MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

UAV Aided Smart Agriculture Networks: A Multi-Agent Reinforcement Learning Approach

Xiong, Guojun; Guo, Jianlin; Parsons, Kieran; Nagai, Yukimasa; Sumi, Takenori; Orlik, Philip V.; Li, Jian

TR2025-082 June 10, 2025

Abstract

This paper explores the transformative potential of the IoT paradigm in promoting smart agriculture. Key challenges lie in how to connect agriculture sensors to remote cloud servers in the absence of feasible communication infrastructure and the unreliable wireless links in rural areas. To address these issues, we propose an innovative two-tier smart agriculture architecture: an Unmanned Aerial Vehicle (UAV) aided agriculture network model, which leverages UAVs as intermediaries to collect and route data from agriculture sensors to cloud servers. This novel architecture leads to two particular problems, i.e., data packet scheduling in the first-tier networks and multi-hop routing in the second-tier UAV mesh network. To that end, we present formal Markov decision process (MDP) based problem formulations for both tiers, with a primary focus on the more challenging multi-hop routing problem in the second-tier network. This problem is approached as a multi-agent reinforcement learning (MARL) framework, for which we introduce a novel distributed algorithm – Focus Coordination: attention-guided Multi-Agent Deep Deterministic Policy Gradient (FC-MADDPG). This algorithm reduces communication overhead and mitigates the risks associated with single-node failures. We evaluated the performance of the proposed FC-MADDPG algorithm, demonstrating its efficacy in enhancing data transmission reliability and efficiency.

IEEE International Conference on Communications Workshops (ICC) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

UAV Aided Smart Agriculture Networks: A Multi-Agent Reinforcement Learning Approach

Guojun Xiong*‡, Jianlin Guo*, Kieran Parsons*, Yukimasa Nagai†, Takenori Sumi†, Philip Orlik* and Jian Li‡
*Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
†Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Kanagawa 2478501, Japan
‡ Department of Computer Science & Applied Mathematics and Statistics, Stony Brook University, 100 Nicolls Road, Stony Brook, NY 11794, USA

***Abstract*—This paper explores the transformative potential of the IoT paradigm in promoting smart agriculture. Key challenges lie in how to connect agriculture sensors to remote cloud servers in the absence of feasible communication infrastructure and the unreliable wireless links in rural areas. To address these issues, we propose an innovative two-tier smart agriculture architecture: an Unmanned Aerial Vehicle (UAV) aided agriculture network model, which leverages UAVs as intermediaries to collect and route data from agriculture sensors to cloud servers. This novel architecture leads to two particular problems, i.e., data packet scheduling in the first-tier networks and multi-hop routing in the second-tier UAV mesh network. To that end, we present formal Markov decision process (MDP) based problem formulations for both tiers, with a primary focus on the more challenging multi-hop routing problem in the second-tier network. This problem is approached as a multi-agent reinforcement learning (MARL) framework, for which we introduce a novel distributed algorithm – Focus Coordination: attention-guided Multi-Agent Deep Deterministic Policy Gradient (FC-MADDPG). This algorithm reduces communication overhead and mitigates the risks associated with single-node failures. We evaluated the performance of the proposed FC-MADDPG algorithm, demonstrating its efficacy in enhancing data transmission reliability and efficiency.**

***Index Terms*—UAV aided smart agriculture, dynamic two-tier network model, MDP problem formulation, MARL based multi-hop routing.**

I. Introduction

As global population grows, sustainable and efficient agricultural practices have become imperative. Internet of Things (IoT) has been playing important role in urban applications such as smart city and smart utility. IoT paradigm also fits many agriculture use cases such as crop monitoring, soil moisture sensing and predictive analytics for smart farming [1]–[3]. However, the realization of smart agriculture faces the challenges as well. The lack of feasible communication infrastructure in rural area is a major issue for agriculture sensors to communicate with remote data centers. Poor communication connectivity in rural area is another issue. Efficient agriculture sensor data delivery presents additional challenge. Hardware and operation cost is also a concern.

Smart agriculture technologies can be divided into three categories: sensing, cloud computing, and networking. There are works such as SoilTech project developing smart agriculture sensing technology. There are also works such as Microsoft FarmBeats developing the cloud computing technology. However, the networking technology, a bridge between agriculture sensors and cloud servers, is less studied.

This paper delves into the promising capabilities of IoT paradigm in revolutionizing smart agriculture networking. To tackle the lack of feasible communication infrastructure and the inefficient data transfer due to unstable wireless connections in rural areas, we explore a new use case of the UAV in smart agriculture, where UAVs are used as dynamic communication infrastructure. We propose a novel two-tier smart agriculture network model that utilizes UAVs as sensor data collectors and relay agents. This UAV aided network model can significantly enhance the reliability and efficiency of data transfer from agriculture sensors to cloud servers. To the best of our knowledge, we are the first to propose such UAV aided two-tier smart agriculture network architecture.

We further propose formal MDP based problem formulations for data packet scheduling in the first-tier multi-point to point (MP2P) networks and multi-hop routing in the second-tier UAV mesh network upon the novel two-tier network architecture. In addition, we introduce the Focus Coordination: attention-guided Multi-Agent Deep Deterministic Policy Gradient (FC-MADDPG) algorithm to perform multi-hop routing in the second-tier UAV mesh network, with reduced communication overhead and mitigating vulnerabilities associated with centralized control systems. The simulation results show significant performance improvements over traditional methods, indicating the potential of our proposed solutions in revolutionizing data delivery in emerging smart farming.

II. Related Works

Smart agriculture draws attention from researchers. IoT, UAV, and machine learning especially reinforcement learning (RL) are the most promising technologies in this field.

The precision agriculture (PA) is a collage of strategies and technologies to optimize operations and decisions in smart farming. Work [4] models PA as a multi-agent patrolling problem, where robots visit subregions requiring immediate attention in the agricultural field to address the area coverage of monitoring crop health.

This work was done while Guojun Xiong was working at Mitsubishi Electric Research Laboratories (MERL) as an intern.

Authors in [2] present a smart agriculture IoT system based on deep RL, which includes four layers, namely agricultural data collection layer, edge computing layer, agricultural data transmission layer, and cloud computing layer. The presented system integrates some advanced information techniques such as artificial intelligence (AI) and cloud computing to increase agricultural production. Paper [1] highlight the modernization of traditional agriculture through IoT paradigm. Automation and IoT technologies via smart GPS-based remote controlled robot are applied to perform tasks like weeding, spraying, and moisture sensing.

Agricultural UAV has attracted remarkable academic attention recently. Work [5] reviews trends and applications of leading technologies related to agricultural UAVs, control technologies, equipment, and development. Authors in [3] conduct a comprehensive review based on bibliometrics to summarize and structure existing academic literature and reveal current research trends. Their analysis indicates that remote sensing, precision agriculture (PA), deep learning, machine learning, and IoT are critical topics related to agricultural UAV. Paper [6] proposes UAV remote sensing to offer high-resolution imagery and extract crop traits from 3D data. Work [7] surveys a detailed understanding of UAV applications in PA. This survey classifies UAV applications into three categories: a) UAV-based applications for tracking, b) UAV-based applications for spraying, and c) Multi-UAV applications, but it is noted that there is a shortage of research studying multi-UAV applications in the agriculture, which drives our propose of the multi-UAV aided smart agriculture network architecture.

The dynamic routing in UAV mesh networks has been studied recently. Paper [8] uses a deep Q-network (DQN) to design a routing scheme in a manned-and-unmanned airborne network. More recent works [9], [10] apply multi-agent deep reinforcement learning (DRL) algorithms to enhance routing decisions. However, these approaches rely on a central coordinator which suffer high communication overhead and risks associated with single-node failures.

III. System Model

This section presents the proposed UAV aided two-tier smart agriculture network model as illustrated in Fig. 1, in which agriculture sensors are divided into clusters, a first-tier MP2P network is dynamically formed by an agriculture sensor cluster and an assigned UAV to transfer data from sensors to the UAV, while the second-tier mesh network is dynamically formed by UAVs and a cloud server (CS) to route the collected sensor data from UAVs to the CS. In this smart agriculture network model, UAVs operate only in sensor data collection phase for cost reduction. Although each sensor cluster is assigned one UAV in Fig. 1, it is possible to assign multiple UAVs to one sensor cluster or assign one UAV to multiple sensor clusters. To the best of our knowledge, we are the first to propose such two-tier dynamic smart agriculture network architecture.

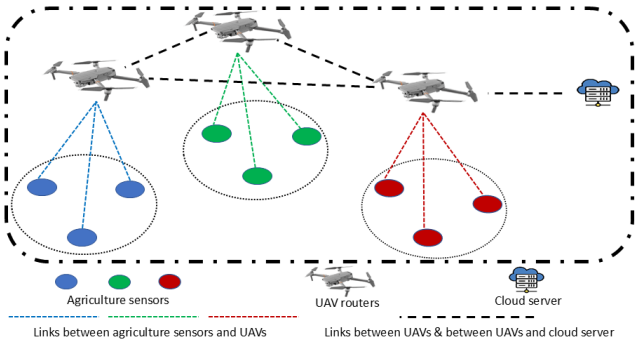

Fig. 1: UAV aided two-tier smart agriculture network model

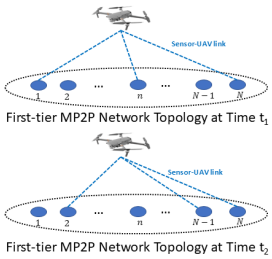

Fig. 2: First-tier MP2P Topology Evolution

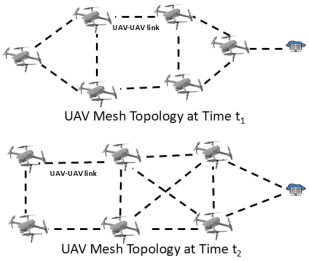

Fig. 3: Second-tier Mesh Topology Evolution

During data collection process, UAVs navigate and hover over their sensor clusters to form dynamic smart agriculture networks, where an UAV can communicate with multiple sensors simultaneously if it is equipped with multiple antennas. UAVs navigate over their sensor clusters to establish reliable communication links with the sensors whose data to be collected. Once reliable communication links are formed, UAVs hover over for data collection. Therefore, both network topologies are varying with respect to time. Fig. 2 illustrates topology variation of a first-tier MP2P network from time t_1 to time t_2 with different communication links and Fig. 3 shows the second-tier mesh topology variation from time t_1 to time t_2, indicating the challenge of agriculture sensor data collection and relay as well as the need of innovative solutions.

This network model considers fact that agriculture sensors are more practical for smart agriculture tasks such as daily monitoring and sensing, and they are typically equipped with short range communication radio operating in unlicensed frequency band for cost reduction. As a result, agriculture sensors may not have direct communication links with remote CS. Accordingly, a set of UAVs are used to bridge agriculture sensors and remote CS. The use of UAVs is a novel and practical approach to address the lack of feasible communication infrastructure issue in rural areas.

We assume there are K sensor clusters and each cluster consists of N sensors. The UAVs are indexed by set $\mathcal{K} = \{1, 2, \ldots, K\}$. It is possible that some UAVs may not be able to directly communicate with CS as well. This requires UAVs to collaboratively route sensor data to CS. Therefore,

UAVs form a mesh network to relay sensor data, indicating that UAVs need to find optimal routes to CS.

As aforementioned, each first-tier MP2P network consists of N agriculture sensors in a sensor cluster and one UAV serving that cluster. A sensor n senses data and generates data packets probabilistically and independently according to its probability p_n, while a UAV with the limited capacity can only collect data from C out of N sensors in each time slot, where the capacity is an integrated factor of communication capability and storage limitation. The challenge is how to select C sensors during each time slot to minimize the average latency. This optimization problem encompasses the need to efficiently allocate resources, consider probabilistic data generation, and prioritize real-time data delivery.

In the second-tier UAV mesh network, UAVs continuously traverse over their respective cluster regions to collect data while collaboratively routing the collected data to CS in regular intervals of T time slots. To effectively manage data relay in this dynamic environment, the dynamic routing is imperative. Traditionally, the routing is performed by protocol and optimization based methods. Recently, the RL based routing techniques have been proposed and can outperform traditional routing methods in complex network environments [8]–[11]. Therefore, we apply RL methodologies to design routing techniques that can efficiently allocate UAV resources and minimize latency.

Although we assume each sensor cluster consists of N sensors, the proposed model works for different cluster sizes.

IV. MDP Based Problem Formulations

This section provides formal problem formulations for the first-tier and the second-tier networks.

A. Data Packet Scheduling in the First-Tier Networks

Consider an agriculture sensor cluster $\mathcal{N} = \{1, \cdots, N\}$ as shown in Fig. 4, where agriculture sensors collect different data and generate different data packets to be transmitted to the remote CS. The goal of the UAV is to decide at each time slot which sensors to serve so that the cumulative value of the average packet delivery latency experienced by agriculture sensors is minimized.

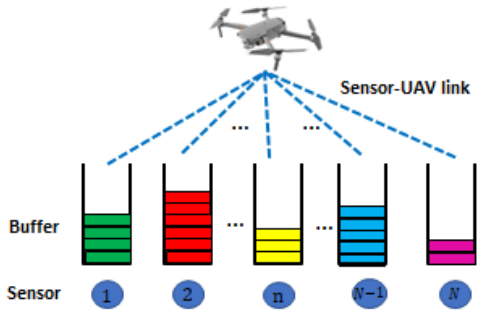

Fig. 4: Sensor queue for data packet buffering and scheduling

Packet Generation and Delivery Model: Consider a scheduling interval with T time slot denoted by $t \in \mathcal{T} = \{1, \ldots, T\}$. The data packets generated by sensor n are buffered in a queue as shown in Fig. 4. The queue length at time t is denoted by $X_{n,t}$. The number of data packets generated by sensor cluster may be larger than the service capacity of the UAV [12], [13]. Hence, the data packets may not be served immediately so that there will be a latency. In addition, the wireless links between the UAV and agriculture sensors can be unreliable. Assume successful data packet transmission probability for sensor n is q_n. This motivates us to consider a queuing model that captures the latency and successful packet transmission probability. The model is formulated as an MDP.

States: Denote the state of sensor cluster at time t as $\mathbf{X}_t := (X_{1,t}, \cdots, X_{N,t}) \in \mathbb{N}^N$, where $X_{n,t}$ is the number of outstanding data packets stored at sensor $n \in \mathcal{N}$. To guarantee the stability of the Markov chain, we assume $X_{n,t} \in [0, X_{max}]$, where X_{max} is the queue capacity, i.e., the maximum number of packets can be buffered by an agriculture sensor.

Actions: At each time slot t, the UAV makes a decision regarding whether or not to serve a sensor. Denote $U_{n,t}$ as action for sensor n with $U_{n,t} = 1$ indicating sensor n being served and $U_{n,t} = 0$ indicating otherwise. Let $\mathbf{U}_t := (U_{1,t}, \cdots, U_{N,t})$ be the vector of decisions for the sensor cluster. The capacity constraint of the UAV implies that $\mathbf{U}_t$ must satisfy the following constraint

$$\sum_{n=1}^{N} U_{n,t} \leq C, \quad \forall t. \tag{1}$$

We aim to design a policy $\pi : \mathbf{X}_t \mapsto \mathcal{U}^N = \{0,1\}^N$ maps the state $\mathbf{X}_t$ of the sensor cluster to decisions $\mathbf{U}_t = \pi(\mathbf{X}_t)$.

Transition Kernel: The state of the n-th agriculture sensor queue can change from $X_{n,t}$ to either $X_{n,t}+1$ or $X_{n,t}-1$ or keep unchanged from time t to $t+1$. Depending on the packet generation probability p_n, action $U_{n,t}$ and successful packet transmission probability q_n of agriculture sensor n, the detailed transitions are as follows

$$X_{n,t+1} = \begin{cases} X_{n,t}+1, & \text{w.p. } p_n(1-U_{n,t}) + p_n U_{n,t}(1-q_n), \\ X_{n,t}, & \text{w.p. } p_n U_{n,t} q_n + (1-p_n)(1-U_{n,t}) \\ & \quad +(1-p_n)U_{n,t}(1-q_n), \\ X_{n,t}-1, & \text{w.p. } (1-p_n)U_{n,t}q_n. \end{cases} \tag{2}$$

It is straightforward to verify that the summation of the probability for three scenarios equals 1.

Data Packet Delivery Problem: Little's Law indicates that minimizing average latency is equivalent to minimizing average number of outstanding data packets [13]. Let $C_{n,t}(X_{n,t}, U_{n,t}) := X_{n,t}$ be the instantaneous cost incurred by sensor n at time t, the cumulative cost incurred by the sensor cluster is given by

$$C_t(\mathbf{X}_t, \mathbf{U}_t) = \sum_{n=1}^{N} C_{n,t}(X_{n,t}, U_{n,t}) = \sum_{n=1}^{N} X_{n,t}. \tag{3}$$

Our objective is to derive a policy π to solve following MDP:

$$\min_{\pi\in\Pi} C_\pi := \limsup_{T\to\infty} \sum_{n=1}^{N} \frac{1}{T}\mathbb{E}_\pi \left[\sum_{t=0}^{T} X_{n,t}\right],$$
$$\text{s.t.} \sum_{n=1}^{N} U_{n,t} \le C, \ \forall t. \tag{4}$$

The problem (4) is an infinite-horizon average-cost problem and can be solved via existing methods such as the relative value iteration [14].

B. Data Routing in the Second-Tier UAV Mesh Network

UAV mesh routing is primary focus of this work. At any time t, the UAV mesh network is structured as an undirected graph $\mathcal{G} := (\tilde{\mathcal{K}}, \mathcal{E}(t))$, where $\tilde{\mathcal{K}} = \mathcal{K} \cup \{CS\}$ is the UAV mesh network node set and $\mathcal{E}(t)$ is the link set. A link (k, j) exists between node k and node j only when they can directly communicate and we denote $\mathcal{N}_k := \{j|(k, j) \text{ exists}\}$ as the neighbor set of node k. The links between UAVs are bidirectional, while the links between UAVs and the CS are one-directional, indicating no data flow from CS to UAVs. The goal of each UAV k is to find a routing policy π_k to minimize average latency, i.e., to maximize expected accumulated reward from RL perspective. Similarly, we formulate this UAV mesh network routing problem as an MDP as shown in Fig. 5.

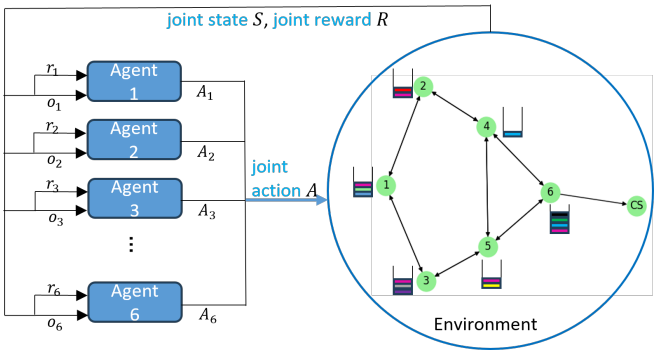

Fig. 5: MDP formulation of routing in UAV mesh network

States: We denote the state of UAV mesh network as $\mathbf{S}_t := (S_{1,t}, \cdots, S_{K,t}) \in \mathbb{N}^K$, where $S_{k,t}$ is the state of UAV $k \in \mathcal{K}$, representing the queue length ($\le S_{max}$) and the hop observed (HO). An UAV sets HO $= 1$ if the CS is its neighbor and HO $= 1+$ the minimal HO otherwise. The HO provides UAVs a reference for next-hop router selection.

Observation space: To reduce communication overhead and computational complexity, UAVs share their states with neighbors only, i.e., a UAV has local state observations only. Denote the observation of the UAV k as

$$O_{k,t} := S_{k,t} \cup \{S_{j,t}|j \ \in \mathcal{N}_k\},$$

where $S_{j,t}$ is an empty set if the neighbor j is the CS. The global state $\mathbf{S}_t$ of the environment is the joint observations, i.e., $\mathbf{S}_t = \cup_{k=1}^{K} O_{k,t}$.

Actions: At each time slot t, a UAV k has to make a decision regarding which neighbor node j it should select as next-hop router. Denote $A_{k,t}$ as the action for UAV k, where $A_{k,t} = j$ indicates neighbor node j is chosen. Let $\mathbf{A}_t := (A_{1,t}, \cdots, A_{K,t})$ be the vector consisting of decisions for all K UAVs at time slot t.

Reward Function. RL can transform optimization problems into maximizing the expected cumulative reward problems through appropriate reward function design. To design reward functions for the routing problem in the UAV mesh network, there are several key aspects to be considered.

- **Queue Length:** The queue length $S_{k,t}$ reflects the number of outstanding data packets waiting to be transmitted at the UAV k. A lower value of $S_{k,t}$ indicates lower latency, and hence we penalize the queue length when designing reward function.
- **End-to-End Delay:** This is quantified by the transmission time $T_{k,j}$ from node k to node j. The end-to-end delay is a critical metric in network performance, reflecting the time taken for a packet to travel across the network from node k to node j. A higher value of $T_{k,j}$ indicates slower packet travel, i.e., negative contribution to the reward function.
- **Packet Delivery Success/Failure:** This is indicated by an ACK signal $\mathbb{1}(ACK)$, which is a binary indicator with 1 representing successful packet delivery and 0 indicating failure. Successful delivery would increase the reward, while failure would decrease the reward. To use ACK mechanism, a UAV acknowledges transmitter when it successfully receives a data packet from a neighbor. Each UAV starts a timer to wait for the ACK when it transmits a data packet. If the timer expires without receiving ACK, the transmission is considered as failure. Otherwise, the transmission succeeds.
- **Congestion:** Congestion is inferred from the case that UAV's buffer length approaches the maximum value S_{max}, denoted by the signal $\mathbb{1}(congestion)$. Congestion can lead to packet delay and even packet loss. Thus, the reward function should be designed to penalize congestion.
- **Destination Arrival:** In the UAV mesh network routing, the destination of all data packets is the CS. Therefore, destination arrival is denoted by $\mathbb{1}(CS)$ with binary value 1 or 0, indicating whether a packet has arrived at CS. $\mathbb{1}(CS) = 1$ indicates a data packet has arrived at CS and thus would positively contribute to the reward. On the other hand, $\mathbb{1}(CS) = 0$ indicates a data packet has not arrived at CS and therefore, would contribute negatively to the reward.

Provided all the key factors, we propose a novel reward function as

$$\begin{aligned} R(S_{k,t}, A_{k,t}) := & -\alpha \cdot S_{k,t} - \beta \cdot \mathbb{1}(ACK) \cdot T_{k,A_{k,t}} \\ & - \kappa \cdot (1 - \mathbb{1}(ACK)) \\ & - \eta \cdot \mathbb{1}(congestion) + \mu \cdot \mathbb{1}(CS), \end{aligned}$$

where $\alpha, \beta, \eta, \kappa$ and μ are positive weight scalars. In the considered routing settings, each UAV aims to maximize its individual reward at the expense of the other UAVs. This is typical in game-theoretic scenarios. For competitive MARL, the objective of each UAV k can be written as

$$\max_{\pi_k} \mathbb{E}_{\pi_k} \left[\sum_{t=0}^{T} \gamma^t R(\mathbf{S}_t, A_{k,t}, A_{-k,t}) \right], \tag{5}$$

where $A_{-k,t}$ denotes the actions of all UAVs other than UAV k. Notice that the optimal policy of UAV k relies on the global state information $\mathbf{S}_t$ and actions of other UAVs, which might be impractical to obtain by UAV k in the multi-hop networks. We propose a MARL solution that only requires partial local observations, i.e., the states of the neighbors.

V. Proposed Attention-Guided MARL Method

Due to space limitation and the fact that the data packet scheduling problem in (4) can be solved by existing methods, we focus on solving the UAV mesh network routing problem (5) in this section, which is a multi-agent cooperative RL task that can be solved using multi-agent DDPG (MADDPG) framework [15], which is a actor-critic approach, where the "actor" is a policy network that decides the best action to take, while the "critic" is a value network that evaluates the action taken by the actor.

However, conventional MADDPG framework often grapples with significant drawbacks. The primary challenge is the substantial communication overhead due to the constant need for information exchange among UAVs, which strains network resources, especially in bandwidth-limited multi-hop scenarios. Another challenge is computational complexity due to the excessive states shared across the network. It is possible that the states of UAVs far away from each other may not contribute positively. Furthermore, these traditional methods typically depend on a central controller for coordination, creating a bottleneck in terms of computational load and posing a risk as a single point of failure, particularly as the network scales. This centralized approach also struggles with adaptability in dynamic environments, limiting the overall efficiency and robustness of the system.

To overcome these limitations, we propose the Focus Coordination: attention-guided Multi-Agent DDPG (FC-MADDPG) algorithm 1. Our approach innovatively integrates a query-key-value attention mechanism, allowing each UAV to autonomously determine the relevance of information from neighboring UAV agents, thus significantly reducing unnecessary communication overhead. FC-MADDPG algorithm also applies Gumbel-Softmax technique to approximate the gradient of discrete stochastic policies. In addition, by decentralizing the decision-making process, FC-MADDPG algorithm eliminates the need of a central controller, enhancing the scalability and resilience of UAV networks. This attention-based model is adept at adapting to changing environments and network topologies, ensuring that UAVs focus on processing the most pertinent information. FC-MADDPG algorithm, therefore, presents a robust, efficient, and scalable solution, bypassing the drawbacks of conventional methods and offering a significant advancement in managing the large sparse UAV networks.

Algorithm 1 FC-MADDPG

Initialization:
for each UAV agent k in the mesh network **do**
  Initialize an actor network $\pi_k(\theta_k)$ with weights θ_k for policy representation, and a critic network $Q_k(\omega_k)$ with weights ω_k for value estimation;
  Copy the initial weights to target networks $\pi'_k(\theta'_k)$, $Q'_k(\omega'_k)$;
  Prepare a local replay buffer $\mathcal{B}_k$ initialized to empty;
end for
For each episode:
  Reset the environment and receive the initial observation for each agent; *//Initial Observations*
for each agent k **do**
  Select action A_k using actor network π_k based on the current state;
  Apply actions, observe new state, rewards, and check episode end; *//Environment Interaction*
  Store transition tuple

$$(\{O_j, j \in \mathcal{N}_k\}, A_k, r_k, \{O'_j, j \in \mathcal{N}_k\})$$

  in $\mathcal{B}_k$; *//Storing Experiences*
  Update critic Q_k by minimizing the loss between predictions and target values;
  Update actor π_k using the sampled policy gradient;
  Update target networks π'_k, Q'_k with a mix of target and main network weights; *//Target Networks Update*
end for
End of Episode Handling:
  Reset the environment for the next episode if the terminal state is reached
Output:
A set of optimized policies for each UAV ensuring efficient routing and network performance

At the beginning of each episode, each UAV agent k prepares its initial observation O_k. At each training step, each agent k receives the observations O_j and actions A_j from its neighbor $j, \forall j \in \mathcal{N}_k$.

The critic network Q_k is updated to consider the effective state and action:

$$Q_k(\{O_j, A_j, \forall j \in \mathcal{N}_k\}; \omega_k), \tag{6}$$

where ω_k are the parameters of the critic network. In particular, inside the critic network, agent k leverages the attention mechanism to compute a weighted sum of observations and actions from each neighboring agent j. The output of the attention layer of agent k is denoted as:

$$c_k = \text{Attention}(\{[O_j, A_j], \forall j \in \mathcal{N}_k\}; \mathbf{W}_q, \mathbf{W}_u, \mathbf{W}_v), \tag{7}$$

where $\mathbf{W}_q, \mathbf{W}_u, \mathbf{W}_v$ are the parameters of attention model.

Similarly, the actor-network π_k is updated based on both the effective output of the Attention module and local observation $\{O_j, \forall j \in \mathcal{N}_k\}$:

$$\pi_k(\{O_j, \forall j \in \mathcal{N}_k\}; \theta_k), \tag{8}$$

where θ_k are the parameters of the actor-network.

The agent k then selects action A_k using actor-network π_k and adds Gumbel exploration noise to encourage policy exploration. After applying action, agent k gets the new state observations and rewards. It stores the transition tuple $(\{O_j, j \in \mathcal{N}_k\}, A_k, r_k, \{O'_j, j \in \mathcal{N}_k\})$ in the replay buffer $\mathcal{B}_k$, from which a mini-batch of transitions is randomly sampled for the critic Q_k, the actor π_k, and the target networks π'_k and Q'_k updating. More specifically, update Q_k by minimizing the loss between its predictions and the target values, update π_k using the sampled policy gradient, and update π'_k and Q'_k with a mix of target and main network weights. The optimized policies ensure efficient routing and robust mesh network performance.

VI. EXPERIMENTS

This section presents performance evaluation of our proposed FC-MADDPG algorithm and observes interesting insights. We compare our algorithm with four benchmark algorithms [1]: (1) The Deep Q-Network (DQN) algorithm; (2) The centralized MADDPG algorithm; (3) The independent MADDPG algorithm in which UAV agents do not share state information, and (4) Random Policy in which each UAV agent randomly selects a neighbor as next hop router.

We simulated algorithms under two distinct UAV mesh network topologies. In the simulation, each UAV functions as a routing agent, is equipped with a buffer of capacity 10 packets and is responsible to deliver 25 data packets. Packet arrival at each UAV is modeled to follow an exponential distribution, introducing stochastic elements to the arrival times, thus emulating realistic network traffic conditions. Congestion occurs when the number of packets surpasses the buffer's capacity. The simulation episode concludes once all packets have been successfully relayed to their respective destination nodes, signifying the completion of a data transmission cycle within the UAV mesh network.

A. Topology with 11 Nodes

The 11-node topology used in Fig. 6 is inspired by the well known Abilene Core Topology. The 11 nodes are deployed within 1 km by 1 km area. This configuration provides a real-world scenario and a baseline scenario for performance assessment.

Fig. 7 shows the number of time steps required to transmit all data packets to the CS and Fig. 8 tracks the accumulated reward, conceptualized as negative cost, to evaluate

[1]The selected benchmarks are highly representative. DQN represents a single-agent DRL algorithm, the centralized MADDPG represents a setting with a central controller, and independent MADDPG represents each agent training independently. These algorithms broadly encompass the characteristics of all existing benchmarks [8]–[10].

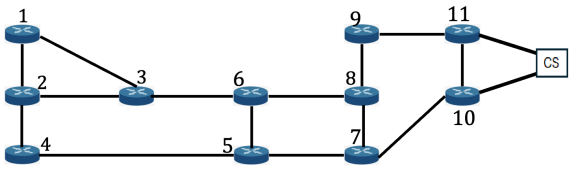

Fig. 6: Topology of a UAV mesh network with 11 nodes

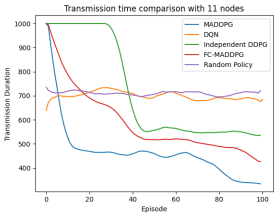

Fig. 7: TX time step

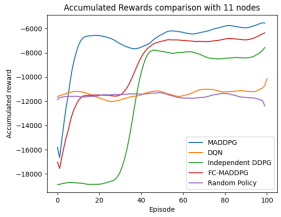

Fig. 8: Accumulated reward

overall performance. With each time step, it denotes one packet transmission from each UAV node to CS through the next-hop router. The results indicate that the centralized MADDPG achieves the best performance due to its access to global information and small network topology. Our FC-MADDPG outperforms the rest of algorithms. Notably, the DQN exhibits a performance comparable to random policy, suggesting that DQN is not applicable to complex systems with heterogeneous action space. Specifically, the proposed FC-MADDPG takes 426 transmission steps to deliver all packets to the CS. The centralized MADDPG takes 364 steps, which is around 85% of FC-MADDPG transmission steps. Independent DDPG takes 536 steps, which is around 126% of FC-MADDPG transmission steps. DQN requires 684 transmission steps, which is around 160% of FC-MADDPG transmission steps. This comparison highlights the effectiveness of the proposed FC-MADDPG algorithm in managing multi-hop UAV routing for emerging smart agriculture applications. Fig. 8 shows that the rewards for five algorithms follow same pattern as the transmission steps.

B. Topology with 30 Nodes

To demonstrate the scalability and effectiveness of the proposed FC-MADDPG algorithm in more complex and demanding network environments, we constructed a complex scenario featuring a large topology with 30 UAV nodes as illustrated in Fig. 9. Specifically, nodes 10 and 29 serve as critical junctions within the network, providing the essential links connecting the entire UAV mesh to the CS. We keep same setting as that in the 11-node topology.

Fig. 10 and Fig. 11 demonstrate the transmission duration and accumulated reward, respectively. Notably, the centralized MADDPG does not yield the best result. Its use of global information appears to be excessive in large topology, leading to inefficiency. In contrast, the proposed FC-MADDPG demonstrates superior performance by outshining all benchmark algorithms. This improvement is attributed to its ability to focus on pertinent information from neighboring

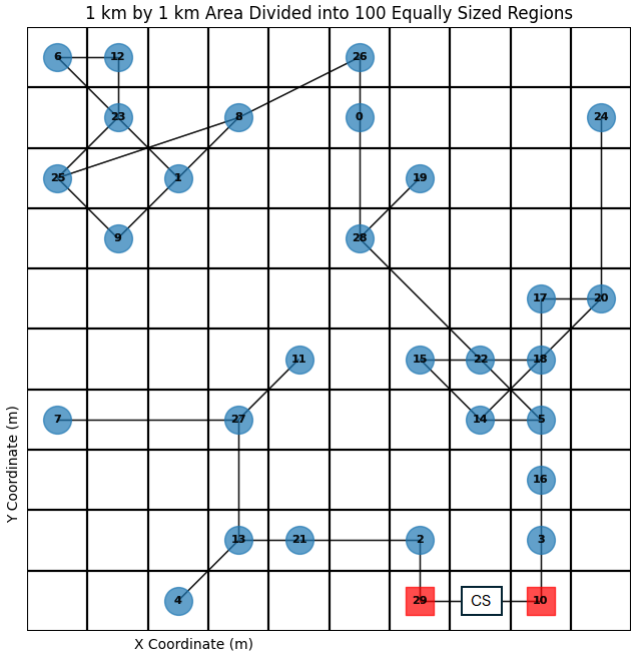

Fig. 9: Topology of UAV mesh network with 30 nodes

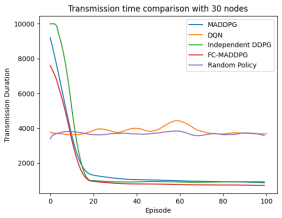

Fig. 10: Transmission time step

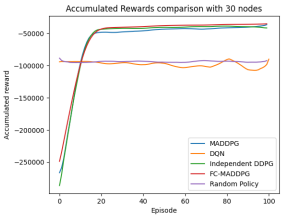

Fig. 11: Accumulated reward

nodes, reducing the unnecessary processing of non-pertinent data. Once again, the DQN exhibits performance akin to random policy. Specifically, FC-MADDPG takes 712 transmission steps to deliver all packets to the CS. The centralized MADDPG requires 874 steps, which is around 123% of FC-MADDPG transmission steps. The independent DDPG takes 917 steps, which is around 129% of FC-MADDPG transmission steps. DQN requires 3684 transmission steps, which is around 517% of FC-MADDPG transmission steps. The FC-MADDPG also outperforms benchmark algorithms in terms of the accumulated reward. Specifically, the accumulated rewards for centralized MADDPG, independent DDPG, FC-MADDPG, DQN and random policy are -36631, -41665, -35233, -90159 and -92404, respectively, which indicates that our FC-MADDPG is more reliable than the benchmarks in large agriculture networks. These findings underscore the effectiveness of the proposed FC-MADDPG algorithm in managing network resources and optimizing data transmissions in UAV-aided agriculture networks.

VII. Conclusion and Future Direction

This study introduces an innovative UAV assisted agricultural network model designed to ensure reliable and efficient data transmission from agricultural sensors to cloud servers. Numerical results demonstrate the superiority of the proposed FC-MADDPG algorithm in enhancing data transmission reliability and efficiency while minimizing communication overhead, marking a significant advancement for emerging smart agriculture practices. However, a key limitation is the lack of optimization for the UAVs' movement patterns. In real-world scenarios, controlling UAV trajectories to maximize overall network utility is crucial. Incorporating trajectory optimization would enhance the proposed two-tier smart agriculture architecture, making it more practical and effective—a promising direction for future research.

References

[1] N. Gondchawar, R. Kawitkar *et al.*, "IoT Based Smart Agriculture," *International Journal of advanced research in Computer and Communication Engineering*, vol. 5, no. 6, pp. 838–842, 2016.

[2] F. Bu and X. Wang, "A Smart Agriculture IoT System Based on Deep Reinforcement Learning," *Future Generation Computer Systems*, vol. 99, 2019.

[3] A. Rejeb, A. Abdollahi, K. Rejeb, and H. Treiblmaier, "Drones in Agriculture: A Review and Bibliometric Analysis," vol. 198, 2022.

[4] A. Din, M. Yousoof, I. B. Shah, M. Babar, F. Ali, and S. U. Baig, "A Deep Reinforcement Learning-based Multi-agent Area Coverage Control for Smart Agriculture," *Computers and Electrical Engineering*, vol. 101, 2022.

[5] J. Kim, S. Kim, C. Ju, and H. Son, "Unmanned Aerial Vehicles in Agriculture: A Review of Perspective of Platform, Control, and Applications," *IEEE Access*, 2019.

[6] S. Zhang, H. Feng, S. Han, Z. Shi, H. Xu, Y. Liu, H. Feng, C. Zhou, and J. Yue, "Monitoring of Soybean Maturity Using UAV Remote Sensing and Deep Learning," *Agriculture*, vol. 13, no. 1, p. 110, 2022.

[7] M. Raj, H. N B, S. Gupta, M. Atiquzzaman, O. Rawlley, and L. Goel, "Leveraging Precision Agriculture Techniques using UAVs and Emerging Disruptive Technologies," *Energy Nexus*, vol. 14, 2024.

[8] A. Koushik, F. Hu, and S. Kumar, "Deep Q-Learning-Based Node Positioning for Throughput-Optimal Communications in Dynamic UAV Swarm Network," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 554–566, 2019.

[9] R. Ding, J. Chen, W. Wu, J. Liu, F. Gao, and X. Shen, "Packet Routing in Dynamic Multi-Hop UAV Relay Network: A Multi-Agent Learning Approach," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 10 059–10 072, 2022.

[10] Z. Wang, H. Yao, T. Mai, Z. Xiong, X. Wu, D. Wu, and S. Guo, "Learning to Routing in UAV Swarm Network: A Multi-agent Reinforcement Learning Approach," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 6611–6624, 2023.

[11] R. Erguna, Kazim andAyoubb, P. Mercatib, and T. Rosinga, "Reinforcement Learning Based Reliability-Aware Routing in IoT Networks," *Ad Hoc Networks*, vol. 132, 2022.

[12] N. Atre, J. Sherry, W. Wang, and D. S. Berger, "Caching with Delayed Hits," in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2020, pp. 495–513.

[13] G. Xiong, S. Wang, G. Yan, and J. Li, "Reinforcement Learning for Dynamic Dimensioning of Cloud Caches: A Restless Bandit Approach," *IEEE/ACM Transactions on Networking*, 2023.

[14] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[15] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative Multi-Agent Control Using Deep Reinforcement Learning," in *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, São Paulo, Brazil*. Springer, 2017, pp. 66–83.