] & if\ u_i^t(x) < u_j^t(x)\ and\ u_i^t(y) = u_j^t(y) \\ \text{mask}_i^{t+1} = [1,1,0,1,1,1],\ \text{mask}_j^{t+1} = [0,1,1,1,1,1] & if\ u_i^t(x) = u_j^t(x)\ and\ u_i^t(y) > u_j^t(y) \\ \text{mask}_i^{t+1} = [0,1,1,1,1,1],\ \text{mask}_j^{t+1} = [1,1,0,1,1,1] & if\ u_i^t(x) = u_j^t(x)\ and\ u_i^t(y) < u_j^t(y) \end{cases} \tag{14}$$

4.3. Field-of-View-Encoding-Based State Representation Mechanism

In the study of multi-UAV collaborative target search using deep reinforcement learning algorithms, UAVs' observable targets' positions are typically incorporated into the neural network inputs. The presence or absence of targets in the UAVs' fields of view

results in dynamic changes in observable information, thus leading to dynamic changes in the neural networks' input dimensions. To effectively tackle the issue of changes in input dimensions due to the uncertainty of UAVs' perceived information, this study introduces a field-of-view-encoding-based state representation mechanism. In this mechanism, the grids within the UAVs' fields of view are classified as either containing or lacking targets and are encoded accordingly. Specifically, grids lacking targets are encoded as 0, while those containing targets are encoded as 1. We denote c_k to represent the existence state of the target within the grid k, i.e., $c_k \in \{0, 1\}$. This method concisely and effectively encapsulates the target information within the UAVs' fields of view, offering an accurate foundation for subsequent decision-making.

Furthermore, considering that the fields of view of UAVs vary across different flight altitude levels, this paper adopts a strategy of uniformly fixing the field-of-view range by representing object categories in terms of 3×3 grids. When the flight level of the UAV is 1 or 2, the object category of any grid exceeding its field of view is uniformly encoded as 0. This processing method ensures the consistency and stability of input data and effectively addresses the challenge of dynamic changes in the volume of information within the field of view. Figure 3 shows the specific implementation of the state representation mechanism. Firstly, based on the fields of view of UAVs at different heights, we obtain the projection results of the UAV's field of view in the 2D search area. Then, we place the UAV at the center of an area of 3×3 grids and encode the target existence attribute of grids within the field of view based on the UAV's target detection results. For grids beyond the field of view, the target existence attribute is encoded as 0.

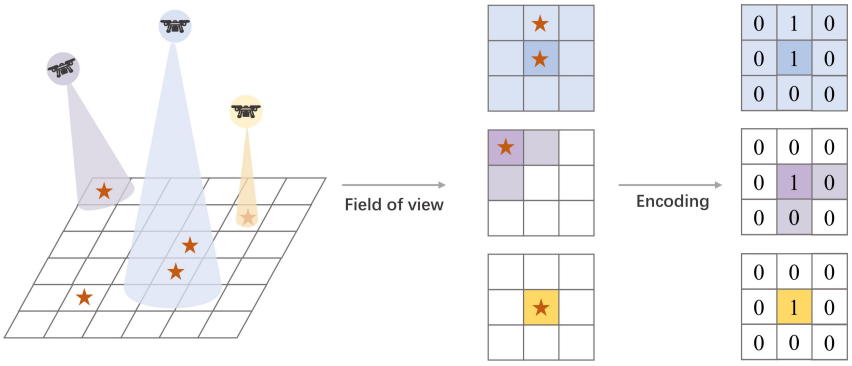

Figure 3. The state representation mechanism. The numbers 0 and 1 represent the UAV's belief that there is no target or a target is present in the grid, respectively.

4.4. DEC-POMDP Formulation

In the collaborative search task involving multiple UAVs, each UAV functions as an independent decision-making unit and is required to collaborate with others to achieve the dual goals of minimizing search area uncertainty and maximizing the number of captured moving targets. Considering the limitations of each UAV's field of view, which precludes it from observing the entire environment, UAVs must make decisions based on their local observations and the information communicated with other UAVs. Based on these considerations, this paper formulates the collaborative search task as a DEC-POMDP. Specifically, the core components of the model comprise:

(1) State space: The environmental state at time t is represented as:

$$\mathcal{S}^t \triangleq \{\mathcal{S}_p^t, \mathcal{S}_c^t, \mathcal{S}_\chi^t\} \tag{15}$$

where $\mathcal{S}_p^t = \{\mathbf{u}_i^t \mid i \in \mathbb{N}\}$ represents the UAVs' positions, $\mathcal{S}_c^t = \{c_k^t \mid k \in F_i, i \in \mathbb{N}\}$ denotes the existence status of targets within the UAVs' fields of view, and $\mathcal{S}_\chi^t = \{\chi_k^t \mid k \in \mathbb{K}\}$ signifies the uncertainty of all grids.

(2) Action space: The joint actions taken by the UAVs at time t are defined as:

$$\mathcal{A}^t \triangleq \{\mathcal{A}_1^t, \mathcal{A}_2^t, \ldots, \mathcal{A}_N^t\} \tag{16}$$

where the set of executable actions for UAV i comprises $\mathcal{A}_i^t$ = {North, East, South, West, Up, Down}.

(3) Observation space: The joint observations of the UAVs at time t are defined as follows:

$$\mathcal{O}^t \triangleq \{\mathcal{O}_1^t, \mathcal{O}_2^t, \ldots, \mathcal{O}_N^t\} \tag{17}$$

The observation space of UAV i comprises $\mathcal{O}_i^t = \{\mathcal{O}_p^t, \mathcal{O}_{i,c}^t, \mathcal{O}_{i,\chi}^t\}$, where, $\mathcal{O}_p^t$ denotes the 3D positions of all UAVs, $\mathcal{O}_{i,c}^t$ indicates the existence status of targets in the UAV's field of view, where 0 signifies the absence of objects in the grid and 1 signifies the presence of suspected targets in the grid, and $\mathcal{O}_{i,\chi}^t$ measures the uncertainty of the grid in the UAV's field of view.

(4) Reward function: This reward function comprises two components: target discovery reward and cognitive reward. Additionally, to underscore the efficacy of the collision avoidance mechanism proposed herein, algorithms that incentivize UAVs towards collision avoidance via reward functions are compared. Consequently, for comparative algorithms lacking the action-mask-based collision avoidance mechanism, collision penalties are employed to steer UAVs away from collisions.

- The target discovery reward aims to incentivize UAVs to locate additional targets, noting that the target reward is exclusive to the initial discovery by any UAV.

$$r_1^t = \sum_{i=1}^{N} \sum_{x=1}^{L_X} \sum_{y=1}^{L_Y} 1_{p_i^{t-1}(x,y)<\xi\leq p_i^t(x,y) \text{ and } p_j^{t-1}(x,y)<\xi(\forall j\neq i)} \tag{18}$$

- The cognitive reward encourages UAVs to achieve comprehensive area coverage. This reward is granted when the grid's uncertainty falls below a predetermined threshold τ for the first time.

$$r_2^t = \sum_{i=1}^{N} \sum_{x=1}^{L_X} \sum_{y=1}^{L_Y} 1_{\chi_i^{t-1}(x,y)<\tau\leq\chi_i^t(x,y)} \tag{19}$$

- The collision penalty serves to deter UAVs from colliding. It is applied when the inter-UAV distance falls below the safety threshold.

$$r_3^t = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1_{\|\mathbf{u}_i^t-\mathbf{u}_j^t\|\leq d_{safe}} \tag{20}$$

In summary, the comprehensive reward function is formulated as follows:

$$r^t = w_1 \times r_1^t + w_2 \times r_2^t + w_3 \times r_3^t \tag{21}$$

where w_1, w_2, and w_3 represent weight coefficients.

4.5. AM-MAPPO Algorithm

When the UAV is located at the boundary of the search area or faces potential risks of collisions with other UAVs, its optional actions are limited and can only be sampled from a reasonable set of actions. In response to the above situation, to ensure the convergence of model training, an action masking technique is introduced to update the policy network of MAPPO so that only effective actions are sampled [44].

The actor network architecture of the AM-MAPPO algorithm is shown in Figure 4. Firstly, the actor calculates each action's non-standardized scores (logits). Then, the Hadamard product of the action masks and logits are computed. Finally, the softmax

operation is used to convert the masked logits into the probability distribution of actions to determine the actions of the UAV.

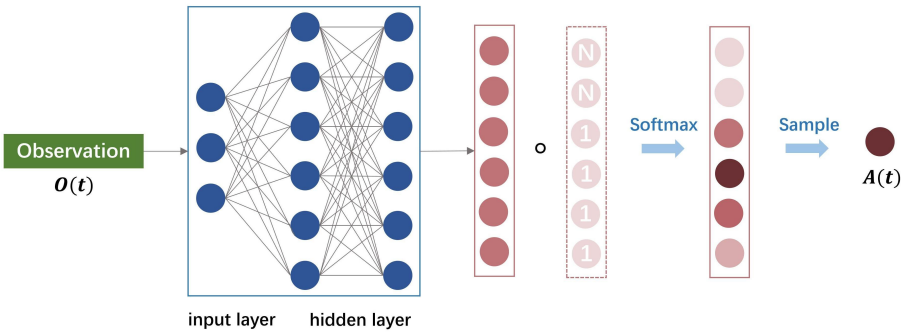

Figure 4. The actor network architecture of the AM-MAPPO algorithm.

(1) Action Selection Stage

Assuming UAV 1 is currently positioned at the boundary of the search area, with coordinates [19, 19], the set of feasible actions it can perform includes $\{2,3,4,5\}$. UAV 2 and UAV 3 currently face a risk of collision, with coordinates [6, 7] and [8, 9], respectively. Therefore, they can execute effective actions of $\{0,3,4,5\}$ and $\{1,2,4,5\}$, respectively. To simplify the description, we define the logits representing the actions for UAV 1, UAV 2, and UAV 3 output by the strategy π_θ as follows:

$$\text{logits}_1 = \left[l_{1,0},\ l_{1,1},\ l_{1,2},\ l_{1,3},\ l_{1,4},\ l_{1,5}\right] = \left[1,1,1,1,1,1\right] \tag{22}$$

$$\text{logits}_2 = \left[l_{2,0},\ l_{2,1},\ l_{2,2},\ l_{2,3},\ l_{2,4},\ l_{2,5}\right] = \left[1,1,1,1,1,1\right] \tag{23}$$

$$\text{logits}_3 = \left[l_{3,0},\ l_{3,1},\ l_{3,2},\ l_{3,3},\ l_{3,4},\ l_{3,5}\right] = \left[1,1,1,1,1,1\right] \tag{24}$$

where $l_{1,i}$, $l_{2,i}$, and $l_{3,i}$ represent the logits obtained by UAV 1, UAV 2, and UAV 3, respectively, when executing action i. Here, $i \in \{0,1,2,3,4,5\}$.

To implement action masks, we use a large negative number M to replace the logits of actions that need to be masked. For example, we can choose $M = -10^8$. The action masks for UAV 1, UAV 2, and UAV 3 at their current positions can be represented as follows:

$$\text{mask}_1 = \left[-10^8, -10^8, 1, 1, 1, 1\right] \tag{25}$$

$$\text{mask}_2 = \left[1, -10^8, -10^8, 1, 1, 1\right] \tag{26}$$

$$\text{mask}_3 = \left[-10^8, 1, 1, -10^8, 1, 1\right] \tag{27}$$

The masking process inv_s for the actions of UAV 1, UAV 2, and UAV 3 can be represented as:

$$\text{inv}_s(\text{mask}_1, \text{logits}_1) = \text{mask}_1 \circ \text{logits}_1 = \left[-10^8, -10^8, 1, 1, 1, 1\right] \tag{28}$$

$$\text{inv}_s(\text{mask}_2, \text{logits}_2) = \text{mask}_2 \circ \text{logits}_2 = \left[1, -10^8, -10^8, 1, 1, 1\right] \tag{29}$$

$$\text{inv}_s(\text{mask}_3, \text{logits}_3) = \text{mask}_3 \circ \text{logits}_3 = \left[-10^8, 1, 1, -10^8, 1, 1\right] \tag{30}$$

The probability distributions of UAV 1, UAV 2, and UAV 3 adopting different actions are represented as:

$$\begin{aligned} \pi_1(\cdot \mid s(t)) &= \text{Softmax}(\text{inv}_s(\text{mask}_1, \text{logits}_1)) \\ &= \text{Softmax}\left(\left[\tilde{l}_{1,0},\ \tilde{l}_{1,1},\ \tilde{l}_{1,2},\ \tilde{l}_{1,3},\ \tilde{l}_{1,4},\ \tilde{l}_{1,5}\right]\right) \\ &= \left[0,\ 0,\ 0.25,\ 0.25,\ 0.25,\ 0.25\right] \end{aligned} \tag{31}$$

$$\begin{aligned}\pi_2(\cdot \mid s(t)) &= \text{Softmax}(\text{inv}_s(\text{mask}_2, \text{logits}_2)) \\ &= \text{Softmax}\left(\left[\tilde{l}_{2,0},\ \tilde{l}_{2,1},\ \tilde{l}_{2,2},\ \tilde{l}_{2,3},\ \tilde{l}_{2,4},\ \tilde{l}_{2,5}\right]\right) \\ &= \left[0.25,\ 0,\ 0,\ 0.25,\ 0.25,\ 0.25\right]\end{aligned} \tag{32}$$

$$\begin{aligned}\pi_3(\cdot \mid s(t)) &= \text{Softmax}(\text{inv}_s(\text{mask}_3, \text{logits}_3)) \\ &= \text{Softmax}\left(\left[\tilde{l}_{3,0},\ \tilde{l}_{3,1},\ \tilde{l}_{3,2},\ \tilde{l}_{3,3},\ \tilde{l}_{3,4},\ \tilde{l}_{3,5}\right]\right) \\ &= \left[0,\ 0.25,\ 0.25,\ 0,\ 0.25,\ 0.25\right]\end{aligned} \tag{33}$$

where $\pi_i(\cdot \mid s(t)) = \frac{\exp(\tilde{l}_{i,j})}{\sum_j \exp(\tilde{l}_{i,j})}$.

Finally, UAV 1, UAV 2, and UAV 3 select their actions based on the values of $\pi_1(\cdot \mid s(t))$, $\pi_2(\cdot \mid s(t))$, and $\pi_3(\cdot \mid s(t))$. This process ensures that the probability of selecting invalid actions (masked with a large negative number M) is zero.

(2) Strategy Update Stage

The objective function of the AM-MAPPO algorithm is represented as:

$$J(\theta) = \mathbb{E}_t\left[min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right] \tag{34}$$

$$\hat{A}_t = \sum_{t<T} \gamma^t r(t) - V(s(t)) \tag{35}$$

$$r_t(\theta) = \frac{\pi_\theta(a(t) \mid s(t),\ \text{mask}(t))}{\pi_{\theta_{old}}(a(t) \mid s(t),\ \text{mask}(t))} \tag{36}$$

where $\mathbb{E}_t$ represents the expected reward at time step t, $\hat{A}_t$ represents the advantage function, $r_t(\theta)$ is the ratio between the current strategy and the original strategy, θ represents the trainable parameters of the strategy function, ε controls the clipping of the policy ratio, and $\pi_\theta(a(t) \mid s(t),\ \text{mask}(t))$ and $\pi_{\theta_{old}}(a(t) \mid s(t),\ \text{mask}(t))$ represent the current policy and the original policy, respectively, after the action masking.

Figure 5 shows the AM-MAPPO architecture for solving the path planning problem in collaborative target search tasks involving multiple UAVs. In this architecture, each agent corresponds to a UAV. Algorithm 1 provides a detailed description of the training process of the AM-MAPPO algorithm.

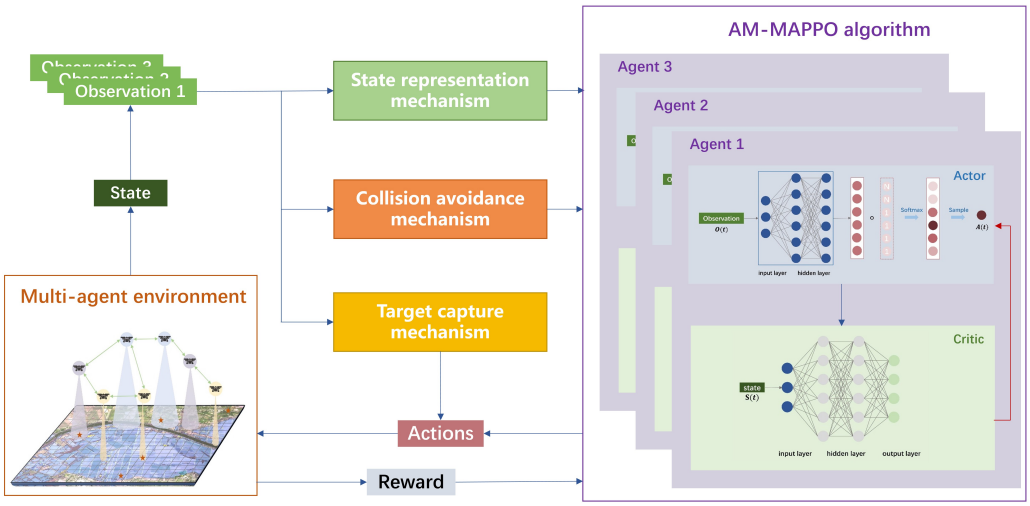

Figure 5. AM-MAPPO architecture for solving multi-UAV collaborative target search problem.

Algorithm 1: AM-MAPPO-based search algorithm

Input: number of training episodes, number of steps in one episode, learning rate, discount factor, clip factor, batch size, generalized advantage estimation lambda

Output: multiple UAVs' actions $\mathbf{a}^t$, actor network π_θ, critic network V_ϕ

Initialize: θ, the parameters for actor π_θ and ϕ, the parameters for critic V_ϕ, data buffer $D = \{\}$;

for *episode* = *1, 2, . . . , episodes* **do**
  initialize states: 3D positions of multiple UAVs, 2D positions and motion directions of targets ;
  set trajectory list $\tau = []$;
  for *step*= *1, 2, . . . , steps* **do**
    for i =*1, 2, . . . , N* **do**
      obtain observation o_i^t based on environmental state s^t and the state representation mechanism based on field-of-view encoding ;
      obtain mask_i^t based on 2D position of the UAV ;
      obtain altitude change instructions $flag$ based on the altitude of the UAV and the 2D positions of targets ;
      if $flag = 0$ **then**
        $f_i^t = \pi(o_i^t, \text{mask}_i^t; \theta)$;
        $a_i^t \sim f_i^t$;
      else
        $a_i^t = \text{Down}$;
      end
      $v_i^t = V(s^t; \phi)$;
    end
    execute actions $\mathbf{a}^t$, observe r^t, s^{t+1}, $\mathbf{o}^{t+1}$;
    update trajectory list $\tau += [s^t, \mathbf{o}^t, \mathbf{a}^t, r^t, s^{t+1}, \mathbf{o}^{t+1}]$;
  end
  Compute advantage estimate $\hat{A}_t$;
  Compute $\nabla_\theta L^{CLIP}$ and update θ ;
  Compute $\nabla_\phi L^{CLIP}$ and update ϕ ;
end

5. Performance Analysis

To evaluate the effectiveness of the proposed multi-UAV collaborative moving target search method, this section constructs a simulation environment for multi-UAV collaborative search. First, the effectiveness and generalization of the improved algorithm across different scenarios are verified through simulation experiments. Second, ablation experiments are designed to evaluate the specific effects of the proposed enhancement mechanism. Finally, to evaluate the AM-MAPPO algorithm's superiority comprehensively, the performances of other multi-agent reinforcement learning algorithms are systematically compared.

5.1. Parameter Setting

We consider a rectangular search area of 2000 m × 2000 m, which is further discretized into 20 × 20 grids. When multiple UAVs begin to perform search tasks, the prior information of the region is zero; thus, the probability of the target's existence in each grid is 0.5, and the uncertainty is 1, i.e., $p^0 = 0.5, \chi^0 = 1$. At the initial moment, the 2D positions of multiple UAVs and targets are randomly distributed within the search area. The heights of each UAV are randomly set to one of three values: 1, 2, or 3, representing low, middle, and high levels, respectively. The parameters of the UAVs' sensors are shown in Table 2 [45]. The search speed of the UAVs is 10 m/s, while the speed of the moving targets

is 1 m/s. In addition, in different simulation experiments, the number of UAVs is set to 3, 5, or 8, while the number of moving targets is set to 10, 15, 20, or 25 to demonstrate the algorithm's generalizability. The number of training episodes for the algorithm is set to 7000, and the simulation time for each episode is 5000 s. The simulation time is discretized into 500 planning steps, with a planning interval of 10 s.

The neural network architecture of the AM-MAPPO algorithm consists of fully connected layers, including an input layer, two hidden layers, and an output layer. The number of neurons in the input layer of the actor network corresponds to the elements in the observation space of a single UAV, while the number of neurons in the output layer corresponds to the elements in the UAV's action space. The input layer of the critic network has a number of neurons equal to the sum of the elements in the joint observation space of multiple UAVs, with the output layer comprising a single neuron. Both the actor and critic networks have 64 neurons in their hidden layers. The remaining simulation parameters for the simulation experiment are shown in Table 3. When setting the weight coefficients of the reward function, we thoroughly consider the value ranges of different reward and penalty terms. Consequently, we multiply them by distinct weight coefficients to ensure that the contributions of each reward and penalty term to the total reward are of comparable magnitude.

Table 2. Parameters of UAVs' sensors.

Flight Altitude	Field-of-View Size	Detection Probability	False Alarm Probability
1	1	0.90	0.10
2	5	0.80	0.20
3	9	0.70	0.30

Table 3. Simulation parameters.

Parameter	Value	Parameter	Value
Threshold (τ_p)	0.99	Discount factor (γ)	0.99
Safe distance (d_{safe})	141.4 m	Batch size (B)	1024
Target capture reward weight (w_1)	1	Learning rate (L_r)	5×10^{-4}
Cognitive reward weight (w_2)	0.1	Clip factor (ε)	0.2
Collision penalty weight (w_3)	0.01	Gae lambda	0.95

5.2. Performance Indicators

We evaluate the performances of algorithms using four indicators: the number of captured targets, the number of covered grids, the average uncertainty of the search area, and the reward obtained by the algorithm. It should be noted that the experimental data represent the average of multiple episodes. Therefore, the number of captured targets and the number of covered grids are presented as precise decimals. The detailed descriptions are as follows:

- Number of captured targets: This indicator evaluates the number of targets successfully captured by the algorithm within a given time. A higher number indicates greater efficiency in target capture.
- Number of covered grids: This indicator focuses on the specific number of grids in the search area where the uncertainty is reduced to a certain level or below. It reflects the coverage effect of multiple UAVs in specific areas. A higher number of low-uncertainty grids indicates better algorithm performance in these regions, reducing search blind spots.
- Average uncertainty: This indicator measures the degree to which the algorithm reduces the uncertainty of the entire region during the search process, reflecting the overall mastery of environmental information by all UAVs. A lower average uncertainty indicates more comprehensive and effective coverage of the search area by the UAVs.

- Reward: This indicator considers the total reward obtained by the algorithm throughout the entire task process. A higher reward typically indicates better performance in balancing broad area coverage and precise target capture.

The number of covered grids and the average uncertainty evaluate the effectiveness of multi-UAV collaborative search methods from two perspectives: coverage range and information mastery level. Although there is correlation between these two indicators, they provide different perspectives that help to evaluate the algorithm's performance comprehensively. For example, lower average uncertainty indicates good overall coverage. However, if the number of covered grids is small, it may indicate the presence of redundant searches and search blind spots in local areas. Therefore, combining these two indicators can provide a more comprehensive evaluation of the effectiveness of the multi-UAV collaborative search method.

5.3. Performance Evaluation

5.3.1. Effectiveness Analysis of the AM-MAPPO Algorithm

A. The Effectiveness Analysis for Different Numbers of UAVs

To evaluate the impact of the number of intelligent agents on the performance of the AM-MAPPO algorithm, this section fixes the number of targets in the search scenario to 10 and sets the number of UAVs to 3, 5, and 8, respectively. Simulation experiments are conducted for these three different search scenarios.

Figure 6 illustrates the search performance of the AM-MAPPO algorithm for three scenarios with varying numbers of UAVs. As the number of UAVs increases, the number of captured targets and the number of covered grids by all UAVs increase. Simultaneously, the average uncertainty of the search area decreases. The average reward increases when the number of UAVs increases from 3 to 5. However, the average reward decreases when the number of UAVs increases to 8. The reasons for the trend above are explained as follows.

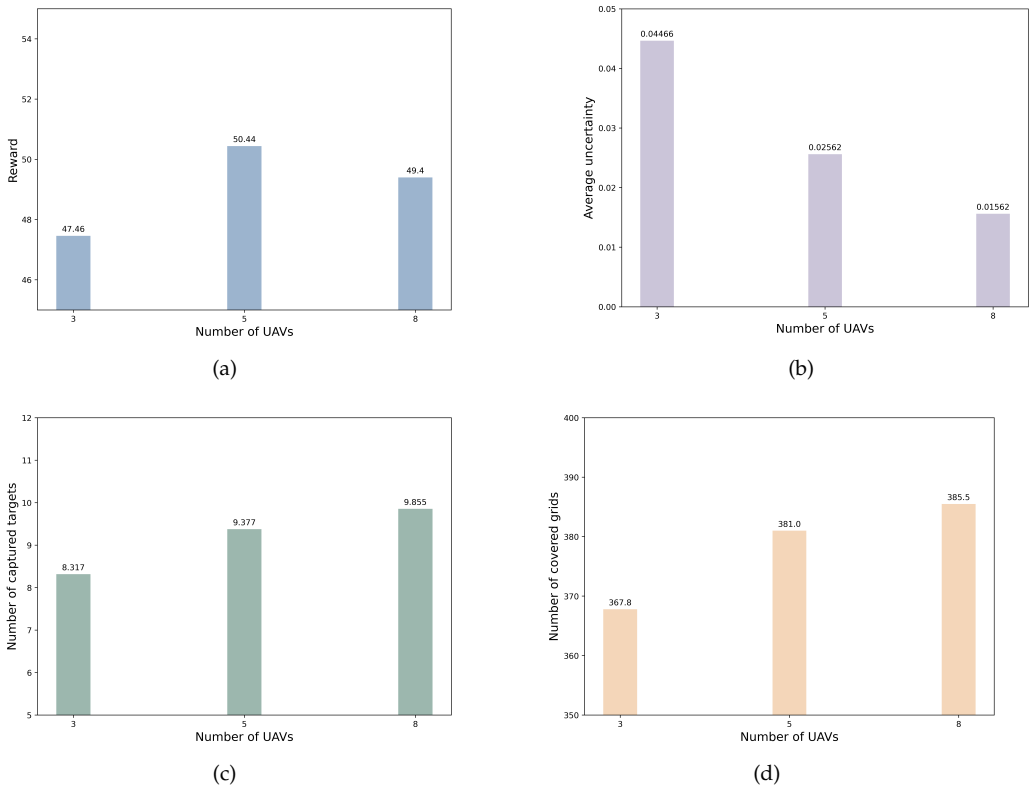

Figure 6. The effectiveness analysis for different numbers of UAVs.(**a**) Reward. (**b**) Average uncertainty. (**c**) Number of captured targets. (**d**) Number of covered grids.

Firstly, an increase in the number of UAVs involved in search tasks implies that more sensors are used for target detection and recognition, resulting in a higher number of captured targets. This indicates that increasing the number of UAVs can enhance the efficiency and success rate of target capture. Secondly, as the number of UAVs increases, the area UAVs can cover expands. More UAVs can scan and monitor additional grids, thereby increasing the number of covered grids and significantly reducing the uncertainty of the overall search area. Finally, the change in reward suggests that a balance should be found between the number of UAVs and task efficiency to avoid the negative impact of resource waste and increased conflicts. An increased number of UAVs operating within the same search area may lead to resource competition and path planning conflicts, thereby reducing overall task efficiency. Furthermore, increasing the number of UAVs may lead to gradually decreasing marginal benefits and even negative benefits. As shown by the above analysis, the AM-MAPPO search method effectively enhances the performance of search tasks within a specific range.

B. The Effectiveness Analysis for Different Numbers of Targets

To explore the impact of the number of moving targets on algorithm performance, this section fixes the number of UAVs in the search scenario to five and sets the number of targets to 10, 15, 20, and 25, respectively. Simulation experiments are conducted for these four different search scenarios.

Figure 7 illustrates the search performance of the AM-MAPPO algorithm for four scenarios with varying numbers of targets. As the number of moving targets increases, the number of targets captured by all UAVs also increases. The reward increases, but the number of covered grids decreases, and the average uncertainty of the search area increases. The reasons for the trend above are explained as follows.

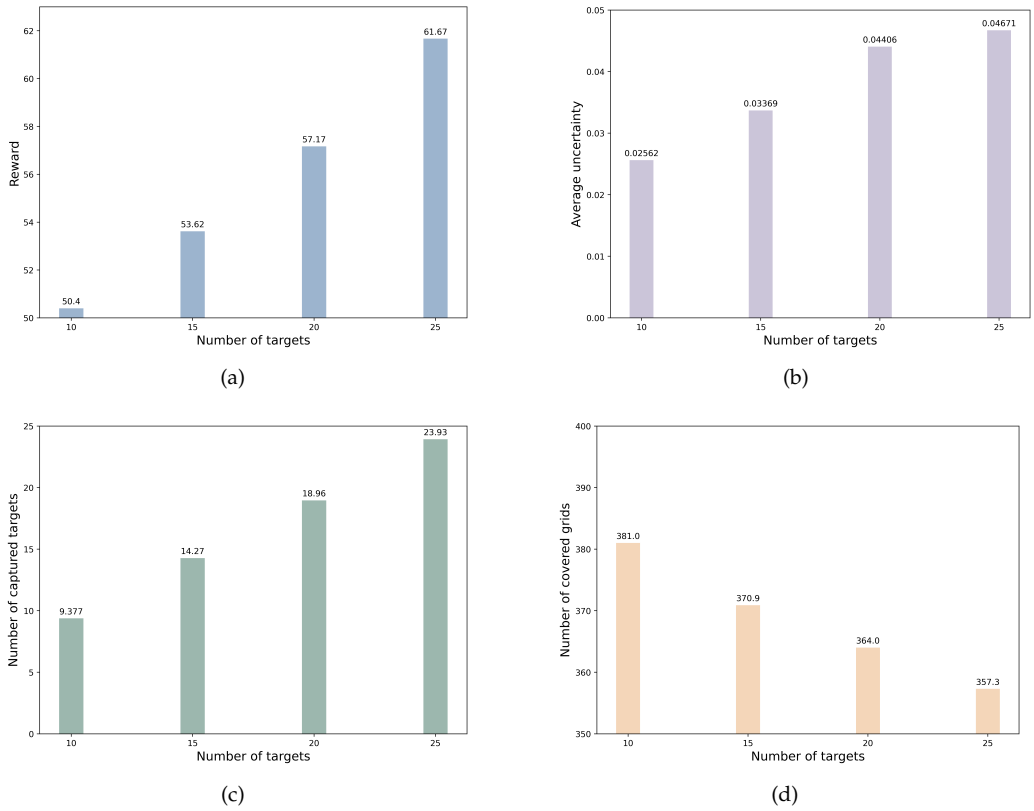

Figure 7. The effectiveness analysis for different numbers of targets. (**a**) Reward. (**b**) Average uncertainty. (**c**) Number of captured targets. (**d**) Number of covered grids.

Firstly, as the number of targets increases, UAVs have more targets to capture in search tasks, thus increasing the number of captured targets. This indicates that the search method proposed in this paper can still effectively recognize and capture targets in dense target scenarios. Secondly, the increase in the number of targets may lead to UAVs focusing more on dense target areas during coverage searches, thereby ignoring searches in other areas, resulting in an increase in the overall search area uncertainty. Meanwhile, due to the concentration of UAVs in dense target grids for precise target capture, the number of covered grids decreases. This indicates that the efficiency of broad area coverage search has decreased in dense target scenarios. However, it also reflects the strategy of all UAVs prioritizing target capture: that is, capturing targets with higher priority than regional coverage. In addition, due to the weight coefficient of the target discovery reward being much greater than that of the regional coverage reward, the reward increases as the number of captured targets increases and the number of covered grids decreases. From the above analysis, the AM-MAPPO search method can still effectively complete search tasks in dense target scenarios, demonstrating high target capture efficiency and task rewards, verifying the effectiveness and generalization of this method.

5.3.2. Trajectory Analysis

Tables 4–8 present the movement trajectories of five UAVs and their successfully captured targets over ten consecutive steps within one episode, as well as the operational status of the target capture mechanism. In the tables, "Y" indicates that the target capture mechanism is operational. The UAV is compelled to perform a "Down" action. In contrast, "N" indicates that the target capture mechanism is inactive. The flight actions of the UAV are determined by the actor network's output using the AM-MAPPO algorithm. The data indicate that UAV1, UAV2, UAV3, UAV4, and UAV5 successfully captured targets at steps 33, 103, 174, 191, and 203, respectively. At step 30, UAV1, flying at high altitude, detected a suspected target within its field of view. At this point, the rule-based target capture mechanism compels UAV1 to lower its flight altitude to improve target detection accuracy, ultimately resulting in successful target capture. Similarly, UAV2, UAV3, UAV4, and UAV5 activated the rule-based target capture mechanism at steps 100, 172, 188, and 200, respectively, and successfully captured targets. The flight trajectories of the UAVs indicate that when a high-altitude UAV detects a target within its field of view, it quickly lowers its altitude, thereby enhancing target capture performance.

Table 4. The process of UAV 1 capturing Target 1.

Step	Location of UAV 1	Location of Target 1	Working Status of the Target Capture Mechanism
24	(2,12,1)	(0,8,0)	N
25	(1,12,1)	(0,8,0)	N
26	(0,12,1)	(0,8,0)	N
27	(0,11,1)	(0,8,0)	N
28	(0,10,1)	(0,8,0)	N
29	(0,10,2)	(0,8,0)	N
30	(0,9,2)	(0,8,0)	Y
31	(0,9,1)	(0,8,0)	Y
32	(0,9,0)	(0,8,0)	N
33	(0,8,0)	(0,8,0)	N

Table 5. The process of UAV 2 capturing Target 2.

Step	Location of UAV 2	Location of Target 2	Working Status of the Target Capture Mechanism
94	(10,9,2)	(13,8,0)	N
95	(10,9,1)	(13,8,0)	N
96	(11,9,1)	(13,8,0)	N
97	(12,9,1)	(13,8,0)	N
98	(12,9,2)	(13,8,0)	N
99	(12,9,1)	(13,8,0)	N
100	(12,9,2)	(13,8,0)	Y
101	(12,9,1)	(13,9,0)	Y
102	(12,9,0)	(13,9,0)	N
103	(13,9,0)	(13,9,0)	N

Table 6. The process of UAV 3 capturing Target 3.

Step	Location of UAV 3	Location of Target 3	Working Status of the Target Capture Mechanism
165	(17,6,1)	(15,3,0)	N
166	(17,5,1)	(15,3,0)	N
167	(17,4,1)	(15,3,0)	N
168	(17,3,1)	(15,3,0)	N
169	(17,2,1)	(15,3,0)	N
170	(17,1,1)	(15,3,0)	N
171	(17,2,1)	(16,3,0)	N
172	(17,3,1)	(16,3,0)	Y
173	(17,3,0)	(16,3,0)	N
174	(16,3,0)	(16,3,0)	N

Table 7. The process of UAV 4 capturing Target 4.

Step	Location of UAV 4	Location of Target 4	Working Status of the Target Capture Mechanism
182	(12,9,1)	(9,9,0)	N
183	(11,9,1)	(9,9,0)	N
184	(11,8,1)	(9,9,0)	N
185	(11,8,0)	(9,9,0)	N
186	(11,8,1)	(9,9,0)	N
187	(11,8,2)	(9,9,0)	N
188	(10,8,2)	(9,9,0)	Y
189	(10,8,1)	(9,9,0)	Y
190	(9,8,1)	(9,9,0)	N
191	(9,8,0)	(9,8,0)	N

Table 8. The process of UAV 5 capturing Target 5.

Step	Location of UAV 5	Location of Target 5	Working Status of the Target Capture Mechanism
194	(16,2,1)	(16,7,0)	N
195	(16,3,1)	(16,7,0)	N
196	(16,4,1)	(16,7,0)	N
197	(16,5,1)	(16,7,0)	N
198	(16,4,1)	(16,7,0)	N
199	(16,5,1)	(16,7,0)	N
200	(16,6,1)	(16,7,0)	Y
201	(16,6,0)	(16,8,0)	N
202	(16,7,0)	(16,8,0)	N
203	(16,8,0)	(16,8,0)	N

5.3.3. Ablation Experiment

In order to explore the impact of various enhancement mechanisms on model performance, different enhancement mechanisms are removed from the AM-MAPPO algorithm to generate four algorithms: namely, NAM-NCA (no action mask and no collision avoidance), NTC (no target capture), NCA (no collision avoidance), and the benchmark algorithm MAPPO [46]. We evaluate the effectiveness of various mechanisms by comparing the performance of these algorithms. The specific settings of the four algorithms in the ablation experiment are shown in Table 9. Due to the collision avoidance mechanism being implemented based on action masks, removing the action mask mechanism from the policy network of the AM-MAPPO algorithm also affects the collision avoidance mechanism.

Table 9. Ablation experiment settings. "●" represents including the mechanism, and "○" represents not including the mechanism.

Algorithm	Action Mask Mechanism	Target Capture Mechanism	Collision Avoidance Mechanism
AM-MAPPO	●	●	●
NAM-NCA	○	●	○
NTC	●	○	●
NCA	●	●	○
MAPPO	○	○	○

Figure 8 illustrates the performance comparison curves of different algorithms in the ablation experiment. Based on MAPPO, the proposed AM-MAPPO introduces action mask technology into the network architecture while taking into account both a rule-based target capture mechanism and an action-mask-based collision avoidance mechanism. As shown in Figure 8a, the AM-MAPPO algorithm achieves the highest reward among all five. The benchmark MAPPO obtains the lowest reward and performs the worst. Figure 8b illustrates that with respect to reducing the average uncertainty, the performance of NTC is the best, followed by MAPPO, AM-MAPPO, NCA, and NAM-NCA. Figure 8d illustrates the number of grids with uncertainty below the set threshold, and the data further confirm the results of Figure 8b. Figure 8c illustrates that in terms of the number of captured moving targets, the performance of AM-MAPPO is the best, followed by NCA, NAM-NCA, NTC, and MAPPO.

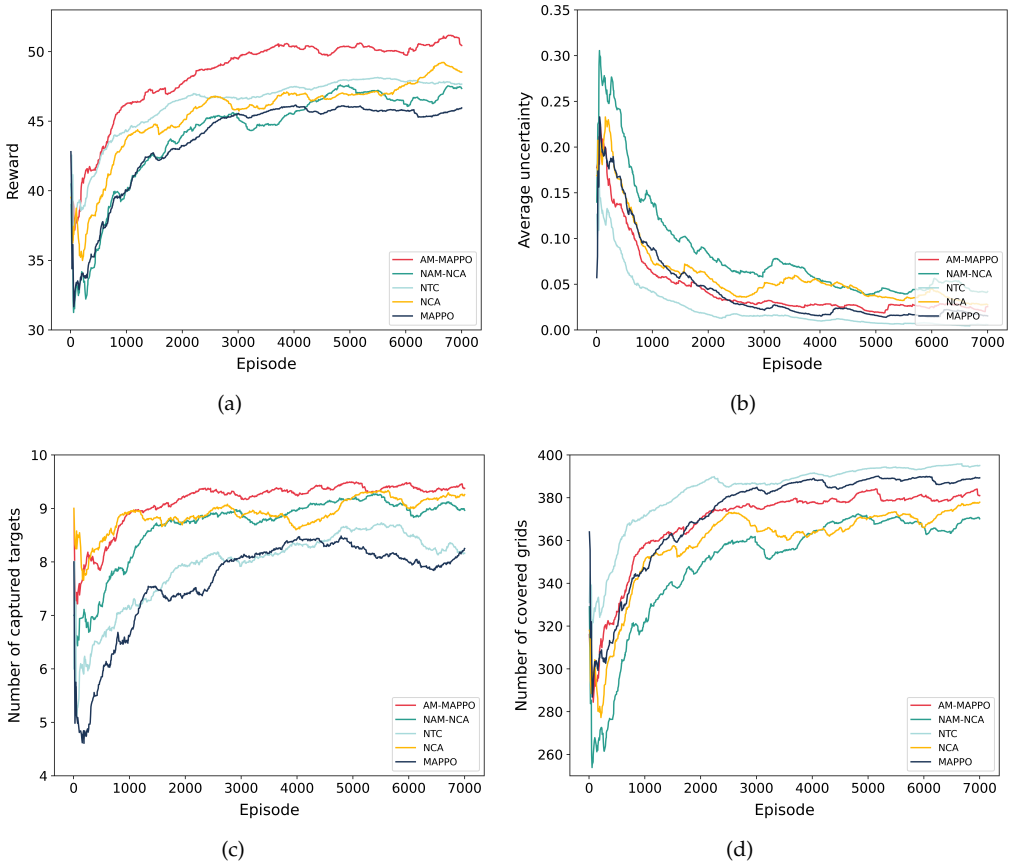

Figure 8. Ablation experiment. (**a**) Reward. (**b**) Average uncertainty. (**c**) Number of captured targets. (**d**) Number of covered grids.

The experimental results indicate that although NTC has a significant advantage in reducing uncertainty, its ability to capture targets is weak. The reason is that in the multi-UAV collaborative search architecture proposed in this paper, high-altitude UAVs have the advantage of covering a wide area, while low-altitude UAVs have the advantage of accurately capturing targets. The rule-based target capture mechanism can lead to UAVs choosing to lower their altitude to accurately capture targets when there are targets in their field of view, thereby affecting their coverage of the area and further affecting the degree of uncertainty reduction in the search area. Therefore, compared to algorithms with target capture mechanisms, such as AM-MAPPO, NAM-NCA, and NCA, NTC sacrifices the performance of the number of captured targets in exchange for the ability to reduce regional uncertainty. It should be noted that in the scenario of this paper, the priority of accurately capturing targets is higher than that of a broad area coverage search. Therefore, the experimental results verify the effectiveness of the target capture mechanism proposed in this paper. By comparing the performance of the AM-MAPPO and the NTC algorithms, it can be found that the rule-based target capture mechanism has significant advantages in reducing the action space dimension. This mechanism compels UAVs to perform specific actions when detecting suspicious targets, thereby improving target capture efficiency. In contrast, the NTC algorithm requires the agent to autonomously learn how to capture targets by lowering its flight altitude, resulting in a larger state and action space dimension and slower convergence speed. Therefore, the target capture mechanism proposed in this paper not only accelerates the convergence speed of the algorithm but also increases the number of captured targets, demonstrating its potential in practical applications.

In terms of four performance indicators, the performance of NCA is slightly inferior to AM-MAPPO, which verifies the effectiveness of the collision avoidance mechanism based on action masks proposed in this paper. In addition, the proposed collision avoidance

mechanism eliminates the risk of collisions between UAVs by avoiding ineffective actions, while NCA guides UAVs to avoid collisions through a reward function. Although the NCA algorithm can effectively reduce the risk of collision, its safety is not as good as the AM-MAPPO proposed in this paper. The performance of NAM-NCA is worse than that of NCA, which verifies the effectiveness of the action-mask-based network architecture designed in this paper. The action mask mechanism reduces the action space dimension of the AM-MAPPO algorithm by avoiding invalid actions when the UAV is at the boundary of the search area or there is a potential risk of collision with other UAVs, thereby improving the convergence of the algorithm. The ablation experiment verifies that the three enhancement mechanisms introduced in this paper can improve the effectiveness, security, and target capture ability of the AM-MAPPO algorithm.

5.3.4. Comparison with Other Algorithms

To evaluate the performance of the proposed AM-MAPPO algorithm, we compare it with three benchmark algorithms: MAPPO, VDN, and QMIX. When selecting the benchmark algorithm, we primarily consider the type of action space and the adaptability of the algorithm. As the Markov decision process established in this paper has a discrete action space, we select multi-agent reinforcement learning algorithms that are suitable for discrete action spaces as the comparative algorithms. Although algorithms such as MADDPG and COMA, mentioned in the literature, perform well in continuous action spaces, they are unsuitable for the optimization problem addressed in this paper. Furthermore, although some improved algorithms, such as DNQMIX and MNF-HDQN, perform well in specific scenarios, their adaptability to the search scenarios discussed in this paper is relatively limited. Therefore, this paper selects mainstream multi-agent reinforcement learning algorithms as benchmark algorithms to ensure the effectiveness and reliability of the comparative analysis. The benchmark algorithms reduce the risk of UAV collisions through the reward function term of collision avoidance. The following is a detailed description of each algorithm:

- MAPPO: a search method for multi-UAVs based on multi-agent proximal policy optimization [47]. Each UAV is considered an intelligent agent for deploying actor and critic network models. The structure of the critic network is the same as the settings of the AM-MAPPO algorithm. The architecture of the actor network is similar to the AM-MAPPO algorithm, but it does not include the action mask mechanism designed in this paper.
- VDN [48]: a search method for multi-UAVs based on a value decomposition network. Each UAV is considered an intelligent agent for deploying Q-network models. The network architecture of VDN is similar to the critic network of the AM-MAPPO algorithm. Its input is the local observation of the intelligent agent, and the global Q value is formed by adding up the local Q values of all intelligent agents.
- QMIX [49]: a search method for multi-UAVs based on a hybrid Q-network. Each UAV is considered an intelligent agent for deploying the Q-network model. The network architecture of QMIX is similar to the critic network of the AM-MAPPO algorithm. Its input is the local observation of the intelligent agent, which forms a global Q value by nonlinearly combining the local Q values of all intelligent agents through a hybrid network.

Figure 9 shows the performance of different multi-agent reinforcement learning algorithms. The experimental results indicate that the AM-MAPPO algorithm performs the best in terms of reward and the number of captured moving targets. Regarding reducing the uncertainty of the search area, its performance is slightly inferior to MAPPO and QMIX. Regarding the number of covered grids, its performance is second only to that of MAPPO. Specifically, combining the data from Figure 9b,d, QMIX shows the most significant performance in reducing regional uncertainty. However, it covers the least number of grids, indicating that QMIX has redundant coverage of some grids. Based on the experimental data in Section 5.3.3, compared to MAPPO, AM-MAPPO, with its target capture mecha-

nism, sacrifices broad area coverage capability to improve the performance of the number of captured targets. In summary, compared to other comparative algorithms, AM-MAPPO has the optimal search efficiency and target capture capability and can effectively reduce the uncertainty of the search area and efficiently perform moving target search tasks in three-dimensional space.

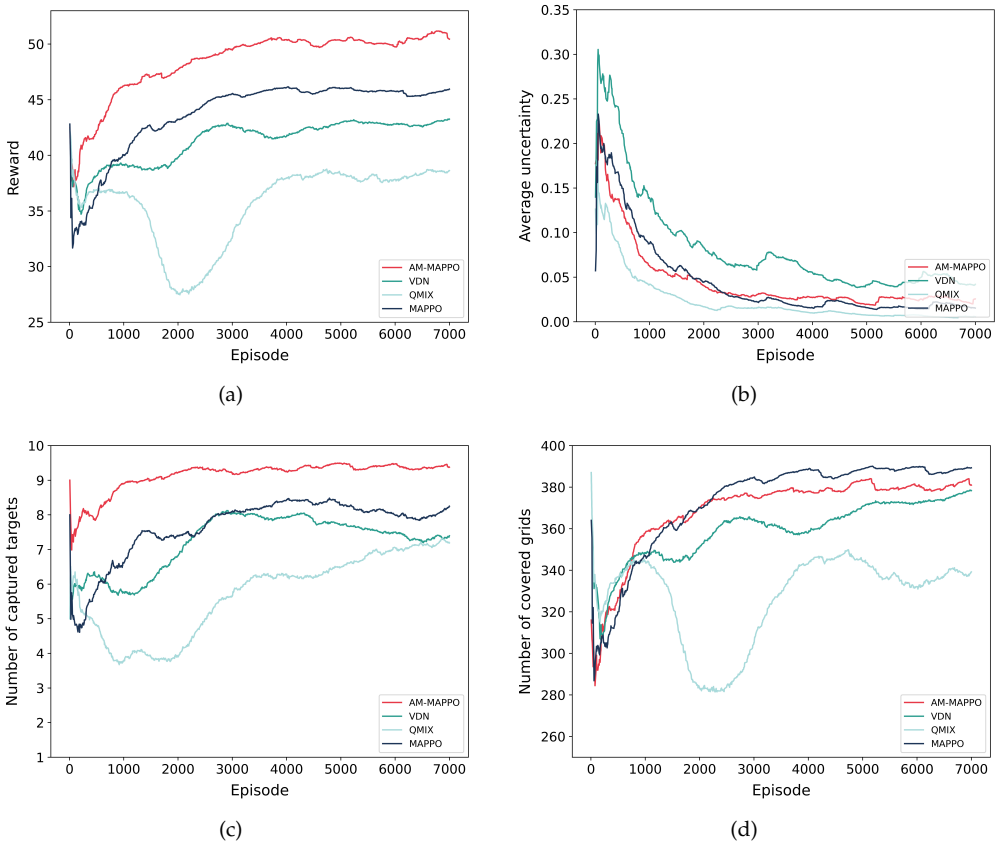

Figure 9. Performances of different multi-agent reinforcement learning algorithms. (**a**) Reward. (**b**) Average uncertainty. (**c**) Number of captured targets. (**d**) Number of covered grids.

6. Conclusions

This paper investigates the path planning problem of multiple UAVs searching for moving targets in a 3D environment and proposes an AM-MAPPO algorithm. A collaborative search architecture is designed for high-altitude and low-altitude UAVs, leveraging the broad field of view of high-altitude UAVs and the high detection quality of low-altitude UAVs, thereby improving the cooperative search efficiency of multiple UAVs. In addition, we introduce a field-of-view-encoding-based state representation mechanism to handle the dynamic changes in the input dimensions of neural networks. We design a rule-based target capture mechanism and an action-mask-based collision avoidance mechanism to effectively reduce the dimension of the action space, thereby improving the convergence speed of the AM-MAPPO algorithm. Numerical results show that the AM-MAPPO algorithm significantly outperforms other existing reinforcement learning algorithms in moving target search tasks. Ablation experiments verify the effectiveness of the action mask mechanism, target capture mechanism, and collision avoidance mechanism designed in this paper, which improves the convergence speed, target capture ability, and search safety of the search method based on the AM-MAPPO algorithm.

In this study, we consider the scenario where the target moves at a constant speed. Future research will expand to scenarios involving targets with varying speeds. In the next

phase of our work, we plan to design a method for multi-UAV collaborative search for moving targets that can effectively track and respond to changes in target speed.

Author Contributions: Conceptualization, Y.L., X.L., J.W. and J.Y.; methodology, Y.L.; software, Y.L. and X.L.; validation, F.W.; formal analysis, Y.L. and X.L; investigation, Y.L., X.L. and J.W.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number [62201601].

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Muchiri, G.; Kimathi, S. A review of applications and potential applications of UAV. In Proceedings of the Sustainable Research and Innovation Conference, Pretoria, South Africa, 20–24 June 2022; pp. 280–283.
2. Hu, X.; Assaad, R.H. The use of unmanned ground vehicles and unmanned aerial vehicles in the civil infrastructure sector: Applications, robotic platforms, sensors, and algorithms. *Expert Syst. Appl.* **2023**, *232*, 120897. [CrossRef]
3. Kats, V.; Levner, E. Maximizing the average environmental benefit of a fleet of drones under a periodic schedule of tasks. *Algorithms* **2024**, *17*, 283. [CrossRef]
4. Baniasadi, P.; Foumani, M.; Smith-Miles, K.; Ejov, V. A transformation technique for the clustered generalized traveling salesman problem with applications to logistics. *Eur. J. Oper. Res.* **2020**, *285*, 444–457. [CrossRef]
5. He, H.; Yuan, W.; Chen, S.; Jiang, X.; Yang, F.; Yang, J. Deep reinforcement learning based distributed 3D UAV trajectory design. *IEEE Trans. Commun.* **2024**, *72*, 3736–3751. [CrossRef]
6. Frattolillo, F.; Brunori, D.; Iocchi, L. Scalable and cooperative deep reinforcement learning approaches for multi-UAV systems: A systematic review. *Drones* **2023**, *7*, 236. [CrossRef]
7. Lyu, M.; Zhao, Y.; Huang, C.; Huang, H. Unmanned aerial vehicles for search and rescue: A survey. *Remote Sens.* **2023**, *15*, 3266. [CrossRef]
8. Qi, S.; Lin, B.; Deng, Y.; Chen, X.; Fang, Y. Minimizing maximum latency of task offloading for multi-UAV-assisted maritime search and rescue. *IEEE Trans. Veh. Technol.* **2024**, 1–14. [CrossRef]
9. Zhu, W.; Li, L.; Teng, L.; Yonglu, W. Multi-UAV reconnaissance task allocation for heterogeneous targets using an opposition-based genetic algorithm with double-chromosome encoding. *Chin. J. Aeronaut.* **2018**, *31*, 339–350.
10. Kim, T.; Lee, S.; Kim, K.H.; Jo, Y.I. FANET routing protocol analysis for Multi-UAV-based reconnaissance mobility models. *Drones* **2023**, *7*, 161. [CrossRef]
11. Li, K.; Yan, X.; Han, Y. Multi-mechanism swarm optimization for multi-UAV task assignment and path planning in transmission line inspection under multi-wind field. *Appl. Soft Comput.* **2024**, *150*, 111033. [CrossRef]
12. Lu, F.; Jiang, R.; Bi, H.; Gao, Z. Order distribution and routing optimization for takeout delivery under drone–rider joint delivery mode. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 774–796. [CrossRef]
13. Lu, F.; Chen, W.; Feng, W.; Bi, H. 4PL routing problem using hybrid beetle swarm optimization. *Soft Comput.* **2023**, *27*, 17011–17024. [CrossRef]
14. Yahia, H.S.; Mohammed, A.S. Path planning optimization in unmanned aerial vehicles using meta-heuristic algorithms: A systematic review. *Environ. Monit. Assess.* **2023**, *195*, 30. [CrossRef] [PubMed]
15. Aljalaud, F.; Kurdi, H.; Youcef-Toumi, K. Bio-inspired multi-UAV path planning heuristics: A review. *Mathematics* **2023**, *11*, 2356. [CrossRef]
16. Wang, X.; Fang, X. A multi-agent reinforcement learning algorithm with the action preference selection strategy for massive target cooperative search mission planning. *Expert Syst. Appl.* **2023**, *231*, 120643. [CrossRef]
17. Yu, X.; Luo, W. Reinforcement learning-based multi-strategy cuckoo search algorithm for 3D UAV path planning. *Expert Syst. Appl.* **2023**, *223*, 119910. [CrossRef]
18. Bai, Y.; Zhao, H.; Zhang, X.; Chang, Z.; Jäntti, R.; Yang, K. Towards autonomous multi-UAV wireless network: A survey of reinforcement learning-based approaches. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 3038–3067. [CrossRef]
19. Adoni, W.Y.H.; Lorenz, S.; Fareedh, J.S.; Gloaguen, R.; Bussmann, M. Investigation of autonomous multi-UAV systems for target detection in distributed environment: Current developments and open challenges. *Drones* **2023**, *7*, 263. [CrossRef]
20. Seuken, S.; Zilberstein, S. Formal models and algorithms for decentralized decision making under uncertainty. *Auton. Agents Multi-Agent Syst.* **2008**, *17*, 190–250. [CrossRef]
21. Zhang, B.; Lin, X.; Zhu, Y.; Tian, J.; Zhu, Z. Enhancing multi-UAV reconnaissance and search through double critic DDPG with belief probability maps. *IEEE Trans. Intell. Veh.* **2024**, *9*, 3827–3842. [CrossRef]
22. Cui, J.; Liu, Y.; Nallanathan, A. Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Trans. Wirel. Commun.* **2019**, *19*, 729–743. [CrossRef]

23. Shen, G.; Lei, L.; Zhang, X.; Li, Z.; Cai, S.; Zhang, L. Multi-UAV cooperative search based on reinforcement learning with a digital twin driven training framework. *IEEE Trans. Veh. Technol.* **2023**, *72*, 8354–8368. [CrossRef]
24. Luo, Q.; Luan, T.H.; Shi, W.; Fan, P. Deep reinforcement learning based computation offloading and trajectory planning for multi-UAV cooperative target search. *IEEE J. Sel. Areas Commun.* **2022**, *41*, 504–520. [CrossRef]
25. Hou, Y.; Zhao, J.; Zhang, R.; Cheng, X.; Yang, L. UAV swarm cooperative target search: A multi-agent reinforcement learning approach. *IEEE Trans. Intell. Veh.* **2023**, *9*, 568–578. [CrossRef]
26. Yang, Y.; Polycarpou, M.M.; Minai, A.A. Multi-UAV cooperative search using an opportunistic learning method. *J. Dyn. Syst. Meas. Control.* **2007**, *129*, 716–728. [CrossRef]
27. Fei, B.; Bao, W.; Zhu, X.; Liu, D.; Men, T.; Xiao, Z. Autonomous cooperative search model for multi-UAV with limited communication network. *IEEE Internet Things J.* **2022**, *9*, 19346–19361. [CrossRef]
28. Zhou, Z.; Luo, D.; Shao, J.; Xu, Y.; You, Y. Immune genetic algorithm based multi-UAV cooperative target search with event-triggered mechanism. *Phys. Commun.* **2020**, *41*, 101103. [CrossRef]
29. Ni, J.; Tang, G.; Mo, Z.; Cao, W.; Yang, S.X. An improved potential game theory based method for multi-UAV cooperative search. *IEEE Access* **2020**, *8*, 47787–47796. [CrossRef]
30. Sun, X.; Cai, C.; Pan, S.; Zhang, Z.; Li, Q. A cooperative target search method based on intelligent water drops algorithm. *Comput. Electr. Eng.* **2019**, *80*, 106494. [CrossRef]
31. Yue, W.; Tang, W.; Wang, L. Multi-UAV cooperative anti-submarine search based on a rule-driven MAC scheme. *Appl. Sci.* **2022**, *12*, 5707. [CrossRef]
32. Pérez-Carabaza, S.; Besada-Portas, E.; López-Orozco, J.A. Minimizing the searching time of multiple targets in uncertain environments with multiple UAVs. *Appl. Soft Comput.* **2024**, *155*, 111471. [CrossRef]
33. Duan, H.; Zhao, J.; Deng, Y.; Shi, Y.; Ding, X. Dynamic discrete pigeon-inspired optimization for multi-UAV cooperative search-attack mission planning. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *57*, 706–720. [CrossRef]
34. Xu, L.; Cao, X.; Du, W.; Li, Y. Cooperative path planning optimization for multiple UAVs with communication constraints. *Knowl.-Based Syst.* **2023**, *260*, 110164. [CrossRef]
35. Cao, X.; Luo, H.; Tai, J.; Jiang, R.; Wang, G. Multi-agent target search strategy optimization: Hierarchical reinforcement learning with multi-criteria negative feedback. *Appl. Soft Comput.* **2023**, *149*, 110999. [CrossRef]
36. Waharte, S.; Trigoni, N. Supporting search and rescue operations with UAVs. In Proceedings of the IEEE 2010 International Conference on Emerging Security Technologies, Canterbury, UK, 6–7 September 2010; pp. 142–147.
37. Gupta, A.; Bessonov, D.; Li, P. A decision-theoretic approach to detection-based target search with a UAV. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5304–5309.
38. Bertuccelli, L.F.; How, J.P. Robust UAV search for environments with imprecise probability maps. In Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 12–15 December 2005; pp. 5680–5685.
39. Millet, T.; Casbeer, D.; Mercker, T.; Bishop, J. Multi-agent decentralized search of a probability map with communication constraints. In Proceedings of the AIAA Guidance, Navigation, and Control Conference, Toronto, ON, Canada, 2–5 August 2010; p. 8424.
40. Zhen, Z.; Chen, Y.; Wen, L.; Han, B. An intelligent cooperative mission planning scheme of UAV swarm in uncertain dynamic environment. *Aerosp. Sci. Technol.* **2020**, *100*, 105826. [CrossRef]
41. Jin, Y.; Liao, Y.; Minai, A.A.; Polycarpou, M.M. Balancing search and target response in cooperative unmanned aerial vehicle (UAV) teams. *IEEE Trans. Syst. Man, Cybern. Part B* **2006**, *36*, 571–587. [CrossRef]
42. Gao, Y.; Li, D. Unmanned aerial vehicle swarm distributed cooperation method based on situation awareness consensus and its information processing mechanism. *Knowl.-Based Syst.* **2020**, *188*, 105034. [CrossRef]
43. Zhang, H.; Ma, H.; Mersha, B.W.; Zhang, X.; Jin, Y. Distributed cooperative search method for multi-UAV with unstable communications. *Appl. Soft Comput.* **2023**, *148*, 110592. [CrossRef]
44. Huang, S.; Ontañón, S. A closer look at invalid action masking in policy gradient algorithms. *arXiv* **2020**, arXiv:2006.14171.
45. Wang, H.; Yufeng, Z.; Wei, L.; Jiaqiang, T. Multi-UAV 3D collaborative searching for moving targets based on information map. *Control. Decis.* **2023**, *38*, 3534–3542.
46. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624.
47. Su, K.; Qian, F. Multi-UAV cooperative searching and tracking for moving targets based on multi-agent reinforcement learning. *Appl. Sci.* **2023**, *13*, 11905. [CrossRef]
48. Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W.M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J.Z.; Tuyls, K.; et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv* **2017**, arXiv:1706.05296.
49. Rashid, T.; Samvelyan, M.; De Witt, C.S.; Farquhar, G.; Foerster, J.; Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–51.