2023 9th International Conference on Computer and Communications

Imitation Learning based Alternative Multi-Agent Proximal Policy Optimization for Well-Formed Swarm-Oriented Pursuit Avoidance

Sizhao Li[1], Yuming Xiang[1], Rongpeng Li[1]*, Zhifeng Zhao[2], Honggang Zhang[2]
[1]*College of Information Science and Electronic Engineering, Zhejiang University* [2]*Zhejiang Lab*
Hangzhou, 310001, Zhejiang, China
{liszh5, xiangym, lirongpeng}@zju.edu.cn, {zhaozf, honggangzhang}@zhejianglab.com

2023 9th International Conference on Computer and Communications (ICCC) | 979-8-3503-1725-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCC59590.2023.10507501

***Abstract*—Multi-Robot System (MRS) has garnered widespread research interest and fostered tremendous interesting applications, especially in cooperative control fields. Yet little light has been shed on the compound ability of formation, monitoring and defence in decentralized large-scale MRS for pursuit avoidance, which puts stringent requirements on the capability of coordination and adaptability. In this paper, we put forward a decentralized Imitation learning based Alternative Multi-Agent Proximal Policy Optimization (IA-MAPPO) algorithm to provide a flexible and communication-economic solution to execute the pursuit avoidance task in well-formed swarm. In particular, a policy-distillation based MAPPO executor is firstly devised to capably accomplish and swiftly switch between multiple formations in a centralized manner. Furthermore, we utilize imitation learning to decentralize the formation controller, so as to reduce the communication overheads and enhance the scalability. Afterwards, alternative training is leveraged to compensate the performance loss incurred by decentralization. The simulation results validate the effectiveness of IA-MAPPO and extensive ablation experiments further show the performance comparable to a centralized solution with significant decrease in communication overheads.**

***Index Terms*—Pursuit Avoidance, Adaptive Formation Control, Multi-Agent Reinforcement Learning, Imitation Learning**

I. Introduction

Nowadays, cooperative control in Multi-Robot Systems (MRS) has been attracting growing interest, since robotic swarm has demonstrated great potentials in both civilian and military tasks [1]. As for decentralized large-scale MRS, it is indispensable to develop the compound ability of Formation, Monitoring and pursuit Avoidance (FMA), which is quite prevalent in real-world. In that regard, [2] presents a two-level flocking control system for static obstacle avoidance and targeted positions navigation. [3] proposes a hybrid system for bypassing a pre-defined predator (attacker) trajectory, by integrating flocking control and Reinforcement Learning (RL), which implements invariant formation, thus lacking the essential flexibility. Furthermore, the considered scenarios are over-simplified and far from the reality [4]. Therefore, the process of FMA should be further calibrated to be more consistent with the practice, and effective formation control belongs to an inevitable ingredient for pursuit avoidance.

Benefiting from the robust adaptability in complex environments, Multi-Agent Reinforcement Learning (MARL) algorithms have been widely used to address formation control issues in MRS systems. For example, [5] combines Multi-Agent Proximal Policy Optimization (MAPPO) [6] with policy distillation [7] to achieve multi-formation assignment with global observations. [8] enhances the formation stability in communication-limited scenarios by incorporating an attention mechanism. Nevertheless, these formation control schemes are responsive to abrupt factors (e.g., battery life depletion) and can not be easily extended to scenarios with malicious actions of predators, which typically are un-predictable, since individual agents in the large-scale MRS can not spontaneously reach a policy consensus in a distributed manner. As a remedy, central scheme for formation control satisfies the demand on stability and synchronism. However, its heavy dependency on costly communications between agents and the central controller often hampers the application in large-scale MRS.

In order to establish a consensus for formation control in a distributed manner, a cooperative sequential decision-making framework is often leveraged, and individual agents attempt to discover beneficial behavior sequences distributively. However, such an attempt is often in vain, due to the difficulty to locate useful actions from the huge-dimensional joint action space from all agents. In other words, current MARL algorithms encounter severe challenges owing to the extreme sparsity of meaningful rewards from the environment [9]. Fortunately, Imitation Learning (IL) [10], which emulates a centralized policy-driven rational trajectories for decentralized execution, manifests itself in converging the policy. However, compared to a centralized policy, such an approach often degrades the performance due to the cumulative errors between practical execution trajectories and the imitated training samples [11].

In this paper, in order to enhance the flexibility of formation control in previous FMA work [2], [3] and ameliorate the practicality issue towards distributed formation control [5], we put forward a communication-efficient algorithm named Imitation learning based Alternative Multi-Agent Proximal Policy Optimization (IA-MAPPO). Compared with the existing work, the contribution of our paper can be summarized as follows.

This work was supported in part by the National Natural Science Foundation of China under Grants 62071425, in part by the Zhejiang Key Research and Development Plan under Grant 2022C01093, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010005.

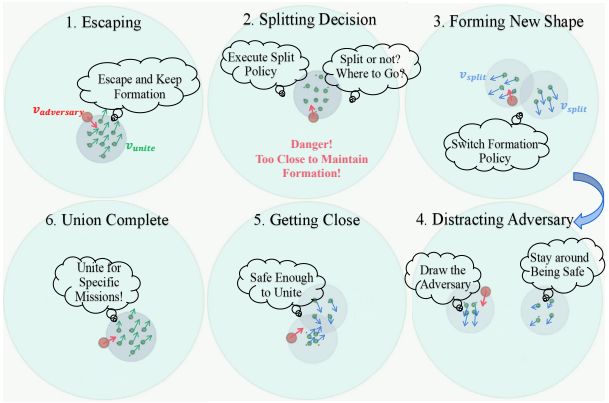


Fig. 1. Overview of well-formed swarm's pursuit avoidance with transformation from $\mathcal{F}_8$ to $\mathcal{F}_4$.

- We develop a hierarchical proactive formation control framework, where the high-level policy determines the division control policy (i.e., the appropriate time for formation switching), while the low-level policy resolves to consensus-oriented distributed mixed-formation policy following our previous works [8].
- As for the high-level policy, we first leverage IL to learn a distributed division policy that generates the formation confidence. After communicating with neighboring agents and aggregating the confidence, observation masking is applied on individual agents to match appropriate neighbors for the next formation. Furthermore, in order to compensate the IL-induced performance degradation, we adopt Alternative Training (AT) to fine-tune low-level policy on the basis of well-trained high-level policy.
- We validate the effectiveness of IA-MAPPO and extensive ablation experiments further show that IA-MAPPO yields competitive performance as a centralized solution and significantly lower communication overheads.

The remainder of this paper is organized as follows. We introduce the system model and formulate the problem in Section II. Afterwards, we elaborate on the details of the proposed IA-MAPPO algorithm in Section III. In Section IV, we present the simulation settings and discuss the experimental results. Finally, Section V concludes this paper.

II. System Model and Problem Formulation

A. System Model

We primarily consider a decentralized formation control problem, wherein agents in $\mathcal{N}$, with $|\mathcal{N}| = N$, are required to stay in the target area $\mathcal{T}$ in a communication-limited decentralized manner and transform within a set of pre-defined formation patterns $\{\mathcal{F}_c|\forall c \in \mathcal{C}\}$ where c represents the agent quantity in formation $\mathcal{F}_c$ and $\mathcal{C}$ denotes the set of possible quantities with $|\mathcal{C}| = C$. At time step t, each agent i needs to spontaneously determine one formation pattern c (i.e., $\mathcal{F}_i(t) = \mathcal{F}_c$) and recognize its $c-1$ neighbors cooperatively, resulting in $\chi(t)$ groups with the number $n_i, i \in \{1, \cdots, \chi(t)\}$ of agents in each group satisfying $n_1 + \cdots + n_{\chi(t)} = N$. An overview of the process is demonstrated in Fig. 1, wherein $c \in \mathcal{C} = \{8, 4\}$, and a formation transforms from F_8 to F_4 (implying $\chi(t)$ becomes 2 from 1) for pursuit avoidance.

In order to accomplish the basic FMA task for fixed formation, we formulate the problem as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [12], which is defined as $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \Omega, R, \mathcal{O}, \gamma \rangle$. In the FMA task, $\mathcal{I}$ represents N agents and an adversary that utilizes default policy for pursuit. Each agent $i \in \mathcal{N}$ maintains a joint policy $\boldsymbol{\pi} = \{\pi_f, \boldsymbol{\pi}_i\}$ where π_f denotes the division policy and $\boldsymbol{\pi}_i = \{\pi_i^c | \forall c \in \mathcal{C}\}$ represents the policy set for predefined formations. $\mathcal{S}$ denotes the global state space while $\mathcal{A}$ is the homogeneous action space for a single agent. Owing to the scant ability of perception against the colossal environment, each agent i obtains a local observation $\mathbf{o}_i \in \Omega$ via the observation function $\mathcal{O}(\mathbf{o}_i \mid \mathbf{s}, i) : \mathcal{S} \times \mathcal{N} \times \Omega \to [0, 1]$ instead of the state $\mathbf{s} \in \mathcal{S}$ at each time-step and adopts an action $\mathbf{a}_i \in \mathcal{A}$ according to mixed-formation policy $\boldsymbol{\pi}_i(\cdot \mid \mathbf{o}_i) : \Omega \times \mathcal{A} \to [0, 1]$. The joint action $\mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N\}$ taken at the current state $\mathbf{s}$ makes the environment transit into the next state $\mathbf{s}'$ according to the function $\mathcal{P}(\mathbf{s}' \mid \mathbf{s}, \mathbf{a}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. All agents share a global reward function $R(\mathbf{s}, \mathbf{a}) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with a discount factor γ. Consistent with the Dec-POMDP framework, we specify the elements as follows.

1) State and Observation: The raw information of agent i is denoted as $\mathbf{o}_i^-(t) = [\mathbf{p}_i(t), \mathbf{v}_i(t), \mathbf{p}_e(t)]$ which encompasses position $\mathbf{p}_i(t) = [p_{x_i}(t), p_{y_i}(t)]$, velocity $\mathbf{v}_i(t) = [v_{x_i}(t), v_{y_i}(t)]$ of agent i and the position of adversary $\mathbf{p}_e(t) = [p_{x_e}(t), p_{y_e}(t)]$ gained from sensors. After communicating with neighbors, agent i augments its observation by $\mathbf{o}_{\boldsymbol{\xi}_i}^-(t)$, where $\mathbf{o}_{\boldsymbol{\xi}_i}^-(t) = \{(\mathbf{p}_j(t), \mathbf{v}_j(t)) | \forall j \in \boldsymbol{\xi}_i(t)\}$ contains position and velocity of neighbors in $\boldsymbol{\xi}_i(t)$, who are within agent i's communication range δ_{com} (i.e., $\|\mathbf{p}_i(t) - \mathbf{p}_j(t)\| < \delta_{\text{com}}$), where $\|\cdot\|$ denotes an Euclidean norm operator. Thus, the observation of agent i is summarized as $\mathbf{o}_i(t) = [\mathbf{o}_i^-(t), \mathbf{o}_{\boldsymbol{\xi}_i}^-(t)]$. On the other hand, the global state includes the positions and velocities of all agents and the adversary, that is, $\mathbf{s}(t) = [\{\mathbf{o}_i^-(t) | \forall i \in \mathcal{N}\}, \{\mathbf{p}_e(t), \mathbf{v}_e(t)\}]$.

2) Action: Based on the local observation $\mathbf{o}_i(t)$, each agent sets its acceleration $\mathbf{a}_i(t) = [a_{x_i}(t), a_{y_i}(t)] \in \mathcal{A}$ following policy $\pi_i^c(\cdot \mid \mathbf{o}_i)$ individually to complete FMA task in a $\pi_f(t)$-determined formation $\mathcal{F}_i(t) = \mathcal{F}_c$[1]. Later in Section III, we shall further discuss the determination of $\boldsymbol{\pi}$ within the framework of IA-MAPPO.

3) Reward: In this part, we introduce the reward designed for a specific number c of swarm in FMA task.

- **Formation reward**. Towards implementing leader-free formation control and enhancing the robustness, we adopt the Hausdorff Distance (HD) [13] to measure the topology distance between the current and the expected formation. We denote the relative positions of agents as $\mathbf{P}(t) = \{\mathbf{p}_j(t) - \bar{\mathbf{p}}(t) | \forall j \in \mathcal{F}_c\}$ where $\bar{\mathbf{p}}(t)$ is the center of the swarm and $\mathcal{F}_c$ is the target formation corresponding

[1]We slightly abuse the terminology here, as the action shall implicitly involves the division control policy π_f as well.

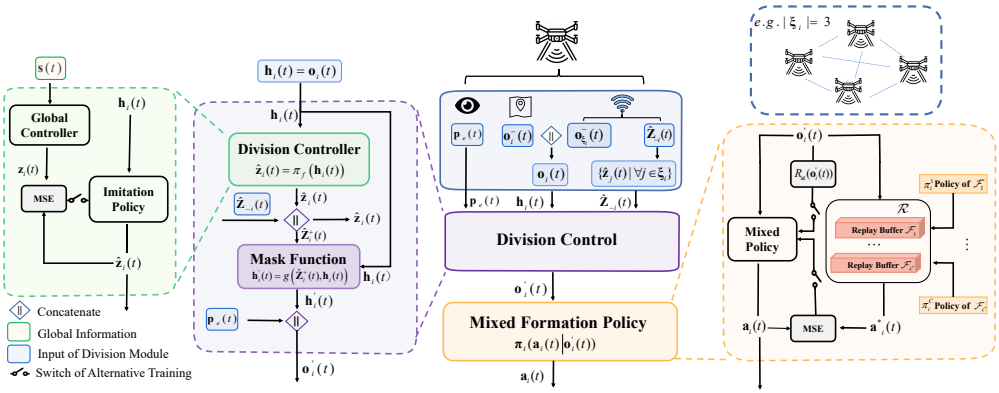


Fig. 2. The illustration of IA-MAPPO with Policy Distillation, Imitation Learning and Alternative Training.

to c. Thus the formation reward can be given by

$$R_f(t) = -k_f \max\{(d_{\text{HD}}(\mathbf{P}(t), \mathcal{F}_c) - \varphi_f, 0\}, \quad (1)$$

where k_f denotes the coefficient of formation punishment and the HD between current and ideal topology is defined as $d_{\text{HD}}(\mathbf{P}(t), \mathcal{F}_c) = \max\{h_d(\mathbf{P}(t), \mathcal{F}_c), h_d(\mathcal{F}_c, \mathbf{P}(t))\}$ with $h_d(\mathbf{P}(t), \mathcal{F}_c) = \max_{\mathbf{x}\in\mathbf{P}(t)} \min_{\mathbf{y}\in\mathcal{F}_c} \|\mathbf{x} - \mathbf{y}\|$. The hyper-parameter φ_f is used to fine-tune the trade-off between the formation accuracy and convergence rapidity.

- **Target area reward**. In order to maintain the swarm in target area, we assume the target area $\mathcal{T}$ as a circular region with radius $r_{\mathcal{T}}$, the center of which is set at the original point (i.e., $\mathcal{T} : \{(x, y) \mid \|(x, y)\| \leqslant r_{\mathcal{T}}\}$), and the target area reward can be formulated as

$$R_a(t) = \sum\nolimits_{i\in\mathcal{N}} R_{a_i}(t), \quad (2)$$

where $R_{a_i}(t)$, which equals $-k_a e^{\alpha\|\mathbf{p}_i(t)\|}$ if $\|\mathbf{p}_i(t)\| > r_{\mathcal{T}}$ and nulls otherwise, represents the punishment for the agent i exceeding the circle. k_a and α are both constants.

- **Evasion reward**. In order to keep agents away from the adversary, a penalty would be added as the evasion reward if any agent collides with the adversary,

$$R_e(t) = -k_e \sum\nolimits_{i\in\mathcal{N}} (\|\mathbf{p}_i(t) - \mathbf{p}_e(t)\| < \delta_{\text{safe}}), \quad (3)$$

where k_e is the coefficient of evasion reward and δ_{safe} is the minimum safe distance between agents and the adversary.

To conclude, the final reward function is as

$$R(t) = R_f(t) + R_a(t) + R_e(t). \quad (4)$$

B. Problem Formulation

Without loss of generality, at time-step t, each agent should unanimously determine the swarm division and decide the belonging group. Dependent on the centralized or decentralized controller, the algorithm takes either the global state $\mathbf{s}(t)$ or the local observation $\mathbf{o}_i(t)$ as the input. Therefore, we generally denote the input as $\mathbf{h}_i(t)$ for simplicity of representation. Correspondingly, the division policy π_f generates the separation confidence at time step t with $\mathbf{z}_i(t) = \pi_f(\mathbf{h}_i(t))$. Furthermore, agents share the separation confidence with neighbors and concatenate these received values as the enhanced confidence (i.e., $\mathbf{Z}_i^+(t)$, $i \in \{1, \cdots, N\}$). After identifying the formation c based on $\mathbf{Z}_i^+(t)$, agent i reshapes its local observation as $\mathbf{o}_i'(t) = g\left(\mathbf{o}_i(t), \mathbf{Z}_i^+(t)\right)$ and computes the action $\mathbf{a}_i(t) = \pi_i^c\left(\mathbf{o}_i'(t)\right)$, so as to realize a new formation c together with new neighbors. Then reward $R(t)$ is computed to evaluate the impact of the joint action $\mathbf{a}(t) = \{\mathbf{a}_1(t), \mathbf{a}_2(t), \cdots, \mathbf{a}_N(t)\}$. As for distributed system, we let ϵ to indicate average compound error over time T relative to centralized system as [11] discussed, while $\epsilon = 0$ in centralized scheme. Here we propose the communication overheads $D(\pi_f, \boldsymbol{\pi}_i)$ as the objective of policy optimization, which can be formulated as

$$\begin{aligned} &\min_{\pi_f, \boldsymbol{\pi}_i} D(\pi_f, \boldsymbol{\pi}_i) \\ \text{s.t. } &\sum\nolimits_{t=1}^{T} (R(t) - \epsilon) \geqslant r_{\text{thre}}, \\ &\mathbf{a}_i(t) = \pi_i^c\left(g\left(\mathbf{o}_i(t), \mathbf{Z}_i^+(t)\right)\right) \\ &\mathbf{Z}_i^+(t) = [\mathbf{z}_i(t) \| \{\mathbf{z}_j(t) | \forall j \in \boldsymbol{\xi}_i(t)\}] \\ &\mathbf{z}_i(t) = \pi_f\left(\mathbf{h}_i(t)\right), \qquad \forall i \in \{1, \cdots, N\} \end{aligned} \quad (5)$$

where r_{thre} denotes the required minimum cumulative reward and $\mathbf{h}_i(t) = \mathbf{s}(t)$, if it adopts a central policy π_f^*; otherwise, $\mathbf{h}_i(t) = \mathbf{o}_i(t)$, where $\|$ indicates the concatenate operation. Notably the communication overheads D heavily depend on the means for controlling (i.e., centralized, leader-follower [14] and decentralized policies denoted as π_f^*, $\pi_f^{*'}$ and π_f). Beforehand, we denote overheads for positions of agent and adversary respectively as Ω_{p_i} and Ω_{p_e} while Ω_z indicates the instruction costs and Ω_s for the expenses of sequence number in decentralized system.

- **Centralized**. N positions in Ω_{p_i}-bit of all agents along with Ω_{p_e}-bit adversary's position are needed for centralized controller, based on which it generates instructions in Ω_z-bit for each agent. Thus overheads are summarized as $D(\pi_f^*, \boldsymbol{\pi}_i) = N\Omega_{p_i} + \Omega_{p_e} + N\Omega_z$.
- **Leader-Follower**. One agent is selected as the leader to generates Ω_z-bit instructions for $N-1$ followers based on $N-1$ positions of followers in Ω_{p_i}-bit while adversary's position is locally obtained. The communication overheads are $D(\pi_f^{*'}, \boldsymbol{\pi}_i) = (N-1)\Omega_{p_i} + (N-1)\Omega_z$.
- **Decentralized**. Division confidence $\hat{\mathbf{z}}_i(t)$ is computed individually without up-link cost. Along with the Ω_s-bit sequence number, the local confidence $\hat{\mathbf{z}}_i(t)$ in Ω_z bits is broadcast to $|\boldsymbol{\xi}_i(t)|$ neighbors for agent i, so the total overheads of fully-decentralized system are $D(\pi_f, \boldsymbol{\pi}_i) = \sum_{i\in\mathcal{N}} |\boldsymbol{\xi}_i(t)|(\Omega_s + \Omega_z)$.

Since the overheads of Ω_z and Ω_s are both in bit level while Ω_{p_i} and Ω_{p_e} cost few bytes to store positions, the overall overheads can be significantly reduced in distributed execution relative to the other two systems. Therefore, we try to develop a decentralized communication-efficient solution.

III. The Framework of IA-MAPPO

In this section, we present details of the hierarchical structure IA-MAPPO, in which the low-level mixed-formation policy determines where to go and how to form different formations, and the up-level division control policy decides when to deconstruct and reform new shapes. Beforehand, we delve into the details of MAPPO [6], which shall constitute the basis of IA-MAPPO.

A. MAPPO-based Specific Formation Control

The conventional MAPPO algorithm [6] aims to obtain a policy π_i^c for fixed formation $\mathcal{F}_c$ in FMA. Specifically, MAPPO adopts the CTDE architecture and utilizes importance sampling to stabilize the learning progress. Notably, we assume that all agents are homogeneous in the environment, and thus each agent executes the same individual policy π_i^c with parameter θ. In particular, MAPPO uses an old-version $\pi_{i,\theta_{\text{old}}}^c$ and value function $V_{\psi_{\text{old}}}$ to interact with the environment, store decision trajectories, and calculate the ratio $\mu_i(t) = \frac{\pi_{i,\theta}^c(\mathbf{a}_i(t)|\mathbf{o}_i(t))}{\pi_{i,\theta_{\text{old}}}^c(\mathbf{a}_i(t)|\mathbf{o}_i(t))}$ by the target policy $\pi_{i,\theta}^c$. Afterwards, parameters θ and ψ of actor networks and critic networks respectively are periodically updated to maximize

$$\begin{aligned} J_{\pi_i}(\theta) =& \min\left(\mu_i(t)\hat{A}(t), \text{clip}\left(\mu_i(t), 1-\varepsilon, 1+\varepsilon\right)\hat{A}(t)\right), \\ J_V(\psi) =& -\left(V_\psi\left(\mathbf{s}(t)\right) - \left(\hat{A}(t) + V_{\psi_{\text{old}}}\left(\mathbf{s}(t)\right)\right)\right)^2, \end{aligned} \tag{6}$$

where $\hat{A}(t) = \sum_{l=0}^{T-t-1}(\gamma\lambda)^l\delta(t+l)$ is the advantage estimation function at time step t with $\delta(t) = R(t) + \gamma V_{\psi_{\text{old}}}(\mathbf{s}(t+1)) - V_{\psi_{\text{old}}}(\mathbf{s}(t))$, within which $R(t)$ is defined in (3), while $\mathbf{s}(t)$ denotes the global state and ε is a hyperparameter. Based on the classical MAPPO, the swarm has the capacity to accomplish the function of monitoring the target area and avoiding an adversary in a fixed formation $\mathcal{F}_c$.

B. Mixed-Formation Policy

Following Section III-A, we can obtain a set of policies respectively for formations $\{\mathcal{F}_c|\forall c \in \mathcal{C}\}$. In this part, we regard these policies as teacher models (i.e., a teacher model π_i^c instructs the formation $\mathcal{F}_c$), and utilize policy distillation [7] to obtain a mixed-formation policy, so as to reduce local memory occupation. As shown in Fig. 2, we collect both inputs and outputs of policies $\{\pi_i^c|\forall c \in \mathcal{C}\}$ for multi formation patterns by constituting a replay memory $\mathcal{B}$ as

$$\mathcal{B} = \langle(\mathbf{o}_1, \mathbf{a}_1), \ldots, (\mathbf{o}_C, \mathbf{a}_C)\rangle_{\times\Lambda}, \tag{7}$$

where $\mathbf{a}_c$ is generated by a learned teacher model π_i^c to form $\mathcal{F}_c$ and Λ is the capacity of replay buffer. Considering the dimension of observation $\mathbf{o}_c$ is determined by agent number c, we align the vectors of observations in different formations by zero-padding operation. During training, memories $\langle\mathbf{o}_c, \mathbf{a}_c\rangle$ from teacher models in buffer are utilized by mixed-formation policy π_s via minimizing Mean-Squared-Error (MSE) loss

$$\mathcal{L}_{\text{PD}}(\theta_s) = \sum\nolimits_{i=1}^{\Lambda}\sum\nolimits_{c=1}^{C}\left\|\pi_i^c(\mathbf{a}_c|\mathbf{o}_c) - \pi_s(\mathbf{a}_s|\mathbf{o}_c;\theta_s)\right\|, \tag{8}$$

where θ_s is the parameter of π_s, and $\mathbf{a}_c$ along with $\mathbf{a}_s$ refer to the actions specific to the same state $\mathbf{o}_c$ respectively from teacher models π_i^c and the student model π_s. After updating θ_s through (8), the mixed-formation policy set $\boldsymbol{\pi}_i = \{\pi_i^c|\forall c \in \mathcal{C}\}$ is simplified to one student policy $\boldsymbol{\pi}_i = \pi_s$, which significantly saves the usage of agents' memory.

C. Division Control

With the help of policy distillation, the low-level policy π_s is capable to realize C formation patterns, assuming the availability of the teammate matrix $\mathbf{M}_i^{\mathcal{C}}$ of agent i, with the size $C \times N$. In other words, agent i can recognize its $c-1$ neighbors when it selects $\mathcal{F}_i(t) = \mathcal{F}_c$ by referring to the teammate vector $\mathbf{m}_i^c$ with size $1 \times N$ in matrix $\mathbf{M}_i^{\mathcal{C}}$. Therefore, we only need to learn a division control policy π_f for formation pattern switching. In particular, the division control consists of policy imitation, confidence communication and masking operation.

1) Policy Imitation: For a completely centralized system, the controller can adopt a policy π_f^* to notify a specific formation pattern $\mathcal{F}_c$ to agent i (i.e., an instruction of a C-length one-hot vector $\mathbf{z}_i(t)$ is given that the c-th element $z_i^c(t)$ is equal to 1 while others being 0). In advance, specific intervals δ_{safe}^c are designed for each pattern $c \in \mathcal{C}$. Mathematically,

$$z_i^c(t) = \begin{cases} 1, & \delta_{\min}^c \le \beta(t) < \delta_{\max}^c; \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $\beta(t) = \|\frac{1}{|\mathcal{F}_i(t)|}\sum_{j\in\mathcal{F}_i(t)}(\mathbf{p}_j(t) - \mathbf{p}_e(t))\|$ represents the distance between the center of formation $\mathcal{F}_i(t)$ and the adversary $\mathbf{p}_e(t)$, while $\delta_{\text{safe}}^c = [\delta_{\min}^c, \delta_{\max}^c)$ denotes the safety distance interval for formation pattern c. Apparently, the centralized control process is energy-consuming in obtaining positions of all agents.

Targeted at getting rid of the central controller and reducing communication overheads, we assume that the featureless agent i obtains $\mathbf{h}_i(t) = \mathbf{o}_i(t)$ with raw information from neighbors $\mathbf{o}_{\boldsymbol{\xi}_i}^-(t)$ by communications. After that, the C-dimension formation confidence vector $\hat{\mathbf{z}}_i(t)$ is generated by a decentralized policy network π_f parameterized by θ_f which is shared among agents, denoted as

$$\hat{\mathbf{z}}_i(t) = \pi_f(\mathbf{h}_i(t)), \tag{10}$$

the element of which $\hat{z}_i^c(t)$ represents the derived confidence to pattern c for agent i in a distributed manner.

We adopt IL to update the parameter θ_f. Specifically, in the data-collecting phase, $\mathbf{z}_i(t)$ is computed by the central policy according to (9) and then the state-action pairs $(\mathbf{h}_i(t), \mathbf{z}_i(t))$ are stored at each time-step t. In the training phase, we define $\mathbf{z}_i(t)$ as the label of $\hat{\mathbf{z}}_i(t)$ and compute the MSE loss as

$$\mathcal{L}(\theta_f) = \sum\nolimits_{t=0}^{T}\sum\nolimits_{i\in\mathcal{N}}\left\|\pi_f\left(\mathbf{h}_i(t);\theta_f\right) - \mathbf{z}_i(t)\right\|, \tag{11}$$

where T is the length of each episode. Thus an estimator of global controller $\hat{\mathbf{z}}_i(t)$ can be attained in a decentralized manner from π_f.

2) Confidence Communication: To enhance the cooperation and stability of the system, it is essential for agents to communicate their separation confidence to understand each other. After calculating $\hat{\mathbf{z}}_i(t)$ by π_f, agent i broadcasts the result to neighbors in δ_{com}. Meanwhile, it receives neighbors' confidence collected as a vector $\{\hat{\mathbf{z}}_j(t)|\forall j \in \boldsymbol{\xi}_i(t)\}$, denoted as $\hat{\mathbf{Z}}_{-i}(t)$. Up to this point, the total division confidence that agent i can partially gather is $\hat{\mathbf{Z}}_i^+(t) = [\hat{\mathbf{z}}_i(t)\|\hat{\mathbf{Z}}_{-i}(t)]$.

3) Masking Operation: The masking operation transfers the raw observation $\mathbf{h}_i(t)$ into $\mathbf{h}_i'(t)$ to meet the input of low-level mix-formation policy π_s, and it involves three sub-procedures.

- **Information Aggregation**. $\hat{\mathbf{Z}}_i^+(t)$ contains all neighbors' intentions of which formation pattern to choose at next time step. Recalling the definition of $\hat{\mathbf{Z}}_{-i}(t)$ and $\hat{\mathbf{z}}_i(t)$, $\hat{\mathbf{Z}}_i^+(t)$ is a matrix with the size of $|\boldsymbol{\xi}_i(t)+1|\times C$. In order to aggregate preferences in the domain $\boldsymbol{\xi}_i(t)$ of agent i, we apply column summation to $\hat{\mathbf{Z}}_i^+(t)$ to figure out the importance of each formation pattern
$$\phi_i^c(t) = \sum\nolimits_{j\in\{i\cup\boldsymbol{\xi}_i(t)\}} \hat{z}_j^c. \tag{12}$$
Correspondingly, we further obtain a C-length importance vector $\boldsymbol{\phi}_i(t)$.
- **Formation Pattern Determination**. After obtaining the importance vector $\boldsymbol{\phi}_i(t)$, we determine the formation pattern $c^\sharp$ for next time-step $t+1$ by locating the index $c^\sharp$ corresponding to the maximum value in $\boldsymbol{\phi}_i(t)$.
- **Observation Reshape**. As preliminary assumed, agent i refers the set of teammates $\mathbf{m}_i^{c^\sharp}$ from the teammates matrix $\mathbf{M}_i^{\mathcal{C}}$ in its memory for current pattern $c^\sharp$. Then it computes $\mathbf{h}_i(t)$ and updates
$$\mathbf{h}_i'(t) = [\mathbf{o}_i^-(t)\|\{\mathbf{o}_j^-(t)|\forall j \in \mathbf{m}_i^{c^\sharp}\}]. \tag{13}$$
Notably, due to the adoption of IL from central policy, the precondition $\mathbf{m}_i^{c^\sharp} \subseteq \boldsymbol{\xi}_i(t)$ of transition $\mathbf{h}_i'(t) \leftarrow \mathbf{h}_i(t)$ is always satisfied.

D. Alternative Training

The IL-induced decentralized division policy π_f degrades the performance relative to central policy π_f^* because of compounding deviations, which leads to the deterioration of $R(t)$ as (5). Theoretically, it is proved in [11] as
$$\mathbb{E}_{s\sim d_{\pi_f}}\left(C_\pi(s)\right) \le \mathbb{E}_{s\sim d_{\pi_f^*}}\left(C_\pi(s)\right) + T^2\epsilon, \tag{14}$$
where $C_\pi(s)$ denotes the expected episodic cost of performing policy π in state s. The distribution d_π encodes the state visitation frequency over T steps with policy π, and ϵ represents the average compounding deviations over time T between π_f and π_f^*. As indicated in (14), the extra cost for imitative policy π_f grows quadratically and can result in tremendous performance loss.

To rectify the adverse effects incurred by IL, we leverage AT by fine-tuning mixed-formation policy π_s after learning the division policy π_f. Therefore, through interacting with the environment by the joint actions $\mathbf{u}_t$ of division instructions $\hat{\mathbf{z}}_i(t)$ and acceleration actions $\mathbf{a}_i(t)$ respectively from π_f and π_s, we first collect trajectories $\langle(\mathbf{o}_1,\mathbf{u}_1),\ldots,(\mathbf{o}_T,\mathbf{u}_T)\rangle$. Different from manual trajectory labeling in [11], we design a task-oriented alternative training reward R_{at} to evaluate the state $\mathbf{o}_t$. Specifically, in order to rectify the undesired situations (e.g., the reluctancy to unite due to the over-long inter-group distance), R_{at} is designed as
$$R_{\text{at}}(\mathbf{o}_t) = -\sum\nolimits_{i\neq j} \|\bar{\mathbf{p}}_{q_i}(t) - \bar{\mathbf{p}}_{q_j}(t)\|, \tag{15}$$
where q_i and q_j are two arbitrary groups in $\{q_1,\ldots,q_{\chi(t)}\}$, and both $\bar{\mathbf{p}}_{q_i}$ and $\bar{\mathbf{p}}_{q_j}$ signify the center of them. (15) evaluates $\mathbf{o}_t$ by computing average distance between two groups. Afterwards, we transfer the tuples into trajectories $\langle(\mathbf{o}_1,\mathbf{u}_1,\mathbf{o}_2,R_{\text{at}}(\mathbf{o}_1)),\ldots,(\mathbf{o}_T,\mathbf{u}_T,\mathbf{o}_{T+1},R_{\text{at}}(\mathbf{o}_T))\rangle$ and fine-tune the parameter θ of π_s according to (6).

IV. Experimental Results and Discussions

In this section, we evaluate the performance of IA-MAPPO in multi-agent particle environment [15] for FMA task, so as to verify the reduction in the communication overheads, on the basis of competitive performance as the centralized system.

Beforehand, we select RAN-MAPPO, CE-MAPPO, AT-MAPPO, LF-MAPPO, and IL-MAPPO as baselines. In particular, RAN-MAPPO refers to the algorithm with randomly generated division instructions on top of a mixed-formation policy π_s, while CE-MAPPO indicates the centralized division policy π_f^* with π_s. On top of CE-MAPPO, AT-MAPPO denotes the CE-MAPPO with π_s being further fine-tuned via AT, while LF-MAPPO substitutes the leader-follower policy $\pi_f^{*'}$ for π_f^* in CE-MAPPO. IL-MAPPO implies the absence of AT in IA-MAPPO. Components of these algorithms are summarized in Table I.

In the simulation, we focus on the FMA task that requires transformation between $\mathcal{F}_8$ in low velocity v_{unite} and $\mathcal{F}_4$ in high velocity v_{split} with limited communication range δ_{com} for each agent. The swarm is expected to avoid attacks from the adversary and stay in a target area with radius $r_{\mathcal{T}}$. An adversary keeps pursing at the speed of $v_{\text{adversary}}$, whose policy is chasing the center of swarm all the time. The performance of system is quantified by the reward defined in (4). The key parameters are summarized in Table II.

A. Communication Overheads

Table III concludes the communication costs of six systems that respectively represents centralized, leader-follower and decentralized architectures, when $[\Omega_{p_i},\Omega_{p_e},\Omega_z,\Omega_s]$ is set as $[64,64,1,3]$ bits, since a single-bit Ω_z is required due to $C=2$ and 3-bit Ω_s stands for sequence number as $n_a=8$. Besides, Ω_{p_i} and Ω_{p_e} represent (p_x,p_y) of agents and adversary in

TABLE I
Component Comparison of Algorithms.

Name(-MAPPO)	RAN	CE	LF	AT	IL	IA
Decentralized	○	○	◑	○	●	●
With AT	○	○	○	●	○	●

TABLE II
THE REMAINING KEY PARAMETER SETTINGS.

Parameters	Value
Number of Agents and Adversary	$n_a = 8, n_e = 1$
Discount Factor	$\gamma = 0.8$
Communication Range	$\delta_{\text{com}} = 2$ m
Radius of Target Area	$r_{\mathcal{T}} = 8$ m
Time Steps each Episode	$T = 200$ s
Agent Number of each Formation	$c_1 = 8, c_2 = 4$
Safe Distance	$\delta^4_{\text{safe}} = [0, 2.8), \delta^8_{\text{safe}} = [2.8, \infty)$ m
Parameters $(k_f, k_a, k_e, \varphi_f, \alpha, \varphi_m)$	$(1, 0.3, 20, 0.75, 0.5, 100)$
Maximum of Velocity	$v_{\text{unite}} = 0.5$ m/s, $v_{\text{split}} = 1.2$ m/s, $v_{\text{adversary}} = 0.6$ m/s

TABLE III
AVERAGE COMMUNICATION OVERHEADS EACH EPISODE IN TYPICAL SYSTEMS FOR DIVISION CONTROL.

Methods	Up-link	Down-link	Overall Cost
CE/RAN-MAPPO	11.7 KB	1.44 KB	13.14 KB
AT-MAPPO	12.47 KB	1.53 KB	14.0 KB
LF-MAPPO	10.02 KB	0.15 KB	10.18 KB
IL-MAPPO	–	4.48 KB	4.48 KB
IA-MAPPO	–	5.15 KB	5.15 KB

32-bit floating-point format. It is consistent with the analysis in Section II-B that the decentralized IL/IA-MAPPO decrease the communication overheads respectively to $34.1\%/39.2\%$ of CE-MAPPO and $44.0\%/50.6\%$ of LF-MAPPO.

B. Pursuit Avoidance Performance

Fig. 3 presents the pursuit avoidance performance of IA-MAPPO and other algorithms. Notably, RAN-MAPPO exhibits worst performance due to severe communication disconnections and collisions with the adversary in Fig. 4(a). It can be observed that CE-MAPPO yields evidently better than the IL-MAPPO in Fig. 3, while it costs heavier communication overheads as Table III presents. Consistent with the discussion in Section III-D, the distributed system (IL-MAPPO) significantly reduces the communication expenses at the sacrifice of connection and safety maintenance in Fig. 4(a).

With the implementation of AT, IA-MAPPO obtains comparable performance as CE-MAPPO in Fig. 3, as collisions with adversary occur less frequently in IA-MAPPO with plunging disconnections between agents and encouraged intention sharing, and similar observations also applies to the centralized AT-MAPPO shows in Fig. 4(a). Specifically, the superiority of AT can be explained as the reflection of the improved R_{at} in the process of fine-tuning the mixed-formation policy π_f in Fig. 4(b).

V. CONCLUSION

In this paper, we have proposed IA-MAPPO to solve the problem of FMA. Specifically, we have enhanced the flexibility of formation by distilled policies and utilize Imitation Learning to obtain a decentralized solution with significantly reduced communication overheads. Afterwards, Alternative Training is put forward to complement the performance loss incurred due to the decentralization. Finally, we have verified the communication efficiency and proven the effectiveness of IA-MAPPO in extensive experiments. In the future, we will enlarge the scale of swarms and enrich formation patterns.

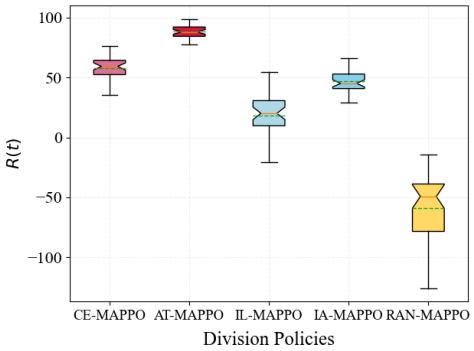
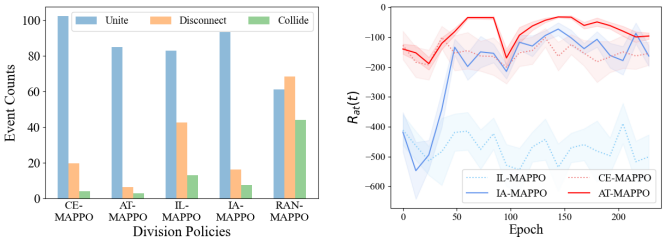


Fig. 3. Average $R(t)$ over 600 episodes in five division policies.



(a) Event counts (b) Variation of $R_{\text{at}}(t)$

Fig. 4. Average event counts over 600 episodes and variation of $R_{\text{at}}(t)$ during fine-tuning CE-MAPPO and IL-MAPPO.

More practical factors (e.g., communication delay) shall be accurately incorporated as well.

REFERENCES

[1] S. Roy, *et al.*, "Iot security and computation management on a multi-robot system for rescue operations based on a cloud framework," *Sensors*, vol. 22, no. 15, Aug. 2022.

[2] R. Konda, *et al.*, "Decentralized function approximated q-learning in multi-robot systems for predator avoidance," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6342–6349, Oct. 2020.

[3] Z. Young, *et al.*, "Consensus, cooperative learning, and flocking for multiagent predator avoidance," *International Journal of Advanced Robotic Systems*, vol. 17, no. 5, Sep. 2020.

[4] A. Nedic, *et al.*, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, May 2018.

[5] Y. Yan, *et al.*, "Relative distributed formation and obstacle avoidance with multi-agent reinforcement learning," in *ICRA*, Philadelphia, USA, May 2022.

[6] C. Yu, *et al.*, "The surprising effectiveness of ppo in cooperative multi-agent games," in *NeurIPS*, New Orleans, USA, Nov. 2022.

[7] S. Green, *et al.*, "Distillation strategies for proximal policy optimization," *arXiv:1901.08128*, Jan. 2019.

[8] Y. Xiang, *et al.*, "Decentralized adaptive formation via consensus-oriented multi-agent communication," *arXiv:2307.12287*, Jul. 2023.

[9] H.-R. Lee, *et al.*, "Improved cooperative multi-agent reinforcement learning algorithm augmented by mixing demonstrations from centralized policy," in *AAMAS*, Montreal, Canada, May 2019.

[10] X. Zhang, *et al.*, "Pretraining deep actor-critic reinforcement learning algorithms with expert demonstrations," *arXiv:1801.10459*, Jan. 2018.

[11] S. Ross, *et al.*, "A reduction of imitation learning and structured prediction to no-regret online learning," in *AISTATS*, Fort Lauderdale, USA, Apr. 2011.

[12] B. Xiao, *et al.*, "Stochastic graph neural network-based value decomposition for marl in internet of vehicles," in *VTC*, Florence, Italy, Jun. 2023.

[13] C. Pan, *et al.*, "Flexible formation control using hausdorff distance: A multi-agent reinforcement learning approach," in *EUSIPCO*, Belgrade, Serbia, Aug. 2022.

[14] L. Najjar, *et al.*, "A leader-follower communication protocol for multi-agent robotic systems," in *JEEIT*, Amman, Jordan, May 2019.

[15] R. Lowe, *et al.*, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *NeurIPS*, Long Beach, USA, Dec. 2017.