2023 IEEE 39th International Conference on Data Engineering (ICDE) | 979-8-3503-2227-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDE55515.2023.00023

Exploring both Individuality and Cooperation for Air-Ground Spatial Crowdsourcing by Multi-Agent Deep Reinforcement Learning

Yuxiao Ye
School of Comp. Sci. and Tech.
Beijing Institute of Technology
Beijing, China
1120181659@bit.edu.cn

Chi Harold Liu
School of Comp. Sci. and Tech.
Beijing Institute of Technology
Beijing, China
chiliu@bit.edu.cn

Zipeng Dai
School of Comp. Sci. and Tech.
Beijing Institute of Technology
Beijing, China
3120215520@bit.edu.cn

Jianxin Zhao
School of Comp. Sci. and Tech.
Beijing Institute of Technology
Beijing, China
jianxin.zhao@bit.edu.cn

Ye Yuan
School of Comp. Sci. and Tech.
Beijing Institute of Technology
Beijing, China
yuan-ye@bit.edu.cn

Guoren Wang
School of Comp. Sci. and Tech.
Beijing Institute of Technology
Beijing, China
wanggr@bit.edu.cn

Jian Tang
AI Labs
Midea Group
Beijing, China
tangjian22@midea.com

***Abstract*—Spatial crowdsourcing (SC) has proven as a promising paradigm to employ human workers to collect data from diverse Point-of-Interests (PoIs) in a given area. Different from using human participants, we propose a novel air-ground SC scenario to fully take advantage of benefits brought by unmanned vehicles (UVs), including unmanned aerial vehicles (UAVs) with controllable high mobility and unmanned ground vehicles (UGVs) with abundant sensing resources. The objective is to maximize the amount of collected data, geographical fairness among all PoIs, and minimize the data loss and energy consumption, integrated as one single metric called "efficiency". We explicitly explore both individuality and cooperation natures of UAVs and UGVs by proposing a multi-agent deep reinforcement learning (MADRL) framework called "h/i-MADRL". Compatible with all multi-agent actor-critic methods, h/i-MADRL adds two novel plug-in modules: (a) h-CoPO, which models the cooperation preference among heterogeneous UAVs and UGVs; and (b) i-EOI, which extracts the UV's individuality and encourages a better spatial division of work by adding intrinsic reward. Extensive experimental results on two real-world datasets on Purdue and NCSU campuses confirm that h/i-MADRL achieves a better exploration of both individuality and cooperation simultaneously, resulting in a better performance in terms of efficiency compared with five baselines.**

***Index Terms*—Air-ground spatial crowdsourcing, Multi-agent deep reinforcement learning, Intrinsic reward**

I. Introduction

Spatial crowdsourcing (SC [1], [2]) is an attractive paradigm where human participants join data collection campaign (e.g., OpenStreetMap [3] and Waze [4]). Different from that, air-ground SC by unmanned vehicles (UVs), including unmanned aerial vehicles (UAVs, e.g., drones) and unmanned ground vehicles (UGVs, e.g., driverless cars), are able to provide ubiquitous sensing services in dangerous or inaccessible task areas, e.g., earthquake and city fire. These controllable UVs are usually equipped with high-speed data receivers like WiFi/5G, able to collect sensory data from Point-of-Interests (PoIs) like CCTV cameras and alarm sensors within a larger range compared to human participants. To achieve higher data rates and quality-of-service (QoS), we consider to employ the air-ground non-orthogonal multiple access (AG-NOMA [5]) technique in the data upload process. Specifically, a UAV collects the uplink data and then relay to a UGV, where the latter not only serves as the mobile base stations (BSs) for decoding but also collects the data from PoIs as well. The overall goal is to maximize the amount of collected data from all PoIs, given the limited energy reserve of UVs when they are moving around and collecting data back and forth.

To achieve this, key challenges are as follows. First, UAVs are responsible for data collection only, however UGVs have another role as the mobile BS; and thus correlation between heterogeneous two types of UVs becomes more complicated for cooperative task completion. Second, UAVs and UGVs have different mobility patterns, i.e., UAVs have a relatively higher movement speed and able to move in any direction, while UGVs need to strictly follow the lane and some of the corner areas are inaccessible, which poses a greater challenge to the spatial division of work between UVs.

To this end, the goal of this paper is to design control algorithms to navigate a group of UVs and explore both individuality and cooperation for air-ground SC tasks. However, it is hard to formulate it as a closed-form optimization problem. Thus, We opt for heuristic solutions by considering a time-slotted system, which can be regarded as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP). Among many existing solutions to Dec-POMDP, multi-agent deep reinforcement learning (MADRL) is becoming increasingly popular for several sequential decision-making scenarios [6] [7] [8] with multiple controllable agents.

2375-026X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDE55515.2023.00023

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

Exemplar methods like QMIX [9], MAPPO [10] and IPPO [11] all assumed that agents are homogeneous with fully cooperative or fully competitive motivations, where agents hardly give up a portion of their short-term individual reward, to increase the long-term reward of the entire team. However, our considered air-ground SC tasks are targeting a mixture of different optimization goals where some UVs may have to give up a portion of their short-term individual reward, but increase the long-term reward of all UVs. For example, to fully explore the mobility benefits brought by UAVs, they need to spend more energy and collect more data if compared with fixed-lane driving UGVs, while UGVs may have to stay within a certain range of UAVs to decode their relayed data as mobile BSs. The contribution of this paper is four-fold:

- We propose a novel scalable and effective framework called h/i-MADRL for air-ground SC tasks by distributed MADRL. It consists of one base module and two plug-in modules. The former can be almost any multi-agent actor-critic algorithms, and the latter effectively exploits the individuality and cooperations among UVs.
- We propose i-EOI, the intrinsic reward based emergence of individuality, that encourages a better spatial division of work between UAVs and UGVs.
- We propose h-CoPO, a heterogeneous Coordinated Policy Optimization that incorporates a social psychology principle to learn neural controller for both UAVs and UGVs, by accurately modeling their cooperation preferences.
- Extensive experimental results and trajectory visualizations based on two real-world datasets on Purdue and NCSU campuses confirm that h/i-MADRL achieves a better exploration of both individuality and cooperation simultaneously, resulting in higher efficiency when compared with five baselines.

The rest of the paper is organized as follows. We review related works in Section II. We present the system model in Section III. Problem definition and formation are given in Section IV and the solution is presented in Section V. Experimental results are supplemented in Section VI and finally, Section VII concludes the paper.

II. Related Work

A. Classical SC and UV-aided SC

Zhao *et al.* in [12] employed more than one workers to cooperate to maximize the overall rewards, and they designed an equilibrium-based method. Zhao *et al.* in [13] proposed two game-theoretic algorithms to achieve fair task assignment by designing several payoff policies. Wang *et al.* in [14] solved data deficiency problems for a time-continuous SC with limited availability of workers, and proposed a polynomial-time task assignment algorithm with entropy-based quality improvement.

On the other hand, UV-aided SC becoming a promising research direction for large-scale urban sensing. Xu *et al.* in [15] proposed a mathematical model of UAV-aided task allocation and a genetic-based algorithm to balance task quality and cost. Zhao *et al.* in [16] proposed a DRL method to navigate multiple UAVs for data collection and improve the energy efficiency. Wang *et al.* in [17] considered the disaster response applications where PoIs are unevenly distributed in a 3-dimensional space. Ding *et al.* in [18] proposed a UAV-aided MCS paradigm for large-scale and high-quality urban sensing by using MADRL. However, none of the existing works jointly considered to deploy heterogeneous UAVs and UGVs simultaneously for cooperative air-ground SC, nor considering both individuality and cooperation among them. To the best of our knowledge, this paper is one of the first along this direction.

B. MADRL

To address the non-stationary issue in multi-agent system, different MADRL solutions are proposed, among which the centralized training and decentralized execution (CTDE) framework has achieved remarkable success, like value decomposition methods [9], [19] that approximate the joint value function by optimizing it for each agent. RMIX [20] investigated risk-sensitive MARL, which obtained more sufficient estimations of future returns. QPLEX [21] achieved a better representation without violating the IGM consistency by using a duplex dueling structure. As a typical continuous control framework of CTDE, MADDPG [6] feeds actions of other agents into the centralized critic for the sake of stabilizing the training. FACMAC [22] combined the merit of QMIX and MADDPG, where expressive non-monotonic factorization and a centralized gradient estimator are used to allow more coordinated policy changes. Policy-based MADRL approaches such as IPPO [11] and MAPPO [10] also show promising results for a wide range of multi-agent tasks [23], [24]. However, how to explore individuality and cooperation simultaneously is not explicitly discussed in those methods, which is non-trivial in air-ground SC tasks. Intrinsic reward (i.e., reward outside the environment) is very useful in MADRL to encourage agents to explore individuality and cooperate with each other. Jaques *et al.* [25] took the action correlation between agents (as how much the change of one agent's action will affect the change of another agent's action) as an intrinsic reward. Du *et al.* [26] expected agents to learn different kinds of policies by introducing intrinsic rewards.

III. System Model

A set $\mathcal{K} \triangleq \{k|1,2,...,K\}$ of UVs (which consists of a set $\mathcal{U} \triangleq \{u|1,2,...,U\}$ of UAVs and $\mathcal{G} \triangleq \{g|1,2,...,G\}$ of UGVs) are jointly navigated to collect data from a set of PoIs $\mathcal{I} \triangleq \{i|1,2,...,I\}$. Initially, each PoI $i \in \mathcal{I}$ contains D_0^i unit data. Without loss of generality, we consider a fixed task duration which can be divided into T timeslots of equal length τ, and we assume each timeslot duration can be divided into two parts, UV movement time cost τ_{move} and data collection time cost τ_{coll}.

A. UAV and UGV Movement

We deploy rotary-wing drones as UAVs in an air-ground SC task. Each UAV u flies or hovers at a constant altitude H^u.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

TABLE I
LIST OF IMPORTANT NOTATIONS USED IN THIS PAPER.

Notation	Explanation
$u, U, \mathcal{U}$	Index, total #, set of UAVs
$g, G, \mathcal{G}$	Index, total #, set of UGVs
$i, I, \mathcal{I}$	Index, total #, set of PoIs
$t, T, \tau_{\text{move}}, \tau_{\text{coll}}$	Index, total # of timeslots, duration for UV movement and data collection in each timeslot
$\vartheta_t^u, v_t^u, v_t^g$	Moving direction of UAV u, movement speed of UAV u and UGV g in timeslot $[t, t+1)$
η_t^u, η_t^g	Energy consumption of UAV u and UGV g during timeslot [t, t+1)
$\gamma_{\text{SINR}}^{i,u}, \gamma_{\text{SINR}}^{i,g}, \gamma_{\text{SINR}}^{u,g}$	SINR of PoI-UAV uplink channel, PoI-UGV uplink channel, UAV-UGV relay channel
$C^{i,u}, C^{i,g}, C^{u,g}$	Capacity of PoI-UAV uplink channel, PoI-UGV uplink channel, UAV-UGV relay channel
$\boldsymbol{s}_t, \boldsymbol{o}_t, \boldsymbol{a}_t, r_t$	State, observation, action and reward over all UVs in timeslot $[t, t+1)$
$\psi, \sigma, \xi, \kappa, \lambda$	Data collection ratio, data loss ratio, energy consumption ratio, geographical fairness, efficiency

In each timeslot $[t, t+1)$, each UAV u spends τ_{move} amount of time moving to a certain direction $\vartheta_t^u \in [0, 2\pi)$, at speed $v_t^u \in [0, v_{\max}^{\text{UAV}}]$.

Different from UAVs, UGV movement is restricted by the roadmap. Thus, given the speed $v_t^g \in [0, v_{\max}^{\text{UGV}}]$, we assume that each UGV g can move to a destination only if the shortest path length between the current position and the destination does not exceed the maximum moving range (i.e., $\tau_{\text{move}} \cdot v_{\max}^{\text{UGV}}$) in each timeslot $[t, t+1)$. Due to the complexity of city roadmap, actual moving range of a UGV is usually much smaller than a UAV in each timeslot.

The energy consumption models of UAV u and UGV g during each timeslot $[t, t+1)$ are set as proportional to their movement speed as:

$$\eta_t^u \propto \tau_{\text{move}} \cdot v_t^u, \quad \eta_t^g \propto \tau_{\text{move}} \cdot v_t^g. \tag{1}$$

B. Data Collection by AG-NOMA Uplink Channel

We consider an AG-NOMA based uplink communications, where UAVs and UGVs stop and receive data uploaded from PoIs. The total available spectrum for these I PoIs are equally divided into Z subchannels, where the unit bandwidth is B and the power spectral density of the noise is N_0. Since UAVs are usually equipped with limited computational resources, we assume that only UGVs can decode the uplink data by themselves, as mobile BSs. That is, UAVs should relay data from PoIs to UGVs. Thus, the channel models for data collection can be divided into three types: (a) uplink channel from PoI i to UAV u, (b) uplink channel from PoI i' to UGV g, and (c) relay channel from UAV u to UGV g. Following [5], we consider air-ground co-channel interference suppression method that pairs the direct links and relay links on the same subchannels. Thus, there exists a set $\mathcal{E} \triangleq \{(u, g, i, i')_{z,t} | u \in \mathcal{U}, g \in \mathcal{G}, i, i' \in \mathcal{I}, i \neq i'\}$ of tuples, each of which represents a data collection event in each subchannel z and timeslot $[t, t+1)$. Then, for simplicity, we temporally omit the timeslot index t and subchannel index z when defining the following channel models.

PoI-UAV uplink channel: considering that a ground-to-air (G2A) link from PoI i to UAV u is either LoS or NLoS, we calculate the corresponding factors by:

$$\omega_{\text{LoS}}^{i,u} = \frac{1}{1 + \omega \exp(-\beta[\text{ang}(i,u)])}, \omega_{\text{NLoS}}^{i,u} = 1 - \omega_{\text{LoS}}^{i,u}, \tag{2}$$

where β and ω are constants related to the network coverage; $\text{ang}(i,u)$ is elevation angle of PoI i respect to UAV u (measured in degree), expressed as $\text{ang}(i,u) = \arcsin(H^u / d[i,u])$; $d[i,u]$ denotes the direct distance between PoI i and UAV u. Then, the channel gain can be computed as:

$$\varsigma^{i,u} = \omega_{\text{LoS}}^{i,u} \eta_{\text{LoS}} \cdot d[i,u]^{-\alpha_1} + \omega_{\text{NLoS}}^{i,u} \eta_{\text{NLoS}} \cdot d[i,u]^{-\alpha_1}, \tag{3}$$

where α_1 is the path loss factor of G2A channels; η_{LoS} and η_{NLoS} are additional attenuation fading factors of LoS and NLoS channels, respectively. We assume that the full-duplex introduced self-interference is fully canceled, then considering both PoI i and i' are in the same subchannel z, the Shannon capacity $C^{i,u}$ in each subchannel can be computed as:

$$C^{i,u} = B \log(1 + \gamma_{\text{SINR}}^{i,u}), \quad \gamma_{\text{SINR}}^{i,u} = \frac{\varsigma^{i,u} \rho^i}{N_0 B + \varsigma^{i',u} \rho^{i'}}, \tag{4}$$

where $\gamma_{\text{SINR}}^{i,u}$ is the uplink signal-to-interference-plus-noise ratio (SINR) from PoI i to UAV u; ρ^i is the transmission power of PoI i; and $\varsigma^{i',u} \rho^{i'}$ is the interference power introduced to UAV u by PoI i'.

PoI-UGV uplink channel: as a ground-to-ground (G2G) channel, based on path loss and Rayleigh fading, the channel gain can be computed as:

$$\varsigma^{i',g} = |h_z|^2 \cdot d[i',g]^{-\alpha_2}, \tag{5}$$

where α_2 is the path loss factor of the G2G channels; h_z is the amplitude gain of the signals on subchannel z. Then, since UGV g has decoded relayed data from UAV u, the capacity of direct G2G uplinks from the PoI i' to UGV g is:

$$C^{i',g} = B \log(1 + \gamma_{\text{SINR}}^{i',g}), \quad \gamma_{\text{SINR}}^{i',g} = \frac{\varsigma^{i',g} \rho^{i'}}{N_0 B}, \tag{6}$$

where $\gamma_{\text{SINR}}^{i',g}$ and $\rho^{i'}$ are the SINR and transmission power for PoI i', respectively.

UAV-UGV relay channel: since data transmission from the hovering UAV u to UGV g is based on air-to-ground (A2G) links, we follow [27] and assume that UAVs relay data with decode-and-forward scheme and perfect full-duplex technology. Thus, similar to Eqn. (2) and Eqn. (3) in G2A channels, we calculate the LoS factor, NLoS factor and channel gain by:

$$\omega_{\text{LoS}}^{u,g} = \frac{1}{1 + \omega \exp(-\beta[\text{ang}(u,g)])}, \omega_{\text{NLoS}}^{u,g} = 1 - \omega_{\text{LoS}}^{u,g}, \tag{7}$$

$$\varsigma^{u,g} = \omega_{\text{LoS}}^{u,g} \eta_{\text{LoS}} \cdot d[u,g]^{-\alpha_1} + \omega_{\text{NLoS}}^{u,g} \eta_{\text{NLoS}} \cdot d[u,g]^{-\alpha_1}, \tag{8}$$

where $\text{ang}(u,g)$ and $d[u,g]$ denote the elevation angle and direct distance between UAV u and UGV g, respectively. When decoding the relayed data from PoI i, we consider that the useful data received by UGV g include not only relayed

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

from UAV u but also from PoI i since UGV g will receive a copy wirelessly as well. As a result, the Shannon capacity is:

$$C^{u,g} = B\log(1+\gamma^{u,g}_{\text{SINR}}), \quad \gamma^{u,g}_{\text{SINR}} = \frac{\varsigma^{u,g}\rho^u + \varsigma^{i,g}\rho^i}{N_0 B + \varsigma^{i',g}\rho^{i'}}, \quad (9)$$

where $\gamma^{u,g}_{\text{SINR}}$ is the received SINR from UAV u to UGV g on subchannel z for relaying PoI i's data, with interference from PoI i'. ρ^u denotes the transmission power of UAV u.

Note that in this paper we assume that our system model is based on NOMA, a promising communication technique to utilize all the time and frequency resources through power domain superposition coding and successive interference cancelation. However, our proposed solution is also applicable to other communication models, such as TDMA and OFDMA, by simply re-defining the data collection and relay models in Section III-B.

IV. Problem Definition and Formulation

A. Problem Definition

We assume that each UAV/UGV moves and chooses the nearest PoIs to access. To enforce QoS, data can be uploaded successfully only if the SINR exceeds the given threshold in a subchannel, otherwise data loss occurs.

Definition 1. ***UAV Data Collection:*** *In timeslot $[t, t+1)$, a UAV u can successfully collect $D^{i,u}_{z,t}$ amount of data from PoI i in subchannel z by:*

$$\Delta D^{i,u}_{z,t} = \begin{cases} 0, & \text{if } \min(\gamma^{i,u}_{\text{SINR}}, \gamma^{u,g}_{\text{SINR}}) < \text{threshold} \\ \tau_{\text{coll}} \cdot \min(C^{i,u}, C^{u,g}), & \text{otherwise.} \end{cases} \quad (10)$$

Definition 2. ***UGV Data Collection:*** *In timeslot $[t, t+1)$, a UGV g can successfully collect $D^{i',g}_{z,t}$ amount of data from PoI i' in subchannel z by:*

$$\Delta D^{i',g}_{z,t} = \begin{cases} 0, & \text{if } \gamma^{i',g}_{\text{SINR}} < \text{threshold} \\ \tau_{\text{coll}} \cdot C^{i',g}, & \text{otherwise.} \end{cases} \quad (11)$$

Then, when timeslot $[t, t+1)$ ends, the remaining amount of data in PoI i is $D^i_{t+1} = \max(D^i_t - \sum_{k\in\mathcal{K}^i_t}\sum_z \Delta D^{i,k}_{z,t}, 0)$, where $\mathcal{K}^i_t$ includes those UAVs/UGVs who access PoI i in that timeslot.

We use five metrics to evaluate the performance of air-ground SC tasks. First is data collection ratio ψ, defined as:

$$\psi = 1 - \frac{\sum_{i\in\mathcal{I}} D^i_T}{\sum_{i\in\mathcal{I}} D^i_0}, \quad (12)$$

where D^i_T is the remaining data from PoI i. Second is data loss ratio σ, due to the impact of low SINR as:

$$\sigma = \frac{\left|\{\mathcal{E} \subset \mathcal{K}^i_t,\ \Delta D^{i,u}_{z,t} = 0 \vee \Delta D^{i',g}_{z,t} = 0\}\right|}{Z \cdot T \cdot (U+G)}. \quad (13)$$

Third is energy consumption ratio ξ, defined as:

$$\xi = \frac{1}{U}\sum_{t=1}^{T}\sum_{u\in\mathcal{U}} \frac{\eta^u_t}{E^u_0} + \frac{1}{G}\sum_{t=1}^{T}\sum_{g\in\mathcal{G}} \frac{\eta^g_t}{E^g_0}, \quad (14)$$

where E^u_0 and E^g_0 are the initial energy reserve for UAV u and UGV g, respectively. Fourth is geographical fairness κ, following Jain's fairness index [28] as:

$$\kappa = \frac{(\sum_{i\in\mathcal{I}}(D^i_0 - D^i_T)/D^i_0)^2}{I\sum_{i\in\mathcal{I}}((D^i_0 - D^i_T)/D^i_0)^2}. \quad (15)$$

The overall objective is to maximize data collection ratio ψ and geographical fairness κ, while minimizing data loss ratio σ and energy consumption ratio ξ simultaneously. To this end, we use a comprehensive, integrated metric λ to measure the performance of the task, called "efficiency", as

$$\lambda = \frac{\psi \cdot (1-\sigma) \cdot \kappa}{\xi}. \quad (16)$$

B. Problem Formulation

We formulate an air-ground SC task as a Dec-POMDP [29], represented by a tuple $(\mathcal{K}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \Pr, \gamma)$, where $\mathcal{K}$, $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ are the set of UVs, states, local observations and actions, and γ is the discount factor. The system works as follows. At the beginning of a task, the global state is initialized as $\boldsymbol{s}_0$. In each timeslot $[t, t+1)$, each UV $k \in \mathcal{K}$ has its own observations $\boldsymbol{o}^k_t$ of state $\boldsymbol{s}_t$, and then decides an action $\boldsymbol{a}^k_t$ sampled from its policy $\boldsymbol{a}^k_t \sim \pi^k(\cdot|\boldsymbol{o}^k_t)$. After all UVs have made decisions, the environment receives the joint action $\mathbf{a}_t = \{\boldsymbol{a}^k_t\}_{k=1}^K$ and then calculate the environmental (i.e, extrinsic) reward $r^k_{t,\text{ext}} = \mathcal{R}^k(\boldsymbol{s}_t, \mathbf{a}_t)$ for each k; followed by transiting to the next state $\boldsymbol{s}_{t+1}$, based on the state transition distribution $\Pr(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \mathbf{a}_t)$.

1) State and observation space: The global state $\boldsymbol{s}_t$ is a vector, concatenating two types of information: each UV k's current 2-D position along with its remaining energy (x^k_t, y^k_t, E^k_t), and each PoI i's position with its remaining amount of data (x^i, y^i, D^i_t). Then, each k's observation $\boldsymbol{o}^k_t$ has the identical size as the global state $\boldsymbol{s}_t$. However, when certain UVs or PoIs are out of the observable range of a UV k in timeslot $[t, t+1)$, their corresponding information will become $(0,0,0)$ in $\boldsymbol{o}^k_t$ as blind.

2) Action space: Since UAVs and UGVs are two types of heterogeneous UVs, let action spaces be $\mathcal{A}^u$ and $\mathcal{A}^g$, respectively. For UAVs, $\mathcal{A}^u$ is continuous and represents the control of moving direction and speed, i.e., $\boldsymbol{a}^u_t = (\vartheta^u_t, v^u_t) \in \mathbb{R}^2$. For UGVs, considering the restrictions of a roadmap, we set $\mathcal{A}^g \subset \mathcal{A}^u$ because any movement outside the road or out of the maximum moving range is forbidden.

3) Reward function: It measures the successfully collected amount of data by a UAV/UGV in timeslot $[t, t+1)$, as:

$$r^k_{t,\text{ext}} = \sum_{i\in\mathcal{I}}\sum_{z\in\mathcal{Z}}\left(\frac{\Delta D^{i,k}_{z,t}}{I \cdot D^i_0} - \omega_{\text{coll}} \cdot \iota^{i,k}_{z,t}\right) - \omega_{\text{move}} \cdot \frac{\eta^k_t}{E^k_0}, \quad (17)$$

where $\iota^{i,k}_{z,t} = 1$ if data loss occurs, i.e., $\Delta D^{i,k}_{z,t} = 0$; otherwise $\iota^{i,k}_{z,t} = 0$. ω_{move} and ω_{coll} are penalties of energy consumption and data loss, respectively. Thus, for each k, the goal of

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

the optimization problem is to find an optimal policy π^k to maximize the accumulated discounted reward:

$$\begin{aligned} \max_{\pi^k} \quad & \mathbb{E}\big[\textstyle\sum_{t=1}^{T} \gamma^t \, r^k_{t,\text{ext}}(\boldsymbol{o}^k_t, \boldsymbol{a}^k_t)\big] \\ \text{s.t.} \quad & \boldsymbol{o}^k_t \in \mathcal{O}, \pi^k(\cdot|\boldsymbol{o}^k_t) \in \mathcal{A}. \end{aligned} \tag{18}$$

V. Proposed Solution: h/i-MADRL

To solve Dec-POMDP in air-ground SC tasks, we propose a MADRL-based solution called "h/i-MADRL". As shown in Fig. 1, it consists of a base module compatible to any multi-agent actor-critic architectures, e.g., MADDPG [6], IPPO [11] and MAPPO [10], as well as two novel plug-ins: (a) intrinsic reward driven exploitation of individuality (i-EOI) as an enhancement to the state-of-the-art approach EOI [30], and (b) heterogeneous coordinated policy optimization (h-CoPO) as an improvement to the state-of-the-art approach CoPO [31]. The former encourages a more obvious division of different UVs by extracting their individuality, and the latter constructs correlations between different UAVs and UGVs to achieve better cooperative patterns. For simplicity, we omit the timeslot index t in this section.

A. Exploring Individuality by Self-supervised Identity Classifier: i-EOI

Fully exploring individually is important to finish an air-ground SC task. This is because that PoIs are unevenly distributed in a task area but each UV can only observe its own partial observation, thus they need to move constantly to achieve spatial division of work, while avoid going to the same area at the same time which results in waste of resources. Furthermore, since UAVs and UGVs are heterogeneous with different capabilities (e.g, movement speed and whether the movement is restricted by the roadmap, data transmission capacity due to interferences), introducing individuality to self-interested UVs is promising to make them aware of their own duty in the task (e.g., UAVs move to remote areas while UGVs stay around UAVs to decode their data) and eventually improve efficiency.

To this end, we propose to extract UV's individuality and import an auxiliary task called "intrinsic reward driven exploitation of individuality (i-EOI)". The goal is to accurately identify a UV k from its given distinct observation $\boldsymbol{o}^k$, by training a global probabilistic classifier p_μ parameterized by μ, whose input is $\boldsymbol{o}^k$ and output $p_\mu(\cdot|\boldsymbol{o}^k)$ is the probability of the observation belonging to each UV. By defining an intrinsic reward $p_\mu(k|\boldsymbol{o}^k)$ that denotes the possibility of a certain UV is accurately predicted/identified from its observation, the compound reward for a UV k is expressed as:

$$r^k = r^k_{\text{ext}} + \omega_{\text{in}} \cdot p_\mu(k|\boldsymbol{o}^k), \tag{19}$$

where $\sum_k p_\mu(k|\boldsymbol{o}^k) = 1$, r^k_{ext} is the extrinsic reward from the environment, and ω_{in} is a tuning hyperparameter to weight the importance between intrinsic and extrinsic rewards. The physical meaning of intrinsic reward is that, since UVs can only partially observe the environment, reaching far away areas can help a certain UV to distinguish its observation from others and identify itself, and thus receiving higher intrinsic reward motivates a clearer division of work.

Since UVs maximize the expected future reward, the difference in terms of UV policies are reinforced, which makes them more identifiable and self-interested, in turn, boosts the training of classifier p_μ. Therefore, the learning process is closed-loop with positive feedback. We also use the regularizer which maximizes the mutual information between a UV's identity and observation, computed as:

$$\begin{aligned} &\text{MI}(K;O) = \mathcal{H}(K) - \mathcal{H}(K|O) \\ &= \mathcal{H}(K) - \mathbb{E}_{\boldsymbol{o}\sim p_\mu(\boldsymbol{o})}[\sum_k -p_\mu(k|\boldsymbol{o}^k)\log p_\mu(k|\boldsymbol{o}^k)]. \end{aligned} \tag{20}$$

During training, the number of samples for each UV is equal, thus $\mathcal{H}(K)$ is a constant when we uniformly sample $< \boldsymbol{o}^k, k >$ from the data buffer in a given actor-critic method (e.g., experience replay used in MADDPG) to train the classifier. As a result, maximizing $\text{MI}(K;O)$ is equivalent to minimize $\mathcal{H}(K|O) = \text{Cross\_Entropy}(p_\mu(\cdot|\boldsymbol{o}^k), p_\mu(\cdot|\boldsymbol{o}^k))$. The significance is that if its observation is more identifiable, it is easier to infer a UV that visits the given observation most, therefore maximizing this mutual information can accelerate the development of UV individuality. We can train the classifier $p_\mu(k|\boldsymbol{o}^k)$ by using a self-supervised loss as:

$$\begin{aligned} \mathcal{L}_{\text{EOI}} = &\,\text{Cross\_Entropy}(p_\mu(\cdot|\boldsymbol{o}^k), \text{one\_hot}(k)) \\ &+ \epsilon \cdot \text{Cross\_Entropy}(p_\mu(\cdot|\boldsymbol{o}^k), p_\mu(\cdot|\boldsymbol{o}^k)), \end{aligned} \tag{21}$$

where ϵ is hyperparameter, $\text{one\_hot}(k)$ refers to the one hot vector with length K where the k-th dimension is 1.

Finally, we discuss the significance of adding i-EOI for air-ground SC. In the air-ground SC task, obtaining a new local observation usually refers to discovering and accessing a group of PoIs that have not been seen before, which likely lead to a new trajectory pattern with higher cumulative extrinsic reward. That is, intrinsic reward can be regarded as a signal to strengthen, rather than weaken the stimulation of extrinsic reward, incurring UVs forgetting to better complete the task itself to obtain individuality.

B. Encouraging UV Cooperations by h-CoPO

Since UAVs and UGVs are heterogeneous UVs with different movement patterns and communication capacities, it is hard to model them by fully cooperative MARL, e.g., VDN [19] and QMIX [9], which share their policy networks and rewards. On the other hand, to achieve our complicated goal, UVs must learn to cooperate under the AG-NOMA communications model. That is, UAVs are able to explore wider areas than UGVs but their received data should have to be relayed to UGVs for decoding. Thus, a fully independent learning scheme like IPPO [11] may trap into the local optima. To this end, we propose h-CoPO, which constructs interactive correlations between UVs to form the appropriate level of coordination. It is reasonable to assume that each UV may not interact with all others simultaneously, we define two kinds of "neighbors" whom an UV should interact with.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

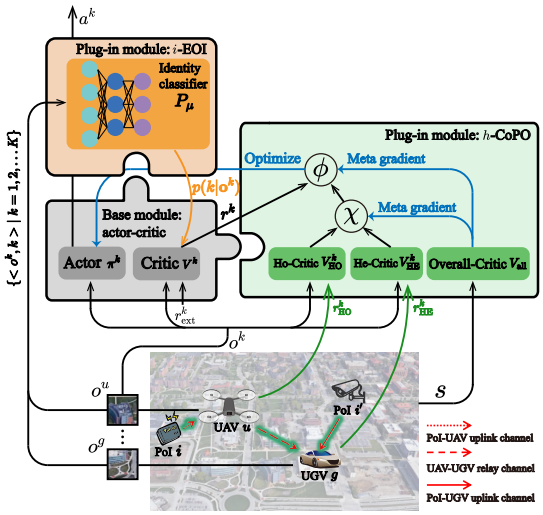


Fig. 1. Overall proposed solution h/i-MADRL.

Heterogeneous relaying neighbors $\mathcal{N}_{\rm HE}$: we define UAV u and UGV g as heterogeneous neighbors in a relayed pair if they are in the same subchannel, and their collected amount of data are affected by each others due to interference.

Homogeneous nearby neighbors $\mathcal{N}_{\rm HO}$: since each UV may be affected by its physically nearby UVs due to task competition [32], we simply consider them as neighbors.

Appropriate level of coordination among all UVs might differ from task to task. For example, in a larger task area, the appropriate distance between UAV and UGV in a relayed pair might be more crucial to prevent data loss (with satisfactory received SINR). Hence, it is challenging to generate their cooperation preferences through a top-down design or a priori knowledge. To this end, we propose a bi-level coordination method. First, each UV chooses to be self-interested or cooperative to neighbors, and then selects the proportion of attention focusing on both heterogeneous and homogeneous neighbors. To depict such characteristics, inspired by the ring measure of social value orientation [33], [34], in this paper we propose a spherical measure to model the bi-level coordination.

Specifically, we import two local coordination factors (LCFs) $\phi, \chi \in [0°, 90°]$ to the current reward function of each UV. We assume that whether a UV is self-interested to others or not is determined by ϕ, while how much attention it pays attention to each kind of neighbors is determined by χ. Thus, weighted by LCFs, we define a cooperation-aware reward as:

$$r^k_{\rm CO} = r^k \cos\phi + \Big(r^k_{\rm HE}\cos\chi + r^k_{\rm HO}\sin\chi\Big)\sin\phi, \tag{22}$$

where $r^k_{\rm HE}$ and $r^k_{\rm HO}$ are average reward of heterogeneous and homogeneous neighbors, calculated by:

$$r^k_{\rm HE} = \frac{\sum_{n\in\mathcal{N}^k_{\rm HE}} r^n}{|\mathcal{N}^k_{\rm HE}|}, \quad r^k_{\rm HO} = \frac{\sum_{n\in\mathcal{N}^k_{\rm HO}} r^n}{|\mathcal{N}^k_{\rm HO}|}. \tag{23}$$

Although we hypothesize that the cooperation preference of UAVs and UGVs exist, the selection of LCFs still remains a major challenge. We conduct a automatic search of optimal LCFs aided by meta learning approach [31] and details will be discussed in Section V-C.

Apart from the existing actor and critic networks in a given base module, our proposed h-CoPO introduces another three kinds of critic networks, namely: heterogeneous neighborhood value network $V^k_{\rm HE}$ and homogeneous neighborhood value network $V^k_{\rm HO}$ for each UV k, which input local observation $\boldsymbol{o}^k$, and one overall value network $V_{\rm all}$ shared by all UVs, which inputs state $\boldsymbol{s}$.

C. Details of Training Process: IPPO as Base Module

To better illustrate our method, we use IPPO [11] as the exemplar base module. In our scenario, the individual value function and advantage function of UV k in timeslot $[t, t+1)$ can be computed as:

$$\begin{aligned} V^k_t &= V^k(\boldsymbol{o}^k_t) = \mathbb{E}\Big[\sum_{t'=t}^{T}\gamma^{t'-t} r^k_{t'}\Big], \\ A^k_t &= A^k(\boldsymbol{o}^k_t, \boldsymbol{a}^k_t) = r^k_t + \gamma V^k_{t+1} - V^k_t. \end{aligned} \tag{24}$$

Note that r^k_t combines intrinsic and extrinsic rewards by Eqn. (19). Then, by adopting a truncated importance sampling factor $\varrho = \frac{\pi^k_{\rm new}(\boldsymbol{a}^k|\boldsymbol{o}^k)}{\pi^k_{\rm old}(\boldsymbol{a}^k|\boldsymbol{o}^k)}$, the surrogate objective of IPPO is written as:

$$J_{\rm IPPO}(\pi^k) = \mathbb{E}_{(\boldsymbol{s},\mathbf{a})}\min(\varrho A^k, \text{clip}(\varrho, 1-\epsilon, 1+\epsilon)A^k), \tag{25}$$

where clip function helps remove the incentive for moving r^k outside of the interval $[1-\epsilon, 1+\epsilon]$. The value function V^k is updated by mean squared error loss function:

$$\mathcal{L}^k = \mathbb{E}[r^k_t + \gamma V^k(\boldsymbol{o}^k_{t+1}) - V^k(\boldsymbol{o}^k_t)]^2. \tag{26}$$

Algorithm 1 shows the pseudo-code of our solution. First, we initialize i-EOI classifier, overall value network $V_{\rm all}$, data buffer shared by all UVs (Line 1) and learnable parameters $\theta^k = \{\pi^k, V^k, V^k_{\rm HE}, V^k_{\rm HO}\}$ for each k (Line 2). ϕ and χ are initialized to $0°$ and $45°$ for each k (Line 3), i.e., all UVs are completely self-interested and do not hold a priori preference between each type of neighbors. Then, loop for sampling and training is started (Line 4). After interacting with the environment (Line 5-10) and storing the sampled data ($\boldsymbol{o}^k_t$, $\boldsymbol{a}^k_t$, $r^k_{t,\rm ext}$) and $\boldsymbol{s}_t$ into the data buffer (Line 11), training started. We consecutively update the i-EOI classifier (Line 12), update the policy (Line 14-20) for M_1 times and update the LCFs for M_2 times (Line 21-23) as sample reuse in PPO [35], in each time the data buffer is traversed to conduct mini-batch training.

For policy updating, reward r^k is computed with the intrinsic reward (Line 16). Then considering the impact brought by neighbors, we compute the corresponding value functions $V^k_{\rm HE}, V^k_{\rm HO}$ and advantage functions $A^k_{\rm HE}, A^k_{\rm HO}$ by Eqn. (24), and then extend cooperation-aware reward into cooperation-aware advantages (Line 17), by:

$$A^k_{\rm CO}(\phi,\chi) = A^k\cos\phi + \Big(A^k_{\rm HE}\cos\chi + A^k_{\rm HO}\sin\chi\Big)\sin\phi. \tag{27}$$

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: h/i-MADRL

1 Initialize environment, i-EOI classifier p_μ, overall value network $V_{\rm all}$, data buffer.
2 Initialize policy network π^k, value network V^k, heterogeneous neighborhood value network $V^k_{\rm HE}$, homogeneous neighborhood value network $V^k_{\rm HO}$ for each UV k.
3 Initialize $\phi^k = 0^\circ$ and $\chi^k = 45^\circ$ for each k.
4 **for** *iteration = 1, ..., N* **do**
5 Clean data buffer
6 **for** *t=0, 1, ... T-1* **do**
7 **for** *k=1, 2, ... K* **do**
8 Get local observation $\boldsymbol{o}^k_t$
9 UVs select action $\boldsymbol{a}^k_t \sim \pi_k$
10 Interact the environment with $\mathbf{a}_t = \{\boldsymbol{a}^k_t | k = 1, 2, ..., K\}$
11 Store the tuple $(\boldsymbol{o}^k_t, \boldsymbol{a}^k_t, r^k_{t,\rm ext})$ for each UV and state $\boldsymbol{s}_t$ in data buffer
12 Update p_μ by Eqn. (21)
13 Store current policy $\theta^k_{\rm old} \leftarrow \theta^k$ for each UV
14 **for** *m = 1, ..., M_1* **do**
15 **for** *k=1, 2, ... K* **do**
16 Compute reward r^k by Eqn. (19)
17 Compute cooperation-aware advantages $A^k_{\rm CO}(\phi, \chi)$ by Eqn. (27)
18 Update π^k by Eqn. (28)
19 Update V^k, $V^k_{\rm HE}$, $V^k_{\rm HO}$ by Eqn. (26)
20 Compute $r_{\rm all}$ and update $V_{\rm all}$ by Eqn. (26)
21 **for** *m = 1, ..., M_2* **do**
22 **for** *k=1, 2, ... K* **do**
23 Update ϕ^k, χ^k by Eqn. (30).

The surrogate objective function of h-CoPO can replace $J_{\rm IPPO}$ to update policy (Line 18) by:

$$J_{\rm CO}(\theta^k, \phi, \chi) = \mathbb{E}_{(\boldsymbol{s},\mathbf{a})} \min\Big(\varrho A^k_{\rm CO}, {\rm clip}(\varrho, 1-\epsilon, 1+\epsilon) A^k_{\rm CO}\Big). \tag{28}$$

Three value networks for each UV and overall value network $V_{\rm all}$ are updated in the inner and outer loop respectively (Line 19-20).

For LCFs updating, let $r_{\rm all}$ be sum of all UV's reward in all timeslots, and we define the overall objective as:

$$J_{\rm all} = \mathbb{E}_{(\boldsymbol{s},\mathbf{a})}[r_{\rm all}] = \mathbb{E}_{(\boldsymbol{s},\mathbf{a})}\Big[\sum_t \sum_k r^k_t\Big]. \tag{29}$$

Then, we can maximize the overall performance of an air-ground SC task by finding optimal LCFs in terms of maximizing $r_{\rm all}$, i.e., computing the gradient of overall objective w.r.t LCFs as in Line 23:

$$\nabla_{\phi,\chi} J_{\rm all}(\theta^k_{\rm new}) = \nabla_{\theta^k_{\rm new}} J_{\rm all}(\theta^k_{\rm new}) \nabla_{\phi,\chi} \theta^k_{\rm new}, \tag{30}$$

where $\theta^k_{\rm new}$ denotes the policy parameters of UV k after optimizing Eqn. (28). Here the first term of Eqn. (30) is analogous to the gradient of IPPO in Eqn. (25) but the objective is replaced with $J_{\rm all}$ as:

$$\mathbb{E}_{(\boldsymbol{s},\mathbf{a})\sim\theta^k_{\rm old}}[\nabla\theta^k_{\rm new} \min(\varrho A_{\rm all}(\boldsymbol{s},\mathbf{a}), {\rm clip}(\varrho, 1-\epsilon, 1+\epsilon) A_{\rm all}(\boldsymbol{s},\mathbf{a})], \tag{31}$$

where the overall advantage $A_{\rm all}$ can be computed by Eqn. (24), by leveraging overall value network $V_{\rm all}$. $\theta^k_{\rm old}$ denotes the policy parameters of UV k before optimizing Eqn. (28). Note that the samples $(\boldsymbol{s}, \mathbf{a})$ are generated by the behavior policy $\theta^k_{\rm old}$. The second term of Eqn. (30) can be computed by first-order Taylor expansion:

$$\begin{aligned}&\nabla_{\phi,\chi}(\theta^k_{\rm old} + \alpha \nabla_{\theta^k_{\rm old}} J_{\rm CO}(\theta^k_{\rm old}, \phi, \chi)) \\ &= \alpha \mathbb{E}_{(\boldsymbol{s},\mathbf{a})\sim\theta^k_{\rm old}}[\nabla_{\theta^k_{\rm old}} \log \pi_{\theta^k_{\rm old}}(\boldsymbol{a}^k|\boldsymbol{o}^k) \nabla_{\phi,\chi} A^k_{\rm CO}(\phi,\chi)],\end{aligned} \tag{32}$$

where $\nabla_{\theta^k_{\rm old}} J_{\rm CO}(\theta^k_{\rm old}, \phi, \chi)$ have the same form with vanilla policy gradient [36], and α is the learning rate of gradient ascent in meta-learning.

Note that the base module of h/i-MADRL can also be MAPPO, to simply replace the input of critic networks V^k (see Eqn. (26)) from the local observation $\boldsymbol{o}^k$ by the global state s.

TABLE II
SIMULATION SETTINGS.

Notation	Value	Notation	Value	Notation	Value
T	100	E^g_0	2000 KJ	α_2	4
$\tau_{\rm move}$	10	$v^{\rm UAV}_{\max}$	18 m/s	$\eta_{\rm LoS}$	0 dB
$\tau_{\rm coll}$	10	$v^{\rm UGV}_{\max}$	10 m/s	$\eta_{\rm NLoS}$	-20 dB
I	100	H_u	60 m	ω	9.6
D^i_0	3 Gbit	Z	3	β	0.16
U	2	B	20 MHz	ρ^u	3 watts
G	2	N_0	5×10^{-20} watt/Hz	ρ^i, ρ^j	0.1 watts
E^u_0	1500 KJ	α_1	2	SINR threshold	0 dB

VI. Performance Evaluation

We conduct the experiments on two real-world student movement trajectories in Purdue [37] (with 59 traces) and NCSU [38] (with 33 traces) campuses, where each trace corresponds to a student. PoIs are considered as places which are frequently visited and we take $I = 100$ most frequently visited PoIs into account for both two campuses. We use Google Map to mark the campus map data, including the specific roadmap in each campus which is crucial for UGV movement.

Default simulation settings are summarized in Table II. $v^{\rm UAV}_{\max} = 18$ m/s is set by referring to the technical report of industrial UAVs like DJI Matrice 600 [39]. By considering the height of the tallest buildings in Purdue and NCSU campus (i.e., 48.768 meters and 55.778 meters respectively), all UAVs' hovering height are set to $H_u = 60$ meters, which is safe for UAVs to fly around without crashing into buildings.

A. Baselines and Evaluation Metrics

We compare h/i-MADRL with five baselines:

- h/i-MADRL(CoPO): It replaces our proposed module h-CoPO with CoPO [31], in which two kinds of neighbors are considered equivalently.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

TABLE III
HYPERPARAMETER TUNING

			w/o SP, w/o CC	w/ SP, w/o CC	w/o SP, w/ CC	w/ SP, w/ CC
Purdue	ω_{in} = 0.001	ψ	0.803	0.719	0.725	0.763
		σ	0.025	0.063	0.150	0.050
		ξ	0.090	0.090	0.091	0.092
		κ	0.815	0.751	0.770	0.789
		λ	7.093	5.620	5.217	6.214
	ω_{in} = 0.003	ψ	**0.834**	0.816	0.810	0.808
		σ	**0.007**	0.117	0.057	0.023
		ξ	**0.092**	0.091	0.092	0.096
		κ	**0.874**	0.852	0.859	0.876
		λ	**7.872**	6.715	7.106	7.210
	ω_{in} = 0.01	ψ	0.777	0.753	0.687	0.728
		σ	0.020	0.091	0.068	0.058
		ξ	0.094	0.092	0.114	0.102
		κ	0.790	0.728	0.807	0.840
		λ	6.407	5.414	4.526	5.652
NCSU	ω_{in} = 0.001	ψ	0.775	0.746	0.790	0.711
		σ	0.069	0.038	0.088	0.054
		ξ	0.106	0.101	0.095	0.095
		κ	0.822	0.833	0.835	0.762
		λ	5.607	5.916	6.305	5.387
	ω_{in} = 0.003	ψ	**0.822**	0.785	0.779	0.777
		σ	**0.019**	0.033	0.035	0.088
		ξ	**0.100**	0.107	0.108	0.095
		κ	**0.888**	0.833	0.831	0.809
		λ	**7.158**	5.903	5.768	6.029
	ω_{in} = 0.01	ψ	0.823	0.766	0.787	0.745
		σ	0.066	0.057	0.025	0.133
		ξ	0.098	0.092	0.100	0.102
		κ	0.877	0.813	0.819	0.803
		λ	6.861	6.421	6.284	5.089

- MAPPO [10]: It is a state-of-the-art MADRL approach as an extension of PPO. It uses a number of practical techniques (e.g., value normalization and incorporating agent-specific features to state) to improve the performance of multi-agent tasks.
- e-Divert [40]: It is a state-of-the-art MADRL approach for SC tasks. It is based on CTDE schemes aided by a distributed prioritized experience replay and an LSTM for sequential modeling.
- Shortest Path: Each UV finds the shortest path by genetic algorithm to visit a sequence of PoIs. Note that shortest paths of UGVs are under the restriction of roadmap.
- Random: For each UV k, action $\boldsymbol{a}_t^k$ is uniformly sampled from its action space at timeslot t.

We use efficiency λ as an integrated, comprehensive metric for comparisons, while showing individual metrics in air-ground SC tasks in Purdue and NCSU. We use Pytorch 1.8.1 for implementation on Ubuntu 18.04.2 LTS with eight GeForce RTX 3090 graphic cards. We train all methods 10,000 iterations, and test each model 50 times to take an average.

B. Hyperparameter Tuning

Suitable hyperparameters in i-EOI and h-CoPO will significantly improve the overall performance of h/i-MADRL. As shown in Table III, We tune ω_{in} to study the impact of introducing different portion of the intrinsic reward in i-EOI. Then, for h-CoPO, we jointly consider whether UVs share the same neural network parameters (SP, i.e., forced to be homogeneous), and whether to use the centralized critic network (CC, i.e., use the global state $\boldsymbol{s}$ as the input to each V^k). For other hyperparameters, we simply use common settings as in [10].

TABLE IV
IMPACT OF LINEARLY DECREASED ω_{in}

ω_{in}	$0.01 \rightarrow 0.001$	$0.003 \rightarrow 0$
λ (Purdue)	7.803	7.744
λ (NCSU)	5.707	6.210

TABLE V
IMPACT OF NEIGHBOR RANGE

% w.r.t task area size	10	25	33	50	66
λ (Purdue)	6.870	**7.872**	6.381	6.800	5.960
λ (NCSU)	6.214	**7.158**	6.258	6.234	5.304

We observe that $\omega_{\text{in}} = 0.003$ yields a peak in terms of efficiency λ. When ω_{in} is too low, UVs tend to visit similar areas since they all start at the same point, which results in waste of resources and many remote PoIs are hard to visited in the limited task duration. This is because UV behaviors may become similar and fail to visit different PoIs, if not considering too much individuality to explore far away distinct areas than others. On the contrary, giving too much weight ω_{in} on the intrinsic reward might do harm to the extrinsic reward from the environment. Besides fixed ω_{in}, we also linearly decrease ω_{in} during training. From Table IV, we find that it makes the intrinsic reward unstable and thus deteriorating the overall performance, compared with results in Table III). This is because individuality does not conflict with the overall goal of air-ground SC tasks (as mentioned in Section V-A), which differs from the original EOI [30].

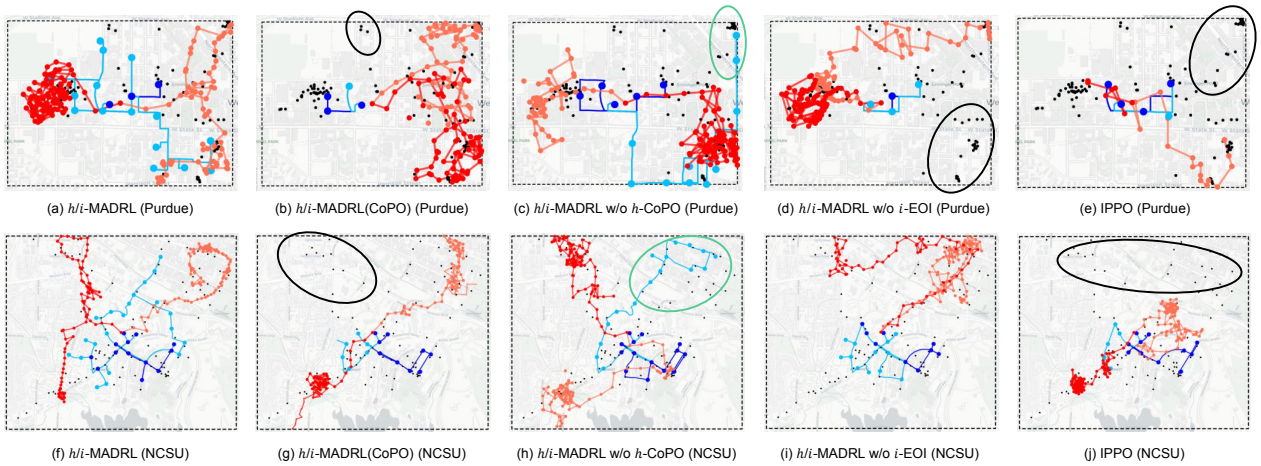
For h-CoPO, we find that it is not a good choice to share the same neural network parameters or utilize the centralized value network. This is because these two structures increase the homogeneity of UVs, in terms of making decisions and estimating the state values, respectively, which are important tricks for QMIX [9] and MAPPO [10] in homogeneous cooperative games [7], [8]. However, they are not applicable to h-CoPO which considers how to take advantages of different UVs to form a heterogeneous coordination pattern in our considered air-ground SC tasks.

Next, we aim to study the impact of distance to distinguish physically nearby homogeneous neighbors, as a percentage w.r.t the size of the task area. From Table V, we see 25% gives the highest efficiency, because a shorter neighbor range may ignores some close-by useful homogeneous UAVs/UGVs for cooperation, while a much longer range includes unnecessary UVs which should not be cooperated at this time. Hence, we observe that hyperparameter set "ω_{in}= 0.003, w/o SP, w/o CC, neighbor range is 25% w.r.t task area size" yields the best performance in terms of efficiency λ which will be used for the rest of performance comparisons.

C. Ablation Study

We perform the ablation study by gradually removing two plug-in modules of h/i-MADRL, under the default simulation settings (see Table VI). When i-EOI is removed, We observe that the data collection ratio in Purdue and NCSU is reduced

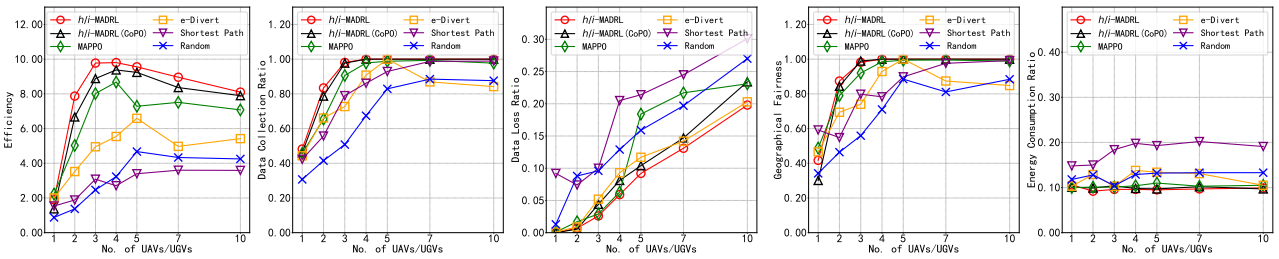
Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.



(a) h/i-MADRL (Purdue) (b) h/i-MADRL(CoPO) (Purdue) (c) h/i-MADRL w/o h-CoPO (Purdue) (d) h/i-MADRL w/o i-EOI (Purdue) (e) IPPO (Purdue)

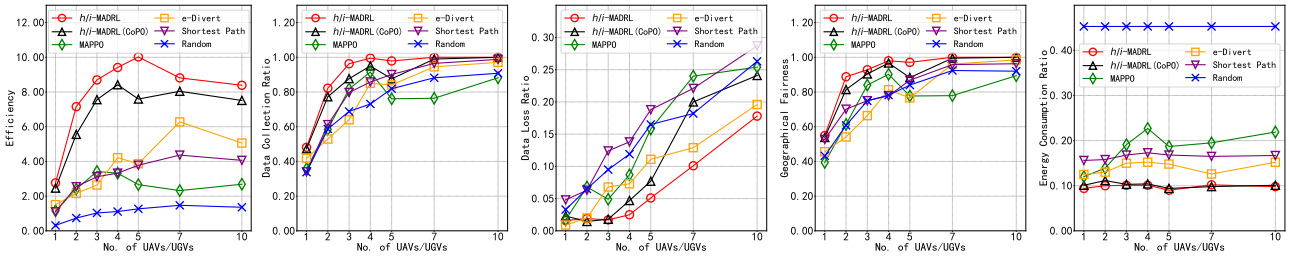
(f) h/i-MADRL (NCSU) (g) h/i-MADRL(CoPO) (NCSU) (h) h/i-MADRL w/o h-CoPO (NCSU) (i) h/i-MADRL w/o i-EOI (NCSU) (j) IPPO (NCSU)

Fig. 2. Different trajectory patterns over ablation study in Purdue and NCSU (UAVs: red/pink, UGVs: blue/light blue, PoIs: black dots; green ellipses indicate self-interested pattern of UGVs, and black ellipses denote areas with unvisited PoIs).



(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 3. Impact of no. of UAVs/UGVs (Purdue).



(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 4. Impact of no. of UAVs/UGVs (NCSU).

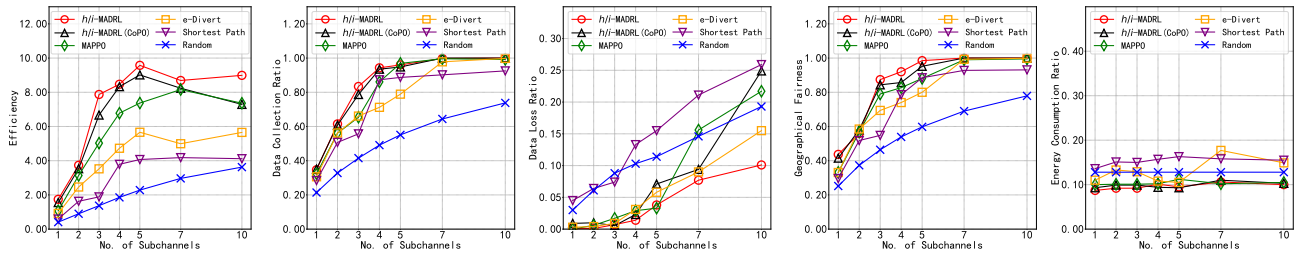
TABLE VI
ABLATION STUDY

		ψ	σ	ξ	κ	λ
Purdue	h/i-MADRL	**0.834**	**0.007**	**0.092**	**0.874**	**7.872**
	h/i-MADRL w/o i-EOI	0.699	0.021	0.096	0.846	6.020
	h/i-MADRL w/o h-CoPO	0.745	0.061	0.102	0.816	5.571
	h/i-MADRL w/o i-EOI, h-CoPO	0.654	0.100	0.101	0.516	3.007
NCSU	h/i-MADRL	**0.822**	**0.019**	**0.100**	**0.888**	**7.158**
	h/i-MADRL w/o i-EOI	0.706	0.056	0.109	0.806	4.922
	h/i-MADRL w/o h-CoPO	0.777	0.088	0.106	0.809	5.404
	h/i-MADRL w/o i-EOI, h-CoPO	0.585	0.068	0.139	0.613	2.404

by 13% and 12%, respectively. This is because that lack of individuality will make different UVs behave similarly and fail to collect data from those PoIs especially in corner areas. When h-CoPO is removed, we observe that the data loss ratio increases by 5% and 7% for two datasets, respectively, resulting in the decrease of efficiency. This is because lack of UV cooperations will make heterogeneous UAVs and UGVs become more self-interested and fail to form the successful data upload by AG-NOMA uplink channel.

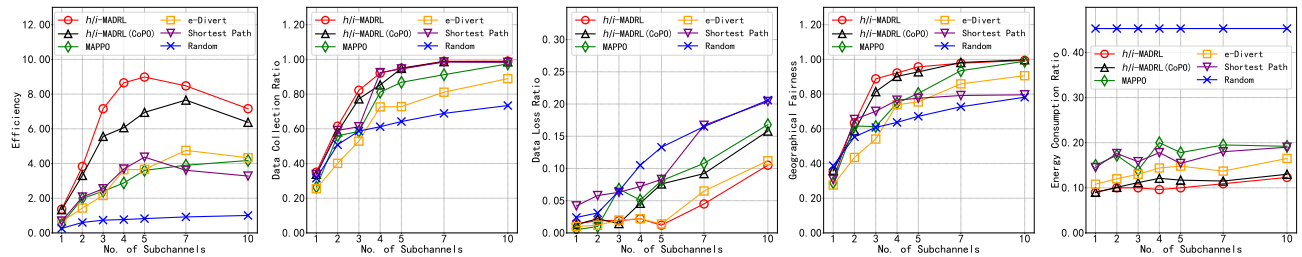
We further visualize the trajectories in both campus datasets (see Fig. 2). By comparing subfigures in Fig. 2 vertically, when using CoPO instead of our proposed h-CoPO, we see that UGVs fail to approach UAVs to serve as mobile BSs. This is because lack of heterogeneity makes it difficult to form an AG-NOMA communications pattern for data collection, where UGVs getting closer to UAVs results in higher data rates. Therefore, UAVs spend more time to catch UGVs to avoid data loss, which eventually lead to some unvisited PoIs as shown in Fig. 2(b) and (g). If entirely removing h-CoPO, we observe selfish behaviors of UGVs in Fig. 2(c), (h). In both datasets,

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.



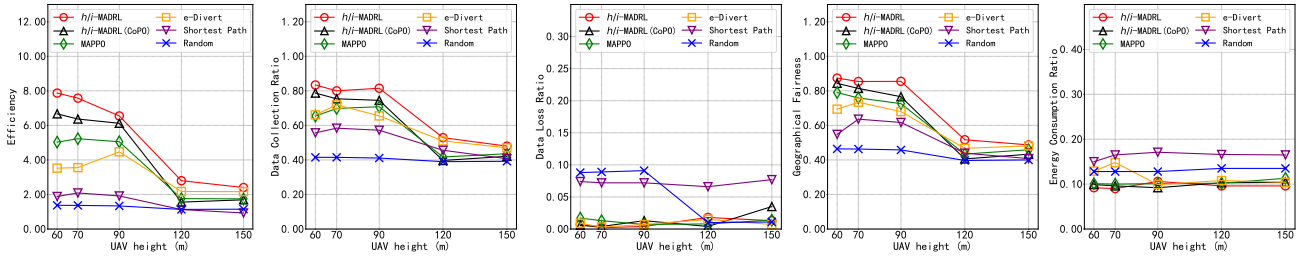
(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 5. Impact of no. of subchannels (Purdue).



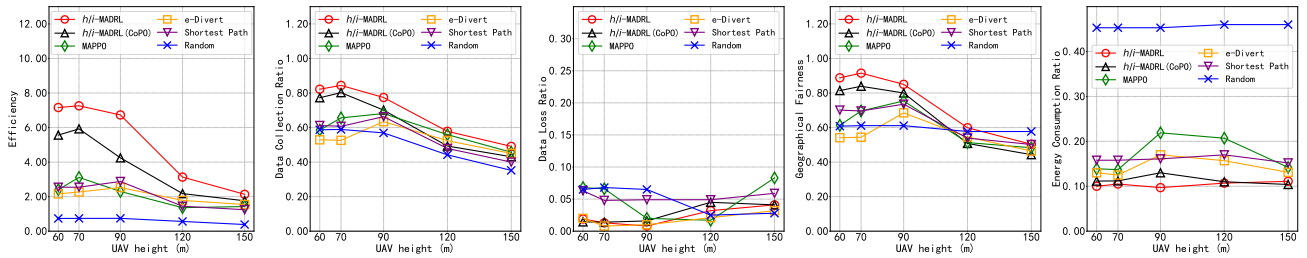
(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 6. Impact of no. of subchannels (NCSU).



(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 7. Impact of UAV hovering height (Purdue).



(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 8. Impact of UAV hovering height (NCSU).

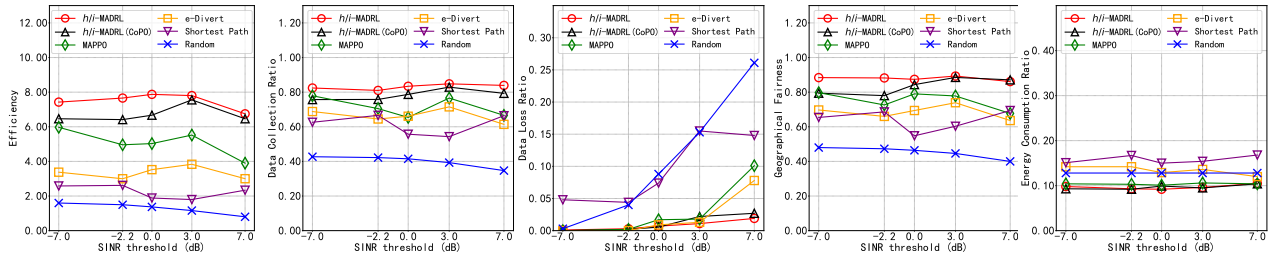
the light blue UGV goes to the upper right corner to collect data and thus receiving more extrinsic reward. This is because that UGVs are exploring the individuality only by using i-EOI and form egocentric behaviors, regardless of their limited mobility compared to UAVs. Next, we discuss the advantages of i-EOI by comparing Fig. 2(a), (f) and Fig. 2(d), (i). We observe too many unvisited PoIs, because all UAVs and UGVs lose individuality and visit similar areas, rather than the proper division of work. The similar results can be seen in Fig. 2(e), (j), where UAVs and UGVs behave similarly around the start point and fail to access PoIs in remote areas.

D. Comparing with Five Baselines

To justify the effectiveness and robustness of h/i-MADRL, we change the number of UAVs/UGVs (of equal amount), SINR threshold, number of subchannels and hovering height of UAVs, respectively.

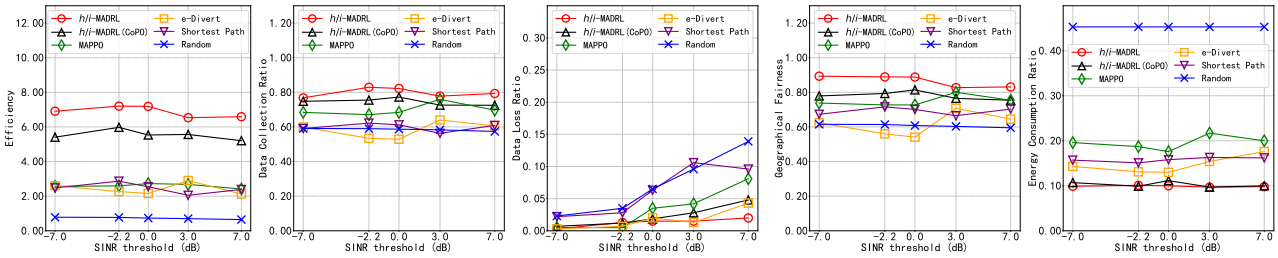
1) Impact of no. of UAVs/UGVs: With fewer UVs, h/i-MADRL achieves higher performance in both datasets (see Fig. 3 and Fig. 4). Taking NCSU for example as a big campus, when 2 UAVs and UGVs are deployed, h/i-MADRL obtains $\lambda = 7.158$ which is 1.97 and 2.31 times higher than MAPPO and e-Divert (see Fig. 4(a)). Without i-EOI, all five baselines

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.



(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 9. Impact of SINR threshold (Purdue).



(a) λ (b) ψ (c) σ (d) κ (e) ξ

Fig. 10. Impact of SINR threshold (NCSU).

cannot navigate UVs to behave in different mobility patterns. As a result, UVs fail to access PoIs in remote areas (see Fig. 4(b) and Fig. 4(d)).

As more UVs are deployed, the attained efficiencies of all methods first rise and then drop, since the capability of data collection becomes saturated and the task starts to suffer from a severe data loss (see Fig. 3(c) and Fig. 4(c)). This is because deploying more UVs increase the density and the interference of AG-NOMA data uplink. As a result, lower received SINR is more likely to happen. When no. of UAVs/UGVs is 5 in NCSU, h/i-MADRL obtains $\lambda = 10.024$, 32% higher than h/i-MADRL(CoPO) (see Fig. 4(a)). The main reason is that CoPO does not make a distinction between them UAVs and UGVs, while h-CoPO forms a more detailed and accurate cooperation by treating them as heterogeneous modalities, which may help to improve the quality of AG-NOMA uplink channel and alleviate the consequence of data loss.

2) Impact of no. of subchannels: As shown in Fig. 5 and Fig. 6, the efficiencies of all methods first increase and then decrease when more subchannels are employed. More subchannels pose the challenge of the larger solution space, which makes it more challenging to optimize UV scheduling policies. However, even when the no. of subchannels is 10, h/i-MADRL still achieves the highest efficiency $\lambda = 8.991$ in Purdue, which is 23% higher than that of the best baseline h/i-MADRL(CoPO) (see Fig. 5(a)).

3) Impact of UAV hovering height: As shown in Fig. 7 and Fig. 8, we observe that h/i-MADRL significantly outperforms all baselines in terms of efficiency when the hovering height of UAVs is relatively low (from 50m to 90m). However, this advantage becomes weaker as the hovering height increases. This is because lower height can decrease the path loss and increase the average capacity of the PoI-UAV uplink channel and the UAV-UGV relay channel, which strengthens the UAVs' engagement in AG-NOMA and brings clearer benefits of considering cooperation and individuality. On the contrary, if UAVs are deployed so high that relayed data from UAVs can be ignored due to the big path loss, our scenarios can be simplified as a ground SC scenario (the reason why MAPPO and e-Divert can achieve performance close to h/i-MADRL when the UAV hovering height is 150m).

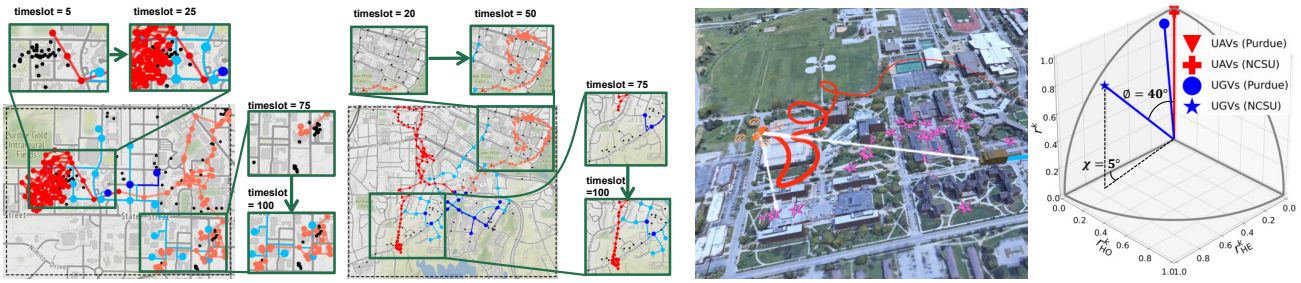
4) Impact of SINR threshold: As shown in Fig. 9 and Fig. 10, we observe that h/i-MADRL is quite robust to SINR threshold (referring to as the QoS constraints in communications systems) and its attained data loss ratio is significantly lower than all baselines. Even when the SINR threshold is 7.0dB (very stringent QoS requirement), data loss ratio of h/i-MADRL is only 14% and 33% of Shortest Path in Purdue and NCSU (see Fig. 9(c) and Fig. 10(c)). Also, Shortest Path severely suffers from a high data loss when SINR threshold increases, since it only optimizes the trajectories based on the total movement distance. As a result, the benefits using UAV and UGV in a relayed pair is not exploited.

Finally, regardless of above settings, h/i-MADRL keeps energy consumption to lowest level. Especially in NCSU of bigger campus, energy consumption ratio is 20% of the Random approach on average (see Fig. 10(e)). Also, without encouraging UV cooperation by h-CoPO, both MAPPO and e-Divert tend to deploy UGVs to access PoIs in remote areas, thus consuming a relatively big amount of energy to obtain a higher data collection ratio. However, their efficiencies are still lower than h/i-MADRL.

E. Visualization of UV cooperation

To better illustrate "air-ground coordination" by h/i-MADRL, we further demonstrate two highlighted trajectories

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.



(a) UV coordination (Purdue) (b) UV coordination (NCSU) (c) Implemented simulator (Purdue) (d) (ϕ, χ) value of UVs

Fig. 11. UV coordination and LCFs value of UVs (roadmaps are visually illustrated in (a) and (b); In (c), pink stars refer to PoIs, and white lines refer to the PoI-UAV uplink channel and the UAV-UGV relay channel).

TABLE VII
COMPUTATIONAL COMPLEXITY

Method	Time Cost (ms)	Graphic Card Mem. Usage (MB)
h/i-MADRL	**1.329**	**686**
h/i-MADRL(CoPO)	1.329	686
MAPPO	1.329	686
e-Divert	2.248	793

of UVs in Fig. 11(a) and Fig. 11(b). we observe a clear coordination pattern of the light blue UGV in both datasets. In NCSU, during timeslots 20∼50, the light blue UGV stays besides the pink UAV in order to receive its relayed data. Then, during timeslots 75∼100, the light blue UGV did not access the remote area until the red UAV arrives. This example confirms that h-CoPO module helps the spatio-temporal synchronization between UAVs and UGVs in the same subchannel. We also implemented a simulator by Unity 2021.3.4f1c1 as shown in Fig. 11(c). In this snapshot, the coordination pattern between PoIs, UAV and UGV during timeslots 20∼50 is presented.

Furthermore, we visualize the learned mean LCF values (ϕ, χ) of UAVs and UGVs in Fig. 11(d). We observe that UGVs' average ϕ is around $10°$ and $40°$ in z-axis in Purdue and NCSU respectively, which indicates UGVs put a certain degree of attention into their neighbors, and UGVs in NCSU are more cooperative than those in Purdue. In addition, UGVs also focus more on collaborating UAVs for data relaying, confirmed by $\chi = 40°$ and $\chi = 5°$ in x/y-axis in Purdue and NCSU campuses, respectively. On the contrary, UAVs' ϕ remains nearly $0°$ in z-axis, since UAVs move quickly and collect more data than UGVs, while UGVs move relatively slowly to become mobile BSs for receiving and decoding UAVs' relayed data. As our defined reward function is largely determined by the collected data amount, UAVs whose only responsibilities are data collection should become more egoistic than UGVs.

F. Computational Complexity

Since h/i-MADRL belongs to the category of multi-agent Markovian actor-critic algorithms which uses mini-batches of experience to train policies by maximizing the discounted reward, we can follow [41] and use the sample complexity to characterize the convergence rate, approximated by achieving $\mathbb{E}\left[\|\nabla_{\theta^k} J_{\mathrm{CO}}(\theta^k)\|^2\right] \leq \epsilon$ (defined in Eqn. (28)) , denoted by: $\mathcal{O}\left(\epsilon^{-2}\log\left(\epsilon^{-1}\right)\right)$, where ϵ is hyperparameter. From our experiments, we obtain that h/i-MADRL requires 225k and 606k samples to achieve $\epsilon = 0.5, 0.4$ on Purdue, and 479k and 659k samples to achieve $\epsilon = 0.7, 0.6$ on NCSU. By contrast, MAPPO requires 493k and 868k on Purdue, 590k and 750k on NCSU, a lot more than h/i-MADRL to achieve the same ϵ value, respectively.

h/i-MADRL only contains fully connected layers, which achieves fast inference speed on both CPUs and GPUs. That is, GPUs are not necessary in h/i-MADRL. Considering GPU is expensive and of big size, which makes it costly to be deployed in UAVs, our proposed solution is cost-effective in real air-ground SC tasks. Table VII shows the time cost results. We observe that the running time to select actions by h/i-MADRL in a timeslot is the same as MAPPO (i.e, same as the exemplar base module IPPO), since the probability classifier introduced by i-EOI and three value network introduced by h-CoPO are only used in training, under the CTDE framework. That is, h/i-MADRL improves performance without introducing additional time and space costs.

VII. CONCLUSION

In this paper, we propose a novel MADRL framework called h/i-MADRL for air-ground SC tasks, which consists of one base module of any actor-critic MARL algorithm (e.g., IPPO, MADDPG, etc.), and two novel plug-in modules i-EOI and h-CoPO. i-EOI helps accomplish a better spatial division of work by adding intrinsic rewards, and h-CoPO enhances the capacity of accurately modeling the cooperation preference among heterogeneous UAVs and UGVs. Extensive experimental results on two real-world datasets from student trajectories in Purdue and NCSU campuses confirm that h/i-MADRL achieves a greater exploration of both individuality and cooperation, resulting in a better performance in terms of efficiency compared with all five baselines.

ACKNOWLEDGMENT

This paper was sponsored by the National Natural Science Foundation of China (No. U21A20519 and 62022017). Ye Yuan is supported by the NSFC (No. 61932004, 62225203 and U21A20516). Guoren Wang is supported by the NSFC (No. 61732003 and U2001211). Corresponding Author: Jianxin Zhao.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Z. Chen, P. Cheng, Y. Zeng *et al.*, "Minimizing maximum delay of task assignment in spatial crowdsourcing," in *IEEE ICDE'19*, 2019, pp. 1454–1465.

[2] Y. Cheng, B. Li, X. Zhou *et al.*, "Real-time cross online matching in spatial crowdsourcing," in *IEEE ICDE'20*, 2020, pp. 1–12.

[3] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org , 2017.

[4] M. Galeso, *Waze: An Easy Guide to the Best Features*, 1st ed. North Charleston, SC, USA: CreateSpace Independent Publishing Platform, 2016.

[5] M. Liu, G. Gui, N. Zhao *et al.*, "Uav-aided air-to-ground cooperative nonorthogonal multiple access," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2704–2715, 2019.

[6] R. Lowe, Y. I. Wu, A. Tamar *et al.*, "Multi-agent actor-critic for mixed cooperative-competitive environments," *NeurIPS'17*, vol. 30, 2017.

[7] M. Samvelyan, T. Rashid, C. S. De Witt *et al.*, "The starcraft multi-agent challenge," *arXiv preprint arXiv:1902.04043*, 2019.

[8] K. Kurach, A. Raichuk, P. Stańczyk *et al.*, "Google research football: A novel reinforcement learning environment," *arXiv preprint arXiv:1907.11180*, 2019.

[9] T. Rashid, M. Samvelyan, C. Schroeder *et al.*, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *ICML'18*, 2018, pp. 4295–4304.

[10] C. Yu, A. Velu, E. Vinitsky *et al.*, "The surprising effectiveness of MAPPO in cooperative, multi-agent games," *CoRR*, vol. abs/2103.01955, 2021.

[11] C. S. de Witt, T. Gupta, D. Makoviichuk *et al.*, "Is independent learning all you need in the starcraft multi-agent challenge?" *arXiv preprint arXiv:2011.09533*, 2020.

[12] Y. Zhao, J. Guo, X. Chen *et al.*, "Coalition-based task assignment in spatial crowdsourcing," in *IEEE ICDE'21*, 2021, pp. 241–252.

[13] Y. Zhao, K. Zheng, J. Guo *et al.*, "Fairness-aware task assignment in spatial crowdsourcing: Game-theoretic approaches," in *IEEE ICDE'21*, 2021, pp. 265–276.

[14] T. Wang, X. Xie, X. Cao *et al.*, "On efficient and scalable time-continuous spatial crowdsourcing," in *IEEE ICDE'21*, 2021, pp. 1212–1223.

[15] S. Xu, J. Zhang, S. Meng *et al.*, "Task allocation for unmanned aerial vehicles in mobile crowdsensing," *Wireless Networks*, pp. 1–13, 2021.

[16] C. H. Liu, Y. Zhao, Z. Dai *et al.*, "Curiosity-driven energy-efficient worker scheduling in vehicular crowdsourcing: A deep reinforcement learning approach," in *IEEE ICDE'20*, 2020, pp. 25–36.

[17] H. Wang, C. H. Liu, Z. Dai *et al.*, "Energy-efficient 3d vehicular crowdsourcing for disaster response by distributed deep reinforcement learning," in *KDD'21*, 2021, pp. 3679–3687.

[18] L. Ding, D. Zhao, M. Cao, and H. Ma, "When crowdsourcing meets unmanned vehicles: Toward cost-effective collaborative urban sensing via deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 12 150–12 162, 2021.

[19] P. Sunehag, G. Lever, A. Gruslys *et al.*, "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.

[20] W. Qiu, X. Wang, R. Yu, R. Wang, X. He, B. An, S. Obraztsova, and Z. Rabinovich, "Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 049–23 062, 2021.

[21] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, "Qplex: Duplex dueling multi-agent q-learning," *arXiv preprint arXiv:2008.01062*, 2020.

[22] B. Peng, T. Rashid, C. Schroeder de Witt, P.-A. Kamienny, P. Torr, W. Böhmer, and S. Whiteson, "Facmac: Factored multi-agent centralised policy gradients," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 208–12 221, 2021.

[23] C. H. Liu, C. Piao, and J. Tang, "Energy-efficient uav crowdsensing with multiple charging stations by deep learning," in *IEEE INFOCOM'20*, 2020, pp. 199–208.

[24] C. Piao and C. H. Liu, "Energy-efficient mobile crowdsensing by unmanned vehicles: A sequential deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6312–6324, 2020.

[25] N. Jaques, A. Lazaridou, E. Hughes *et al.*, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *ICML'19*, 2019, pp. 3040–3049.

[26] Y. Du, L. Han, M. Fang *et al.*, "Liir: Learning individual intrinsic reward in multi-agent reinforcement learning," *NeurIPS'19*, vol. 32, 2019.

[27] L. Dai, B. Wang, Y. Yuan *et al.*, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[28] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 1984.

[29] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*, ser. Springer Briefs in Intelligent Systems. Springer, 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-28929-8

[30] J. Jiang and Z. Lu, "The emergence of individuality," in *ICML'21*, 2021, pp. 4992–5001.

[31] Z. Peng, K. M. Hui, C. Liu *et al.*, "Learning to simulate self-driven particles system with coordinated policy optimization," *NeurIPS'21*, vol. 34, 2021.

[32] Y. Yang, R. Luo, M. Li *et al.*, "Mean field multi-agent reinforcement learning," in *ICML'18*, 2018, pp. 5571–5580.

[33] W. B. Liebrand, "The effect of social motives, communication and group size on behaviour in an n-person multi-stage mixed-motive game," *European journal of social psychology*, vol. 14, no. 3, pp. 239–264, 1984.

[34] W. Schwarting, A. Pierson, J. Alonso-Mora *et al.*, "Social behavior for autonomous vehicles," *Proceedings of the National Academy of Sciences*, vol. 116, no. 50, pp. 24 972–24 978, 2019.

[35] J. Schulman, F. Wolski, P. Dhariwal *et al.*, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[36] R. S. Sutton, D. A. McAllester, S. P. Singh *et al.*, "Policy gradient methods for reinforcement learning with function approximation," in *NeurIPS'00*, 2000, pp. 1057–1063.

[37] H. Zhang, M. A. Roth, R. K. Panta *et al.*, "Crowdbind: Fairness enhanced late binding task scheduling in mobile crowdsensing," *EWSN'20*, pp. 1–12, 2020.

[38] I. Rhee, M. Shin, S. Hong *et al.*, "CRAWDAD dataset ncsu/mobilitymodels (v. 2009-07-23)," Downloaded from https://crawdad.org/ncsu/mobilitymodels/20090723, Jul. 2009.

[39] DJI, "Matrice 600 pro - product information - dji," https://www.dji.com/cn/matrice600-pro/infospecs.

[40] C. H. Liu, Z. Dai, Y. Zhao *et al.*, "Distributed and energy-efficient mobile crowdsensing with charging stations by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 20, no. 1, pp. 130–146, 2019.

[41] F. Hairi, J. Liu, and S. Lu, "Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward," in *ICLR'22*, 2022.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on August 02,2025 at 13:36:44 UTC from IEEE Xplore. Restrictions apply.