Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.069128

ARTICLE

Recurrent MAPPO for Joint UAV Trajectory and Traffic Offloading in Space-Air-Ground Integrated Networks

Zheyuan Jia, Fenglin Jin*, Jun Xie and Yuan He

College of Command & Control Engineering, Army Engineering University of PLA, Nanjing, 210044, China

*Corresponding Author: Fenglin Jin. Email: fljin@aeu.edu.cn

Received: 15 June 2025; Accepted: 24 July 2025; Published: 10 November 2025

ABSTRACT: This paper investigates the traffic offloading optimization challenge in Space-Air-Ground Integrated Networks (SAGIN) through a novel Recursive Multi-Agent Proximal Policy Optimization (RMAPPO) algorithm. The exponential growth of mobile devices and data traffic has substantially increased network congestion, particularly in urban areas and regions with limited terrestrial infrastructure. Our approach jointly optimizes unmanned aerial vehicle (UAV) trajectories and satellite-assisted offloading strategies to simultaneously maximize data throughput, minimize energy consumption, and maintain equitable resource distribution. The proposed RMAPPO framework incorporates recurrent neural networks (RNNs) to model temporal dependencies in UAV mobility patterns and utilizes a decentralized multi-agent reinforcement learning architecture to reduce communication overhead while improving system robustness. The proposed RMAPPO algorithm was evaluated through simulation experiments, with the results indicating that it significantly enhances the cumulative traffic offloading rate of nodes and reduces the energy consumption of UAVs.

KEYWORDS: Space-air-ground integrated networks; UAV; traffic offloading; reinforcement learning

1 Introduction

SAGIN is an emerging network architecture and a key enabler for future sixth-generation (6G) mobile communication systems, supporting full-area coverage, ultra-high speed, ultra-low latency, high reliability, and intelligent services [1]. By integrating air-based networks (e.g., UAVs and high-altitude platforms), space-based networks (e.g., satellite communication systems) and ground-based networks (e.g., terrestrial base stations), SAGIN establishes a seamless global communication infrastructure [2].

The exponential growth in network devices has led to a dramatic surge in data traffic, particularly in densely populated areas or regions with limited ground communication infrastructure, resulting in significantly increased network load pressures. As a crucial component of SAGIN, UAV networks offer rapid deployment capabilities and exceptional environmental adaptability, serving as effective supplements to ground-based communication nodes. These characteristics not only enhance user experience, but also enable reliable communication services in remote areas. Furthermore, UAV networks contribute to improved overall network performance, supporting the comprehensive coverage and ubiquitous connectivity requirements essential for 6G mobile networks [3].

UAV networks typically comprise one or more UAVs equipped with configurable sensors and communication modules to accommodate diverse mission requirements. This adaptability enables their widespread application in edge computing, traffic offloading, and relay communication systems. Within SAGIN architectures, UAV nodes assume particular significance owing to their rapid deployment capabilities and controllable mobility. Network administrators can further improve overall system performance through dynamic resource allocation and intelligent scheduling of UAV networks.

While UAV networks offer significant advantages and potential for traffic offloading, their implementation presents several challenges [4]. As a novel networking platform, UAVs exhibit high mobility and energy sensitivity, requiring careful consideration of multiple factors including flight duration, coverage range, energy consumption, communication capacity, and compatibility with existing network infrastructure. Although extensive research has addressed optimization problems in UAV networks, such as trajectory planning, deployment location, and resource allocation, relatively few studies have investigated trajectory optimization in SAGIN traffic offloading scenarios that simultaneously account for both energy consumption and data throughput in UAV networks [5].

In response to these issues, this study first formulates a traffic offloading model for SAGIN and defines the problem of offloading efficiency optimization. We then propose a Multi-Agent Proximal Policy Optimization (MAPPO) framework, enhanced with RNNs, to jointly optimize UAV flight trajectories and traffic offloading strategies. This paper makes the following key contributions:

1) We formulate a traffic offloading scenario in SAGIN and propose an efficiency optimization problem for traffic offloading. By jointly optimizing the drone flight trajectories and offloading strategies, we maximize both the user offloading rate and fairness among offloading nodes, while minimizing drone energy consumption.

2) To optimize traffic offloading efficiency, we formulate the problem as a Partially Observable Markov Decision Process (POMDP) and propose an RMAPPO-based algorithm for joint drone trajectory control and traffic offloading decisions. Unlike prior works employing DDPG frameworks, our RMAPPO approach introduces two key innovations: a recurrent architecture that captures temporal dependencies in UAV trajectories, enabling more informed decision-making; a fully decentralized training paradigm that reduces communication overhead while maintaining cooperation. By incorporating RNNs [6], our approach leverages historical observations to improve training efficiency.

3) We designed multiple simulation experiments to evaluate the proposed RMAPPO algorithm. The experimental results show that the algorithm significantly improves the cumulative traffic offloading rate of nodes and reduces the energy consumption of the drones.

2 Related Work

In recent years, SAGIN architecture has employed drones and satellites as traffic offloading nodes to mitigate the challenges posed by complex network topologies and resource management. This technology leverages the complementary advantages of heterogeneous networks, improving overall traffic transmission efficiency. However, substantial disparities in communication environments, latency, and coverage among these networks impose strict requirements on traffic offloading techniques.

The intelligentization of traffic offloading algorithms has emerged as a pivotal trend in network technology [7]. Researchers are increasingly using artificial intelligence (AI) to address the complex challenges of air-ground-space networks, optimizing key performance metrics such as throughput, latency, and energy consumption [8]. Depending on the specific offloading problem, different AI techniques are employed. Among existing studies, reinforcement learning (RL) has become the predominant intelligent approach for traffic offloading in such networks.

Recent advances in UAV-assisted wireless networks have seen significant methodological innovations. Hu et al. [9] proposed a two-stage alternating optimization algorithm for simultaneous traffic and computation offloading in UAV-assisted wireless networks. Their approach maximizes user satisfaction through joint optimization of (1) resource allocation (bandwidth and power) and (2) UAV trajectory. For a given trajectory, the resource allocation problem is transformed into a convex optimization via variable substitution, while non-convex trajectory constraints are addressed using successive convex approximation (SCA). Li et al. [10] addressed cellular network congestion caused by exponential data growth through a hierarchical intelligent framework leveraging deep federated learning. This multi-UAV system optimizes nergy-efficient flight paths, deployment locations, and dynamic resource allocation, significantly reducing energy consumption while extending operational endurance. Luo et al. [11] developed a MADRL approach for UAV relay systems serving multiple ground user pairs. Their method simultaneously ensures fairness among users, throughput maximization, and connectivity maintenance, while accounting for UAV payload and energy constraints. In the domain of UAV trajectory optimization, Chapnevis and Bulut [12] investigated a data collection scenario for Internet of Things (IoT) networks employing UAVs. The authors developed both an integer linear programming model and a heuristic algorithm to optimize UAV flight paths. Their approach incorporates the Age of Information (AoI) metric while simultaneously minimizing both mission duration and flight distance. In mobile edge computing, Zhang et al. [13] tackled the challenge of resource-constrained devices handling bandwidth-intensive applications. Their UAV-assisted solution employs a Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm to optimize computational task offloading in edge environments. For space-air-ground integrated networks, An et al. [14] proposed a Deep Deterministic Policy Gradient (DDPG)-based framework that jointly minimizes energy consumption and latency by optimizing UAV trajectories, transmission power, offloading rates, and destination selection.

Current research in UAV network optimization has primarily focused on two key areas: (1) computational offloading strategies [15] and (2) UAV deployment and trajectory optimization [16]. While these approaches effectively address computational capacity and energy efficiency, they often overlook critical requirements for effective traffic offloading. Comprehensive traffic offloading solutions must additionally consider quality-of-service (QoS) guarantees and fair resource allocation. Furthermore, in traffic offloading scenarios, UAV trajectory planning and deployment locations directly impact user experience metrics [17].

Take into account the above questions. First, research on drone networks for traffic offloading must account for their high dynamics. While the dynamic nature of drones enhances node adaptability and network robustness, it also increases network complexity. Consequently, when addressing traffic offloading, it is essential not only to optimize offloading strategies for the existing network topology but also to leverage the drones' high mobility to improve transmission efficiency and enhance user experience [18].

Second, in scenarios where multiple candidate nodes are available for traffic offloading, the node selection process becomes a critical optimization challenge. This dynamic selection mechanism must account for the rapidly changing network topology while balancing computational overhead with decision accuracy.

3 System Model

3.1 Mission Scenario

As illustrated in Fig. 1, in the SAGIN traffic offloading scenario, satellites and multiple UAVs collaboratively provide network traffic offloading services for ground users. Both UAVs and users can obtain their location information through satellite positioning. Each node can function as an offloading node to alleviate user data transmission pressure. In UAV network traffic offloading, the movement of drones can be controlled to optimize their positions, ensuring optimal offloading service provision. Users can select the most suitable offloading node based on QoS offered by each node. UAVs do not enforce centralized offloading assignments. Instead, they dynamically adjust their trajectories to improve channel conditions, thereby indirectly influencing user decisions. When adjusting their positions, UAVs must consider both the overall network communication efficiency and fairness among offloading nodes.

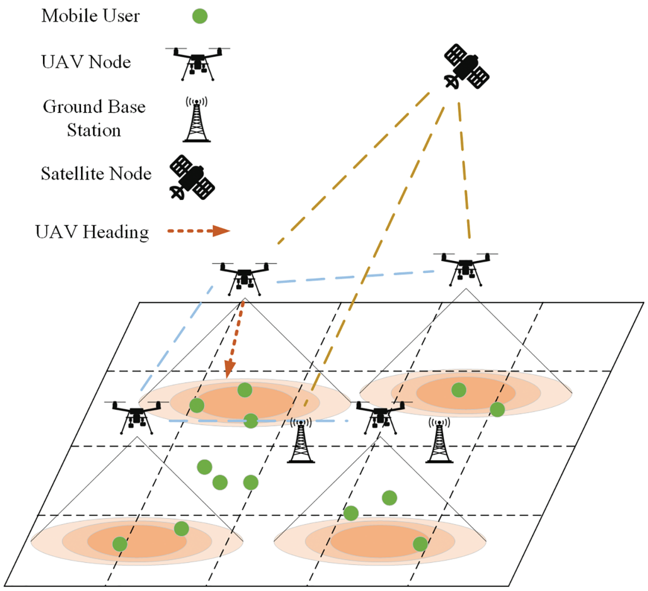


Figure 1: SAGIN architecture

Ground users outside UAV coverage areas or those experiencing poor communication efficiency with UAVs can utilize satellite nodes for traffic offloading. However, given the limited availability of satellite communication resources and their significant propagation delays, ground users preferentially employ UAV networks for traffic offloading when possible.

3.2 Communications Model

For the satellite-to-ground communication model, since the satellite altitude significantly exceeds ground users' movement range, ground nodes can be treated as relatively stationary when analyzing satellite-ground communications. This paper employs a Weibull distribution-based channel model [19] to characterize the satellite-ground communication link, where the channel gain is expressed as:

$$L_S = \frac{G_S G_O \lambda^2}{(4\pi d)^2} 10^{-\frac{W}{10}} \tag{1}$$

where G_S and G_O are the antenna gains of the satellite node and the UAV node, respectively, λ denotes the wavelength of the electromagnetic wave, d is the distance between the communicating nodes, and W indicates the communication bandwidth. The communication rate between the satellite and ground user can then be expressed as [20]:

$$R_{GS} = W\log_2\left(1 + \frac{P_t \cdot |L_s|^2}{\sigma^2}\right) \tag{2}$$

where P_t represents the device transmission power, σ^2 denotes the Gaussian noise power, and W indicates the available network bandwidth.

The UAV-user communication system employs both uplink and downlink channels. To minimize interference, distinct frequency bands are allocated for uplink and downlink transmission, supplemented by Orthogonal Frequency Division Multiple Access (OFDMA) technology. Given that UAV-ground user communication primarily occurs through line-of-sight (LoS) links while being susceptible to terrestrial obstructions, we adopt a probabilistic LoS channel model [21]. The *Los* probability between user m and UAV k at time t is given by:

$$P_{m,k}^{Los}(t) = \frac{1}{1 + a \cdot e^{-b(\theta_{m,k} - a)}} \tag{3}$$

where a and b are environment-dependent constants, $\theta_{m,k}$ represents the elevation angle between user m and UAV k, and $PL_{m,k}$ denotes the average path loss between ground user m and UAV k [22], expressed as:

$$PL_{m,k}(t) = P_{m,k}^{Los} \times PL_{m,k}^{Los}(t) + (1 - P_{m,k}^{Los}(t)) \times PL_{m,k}^{NLos}(t) \tag{4}$$

Path loss for line-of-sight communication $PL_{m,k}^{Los}(t)$ and average path loss for non-line-of-sight communication $PL_{m,k}^{NLos}(t)$, respectively:

$$\begin{cases} PL_{m,k}^{Los}(t) = 20\lg\left(\dfrac{4\pi f d_{m,k}(t)}{c}\right) + \eta_{Los} \\ PL_{m,k}^{NLos}(t) = 20\lg\left(\dfrac{4\pi f d_{m,k}(t)}{c}\right) + \eta_{NLos} \end{cases} \tag{5}$$

where f denotes the carrier frequency, $d_{m,k}(t)$ represents the Euclidean distance between user m and UAV k at time t, c is the speed of light, η_{LoS} indicates the additional path loss for line-of-sight propagation, and η_{NLoS} denotes the additional path loss for non-line-of-sight propagation. The resulting communication rate between UAV k and ground user m is given by:

$$R_{GU} = W\log_2\left(1 + \frac{P_t \cdot 10^{\frac{PL_{m,k}(t)}{10}}}{\sigma^2}\right) \tag{6}$$

3.3 Model Architecture

In a mission area of $L \times L$, there are K UAVs performing traffic offloading tasks, denoted by the set $\mathbf{K} = \{1, 2, \ldots, K\}$, and these UAVs provide traffic offloading services to M ground users, denoted by the set $\mathbf{M} = \{1, 2, \ldots, M\}$. The total mission duration is divided into T uniform time slots, with each user restricted to selecting only one UAV for offloading services per time slot. Each UAV k's position $p_k^{UAV}(t)$ is specified by coordinates $(x_k(t), y_k(t), h_k)$, where h_k represents the constant flight altitude. After each time slot t, the UAV will travel to the next location to ensure the best offloading service for the user, between each time slot, the movement of the UAV is simplified as a straight line movement in the horizontal direction, and the maximum speed of the UAV is set to $V_{\max}$, and between two time slots, the UAV's moving speed will not exceed $V_{\max}$, therefore, the UAV's position coordinates are updated by:

$$\begin{cases} x_m(t+1) = x_m(t) + v_m \times t \times \cos\theta_u \\ y_m(t+1) = y_m(t) + v_m \times t \times sin\theta_u \end{cases} \tag{7}$$

Let θ_u denote the flight angle of UAV node m, and v_m represents its flight speed. The ground user node follows a random walk model with a constant velocity v_{user}. The maximum speed of the UAV ($V_{\max}$) significantly exceeds the user's movement speed, i.e., $V_{\max} \gg v_{\text{user}}$.

Each ground user has distinct communication demands and traffic offloading requirements. Let w_k denote the traffic offloading size required by ground user k per unit time. In practical scenarios, UAVs can be deployed at predetermined locations based on a preconfigured network topology. However, for this study, we initialize the system with randomly generated UAV deployment locations, corresponding to a random network topology configuration.

3.4 Problem Description

During mission execution, the UAV dynamically adjusts its trajectory within a predefined operational area according to the real-time positions of user nodes, enabling it to identify and navigate toward optimal data offloading points. This study optimizes the UAV's movement direction and flight path to enhance traffic offloading efficiency while minimizing energy consumption and maintaining fair resource allocation.

At time t, ground user i selects the optimal service provider UAV j from the available set of offloading-capable UAV nodes based on a service quality metric. The user then offloads its traffic to UAV j for processing. In practice, this optimal selection can be derived through channel quality detection and related techniques. For our experimental setup, we determine the optimal UAV based on: (1) the user's position (x_i, y_i), (2) the UAV node's coordinates (x_j, y_j), and (3) the air-to-ground channel model, which yields the achievable communication rate R_{ij} between user i and UAV j. The user selects the UAV node offering the highest R_{ij} as its optimal service provider.

Let r denote the effective coverage radius of a UAV. User nodes outside this coverage area (i.e., where $d_{ij} > r$ for all UAVs j) achieve better offloading performance by utilizing satellite nodes instead of UAV-based service.

To prevent user distribution from becoming overly centralized—which could lead to multiple users simultaneously selecting the same subset of UAVs as offloading nodes and causing UAV congestion—we employ the Jain fairness index [23] to define a fairness metric for each UAV. This approach ensures balanced load distribution across all available UAV nodes.

$$F_j = \frac{\left(\sum_{j=1}^{K} W_j\right)^2}{n \sum_{j=1}^{K} W_j^2} \tag{8}$$

Let W_j denote the total traffic offloading demand per unit time for all ground users selecting UAV j, while F_j takes values in the interval $[0, 1]$.

In SAGIN traffic offloading, it is essential to enhance the overall network offloading efficiency while maximizing the user throughput rate. The offloading rate for each user node is derived from the communication model. Throughout the task cycle, it is necessary to maximize optimization target Z, which is expressed as follows:

$$Z = \alpha \cdot T + (1 - \alpha) \cdot F \tag{9}$$

Let $\alpha \in [0, 1]$ be a weighting parameter that balances the trade-off between throughput rate and fairness, where $T = \sum_{i=0}^{M} v_i$ represents the total throughput rate, and F denotes the fairness metric across UAVs. To ensure problem feasibility, we impose the following constraints, yielding the final optimization formulation:

$$Z = \alpha \cdot \sum_{i=1}^{M} v_i + (1-\alpha) \cdot \frac{\left(\sum_{j=1}^{K} W_j\right)^2}{n \sum_{j=1}^{K} W_j^2} \quad \text{s.t.} \begin{cases} C1 : x_{\min} \le q_{j,x}(t) \le x_{\max}, & \forall j, t \\ C2 : y_{\min} \le q_{j,y}(t) \le y_{\max}, & \forall j, t \\ C3 : \|q_j(t) - q_k(t)\| \ge d_s, & \forall j \ne k, t \\ C4 : \|v_j(t)\| \le V_{\max}, & \forall j, t \\ C5 : \|p_i - q_j(t)\| \le r, & \forall i, j, t \\ C6 : v_i \ge 0, & \forall i, t \end{cases} \tag{10}$$

Let p_i denote the location of user node i, and $q_j(t)$ represents the location of UAV node j at time t. The speed of UAV j at time t is denoted by $v_j(t)$, while r represents the effective service range of the UAV node. The throughput rate of user i is given by R_i. The optimization problem is subject to the following constraints:

- C1 and C2 restrict the UAV's movement range.
- C3 ensures collision avoidance between UAVs by maintaining a safe distance.
- C4 limits the UAV's speed to the maximum allowable speed $V_{\max}$.
- C5 guarantees that the selected offloading UAV must be within its service range r.
- C6 requires the user throughput rate to be strictly positive ($R_i > 0$).

4 RMAPPO Method

4.1 Problem Analysis

To ensure all ground users in the mission area receive high-quality service and enhance the overall data traffic offloading efficiency in SAGIN, the system must not only direct ground users to select optimal UAV or satellite nodes for data offloading based on channel conditions and QoS requirements, but also enable UAV nodes to dynamically adjust their flight trajectories through mobility optimization for deployment at optimal service locations.

There are m users and k UAVs in the mission area, and each UAV node has a certain coverage radius when offloading traffic r, and it needs to cover as many ground users as possible in the mission area in order to provide traffic offloading service for them. When solely maximizing user coverage without considering constraints and optimization objectives, the problem can be reduced to a circular coverage problem, which has been proven to be NP-hard [24]. However, introducing multiple constraints and multi-objective optimization requirements in practical systems substantially increases the problem's complexity. Consequently, since obtaining an optimal solution through exact algorithms in polynomial time remains infeasible, research efforts should prioritize developing efficient approximation or heuristic algorithms to achieve high-quality suboptimal solutions within reasonable timeframes.

Moreover, the target solution for this problem includes not only the UAV's static deployment locations but also its optimized dynamic trajectory during mission execution, ensuring real-time adaptation to dynamic user distribution for service optimization.

4.2 Prescription

The movement of UAVs is temporally discrete, allowing only one action per time slot. However, its action selection is continuous, and the objective function is a time-varying non-convex optimization problem, which traditional methods struggle to solve. To address this, we employ reinforcement learning (RL) to derive an optimal strategy. Given the cooperative nature of drone networks, a multi-agent reinforcement learning (MARL) approach is adopted. Since the UAV's action space is continuous, policy-based RL methods are particularly suitable. While Proximal Policy Optimization (PPO) is typically applied in single-agent settings, its core principles can be extended to multi-agent systems.

The basic MAPPO training framework is shown in Fig. 2, and the Clip clipping method used in this paper updates the policy gradient, and the network parameters are updated with the formula [25]:

$$L^{CLIP}(\theta) = E_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1- \epsilon, 1+ \epsilon)\hat{A}_t)] \tag{11}$$

Since PPO is designed for single-agent reinforcement learning, it requires centralized training to access global network information. However, in our problem, UAV nodes cannot share all observations, making single-agent PPO unsuitable. To enable effective cooperation among nodes, we adopt a multi-agent architecture.

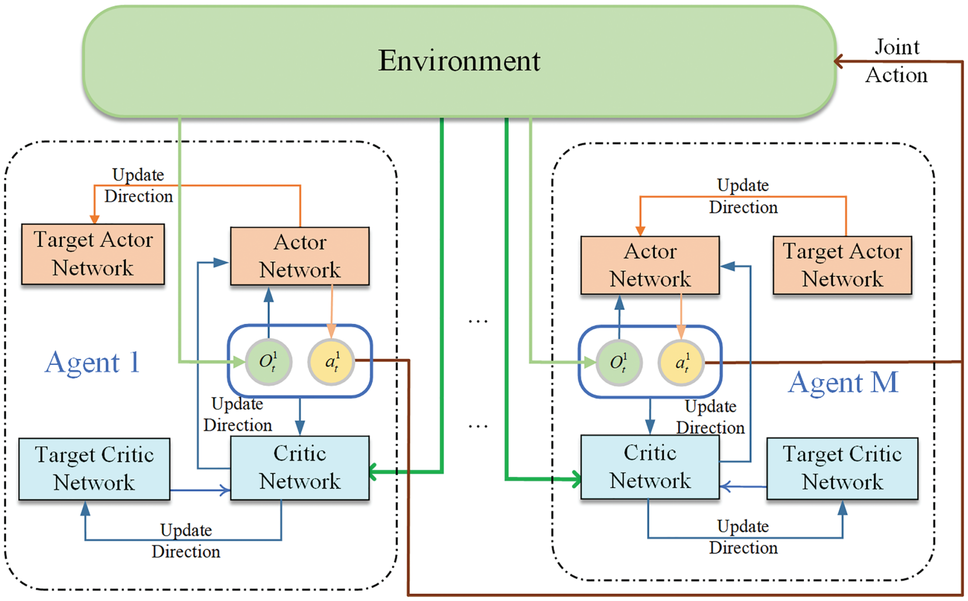


Figure 2: Training framework diagram

Traditional MAPPO employs centralized training with decentralized execution (CTDE), where agents share a centralized Critic network that estimates the global value function using the global state and all agents' actions, while each agent's Actor network generates actions based on local observations. In contrast, our algorithm uses a fully distributed training approach: each agent maintains its own Critic and Actor networks, updating parameters based solely on local observations. This design reduces reliance on global information, enhancing user privacy and scalability for larger networks.

4.3 RMAPPO Model

In path planning problems, trajectories exhibit temporal continuity, where current actions depend on past states and actions. To enable agents to capture these temporal dependencies, we integrate a recurrent neural network (RNN) with reinforcement learning, endowing them with memory capabilities. The architecture of our proposed RMAPPO network is illustrated in Fig. 3.

Each agent's state is initially processed by a Gated Recurrent Unit (GRU) layer with 64 hidden units, which maintains historical information through its hidden state mechanism. The processed features are then fed into the Actor network, comprising two fully-connected hidden layers (64 and 32 units, respectively), with actions ultimately generated by an output layer using $Tanh()$ activation. Similarly, the Critic network also incorporates GRU processing.

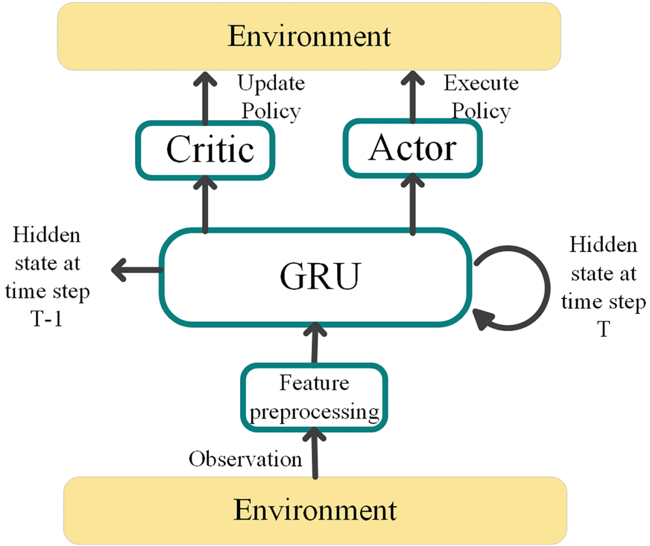


Figure 3: Optimize actor-critic algorithms

4.3.1 Observation Space

Each UAV node has its own observation space, which includes UAV node ID, node position, current user position, node resource consumption. Agent will take appropriate actions according to the local observation of the UAV to guide the UAV to take the best flight trajectory.

4.3.2 Action Space

The action of the agent includes the flight direction of the UAV $\theta \in [-\pi, \pi]$ and the flight distance $d \in [0, d_{\max}]$, and the joint action of all the UAVs at the time of t is represented by $A_t = [\theta_t^1, d_t^1, \ldots, \theta_t^K, l_t^K]$, and the dimension of the action space is $[2K]$.

4.3.3 Reward Function

The design of the reward function must consider both the optimization objectives and problem constraints, while carefully scaling different reward components to prevent dominance by large numerical values. For the UAV trajectory optimization problem, we define the following reward and penalty structure:

1. **Throughput Reward:** The reward is based on the throughput rate achieved by the ground user upon selecting an offloading node, formulated as:

$$R_1 = \sum_{i=0}^{M} v_i \tag{12}$$

2. **Fairness Reward:** This reward is determined by the fairness of resource allocation across drone nodes when performing traffic offloading, defined as:

$$R_2 = \frac{\left(\sum_{j=1}^{K} W_j\right)^2}{n \sum_{j=1}^{K} W_j^2} \tag{13}$$

3. **Out-of-Bounds Penalty:** This penalty constrains the drone's service range. If drone node k exceeds its designated service area, it incurs an out-of-bounds penalty ν, expressed as:

$$P_k^{\text{out}} = \begin{cases} -\nu & otherwise \\ 0 & \mathbf{p}_{\mathrm{k}}^{UAV} \in \mathrm{P}_{\text{safe}} \end{cases} \tag{14}$$

   where ν denotes the out-of-bounds penalty factor for UAVs and $\mathcal{P}_{\text{safe}}$ represents the safe communication region for UAV nodes.

$$P_{ij} = \begin{cases} -\alpha\left(d^{\text{safe}} - \|\mathbf{p}_i^{UAV} - \mathbf{p}_j^{UAV}\|\right) & \text{if } \|\mathbf{p}_i^{UAV} - \mathbf{p}_j^{UAV}\| < d_{\text{safe}} \text{ and i} \neq \text{j} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

   where α denotes the penalty coefficient and d^{safe} denotes the safe distance between the UAVs. When the distance $|\mathbf{p}_i^{UAV} - \mathbf{p}_j^{UAV}|$ between UAV i and UAV j is less than d^{safe} and $i \neq j$, a penalty P_{ij} is applied.

4. **UAV Energy Consumption Penalty:** The energy consumption of the UAV directly affects its endurance and serves as a crucial performance metric. To minimize energy waste, we impose a movement penalty proportional to the displacement distance:

$$P_{energy}^{\mathrm{k}} = -\mu \cdot \Delta d_{\mathrm{k}} \tag{16}$$

The fairness-throughput tradeoff parameter α was set to 0.7 through empirical testing, prioritizing throughput while maintaining reasonable fairness. The penalty weights ν (out-of-bounds penalty) and μ (energy consumption penalty) were tuned via grid search, with final values set to 100 and 0.001, respectively.

4.4 Algorithm Workflow

Algorithm 1 employs a fully distributed training paradigm, wherein each agent maintains its own Critic network and independently updates the network parameters based on local observations. However, agents can share partial observations with each other through communication channels. This design facilitates straightforward implementation for interconnected nodes. The training phase of the algorithm proceeds as follows:

Algorithm 1: RMAPPO training.

```
Input: Initialized environment, Actor/Critic networks, Training hyperparameters
Output: Trained model and training data
Initialize environment, network parameters, and hyperparameters;
while training not terminated do
    for each agent do
        Select action from Actor network based on local observations;
    end
    Execute joint action in environment;
    Observe reward and next state;
    Store transition (state, action, reward, next state) in replay buffer;
    Sample minibatch from replay buffer;
    for each agent do
        Compute loss using objective function;
        Update network parameters;
    end
```

(Continued)

Algorithm 1 (continued)

```
    if termination condition met then
        Save final model and training data;
    end
end
```

5 Simulation Experiments

We evaluate the proposed algorithm through comprehensive simulation experiments implemented using the PyTorch framework. Table 1 shows the experimental parameters. The experimental scenario consists of a 5000 m × 5000 m flat terrain area with UAVs operating at a maximum altitude of 800 m. The UAVs are initially deployed according to a predefined network topology, while ground user nodes are randomly distributed throughout the area. To demonstrate the algorithm's effectiveness, we conduct comparative evaluations against two benchmark methods: (1) Independent Proximal Policy Optimization (IPPO) proposed in [11] and (2) the K-means clustering-based deployment algorithm.

Table 1: Experimental parameters

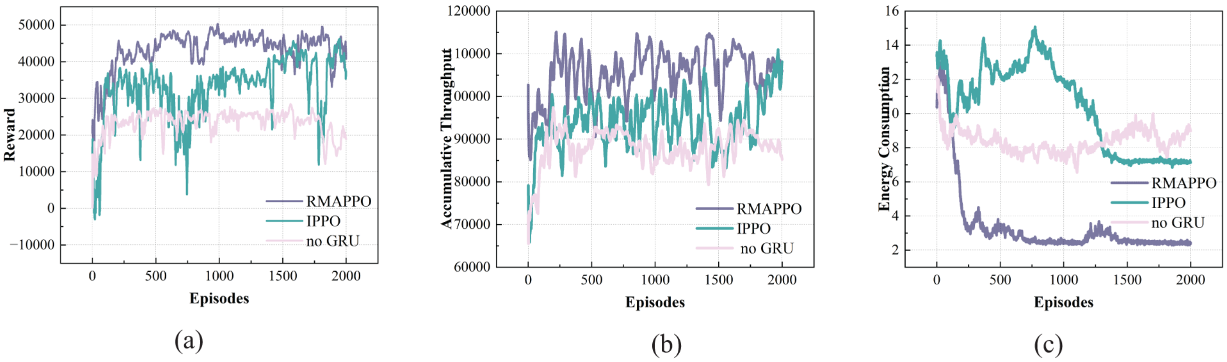
Parameter	Significance	(Be) Worth
S	Number of satellites	1
$V_{\max}$	Maximum drone speed	30 m/s
v_{user}	Terrestrial user rate	2 m/s
r	UAV communications radius	1000 m
d_{safe}	Safe distance between drones	300 m
W	Bandwidths	2 MHz
d_{sate}	Delay in satellite communications	6 ms
n_0	Noise power spectral density	−110 dBm/Hz
η_{Los}	Line-of-sight loss	1 dB
η_{NLos}	Non-line-of-sight loss	20 dB
p^{UAV}	UAV launch power	5 W
α	Fairness-throughput weighting parameter	0.7
T	Total training steps	2000
lr	Learning rate	0.0003
$clip$	Trimming factor	0.2

The experiment is conducted in a rectangular task area measuring 5000 m × 5000 m, with 4 UAV nodes and 30 ground users. The offloading task duration is set to 500 s. The UAVs operate at altitudes ranging from 500 to 800 m. During the initialization phase, both the positions of ground users and their required offloading rates are randomly generated, with the latter constrained between 1 and 20 MB/s. Two approaches are compared under identical experimental conditions.

The first approach implements the IPPO algorithm from [11], which addresses trajectory optimization for multi-UAV transmission services to ground user pairs. Notably, this method does not employ RNN and is also a MARL algorithm. The second approach is a K-means-based heuristic algorithm that: (1) clusters users by location into K groups (where K equals the number of UAVs), (2) directs each UAV toward its cluster centroid, and (3) iteratively updates cluster centroids and UAV trajectories as user positions change. Additionally, we conducted ablation studies to isolate the impact of the GRU layer within the RMAPPO architecture.

5.1 Convergence Analysis

As illustrated in Fig. 4, both methods exhibit steadily increasing cumulative rewards with growing iterations, ultimately achieving stable convergence. Notably, the proposed RMAPPO algorithm demonstrates superior performance to IPPO across three key metrics: cumulative reward, throughput, and energy consumption. By integrating recurrent neural networks with reinforcement learning within a distributed multi-agent framework, RMAPPO enhances agents' perception of local environmental states, thereby facilitating faster policy optimization. Experimental results validate the convergence of RMAPPO and demonstrate its significantly faster convergence rate compared to IPPO under identical iteration counts.

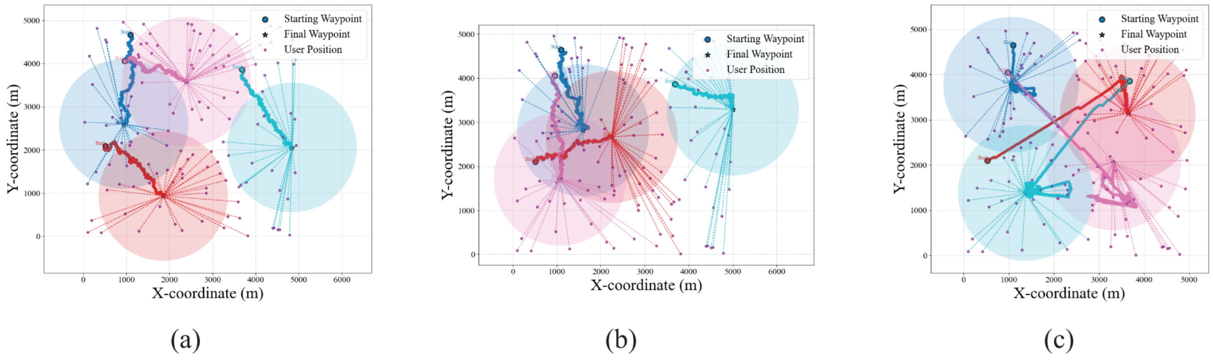


(a) (b) (c)

Figure 4: (**a**) Cumulative reward; (**b**) Cumulative throughput reward; (**c**) Cumulative energy penalty

5.2 UAV Trajectories

As illustrated in Fig. 5, the scenario comprises four UAVs and 130 ground users. Fig. 5a presents the trajectory generated by the RMAPPO method, demonstrating more comprehensive user coverage compared to other approaches. Fig. 5b displays the trajectory produced by IPPO, revealing that a significant number of users still rely on satellite nodes for traffic offloading during UAV movement. Fig. 5c exhibits the trajectory generated by the K-means algorithm, where substantial offsets between consecutive clustering centers—caused by ground node mobility—result in oscillatory UAV movement patterns.



(a) (b) (c)

Figure 5: (**a**) RMAPPO; (**b**) IPPO; (**c**) K-means

5.3 Performance Analysis

To further validate the effectiveness of RMAPPO, we evaluated all three methods under varying numbers and spatial distributions of ground users.

As shown in Fig. 6, the RMAPPO algorithm outperforms the other two methods in both fairness metrics and average throughput rate. In contrast, the K-means algorithm fails to account for user mobility, heterogeneous offloading demands, varying satellite-assisted offloading capabilities, and fairness among drone nodes, making it unsuitable for achieving the optimization objective. Furthermore, UAV states exhibit strong temporal dependencies in their continuous position, direction, and velocity, which enables RMAPPO to learn more effective trajectory optimization strategies. However, since RMAPPO incorporates an additional GRU layer compared to IPPO, it exhibits a lower computational efficiency than IPPO. The RMAPPO algorithm exhibits a 10.65% longer execution time compared to IPPO when implemented on identical hardware configurations.

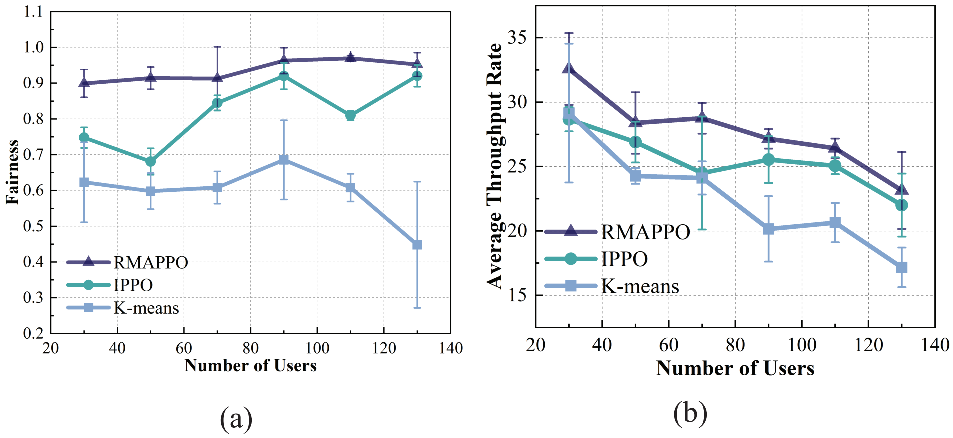


Figure 6: (**a**) Fairness comparison; (**b**) Throughput comparison

6 Conclusion

This paper proposed the RMAPPO algorithm to optimize UAV trajectory and traffic offloading in SAGIN, achieving higher throughput, lower energy consumption, and fairer resource allocation. Key innovations include integrating RNNs for temporal dependency modeling and a decentralized training framework for reduced overhead. Simulations confirmed RMAPPO's superiority over IPPO and K-means in performance metrics. The work provides a practical solution for dynamic SAGIN environments, paving the way for efficient 6G network offloading. Future research may explore scalability and diverse scenario designs.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by Army Engineering University of PLA.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Zheyuan Jia and Fenglin Jin; analysis and interpretation of results: Jun Xie; draft manuscript preparation: Yuan He. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Fenglin Jin, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Cui H, Zhang J, Geng Y, Xiao Z, Sun T, Zhang N, et al. Space-air-ground integrated network (SAGIN) for 6G: requirements, architecture and challenges. China Commun. 2022;19(2):90–108. doi:10.23919/jcc.2022.02.008.
2. Liu J, Shi Y, Fadlullah ZM, Kato N. Space-Air-ground integrated network: a survey. IEEE Commun Surv Tut. 2018;20(4):2714–41.
3. Mahboob S, Liu L. Revolutionizing future connectivity: a contemporary survey on ai-empowered satellite-based non-terrestrial networks in 6G. IEEE Commun Surv Tut. 2024;26(2):1279–321. doi:10.1109/comst.2023.3347145.
4. Lyu J, Zeng Y, Zhang R. UAV-aided offloading for cellular hotspot. IEEE Trans Wireless Commun. 2018;17(6):3988–4001. doi:10.1109/twc.2018.2818734.
5. Owaid SA, Miry AH, Salman TM. A survey on UAV-assistedwireless communications: challenges, technologies, and application. In: 2024 11th International Conference on Electrical and Electronics Engineering (ICEEE). Marmaris, Turkiye; 2024. p. 333–40.
6. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv:1409.2329. 2014.
7. Qin Y, Yang Y, Tang F, Yao X, Zhao M, Kato N. Differentiated federated reinforcement learning based traffic offloading on space-air-ground integrated networks. IEEE Trans Mobile Comput. 2024;23(12):11000–13. doi:10.1109/tmc.2024.3389011.
8. Huang C, Chen P. Mobile traffic offloading with forecasting using deep reinforcement learning. arXiv:1911.07452. 2019.
9. Hu X, Zhuang X, Feng G, Lv H, Wang H, Lin J. Joint optimization of traffic and computation offloading in UAV-assistedwireless networks. In: 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS). Chengdu, China; 2018. p. 475–80.
10. Li F, Zhang K, Wang J, Li Y, Xu F, Wang Y, et al. Multi-UAV hierarchical intelligent traffic offloading network optimization based on deep federated learning. IEEE Internet of Things J. 2024;11(12):21312–24. doi:10.1109/jiot.2024.3363188.
11. Luo X, Xie J, Xiong L, Wang Z, Liu Y. UAV-assisted fair communications for multi-pair users: a multi-agent deep reinforcement learning method. Comput Netw. 2024;242(3):110277. doi:10.1016/j.comnet.2024.110277.
12. Chapnevis A, Bulut E. AoI-optimal cellular-connected UAV trajectory planning for IoT data collection. In: 2023 IEEE 48th Conference on Local Computer Networks (LCN). Daytona Beach, FL, USA; 2023. p. 1–6.
13. Zhang X, Wang J, Wang B, Jiang F. Offloading strategy for UAV-assisted mobile edge computing based on reinforcement learning. In: 2022 IEEE/CIC International Conference on Communications in China (ICCC). Sanshui, China; 2022. p. 702–7.
14. An P, Du L, Chen Y. Learning-based task offloading and UAV trajectory optimization in SAGIN. In: 2024 33rd Wireless and Optical Communications Conference (WOCC). Hsinchu, Taiwan; 2024. p. 12–6.
15. Gao Y, Ye Z, Yu H. Cost-efficient computation offloading in SAGIN: a deep reinforcement learning and perception-aided approach. IEEE J Selected Areas Commun. 2024;42(12):3462–76. doi:10.1109/jsac.2024.3459073.
16. Lakew DS, Masood A, Cho S. 3D UAV placement and trajectory optimization in UAV assisted wireless networks. In: 2020 International Conference on Information Networking (ICOIN). Barcelona, Spain; 2020. p. 80–2.
17. Chapnevis A, Güvenç I, Njilla L, Bulut E. Collaborative trajectory optimization for outage-aware cellular-enabled UAVs. In: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021). Helsinki, Finland; 2021. p. 1–6.
18. Fan B, Jiang L, Chen Y, Zhang Y, Wu Y. UAV Assisted traffic offloading in air ground integrated networks with mixed user traffic. IEEE Trans Intell Transp Syst. 2022;23(8):12601–11. doi:10.1109/tits.2021.3115462.
19. Kanellopoulos SA, Kourogiorgas CI, Panagopoulos AD, Livieratos SN, Chatzarakis GE. Channel model for satellite communication links above 10GHz based on weibull distribution. IEEE Commun Lett. 2014;18(4):568–71. doi:10.1109/lcomm.2014.013114.131950.
20. Tan J, Tang F, Zhao M, Kato N. Performance analysis of space-air-ground integrated network (SAGIN): UAV altitude and position angle. In: 2023 IEEE/CIC International Conference on Communications in China (ICCC). Dalian, China; 2023. p. 1–6.

21. Al-Hourani A, Kandeepan S, Lardner S. Optimal LAP altitude for maximum coverage. IEEE Wireless Commun Lett. 2014;3(6):569–72. doi:10.1109/lwc.2014.2342736.
22. Seid AM, Boateng GO, Mareri B, Sun G, Jiang W. Multi-agent DRL for Task offloading and resource allocation in multi-UAV enabled IoT edge network. IEEE Trans Netw Service Manage. 2021;18(4):4531–47. doi:10.1109/tnsm.2021.3096673.
23. Sediq AB, Gohary RH, Schoenen R, Yanikomeroglu H. Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation. IEEE Trans Wireless Commun. 2013;12(7):3496–509. doi:10.1109/twc.2013.061413.121703.
24. Acharyya R, Basappa M, Das GK. Unit disk cover problem in 2D. In: Murgante B, Misra S, Carlini M, Torre CM, Nguyen HQ, Taniar D, et al., editors. Computational Science and Its Applications-ICCSA 2013. Berlin/Heidelberg, Germany: Springer; 2013. p. 73–85 doi:10.1007/978-3-642-39643-4_6.
25. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347. 2017.