drones

Article

Multi-UAV Roundup Inspired by Hierarchical Cognition Consistency Learning Based on an Interaction Mechanism

Longting Jiang [1,*], Ruixuan Wei [2] and Dong Wang [2]

[1] Graduate School, Air Force Engineering University, Xi'an 710051, China
[2] Aviation Engineering School, Air Force Engineering University, Xi'an 710038, China
* Correspondence: kgdjltsmile@163.com

Abstract: This paper is concerned with the problem of multi-UAV roundup inspired by hierarchical cognition consistency learning based on an interaction mechanism. First, a dynamic communication model is constructed to address the interactions among multiple agents. This model includes a simplification of the communication graph relationships and a quantification of information efficiency. Then, a hierarchical cognition consistency learning method is proposed to improve the efficiency and success rate of roundup. At the same time, an opponent graph reasoning network is proposed to address the prediction of targets. Compared with existing multi-agent reinforcement learning (MARL) methods, the method developed in this paper possesses the distinctive feature that target assignment and target prediction are carried out simultaneously. Finally, to verify the effectiveness of the proposed method, we present extensive experiments conducted in the scenario of multi-target roundup. The experimental results show that the proposed architecture outperforms the conventional approach with respect to the roundup success rate and verify the validity of the proposed model.

Keywords: multi-target roundup; neighborhood cognitive consistency; opponent graph reasoning network; hierarchical cognitive consistency learning

Citation: Jiang, L.; Wei, R.; Wang, D. Multi-UAV Roundup Inspired by Hierarchical Cognition Consistency Learning Based on an Interaction Mechanism. *Drones* **2023**, *7*, 462. https://doi.org/10.3390/drones7070462

Academic Editors: Mou Chen, Bin Jiang, Youmin Zhang, Zixuan Zheng and Shuyi Shao

Received: 17 May 2023
Revised: 11 June 2023
Accepted: 28 June 2023
Published: 11 July 2023

1. Introduction

In recent years, inspired by the self-organizing behavior of biological swarms in nature, the collaborations and information interactions of multi-agent systems have attracted increasing attention among researchers. The core idea behind such systems is information interaction between individuals and between individuals and the environment, from which orderly, collective, organized behavior emerges with a certain level of robustness. Inspired by this, improving the intelligence level of Unmanned Aerial Vehicle (UAV) clusters through the convergence of individual intelligence has become a popular research topic in the field of UAVs. Currently, UAV clusters are mostly used in target search, reconnaissance surveillance, and target roundup tasks. For target roundup tasks, UAV clusters need to dynamically generate suitable group aggregation formations to achieve effective roundup behavior. Based on the number of targets, roundup missions can be divided into single-target roundups [1,2] and multi-target roundups [3]. The key problem for multi-target encirclement is controlling the UAV swarm to cooperatively encircle multiple targets in a special formation through local information interactions. At the same time, due to the development of intelligence, the targets may have the ability to formulate highly intelligent escape strategies. Hence, how to solve the multi-target roundup problem in an adversarial environment remains an open question.

Although there has been much success in the field of single-target roundup, multi-target roundup needs more in-depth research and improvement. Unlike the existing work on single-target roundup, current studies on multi-target roundup are restricted to processing relatively little information and are usually based on an assumption that the targets are stationary or employ a relatively simple escape strategy. Existing multi-target

roundup models also give less consideration to the adversarial nature of the targets and make stronger assumptions about the adversarial environment. In Reference [2], a target is dynamically assigned to each agent, enabling a multi-agent system to self-organize to round up dynamic targets, but a limitation of this is that the number of targets needs to be known in advance. Ref. [3] investigated the self-organized multi-target roundup problem for a given task region, but the targets were assumed to be stationary, and the communication between individuals was global and computationally intensive. In multi-target roundup, the agents must not only adapt to a dynamic environment to complete the assignment of targets but also learn the opponents' time-varying strategies. The above issues pose significant challenges to research in this field.

To improve the success rate of multi-target roundup, some scholars have considered the communication relationships of the agents, combined with multi-agent reinforcement learning (MARL) theory, to improve the efficiency of rounding up targets through information interaction. Recently, the problem of communication present in MARL has aroused researchers' interest. Related methods can be divided into "predesigned" methods and "learning-based" methods. Although the traditional "predesigned" approaches [4–9] can find good solutions to the issues with the communication strategies of MARL, they are limited by several factors; for example, most approaches require prior knowledge, and the topological relationships between intelligent agents cannot vary over time. Unlike traditional "predesigned" methods, "learning-based" methods mainly apply a framework of centralized training with decentralized execution (CTDE) based on a deep neural network. However, some of these methods rely on the simple stacking of the states of the intelligent agents, and some involve the simple summation of information exchange. In recent years, attention mechanisms [10], which are suitable for application in communication models, have endowed communication strategies with powerful learning capabilities in complex and realistic scenarios. For instance, an attention network can help to determine when and with whom to communicate. The Individualized Controlled Continuous Communication Model (IC3Net) [11] controls communication with a gating mechanism to make decisions. Attention Multi-agent Deep Deterministic Policy Gradient (ATT-MADDPG) was proposed in Reference [12] based on the Multi-agent Deep Deterministic Policy Gradient (MADDPG) [13] framework to improve the learning efficacy by introducing a special network to explicitly model the dynamic joint policy of teammates. An attention-based communication neural network (CommNet) [14] can additionally precisely calculate whether communication is necessary for each pair of agents by considering the relevance of each received message.

The existing communication methods construct an explicit communication paradigm in which all agents make decisions regarding when and with whom communication occurs; however, this paradigm entirely ignores the information utility of different agents in some adversarial multi-agent games. For instance, in the multi-target roundup task, agents need to obtain messages from neighboring agents in order to appropriately cooperate. However, in the process of information exchange, if an agent gives the same weight to each neighboring agent, then all agents will have difficulty deducing their own contributions to the team's success. Moreover, the communication topology is dynamic and varies with time, and using a static communication topology will reduce the efficiency of cooperation.

To address the limitations mentioned above, we propose a hierarchical attentional communication mechanism strategy network based on the multi-agent reinforcement learning framework to solve the problem of cooperative multi-target roundup by a UAV swarm in an adversarial environment. Specifically, in this paper, we propose a novel target assignment model based on neighborhood cognitive consistency (NCC). Then, a dynamic communication topology is constructed, and an information utility model based on an attention mechanism is developed to achieve efficient communication among the UAV swarm.

The main contributions of this article are as follows:

1. We model the multi-target roundup problem based on neighborhood consistency theory to promote cognitive consistency during group tasks and realize coordinated behavior among predators.
2. We propose a novel communication framework for MARL to explicitly quantify the information effectiveness among UAVs using graph attention neural networks; this is more in line with the information exchange that occurs in multi-agent systems. This is also the main innovation of this work, as previous methods of information interaction have been mainly based on a preset non-time-varying communication topology established at the beginning of the task.
3. Unlike in previous studies, which have mostly focused on single-target roundup tasks, extensive experiments are conducted in the scenario of multitarget roundup to verify the superiority of our proposed method and the effectiveness of the components of the proposed model.

The outline for the remainder of the article is as follows. Section 2 summarizes related work on the multi-agent roundup problem and multi-agent communication strategies. Section 3 gives a brief introduction to the theory of neighborhood cognitive consistency. Section 4 details the proposed method and the cooperative roundup model. Section 5 presents and discusses the simulation results. Finally, concluding remarks are provided in Section 6.

2. Related Work

To address the abovementioned issues, e.g., multi-agent roundup, communication strategies, and communication modeling based on attention mechanisms, a tremendous number of studies have recently been reported. Here, we focus on the differences between the previous literature and our proposed method.

2.1. Multi-Target Roundup

Recently, scholars have conducted much research on multi-agent systems in the task scenario of target roundup. Their research can be mainly divided into cooperation and game strategies [15,16], coverage control [17], and circular tracking strategies [18]. Awheda M. D. et al. [19] transformed the roundup problem into an Apollonius circle solving problem to achieve control of the agents through fuzzy logic. Wang X. et al. [20] extended the target roundup problem to three-dimensional space and designed a seizure formation to achieve target roundup in a three-dimensional environment.

However, most previous studies have focused on single-target roundup, and multi-target roundup has rarely been investigated. Some scholars have recently made attempts to address this issue. Yasuda T. et al. [21] investigated multi-target roundup in a two-dimensional environment based on evolutionary artificial neural networks but did not consider an environment with the presence of obstacles. Dutta K. et al. [22] proposed a multi-target discrete roundup model but employed strong assumptions: no obstacles, no confrontation, and specific paths. Hongqiang Z. et al. [2] proposed a multi-target simplified virtual force model to achieve multi-target roundup from a cybernetic point of view, but their work was limited to certain assumed target motion laws. Fan et al. [23] proposed a consensus initiative-based multi-target roundup framework. Their model is limited to ideal assumptions about the communication conditions of the agents and cannot achieve dynamic adjustment of the roundup strategy.

In this article, we propose a roundup strategy for multiple targets that explicitly considers the adversarial strategies of the targets as well as the dynamics of the obstacles in the mission environment.

2.2. Multi-Agent Reinforcement Learning with Communication

Communication learning presents a challenging problem for multiple agents. In recent decades, various approaches have been proposed, such as “predesigned” communication

strategies and "learning-based" communication strategies. "Predesigned" communication strategies are mainly based on predefined rules or strong assumptions about the communication topology among agents. Chu T. et al. [24] and Gupta S. et al. [4] proposed an interaction model based on a specified communication range. Only agents that are within the communication range are considered neighbors, and then they can communicate with each other. Deshpande A. M. et al. [5] and Pakizeh E. et al. [6] proposed predetermined communication rules to solve the formation control problem for multiple agents. To address the communication problem, Wang B. et al. [7] proposed weighted mean field reinforcement learning [8], in which the pairwise communication between any UAV and its neighbors is modeled as communication between a central UAV and a virtual UAV, abstracted from the weighted mean effect of the neighboring UAVs. While all of the above methods improve the efficiency of multi-agent collaboration to some extent, such "predesigned" communication strategy methods are limited by prior knowledge, and once set up, they cannot adjust to changes.

"Learning-based" communication strategies are mainly based on neural network learning and the establishment of relationships between agents. CommNet [8] uses only the average of other intelligent agents' states as information for communication rather than constructing a specific model of information interaction. IC3Net [5] can determine when agents communicate with others based on a gating mechanism, which can be regarded as the original hard attention mechanism. CommNet and IC3Net both process messages through simple averaging, which may result in the loss of some information. Much like with the weighted average field, CommNet and IC3Net also process information generated by the communication network. Jiang J. et al. [9] proposed the ATOC model based on an attention mechanism to decide when and with whom an agent should communicate within its observation range. The disadvantage is that, in this model, an agent assigns the same weight to every other agent with which it communicates, so it is not possible to determine which agents contribute more to the completion of the task. Based on the CTDE framework, the SchedNet [25] algorithm was proposed, which selects a preset number of agents for communication based on the weights of different information. The above methods use various techniques to determine the objects of communication. MADDPG [7] extends actor–critic algorithms to the multi-agent setting based on the CTDE framework without an explicit communication model. This method is not feasible for large-scale multi-agent problems in which the states of others are introduced directly during evaluation. ATT-MADDPG [6] introduces an attention mechanism [26–28] to explicitly model the joint strategy of a multi-agent system, thereby enhancing the effect of centralized evaluation and achieving the efficient processing of information. Based on the MADDPG algorithm, the Multi-Actor-Attention-Critic algorithm (MAAC) [29] introduces a soft attention mechanism into the construction of Q functions, assigns different weights to local observations, and dynamically selects agents for communication. These methods are limited by the characteristics of MADDPG, the biggest drawback of which is that the complexity of the algorithm grows exponentially with the number of intelligent agents. All of the above methods are based on centralized evaluation processing and rely on the simple stacking of communication information.

2.3. Multi-Agent Communication with a Graph Attention Network Mechanism

With the development of graph neural networks, recent works have converted communication topologies into graph neural networks to model the interactions between agents. Graph Attention Multi-agent reinforcement learning (GAMA) [30] constructs the communication model between agents based on MADDPG and a graph attention mechanism; however, this method is limited by the number of agents present during centralized evaluation. The Multi-Agent Graph-attention Communication (MAGIC) [31] network uses schedulers to solve the problem of when and with whom agents should communicate, and message processors use a dynamic graph attention network to process the information. This framework achieves improved communication efficiency and can scale to larger state–

action spaces. Both the distributional multiagent cooperation algorithm (DMAC) [32] and Deep Graph Q-learning algorithm (DGQ) [33] essentially introduce an attention mechanism into the framework of value function decomposition, which facilitates collaboration among agents to accomplish a task. However, there are some differences between this simple summation of value functions and the intelligent aggregation of swarms. Graph-Based Coordination Strategy (GCS) [34] constructs graph generator models and graph-based coordination policies to achieve behavior coordination among agents, and the outputted directed acyclic graphs can capture the interdependencies of dynamic decisions among agents and facilitate behavior learning. DGN [35] models the environment as a graph, employs a multi head attention mechanism to extract the relationships between agents, and uses convolutional networks to represent the Q function for centralized evaluation. Although all of the above approaches achieve information exchange, they all face constraints regarding the differentiability of graphs. Yali Du et al. [36] took the dynamic nature of communication into consideration and generalized the coupling flow to model the interaction graph. The resulting dynamic communication topology reflects the correlations between agent interactions. However, the information utility of the agents is less considered.

In contrast to the above literature, we aimed to account for both group cognitive consistency and communication utility. Therefore, our motivation was quite different from that of the previous works. In our work, we developed a modified graph attention network, designed a dynamic communication model that takes the influence of agents into consideration, and designed an opponent reasoning graph network [37–39] to predict the ideal roundup position. With the help of these components, our objective was to improve the efficiency of target roundup in a multi-target scenario.

3. Preliminaries

3.1. Partially Observable Markov Decision Process (POMDP)

MARL can be modeled as a partially observable Markov decision process (POMDP) with multiple agents. The POMDP can be described as a tuple $\langle S, O, A, R, P, Z, \gamma, N \rangle$. At each time step t, $s_t \in S$ denotes the global state, $o_t^i \in O_t$ is the partial observation of agent i of the global state, agent i chooses its own action $a_t^i \in A$, the joint actions for N agents are represented by $\boldsymbol{a}_t$, and the next state is determined in accordance with the transition probability $P(s_{t+1}|s_t, \boldsymbol{a}_t) : S \times A^N \to [0, 1]$. At each transition, agent i will obtain a reward $r_t^i : S \to R$, and $\gamma \in (0, 1)$ is a corresponding discount factor. In the partially observable Markov decision process, each agent only has access to a local observation $z \in Z$ in accordance with the observation function $O(s_t) : S \to Z$. Agent i aims to learn a policy $\pi^i(a^i|o^i) : o^i \to a^i$ that will maximize the accumulated discounted reward $R^i = \sum_{t=0}^{T} \gamma^t r_t^i$. The joint policy of all agents can be expressed as $\boldsymbol{\pi} = [\pi_1, \pi_2 \cdots \pi_N]$. The state–action value function $Q^{\boldsymbol{\pi}}$ is defined as follows:

$$Q^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) = E_{s_{t+1}, \boldsymbol{a}_{t+1}, \cdots}[R_t|s_t, \boldsymbol{a}_t]. \tag{1}$$

The state value function $V^{\boldsymbol{\pi}}$ is defined as the expected cumulative discounted future reward:

$$V^{\boldsymbol{\pi}}(s_t) = E_{s_{t+1}, \boldsymbol{a}_{t+1}, \cdots}[R_t|s_t]. \tag{2}$$

The advantage function $A^{\boldsymbol{\pi}}$ is described as follows:

$$A^{\boldsymbol{\pi}} = Q^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) - V^{\boldsymbol{\pi}}(s_t). \tag{3}$$

3.2. Graph Attention Network (GAT)

A graph attention network is an effective model for processing structured data that are represented as a graph. It can extract the relationships between agents and other related agents. Based on the attention mechanism, the agents can preferentially obtain valuable information rather than processing all information with the same weight, thereby improving communication efficiency. Given the fact that the communication topology is

similar to a graph, the communication between agents can be described as an undirected graph $G = (\boldsymbol{V}, \boldsymbol{E})$, where the variable $\boldsymbol{V}$ is a set of nodes and the variable $\boldsymbol{E} \subseteq \boldsymbol{V} \times \boldsymbol{V}$ is a set of edges $\boldsymbol{E} = \left\{ e_{ij} \middle| i, j \in \boldsymbol{V} \right\}$, for which the edge between agent i and agent j is denoted by e_{ij} when agent i can communicate with agent j. For simplicity, edge e_{ij} is considered to be determined by its two endpoints. Therefore, the communication topology can be defined as follows:

$$\boldsymbol{E}_t = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1N} \\ e_{21} & e_{22} & \cdots & e_{2N} \\ \cdots & \cdots & \ddots & \cdots \\ e_{N1} & e_{N2} & \cdots & e_{NN} \end{bmatrix}, \tag{4}$$

where $e_{ij} \in \{0, 1\}$ indicates whether communication is possible between agents, with $e_{ij} = 1$ indicating that agent i and agent j can communicate with each other; otherwise, $e_{ij} = 0$.

Each node $i \in \boldsymbol{V}$ computes the node-embedding vector h_k^i of other graph nodes by aggregating the nodes' representations h_{k-1}^j in the fully connected mode. The computation can be described as follows:

$$h_k^i = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} \boldsymbol{W} h_{k-1}^j\right), \tag{5}$$

where $\{j \in N_i\}$ denotes that agent j is connected to agent i. The attention weights can be defined as follows:

$$\alpha_{ij} = \frac{\exp\left(f\left(\boldsymbol{W}h_i, \boldsymbol{W}h_j\right)\right)}{\sum_{j=1}^{K} \exp\left(f\left(\boldsymbol{W}h_i, \boldsymbol{W}h_j\right)\right)}. \tag{6}$$

In this article, the feature vector h^i is the hidden state of observation o_t^i of agent i.

4. Methodology

In a multi-target roundup scenario, all agents strategically and simultaneously take their actions based on their own policies. Suppose that there are N^{uav} agents and N^{tar} targets in multi-target roundup task $M = \left\{ target_{left} \middle| target_{left} = 0 \right\}$. The multi-target roundup task can be divided into multiple subtasks $M = \{m_1, m_2, \cdots m_n\}$ in accordance with the target assignment strategy of the agents, where n denotes the number of subtasks. Each group finishes a subtask so as to maximize the rewards, as optimized through the training process. The objective in this article is to learn a policy for agent i to efficiently accomplish a multitarget roundup task in an adversarial environment.

4.1. Overall Structure and Training Method

In this section, we outline the overall structure of the hierarchical cognitive consistency learning (HCCL) method in Figure 1. We apply cognitive consistency and an opponent relation graph in combination with the typical MADDPG algorithm. As shown in Figure 1, the proposed framework includes a multi-agent dynamic communication model and a hierarchical cognitive consistency model. The opponent state prediction model is shown in Figure 5, and we elaborate upon this model below.

As shown in Figure 1, the multi-agent dynamic communication model is mainly composed of a hard-attention module and a soft-attention module. Through a hard attention mechanism and a soft attention mechanism, the communication topology of the agents is simplified, and the information utility values are calculated; then, the communication messages between agents are determined. At the same time, the dynamic attention communication network makes it more likely that invalid connections between agents can be avoided.

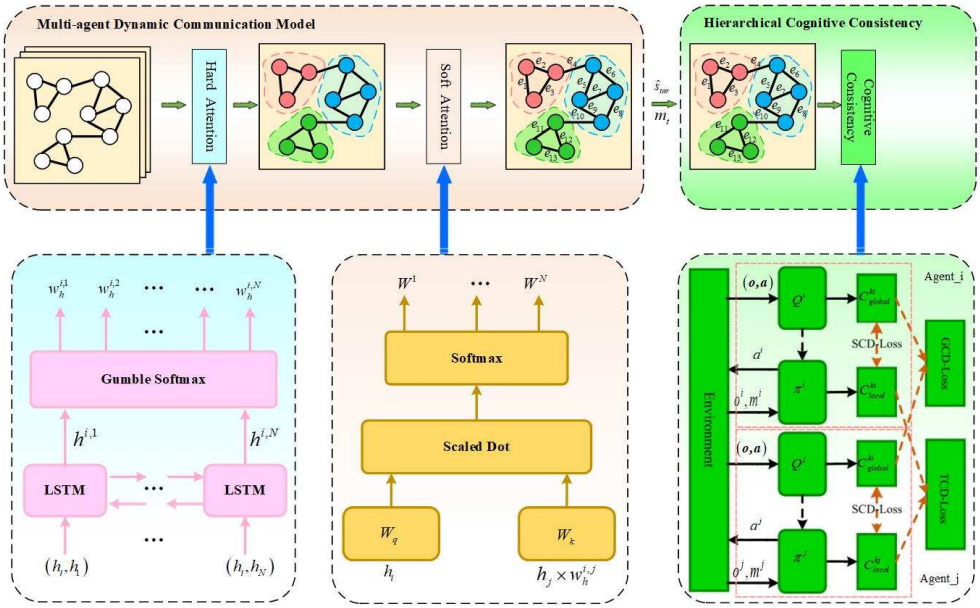

Figure 1. The framework of the proposed hierarchical attention network.

As shown in Figure 1, the hierarchical cognitive consistency model mainly includes three levels: self-cognition, teammate cognition and group cognition. Each agent forms its own local self-cognition and makes decision based local observation and received messages, and global cognition is formed based on global observation and actions. In accordance with the communication messages and the observation states of the agents, different levels of task cognition are formed. Furthermore, the proposed framework is designed based on the well-known MADDPG algorithm. Under the framework of the multi-agent reinforcement learning MADDPG network, combined with the constraints of hierarchical cognitive consistency, the action policy network and global evaluation network of MADDPG are updated, and multi-target roundup is finally realized.

4.2. Multi-Agent Dynamic Communication Model

The communication relationships among agents can be constructed as an undirected graph $G = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ denotes the agents and $\boldsymbol{E}$ represents the communication connections between agents. For a communication topology with N agents, the edges can be expressed as follows:

$$\boldsymbol{E} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}, \tag{7}$$

where $\boldsymbol{E}$ satisfies the condition $a_{ij} = \{0, 1\}$; if there is a connection, then $a_{ij} = 1$, and otherwise, this value is zero.

A hard attention mechanism can select one of the input vectors as the network output, whereas a soft attention mechanism can assign different weights to each input vector in accordance with their correlations. Accordingly, a combination of a hard attention mechanism and a soft attention mechanism is adopted as described in this section to establish the dynamic communication model of the multiple agents.

In a multi-agent environment with partial observability, each agent obtains a partial observation o_i^t about the environment at time t in accordance with the environmental information perceived by its sensors. To eliminate the influence of different observations and each agent's own performance on the information acquisition error, it is necessary to normalize the partial observations. Then, the partial observations o_i^t are input into a

fully connected neural network and encoded to form feature vectors h_i^t. On this basis, the communication topology is simplified by means of an attention mechanism.

The combination of feature vector h_i^t and feature vector h_j^t can be expressed as $\left(h_i^t, h_j^t\right)$. Considering the influence of different observations, the feature vector combination $\left(h_i^t, h_j^t\right)$ is inputted into a Bi-LSTM network and a fully connected layer $f(\cdot)$ and the relationship feature vector $h^{i,j}$ is outputted as follows:

$$h^{i,j} = f\left(BiLSTM\left(h_i^t, h_j^t\right)\right). \tag{8}$$

To overcome the deficiencies of the hard attention mechanism, the Gumbel soft-max function is used to process the input vector $h^{i,j}$; then, gradient backpropagation can be realized. The output feature vector can be expressed as follows:

$$W_h^{i,j} = gum\left(h^{i,j}\right), \tag{9}$$

where $gum(\cdot)$ denotes the Gumbel soft-max function and $W_h^{i,j}$ denotes the relationship vector.

To measure the utility values of different pieces of information and avoid unnecessary information interference, the simplified communication subgraph of the multiple agents is reprocessed by the soft attention mechanism. Accordingly, the information utility value of the other agents j with respect to agent i can be expressed as follows:

$$W_s^{i,j} = \frac{\exp\left(h_j^T W_k^T W_q h_i W_h^{i,j}\right)}{\sum_{j=1}^{N} \exp\left(h_j^T W_k^T W_q h_i W_h^{i,j}\right)}, \tag{10}$$

where $W_h^{i,j}$ denotes the relationship vector output by the hard attention mechanism, $W_s^{i,j}$ denotes the relationship vector output by the soft attention mechanism, W_k denotes the transform vector of h_j, and W_q denotes the transform vector of h_i. Therefore, in the simplified communication topology, the information m^i received by agent i can be expressed as shown in Formula (11):

$$m^i = \sum_{i \neq j} W_h^{i,j} W_s^{i,j} h_j. \tag{11}$$

Once the communication message m^i between agents has been obtained, the action policy of agent i can be expressed as follows:

$$a^i = \pi\left(m^i, o^i\right) \tag{12}$$

where m^i is the communication message of the agent, o^i is the observation of agent i, and a^i denotes the action that the agent takes.

4.3. Hierarchical Cognitive Consistency Model

Compared with the single-target roundup task, the need for target assignment for multi-target roundup makes the problem more complex. To address this challenge, inspired by cognitive consistency theory [40,41], we propose a hierarchical cognitive consistency model.

The cognition of an agent is defined as its understanding of the local environment. The partial observation of an agent includes the positions of neighboring agents and obstacles as well as high-level knowledge extracted from the environment. In Reference [42], a cognition network was proposed to output cognition variables, which takes the hidden state features of an agent as input. In a multi-target roundup scenario, the multi-agent system can only capture partial observations due to the limitations of each agent's own local view. Therefore, when different agents face the same task environment, their cognitions

are not identical. This phenomenon is referred to as cognitive difference and is intuitively illustrated in Figure 2. As the knowledge of the task environment gradually deepens, the cognitive differences between agents will gradually decrease, and eventually, their individual cognitions will converge. The individual cognitive differences of the agents play an important role in the cognition of the task environment, on the one hand increasing the efficiency of environmental exploration and on the other hand enhancing the synergistic ability among the agents. Hence, if agents of the same type or with the same task have a consistent cognition about a subtask, they will interact more, exhibit more similar behavior, and achieve better cooperation.

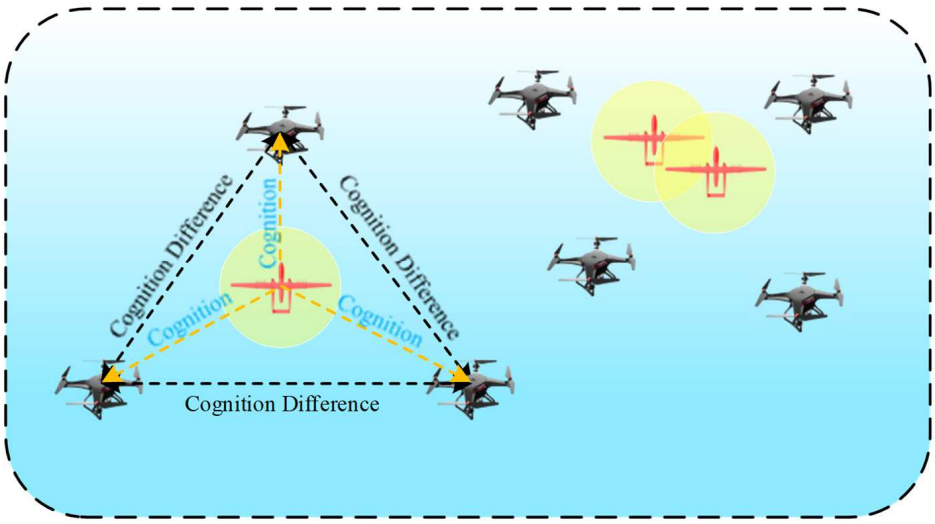

Figure 2. Cognitive differences between agents.

As described in Reference [40], neighborhood cognitive consistency describes the phenomenon where neighboring agents have formed similar cognitions about subtasks. In a multi-target roundup environment, each agent will form a cognition of its surroundings. When faced with the same subtask, neighboring agents usually have closer relationships and similar perceptions, so they are more likely to maintain consistent cognitions and achieve better cooperation. One key factor in determining how agents take action is their deviation from consistency. Hence, agents with similar cognitions will be more closely connected.

In this article, it is assumed that in the multi-target roundup scenario, each agent i can perceive only a partial observation o^i. Based on variational autoencoder (VAE) theory, this partial observation can be used to infer a set of hidden cognition variables C. Given a partial observation o^i, the conditional probability $p\left(C|o^i\right)$ for cognitive variables C can be expressed as shown in Formula (13):

$$p\left(C|o^i\right) = \frac{p\left(o^i|C\right)p(C)}{p\left(o^i\right)} = \frac{p\left(o^i|C\right)p(C)}{\int p\left(o^i|C\right)p(C)dC}. \tag{13}$$

In fact, however, the distribution of the potential cognitive variables C in Formula (13) is unknown, so $p\left(C|o^i\right)$. is difficult to calculate. According to the VAE principle, it can be approximated by an easy-to-handle probability distribution $q\left(C|o^i\right)$. The two probability distributions need to satisfy the divergence constraint shown in Formula (14):

$$\min \quad D_{KL}\left(q\left(C|o^i\right)\|p\left(C|o^i\right)\right). \tag{14}$$

The potential cognitive variables can be modeled by a VAE, as shown in Figure 3.

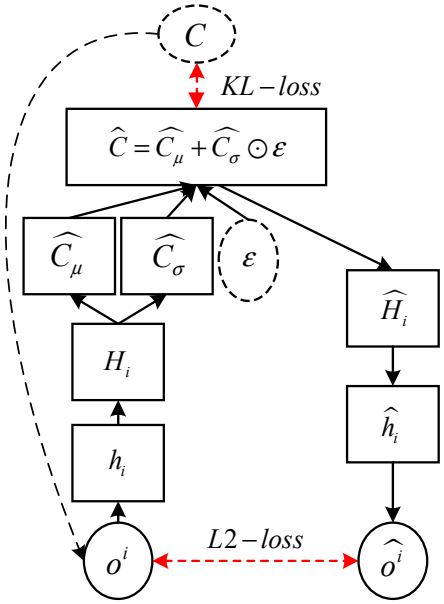

Figure 3. Schematic diagram of VAE of potential cognitive variables.

As shown in Figure 3, the cognitive variable C can be reconstructed by parameters based on the observation o^i. h_i and H_i denote the outputs of hidden layers of the neural networks. Specifically, the encoder of VAE learns a probability distribution $\left(q\left(\hat{C}\middle|o^i\right)\right.$ from observation o^i to cognitive variable $\hat{C}$, where $\hat{C} = \hat{C}_\mu + \hat{C}_\sigma \odot \varepsilon, (\varepsilon \sim N(0,1))$; $\hat{C}_\mu$ denotes the latent distribution's mean; and $\hat{C}_\sigma$ denotes the latent distribution's variance. Then, Formula (14) can be denoted as follows:

$$\min \quad D_{KL}\left(q\left(\hat{C}\middle|o^i\right)\middle\|p\left(C\middle|o^i\right)\right). \tag{15}$$

Inspired by above, in this article, we generalize the VAE process and construct the hierarchical cognitive consistency model. According to the observation o^i and the received message m^i, the agent forms local cognition C^{ki}_{local}. Based on the global joint state $\boldsymbol{s}^k$ and joint action $\boldsymbol{a}^k$, the cognition C^{ki}_{global} of the global task is formed. Based on the theory of VAE, the cognition of subtask C^{ki}_{local} should be consistent with the cognition of the global task C^{ki}_{global}. If two subtasks are members of same group with multiagent dynamic communication model learning, the cognition of these subtask C^{ki}_{local} and C^{kj}_{local} should remain consistent. Furthermore, they form similar global task cognitions, so C^{ki}_{global} and C^{kj}_{global} remain consistent. As the schematic example shown in Figure 4, the multi-agent system is divided into three communication subgraphs. Here, it is assumed that one subtask corresponds to one target, i.e., the number of subtasks is equal to the number of targets. For a task group $\mathcal{T}_k$, the group members can be denoted by the set $Group\_k = \left\{\text{Agent}_{k1}, \text{Agent}_{k2} \cdots \text{Agent}_{kN}\right\}$, and the observations of group k can be denoted by $\boldsymbol{o}^k = \left\{o^{k1}, o^{k2} \cdots o^{kN}\right\}$.

Based on its view of the task, each agent will form a consistent cognition, including its own self-cognition, the cognition of its teammates, and group cognition. Hence, there will be three levels of cognitive consistency. As the network learns and trains, the cognition among the agents will gradually converge. Then, the proposed approach divides the agents into several groups exactly as many groups as there are targets. The details of the above three cognitive consistencies are defined as follows.

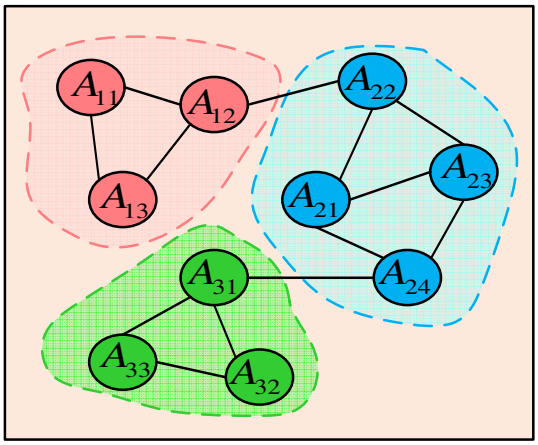

Figure 4. Schematic diagram of multi-agent communication subgraph division.

4.3.1. Consistency through Self-Supervised Learning

In a subtask group, from the perspective of the agents, each agent forms a cognition of the subtask based on its local view, which should be similar to the cognition of the group. To achieve local subtask consistency, similar to the consistency constraint using the KL divergence in reference [42], the local cognition C^{ki}_{local} should be similar to the group cognition C^{ki}_{global} based on the joint state of the group members. Self-supervised learning is achieved by minimizing the following objective:

$$\min \quad D_{KL}\left(q\left(C^{ki}_{local}\middle|o^i, m^i\right)\|p\left(C^{ki}_{global}\middle|\boldsymbol{s}^k, \boldsymbol{a}^k\right)\right), \tag{16}$$

where $\boldsymbol{s}^k$ represents the group state $\boldsymbol{s}^k = \left(s^{k1}, s^{k2}\cdots s^{kN}\right)$. We assume that in the group view, the subtask is referred to based on the joint observation $\boldsymbol{o}^k$ and joint action $\boldsymbol{a}^k$. Therefore, the group state is replaced to supervise the learning of subtask cognition and the self-cognitive dissonance loss (SCD-Loss) is defined as follows:

$$\min \quad D_{KL}\left(q\left(C^{ki}_{local}\middle|o^i, m^i\right)\|p\left(C^{ki}_{local}\middle|\boldsymbol{o}^k, \boldsymbol{a}^k\right)\right). \tag{17}$$

With the help of Formula (17), the actor network will eventually learn better self-cognition.

4.3.2. Consistency through Aligning Teammates in a Group

From the perspective of the group, if agents perform the same subtask, they are more likely to form the same cognition. To achieve cognitive consistency among teammates, each agent learns its subtask cognition in the local view, during which the agents should keep their cognitions consistent with those of their teammates.

For agent i, the team cognitive dissonance loss (TCD-Loss) is minimized in the local view, which is defined as shown in Formula (18):

$$\min \quad \sum_{j\in N(i)\cap j\neq i} D_{KL}\left(q\left(C^{ki}_{local}\middle|o^i, m^i; \theta^i\right)\|q\left(C^{kj}_{local}\middle|o^j, m^j; \theta^j\right)\right), \tag{18}$$

where o^i denotes the partial observation, m^i denotes the received message, C^{ki}_{local} denotes the cognition of the subtask based on partial observation o^i and message m^i, and $j \in N(i) \cap j \neq i$ denotes the neighboring agents in the same group.

4.3.3. Consistency through Global Task Cognition

From the perspective of the group, each agent has the same global goal—that is, to achieve multi-target roundup—and its own global cognitive consistency is achieved by minimizing the global cognitive dissonance loss (GCD-Loss), as shown in Formula (19):

$$\min \quad \sum D_{KL}\left(q\left(C_{global}^{ki}\middle|\boldsymbol{o}^k,\boldsymbol{a}^k;w^i\right)\|q\left(C_{global}^{kj}\middle|\boldsymbol{o}^k,\boldsymbol{a}^k;w^j\right)\right), \tag{19}$$

where C_{global}^{ki} denotes the cognition of the global task, $\boldsymbol{o}^k$ denotes the joint observation of the team, and $\boldsymbol{a}^k$ denotes the joint action of the team.

As shown in Figure 1, the proposed method uses the above definitions of cognitive consistency to train the network. Self-cognitive dissonance loss, team cognitive dissonance loss, and global cognitive dissonance loss are used to constrain agents in the same group to form neighborhood cognitive consistency and global task cognitive consistency in both the local and global views. Self-cognitive dissonance loss helps to improve the learning of an agent's self-cognition in the local view.

The critic network of the proposed method is trained by minimizing the total loss, as expressed in (20). Specifically, the global task cognition network parameters are realized by minimizing the sum of the dissonance losses L_{GCD}^i and the temporal difference error L_{TD}^i.

$$L_{Total}^i\left(w^i\right) = L_{TD}^i\left(w^i\right) + \alpha\sum_{i=1}^{N} L_{GCD}^i\left(w^i\right), \tag{20}$$

where the temporal difference error L_{TD}^i is expressed as shown in Formula (21):

$$L_{TD}^i = \mathbb{E}_{(\boldsymbol{o},\boldsymbol{a},r,\boldsymbol{o}')\sim D}\left(\left(y^i - Q^i\left(\boldsymbol{o},\boldsymbol{a};w^i\right)\right)^2\right), \tag{21}$$

where $\overline{w}^i$ denotes the parameters of the target critic network in MADDPG. Then, the output of the target critic network can be represented by Formula (22):

$$y^i = r^i + \gamma Q^i\left(\boldsymbol{o}',\boldsymbol{a}';\overline{w}^i\right)\Big|_{\boldsymbol{a}'=\pi(\boldsymbol{o}',m')} \tag{22}$$

The GCD-Loss for parameters θ^i can be expressed as shown in Formula (23):

$$L_{GCD}^i = \sum D_{KL}\left(q\left(C_{global}^{ki}\middle|\boldsymbol{o}^k,\boldsymbol{a}^k;w^i\right)\|q\left(C_{global}^{kj}\middle|\boldsymbol{o}^k,\boldsymbol{a}^k;w^j\right)\right). \tag{23}$$

For the actor network in MADDPG, the task cognitive network parameters θ^i of an agent can be obtained by deriving the self-cognitive dissonance loss L_{SCD}^i and the team cognitive dissonance loss L_{TCD}^i and then solving optimally.

The SCD-Loss $L_{SCD}^i\left(\theta^i\right)$ for parameters θ^i can be expressed as shown in Formula (24):

$$L_{SCD}^i\left(\theta^i\right) = D_{KL}\left(q\left(C_{local}^{ki}\middle|o^i,m^i;\theta^i\right)\|p\left(C_{local}^{ki}\middle|\boldsymbol{o}^k,\boldsymbol{a}^k;\theta^i\right)\right). \tag{24}$$

The TCD-Loss $L_{TCD}^i\left(\theta^i\right)$ for parameters θ^i can be expressed as shown in Formula (25):

$$L_{TCD}^i\left(\theta^i\right) = \sum_{j\in N(i)\cap j\neq i} D_{KL}\left(q\left(C_{local}^{ki}\middle|o^i,m^i;\theta^i\right)\|q\left(C_{local}^{kj}\middle|o^j,m^j;\theta^j\right)\right). \tag{25}$$

For the actor network of the proposed method, the derivative of the state–action value function $Q^i\left(\boldsymbol{o},\boldsymbol{a};w^i\right)$ can be expressed as shown in Formula (26):

$$\nabla_{\theta^i}L\left(\theta^i\right) = \mathbb{E}_{(o,a)\sim D}\left[\nabla_{\theta^i}\pi^i\left(o^i,m^i;\theta^i\right)\nabla_{a^i}Q^i\left(\boldsymbol{o},\boldsymbol{a};w^i\right)\Big|_{a^i=\pi^i(o^i,m^i;\theta^i)}\right]. \tag{26}$$

Then, the gradient of the actor network can be expressed as shown in Formula (27):

$$\nabla_{\theta^i} L^i_{Total}\left(\theta^i\right) = \nabla_{\theta^i} L\left(\theta^i\right) + \nabla_{\theta^i} L^i_{TCD}\left(\theta^i\right) + \nabla_{\theta^i} L^i_{SCD}\left(\theta^i\right). \tag{27}$$

Once the gradients of the policy network ($\nabla_{\theta^i} L^i_{Total}\left(\theta^i\right)$) and critic network ($\nabla_{w^i} L^i_{Total}\left(w^i\right)$) are obtained, the parameters of the MADDPG networks can be updated, and the task cognitive convergence of the multi-agent system can finally be realized by minimizing the dissonance losses of task cognition at all levels.

4.4. Opponent Graph Reasoning

Although some methods have attempted to model and predict targets' behaviors in an adversarial environment, they cannot clearly capture the logic of the targets' behaviors and their intentions. In this article, we propose an opponent graph reasoning method based on the observations of UAVs to improve the accuracy of prediction. Previous works have not taken the influence of the opponents into consideration. We modelled the relationship between the opponents and agents in a multi-target roundup scenario as a directed graph network G^o. The graph network G^o can be used to learn the relations among the agents and opponents and then predict the future state of the opponents. The opponent graph reasoning network architecture is shown in Figure 5.

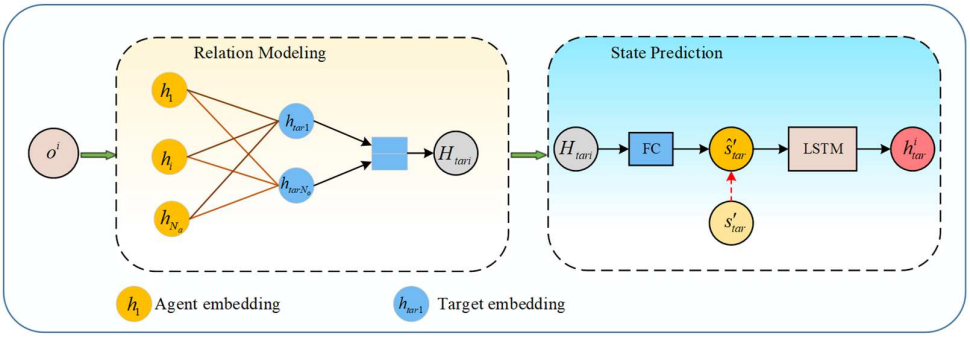

Figure 5. Opponent graph reasoning.

Figure 5 presents the details of the opponent graph reasoning network used to learn the best response in the adversarial environment. The input to the proposed network is the observation o^o_i of agent i, including the states of itself, its neighborhood, and the targets. Then, the hidden state h^o_i is extracted with an attention mechanism. Specifically, the agent will aggregate the states of the neighboring agents and targets and predict the future state of the opponents.

In accordance with the above definition of a graph attention network, the directed graph network used here, which we call the opponent relational graph, can be defined as follows:

$$G^o = (\boldsymbol{V}^o, \boldsymbol{E}^o), \tag{28}$$

where the nodes of the opponent relational graph are denoted by $|\boldsymbol{V}^o| = N^{uav} + N^{tar}$. Each edge e_{ji} represents the influence of agent i on target j. Hence, the different agents will have different influence utilities.

As shown in Figure 5, in the opponent reasoning model, different, fully connected networks are used to compute the hidden states h^o_i and $h^o_{tar\_j}$ by inputting the state s^i of agent i and the state s^{tar_j} of target j. Then, with the help of another fully connected network f^0_a, the weight e^o_{ji}, which describes the influence of neighborhood agent i on target j, is calculated as shown in Equation (29):

$$e^o_{ji} = f^o_a\left(h^o_{tar\_j}, h^o_i, W^o\right). \tag{29}$$

Then, the normalized weight a_{ji}^{o} is calculated using Formula (30):

$$a_{ji}^{o} = \frac{\exp\left(Leaky\ \mathrm{Re}\ LU\left(e_{ji}^{o}\right)\right)}{\sum_{i \in N_a^{tar}+N^{uav}} \exp\left(Leaky\ \mathrm{Re}\ LU\left(e_{ji}^{o}\right)\right)}, \tag{30}$$

where $Leaky\mathrm{Re}LU$ is a nonlinear activation function. Based on the embedded features h_i^o of agent i and the normalized weight a_{ji}^{o}, the state representation vector h_j^k is calculated by aggregating the state representations of the targets with different attention weights, as shown in Formula (31):

$$h_j^k = \sigma\left(\sum_{i \in (N_a^{tar}+N^{uav})} \exp\left(a_{ji}^{o}(W^o)^T h_i^o\right)\right), \tag{31}$$

where h_j^k denotes the aggregated embedding vector from the perspective of agent k and $\sigma(\cdot)$ denotes the nonlinear activation function. Then, by concatenating the state representations h_j^k, the opponent state representation matrix for agent k can be obtained, denoted by H_{tar}^k.

Once the opponent state representation matrix H_{tar}^k has been obtained, a state prediction network f_p is designed to predict the future state of the opponents. Specifically, the next state is expressed as follows:

$$\hat{S}_{tar}^{k'} = f_p\left(H_{tar}^k\right), \tag{32}$$

where $\hat{S}_{tar}^{k'}$ denotes the predicted future state of the opponents from the perspective of agent k.

To train the prediction network, prediction error is used as an intrinsic reward for predicting the future states of the opponents. The prediction error is described as follows:

$$R_k^{in} = E_{\hat{s}_{tar}^{k'} \sim S_{tar}^{a}}\left[\left(\hat{s}_{tar}^{k'} - s_{tar}^{k}\right)^2\right], \tag{33}$$

where $s_{tar}^{k'}$ is the real state of the opponents and $\hat{s}_{tar}^{k'}$ is the predicted state of the opponents. By combining this intrinsic reward with an extrinsic reward, the total reward can be calculated as follows:

$$R = R_k^{ex} - \alpha R_k^{in}, \tag{34}$$

where R_k^{ex} denotes the extrinsic reward. The details of the extrinsic reward will be described as part of the experimental setting.

In a multi-target roundup task, which is a cooperative and competitive mission, the number of targets may dynamically change during the process of performing the mission. To solve this issue, a long short-term memory (LSTM) network was designed to encode the opponent team and output the predicted opponent state representation h_{tar}^k. This will promote cooperation in an adversarial environment while taking the dynamic nature of the opponents into account.

5. Simulations

Our experiments aimed to answer the following questions.

RQ (1): Can the HCCL method consider the dynamic environment and the adversarial nature of the targets more effectively than state-of-the-art MARL algorithms that consider only non-time-varying communication?

RQ (2): Are the main components of the HCCL method, such as the opponent relation graph and the cognitive consistency model, necessary?

RQ (3): How do the key hyperparameters in the HCCL method affect the efficiency in accomplishing tasks?

To address these questions, we conducted simulations based on different multi-target roundup scenarios to evaluate the performance and validate the effectiveness and generalizability of the proposed method.

5.1. Experimental Setting and Baselines

5.1.1. Motion Model of a UAV

There were at least two participating teams. In a two-dimensional continuous scenario, we considered one ally coalition including N^{uav} agents and another target coalition with N^{tar} targets. At time t, the position of an entity can be denoted by $p_i^t = [x_i^t, y_i^t]$, and its velocity is $v_i^t = \left[v_{i\_x}^t \cos\theta, v_{i\_y}^t \sin\theta\right]$.

The dynamic model of each UAV in this article can be expressed as follows:

$$\begin{aligned} x_i^t &= x_i^t + v_i^t \cdot \cos\theta dt \\ y_i^t &= y_i^t + v_i^t \cdot \sin\theta dt \\ v_i^t &= v_i^t + \left(\frac{F_i^t}{m} + \varepsilon\right) dt \end{aligned} . \tag{35}$$

5.1.2. Extrinsic Reward Function

As described in Formula (34) in Section 4 above, the proposed network needs an additional extrinsic reward, which is defined as follows. In the multi-target scenario, there were N^{uav} UAVs and N^{tar} targets. The goal of the N^{uav} homogeneous UAVs was to round up all the targets, while the goal of the targets was to escape from the UAVs. For the UAVs to round up the multiple targets, they needed to cooperate with each other because the targets had superior maneuverability. The detection and communication ranges of the UAVs were set in advance. In detail, the compound extrinsic reward for multi-target roundup is defined as follows:

$$R_i^{ex} = R_{dist}^i + R_{coll}^i + R_{cross}^i + \alpha R_{round}^i + \beta R_{succ}, \tag{36}$$

where R_i^{ex} denotes the total extrinsic reward of agent i, R_{dist}^i is the distance reward of agent i, R_{coll}^i is the collision reward of agent i, R_{cross}^i denotes the boundary crossing reward, R_{round}^i is the roundup reward, and R_{succ} denotes the mission success reward, which was received by the team of UAVs only when all targets were rounded up. α and β represent weight parameters of the rewards. The larger α is, the more selfish each UAV will be, paying more attention to its individual reward. The larger β is, the more the UAVs will be united and behave cooperatively.

The distance reward is defined as follows:

$$R_{dist}^i = -dist\left(p_{u\_i}^t, p_{t\_j}^t\right), \tag{37}$$

where $p_{u\_i}^t$ and $p_{t\_j}^t$ denote the positions of UAV i and target j, respectively.

The boundary crossing reward is defined as follows:

$$R_{cross}^i = \begin{cases} 0, & if \quad x < 1.8 \\ 10(x - 1.8) & if \quad 1.8 \le x < 2 \\ \min\left(e^{2x-1.8}, 10\right) & if \quad x \ge 2 \end{cases}, \tag{38}$$

where x denotes the abscissa or ordinate of agent i.

The collision reward is defined as follows:

$$R_{coll}^i = \begin{cases} 0, & if \quad dist\left(p_{u\_i}^t, p_{t\_j}^t\right) > d_i + d_j \\ -2, & else \end{cases}, \tag{39}$$

where d_i and d_j denote the safe distances of UAV i and UAV j, respectively.

The mission success reward is defined as follows:

$$R^i_{succ} = \begin{cases} 20, & if \quad target_{left} = 0 \\ 0, & else \end{cases}, \tag{40}$$

where $target_{left}$ represents the number of remaining targets. Only when all targets were rounded up would the UAVs obtain this reward.

The roundup reward is defined as follows:

$$R^i_{round} = \begin{cases} 0 & if \quad dist\left(p^t_{u\_i}, p^t_{t\_j}\right) > d_i + d_j \\ 10, & else \end{cases} \tag{41}$$

5.1.3. Baselines

In this section, we list the frequently used baselines in the multi-target roundup scenario that we chose for comparison with our model. The abbreviations used in this paper are listed in Table 1.

Table 1. The list of abbreviations.

NO.	Abbreviations	Algorithms
1	MARL	Multi-agent Reinforcement Learning
2	CTDE	Centralized Training with Decentralized Execution
3	MADDPG	Multi-agent Deep Deterministic Policy Gradient
4	CommNet	Communication neural network
5	NCC	Neighborhood Cognitive Consistency
6	HCCL	Hierarchical Cognitive Consistency Learning
7	SCD-Loss	Self-Cognitive Dissonance loss
8	TCD-Loss	Team Cognitive Dissonance loss
9	GCD-Loss	Global Cognitive Dissonance loss
10	LSTM	Long Short-Term Memory
11	IDQN	Independent Deep Q-learning Network
12	VDN	Value Decomposition Networks
13	QMIX	Q-mixing network
14	NCC-VDN	Neighborhood Cognitive Consistency based on Value Decomposition Networks

The details of these baselines are specified as follows. IDQN: Each agent acts independently based on the DQN algorithm, and there is no intercommunication.

The VDN and QMIX both generate an individual Q_i for each agent and share the total Q_{total} with all agents, without taking communication into consideration.

VDN: VDN is a value decomposition algorithm without communication that sums the individual Q_i.

QMIX: The QMIX algorithm is a generalization of the VDN algorithm that uses the concept of neural network representation instead of the linear summation of the VDN and imposes a constraint of monotonicity of the value function on the value decomposition algorithm.

CommNet: CommNet adopts a communication method in which an agent is provided with the average of the hidden state representations of other agents as a communication signal.

NCC-VDN: This method combines the cognitive consistency between agents and neighboring agents with the value function decomposition method VDN to realize cooperation among multiple agents.

5.2. Validation Results

We carried out experiments in various multi-target roundup scenarios to validate the effectiveness of the proposed MARL algorithm. For simplicity, we refer to the hierarchical task cognitive consistency learning method proposed in this paper as HCCL.

(1) We conducted target assignment experiments based on cognitive consistency. In this section, we illustrate the effectiveness of the proposed method in the two-target scenario, in which the targets adopted a simple escape strategy or a randomized escape strategy. Figures 6 and 7 depict the trajectories for roundup, while Figure 8 depicts the dynamic evolution of the agents' cognitions in the scenario where targets adopted a simple escape strategy. Another scenario is shown in Figures 9–11. As Figures 6, 7, 9 and 10 show, the blue dots represent the agents, the red dots represent the targets, the black circles represent the obstacles, and the dashed triangles represent the roundup formation.

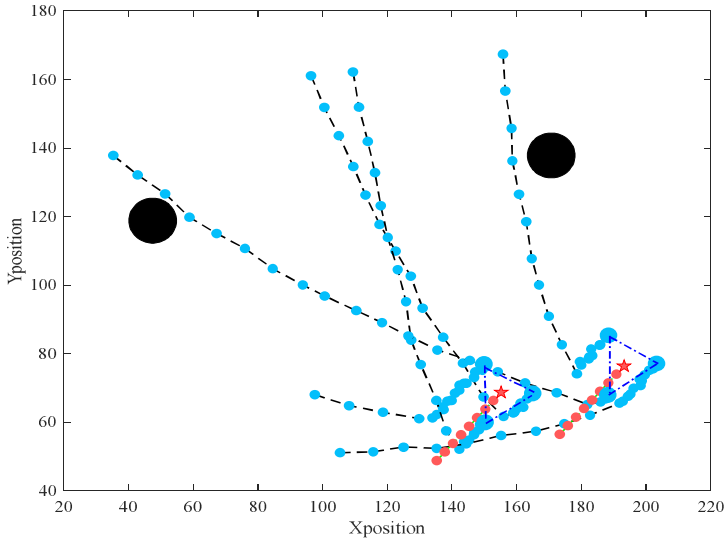

Figure 6. The trajectories for rounding up in the simple scenario with two targets.

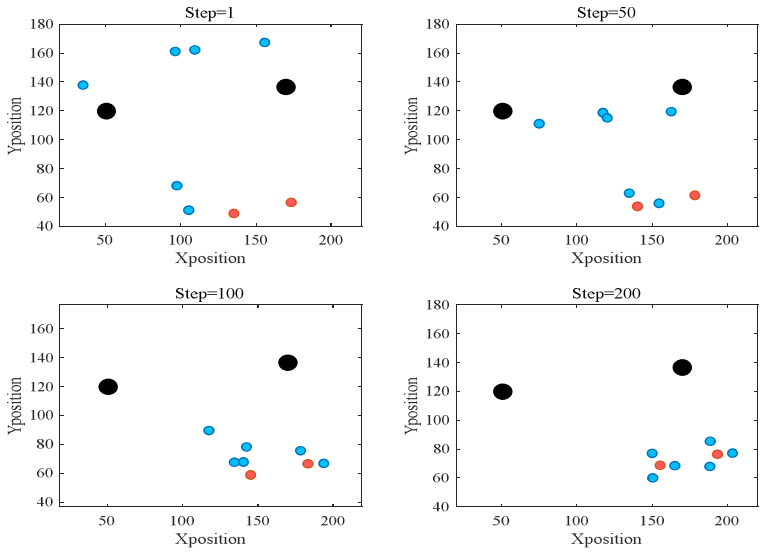

Figure 7. The process of rounding up in the simple scenario with two targets.

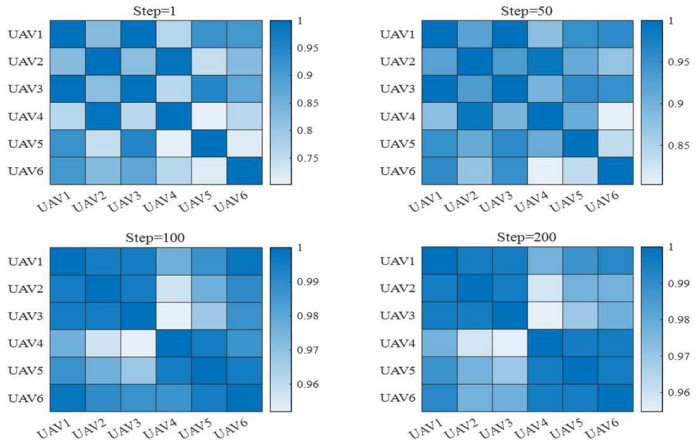

Figure 8. The evolutionary processes of the agents' cognitions in the simple two-target scenario.

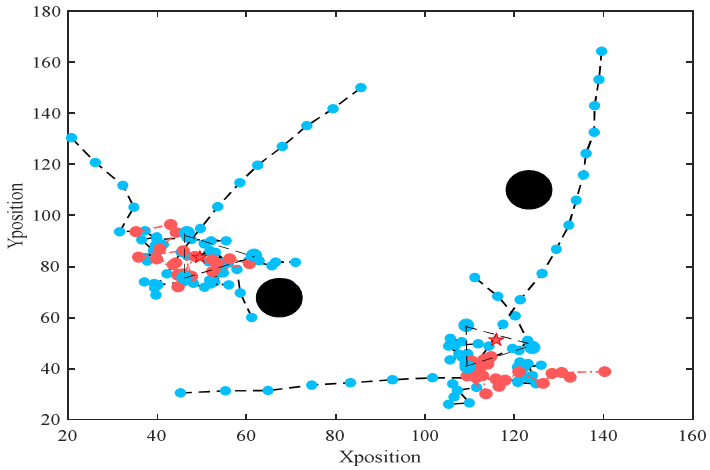

Figure 9. The trajectories for rounding up in the complicated scenario with two targets.

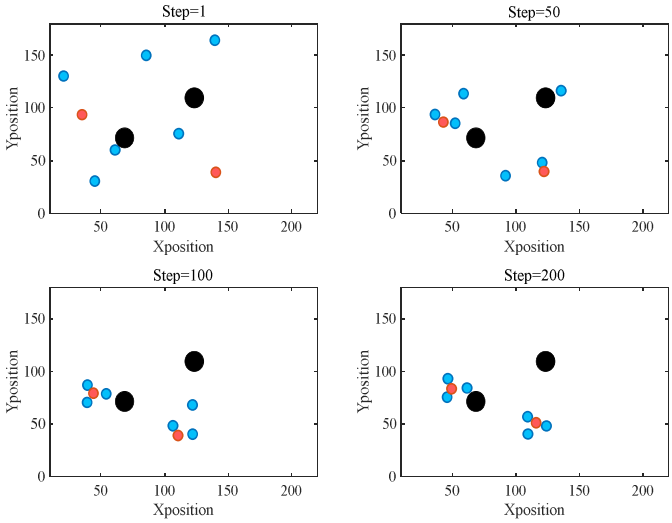

Figure 10. The process of rounding up in the complicated scenario with two targets.

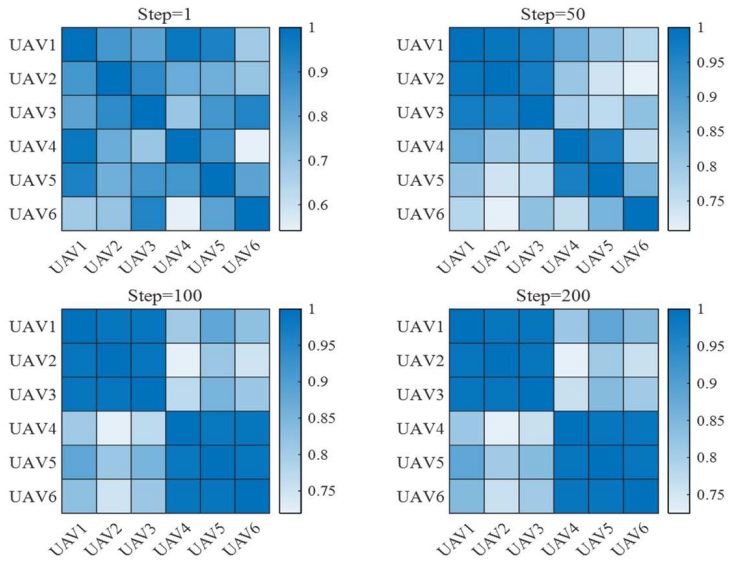

Figure 11. The evolutionary process of the agents' cognitions in the complicated two-target scenario.

As Figures 6 and 7 show, for the simple two-target scenario, the UAVs could successfully complete the task of rounding up the targets while avoiding collision between UAVs and between UAVs and obstacles. Moreover, the cognitive evolution of the UAVs is presented in Figure 8. As Figures 9 and 10 show, for the complicated two-target scenario, the UAVs can successfully complete the required task, Figure 11 depicts the cognitive evolution of the UAVs. Our results also convey that the UAVs were able to learn complex roundup strategies using the proposed method.

In Figures 8 and 11, the color changes in the heatmaps represent the evolution of the cognitive correlations between UAVs with regard to the mission. Initially, the positions of the UAVs and targets were randomly set, and the color distribution in the heatmap thus has no significant characteristics. As time passes, the UAVs' cognitions of the task gradually deepened. After 200 steps, the color distribution in the heatmap exhibits an obvious block distribution. This shows that the UAVs' cognitions of the subtasks gradually became consistent, and UAVs with similar cognitions were assigned to the same subtask to form a subtask alliance and cooperate to round up the same target. The above experiments verify the effectiveness of the task cognitive consistency method for a UAV group with a dynamic communication topology regardless of whether the simple scenario or the complicated scenario is utilized.

(2) In this section, we present two roundup scenarios designed to evaluate the performance of the proposed method: namely, scenario (a) with two targets and scenario (b) with three targets. In the experimental setting, the targets showed better performance than the UAVs. Specifically, the detection and communication ranges of the UAVs were restricted. To round up all targets, the UAVs needed to cooperatively take action to utilize the advantage of their greater quantity. The effectiveness of the proposed method in the two-target scenario has been demonstrated in Figures 6–11.

To validate the effectiveness of the proposed method in a three-target scenario, the global trajectories in this scenario are presented in Figures 12 and 13. Figure 14 shows the dynamic cognitive evolution. As Figures 12 and 13 show, the blue dots represent the agents, the red dots represent the targets, the black circles represent the obstacles, and the dashed triangles represent the roundup formation.

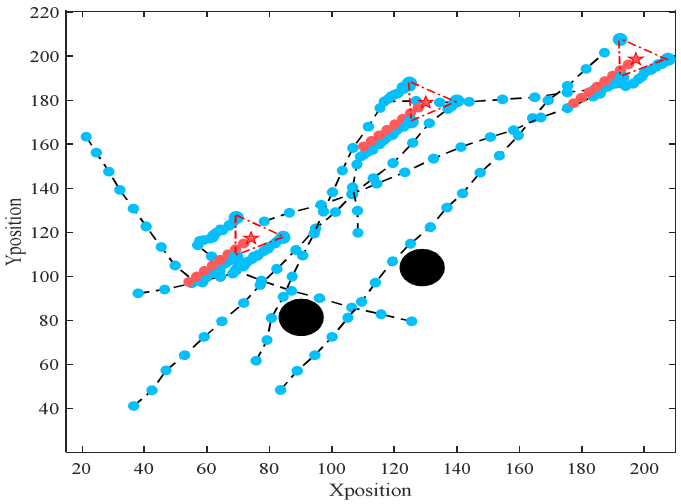

Figure 12. The trajectories for rounding up in the scenario involving three targets.

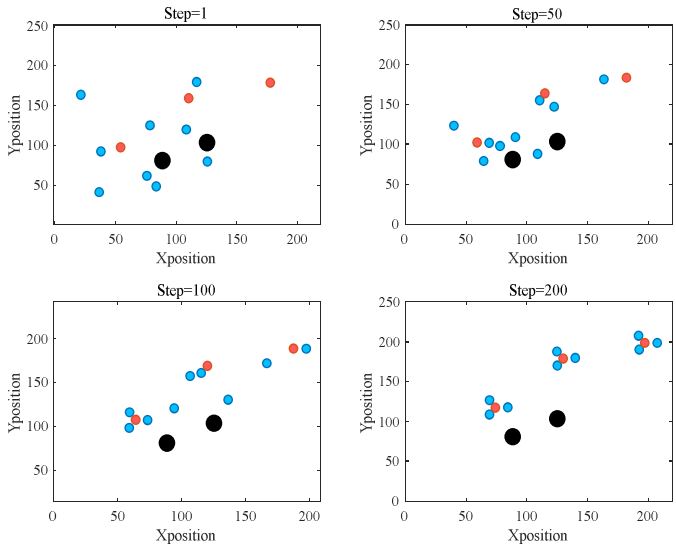

Figure 13. The process of rounding up in the scenario involving three targets.

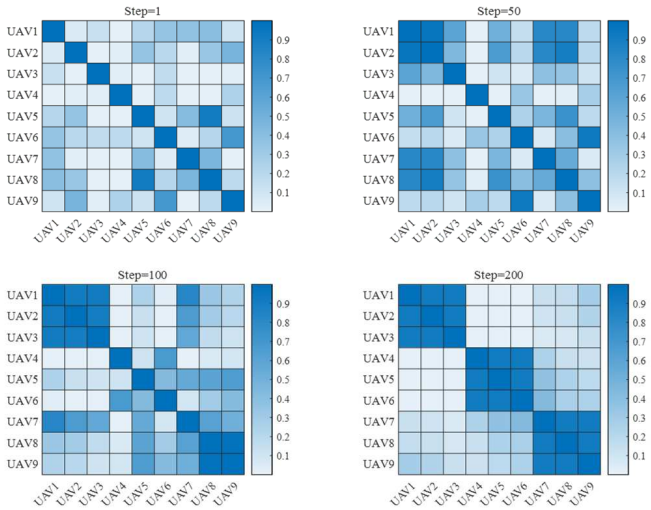

Figure 14. The evolutionary process of the agents' cognitions in the three-target scenario.

As Figures 12 and 13 show, when three targets were in play, the UAVs could successfully complete the task of rounding up the targets while avoiding collisions between UAVs and between UAVs and obstacles. Our results also convey that the UAVs were able to learn complex roundup strategies using the proposed method. Moreover, the cognitive evolution process is presented in the form of heatmaps in Figure 14. As Figure 14 shows, at step = 1, the cognitions varied among the agents. As time passed, the cognitions gradually converged and eventually exhibited cognitive consistency for each subtask, thus achieving the assignment of targets. At step = 200, the heatmap contains three partial blocks of darker colors.

(3) In this section, we present a cross-comparison between the HCCL method and the baselines introduced above to demonstrate the superior performance of our method.

During the algorithm learning, we trained the methods for 2000 episodes on a multi-target roundup scenario with two targets. We used the same hyperparameter settings as those of the VDN, and the detailed hyperparameters of all methods are shown in Table 2. The hyperparameter settings refer to Reference [42].

Table 2. The hyperparameter settings of the algorithms.

No.	Variable	Value
1	lr	0.001
2	γ	0.9
3	Episodes	2000
4	Batch_size	64
5	max_episode_length	200
6	α	0.2
7	β	0.8

To compare the different algorithms, the roundup success rate was used as an index to evaluate their performance. The higher the success rate is, the better the performance of the algorithm. The success rates of IDQN, VDN, CommNet, NCC-VDN, and HCCL for multi-target roundup are compared in Figure 15.

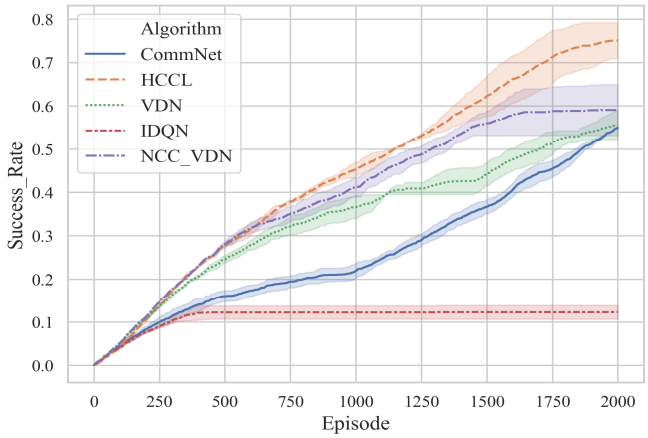

Figure 15. Success rates of different algorithms in a two-target roundup scenario.

Figure 15 presents the success rates of the different algorithms over 2000 episodes in the roundup task with two targets. As seen in Figure 15, the success rate of the IDQN algorithm is low. With an increasing number of simulation rounds, the success rate of the CommNet algorithm gradually increases, but it remains lower than those of the VDN, NCC-VDN, and HCCL, which shows that the latter three algorithms have better learning performance. The task success rate of the NCC-VDN algorithm converges to 59.1% at approximately 1600 rounds, while the HCCL method proposed in this paper continues

to show small increases in success rate and reaches a task success rate of 75.1% after approximately 2000 rounds. The performance of different algorithms is explored below.

As shown in Table 3, the IDQN lacks cooperation among its agents, and each agent only makes decisions according to its own reward, so the success rate remains at a low level. NCC-VDN is an improved algorithm based on the VDN. Because of the cognition consistency model, the task success rate of NCC-VDN is improved compared over that of the VDN algorithm. HCCL adopts a multiagent dynamic communication model, a hierarchical cognitive consistency model and an opponent graph reasoning model, and realizes cognitive task convergence through three levels of cognitive consistency constraints; thus, it produces the highest task success rate among the tested methods.

Table 3. The performance achieved by different algorithms after 2000 rounds.

Algorithms	IDQN	VDN	CommNet	NCC-VDN	HCCL
Success rate	12.3%	55.6%	55.1%	59.1%	75.1%

The reason for this continuing improvement may be that the hard attention mechanism discards some unimportant information, while the soft attention mechanism gives different weights to information of different levels of importance, which results in more accurate decisions made by the UAVs. Therefore, in these experiments, the proposed method obtained a substantially higher success rate than the baselines. From the above results, it can be seen that the hierarchical cognitive consistency learning method exhibits the best performance among the compared algorithms, which means that the proposed method can efficiently explore information and promote cooperation. Through cognitive consistency, the UAVs are able to cooperate efficiently and realize a consistent common belief.

With the aim of making quantitatively comparing the task performances achieved in different scenarios, the average number of steps is counted for each algorithm. The average number of steps is the number of steps required before all targets are rounded up in an episode. The simulations are conducted for 200 episodes. The task performance achieved in different scenarios is shown in Figure 16.

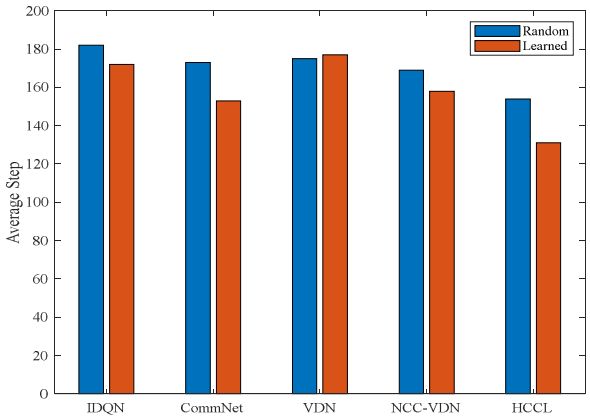

Figure 16. The task performance achieved in different scenarios.

Figure 16 shows the average numbers of steps required for successful round-ups of different scenarios. As shown in Figure 16, the HCCL algorithm requires the fewest steps to complete task, whether the targets adopt the randomized strategy or learned strategy. The reason for this finding may be that the HCCL combines the dynamic communication mechanism with an opponent reasoning graph. In contrast, the VDN, CommNet and NCC-VDN do not take the action of opponent into consideration, so those algorithms need more steps to round up. Because the IDQN does not possess intercommunication and acts independently, it needs the most steps to complete the task. In addition, the

above algorithms all require fewer steps in the simple scenario than a complicate scenario, which targets adopt the randomized strategy. In summary, the proposed HCCL algorithm enhances the effectiveness of the round-up process, which also verifies that the proposed method has better performance in the round-up task.

5.3. Ablation Study

The experiments reported in this section were conducted to address RQ (2). To investigate the effectiveness of the different components of our proposed method, we conducted the following ablation studies. There were three main components to our model: (1) the parameters of the reward function; (2) the target assignment; and (3) opponent graph reasoning. Ablation studies on these three major components were conducted under various multi-target roundup scenarios. In this section, we further verify how our contributions affect the learning process for multi-target roundup.

(1) The influence of the parameter settings of the reward function

From the previous analysis, we know that the larger the adjustment factor α, the more effort the UAVs will allocate to completing their own subtasks, whereas the larger the adjustment factor β, the more the UAVs will focus on global task completion efficiency. To further validate the effects of different values applied to the adjustment factors α and β on success rate in roundup tasks, we conducted simulations in the simple roundup scenario with two targets and validated the influence of the reward function parameter settings in three cases: (1) $\alpha = 0.8$ and $\beta = 0.2$; (2) $\alpha = 0.5$ and $\beta = 0.5$; (3) $\alpha = 0.2$ and $\beta = 0.8$.

As shown in Figure 17, when the adjustment factors were $\alpha = 0.2$ and $\beta = 0.8$, the success rate in the roundup task was the highest; when the adjustment factors were $\alpha = 0.5$ and $\beta = 0.5$, the success rate in the roundup task was second highest; and when the adjustment factors were $\alpha = 0.8$ and $\beta = 0.2$, the success rate was lower than when the adjustment factors were equal. The performance achieved after the 2000th round with different parameter settings is shown in Table 4.

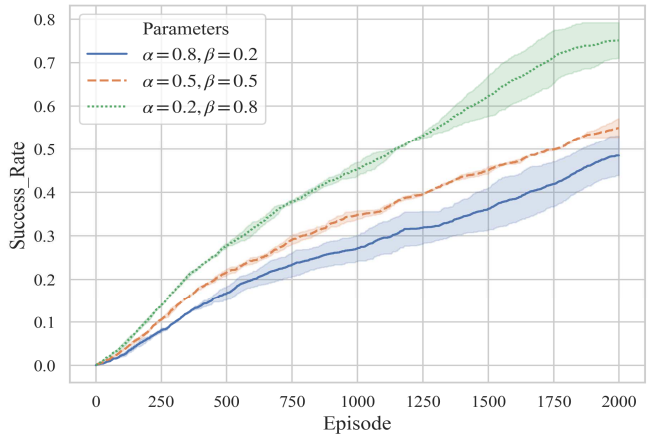

Figure 17. Success rates in roundup scenarios with different parameters.

Table 4. The performance achieved with different parameter settings after 2000 rounds.

Parameters	**α=0.8 β=0.2**	**α=0.5 β=0.5**	**α=0.2 β=0.8**
Success rate	48.5%	54.9%	75.1%

As shown in Table 4, in the 2000th round, the different parameter settings yield different performances. When the adjustment factors were $\alpha = 0.2$ and $\beta = 0.8$, the success rate in the roundup task was approximately 75.1%; when the adjustment factors were $\alpha = 0.5$ and $\beta = 0.5$, the success rate in the roundup task was approximately 54.9%; and when the adjustment factors were $\alpha = 0.8$ and $\beta = 0.2$, the task success rate was

approximately 48.5%. When the adjustment factors assign a larger weight to parameter β, it means that the reward pays more attention to the global task and achieves a higher success rate for the roundup task.

As shown in Figure 17 and Table 4, increasing the α or β can improve the success rate in the roundup task. A possible reason for this is that an increase in α or a reduction in β makes the multi-agent system pay more attention to global cooperation among agents to achieve an improved roundup effect.

(2) The influence of the target assignment model

We considered the influence of target assignment on the final performance. We replaced the proposed target assignment model with the assignment of neighboring UAVs based on distance; namely, "with neighborhoods", which meant that the UAVs closest to a target would be assigned to that target.

As seen in Figure 18, the clustering of the UAVs based on cognitive consistency achieved significantly better performance than clustering based on neighborhoods. Possible reasons for this may be that the target assignment model with neighborhoods can result in an unbalanced distribution of the subtasks, meaning that there may be large differences in the numbers of UAVs assigned to different targets, while the cognitive consistency assignment model results in a relatively balanced distribution. Since we assume that at least three UAVs are required to complete the roundup of each target, the allocation of a balanced number of UAVs makes a notable difference in the success rate, and the experimental results of this simulation verify the important impact of the target assignment method.

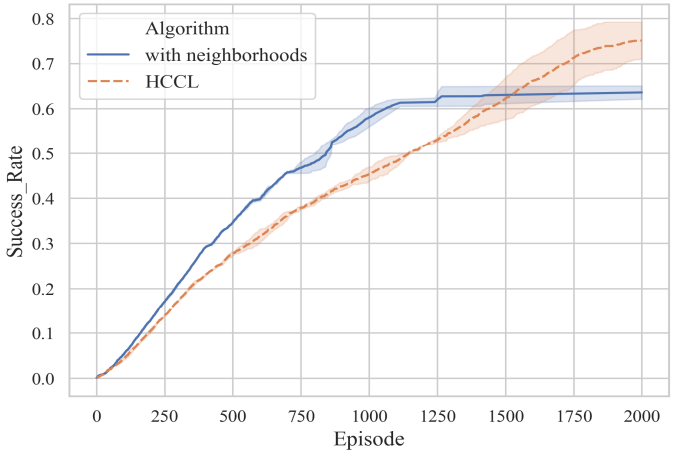

Figure 18. Success rates of rounding up with different target assignment models.

(3) The influence of the opponent graph reasoning model

In this section, we consider the influence of opponent graph reasoning on the final performance. To do so, we verified the variation in the success rate with and without the opponent graph reasoning model in an adversarial environment with two targets.

As shown in Figure 19, the method using the opponent graph reasoning model achieved a higher roundup success rate after 2000 episodes. In the initial stage of simulation, the network framework without the opponent graph reasoning model maintained a high task success rate. However, after approximately 1000 episodes, the growth in the success rate tended to slow, while the success rate with the opponent graph reasoning model continued to gradually increase. With an increasing number of episodes, after approximately 1250 episodes, the task success rate became higher than that without the opponent graph reasoning model. In the 2000th episode, the task success rate using the opponent graph reasoning model was 75.1%, and the task success rate without the opponent graph reasoning model was 59.3%. The curves indicate that the opponent graph reasoning model is beneficial to improving the success rate of roundup.

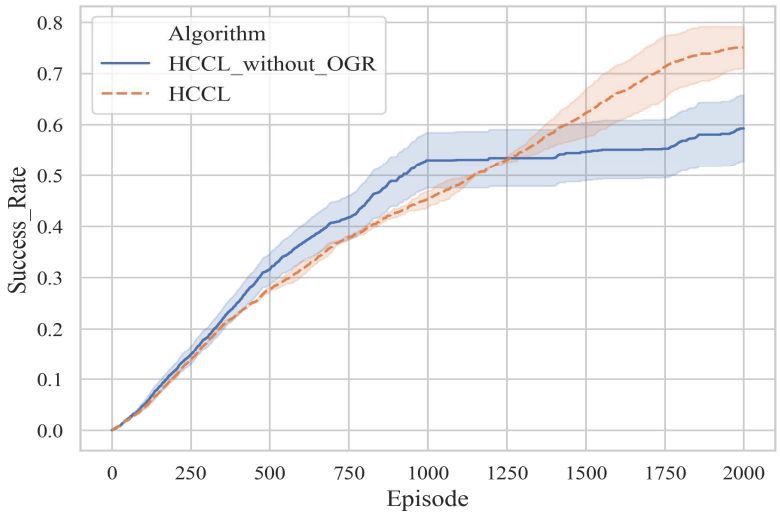

Figure 19. Success rates of rounding up with and without the opponent graph reasoning model.

A possible reason for the above observations may be that in the initial stage of training, the prediction of the target state is not sufficiently accurate, but due to the opponent graph reasoning model, the network complexity has a detrimental effect on the success rate. However, with an increasing number of episodes, the target state prediction becomes more accurate, and the UAVs can predict the target motion state and move closer to the ideal roundup positions in advance, thus improving the roundup efficiency. In contrast, if the opponent graph reasoning model is not adopted, the target positions at any given moment will be observed and responded to step by step, so target roundup cannot be realized as quickly as possible. The curves indicate that the opponent graph reasoning model is beneficial for improving the success rate of the roundup process.

In summary, all three components contribute to the superior performance of the HCCL method.

6. Conclusions

In this paper, we propose the HCCL method, a novel deep MARL method with decentralized policies and a centralized training setting, to solve the multi-target roundup problem for multi-UAV systems. For multi-target roundup, the HCCL method incorporates an inferential model for predicting the target states in an adversarial environment, and this method outperforms several baseline algorithms. In particular, the scalability of the HCCL method was verified by conducting simulation experiments for roundup scenarios with different numbers of targets. The superiority of the proposed target assignment model for multi-target roundup was verified by simulating different target assignment methods. In addition, the importance of the inferential target prediction model was verified by testing methods with and without the proposed opponent graph reasoning model. These ablation experiments further demonstrate the flexibility of the proposed roundup method and its ability to form consistent cognitions among multiple UAVs.

Regarding further work, the problems of further enhancement of the adversarial strategies as well as scaling up the number of agents await future theoretical and empirical analyses. Moreover, we are interested in exploring how UAVs can reach consensus through a dynamic communication topology in an adversarial environment to reduce the influence of dynamic environments.

Author Contributions: Conceptualization, investigation, methodology, writing—original draft preparation, resources, software, visualization, and validation, L.J.; project administration, funding acquisition and data curation, R.W.; formal analysis and writing—review and editing, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science and Technology Innovation 2030 Key Project of "New Generation Artificial Intelligence", China (No. 2018AAA0102403).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge Ministry of science and technology of China.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Tamakoshi, H.; Ishii, S. Multiagent reinforcement learning applied to a chase problem in a continuous world. *Artif. Life Robot.* **2001**, *5*, 202–206. [CrossRef]
2. Zhang, H.; Wu, L.; Zhou, Y.; Zhang, J.; Zhou, S.W.; Liu, Z.H. Self-organizing cooperative multi-target hunting by swarm robots in complex environments. *Control Theory Appl.* **2020**, *37*, 1054–1062.
3. Yang, B.; Ding, Y.; Jin, Y.; Hao, K. Self-organized swarm robot for target search and trapping inspired by bacterial chemotaxis. *Robot. Auton. Syst.* **2015**, *72*, 83–92. [CrossRef]
4. Gupta, S.; Hazra, R.; Dukkipati, A. Networked multi-agent reinforcement learning with emergent communication. *arXiv* **2020**, arXiv:2004.02780.
5. Deshpande, A.M.; Kumar, R.; Radmanesh, M.; Veerabhadrappa, N.; Kumar, M.; Minai, A.A. Self-organized circle formation around an unknown target by a multi-robot swarm using a local communication strategy. In Proceedings of the 2018 Annual American Control Conference (ACC), Milwaukee, WI, USA, 27–29 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4409–4413.
6. Pakizeh, E.; Palhang, M.; Pedram, M.M. Multi-criteria expertness based cooperative Q-learning. *Appl. Intell.* **2013**, *39*, 28–40. [CrossRef]
7. Wang, B.; Li, S.; Gao, X.; Xie, T. Weighted mean field reinforcement learning for large-scale UAV swarm confrontation. *Appl. Intell.* **2023**, *53*, 5274–5289. [CrossRef]
8. Luo, G.; Zhang, H.; He, H.; Li, J.; Wang, F.Y. Multiagent Adversarial Collaborative Learning via Mean-Field Theory. *IEEE Trans. Cybern.* **2021**, *51*, 4994–5007. [CrossRef]
9. Jiang, J.; Lu, Z. Learning attentional communication for multi-agent cooperation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*. [CrossRef]
10. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
11. Singh, A.; Jain, T.; Sukhbaatar, S. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv* **2018**, arXiv:1812.09755.
12. Mao, H.; Zhang, Z.; Xiao, Z.; Gong, Z. Modelling the Dynamic Joint Policy of Teammates with Attention Multi-agent DDPG. *arXiv* **2019**, arXiv:1811.07029.
13. Lowe, R.; Wu, Y.I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6382–6393.
14. Geng, M.; Xu, K.; Zhou, X.; Ding, B.; Wang, H.; Zhang, L. Learning to cooperate via an attention-based communication neural network in decentralized multi-robot exploration. *Entropy* **2019**, *21*, 294. [CrossRef] [PubMed]
15. Chen, S.; Chen, X.; Mei, Y.; Xie, J.; Fang, H. A Cooperative Hunting Algorithm of Multi-robot Based on Dynamic Prediction of the Target via Consensus-based Kalman Filtering. *J. Inf. Comput. Sci.* **2015**, *12*, 1557–1568. [CrossRef]
16. Chen, J.; Zha, W.; Peng, Z.; Gu, D. Multi-player pursuit–evasion games with one superior evader. *Automatica* **2016**, *71*, 24–32. [CrossRef]
17. Wu, S.; Pu, Z.; Liu, Z.; Qiu, T.; Yi, J.; Zhang, T. Multi-target coverage with connectivity maintenance using knowledge-incorporated policy framework. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 8772–8778. [CrossRef]
18. Kim, T.H.; Hara, S.; Hori, Y. Cooperative control of multi-agent dynamical systems in target-enclosing operations using cyclic pursuit strategy. *Int. J. Control* **2010**, *83*, 2040–2052. [CrossRef]
19. Awheda, M.D.; Schwartz, H.M. A decentralized fuzzy learning algorithm for pursuit-evasion differential games with superior evaders. *J. Intell. Robot. Syst.* **2016**, *83*, 35–53. [CrossRef]
20. Wang, X.; Xuan, S.; Ke, L. Cooperatively pursuing a target unmanned aerial vehicle by multiple unmanned aerial vehicles based on multiagent reinforcement learning. *Adv. Control Appl. Eng. Ind. Syst.* **2020**, *2*, e27. [CrossRef]
21. Yasuda, T.; Ohkura, K.; Nomura, T.; Matsumura, Y. Evolutionary swarm robotics approach to a pursuit problem. In Proceedings of the 2014 IEEE Symposium on Robotic Intelligence in Informationally Structured Space (RiiSS), Orlando, FL, USA, 9–12 December 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1–6.
22. Dutta, K. Hunting in groups. *Resonance* **2014**, *19*, 936–957. [CrossRef]
23. Fan, H.; Sun, F.Z.; Ma, P.L.; Li, W.J.; Shi, Z.; Wang, Z.J.; Zhu, G.J.; Li, K.; Yin, B. Stigmergy-Based Swarm Robots for Target Search and Trapping. *J. Beijing Inst. Technol.* **2022**, *42*, 158–167. [CrossRef]

24. Chu, T.; Chinchali, S.; Katti, S. Multi-agent reinforcement learning for networked system control. *arXiv* **2020**, arXiv:2004.01339.
25. Kim, D.; Moon, S.; Hostallero, D.; Kang, W.J.; Lee, T.; Son, K.; Yi, Y. Learning to schedule communication in multi-agent reinforcement learning. *arXiv* **2019**, arXiv:1902.01554.
26. Pu, Z.; Wang, H.; Liu, Z.; Yi, J.; Wu, S. Attention Enhanced Reinforcement Learning for Multi agent Cooperation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
27. Wang, H.; Pu, Z.; Liu, Z.; Yi, J.; Qiu, T. A Soft Graph Attention Reinforcement Learning for Multi-Agent Cooperation. In Proceedings of the 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), Hong Kong, China, 20–21 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1257–1262.
28. Wang, H.; Liu, Z.; Pu, Z.; Yi, J. STGA-LSTM: A Spatial-Temporal Graph Attentional LSTM Scheme for Multi-agent Cooperation. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 23–27 November 2020; Springer: Cham, Switzerland, 2020; pp. 663–675.
29. Iqbal, S.; Sha, F. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. *arXiv* **2018**, arXiv:181002912.
30. Chen, H.; Liu, Y.; Zhou, Z.; Hu, D.; Zhang, M. Gama: Graph attention multi-agent reinforcement learning algorithm for cooperation. *Appl. Intell.* **2020**, *50*, 4195–4205. [CrossRef]
31. Niu, Y.; Paleja, R.R.; Gombolay, M.C. Multi-Agent Graph-Attention Communication and Teaming. In Proceedings of the AAMAS, Online, 3–7 May 2021; pp. 964–973.
32. Huang, L.; Fu, M.; Rao, A.; Irissappane, A.A.; Zhang, J.; Xu, C. A Distributional Perspective on Multiagent Cooperation with Deep Reinforcement Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef]
33. Yan, L.; Zhu, L.; Song, K.; Yuan, Z.; Yan, Y.; Tang, Y.; Peng, C. Graph cooperation deep reinforcement learning for ecological urban traffic signal control. *Appl. Intell.* **2022**, *53*, 6248–6265. [CrossRef]
34. Ruan, J.; Du, Y.; Xiong, X.; Xing, D.; Li, X.; Meng, L.; Zhang, H.; Wang, J.; Xu, B. GCS: Graph-Based Coordination Strategy for Multi-Agent Reinforcement Learning. *arXiv* **2022**, arXiv:2201.06257.
35. Jiang, J.; Dun, C.; Huang, T.; Lu, Z. Graph convolutional reinforcement learning. *arXiv* **2018**, arXiv:1810.09202.
36. Du, Y.; Liu, B.; Moens, V.; Liu, Z.; Ren, Z.; Wang, J.; Chen, X.; Zhang, H. Learning correlated communication topology in multi-agent reinforcement learning. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, Online, 3–7 May 2021; pp. 456–464.
37. Wu, S.; Pu, Z.; Qiu, T.; Yi, J.; Zhang, T. Deep Reinforcement Learning based Multi-target Coverage with Connectivity Guaranteed. *IEEE Trans. Ind. Inform.* **2022**, *19*, 121–132. [CrossRef]
38. Rădulescu, R.; Verstraeten, T.; Zhang, Y.; Mannion, P.; Roijers, D.M.; Nowé, A. Opponent learning awareness and modelling in multi-objective normal form games. *Neural Comput. Appl.* **2022**, *34*, 1759–1781. [CrossRef]
39. Wu, S.; Qiu, T.; Pu, Z.; Yi, J. Multi-agent Collaborative Learning with Relational Graph Reasoning in Adversarial Environments. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 5596–5602. [CrossRef]
40. Ge, H.; Ge, Z.; Sun, L.; Wang, Y. Enhancing cooperation by cognition differences and consistent representation in multi-agent reinforcement learning. *Appl. Intell.* **2022**, *52*, 9701–9716. [CrossRef]
41. Wang, H.; Qiu, T.; Liu, Z.; Pu, Z.; Yi, J.; Yuan, W. Multi-Agent Cognition Difference Reinforcement Learning for Multi-Agent Cooperation. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–7. [CrossRef]
42. Mao, H.; Liu, W.; Hao, J.; Luo, J.; Li, D.; Zhang, Z.; Wang, J.; Xiao, Z. Neighborhood cognition consistent multi-agent reinforcement learning. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 7219–7226. [CrossRef]