2024 IEEE International Workshop on Radio Frequency and Antenna Technologies (iWRF&AT) | 979-8-3503-8476-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/iWRFAT61200.2024.10594249

DRL-Based Joint Trajectory Planning and Beamforming Optimization in Aerial RIS-Assisted ISAC System

Xipeng Chen[1], Xiaowen Cao[1], Lifeng Xie[2], and Yejun He[1]
[1] State Key Laboratory of Radio Frequency Heterogeneous Integration
Sino-British Antennas and Propagation Joint Laboratory, MOST
Guangdong Engineering Research Center of Base Station Antennas
Shenzhen Key Laboratory of Antennas and Propagation
College of Electronics and Information Engineering, Shenzhen University, 518060, China
[2] Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen, China
Email: 2164486229@qq.com, caoxwen@szu.edu.cn, xielf@pcl.ac.cn, heyejun@126.com

***Abstract*—In this paper, we consider a reconfigurable intelligent surface (RIS) enabled integrated sensing and communication (ISAC) system, where the RIS is mounted on an unmanned aerial vehicle (UAV) to enhance signal quality and connectivity coverage by exploring the flexible mobility and adjusting the phase and amplitude of reflected signals. Towards this end, a communication rate max-min problem is formulated by jointly optimizing the beamforming vector, RIS phase shift and UAV trajectory under the constrains of power consumption and sensing requirement. However, due to the coupled variables, this problem is a non-convex problem and hard to be optimally solved. Hence, we reformulate the primary problem as a sequential decision-making problem and exploit a deep reinforcement learning (DRL)-based solution to find a tractable solution. Numerical results validate the effectiveness and the superiority of the proposed algorithm compared with the benchmark schemes.**

***Index Terms*—Reconfigurable intelligent surface, unmanned aerial vehicle, integrated sensing and communication, Deep reinforcement learning.**

I. Introduction

As the sixth generation (6G) networks develop, various intelligent applications have emerged like smart cities, smart transportation, and industrial Internet of Things, which requires high-precision sensing capabilities and high-throughput wireless communication. However, the surge in the demand for connectivity and the proliferation of devices including sensors and smart devices, intensify the congestion within the available frequency bands, leading to a degradation on network performance [1].

In view of this, the integrated sensing and communications (ISAC) has emerged as a promising approach to enhance the sensing capability and improve spectrum utilization [2], which has recognized by International Mobile Telecommunications (IMT)-2030 as one of the core usage scenarios for 6G. It allows simultaneous data transmission and wireless sensing on the same spectrum. On the other hand, owing to the capability to create a more favorable propagation environment, reconfigurable intelligent surface (RIS) has also emerged as an innovative technique in 6G system [3]. As RIS could assist to establish additional non-line-of-sight (NLoS) links for exploiting more degrees of freedom (DoFs) to promote communication performance, the introduction of RIS technology into ISAC system may offer a solution to overcome limitations caused by traditional propagation environment, which has attracted extensive research [4]–[6]. Especially, in a radar-communication coexistence system, the authors in [4] formulated a joint beamforming and communication covariance matrix optimization problem to maximize the radar signal-to-interference-plus-noise ratio (SINR). In a dual-functional radar communication (DFRC) system, the researchers in [5] reduced the Cramér-Rao bound (CRB) on the estimation error and expanded it to the extended target models. Besides, the authors in [6] considered two scenarios with/without the interference at the receiver end to enhance target detection probability.

While current studies have tackled the joint optimization of RIS-aided ISAC in multiple ways, the efficiency might be limited by the fixed location of the RIS in the network. The rapid development of aerial platforms, particularly unmanned aerial vehicles (UAVs), has introduced a dynamic element that can serve as temporary base stations (BS) or relays, offering a flexible deployment solution to enhance ISAC performance [7], which has attracted many attentions. For instance, the authors in [8] jointly designed the UAV maneuver and the beamforming to maximize the weighted sum-rate, while a joint ground user-UAV association and real-time trajectory problem was solved to satisfy the sensing or communication performance in [9]. On the other side, to fully unlock the potential of RIS, researchers have explored the concept of Aerial RIS (ARIS) with flexible deployment by introducing UAV-mounted RIS. In [10], an

AIRS-assisted secure transmission design was proposed to maximize the worst-case sum secrecy rate, while the authors in [11] considered an ARIS-assisted edge computing scheme to facilitate offloading computing tasks from ground user to access point. Additionally, it is worthy to notice that the UAV's trajectory optimization is always a hot research topic but often poses several challenges under specific application and environmental conditions due to its limitation in weight and battery. Compared with traditional solutions, deep reinforcement learning (DRL) has been deemed as a dynamic and adaptive solution to solve such trajectory optimization problems [12], by leveraging neural networks to learn optimal policies through interactions with the environment.

Despite the existing progress on UAV-enabled ISAC and ARIS system, it's notable that the intersection of ARIS and ISAC has largely remained unexplored, which thus motivates our work. In this paper, we propose an ARIS-assisted ISAC (ARIS- ISAC) system. The objective is to maximize the sum of the minimum achievable rate through the joint optimization of UAV trajectory, RIS phase shift and beamforming to guarantee the communication performance of the users. The mobility of the UAV makes the optimal beamforming and RIS phase dynamically change, which further makes the optimization design more complicated. Thus, we exploit the DRL-based algorithm to solve the primary problem by reformulating it as a sequential decision-making problem.

II. System Model

As shown in Fig. 1, an ARIS-assisted ISAC downlink system is proposed. Specifically, an ISAC BS with M transmitting/receiving antennas serves K single-antenna users and simultaneously detects J targets, where user set and target set are denoted as $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{J} = \{1, 2, \ldots, J\}$. In order to lessen the power consumption of UAV and improve the flexibility of RIS, a RIS containing N elements is assembled on the UAV to assist the downlink communication and sensing simultaneously.

According to whether direct link is blocked by obstacles, users are divided into the blocked users with set $\mathcal{K}_{\mathcal{B}}$ and the unblocked users with set $\mathcal{K}_{\mathcal{U}}$ with $\mathcal{K} = \mathcal{K}_{\mathcal{B}} \cup \mathcal{K}_{\mathcal{U}}$. It is assumed that the BS-target link is unblocked due to the fact that the pathloss of the BS-ARIS-target link exists the multiple effect.

In addition, the UAV flies in 2D space. And the UAV departs from a initial point $\mathbf{q}_i = [x_i, y_i]$ and flies at a fixed altitude H. The ISAC period T is divided into L time slots with length being τ and the UAV is at $\mathbf{q}_t = [x_t, y_t]$ at the t-th time slot. We also assume that the UAV operates within a specific area $\mathcal{D}$, that is, $\mathbf{q}_t \in \mathcal{D}$, $\forall t$. Due to the limited velocity of the UAV, the velocity is subject to

$$\frac{\| \mathbf{q}_t - \mathbf{q}_{t-1} \|}{\tau} \leq \nu_{\max}, \tag{1}$$

where $\nu_{\max}$ represents the maximum velocity of the UAV.

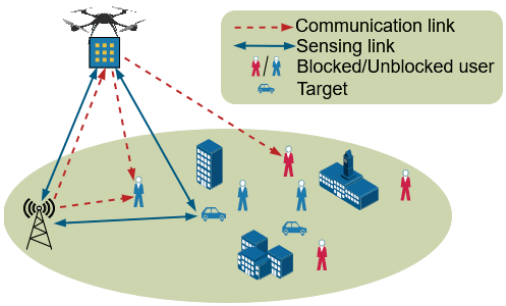

Fig. 1. The ARIS-assisted ISAC system.

The flight energy consumption of UAV at time slot t is modeled as [13]

$$e_t^{\mathrm{UAV}} = \tau \left(P_0 \left[1 + \frac{3(V_t^{\mathrm{xy}})^2}{U_{\mathrm{tip}}^2} \right] + P_1 \left(\sqrt{1 + \frac{(V_t^{\mathrm{xy}})^4}{4V_0^4}} - \frac{(V_t^{\mathrm{xy}})^2}{2V_0^2} \right)^{\frac{1}{2}} + C_0 (V_t^{\mathrm{xy}})^3 \right). \tag{2}$$

where $V_t^{\mathrm{xy}} = \frac{\|\mathbf{q}_t - \mathbf{q}_{t-1}\|}{\tau}$ is the velocity of the UAV.

A. Transmitting Signal

The signal transmitted by the ISAC BS at time slot t is

$$\mathbf{x}_t = \sum_{k=1}^{K} \mathbf{w}_{k,t}^c s_{k,t}^c + \sum_{j=1}^{J} \mathbf{w}_{j,t}^s s_{j,t}^s, \tag{3}$$

where $\mathbf{w}_{k,t}^c \in \mathbb{C}^{M \times 1}$ and $\mathbf{w}_{j,t}^s \in \mathbb{C}^{M \times 1}$ denote the beamforming vectors for communication with the user k and sensing with the target j at time slot t, respectively. The according data streams are $s_{k,t}^c$ and $s_{j,t}^s$.

Let $\mathbf{W}_t^c = \left[\mathbf{w}_{1,t}^c, \cdots, \mathbf{w}_{K,t}^c\right] \in \mathbb{C}^{M \times K}$, $\mathbf{W}_t^s = \left[\mathbf{w}_{1,t}^s, \cdots, \mathbf{w}_{J,t}^s\right] \in \mathbb{C}^{M \times J}$, $\mathbf{s}_t^c = \left[s_{1,t}^c, \cdots, s_{K,t}^c\right]^T \in \mathbb{C}^{K \times 1}$, and $\mathbf{s}_t^s = \left[s_{1,t}^s, \cdots, s_{J,t}^s\right]^T \in \mathbb{C}^{J \times 1}$. Thus, the equation (3) could be revised as

$$\mathbf{x}_t = \mathbf{W}_t^c \mathbf{s}_t^c + \mathbf{W}_t^s \mathbf{s}_t^s. \tag{4}$$

Note that we assume that the ISAC symbols satisfy $\mathbb{E}\{\mathbf{s}_t^c \mathbf{s}_t^{cH}\} = \mathbf{I}$ and $\mathbb{E}\{\mathbf{s}_t^s \mathbf{s}_t^{sH}\} = \mathbf{I}$ while $\mathbf{s}_t^c$ and $\mathbf{s}_t^s$ are statistically independent and uncorrelated, i.e., $\mathbb{E}\{\mathbf{s}_t^c \mathbf{s}_t^{sH}\} = \mathbf{0}$. In this paper, $(\cdot)^T$ and $(\cdot)^H$ denote transpose and conjugate transpose operations, respectively. $\mathbf{I}$ and $\mathbf{0}$ denote the identity matrix and all zero matrix, respectively.

B. Communication Model

The signal received by the blocked user $k \in \mathcal{K}_{\mathcal{B}}$ at time slot t is represented as

$$y_{k,t}^c = \mathbf{h}_{k,t}^c \mathbf{\Phi}_t \mathbf{G}_t \mathbf{x}_t + n_{k,t}, \tag{5}$$

where $\mathbf{h}_{k,t}^c \in \mathbb{C}^{1 \times N}$ is the channel coefficient from the ARIS to the blocked user k and $\mathbf{G}_t \in \mathbb{C}^{N \times M}$ denotes the channel of the BS-ARIS link. In addition, $\mathbf{\Phi}_t = \mathrm{diag}\left(e^{j\theta_{1,t}}, e^{j\theta_{2,t}}, \ldots, e^{j\theta_{N,t}}\right) \in \mathbb{C}^{N \times N}$ is the RIS phase matrix with $\theta_{i,t} \in [0, 2\pi), \forall i, t$ and $n_{k,t} \sim \mathcal{CN}(0, \sigma_k^2)$ is the noise.

The signal received by the unblocked user $k \in \mathcal{K}_{\mathcal{U}}$ at time slot t is represented as

$$y_{k,t}^{uc} = \left(\mathbf{h}_{k,t}^{uc} \mathbf{\Phi}_t \mathbf{G}_t + \mathbf{d}_{k,t}^{uc}\right) \mathbf{x}_t + n_{k,t}. \tag{6}$$

where $\mathbf{h}_{k,t}^{uc} \in \mathbb{C}^{1\times N}$ is the channel from the ARIS to the unblocked user k and $\mathbf{d}_{k,t}^{uc} \in \mathbb{C}^{1\times M}$ the channel from the BS to the unblocked user k.

The SINR of the user k at time slot t is expressed as (7) on the top of next page. Accordingly, the communication rate of the user k at slot t is

$$\Upsilon_{k,t} = \log_2(1 + \gamma_{k,t}^{com}). \tag{8}$$

C. Sensing Model

With the assumption that the interferences among all received signal are perfectly eliminated, the echo signal from the target j at time slot t is denoted as

$$\begin{aligned} \mathbf{y}_{j,t}^{s} =& \mathbf{V}_t (\mathbf{h}_{j,t}^{s} \mathbf{\Phi}_t \mathbf{G}_t + \mathbf{d}_{j,t}^{s})^H \\ & \left(\mathbf{h}_{j,t}^{s} \mathbf{\Phi}_t \mathbf{G}_t + \mathbf{d}_{j,t}^{s}\right) \mathbf{x}_t + \mathbf{V}_t \mathbf{n}_{j,t}, \end{aligned} \tag{9}$$

where $\mathbf{h}_{j,t}^{s} \in \mathbb{C}^{1\times N}$ is the channel from the ARIS to the target j, $\mathbf{d}_{j,t}^{s} \in \mathbb{C}^{1\times M}$ denotes the channel from the BS to the target j, $\mathbf{V}_t \in \mathbb{C}^{M\times M}$ is the receiving beamforming matrix, and $\mathbf{n}_{j,t} \in \mathbb{C}^{M\times 1}$ is the noise.

Thus, the signal-to-noise ratio (SNR) of target j at time slot t is

$$\gamma_{j,t}^{sen} = \frac{\left|\mathbf{V}_t {\mathbf{f}_{j,t}^{s}}^H \mathbf{f}_{j,t}^{s} \mathbf{w}_{j,t}^{s}\right|^2}{\left|\mathbf{V}_t \mathbf{n}_{j,t}\right|^2}. \tag{10}$$

where $\mathbf{f}_{j,t}^{s} = \mathbf{h}_{j,t}^{s} \mathbf{\Phi}_t \mathbf{G}_t + \mathbf{d}_{j,t}^{s} \in \mathbb{C}^{1\times M}$. In order to guarantee the sensing accuracy of each targets, radar SNR is constrained to

$$\gamma_{j,t}^{sen} \geq \gamma_{th}^{sen}, \forall j \in \mathcal{J}, \tag{11}$$

where γ_{th}^{sen} is the threshold of radar SNR.

D. Problem Formulation

The objective is to maximize the sum of the minimum rate within the whole ISAC period through the joint optimization of the UAV trajectory, RIS phase, and BS transmitting/receiving beamforming vectors under the constraint of ensuring the sensing accuracy of each target. Accordingly, the optimization problem is formulated as (12), where $P_{\max}^{BS}$ denotes the maximum BS transmitting power and $e_{\max}^{UAV}$ is the maximum UAV energy consumption. With respect to problem $(\mathrm{P1})$, the optimization variables are highly coupled. Besides, the unimodular constrain in (12f) is intractable. Consequently, problem $(\mathrm{P1})$ is difficult to solve by traditional optimization techniques. As an online DRL method, the proximal policy optimization (PPO) algorithm is adopted as the optimization technique in this paper.

$$(\mathrm{P1}): \max_{\{\mathbf{W},\mathbf{q},\mathbf{\Phi}_t,\mathbf{V}_t\}} \sum_{t=1}^{T} \min_k \Upsilon_{k,t} \tag{12}$$

$$\mathbf{s.t.} \gamma_{j,t}^{sen} \geq \gamma_{th}^{sen}, \forall j \tag{12a}$$

$$\mathrm{Tr}(\mathbf{W}_t^c \mathbf{W}_t^{c\mathrm{H}}) + \mathrm{Tr}(\mathbf{W}_t^s \mathbf{W}_t^{s\mathrm{H}}) \leqslant P_{\max}^{BS}, \forall t \tag{12b}$$

$$\sum_{t=1}^{T} e_t^{UAV}(\mathbf{q}) \leqslant e_{\max}^{UAV} \tag{12c}$$

$$\frac{\parallel \mathbf{q}_t - \mathbf{q}_{t-1} \parallel}{\tau} \leq \nu_{\max}, \forall t \tag{12d}$$

$$\mathbf{q}_t \in \mathcal{D}, \forall t \tag{12e}$$

$$|\mathbf{\Phi}_t| = 1, \forall t \tag{12f}$$

$$\|\mathbf{V}_t\| = 1, \forall t, \tag{12g}$$

III. Joint optimization via PPO-based approach

In this section, problem $(\mathrm{P1})$ is represented as the Markov decision process (MDP) and solved by the PPO-based algorithm based on the clipping function.

A. Problem Reformulation based on MDP

As well known, the goal in a MDP is to find the optimal policy, which is a strategy that prescribes the best action to take in each state to maximize the expected cumulative long-term reward. We first reformulate the primary optimization problem as a MDP. To be specific, the ISAC BS is regarded as the agent with the proposed ARIS-ISAC system being the environment. The agent interacts with the environment to generate a 3-turple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$ are the set of states, actions and immediate reward, respectively. And the detailed descriptions are as follows.

1) State Space $\mathcal{S}$: The state s_t at the slot t consists of the UAV position and the channel state information. For simplicity, the channel information is expressed as $\mathbf{H}_t$ which is equivalent to the set $\{\mathbf{d}_{k,t}^{uc}, \mathbf{h}_{k,t}^{uc}, \mathbf{h}_{k,t}^{c}, \mathbf{d}_{j,t}^{s}, \mathbf{h}_{j,t}^{s}, \mathbf{G}_t\}$. Since the complex variables are hard to be supported in the neural network, the channel information is expressed as the channel gain $|\mathbf{H}_t|^2$. The state s_t is given by

$$s_t = \left\{|\mathbf{H}_t|^2, \mathbf{q}_t\right\}. \tag{13}$$

2) Action Space $\mathcal{A}$: The agent takes an action a_t after observing the state $s_t \in \mathcal{S}$ of the environment. The action set a_t includes the transmit/receive beamforming vectors, the phase shift matrix and UAV displacement. The phase shift matrix is simplified as the phase angle Θ_t. The UAV displacement is denoted by the flight distance d_t as well as the flight angle θ_t. The action a_t taken at the state s_t is expressed by

$$a_t = \left\{\mathrm{Re}\left\{\mathbf{W}_t\right\}, \mathrm{Im}\left\{\mathbf{W}_t\right\}, \mathrm{Re}\left\{\mathbf{V}_t\right\}, \mathrm{Im}\left\{\mathbf{V}_t\right\}, \Theta_t, d_t, \theta_t\right\}. \tag{14}$$

3) Reward Function $\mathcal{R}$: The reward function is the instant reward. As a result, the reward function r_t is defined as

$$r_t = \min_k \Upsilon_{k,t}. \tag{15}$$

$$\gamma_{k,t}^{com}=\begin{cases}\dfrac{\left|\mathbf{h}_{k,t}^{c}\boldsymbol{\Phi}_t\mathbf{G}_t\mathbf{w}_{k,t}^{c}\right|^2}{\sum\limits_{j=1}^{J}\left|\mathbf{h}_{k,t}^{c}\boldsymbol{\Phi}_t\mathbf{G}_t\mathbf{w}_{j,t}^{s}\right|^2+\sum\limits_{i=1,i\neq k}^{K}\left|\mathbf{h}_{k,t}^{c}\boldsymbol{\Phi}_t\mathbf{G}_t\mathbf{w}_{i,t}^{c}\right|^2+\sigma_k^2} & k\in\mathcal{K}_{\mathcal{B}}\\ \dfrac{\left|\left(\mathbf{h}_{k,t}^{uc}\boldsymbol{\Phi}_t\mathbf{G}_t+\mathbf{d}_{k,t}^{uc}\right)\mathbf{w}_{k,t}^{c}\right|^2}{\sum\limits_{j=1}^{J}\left|\left(\mathbf{h}_{k,t}^{uc}\boldsymbol{\Phi}_t\mathbf{G}_t+\mathbf{d}_{k,t}^{uc}\right)\mathbf{w}_{j,t}^{s}\right|^2+\sum\limits_{i=1,i\neq k}^{K}\left|\left(\mathbf{h}_{k,t}^{uc}\boldsymbol{\Phi}_t\mathbf{G}_t+\mathbf{d}_{k,t}^{uc}\right)\mathbf{w}_{i,t}^{c}\right|^2+\sigma_k^2} & k\in\mathcal{K}_{\mathcal{U}}\end{cases} \tag{7}$$

However, the goal of deep reinforcement learning is to find the optimal policy for the sake of maximizing expected discounted reward. Thus, the discount factor γ is introduced and the cumulative discounted reward at the slot t is expressed as

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'}, \tag{16}$$

whose expectation is the value function $V(s_t)=\mathbb{E}[R_t]$.

B. PPO-based Algorithm

Based on the definition of the MDP problem, the framework of the PPO-based scheme is shown in Fig. 2. Different

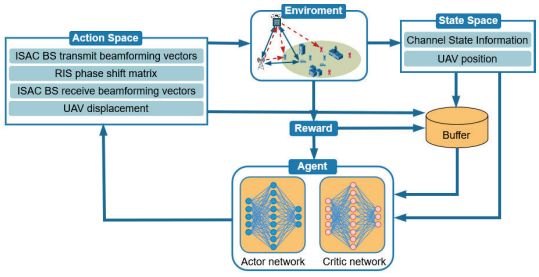

Fig. 2. The structure of the PPO-based algorithm.

from the policy gradient, PPO agent utilizes the old policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$ to interact with the environment to collect samples, while the new policy $\pi_{\boldsymbol{\theta}}$ is gained from training, which can improve sample efficiency [14]. The objective function is expressed as

$$\mathcal{J}(\boldsymbol{\theta})=\mathbb{E}_{\pi_{\boldsymbol{\theta}_{\text{old}}}}\left[\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|s_t)}A^{\boldsymbol{\theta}_{\text{old}}}(s_t,a_t)\right]. \tag{17}$$

The advantage function $A^{\boldsymbol{\theta}_{old}}(s_t,a_t)$ is a crucial metric to quantify whether an action is better or worse than the average action, which is given by

$$A^{\boldsymbol{\theta}_{old}}(s_t,a_t)=r_t+\gamma V^{\boldsymbol{\theta}_{\text{old}}}(s_{t+1})-V^{\boldsymbol{\theta}_{\text{old}}}(s_t). \tag{18}$$

To ensure the similarity between new and old policy, the trust region is introduced. Therefore, the objective function with the clipping function is denoted as

$$\mathcal{J}^{\text{clip}}(\boldsymbol{\theta})=\mathbb{E}_{\pi_{\boldsymbol{\theta}_{\text{old}}}}\left[\min\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|s_t)}A^{\boldsymbol{\theta}_{\text{old}}}(s_t,a_t),\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|s_t)},1-\varepsilon,1+\varepsilon\right)A^{\boldsymbol{\theta}_{\text{old}}}(s_t,a_t)\right)\right]. \tag{19}$$

where $\text{clip}(x,y,z)$ means that $x=y$ when $x<y$ and $x=z$ when $x>z$. The parameter $\boldsymbol{\theta}$ of the actor network is updated by

$$\boldsymbol{\theta}=\arg\max_{\boldsymbol{\theta}}\mathcal{J}^{\text{clip}}(\boldsymbol{\theta}). \tag{20}$$

The critic network is parameterized as $\boldsymbol{\phi}$ which is updated by

$$\boldsymbol{\phi}=\arg\min_{\boldsymbol{\phi}}\frac{1}{2}\left[V^{\boldsymbol{\theta}}(s_t)-R_t\right]^2. \tag{21}$$

The PPO-based approach is summarized in Algorithm 1.

Algorithm 1 The PPO-Based Algorithm

Input: Channels $|\mathbf{H}_t|^2$ and UAV positions $\mathbf{q}_t$
Output: Actions $\text{Re}\{\mathbf{W}_t\}$, $\text{Im}\{\mathbf{W}_t\}$, $\text{Re}\{\mathbf{V}_t\}$, $\text{Im}\{\mathbf{V}_t\}$, Θ_t, d_t, θ_t
Initialize: Initialize s_0, $\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{old}}$ and $\boldsymbol{\phi}$

```
for each episode do
    Receive initial state s_0
    for step = 1, 2, ..., L do
        Take the action a_t by the actor network θ_old
        Store the transition into the experience pool
        if arrives mini-batch size B then
            Compute the advantage function
            for epoch = 1, 2, ..., m do
                Update θ, φ, θ_old
            end for
        end if
        if UAV flies out of the area D then
            break
        end if
    end for
end for
```

IV. Simulation Results

We set the simulation parameters and provide comparisons and then present simulations to verify the effectiveness of PPO-based algorithm.

A. Simulation Scenario and Parameter Setting

In our simulation setup, it is assumed that 4 users (including 3 unblocked users and 1 blocked user) and 2 targets are dispersed. The UAV flies from the initial point (50 m, 30 m) at the fixed height 50 m in the given area 100×100 m^2 and the BS is at (30 m, 100 m) on the ground. The maximum velocity of the UAV is 10 m/s. There are two hidden layers whose sizes are 64×64 in the neural network.

The following four baseline schemes are considered.

i) Fixed trajectory: The UAV flies at the fixed speed 6 m/s and flies towards x-axis before 10 steps as well as towards y-axis after 10 steps as shown in Fig. 4.

ii) Maximum ratio transmission (MRT): The MRT method precodes the signal in the direction of the channel vector, therefore, the MRT precoding vector to the k-th user is given by $\mathbf{w}_{k,t}^{c} = \frac{\mathbf{f}_{k,t}^{H}}{\|\mathbf{f}_{k,t}\|}$, where $\mathbf{f}_{k,t}$ is the whole channel of the BS-user link and BS-ARIS-user link. It is assumed that the targets are regarded as the users and the MRT precoding vector to the j-th target is $\mathbf{w}_{j,t}^{s} = \frac{\mathbf{f}_{j,t}^{H}}{\|\mathbf{f}_{j,t}\|}$, where $\mathbf{f}_{j,t}$ is the whole channel of the BS-target link and BS-ARIS-target link.

iii) Random RIS: The phase shift is chosen from $(0, 2\pi]$ uniformly.

iv) Fixed RIS: A fixed phase shift is adopted.

For simplicity, it is assumed that the channel characteristics are invariant within each time slot and the channel is modeled as Rician channel. According to distance-dependent loss model, the path loss is

$$C = C_0 \left(\frac{d}{d_0}\right)^{-\alpha}, \tag{22}$$

where $C_0 = -30$ dB is the reference path loss at $d_0 = 1$ m and α is the path loss exponent. The channel is modeled as a Rician channel given by

$$\mathbf{H} = \sqrt{C}\left(\sqrt{\frac{\beta}{1+\beta}}\mathbf{H}^{LoS} + \sqrt{\frac{1}{1+\beta}}\mathbf{H}^{NLoS}\right), \tag{23}$$

where $\mathbf{H}^{LoS}$ and $\mathbf{H}^{NLoS}$ are LoS component and NLoS components as well as β is the Rician factor. Generally speaking, the NLoS components are modeled as Rayleigh fading and the LoS components are gained from the steering vector. The channel setups are as follows. The path loss exponents and Rician factors of the BS-ARIS, BS-user, BS-target, ARIS-user and ARIS-target link are set as $\alpha_{BA} = 2$, $\alpha_{BU} = 3$, $\alpha_{BT} = 3$, $\alpha_{AU} = 2.8$, $\alpha_{AT} = 2.8$, $\beta_{BA} = \infty$, $\beta_{BU} = 0$, $\beta_{BT} = 0$, $\beta_{AU} = 10$, and $\beta_{AT} = 10$, respectively. The key simulation parameter settings are shown in I.

B. Performance of Proposed Algorithm and Baseline

We first plot the smooth reward of different schemes when $M = 2, N = 16$ and $P_{\max}^{BS} = 10$ W as shown in Fig. 3. It is shown that at the beginning of the training, the reward is little even negative value. This is because some wrong actions make the constrains not satisfied at the beginning of the training, which generates the penalty to the reward. Then it is observed that the proposed algorithm outperforms the other benchmarks. This is due to the fact that the proposed algorithm jointly optimizes the UAV trajectory, RIS phase shifts and beamforming. In addition, all schemes convergence at the stable value.

The UAV trajectory is depicted in Fig. 4 with $M = 2, N = 16$ and $P_{\max}^{BS} = 10$ W. It can be observed that the UAV trends to fly closer to the blocked user. For the reason

Table I: Simulation Parameter Settings

Parameters	Values
Max training steps	800000
Maximum steps per episode L	20
Mini-batch size B	20
Discount factor γ	0.99
Smooth factor λ	0.95
Clipping factor ϵ	0.2
Learning rate lr	10^{-4}
Communication noise	-160 dBm
Sensing noise	-140 dBm
Max energy comsumption of UAV	39 dB
Sensing threshold	10 dB
Aerodynamics parameters	80 W, 31.43 W, 120 m/s
P_0, P_1, U_{tip}, V_0, and C_0	40 m/s and 0.0046 kg/m

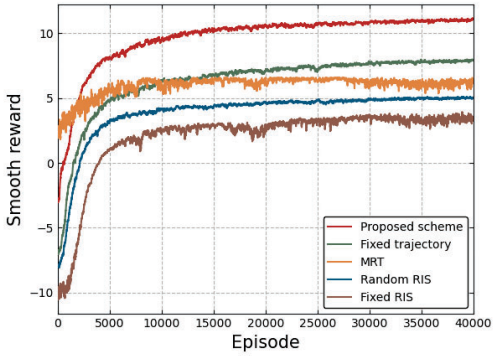

Fig. 3. The convergence of different scheme. (The smooth episode reward is obtained by averaging the episode reward curve using a sliding window of size 60.)

that the achievable rate of the blocked user is generally less than the unblocked user without a direct link. The movement of the UAV guarantees the performance of the worst user to maximize the sum of the minimum achievable rate.

The rate versus the transmitting power is compared as shown in Fig. 5 when $M = 2$ and $N = 16$. It is expected that the rate of all schemes increases with the increasing

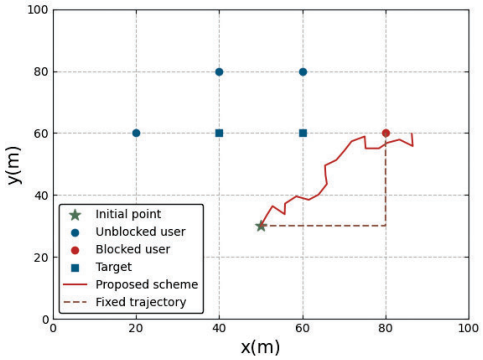

Fig. 4. The 2D UAV-trajectory.

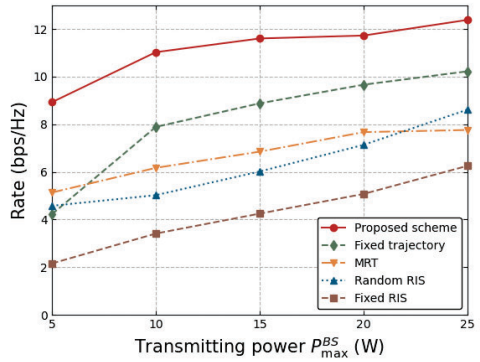

Fig. 5. The rate of the different schemes versus the transmitting power $P_{\max}^{BS}$.

power. In addition, the proposed scheme outperforms in rate maximization compared to other benchmark schemes, which shows the effectiveness of the joint optimization of trajectory and beamforming design. Then the fixed-trajectory scheme is observed to achieve higher rate than other scheme without beamforming optimization when the transmission power is large, which indicates its importance.

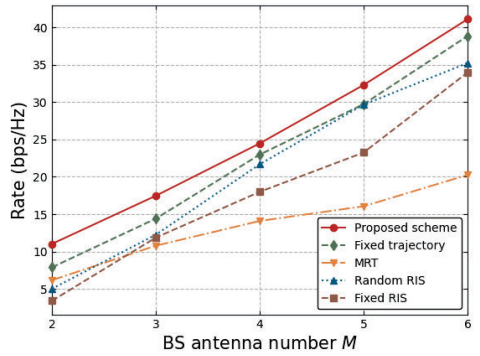

Fig. 6. The rate of the different schemes versus the BS antenna numbers M.

In Fig. 6, we further compare the performance of each solution in terms of BS antenna numbers when $N = 16$ and $P_{\max}^{BS} = 10$ W. It can be observed that the rate of all solutions improves along the increasing BS antenna numbers. This is because more antennas can send signal energy more concentrated to the specific user and gains higher beamforming gain.

V. Conclusions

We proposed a novel ARIS-ISAC system, where the ARIS was deployed to serve the communication users and assist radar sensing. A joint optimization problem over beamforming, RIS phase and UAV trajectory was formulated with the aim of maximizing the sum of the minimum achievable rate during the UAV flight. We introduced the PPO-based algorithm for the problem. Simulation results were presented to demonstrate the effectiveness of the PPO-based algorithm and the ARIS could ensure communication performance of the users via the movement of the UAV.

Acknowledgement

This work was supported in part by National Key Research and Development Program of China under Grant 2023YFE0107900, in part by the National Natural Science Foundation of China under Grant 62071306, and in part by the Shenzhen Science and Technology Program under Grants JCYJ20200109113601723, JSGG20210420091805014, and JSGG20210802154203011.

References

[1] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Trans. Wireless Commun.*, vol. 68, no. 6, pp. 3834-3862, Jun. 2020.

[2] Z. Wei, H. Qu, Y. Wang, X. Yuan, H. Wu, Y. Du, K. Han, N. Zhang, and Z. Feng, "Integrated sensing and communication signals towards 5G-A and 6G: A survey," *IEEE IoT J.*, vol. 10, no. 13, pp. 11068-11092, Jul. 2023.

[3] H. Zhang, B. Di, K. Bian, Z. Han, H. V. Poor, and L. Song, "Toward ubiquitous sensing and localization with reconfigurable intelligent surfaces," *Proc. IEEE*, vol. 110, no. 9, pp. 1401-1422, Sept. 2022.

[4] M. Rihan, A. Zappone and S. Buzzi, "Robust RIS-Assisted MIMO Communication-Radar Coexistence: Joint Beamforming and Waveform Design," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6647-6661, Nov. 2023.

[5] X. Song, J. Xu, F. Liu, T. X. Han, and Y. C. Eldar, "Intelligent Reflecting Surface Enabled Sensing: Cramr-Rao Bound Optimization," *IEEE Trans. Signal Process.*, vol. 71, pp. 2011-2026, 2023.

[6] G. Cheng, Y. Fang, J. Xu, and D. W. K. Ng, "Optimal Coordinated Transmit Beamforming for Networked Integrated Sensing and Communications," *IEEE Trans. Wireless Commun.*, 2024.

[7] J. Mu, R. Zhang, Y. Cui, N. Gao, and X. Jing, "UAV Meets Integrated Sensing and Communication: Challenges and Future Directions," *IEEE Commun. Mag.*, vol. 61, no. 5, pp. 62-67, May. 2023.

[8] Z. Lyu, G. Zhu and J. Xu, "Joint Maneuver and Beamforming Design for UAV-Enabled Integrated Sensing and Communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2424-2440, Apr. 2023.

[9] J. Wu, W. Yuan and L. Bai, "On the Interplay Between Sensing and Communications for UAV Trajectory Design," *IEEE IoT J.*, vol. 10, no. 23, pp. 20383-20395, Dec. 2023.

[10] W. Wei, X. Pang, J. Tang, N. Zhao, X. Wang, and A. Nallanathan, "Secure Transmission Design for Aerial IRS Assisted Wireless Networks," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3528-3540, Jun. 2023.

[11] B. Duo, M. He, Q. Wu, and Z. Zhang, "Joint Dual-UAV Trajectory and RIS Design for ARIS-Assisted Aerial Computing in IoT," *IEEE IoT J.*, vol. 10, no. 22, pp. 19584-19594, Nov. 2023.

[12] Y. Qin, Z. Zhang, X. Li, W. Huangfu, and H. Zhang, "Deep Reinforcement Learning Based Resource Allocation and Trajectory Planning in Integrated Sensing and Communications UAV Network," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8158-8169, Nov. 2023.

[13] C. Deng, X. Fang and X. Wang, "Beamforming Design and Trajectory Optimization for UAV-Empowered Adaptable Integrated Sensing and Communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8512-8526, Nov. 2023.

[14] X. Liu, H. Zhang, K. Long, M. Zhou, Y. Li, and H. V. Poor, "Proximal policy optimization-based transmit beamforming and phase-shift design in an IRS-aided ISAC system for the THz band," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2056-2069, Jul. 2022.