Deep Reinforcement Learning in Multiple UAV-and-RIS Assisted Cognitive Radio System

Siyu QIAN, Linzi HU, Yuwen QIAN, Long SHI

School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Xiao Ling Wei 200, 210094 Nanjing, China

siyvqian@gmail.com, {linzihu, admon}@njust.edu.cn, slong1007@gmail.com

Submitted September 8, 2025 / Accepted January 11, 2026 / Online first February 13, 2026

Abstract. *Cognitive radio (CR) systems enable dynamic spectrum sharing but face substantial challenges in optimizing the rates of secondary users (SUs), particularly in scenarios where multiple SUs compete for the limited resources of the primary user (PU). To address this issue, we propose a multi-unmanned aerial vehicle (UAV)-assisted CR system in which reconfigurable intelligent surfaces (RISs) are mounted on UAVs to enhance spectral efficiency. Furthermore, we cast this challenge as a multi-agent Markov decision process (MDP), providing a formal framework to explore the critical trade-off between independent decision-making and centralized coordination. Consequently, we leverage established deep reinforcement learning algorithms to probe this trade-off. To provide a comprehensive performance evaluation, we adopt a Multi-Agent Proximal Policy Optimization (MAPPO) algorithm to maximize the sum rate of the proposed system. Numerical results demonstrate that the developed UAV-RIS-assisted system adopting the MAPPO algorithm can achieve a faster convergent speed and higher sum rate when compared with that adopting Independent Proximal Policy Optimization (IPPO) and MAPPO with a clipping scheme. In addition, for the MAPPO with a clipping scheme, a selected moderate clipping parameter can effectively balance the trade-off between training stability and learning efficiency.*

Keywords

Cognitive radio, UAV, RIS, reinforcement learning, multi-agent, MAPPO, spectrum sharing

1. Introduction

The rapid proliferation of wireless communication services has led to an unprecedented demand for spectrum resources, resulting in severe spectrum scarcity that constrains the advancement of next-generation communication systems. Cognitive radio (CR) technology emerged as a promising paradigm to address this challenge by enabling secondary users (SUs) to dynamically access temporarily unutilized licensed spectrum without causing interference to primary users (PUs) [1]. With this dynamic spectrum sharing mechanism, spectrum efficiency can be significantly improved, particularly in resource-constrained environments. As a result, CR has been considered highly beneficial in mission-critical applications such as military and emergency communications, where reliability and security requirements are stringent. Moreover, recent advances in signal processing and cooperative spectrum sensing have further enhanced the adaptability of CR systems to heterogeneous traffic patterns and varying interference levels.

Despite these advantages, the practical deployment of CR systems in large-scale or complex scenarios remains challenging. One major limitation stems from the reliance of conventional CR systems on static spectrum sensing and allocation strategies [2], which severely restricts adaptability to rapidly changing spectrum demands and environmental conditions [3]. Such static approaches often result in inefficient spectrum utilization, particularly under highly dynamic conditions. Furthermore, the high sensing latency and the limited spatial diversity of static sensors aggravate detection errors in mobile environments. Another critical issue arises from the fact that traditional CR architectures are tailored for single-user or small-scale settings. These designs lack robust coordination and optimization mechanisms required for multi-user environments [4]. Therefore, efficiently managing spectrum sharing and controlling interference in resource-constrained environments remains a critical challenge for contemporary CR systems [5].

The integration of unmanned aerial vehicles (UAVs) into CR systems has recently attracted significant research interest owing to their capability to enhance wireless communication performance [6], [7]. UAVs can be deployed as aerial base stations or relays to overcome obstacles and extend coverage to otherwise inaccessible regions, thereby improving spectrum utilization in CR networks [8], [9]. In addition, the rapid deployment and mobility make UAVs well-suited for on-demand communication scenarios, enabling quick responses in dynamic environments [10]. Similarly, the application of reconfigurable intelligent surfaces (RISs) in CR

DOI: 10.13164/re.2026.0117

systems has emerged as a promising approach to enhance spectrum access and communication reliability [11], [12]. By dynamically adjusting the phase shifts of the elements, RISs can reconfigure wireless propagation conditions, thus facilitating more efficient spectrum utilization [13]. Moreover, RISs can be strategically positioned to suppress interference at specific locations and improve communication quality for both PU and SU, while incurring low hardware costs and power consumption [14].

Motivated by these inherent advantages, recent works have proposed the joint integration of UAV and RIS in CR systems, leading to hybrid UAV-RIS-assisted architectures [15], [16]. Such architectures have provided new opportunities for three-dimensional coverage and interference management beyond what ground-only or stationary deployments can achieve [18], [17]. For example, in a downlink multi-antenna base station (BS) and single-antenna user setting, equipping a UAV with an RIS and jointly optimizing BS beamforming and RIS phase shifts can improve system energy efficiency by nearly 50% compared to conventional amplify-and-forward relaying schemes [19]. Despite this potential, realizing such improvements in practice remains challenging. In particular, the dynamic nature of UAV channels, imperfect channel feedback, and the requirement for joint optimization of UAV trajectories, RIS phase configurations, and transmit power allocation across multiple agents demand sophisticated control strategies. However, research addressing these challenges remains scarce. Furthermore, even with RIS assistance, outage probability remains considerably high under low transmit power conditions, leading to substantial resource consumption [20].

Reinforcement learning (RL) has recently been introduced into CR systems as a powerful tool for performance optimization [21], [22]. For example, a cooperative spectrum sensing algorithm based on deep reinforcement learning (DRL) was proposed in [23] to enhance the quality of spectrum sensing accuracy. To further improve reliability, a partially cooperative multi-agent RL (PCMARL) spectrum sensing optimization algorithm was developed in [24] to address the limitations of incomplete statistical information regarding PU. In addition, RL has been applied to power control within CR systems, where SUs are capable of learning energy harvesting and transmit power adaptation strategies under dynamic environmental conditions. Specifically, a deep Q-network (DQN) is employed to regulate SU transmit power [25]. Nevertheless, the application of RL in CR systems still faces the following challenge: the learning process is inherently complex and often constrained by the limited availability of training data, which hampers robust policy learning in practical deployments.

In this paper, we propose a UAV- and RIS-assisted CR system based on DRL to maximize the sum rate of all SUs while satisfying the minimum rate requirement of the PU. The system model assumes that the instantaneous spectrum occupancy state of the PU is unavailable, and only outdated state information can be inferred from the automatic repeat request (ARQ) feedback at the PU during the previous transmission round. The proposed work lies in the area of radio communication systems and signal processing, focusing on UAV- and RIS-assisted cognitive radio with learning-based resource optimization. We study spectrum sharing and interference-constrained resource allocation in such UAV/RIS-enabled networks within a multi-agent reinforcement learning framework. The contributions of this paper are listed as follows:

- We propose a CR system assisted by UAV-mounted RISs and further extend the system from a single UAV–RIS pair to a multi-UAV and multi-RIS scenario. This three-dimensional architecture differs from most existing works that consider either a single UAV or a single RIS, and it is more suitable for large-scale multi-cell wireless networks.
- We adopt a practical assumption that the instantaneous spectrum occupancy state of the PU cannot be perfectly observed and can only be inferred through ARQ feedback (ACK/NACK/no feedback). Under this partially observable condition, we formulate the joint optimization of RIS phase shifts, UAV altitudes, and SU transmit powers as a multi-agent Markov decision process (MDP) to maximize the SU sum rate while ensuring the PU rate requirement.
- We develop a Multi-Agent Proximal Policy Optimization (MAPPO) algorithm under the Centralized Training Decentralized Execution (CTDE) framework to solve the continuous-action multi-agent problem. To provide a comprehensive performance evaluation, we further compare the MAPPO with Independent Proximal Policy Optimization (IPPO) and a MAPPO variant with clipping (MAPPO-pure). Simulation results show that the MAPPO-based scheme achieves faster convergence and a higher SU sum rate compared with these baseline approaches.

The remainder of the paper is organized as follows. Section 2 provides a comprehensive introduction to the system model. Section 3 presents the problem formulation. Section 4 details the proposed algorithms. Section 5 presents the experimental results. Section 6 provides detailed discussions and analysis of the obtained results. Finally, Section 7 concludes the paper.

2. System Model

As illustrated in Fig. 1, we consider a CR system assisted by an UAV and an RIS, where the UAV, equipped with the RIS, serves as a relay for the SU link. This architecture aims to enhance the data rate of the SU while mitigating interference to the PU. Specifically, we adopt a multi-cell CR framework, where each SU cell consists of a secondary transmitter (ST), a UAV equipped with an RIS, and a secondary receiver (SR). Additionally, a PU cell is considered, comprising a primary transmitter (PT) and a primary receiver (PR).

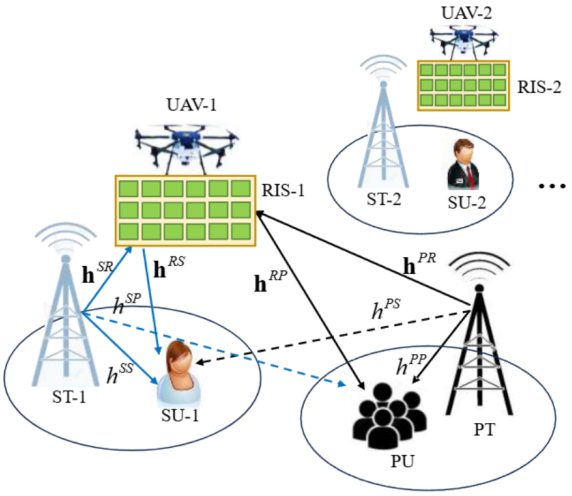

Fig. 1. The UAV-RIS-assisted CR communication system, where each SU is equipped with a UAV and RIS, and the RIS is deployed on the UAV.

Let the number of SU cells be M, and let the number of UAVs and RISs be L. For simplicity, we assume $M = L$. Each RIS is composed of N passive reflecting elements. We respectively define the channel gains of PT-RIS, ST-RIS, RIS-PR, RIS-SR, PT-PR, PT-SR, ST-SR, and ST-PR at the t-th time slot as $\mathbf{h}_t^{PR_l} \in \mathbb{C}^{N\times 1}$, $\mathbf{h}_t^{S_m R_l} \in \mathbb{C}^{N\times 1}$, $\mathbf{h}_t^{R_l P} \in \mathbb{C}^{N\times 1}$, $\mathbf{h}_t^{R_l S_m} \in \mathbb{C}^{N\times 1}$, h_t^{PP}, $h_t^{PS_m}$, $h_t^{S_m S_m}$, and $h_t^{S_m P}$. Due to significant path loss, we neglect signals that undergo two or more reflections via the RIS. The transmit power of the PT is fixed and denoted as P^P, while the transmit power of the m-th ST is denoted as $P_t^{S_m} \in [0, P_{\max}]$.

We define the spectrum state of the PU link in the t-th time slot as $s_t^P \in \{0, 1\}$, where $s_t^P = 1$ indicates that the PU is active, and $s_t^P = 0$ indicates that the PU is idle. Furthermore, the spectrum occupancy is modeled as a Markov process, a common approach in CR networks [1], [2], where the transition probability from state i to state j is denoted by P_{ij}. Accordingly, the transition probability matrix is given by

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}. \tag{1}$$

2.1 Channel Model

For the t-th time slot, the distance between the m-th ST and the l-th RIS can be given by

$$d_t^{(\mathrm{ST})_m R_l} = \sqrt{\left(x_t^{R_l} - x^{(\mathrm{ST})_m}\right)^2 + \left(y_t^{R_l} - y^{(\mathrm{ST})_m}\right)^2 + \left(H_t^{R_l}\right)^2} \tag{2}$$

where the coordinates of the RIS are assumed to be identical to those of the UAV. Similarly, the distance between the SR and the RIS is defined in the same manner. Note that $x_{R_l}^t$ and $y_{R_l}^t$ represent the x and y coordinates of the l-th RIS at time t, while $H_{R_l}^t$ denotes the height of the RIS, and the coordinates of the m-th ST are denoted by $x_{(\mathrm{ST})_m}$ and $y_{(\mathrm{ST})_m}$.

The channels between the PT and the RIS, the ST and the RIS, the RIS and the PR, and the RIS and the SR are modeled as probabilistic channels. As an example, we consider the fading of the channel between the m-th ST and the l-th RIS, expressed as

$$\mathbf{h}_t^{S_m R_l} = \sqrt{\frac{\rho}{\left(d_t^{(\mathrm{ST})_m R_l}\right)^2}} \left[P_t^{S_m R_l,\mathrm{LoS}} \mathbf{h}_t^{S_m R_l,\mathrm{LoS}} + \kappa \left(1 - P_t^{S_m R_l,\mathrm{LoS}}\right) \mathbf{h}_t^{S_m R_l,\mathrm{NLoS}} \right] \tag{3}$$

where ρ denotes the channel power gain at a reference distance of 1 meter, which is related to the reference path loss ρ_0 (in dB) by $\rho = 10^{-\rho_0/10}$, and κ represents the additional attenuation factor for the non-line-of-sight (NLoS) component. In addition, $P_t^{S_m R_l,\mathrm{LoS}}$ is the line-of-sight (LoS) probability in t-th time slot, given by [19]

$$P_t^{S_m R_l,\mathrm{LoS}} = \frac{1}{1 + C \exp\left(-D\left[\theta_t^{S_m R_l} - C\right]\right)} \tag{4}$$

where C and D are the environment parameters, $\theta_t^{S_m R_l} = \frac{180}{\pi} \sin^{-1}\left(\frac{H}{d_t^{(\mathrm{ST})_m R_l}}\right)$ is the elevation from the ST to RIS. $\mathbf{h}_t^{S_m R_l,\mathrm{LoS}} \in \mathbb{C}^{N\times 1}$ denotes the LoS link and $\mathbf{h}_t^{S_m R_l,\mathrm{NLoS}} \in \mathbb{C}^{N\times 1}$ is the NLoS link.

Assume that the small-scale fading of the ground-to-ground links between the PT and PR, PT and SR, ST and SR, and ST and PR can be modeled by the Rician fading channel, which is suitable for scenarios with a dominant line-of-sight path [18]. For example, the fading of the channel between the ST and SR can be expressed as

$$h_t^{S_m S_m} = \sqrt{\rho \left(d^{(\mathrm{ST})_m S_m}\right)^{-2}} \left(\sqrt{\frac{K}{K+1}}\, h_t^{S_m S_m,\mathrm{LoS}} + \sqrt{\frac{1}{K+1}} h_t^{S_m S_m,\mathrm{NLoS}} \right) \tag{5}$$

where $d^{(\mathrm{ST})_m S_m}$ denotes the distance between the ST and the SR, and K is the Rician factor.

2.2 Data Rate Analysis

To analyze the performance of the proposed system, we first derive its signal-to-interference-plus-noise ratio (SINR), which is a key performance metric to quantify the quality of a communication link. The SINRs at the PR and the m-th SR in the t-th time slot are given by

$$\mathrm{SINR}_t^{P^R} = \frac{s_t^P \left| h_t^{P^T P^R} + \sum_{l=1}^{L} \mathbf{h}_t^{R_l P^R} \mathbf{\Phi}_t^l \mathbf{h}_t^{P^T R_l} \right|^2 \sqrt{\zeta^l}}{I_{S^T}(t) + (\sigma^{P^R})^2}, \tag{6}$$

and

$$\mathrm{SINR}_m^{S^R} = \frac{\left| h_t^{S^T S_m^R} + \sum_{l=1}^{L} \mathbf{h}_t^{R_l S_m^R} \mathbf{\Phi}_t^l \mathbf{h}_t^{S_m^T R_l} \right|^2 \sqrt{\zeta_t^{S_m^T}}}{I_{P^T}(t) + I_{S_j^T}(t) + (\sigma^{S_m^R})^2} \tag{7}$$

where $\mathbf{\Phi}_t^l = \text{diag}\left(e^{j\theta_t^1}, e^{j\theta_t^2}, \ldots, e^{j\theta_t^N}\right)$ denotes the phase shift matrix of the l-th RIS, with $\theta_t^n \in [0, 2\pi)$ being the phase shift of the n-th element. In addition, $\left(\sigma^{PR}\right)^2$ and $\left(\sigma^{SR_m}\right)^2$ denote the additive white Gaussian noise (AWGN) powers at the PR and the m-th SR, respectively. Furthermore, $I_{ST}(t)$ is the total interference at the PR from all STs, while $I_{PT}(t)$ and $I_{ST_j}^T(t)$ represent the interference at the SR from the PT and from all STs except the m-th one, respectively.

Therefore, the data rate that can be achieved at PR is

$$C_t^{P^R} = \log_2\left(1 + \text{SINR}_t^{P^R}\right). \tag{8}$$

Similarly, the data rate of the m-th SR can be expressed as

$$C_t^{S_m^R} = \log_2\left(1 + \text{SINR}_t^{S_m^R}\right). \tag{9}$$

According to (9), the sum rate of all SRs is

$$C_t^{S^R} = \sum\nolimits_{m=1}^{M} C_t^{S_m^R}. \tag{10}$$

3. Problem Formulation

To maximize the achieved sum rate of all SRs, we formulate the joint optimization problem by optimizing the phase shift matrix of the RIS, the altitude of the UAV, and the transmit power of ST $P_t^{S_m}$, given by

$$\max_{\mathbf{\Phi}_t^l, H_t^l, \zeta_t^{S_m^T}} \sum_{t=0}^{T} C_t^{S^R}\left(\mathbf{\Phi}_t^l, H_t^l, \zeta_t^{S_m^T}\right) \tag{11a}$$

$$\text{s.t.} \quad \left|\mathbf{\Phi}_t^{l,i}\right| = 1, \quad \forall i = 1, 2, \ldots, N, \tag{11b}$$

$$\zeta_t^{S_m^T} \in [0, \zeta_{\max}], \tag{11c}$$

$$H_t^l \in [0, H_{\max}], \tag{11d}$$

$$\text{if } s_t^P = 1: \quad C_t^{P^R}\left(\mathbf{\Phi}_t^l, H_t^l, \zeta_t^{S_m^T}\right) \geq \alpha^P, \tag{11e}$$

$$C_t^{S_m^R} \geq C_t^{S_m^{\text{req}}} \tag{11f}$$

where $\Phi_t^{l,i}$ represents the i-th diagonal element of the phase shift matrix. The maximum transmit power of each ST is denoted by $\zeta_{\max}$, and the UAV altitude is constrained within the range $[0, H_{\max}]$. Furthermore, (11e) ensures that the achievable rate of the PU is not lower than the minimum requirement α^P when the PR is active. Meanwhile, the sum rate of all SUs satisfies $C_t^{S_m} \geq C_t^{S_m^{\text{req}}}$, as specified in (11f). However, in practical systems without spectrum sensing, the spectrum occupancy state s_t^P of the PU remains unknown.

Even though the spectrum state is obtained, the achievable rate of PU and the interference threshold typically remain unknown to SU due to the non-cooperative nature of the PU. To tackle these challenges, we formulate the optimization problem as a MDP, characterized by the tuple $(\mathcal{S}, \mathcal{A}, r_t, p, \gamma)$. Specifically, $\mathcal{S}$ denotes the state space, $\mathcal{A}$ represents the action space, r_t is the reward function, $p = \Pr(s_{t+1} \mid s_t, a_t)$ defines the transition probability from state $s_t \in \mathcal{S}$ to $s_{t+1} \in \mathcal{S}$ under action $a_t \in \mathcal{A}$, and $\gamma \in [0, 1]$ is the discount factor. In particular, the components of this MDP are detailed as follows.

- **Agent:** The agent consists of an ST, an RIS, and a UAV, which are jointly coordinated by an intelligent decision-making unit. The agent interacts with the environment to learn an optimal policy that maximizes the sum rate of the SUs while ensuring reliable communication for the PU.

- **State s_t:** The agent infers the spectrum occupancy state by monitoring the ARQ feedback from the PR. The state can be categorized into three cases: 1) *ACK*: Indicates successful communication, i.e., $C_t^P \geq \alpha^P$, implying that the interference caused by SU spectrum sharing does not degrade PU communication performance. 2) *NACK*: Indicates communication failure, i.e., $C_t^P < \alpha^P$, which is attributed to excessive interference from the SU. 3) *No Feedback (NF)*: Indicates that the PU is idle.

  Thus, the state of PU at the t-th time slot can be denoted as $\tilde{s}_t^P \in \{\text{NF}, \text{ACK}, \text{NACK}\}$.

  In addition, the agent has access to the states of the SU, RIS, and UAV, which collectively define the local observation of the m-th agent at the t-th time slot, given by

$$s_t^m = \left\{\tilde{s}_{t-1}^P, \tilde{s}_t^{S_m}, \mathbf{h}_t, H_{t-1}^l \mid \forall t, \forall m, \forall l\right\} \tag{12}$$

  where $\tilde{s}_t^{S_m} = 1$ indicates that the rate requirement of the m-th SU is met; otherwise, $\tilde{s}_t^{S_m} = 0$. $\mathbf{h}_t$ denotes the channel state information and H_{t-1}^l represents the altitude of the l-th UAV at the previous time slot.

  Accordingly, the global system state at the t-th time slot is expressed as

$$s_t = \left\{s_t^1, s_t^2, \ldots, s_t^M \mid \forall t\right\}. \tag{13}$$

- **Action a_t:** For the t-th time slot, the agent selects an action to interact with the environment, which includes the phase shift of the RIS $\mathbf{\Phi}_t^l$, the transmit power of the ST $P_t^{S_m}$, and the altitude of the UAV H_t^l. To simplify the analysis, it is assumed that the RIS elements operate independently, i.e., there is no mutual coupling between elements. Under this assumption, the RIS phase shift matrix $\mathbf{\Phi}_t^l$ can be equivalently represented as a vector, denoted as $\theta_t^l = \left[\theta_t^{l,1}, \theta_t^{l,2}, \ldots, \theta_t^{l,N}\right]$. The action space can be defined by

$$a_t^m = \left\{\theta_t^l, P_t^{S_m}, H_t^l \mid \forall t, \forall m, \forall l\right\}. \tag{14}$$

  Thus, the joint action space of all agents can be expressed as

$$a_t = \left\{a_t^1, a_t^2, ..., a_t^M \mid \forall t\right\}. \tag{15}$$

- **Reward function r_t:** The reward function is developed to maximize the sum rate of all SUs. Assuming a global reward is shared for all agents, the instantaneous reward at time slot t is defined as

$$r_t = \begin{cases} C_t^S + A\,, & \text{if } \tilde{s}_t^P = \text{NF or ACK} \\ -K, & \text{if } \tilde{s}_t^P = \text{NACK} \end{cases} \tag{16}$$

where $A = \xi \sum_{m=1}^{M} \min\left\{C_t^{S_m} - C_t^{S_m^{\text{req}}}, 0\right\}$ serves a penalty term derived from the constraint (11f). Specifically, when the minimum sum rate requirement for SUs is not met, the agent incurs a penalty scaled by the adjustment parameter ξ. Furthermore, a fixed penalty $-K$ is imposed whenever the feedback is NACK.

Therefore, the optimization problem can be solved by identifying the optimal policy π^* for the Markov process, which maximizes the expected cumulative reward of all SUs, given by

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[U_t \mid \pi\right] \tag{17}$$

where U_t denotes the discounted cumulative reward, given by

$$U_t = \sum_{t=0}^{T} \gamma^t r_{t+1}. \tag{18}$$

4. Algorithm Design for Multi-Agent UAV-RIS-Assisted CR Systems

For the multi-agent scenario with a continuous variable space, we adopt two algorithms, i.e., IPPO and MAPPO, to solve the formulated problem by jointly optimizing the phase shift of the RIS, the UAV altitude, and the transmit power of the SU.

4.1 IPPO

The IPPO algorithm extends Proximal Policy Optimization (PPO) to multi-agent tasks using a decentralized training and execution framework, where each agent independently operates its own PPO instance. IPPO employs an Actor-Critic (AC) architecture [26], where the actor network selects actions to maximize cumulative rewards while the critic network estimates state-value functions to enable policy updates.

In the IPPO algorithm, the advantage function at the t-th time slot can be defined as [27]

$$A\left(s_t^m, a_t^m\right) = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}^m \tag{19}$$

where γ is the discount factor, λ is the Generalized Advantage Estimation (GAE) parameter, and δ_t^m is the temporal-difference (TD) error for the t-th time slot, given by

$$\delta_t^m = r_t^m + \gamma V_{\varphi^m}\left(s_{t+1}^m\right) - V_{\varphi^m}\left(s_t^m\right) \tag{20}$$

where $V_{\varphi^m}\left(s_t^m\right) = \mathbb{E}\left[U_t \,|s_t\right]$ is the state value function and φ^m denotes the critic-network parameters of the agent.

The objective function of the actor-network for the m-th SU is expressed as [28]

$$J^m(\mu) = \mathbb{E}_{s_t^m, a_t^m}\Big[\min\Big(p_\mu^t A_{\pi_{\text{old}}^m}\left(s_t^m, a_t^m\right), \text{clip}\left(p_\mu^t, 1-\varepsilon, 1+\varepsilon\right) A_{\pi_{\text{old}}^m}\left(s_t^m, a_t^m\right)\Big)\Big] \tag{21}$$

where μ denotes the actor-network parameter, ε is the clip parameter, and p_μ^t is the importance sampling ratio of the current policy and the previous policy, expressed as

$$p_\mu^t = \frac{\pi_{\mu_{\text{new}}^m}\left(a_t^m \left| s_t^m\right.\right)}{\pi_{\mu_{\text{old}}^m}\left(a_t^m \left| s_t^m\right.\right)}. \tag{22}$$

With the clip function, the magnitude of the policy update can be controlled in $(1-\varepsilon, 1+\varepsilon)$ to avoid over-adjustment, thus improving the stability and convergence of the algorithm.

The loss function of the critic network for m-th agent can be defined as

$$L^m(\varphi) = \mathbb{E}_{s_t^m}\Big[\max\Big(\left(V_{\varphi_{\text{new}}^m}(s_t^m) - \sum_{\tau>t} \gamma^{\tau-t} r_\tau\right)^2, \left(\text{clip}\left(V_{\varphi_{\text{new}}^m}(s_t^m), V_{\varphi_{\text{old}}^m}(s_t^m) - \varepsilon, V_{\varphi_{\text{old}}^m}(s_t^m) + \varepsilon\right) - \sum_{\tau>t} \gamma^{\tau-t} r_\tau\right)^2\Big)\Big] \tag{23}$$

where φ_{new}^m and φ_{old}^m denote the parameters of the current and previous critic network, respectively. The update of the loss function is restricted within a trust region, thereby enhancing training stability by preventing overly aggressive parameter shifts in response to individual samples.

4.2 MAPPO

For multi-agent reinforcement learning, a widely adopted framework is CTDE [29]. This paradigm is particularly effective as it enables coordination during the training phase while maintaining agent autonomy during execution, facilitating efficient real-time decision-making. Our approach leverages this framework [30], [31], and the specific architecture is illustrated in Fig. 2.

As depicted in Fig. 3, the centralized critic processes the global state, denoted as S, which encompasses information from all agents [32]. This input is passed through three fully-connected (FC) layers with neuron counts of 256, 256, and 128 [33], to output a single state-value, $V(S)$. In contrast, each decentralized actor operates solely on its local observation o_i (e.g., ARQ, SU state, channel state information (CSI)). It utilizes an identical network structure to produce a multi-dimensional action distribution $\pi(a_i|o_i)$, which determines the agent's next action (e.g., Phase, Alt, Power) [34].

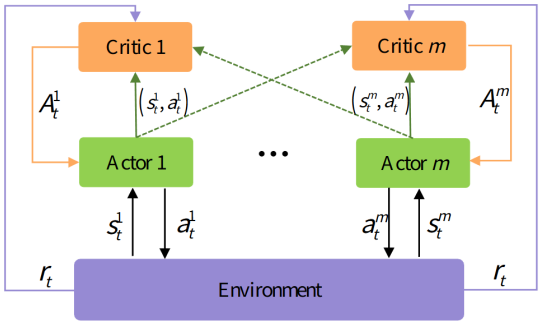

Fig. 2. The framework of MAPPO, where each agent contains a critic and actor network, respectively, interacting with the environment to maximize or minimize the objective function.

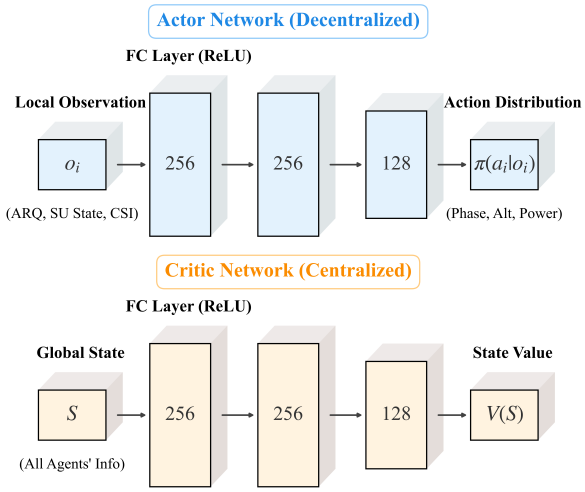

Fig. 3. The Actor-Critic architecture of the proposed MAPPO algorithm under the CTDE framework.

Unlike IPPO, which exclusively utilizes local inputs, MAPPO incorporates a centralized value function that integrates global information potentially unavailable in agents' local observations, thereby enabling PPO to effectively implement the CTDE framework within multi-agent systems. Consequently, the advantage function of MAPPO is defined as [35]

$$A\left(s_t, a_t^m\right) = \sum_{t=0}^{T} (\gamma\lambda)^{-t} \delta_t \tag{24}$$

where δ_t represents the TD error, calculated by

$$\delta_t = r_t + \gamma V_{\varphi^m}(s_{t+1}) - V_{\varphi^m}(s_t) \tag{25}$$

where r_t denotes the immediate reward received at time step t, γ represents the discount factor quantifying the importance of future rewards, and $V_{\varphi^m}(s_t)$ is the centralized value function parameterized by φ^m.

In (24), $(\gamma\lambda)^{-t}$ serves as a discount factor that reduces the impact of past rewards on the current advantage estimation, where $\lambda \in [0, 1]$ is an additional damping factor. Moreover, with (24), the weighted TD errors can be aggregated over the episode, quantifying how much better or worse taking action a_t in global state s_t is compared to the expected action under the current policy [36].

Algorithm 1. Training procedure of MAPPO.

Initialize the actor-network μ and critic network φ.
Initialize the policy π.
for $e = 1 \to E$ **do**
  Each agent obtains the initial state s_0^m.
  for $t = 1 \to T$ **do**
    Each agent executes action a_t^m according to $\pi_{\mu_{\text{old}}^m}\left(a_t^m \middle| s_t^m\right)$.
    Each agent obtains the reward r_t according to (16) and the next state s_{t+1}^m.
    Store trajectory (s_t, a_t, r_t, s_{t+1}) in the experience pool $\mathcal{D}$.
    if Experience pool is full **then**
      Sample a small batch from $\mathcal{D}$.
      Estimate the advantage function (24).
      Each agent updates the actor-network μ_{new}^m with policy gradient.
      Each agent updates the critic-network φ_{new}^m by minimizing the loss function.
      Update the policy $\pi_{\text{old}}^m \leftarrow \pi_{\text{new}}^m$.
    end if
  end for
end for

According to [37], the objective functions of the actor and critic networks in MAPPO align with those in IPPO, differing only in their reliance on the global state, expressed as

$$J^m(\mu) = \mathbb{E}_{s_t, a_t^m}\left[\min\left(p_\mu^t A_{\pi_{\text{old}}^m}\left(s_t, a_t^m\right), \text{clip}\left(p_\mu^t, 1-\varepsilon, 1+\varepsilon\right) A_{\pi_{\text{new}}^m}\left(s_t, a_t^m\right)\right)\right] \tag{26}$$

and

$$L^m(\varphi) = -\mathbb{E}_{s_t}\left[\min\left(\left(V_{\varphi_{\text{new}}^m}(s_t) - \sum_{\tau>t} \gamma^{\tau-t} r_\tau\right)^2, \left(\text{clip}\left(V_{\varphi_{\text{new}}^m}(s_t) - V_{\varphi_{\text{old}}^m}(s_t), -\varepsilon, +\varepsilon\right) - \sum_{\tau>t} \gamma^{\tau-t} r_\tau\right)^2\right)\right] \tag{27}$$

where s_t is the global state of all agents.

The MAPPO algorithm is detailed in Algorithm 1. Initially, all network parameters, including those of the actor and critic networks, are randomly initialized, and the agent policy is set accordingly. The algorithm then iteratively executes for K episodes, each comprising T time steps. At the beginning of each episode, every agent observes its initial state s_0^m.

4.3 MAPPO-pure

In addition to the MAPPO algorithm, we develop a novel MAPPO algorithm named MAPPO-pure, which eliminates redundant PU state information from the global state input. This modification aims to evaluate the effect of reducing the input state space on the learning efficiency and overall performance of the algorithm. Both MAPPO and

MAPPO-pure employ a clipping parameter ε, which is crucial for controlling convergence and ensuring stability during the training process [38].

4.4 Computational Complexity and Onboard Resource Requirements

To evaluate the computational burden of the proposed multi-agent UAV-RIS cooperation algorithm in the online phase [39], let M denote the number of agents, d_s the state dimension of a single agent, d_a the action dimension, and consider an actor network with architecture

$$d_s \to h_1 \to h_2 \to h_3 \to d_a.$$

The total number of multiply-accumulate (MAC) operations required for one multi-agent decision can be approximated by

$$C_{\text{online}} \approx M\left(d_s h_1 + h_1 h_2 + h_2 h_3 + h_3 d_a\right). \tag{28}$$

In our setting, we have $M = 3$, $d_s \approx 9$, $d_a \approx 12$, $h_1 = h_2 = 256$, and $h_3 = 128$. Substituting these values into (28) yields

$$C_{\text{online}} \approx 3.1 \times 10^5 \text{ MACs per time slot}$$

which corresponds to the computation scale of a compact deep neural network (DNN). The complexity of channel and SINR calculations in the environment is further limited by the number of cells M and RIS elements $N = 10$, and is on the order of 10^3 operations per time slot under the current configuration, thus constituting a secondary overhead compared with actor inference.

Let f_{dec} (in Hz) denote the decision frequency. Approximating each MAC as two floating-point operations (FLOPs), the required floating-point throughput of the online inference phase can be written as

$$C_{\text{FLOPs}} \approx 2\, C_{\text{online}}\, f_{\text{dec}}. \tag{29}$$

For a conservative upper-bound decision frequency of $f_{\text{dec}} = 100$ Hz, substituting into (29) gives

$$C_{\text{FLOPs}} \approx 6 \times 10^7 \text{ FLOPs/s} \approx 0.06 \text{ GFLOPS}.$$

By comparison, the NVIDIA Jetson Nano, a standard embedded platform for UAVs, delivers a peak throughput of 472 GFLOPS (FP16) within a 5–10 W power envelope [40]. Consequently, the proposed algorithm utilizes a negligible fraction of the available onboard resources, even at a 100 Hz decision rate. Furthermore, given the offline execution of the critic network via CTDE and the passive nature of the RIS, the proposed scheme imposes minimal hardware overhead, ensuring practical feasibility [41].

5. Numerical Results

This section evaluates the proposed IPPO and MAPPO algorithms within the UAV-RIS-assisted CR system, focusing on the impact of the clipping parameter and RIS element count on the SU sum rate. To investigate input dimensionality, we introduce MAPPO-pure, a streamlined variant excluding redundant PU state information. The simulation deploys three SUs with a maximum UAV altitude of $H_{\max} = 30$. The UAV coordinates are set as $\left[80, 40, H_t^1\right]$, $\left[120, 100, H_t^2\right]$, and $\left[80, 160, H_t^3\right]$, where H_t^m denotes the dynamic altitude of the m-th UAV at time slot t.

To ensure simulation realism, the system parameters align with standard urban micro-cell scenarios. We set the carrier frequency to 2.4 GHz, yielding a theoretical reference path loss of 40 dB. For simplicity, we assume that all transmitters and receivers are equipped with isotropic antennas, corresponding to an antenna gain of 0 dBi. Standard urban blockage is modeled using probabilistic LoS parameters ($C = 10, D = 0.6$) and an attenuation factor $\kappa = 0.1$, while a Rician factor of $K = 10$ captures the strong LoS component of low-altitude UAVs. Furthermore, considering the shared spectrum nature, the noise floor is set to -80 dBm to account for the aggregate interference in an interference-limited environment. Table 1 details these settings.

Figure 4 illustrates the training dynamics of the Actor and Critic networks within the MAPPO framework. The Actor loss (blue curve) exhibits an initial upward trend, reflecting the active optimization of the policy surrogate function. Subsequently, the curve flattens and stabilizes, indicating that the policy has converged to a robust strategy. Meanwhile, the Critic loss (red curve) displays high-frequency oscillations within a bounded range. These fluctuations are attributed to the stochastic nature of the wireless environment, which introduces variance. Despite this variance, the Critic loss remains stable without divergence, confirming that the network effectively maintains a reliable estimate of the global state value throughout the training process.

Symbol	Description	Value	Unit
f_c	Carrier frequency	2.4	GHz
C, D	Environmental parameters (LoS probability)	10, 0.6	–
ρ_0	Reference path loss at 1 m	40	dB
G_t, G_r	Antenna gain of transmitter and receiver	0	dBi
κ	NLoS attenuation factor	0.1	–
K	Rician factor of ground links	10	–
d_{RIS}	Spacing between adjacent RIS elements	$\lambda/2$	m
σ^2	Effective noise power (interference-limited)	-80	dBm
P_P	Transmit power of PT	40	dBm
$\zeta_{\max}$	Max transmit power of STs	35	dBm
N	Number of elements per RIS	10	–
$H_{\max}$	Maximum UAV altitude	30	m
α	Learning rate (Actor & Critic)	0.0005	–
γ	Discount factor	0.99	–

Tab. 1. Simulation and system parameters.

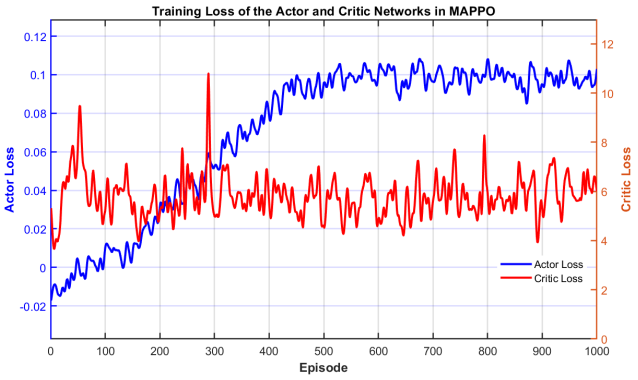

Fig. 4. Training loss of the Actor and Critic networks. The left axis represents the Actor loss, while the right axis indicates the Critic loss, allowing for a simultaneous visualization of their convergence behaviors.

Figures 5 and 6 demonstrate the cumulative reward and average achievable user rate versus episode to evaluate the effectiveness of policy optimization and communication quality. As shown in Fig. 5, MAPPO achieves rapid policy learning during the initial training phase (approximately the first 200 episodes), as evidenced by a steep increase in cumulative reward. The performance curve becomes stable around the 400th episode, converging to a reward level substantially higher than that of IPPO.

Figure 6 further shows the data rate of the algorithms over episodes. First, we can observe that MAPPO consistently increases the achievable user rate during the training process and maintains a high, stable rate upon convergence. In contrast, IPPO demonstrates limited improvement, accompanied by greater fluctuation in the later stages and a significantly lower final average rate. These results indicate that MAPPO is more effective in enhancing the data rate and communication stability, benefiting from its exploitation of shared global state information.

Figure 7 illustrates the sum rate of secondary users as a function of the number of RIS reflecting elements, N, for three different algorithms: IPPO, MAPPO, and MAPPO-pure. The sum rate increases monotonically with N for all methods, which demonstrates the performance benefit of enlarging the RIS reflecting aperture. Notably, MAPPO consistently outperforms IPPO across all values of N, with the performance gap becoming more pronounced as N increases. This trend highlights the superior capability of MAPPO to leverage additional beamforming degrees of freedom offered by the RIS. Among the three approaches, MAPPO-pure consistently achieves the highest sum rate across all configurations, demonstrating that eliminating redundant PU-related inputs from the global state improves learning efficiency, particularly in high-dimensional settings.

Figure 8 illustrates the variations of the average reward for different clipping parameters, revealing an inverse relationship between the coefficient value and convergence speed. Specifically, when $\varepsilon = 0.1$, the system achieves the highest average rewards but converges more slowly. In contrast, $\varepsilon = 0.3$ accelerates convergence at the cost of reduced rewards.

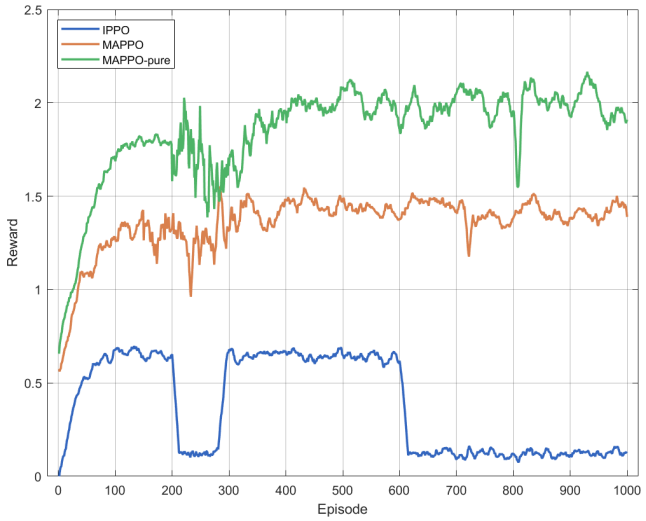

Fig. 5. The comparison of convergence performance under different algorithms.

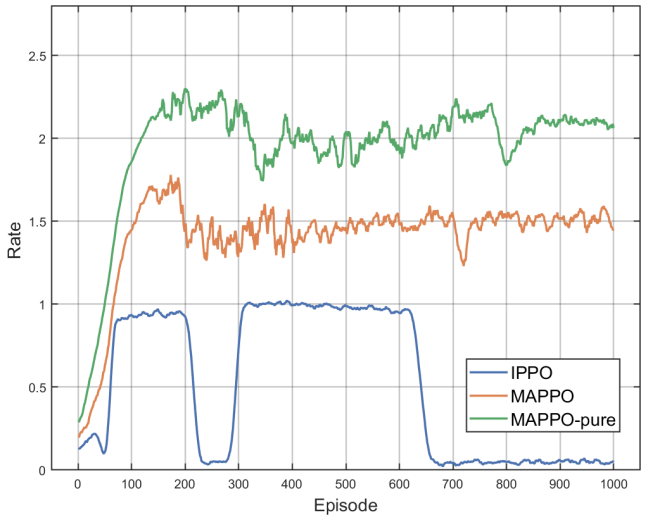

Fig. 6. The achievable sum rate of SUs with different algorithms.

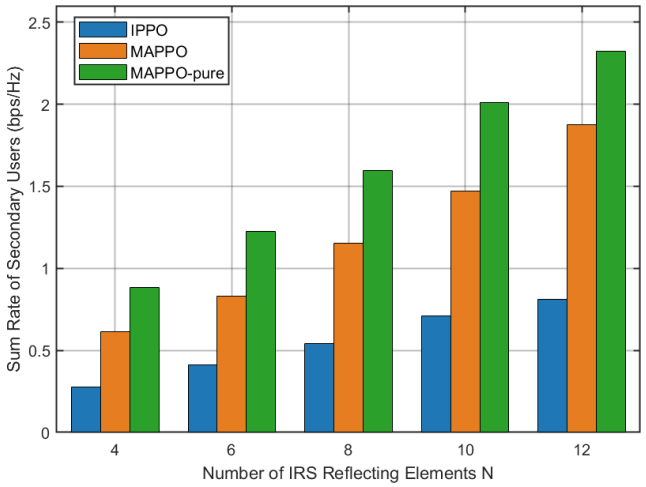

Fig. 7. Impact of the Number of RIS Elements on the System Sum Rate.

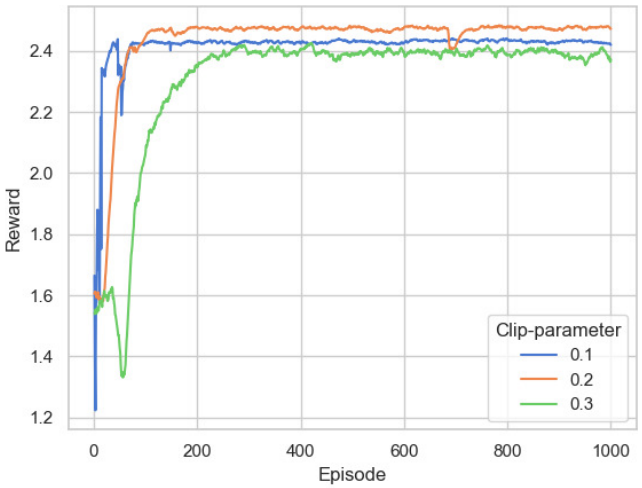

Fig. 8. The cumulative reward of SUs versus different clip parameters in the MAPPO algorithm.

Numerical results indicate that a clipping coefficient of $\varepsilon = 0.2$ achieves an optimal trade-off between stability and convergence speed, which is crucial for ensuring reliable and efficient spectrum sharing in dynamic cognitive radio environments.

Figure 9 plots a spatio-temporal visualization of the phase shift configurations across the RIS elements for Agent 2 over the training process. The heatmap utilizes a chromatic spectrum ranging from deep blue to vibrant red, corresponding to phase values within the interval $[0, 2\pi]$ radians. Each row denotes an individual reflecting element, indexed by $n = 1, 2, \ldots, 10$. At the initial training stages, the phase configurations exhibit high-frequency oscillations, reflecting the exploratory behavior inherent to the initial learning process. As training progresses, the phase distribution transitions from randomness to a more structured configuration.

To evaluate the performance of our approach, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm was selected as the baseline for our experiments. MADDPG is a pioneering work that introduced the CTDE framework, which has since become a standard paradigm in the field of multi-agent systems. A direct comparison with MADDPG under this framework allows for a clear and meaningful assessment of the practical gains achieved by the improvements introduced in our study.

Figure 10 illustrates the reward progression of both algorithms during training. The results unequivocally demonstrate that the MAPPO-based agent significantly outperforms MADDPG. In contrast, the reward curve for MADDPG not only exhibits slow convergence and a lower performance ceiling but also displays severe oscillations throughout the training process, revealing inherent instability in its learning procedure.

From a theoretical standpoint, MADDPG's suboptimal performance can be attributed to intrinsic limitations. First, the centralized critic, based on Q-learning, suffers from significant estimation errors when dealing with the

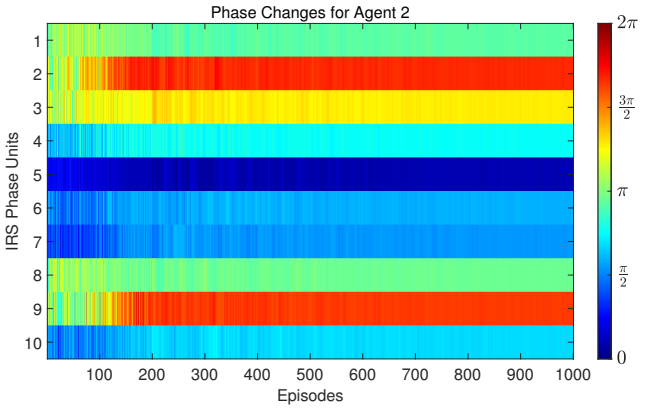

Fig. 9. RIS Phase Variations for Agent 2. Color indicates the phase value $[0, 2\pi]$.

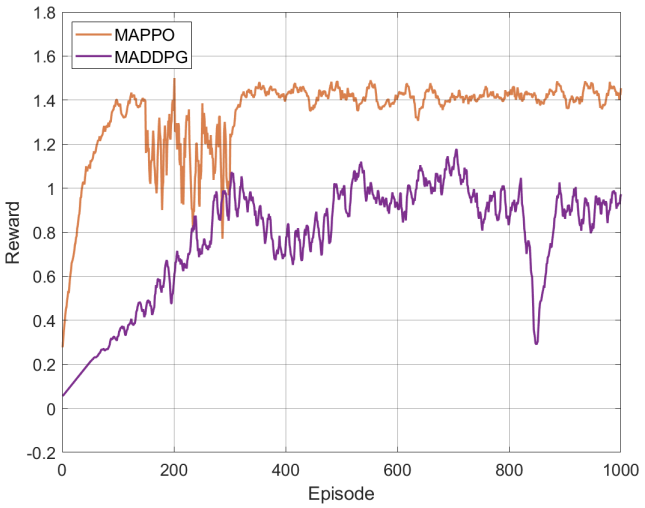

Fig. 10. Convergence performance comparison between the MAPPO algorithm and the MADDPG baseline.

high-dimensional joint action space. These inaccurate value estimates propagate through the policy network, resulting in unreliable gradients that misguide the actor's updates and destabilize the learning process. Second, MADDPG is notoriously sensitive to hyperparameter settings. The algorithm lacks the robustness to adapt to the dynamic wireless environment, where slight deviations in tuning can trigger the severe oscillations and the catastrophic performance collapse observed in the later stages of training.

6. Discussion

In this section, we aim to provide a comprehensive analysis of the simulation results, delving into the underlying mechanisms that contribute to the observed performance differences between the IPPO and MAPPO algorithms. We focus on several critical factors that impact system performance, such as the advantages of the CTDE framework in multi-agent systems, the sensitivity of the algorithms to the clipping coefficient, and the influence of the chosen channel models on the robustness of the algorithms. Through this

discussion, we seek to understand the role of each element in shaping the learning dynamics and overall efficiency of UAV-RIS-assisted cognitive radio systems. Additionally, we explore potential challenges and areas for improvement.

6.1 MAPPO Framework Analysis

The defining advantage of MAPPO lies in its inherent capability for multi-agent collaborative optimization. Whereas IPPO adopts a decentralized training paradigm that constrains each agent to operate solely on localized observations, which is a computationally efficient but inherently suboptimal approach, MAPPO's CTDE architecture fundamentally addresses the partial observability challenge in multi-agent systems. Specifically, the CTDE framework enables agents to leverage global state information during policy optimization while maintaining decentralized decision-making, thereby achieving a principled balance between exploration and exploitation in cooperative tasks.

In spectrum sharing scenarios characterized by non-stationary channel conditions and competitive resource allocation, IPPO's independent learning mechanism inevitably incurs coordination inefficiencies due to the absence of joint action-value estimation. Conversely, MAPPO's centralized critic network facilitates coordinated policy updates through shared value function approximation, effectively mitigating the curse of dimensionality in multi-agent reinforcement learning. Empirical evaluations demonstrate that MAPPO achieves significantly faster convergence and higher cumulative reward compared to IPPO, with performance gains in high-interference operational regimes. These quantitative results validate the theoretical premise that global state awareness is essential for optimizing cooperative strategies in electromagnetic spectrum-sharing applications.

6.2 Clipping Coefficient Effects

The clipping coefficient serves as a critical hyperparameter in MAPPO, mediating the trade-off between policy update stability and learning efficiency. The experimental results indicate that this parameter exerts a non-linear influence on the magnitude of policy gradient updates. Specifically, a smaller clipping coefficient effectively constrains policy divergence, thereby enhancing stability but at the cost of slower convergence. Conversely, larger values accelerate learning by permitting more substantial updates, yet at the expense of increased reward variance and potential instability.

An appropriately tuned clipping coefficient achieves a Pareto-optimal balance between these competing objectives, reducing excessive policy change while maintaining sample efficiency. This observation is consistent with theoretical predictions from trust region policy optimization, where bounded policy updates prevent destructive gradient steps. Moreover, future implementations may benefit from adaptive scheduling of the clipping coefficient, such as by employing annealing strategies that transition from exploratory to exploitative phases based on policy entropy metrics.

6.3 Channel Modeling Validation

The developed environmental simulation framework incorporates a hybrid channel model to capture the complexity of real-world wireless communication environments. For air-to-ground links, a probabilistic LoS/NLoS model is adopted to capture altitude-dependent signal propagation dynamics, accounting for the varying likelihood of unobstructed paths as a function of UAV elevation. For ground-to-ground links, a Rician fading model is employed to characterize dominant LoS components superimposed with multipath scattering effects. This dual-model design achieves a balance between computational tractability and physical-layer realism by jointly modeling deterministic path loss and stochastic fading characteristics. Moreover, the integration of environmental noise and frequency-selective attenuation further enhances scenario realism, thereby facilitating robust validation of DRL-based resource allocation algorithms under diverse channel conditions.

6.4 Future Research Directions

The proposed framework faces two fundamental challenges requiring further investigation. First, ensuring long-term adaptability is essential for real-world deployment. While the current algorithms optimize short-term spectrum sharing, they lack mechanisms to handle non-stationary user behaviors and dynamic interference patterns. Second, scalability limitations become pronounced in large-scale multi-agent systems, where coordination complexity grows exponentially with agent count, potentially degrading performance through increased interference and computational overhead. Addressing these challenges calls for lightweight yet robust coordination mechanisms and adaptive learning architectures capable of maintaining a balance between global cooperation and decentralized execution. Future research directions include the integration of domain-specific priors, such as integrating domain-specific knowledge into the DRL framework to enhance scalability.

7. Conclusion

In this paper, we propose the IPPO and MAPPO algorithms for a multi-UAV-and-RIS-assisted CR system to address the challenge of multiple SUs attempting to share the PU spectrum. Furthermore, we formulate the problem as an MDP process to maximize the sum rate of SUs by jointly optimizing the phase shift of the RIS, the altitude of the UAV, and the transmit power of the SU. In addition, we introduce a variant of the MAPPO algorithm termed MAPPO-pure to further investigate the impact of global state information on system performance. Simulation results demonstrate that the MAPPO algorithm, which employs a centralized training and decentralized execution framework, achieves higher system performance compared to the IPPO algorithm. Moreover, the clipping coefficient on the stability and convergence speed of the training of the proposed algorithm is investigated.

The source codes for the simulations are available at: https://github.com/piko-mo/CR-with-UAV-and-RIS

Acknowledgments

This work was supported by the Hainan Province Science and Technology Special Fund under Grant ZDYF2024GXJS292.

References

[1] SANDYA, B. H., NAGAMANI, K., SHAVANTHI, L. A review of cognitive radio spectrum sensing methods in communication networks. In *International Conference on Communication and Signal Processing (ICCSP)*. Chennai (India), 2018, p. 457–461. DOI: 10.1109/ICCSP.2018.8524489

[2] SALAMEH, H. B., ABDEL-RAZEQ, S., AL-OBIEDOLLAH, H. Integration of cognitive radio technology in NOMA-based B5G networks: State of the art, challenges, and enabling technologies. *IEEE Access*, 2023, vol. 11, p. 12949–12962. DOI: 10.1109/ACCESS.2023.3242645

[3] BOSTIAN, C. W., YOUNG, R. A. *The Application of Cognitive Radio to Coordinated Unmanned Aerial Vehicle (UAV) Missions*. Final Technical Report, Virginia Polytechnic Institute and State University, 2011, p. 1–33.

[4] CHATTERJEE, S., DE, S. QoE-aware cross-layer adaptation for D2D video communication in cooperative cognitive radio networks. *IEEE Systems Journal*, 2022, vol. 16, no. 2, p. 2078–2089. DOI: 10.1109/JSYST.2021.3123463

[5] MUSA, A., HALLOUSH, R., SALAMAH, H. B., et al. Exploiting MIMO and cognitive radio for improved performance in indoor communication systems. In *International Conference on Multimedia Computing, Networking and Applications (MCNA)*. Valencia (Spain), 2023, p. 80–84. DOI: 10.1109/MCNA59361.2023.10185644

[6] HASAN, K. M., YU, S., SONG, M. Secured full-duplex UAV-aided spectrum-sharing network based on NOMA. In *IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. Toronto (Canada), 2023, p. 125–133. DOI: 10.1109/MASS58611.2023.00023

[7] FIROUZJAEI, H. M., ZERAATKAR, J. M., ARDEBILIPOUR, M. A virtual MIMO communication for a UAV-enabled cognitive relay network. *IEEE Sensors Journal*, 2023, vol. 23, no. 17, p. 20267–20274. DOI: 10.1109/JSEN.2023.3294280

[8] WANG, Z., ZHOU, F., WANG, Y., et al. Joint 3D trajectory and resource optimization for a UAV relay-assisted cognitive radio network. *China Communications*, 2021, vol. 18, no. 6, p. 184–200. DOI: 10.23919/JCC.2021.06.015

[9] BHOWMICK, A., ROY, S. D., KUNDU, S. Throughput of an energy-harvesting UAV-assisted cognitive radio network. In *National Conference on Communications (NCC)*. Kharagpur (India), 2020, p. 1–6. DOI: 10.1109/NCC48643.2020.9056090

[10] KRAYANI, A., ALAM, S. A., MARCENARO, L., et al. An emergent self-awareness module for physical layer security in cognitive UAV radios. *IEEE Transactions on Cognitive Communications and Networking*, 2022, vol. 8, no. 2, p. 888–906. DOI: 10.1109/TCCN.2022.3161937

[11] NIU, H., LIN, Z., AN, K., et al. Active RIS-assisted secure transmission for cognitive satellite terrestrial networks. *IEEE Transactions on Vehicular Technology*, 2023, vol. 72, no. 2, p. 2609–2614. DOI: 10.1109/TVT.2022.3208268

[12] ZHANG, L., WANG, Y., TAO, W., et al. Intelligent reflecting surface aided MIMO cognitive radio systems. *IEEE Transactions on Vehicular Technology*, 2020, vol. 69, no. 10, p. 11445–11457. DOI: 10.1109/TVT.2020.3011308

[13] HE, J., YU, K., ZHOU, Y., et al. Reconfigurable intelligent surface enhanced cognitive radio networks. In *IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. Victoria (Canada), 2020, p. 1–5. DOI: 10.1109/VTC2020-Fall49728.2020.9348788

[14] ALLU, R., TAGHIZADEH, O., SINGH, S. K., et al. Robust beamformer design in active RIS-assisted multiuser MIMO cognitive radio networks. *IEEE Transactions on Cognitive Communications and Networking*, 2023, vol. 9, no. 2, p. 398–413. DOI: 10.1109/TCCN.2023.3235788

[15] YU, Y., LIU, X., LIU, Z., et al. Joint trajectory and resource optimization for RIS-assisted UAV cognitive radio. *IEEE Transactions on Vehicular Technology*, 2023, vol. 72, no. 10, p. 13643–13648. DOI: 10.1109/TVT.2023.3270313

[16] VO, N. V., LONG, N. Q., DANG, V.-H., et al. Physical layer security in cognitive radio networks for IoT using UAV with reconfigurable intelligent surfaces. In *International Joint Conference on Computer Science and Software Engineering (JCSSE)*. Lampang (Thailand), 2021, p. 1–5. DOI: 10.1109/JCSSE53117.2021.9493817

[17] HU, H., DA, X., HUANG, Y., et al. SE and EE optimization for cognitive UAV network based on location information. *IEEE Access*, 2019, vol. 7, p. 162115–162126. DOI: 10.1109/ACCESS.2019.2951702

[18] ALI, M., YASIR, M. N., BHATTI, M. S. D., et al. Optimization of spectrum utilization efficiency in cognitive radio networks. *IEEE Wireless Communications Letters*, 2023, vol. 12, no. 3, p. 426–430. DOI: 10.1109/LWC.2022.3229110

[19] POGAKU, A. C., DO, D.-T., LEE, B. M., et al. UAV-assisted RIS for future wireless communications: A survey on optimization and performance analysis. *IEEE Access*, 2022, vol. 10, p. 16320–16336. DOI: 10.1109/ACCESS.2022.3149054

[20] YANG, P., YANG, L., KUANG, W., et al. Outage performance of cognitive radio networks with a coverage-limited RIS for interference elimination. *IEEE Wireless Communications Letters*, 2022, vol. 11, no. 8, p. 1694–1698. DOI: 10.1109/LWC.2022.3174639

[21] NGUYEN, D. C., LOVE, D. J., BRINTON, C. G. Intelligent spectrum sensing and resource allocation in cognitive networks via deep reinforcement learning. In *IEEE International Conference on Communications (ICC)*. Rome (Italy), 2023, p. 4603–4608. DOI: 10.1109/ICC45041.2023.10279539

[22] SHAN, Z., LIU, P., WANG, L., et al. A cognitive multi-carrier radar for communication interference avoidance via deep reinforcement learning. *IEEE Transactions on Cognitive Communications and Networking*, 2023, vol. 9, no. 6, p. 1561–1578. DOI: 10.1109/TCCN.2023.3306854

[23] SARIKHANI, R., KEYNIA, F. Cooperative spectrum sensing meets machine learning: Deep reinforcement learning approach. *IEEE Communications Letters*, 2020, vol. 24, no. 7, p. 1459–1462. DOI: 10.1109/LCOMM.2020.2984430

[24] KHAF, S., ALKHODARY, M. T., KADDOUM, G. Partially cooperative scalable spectrum sensing in cognitive radio networks under SDF attacks. *IEEE Internet of Things Journal*, 2022, vol. 9, no. 11, p. 8901–8912. DOI: 10.1109/JIOT.2021.3116928

[25] XIE, H., LIN, R., WANG, J., et al. Power allocation of energy harvesting cognitive radio based on deep reinforcement learning. In *International Conference on Communication and Information Systems (ICCIS)*. Chongqing (China), 2021, p. 45–49. DOI: 10.1109/ICCIS53528.2021.9645987

[26] SUTTON, R. S., MCALLESTER, D., SINGH, S., et al. Policy gradient methods for reinforcement learning with function approximation. In *International Conference on Neural Information Processing Systems (NIPS)*. Denver (USA), 1999, p. 1057–1063.

[27] SCHULMAN, J., MORITZ, P., LEVINE, S., et al. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*. San Juan (Puerto Rico), 2016, p. 1–14. DOI: 10.48550/arXiv.1506.02438

[28] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., et al. Proximal policy optimization algorithms. *arXiv*, 2017, p. 1–12. DOI: 10.48550/arXiv.1707.06347

[29] CHAI, J., LI, W., ZHU, Y., et al. UNMAS: Multiagent reinforcement learning for unshaped cooperative scenarios. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, vol. 34, no. 4, p. 2093–2104. DOI: 10.1109/TNNLS.2021.3105869

[30] LUONG, N. C., HOANG, D. T., GONG, S., et al. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys and Tutorials*, 2019, vol. 21, no. 4, p. 3133–3174. DOI: 10.1109/COMST.2019.2916583

[31] FOERSTER, J., FARQUHAR, G., AFOURAS, T., et al. Counterfactual multi-agent policy gradients. In *International Conference on Artificial Intelligence (AAAI)*. New Orleans (USA), 2018, p. 1–11. DOI: 10.48550/arXiv.1705.08926

[32] YU, C., VELU, A., VINITSKY, E., et al. The surprising effectiveness of PPO in cooperative multi-agent games. In *International Conference on Neural Information Processing Systems (NIPS)*. New Orleans (USA), 2022, p. 1–30. DOI: 10.48550/arXiv.2103.01955

[33] MNIH, V., BADIA, A. P., MIRZA, M., et al. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning (ICML)*. New York (USA), 2016, p. 1928–1937. DOI: 10.48550/arXiv.1602.01783

[34] HU, J., HU, S., LIAO, S.-W. Policy regularization via noisy advantage values for cooperative multi-agent actor-critic methods. *arXiv*, 2021, p. 1–10. DOI: 10.48550/arXiv.2106.14334

[35] LYU, G., LI, M. Multi-agent cooperative control in neural MMO environment based on MAPPO algorithm. In *IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. Hangzhou (China), 2023, p. 1–4. DOI: 10.1109/AICAS57966.2023.10168653

[36] ZHANG, B., YANG, K. Multi-UAV searching trajectory optimization algorithm based on deep reinforcement learning. In *IEEE 23rd International Conference on Communication Technology (ICCT)*. Wuxi (China), 2023, p. 640–644. DOI: 10.1109/ICCT59356.2023.10419808

[37] KANG, H., CHANG, X., MISIC, J., et al. Cooperative UAV resource allocation and task offloading in hierarchical aerial computing systems: A MAPPO-based approach. *IEEE Internet of Things Journal*, 2023, vol. 10, no. 12, p. 10497–10509. DOI: 10.1109/JIOT.2023.3240173

[38] FENG, Z., HUANG, M., WU, D., et al. Multi-agent reinforcement learning with policy clipping and average evaluation for UAV-assisted communication Markov game. *IEEE Transactions on Intelligent Transportation Systems*, 2023, vol. 24, no. 12, p. 14281–14293. DOI: 10.1109/TITS.2023.3296769

[39] SZE, V., CHEN, Y.-H., YANG, T.-J., et al. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 2017, vol. 105, no. 12, p. 2295–2329. DOI: 10.1109/JPROC.2017.2761740

[40] FRANKLIN, D. *Jetson Nano Brings AI Computing to Everyone*. NVIDIA Technical Blog, 2019. [Online]. Available at: https://blogs.nvidia.com/blog/2019/03/18/jetson-nano-aicomputing/

[41] LANE, N. D., BHATTACHARYA, S., GEORGIEV, S., et al. DeepX: A software accelerator for deep learning on mobile devices. In *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. Vienna (Austria), 2016, p. 1–12. DOI: 10.1109/IPSN.2016.7460664

About the Authors ...

Siyu QIAN received the B.E. degree in Communication Engineering from Nanjing Institute of Technology, Nanjing, China in 2024, and is currently pursuing the M.E. degree in Communication Engineering at Nanjing University of Science and Technology, Nanjing, China. His current research interests with intelligent reflecting surface, unmanned aerial vehicle communications, and Covert communication system design.

Linzi HU received the B.E. degree in Communication Engineering from Nanjing University of Science and Technology, Nanjing, China in 2023. She is currently working toward the M.E. degree in Communication Engineering at Nanjing University of Science and Technology, Nanjing, China. Her research interests include wireless communications, signal processing, and information security.

Yuwen QIAN (corresponding author) received the Ph.D. degree in Automatic Engineering from Nanjing University of Science and Technology, Nanjing, China, in 2011. From Jul. 2002 to Jun. 2011, he was a Lecturer in Automation School of Nanjing University of Science and Technology. Since May 2019, he has been an Associate Professor in School of Electronic and Optical Engineering, Nanjing University of Science and Technology, China.

Long SHI received the Ph.D. degree in Electrical Engineering from the University of New South Wales, Sydney, Australia, in 2012. From 2013 to 2016, he was a Postdoctoral Fellow at the Institute of Network Coding, Chinese University of Hong Kong, China. From 2014 to 2017, he was a Lecturer at Nanjing University of Aeronautics and Astronautics, Nanjing, China. From 2017 to 2020, he was a Research Fellow at the Singapore University of Technology and Design. Now he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include blockchain networks, wireless communications, and federated learning.