Empowering ISAC Systems with Federated Learning: A Focus on Satellite and RIS-Enhanced Terrestrial Integrated Networks

Sonia Pala, *Member, IEEE*, Keshav Singh, *Member, IEEE*, Chih-Peng Li, *Fellow, IEEE,* and Octavia A. Dobre, *Fellow, IEEE*

***Abstract*—This paper presents a state-of-the-art analytical framework aimed to enhance spectral efficiency in satellite and terrestrial integrated networks (STINs), utilizing reconfigurable intelligent surface (RIS) within the realm of integrated sensing and communication (ISAC). Our methodology pivots on a pioneering federated deep reinforcement learning strategy that introduces new ground beyond conventional optimization techniques to tackle the intricate problem of non-convex resource allocation. The approach leverages federated learning to dynamically adapt to network changes, enabling efficient resource management and ensuring compliance with beamforming designs, multiple target signal-to-interference-plus-noise ratio thresholds, and RIS phase-shift requirements through an effective feedback loop. In particular, we propose a federated deep deterministic policy gradient (F-DDPG) algorithm across multi-agent systems that outperforms existing federated deep Q-network (F-DQN), centralized, and traditional DDPG and DQN methods. The empirical findings underscore the efficiency of the federated algorithms, which closely align with the performance of centralized models while markedly reducing execution time, thus achieving an optimal synergy between operational efficiency and system performance. Simulation results highlight the remarkable advantages of optimal RIS configurations, showcasing a performance increase of 54.2% over random RIS setups and a remarkable 76.8% enhancement compared to scenarios without RIS, underscoring the transformative impact of our federated learning approach. Additionally, our study evaluates the impact of channel estimation errors and interference, confirming the robustness of our approach and its potential to optimize ISAC-enabled STINs.**

***Index Terms*—Integrated sensing and communication (ISAC), satellite-terrestrial integrated network (STIN), reconfigurable intelligent surfaces (RIS), beamforming design, federated learning, deep reinforcement learning (DRL).**

I. Introduction

SENSING technologies are pivotal for advancing the capabilities of future wireless networks, such as those anticipated in 6G, facilitating applications like support for navigation, detecting activities, tracking movements, and monitoring environmental conditions [1], [2]. With the rapid advancements in both wireless communication and radar sensing fields, the demand for spectrum resources increasingly exceeds supply, underscoring their scarcity and value. To address these challenges, the concept of integrated sensing and communication (ISAC) emerges as a strategic approach to efficiently harness spectral, hardware, and energy resources. ISAC achieves this by employing a unified approach to signal processing and leveraging a common hardware infrastructure for both sensing and communication tasks. Despite its potential, the majority of ISAC research has been concentrated on terrestrial applications, which narrows its capability for offering services on a global scale [3].

Satellite communication networks offer a viable solution to overcoming coverage limitations due to their wide coverage capability [4]. In earlier years, the deployment of large-scale satellite networks was hindered by substantial setup costs and a comparative deficiency in capacity against ground-based networks. Nevertheless, the growing demands for global communication services and breakthroughs in technology have recently made the concept of satellite constellations a focal point of interest in both the academic realm and the industry. This shift led to the initiation of various satellite constellation projects aimed at furnishing global coverage, highlighting projects such as Starlink, Telesat, and OneWeb [5]. Integrating satellite with terrestrial networks presents a viable strategy for achieving widespread broadband access, leveraging the extensive coverage of satellite systems alongside the high-capacity infrastructure of terrestrial networks [6]. Notably, the 3rd Generation Partnership Project (3GPP) has explored this synergistic approach in Releases 15, 16, and 17, examining the integration of terrestrial and non-terrestrial networks to extend network services to underserved areas, enhance service continuity, and optimize multicast/broadcast communications [7], [8]. According to the vision for future wireless networks outlined in the 6G White paper, the seamless integration between satellite and terrestrial networks is crucial, signaling a pivotal shift towards a more inclusive and versatile communication network infrastructure [9], [10].

A. Related Works

In recent years, satellite-terrestrial integrated networks (STIN) have gained significant attention for their potential to

The work of K. Singh and C.-P. Li was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 113-2218-E-110-009, Grant NSTC 112-2221-E-110-029-MY3, and in part by the Sixth Generation Communication and Sensing Research Center funded by the Higher Education SPROUT Project, the Ministry of Education of Taiwan. The work of O. A. Dobre was supported in part by the Canada Research Chairs Program-Canada Research Chair CRC-2022-00187. (*Corresponding author: Keshav Singh*).

Sonia Pala, Keshav Singh, and Chih-Peng Li are with Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan (Email: sony.pj12@gmail.com, keshav.singh@mail.nsysu.edu.tw, cpli@faculty.nsysu.edu.tw).

O. A. Dobre is with Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada (E-mail: odobre@mun.ca).

enhance spectral efficiency by merging terrestrial and satellite systems [11], [12]. This integration has spurred numerous studies focusing on STIN functionalities, particularly in resource management [13], [14], cooperative coordination [15], [16], and beamforming techniques. Beamforming is especially noteworthy for improving signal quality for target users while minimizing interference, thus enabling effective coexistence of satellite and terrestrial networks [17], [18]. In [19], a hybrid analog-digital beamforming design was proposed for spectrum-sharing, introducing an optimized approach. Similarly, [20] developed a beamforming strategy to enhance the SINR for terrestrial users while reducing interference for satellite users. The work in [21] furthered this by proposing beamforming methods to maximize cellular data rates under satellite interference constraints. This research also extended to joint beamforming strategies for secure communication in scenarios with multiple satellite users and potential eavesdroppers [22].

On the other hand, a significant body of research has explored the concept of ISAC within terrestrial network frameworks. The adoption of multiple-input multiple-output (MIMO) technology within terrestrial ISAC systems has been leveraged to notably improve the spectral and energy efficiencies of communication systems, a development thoroughly investigated in [23]. Moreover, the research outlined in [24]–[27] focused on the development of hybrid beamforming methods, employing diverse MIMO radar techniques to boost the performance of dual-functional radar communication systems. Recent advancements have adopted orthogonal frequency division multiplexing in terrestrial ISAC systems to mitigate inter-symbol interference, enhancing target sensing capabilities [28]. However, these systems are limited by regional coverage and data processing constraints. To overcome these, integrating ISAC with low Earth orbit (LEO) satellites has been explored, supported by progress in satellite onboard processing. A strategy for hybrid beamforming in ISAC-LEO systems is detailed in [29], which also considers the impact of beam squint. Further, a novel ISAC-aided dynamic resource allocation strategy that enhances random access efficiency and system throughput within satellite-terrestrial relay networks is detailed in [30]. Even though ISAC leverages the larger bandwidths of millimeter wave frequencies for enhanced data rates and radar resolution, the higher frequencies introduce significant signal blockage issues, adversely affecting performance. To mitigate this, reconfigurable intelligent surfaces (RISs) can establish effective virtual connections between the ISAC base station (BS) or satellite and sensing targets, offering a promising strategy to overcome the challenge [31].

Recent studies have extensively explored the synergy between RIS and ISAC systems [32], [33], highlighting two predominant approaches in RIS-enhanced ISAC systems [34]. The first approach utilized RIS primarily to enhance communication capabilities while maintaining direct links from the transceiver to the target for sensing purposes [35], [36]. Specifically, the authors in [35] explored designing both transmit and receive beamforming strategies alongside RIS phase adjustments for multi-user settings. Conversely, [36] focused on reducing the transmit power of dual-function radar-communication (DFRC) BS by concurrently optimizing both active and passive beamforming in light of RIS-induced interference. Leveraging the advantages of RIS, the authors in [37] adopted deep reinforcement learning (DRL) to explore the integration of RIS within satellite networks, presenting promising solutions to latency, dynamic channel conditions, and energy constraints in 6G Internet of Things (IoT) environments. In [38], research focused on optimizing beamforming for RIS-enhanced hybrid satellite-terrestrial networks with blocked satellite and BS-user links.

B. Motivation

The potential of ISAC in revolutionizing satellite and RIS-enhanced terrestrial networks is vast, yet a thorough examination of its full potential remains unexplored. Research carried out in [11]–[22] focused on the nuances of STINs without delving into the integration of ISAC or the innovative use of RIS. Further, while studies [24]–[28] have investigated ISAC within terrestrial contexts, their scope does not extend to achieving global coverage or overcoming the data processing and reception challenges inherent to terrestrial networks. Efforts to incorporate ISAC within satellite frameworks, as carried out in [29], [30], have not considered the integration with RIS. Although the synergy between RIS and ISAC was examined in [32]–[36], such research remained limited to terrestrial implementations. However, the works in [37], [38] explored the benefits of RIS in satellite and STINs, yet overlooked ISAC applications. Given these gaps, the challenge of optimal resource allocation within ISAC in satellite and RIS-enhanced terrestrial networks emerges, underscoring the complexity of such an endeavor where traditional optimization methods might not suffice. Our motivation for choosing federated deep deterministic policy gradient (F-DDPG) and federated deep Q-network (F-DQN) is to compare policy-based and value-based learning paradigms in a federated setting. To the best of the authors' knowledge, a comprehensive study leveraging federated learning for ISAC in satellite and RIS-enhanced terrestrial networks has not yet been significantly pursued in the existing literature.

C. Contribution

Motivated by the aforementioned discussion, we introduce a novel analytical framework to evaluate the performance of the ISAC-enabled satellite and RIS-enhanced terrestrial integrated network and address the spectral-efficiency optimization problem through a federated learning approach. The key contributions are outlined as follows:

- We develop an intricate framework that seamlessly combines ISAC functionalities across both satellite and terrestrial domains, enhanced by RIS technology. This framework is meticulously designed to enhance the sum rate for all satellite users (SUs) and cellular users (CUs), while simultaneously guaranteeing the sensing performance through a minimum SINR requirement at the sensing targets.
- We tackle the non-convex resource allocation problem through a federated DRL (F-DRL) approach. This method

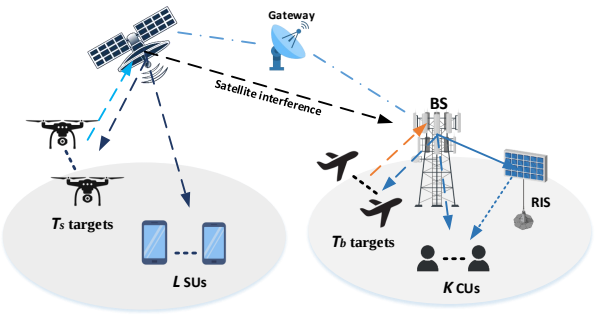

Fig. 1: Illustration of ISAC enabled satellite and RIS-enhanced terrestrial integrated network.

simplifies optimization by dynamically adapting to real-time changes and employing multi-agent reinforcement learning for effective resource management in both satellite and terrestrial segments. It enables agents to fine-tune transmit power and adhere to SINR and RIS phase-shift requirements, enhancing network performance with a strategic feedback mechanism.

- The proposed F-DDPG algorithm outperforms benchmarks such as F-DQN, as well as centralized and traditional DDPG and DQN approaches. Integrating RIS within our framework significantly boosts overall network performance. Our results demonstrate that federated algorithms perform comparably to centralized models while reducing execution time, thus providing an optimal balance between efficiency and performance. Additionally, we have analyzed the impact of channel state information (CSI) errors and interference, further validating the robustness of our approach.

D. Structure of the Paper

Section II outlines the system model, Section III formulates the optimization problem, and Section IV introduces the proposed federated multi-agent (MA) DRL (FMA-DRL) modeled by the MDP framework. Numerical simulations are discussed in Section V, followed by conclusion remarks in Section VI. Table I outlines the key notations employed throughout this paper.

TABLE I: Key notations.

Notation	Definition	Notation	Definition
$(\mathbf{A})^{\mathrm{H}}$	Hermitian	$(\mathbf{A})^{\mathrm{T}}$	Transpose
$\mathrm{diag}(\cdot)$	Diagonalization operator	$\lvert\mathbf{A}\rvert$	Modulus operator
$\mathbb{C}^{M\times M}$	Complex matrix	$\mathbb{C}^{M\times 1}$	Complex vector
$\lvert\lvert\mathbf{A}\rvert\rvert$	Norm operator	$\mathbf{A}^*$	Optimal $\mathbf{A}$

II. System Model

We investigate a RIS-aided downlink (DL) ISAC system utilizing a STIN, where an ISAC-geostationary orbit (GEO)[1] satellite is equipped with N_t antennas for transmission and N_r antennas for reception in a monostatic configuration. This system serves L single-antenna satellite users (SUs), denoted by the set $\mathcal{L} = \{1, ..., L\}$, and detects multiple targets, represented by the set $\mathcal{T}_s = \{1, ..., T_s\}$. All antennas use uniform linear arrays (ULAs) with half-wavelength spacing.

Expanding to the terrestrial segment, a BS with dual ULAs handles both communication and sensing. The BS directs signals to K single-antenna DL cellular users (CUs) ($\mathcal{K} = \{1, ..., K\}$) via a RIS with N elements, providing both direct and RIS-reflected connections. The BS, with an M_t-antenna ULA, facilitates communication within a shared frequency band, simultaneously transmitting data to K CUs and detecting multiple terrestrial targets indexed by $\mathcal{T}_b = \{1, ..., T_b\}$. Echo signals of these targets are captured by an M_r-element receiving ULA. This design ensures that the terrestrial base station optimally controls the RIS for cellular communication, while the satellite system effectively serves users and covers targets outside terrestrial BS coverage, thereby enhancing the overall system efficiency and effectiveness. Dividing into two sub-ISAC systems optimizes performance and resource allocation for communication and sensing tasks in their respective zones.

A. Satellite Channel Model

The communication channel from the ISAC-GEO satellite to the l^{th} SU encompasses free space loss, the radiation pattern of the antenna, and rain attenuation, resulting in the modeling of the DL channel as [41]

$$\mathbf{g}_l = \mathbf{b}_l \circ \mathbf{q}_l \circ \exp\{j\boldsymbol{\varphi}_l\}, \tag{1}$$

where $\mathbf{b}_l = [b_{l,1}, b_{l,2}, ..., b_{l,N_t}] \in \mathbb{C}^{N_t}$ encapsulates both the radiation pattern of the satellite beam and the losses due to free space. The approximation for the n_t^{th} entry in b_k is as follows

$$b_{l,n_t} = \frac{\sqrt{F_g F_{l,n_t}}}{4\pi \frac{d_l}{\lambda}\sqrt{\kappa T_{sys} B_g}}, \tag{2}$$

where F_g denotes the gain of the antenna at the user's location, d_l represents the distance from the satellite to the l^{th} user, λ signifies the wavelength of the carrier signal, κ stands for Boltzmann's constant, T_{sys} indicates the temperature associated with receive noise, and B_g defines the bandwidth. The beam gain from the n_t^{th} feed to the l^{th} user, denoted as F_{l,n_t}, can be estimated by

$$F_{l,n_t} = F_{max}\left[\frac{J_1(u_{l,n_t})}{2u_{l,n_t}} + 36\frac{J_3(u_{l,n_t})}{u_{l,n_t}^3}\right]^2, \tag{3}$$

where F_{max} is the maximum beam gain calculated using $u_{l,n_t} = 2.07123\sin(\theta_{l,n_t})/\sin(\theta_{3\ dB})$, where θ_{l,n_t} is the angle to the n_t^{th} beam's center, and $\theta_{3\ dB}$ is where the gain falls by 3 dB. J_1 and J_3 are the first and third-order Bessel functions, respectively. Rain attenuation coefficients $q_l = [q_{l,1}, q_{l,2}, ..., q_{l,N_t}] \in \mathbb{C}^{N_t}$ with each $q_{l,n_t} = \sqrt{\xi_{l,n_t}}$ and $\xi_{l,n_t}(\mathrm{dB})$ following a log-normal distribution[2] $\ln(\xi_{l,n_t}(\mathrm{dB}))\sim$

[1]The stationary nature and continuous coverage of GEO satellites simplify the channel estimation process and enhance the stability of the system, making them more suitable for our application compared to the rapidly moving LEO satellites [39], [40].

[2]The log-normal distribution, chosen based on empirical evidence and theoretical justification, realistically represents satellite channel behavior under varying rain conditions, crucial for accurate satellite communication system simulations [42].

$\mathcal{N}(\mu,\sigma)$. Further, the phase vectors $\boldsymbol{\varphi}_l=[\varphi_{l,1},\varphi_{l,2},..,\varphi_{l,N_t}]\in \mathbb{C}^{N_t}$ are uniformly distributed as $\varphi_{l,n_t}\sim\mathcal{U}(0,2\pi)$, and $\mathbf{G}=[\mathbf{g}_1,...,\mathbf{g}_L]\in\mathbb{C}^{N_t\times L}$ represents satellite channels to SUs.

B. Radar and Satellite Communication Model:

Initially, we focus on the DL transmission in the ISAC-satellite system using a narrowband ISAC signal $\mathbf{x}^{sat}\in\mathbb{C}^{M_t\times 1}$ for radar sensing and multi-user communication, leveraging multi-antenna beamforming with multiple targets given as

$$\mathbf{x}^{sat}=\sum\nolimits_{l=1}^{L}\mathbf{w}_l s_l^{sat}+\sum\nolimits_{t_s=1}^{T_s}\mathbf{s}_{t_s}^{sat}, \tag{4}$$

where $\mathbf{w}_l\in\mathbb{C}^{N_t\times 1}$ is the beamformer vector and $s_l^{sat}\in\mathbb{C}$ is unit power data symbol normalized to $\mathbb{E}\{|s_l^{sat}|^2\}$=1. Additionally, $\mathbf{s}_{t_s}^{sat}\in\mathbb{C}^{N_t\times 1}$ is the unique radar signal aimed at target t_s, characterized by its covariance matrix $\mathbf{W}_{t_s}\triangleq\mathbb{E}\{\mathbf{s}_{t_s}^{sat}\mathbf{s}_{t_s}^{sat\ H}\}$ [43]. The transmit signal $\mathbf{x}^{sat}$ enhances sensing accuracy by optimizing degrees of freedom (DoF) through $\mathbf{W}_{t_s}$, adhering to the power constraint $\sum_{l=1}^{L}\|\mathbf{w}_l\|^2+\sum_{t_s=1}^{T_s}\mathrm{Tr}(\mathbf{W}_{t_s})\le P_{\max}^{sat}$, the peak satellite power budget.

Subsequently, we formulate the echo signal for the targets assuming LoS radar channels. The transmit and receive ULAs at the satellite are spaced at half wavelength intervals using steering vectors $\mathbf{c}_t(\psi)\triangleq 1/\sqrt{N_t}[1,e^{j\pi\sin(\psi)},...,e^{j\pi(N_t-1)\sin(\psi)}]^T$ and $\mathbf{c}_r(\psi)\triangleq 1/\sqrt{N_r}[1,e^{j\pi\sin(\psi)},...,e^{j\pi(N_r-1)\sin(\psi)}]^T$ towards the direction ψ, respectively. Assuming that the target t_s positioned at an angle ψ_{t_s}, the reflection from the target is given as $\beta_{t_s}\mathbf{C}(\psi_{t_s})\mathbf{x}^{sat}$. Here, $\beta_{t_s}\in\mathbb{C}$ represents the complex amplitude of the target, which is primarily influenced by factors like path loss and radar cross section and $\mathbf{C}(\psi_{t_s})\triangleq\mathbf{c}_r(\psi_{t_s})\mathbf{c}_t^H(\psi_{t_s})$. We make the assumption that ψ_{t_s} and β_{t_s} of the target are known or previously estimated at the satellite and thus with the given target echo, the received signal at the satellite can be expressed as

$$\mathbf{y}_{t_s}^{sat}=\underbrace{\mathbf{H}_{t_s}\mathbf{x}^{sat}}_{\text{Target reflection}}+\underbrace{\sum\nolimits_{t'\neq t_s,t'=1}^{T_s}\mathbf{H}_{t'}\mathbf{x}^{sat}}_{\text{Echo signal of interferer targets}}+\tilde{\mathbf{n}}_{tar},\forall t_s, \tag{5}$$

where $\mathbf{H}_{t_s}(t)=\beta_{t_s}\mathbf{c}_r(\psi_{t_s})\mathbf{c}_t^H(\psi_{t_s})$ is the radar channel and $\tilde{\mathbf{n}}_{tar}\in\mathbb{C}^{N_r\times 1}$ indicates the additive white Gaussian noise (AWGN) with covariance $\tilde{\sigma}_{tar}^2\mathbf{I}_{N_r}$. In practice, a receive beamformer $\tilde{\mathbf{U}}=[\tilde{\mathbf{u}}_1,..,\tilde{\mathbf{u}}_{t_s}]\in\mathbb{C}^{N_r\times t_s}$ captures the desired reflected signal of the target t_s from $\mathbf{y}_{t_s}^{sat}$. Using this information, the SINR of the target t_s is expressed as

$$\gamma_{t_s}^{\text{tar}}=\frac{\mathbb{E}\{|\tilde{\mathbf{u}}_{t_s}^H\beta_{t_s}\mathbf{C}(\psi_{t_s})\mathbf{x}^{sat}|^2\}}{\mathbb{E}\{|\tilde{\mathbf{u}}_{t_s}^H\mathbf{D}\mathbf{x}^{sat}|^2\}+\mathbb{E}\{|\tilde{\mathbf{u}}_{t_s}^H\tilde{\mathbf{n}}_{tar}|^2\}},\forall t_s, \tag{6}$$

where $\mathbf{D}=\sum_{t'\neq t_s}\beta_{t'}\mathbf{C}(\psi_{t'})$. Next, we represent the communication channel between the l^{th} DL SU and the satellite as $\mathbf{g}_l\in\mathbb{C}^{1\times N_t}$. Further, we assume that the SUs located outside the terrestrial service area may experience mutual interference from other SUs but no interference from the BS [44]. The dominant interference from GEO satellites to terrestrial networks is due to higher transmission power, line-of-sight propagation, and stronger signal strength, justifying our focus on mitigating this interference in our proposed system [45]. Thus the received signal at each DL SU as

$$y_l^{sat}=\underbrace{\mathbf{g}_l\mathbf{w}_l s_l^{sat}}_{\text{Desired signal}}+\underbrace{\sum\nolimits_{l'=1,l'\neq l}^{L}\mathbf{g}_l\mathbf{w}_{l'}s_{l'}^{sat}}_{\text{Multi SU interference}}+\underbrace{\sum\nolimits_{t_s=1}^{T_s}\mathbf{g}_l\mathbf{s}_{t_s}^{sat}}_{\text{Interfering sensing signal}}+n_l^{sat},\forall l, \tag{7}$$

where n_l^{sat} indicates the AWGN with $\sigma_l^{sat^2}$ variance. Based on this, the SINR of the DL l^{th} SU can be formulated by referring to (7) as

$$\gamma_l^{sat}=\frac{|\mathbf{g}_l\mathbf{w}_l|^2}{\sum_{l'=1,l'\neq l}^{L}|\mathbf{g}_l\mathbf{w}_{l'}|^2+\sum_{t_s=1}^{T_s}\mathbf{g}_l\mathbf{W}_{t_s}\mathbf{g}_l^H+\sigma_l^{sat^2}},\forall l. \tag{8}$$

C. Radar and Terrestrial Communication Model:

Given the similarity to radar and satellite communication models, we have opted to omit detailed explanations. The terrestrial node controls the RIS to optimize performance and resource allocation within its coverage area, leveraging the satellite's clear line-of-sight (LoS) to minimize signal attenuation and interference. This design ensures efficient system operation and reduces interference from terrestrial obstacles [46]. The downlink (DL) transmission employs a narrowband ISAC signal, $\mathbf{x}^{bs}\in\mathbb{C}^{M_t\times 1}$, for simultaneous communication and target detection, with the signal given by

$$\mathbf{x}^{bs}=\sum\nolimits_{k=1}^{K}\mathbf{v}_k s_k+\sum\nolimits_{t_b=1}^{T_b}\mathbf{s}_{t_b}, \tag{9}$$

where $\mathbf{v}_k\in\mathbb{C}^{M_t\times 1}$: beamformer vector. $s_k\in\mathbb{C}$ is unit power data symbol ($\mathbb{E}|s_k|^2=1$). $\mathbf{s}_{t_b}\in\mathbb{C}^{M_t\times 1}$ is a radar signal for target t_b. $\mathbf{V}_{t_b}=\mathbb{E}\mathbf{s}t_b\mathbf{s}_{t_b}^H$ is the covariance matrix of the radar signal. Thus, the received signal at the BS considering target reflections is

$$\mathbf{y}_{t_b}^{bs}=\bar{\mathbf{H}}_{t_b}\mathbf{x}^{bs}+\sum\nolimits_{t'\neq t_b,t'=1}^{T_b}\bar{\mathbf{H}}_{t'}\mathbf{x}^{bs}+\mathbf{n}_{rad},\forall t_b, \tag{10}$$

where $\mathbf{A}(\vartheta_{t_b})\triangleq\mathbf{a}_r(\vartheta_{t_b})\mathbf{a}_t^H(\vartheta_{t_b})$ with steering vectors for angle ϑ_{t_b}. $\bar{\mathbf{H}}t_b=\alpha t_b\mathbf{A}(\vartheta_{t_b})$is the radar channel. $\alpha_{t_b}\in\mathbb{C}$ is the complex amplitude of the target. $\mathbf{n}_{rad}\in\mathbb{C}^{M_r\times 1}$ is the AWGN with covariance $\sigma_{rad}^2\mathbf{I}_{M_r}$. The SINR for target t_b with $\mathbf{U}\in\mathbb{C}^{M_r\times t_b}$ as the receive beamformer is expressed as

$$\gamma_{t_b}^{\text{rad}}=\frac{\mathbb{E}\{|\mathbf{u}_{t_b}^H\alpha_{t_b}\mathbf{A}(\vartheta_{t_b})\mathbf{x}^{bs}|^2\}}{\mathbb{E}\{|\mathbf{u}_{t_b}^H\mathbf{B}\mathbf{x}^{bs}|^2\}+\mathbb{E}\{|\mathbf{u}_{t_b}^H\mathbf{n}_{rad}|^2\}},\forall t_b, \tag{11}$$

where $\mathbf{B}=\sum_{t'\neq t_b}\alpha_{t'}\mathbf{A}(\vartheta_{t'})$. Further, we assume that each CU encounters mutual interference from other CUs and satellite interference. The received signal at the k^{th} CU is

$$y_k^{bs}=\mathbf{h}_k\mathbf{v}_k s_k+\sum\nolimits_{k'=1,k'\neq k}^{K}\mathbf{h}_k\mathbf{v}_{k'}s_{k'}+\sum\nolimits_{t_b=1}^{T_b}\mathbf{h}_k\mathbf{s}_{t_b}+\sum\nolimits_{l=1}^{L}\mathbf{z}_l\mathbf{w}_l s_l^{sat}+n_k,\forall k, \tag{12}$$

where $\mathbf{h}_k=\mathbf{h}_{b,k}+\mathbf{h}_{r,k}\mathbf{\Phi}\mathbf{H}_{b,r}$. $\mathbf{h}_{b,k}\in\mathbb{C}^{1\times M_t}$ is the direct channel gain. $\mathbf{h}_{r,k}\in\mathbb{C}^{1\times N}$ is the RIS-to-CU channel gain. $\mathbf{H}_{b,r}\in\mathbb{C}^{N\times M_t}$ is the BS-to-RIS channel gain. $\mathbf{z}_l\in\mathbb{C}^{1\times N_t}$ is the satellite interference[3] channel to CUs. $\mathbf{\Phi}=\mathrm{diag}(\boldsymbol{\phi})$ is

[3]Satellite communication interferes with ground communication due to higher transmission power and clear LoS propagation, while ground communication does not interfere with satellite communication because terrestrial signals are attenuated by obstacles and spatial separation [44].

the RIS phase shift matrix with elements $\phi_n \in [0, 2\pi]$. n_k is the AWGN with variance σ_k^2. Based on this, the SINR of each DL CU can be formulated as

$$\gamma_k^{bs} = \frac{|\mathbf{h}_k \mathbf{v}_k|^2}{\sum\limits_{k'=1, k' \neq k}^{K} |\mathbf{h}_k \mathbf{v}_{k'}|^2 + \sum\limits_{t_b=1}^{T_b} \mathbf{h}_k^H \mathbf{V}_{t_b} \mathbf{h}_k + \sum\limits_{l=1}^{L} |\mathbf{z}_l \mathbf{w}_l|^2 + \sigma_k^2}, \forall k. \quad (13)$$

III. Problem Formulation

In this section, we aim to optimize the overall sum rate in the RIS-aided ISAC STIN by jointly optimizing the transmit beamforming vectors $\{\mathbf{w}_l\}_{l=1}^{L}$, $\{\mathbf{v}_k\}_{k=1}^{K}$, $\{\mathbf{W}_{t_s}\}_{t_s=1}^{T_s}$ at the satellite, $\{\mathbf{V}_{t_b}\}_{t_b=1}^{T_b}$ at the BS, the phase shift matrix $\mathbf{\Phi}$ at the RIS, and the receive beamformers $\tilde{\mathbf{U}}$ and $\mathbf{U}$. The optimization set is $\mathcal{J} \triangleq \{\{\mathbf{w}_l\}_{l=1}^{L}, \{\mathbf{v}_k\}_{k=1}^{K}, \{\mathbf{W}_{t_s}\}_{t_s=1}^{T_s} \succeq 0, \{\mathbf{V}_{t_b}\}_{t_b=1}^{T_b} \succeq 0, \mathbf{U}, \tilde{\mathbf{U}}, \mathbf{\Phi}\}$. Our goal is to maximize the sum rate for all SUs and CUs, considering limited transmit power and phase shift constraints at the RIS while ensuring a minimum SINR for the sensing targets. Based on these criteria, the optimization problem is formulated as follows:

$$\underset{\mathcal{J}}{\text{maximize}} \ \hat{R} = \sum_{l=1}^{L} \log_2\left(1 + \gamma_l^{sat}\right) + \sum_{k=1}^{K} \log_2\left(1 + \gamma_k^{bs}\right) \quad (14a)$$

$$\text{s.t} \quad \gamma_{t_s}^{\text{tar}} \geq \tau_{t_s}^{\text{tar}}, \ \forall \ t_s \in \mathcal{T}_s, \quad (14b)$$

$$\gamma_{t_b}^{\text{rad}} \geq \tau_{t_b}^{\text{rad}}, \ \forall \ t_b \in \mathcal{T}_b, \quad (14c)$$

$$\sum\nolimits_{l=1}^{L} \|\mathbf{w}_l\|^2 + \sum\nolimits_{t_s=1}^{T_s} \text{Tr}(\mathbf{W}_{t_s}) \leq P_{\max}^{sat}, \quad (14d)$$

$$\sum\nolimits_{k=1}^{K} \|\mathbf{v}_k\|^2 + \sum_{t_b=1}^{T_b} \text{Tr}(\mathbf{V}_{t_b}) \leq P_{\max}, \quad (14e)$$

$$|\phi_n| = 1, \forall n \in \mathcal{N}, \quad (14f)$$

where the constraints (14b) and (14c) ensure the SINR at the targets meets a minimum threshold, critical for maintaining the quality of sensing operations within the STIN framework. Constraints (14d) and (14e) limit the maximum allowed transmit power for both the ISAC-satellite and the terrestrial BS. Additionally, constraint (14f) is the unit modulus constraint of the phase-shift matrix at the RIS. The coupling relationship between the satellite-based and terrestrial RIS-enhanced ISAC systems is crucial for optimal performance, characterized by interference management and resource allocation strategies that ensure seamless integration and coordination.

The resource allocation problem defined in (14) is non-convex and coupled, making it challenging to solve with standard algorithms. The need for computationally intensive tasks like matrix inversions and singular value decompositions further complicates real-time implementation. To address these issues, we introduce a FMA-DRL algorithm designed to handle non-convex optimization. Our approach dynamically adapts to real-time scenarios, ensuring optimal performance with partial environment observations. We chose DRL with federated MA learning over convex optimization because it handles non-convexity and complexity efficiently while enabling decentralized, collaborative optimization. Thus, we redefine the problem using MA-DRL, assigning independent agents to non-terrestrial and terrestrial zones. Each agent (BS or satellite module) adjusts its operational parameters within each time slot, dynamically optimizing transmit power and other variables to meet SINR requirements and RIS phase-shift design constraints.

IV. The Proposed Federated Multi-Agent Deep Reinforcement Learning Algorithm

In this approach, we reinterpret the previously mentioned problem within the framework of DRL, with the BS and the satellite functioning as agents. The objective of the satellite agent is to enhance the overall user throughput while reducing interference. In the DRL scenario, base stations are required to develop deep neural network (DNN) models that output either Q-values or direct control measures. A pivotal challenge is the expedited training of these DNN models to align with the dynamic nature of network conditions. To address this, we suggest the adoption of an FMA-DRL strategy. Federated learning[4] enhances our system by addressing privacy concerns, enabling localized training, lever aging shared model updates, and optimizing the global objective function across heterogeneous network environments, thus offering significant advantages over independent DRL approaches [48]. In our framework, dedicated agents are assigned to both the satellite zone and the terrestrial base station area, thereby achieving swift adaptation to network changes while safeguarding user data privacy.

In summary, DRL's ability to effectively handle non-convex optimization problems, coupled with its computational efficiency, flexibility, adaptability, and suitability for real-time implementation, makes it a superior choice over traditional convex optimization methods for our system. In subsequent sections, we initially outline the RL issue by detailing the state space, action space, and reward function relevant to the problem (14). Following this, we introduce two variants of FMA-DRL: the FMA-DDPG and the FMA-DQN, both aimed at addressing the challenge of the formulated problem.

A. MDP Formulation

DRL methodologies are typically framed within the structure of Markov decision processes (MDP), which are characterized by a 5-element tuple: $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \bar{\gamma}\}$. By framing the problem as an MDP and including previous beamforming vectors in the state space, our approach leverages historical data for long-term planning, demonstrating the strength of our DRL algorithm in optimizing cumulative network rate. Here, $\mathcal{S}$ denotes the set of all possible states, $\mathcal{A}$ represents the set of all possible actions, $\mathcal{P}$ signifies the probabilities of moving from one state to another state given an action, expressed as $Pr(s^{t+1}|s^t, a^t)$, $r^t(s^t, a^t) \in \mathcal{R}$ is the function that assigns a reward based on the state and action at time t, and $\bar{\gamma} \in [0, 1)$ is the discount factor which adjusts the value of future rewards. At any given time step t, given the current state s^t, the agent selects an action $a^t \in \mathcal{A}$, which leads to

[4]While the proposed framework offers significant potential for improving federated learning in real-world applications, there are several deployment challenges to consider. These include device heterogeneity, communication overhead, data privacy, resource constraints, and scalability [47].

a transition to a new state s^{t+1} with a transition probability of $Pr(s^{t+1}|s^t, a^t) \in \mathcal{P}$, and the agent receives a reward r^t. The decision-making strategy of the agent, known as a policy $\pi(s, a)$, defines the likelihood of choosing action $a^t = a$ when in state $s^t = s$, or formally, $\pi(s, a) = Pr(a^t = a|s^t = s)$. In particular, agent, $\mathcal{S}, \mathcal{A}$, and $\mathcal{R}$ are designed as follows:

1) Agent: In our framework, we allot distinct agents to handle operations in the satellite zone as well as the terrestrial zone.

2) State & Observation Space: The aim is to integrate a comprehensive array of environmental information pertaining to the problem ($\mathbf{P0}$) within the state space. Denote $\mathcal{S} = \{\mathcal{S}_{sat}, \mathcal{S}_{bs}\}$ as the state space of the system, which includes the overall channel conditions and the behavior of all the agents involved. This state space is devised from the information available to the BS and satellite, whether obtained directly or through inference, and it is pivotal in formulating the reward function. The observation state feature set at ISAC-satellite encompasses all channel data within the satellite area, and the previous beamformers and transmission rate [49]–[51]. Therefore, the designated observation state space for the ISAC-satellite is described as

$$\mathcal{S}_{sat} = \{\mathbf{g}_l^t, \mathbf{H}_{t_s}^t, \{\mathbf{w}_l\}^{t-1}, \{\mathbf{W}_{t_s}\}^{t-1}, \tilde{\mathbf{U}}^{t-1}, \hat{R}^{t-1}\}. \quad (15)$$

Meanwhile, the observation state space at the BS includes the current state feature set, which comprises all channel information within the terrestrial area, the phase shift matrix at the RIS, and the previous beamformers and transmission rate [49], [50]. Therefore, the observation state space feature set at the BS is defined as

$$\mathcal{S}_{bs} = \{\mathbf{h}_k^t, \bar{\mathbf{H}}_{t_b}^t, \mathbf{\Phi}^{t-1}, \{\mathbf{v}_k\}^{t-1}, \{\mathbf{V}_{t_b}\}^{t-1}, \{\mathbf{w}_l\}^{t-1}, \mathbf{U}^{t-1}, \hat{R}^{t-1}\}. \quad (16)$$

The observations collected from the agents are saved in a centralized buffer, where each agent retrieves information through its unique control channel. In the training process, updates to the neural network are carried out offline through a random selection of observations from this repository. Following this, agents utilize archived observations to shape their decision-making for forthcoming actions. This process facilitates efficient adaptation and learning in a constantly changing network setting, steadily advancing the decision-making skills of the agents.

3) Action Space: An action results from the policy outputs (either from DQN or the actor-network in actor-critic schemes). Let the action space for the system be denoted as $\mathcal{A} = \{\mathcal{A}_{sat}, \mathcal{A}_{bs}\}$, designed by integrating a policy incorporating both the comprehensive beamforming matrices and the phase-shift matrix at the RIS [49]–[51]. With multiple agents involved, the action space is designed to include the individual actions of each agent. Specifically, the subset of the action space $\mathcal{A}_{sat} \in \mathcal{A}$ that pertains to the satellite area is defined as

$$\mathcal{A}_{sat} = \{\{\mathbf{w}_l\}^t, \{\mathbf{W}_{t_s}\}^t, \tilde{\mathbf{U}}^t\}. \quad (17)$$

Likewise, the subset of the action space $\mathcal{A}_{bs} \in \mathcal{A}$ associated with the terrestrial area is defined as

$$\mathcal{A}_{bs} = \{\{\mathbf{v}_k\}^t, \{\mathbf{V}_{t_b}\}^t, \mathbf{U}^t, \mathbf{\Phi}^t\}. \quad (18)$$

4) Reward Function: The reward function in our study is designed to calculate the instant reward received when an action is performed at state s^t, targeting the maximization of the system's sum rate as detailed in (14) by optimizing the action selection comprehensively. In the realm of DRL, the objective is for the agent to identify actions that lead to the maximization of aggregate rewards over time through discrete-time interactions with the environment. For this purpose, we assign to each agent a reward $r_{\mathcal{F}}^t$, where $\mathcal{F} \in \{sat, bs\}$. To elaborate, "$sat$" refers to the satellite zone, and "bs" signifies the terrestrial BS zone. Consequently, the reward function at any given time step t is determined as $r^t = \sum_{j \in \mathcal{F}} r_j^t = \hat{R}$.

The goal of the learning process is to identify the optimal policy, π^*, which maximizes the expected cumulative reward starting from any state s given by

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_\pi\{R^0|s^0 = s\}. \quad (19)$$

Through the establishment of specific reward functions for both participating agents, our FMA-DRL model is designed to enhance the total rewards accumulated by agents operating in both satellite and terrestrial zones over the duration of their interactions. In this context, the term reward function captures the aggregate of all discounted rewards expressed as $R^t = \sum_{i=0}^{\infty} \bar{\gamma}^{t+i} r^{t+i}$.

B. Federated Learning Model

In this section, we discuss the utilization of DRL algorithms that employ DNNs to determine action probabilities for optimizing returns. The fundamental concept of federated learning is to develop a unified statistical model (here, a DNN) using data collected across numerous devices. Our approach ensures each agent processes and keeps its state data locally, sending periodic updates to the gateway. As the central server for federated learning, the gateway facilitates seamless communication, optimizes resource allocation, and ensures efficient coordination and data aggregation. The objective of this training methodology is to optimize a predefined objective function by minimizing its value which is given by

$$\min_{\varrho} F(\varrho) = \sum\nolimits_{j \in \mathcal{F}} \kappa^j F_j(\varrho^j). \quad (20)$$

The global model is deployed at a central server located in the network gateway, which connects the terrestrial and satellite zones. For optimizing the global model, we aim to fine-tune the global objective function, $F(\varrho)$, alongside its weights, ϱ, and adjust the local loss function, F_j, with their respective weights, ϱ^j, for each agent across different zones j. The share of each zone, j, in the overarching network is determined by κ^j, defined as $\kappa^j = \frac{\mathcal{U}_j}{\sum_{j \in \mathcal{F}} \mathcal{U}_j}$, where $\mathcal{U}_j$ denotes the number of users within zone j of the network set $\mathcal{F}$. Specifically, for the satellite zone, represented as $j = sat$, $\mathcal{U}_j = L$, whereas for the terrestrial zone, denoted as $j = bs$, $\mathcal{U}_j = K$. Our FMA-DRL framework optimizes privacy, efficiency, and scalability in distributed communication by coordinating multiple agents via a central server, defining unique state-action spaces, and using a shared reward function for stable, efficient learning.

C. Federated Multi-Agent Deep Deterministic Policy Gradient

DDPG enhances the actor-critic framework by incorporating DNNs to model policy and value functions. This approach provides a more sophisticated solution to the challenges of handling extensive state and action spaces, characteristic of high-dimensional scenarios. DDPG stands out for its capability to handle continuous action spaces, making it adept at decision-making in such environments. Essentially, the proposed DDPG employs two key DNN components in its architecture:

1) Critic Network: The model, also known as a Q-network and characterized by the parameter ϱ_c, processes an input comprising a state s and an action a as inputs, subsequently yielding the Q-value, $Q(s^t, a^t; \varrho_c)$. Further, the action-value function, often referred to as the Q-function is defined as

$$Q_\pi\left(s^t, a^t\right) = \mathbb{E}_\pi\left[R^t | s^t = s, a^t = a\right]. \tag{21}$$

This function can be updated using the Bellman expectation equation [52]. Furthermore, the optimal Q-function in (21) can be determined through the Bellman optimality equation, expressed as

$$Q^*\left(s^t, a^t\right) = r^t + \bar{\gamma} \max_{a^{t+1} \in \mathcal{A}} Q^*\left(s^{t+1}, a^{t+1}\right). \tag{22}$$

Accordingly, the optimal action a^* is derived by

$$a^* = \arg\max_{a \in \mathcal{A}} Q^*(s, a). \tag{23}$$

2) Actor Network: This is also referred to as a policy network, which accepts a state s as input and produces a continuous action a, denoted by $a^t = \pi(s^t; \varrho_\mu)$ while updating the network parameter ϱ_μ. The actor-network undergoes training through (23), intending to optimize the state-value function.

Besides, DDPG employs both a target actor-network, denoted as $\pi(s^t; \varrho'_\mu)$, and a target critic network, represented by $Q(s^t, a^t; \varrho'_c)$, to enhance training stability. Here, ϱ'_μ and ϱ'_c signify the parameters for the target actor and critic networks, respectively. Further, the actor undergoes training to enhance the objective function by employing a policy gradient method

$$\nabla_{\varrho_\mu} J(\varrho_\mu) \approx \mathbb{E}[\nabla_a Q(s^t, a; \varrho_c)|_{a=\pi(s^t;\varrho_\mu)} \nabla_{\varrho_\mu} \pi(s^t; \varrho_\mu)]. \tag{24}$$

Here, $J(\varrho_\mu) = \mathbb{E}_{s\sim\varrho_c, a\sim\varrho_\mu} R(s, a)$ typically represents the expected cumulative return. Meanwhile, the critic undergoes iterative optimization aimed at reducing the loss function, which is characterized as

$$L(\varrho_c) = \mathbb{E}\left[(y^t - Q(s^t, a^t; \varrho_c))^2\right], \tag{25}$$

where $y^t = R^t + \bar{\gamma} Q(s^{t+1}, \pi(s^{t+1}; \varrho'_\mu); \varrho'_c)$ denotes the expected return. The stability of y^t throughout the training is ensured by incrementally adjusting the parameters of the target networks with a minor coefficient, $\zeta \in [0, 1]$, thus updating as $\varrho'_\mu = \zeta \varrho_\mu + (1 - \zeta)\varrho'_\mu$ and $\varrho'_c = \zeta \varrho_c + (1 - \zeta)\varrho'_c$.

Unlike value-based approaches like Q-learning, policy gradient techniques optimize the policy directly, bypassing the need to calculate Q-values. This strategy avoids the overestimation bias inherent in value-based methods. The update of parameters favors actions leading to more rewarding outcomes. During testing, the best policy is identified by choosing the action that has the highest probability in a deterministic manner. Thus, by incorporating the cost function into (20), we obtain the cost associated with the F-DDPG algorithm as

$$\min_{\varrho} J(\varrho_\mu) = \sum\nolimits_{j \in \mathcal{F}} \kappa^j J_j(\varrho_\mu^j). \tag{26}$$

Furthermore, within the context of the MA-DRL system framework, each agent independently employs a F-DDPG based algorithm, enabling personalized policy optimization and adaptation. This specific methodology is described in **Algorithm 1**.

Algorithm 1 Federated DDPG Algorithm for Each Agent

Input: Initialize the parameter settings for the proposed system model, neural networks at $t = 0$
Input: The aggregation frequency ρ, exploration parameter ϵ, learning rate Ω, number of episodes E
Initialize the actor-network, $\pi(s^t; \varrho_\mu)$ and the critic network $Q(s^t, a^t; \varrho_c)$ with the weights ϱ_μ and ϱ_c.
Create the target DNNs by setting $\varrho'_\mu \leftarrow \varrho_\mu$ and $\varrho'_c \leftarrow \varrho_c$
Initialize a replay buffer
Initialization: get initial ϱ_μ from server
for $ep = 1 \rightarrow E$ **do**
    Initialize a random process η for action exploration
    Receive initial observation state s^1
    for $t = 1 \rightarrow T$ **do**
        Obtain action a^t from the actor-network;
        Add exploration noise to a^t as $a^t = a^t + \eta$
        Calculate the instant reward r^t
        Observe the new state s^{t+1}
        Store experiences in the buffer and sample random mini-batches of experiences to train the DNNs
        Set the expected return y^t
        Update the actor policy via (24) and critic via (25)
        Update the target actor ϱ'_μ and the target critic ϱ'_c
    end for
end for
update $\varrho_\mu^{ep+1} = \varrho_\mu^{ep} + \Omega \nabla_{\varrho_\mu} J(\varrho_\mu)$
if ep mod $\rho = 0$ **then**
    send ϱ_μ^{ep} to server for aggregation
    get aggregated ϱ_μ^{ep} from server
end if

Algorithm 2 Federated DQN Algorithm for Each Agent

Input: Initialize the parameter settings for the proposed system model, neural networks at $t = 0$
Input: ρ, ϵ, Ω, E
Initialization: get initial ϱ_n from server
for $ep = 1 \rightarrow E$ **do**
    Receive initial observation state s^1
    for $t = 1 \rightarrow T$ **do**
        Select random $r \in [0, 1]$. Obtain action a^t using

$$a^t \triangleq \begin{cases} \text{argmax}_a\, Q(s^t, a; \varrho_n) & \text{if } r > \epsilon \\ \text{pick uniformly action} & \text{else} \end{cases}$$

        Take action a^t, go to state s^{t+1} and get reward r^{t+1}
        Store the tuple $\mathcal{B} = \{a^t, s^t, r^{t+1}, s^{t+1}\}$
    end for
    update $\varrho_n^{ep+1} = \varrho_n^{ep} - \Omega \nabla_{\varrho_n} L(\varrho_n)$
    if ep mod $\rho = 0$ **then**
        send ϱ_n^{ep} to server for aggregation
        get aggregated ϱ_n^{ep} from server
    end if
end for

D. Federated Multi-Agent Deep Q Network

Transitioning from DDPG, we now explore the value-based RL methodologies to estimate the expected future rewards from a particular state s upon executing an action a, employing the action-value function $Q(s,a)$ as follows

$$Q_\pi(s^t,a) = \mathbb{E}_\pi\left\{\sum_{j=1}^{\infty}\bar{\gamma}^{t+j}r^{t+j}|s^t,a\right\} \tag{27}$$

$$= \mathbb{E}_{s^{t+1},a}\{r^t + \bar{\gamma}Q_\pi(s^{t+1},a)|s^t,a^t\}. \tag{28}$$

The RL agent aims to identify the optimal action-value function $Q^*(s^t,a)$, which calculates the maximum expected sum of discounted returns from a given state s^t

$$Q^*(s^t,a) = E_{s^{t+1}}\{r^t + \bar{\gamma}\max_a Q^*(s^t,a)|s^t,a\}. \tag{29}$$

In DRL, a method of function approximation specifically, a DNN is employed to learn a parameterized value function $Q(s,a;\varrho_n)$ with the aim of closely approximating the optimal Q-values. The target value for the function $Q(s^t,a;\varrho_n)$ is derived from the one-step lookahead formula $r^t + \bar{\gamma}\max_a Q(s^{t+1},a;\varrho_n)$, making the function contingent upon the parameters ϱ_n. The effectiveness of selecting an appropriate action is contingent upon the precise estimation of action values. Consequently, DQN focuses on identifying the optimal parameters ϱ_n^*, to minimize its loss function

$$L(\varrho_n) = (r^t + \bar{\gamma}\max_a Q(s^{t+1},a;\varrho_n) - Q(s^t,a;\varrho_n))^2. \tag{30}$$

At the outset of Q-learning, agents accrue experience by engaging with their surroundings. This process involves compiling a dataset $\mathcal{B}$, which is essentially a collection of experiences up to a given time t, structured as tuples $(s^{t-1},a^{t-1},r^t,s^t)$. Subsequently, this dataset $\mathcal{B}$ is utilized to refine the loss function $L(\varrho_n)$ through optimization. During the initial phase of training, the precision of the agent's predictions is typically low. To mitigate this, an adaptive ϵ-greedy strategy is employed for decision-making, allowing the agent to experiment with various actions at a certain probability level, irrespective of their immediate rewards. This approach fosters more precise estimations as training progresses and mitigates the likelihood of the model becoming overly fitted to actions that yield high rewards early in the training period. Thus, by incorporating the cost function into (20), we obtain the cost associated with the federated DQN algorithm as

$$\min_{\varrho_n} L(\varrho_n) = \sum\nolimits_{j\in\mathcal{F}} \kappa^j L_j(\varrho_n^j). \tag{31}$$

Additionally, in the framework of the MA-DRL system, every agent individually utilizes a federated algorithm based on DQN, allowing for tailored optimization and adjustment of optimization and adjustment of action-value functions. This particular process is outlined in **Algorithm 2**.

E. Complexity Analysis

For the computation at the t^{th} iteration, we categorize the dimensions of the action and state spaces with the notations $|a^t|$ and $|s^t|$, correspondingly. Algorithm 1 breaks down the process into two primary segments: 1) Calculation of rewards, which holds a computational complexity of $\mathcal{O}(|s^t|)$. 2) Selection of actions, where the actor and critic networks' complexity is determined by the neuron count in each layer and the total number of layers. For the actor-network, let the number of neurons in its m^{th} layer be represented by W^m, and the total layer count by L^a. Consequently, the complexity for a single layer m is given by $\mathcal{O}(W^{m-1}W^m + W^mW^{m+1})$, leading to a total actor network computational complexity of $C_t^a = \mathcal{O}(|s^t|\cdot W^1 + \sum_{m=2}^{L^a-1}(W^{m-1}W^m + W^mW^{m+1}) + W^{L^a-1}\cdot|a^t|)$. In the critic network, with V^q as the neuron count for layer q and L^c for the total layers, the complexity for layer q is $\mathcal{O}(V^{q-1}V^q + V^qV^{q+1})$, leading to a total critic network computational complexity of $C_t^c = \mathcal{O}(|s^t|\cdot V^1 + \sum_{q=2}^{L^c-1}(V^{q-1}V^q + V^qV^{q+1}) + V^{L^c-1})$. Thus, the overall complexity for choosing actions is denoted by $C_t = C_t^a + C_t^c$. The overarching computational complexity of the algorithm across all iterations is therefore expressed as $\mathcal{O}(E\cdot T\cdot C_t)$.

TABLE II: Simulation Parameters [44], [53]

Parameters	**Value**
Carrier frequency	28 GHz (Ka-band)
Bandwidth	500 MHz
3 dB angle	0.4°
Height of GEO satellite	35786 km
Maximum beam gain	52 dBi
User terminal antenna gain F_g	42.7 dBi
UPA inter element spacing	$\lambda/2$
Rain fading parameters	$(\mu^{rain},\sigma^{rain}) = (-2.6, 1.63)$
Boltzmann's constant, κ	1.38×10^{-23}J/m
Noise temperature of system, $Tsys$	517K

V. Numerical Simulations and Discussion

A. Parameter Setup

In this segment, we delve into the performance analysis of our proposed federated learning algorithms, leveraging PyTorch for model development and employing the Adam optimization technique for model training. The architectural setup for both the proposed F-DDPG and the F-DQN across MA systems is similar, employing two hidden layers each with 256 neurons [54]. Furthermore, the neural network parameters are updated using the Adam optimizer and the activation function used is ReLU [54]. Moreover, we set the hyperparameters as follows: $\bar{\gamma} = 0.9$ [55], the learning rates for both actor and critic networks, $\Omega = 0.001$, the memory buffer is $W = 10000$, the size of minibatch = 32 [56], the number of episodes is $E = 5000$, with each episode a horizon of $T = 10$ time slots[5], and the aggregate frequency is $\rho = 100$ [57].

Moreover, Table II outlines the parameters used in the simulation [44], [53]. The satellite is configured with $N_t = N_r = \bar{N} = 4$ antennas for transmission and reception. It includes $L = 4$ SUs and $\mathcal{T}_s = 2$ targets within the satellite zone, with these targets positioned at angles $\psi_1 = 35°$

[5]In our proposed scenario, dynamic factors like channel fading and adjustments to the RIS reflection matrix and beamforming vectors ensure that the communication environment varies across the 10-time slots, creating dependencies between them.

and $\psi_2 = 50°$. On the terrestrial side, the BS features $M_t = M_r = M = 4$ transmit and receive antennas, serving $K = 4$ CUs and integrating $N = 64$ elements of a RIS. We consider $\mathcal{T}_b = 2$ targets in the terrestrial zone, located at angles $\vartheta_1 = 0°$ and $\vartheta_2 = 20°$. For each CU channel, we model the transmission path using a LoS approach [58], characterized by the channel representation $\mathbf{h}_k = \sqrt{\xi_k M_r}\mathbf{a}_r(\tilde{\vartheta}_k), \forall k$. This notation includes ξ_k to denote the path loss and $\tilde{\vartheta}_k$ for the angular direction of the user. A standard path loss value of -103.6 dB is applied to model the link between each CU and the BS. The directional angles for the DL CUs, listed as $\{\tilde{\vartheta}_1, \tilde{\vartheta}_2\}$ and $\{\tilde{\vartheta}_3, \tilde{\vartheta}_4\}$ are configured to $\{-40°, 60°\}$ and $\{45°, -65°\}$, respectively [59]. The simulation does not incorporate any form of user grouping. Both the satellite and BS adhere to a maximum transmit power limit of $P_{max}^{sat} = P_{max} = 40$ dBm. For the satellite channel model, the noise power is normalized by $\kappa T_{sys} B_g$ which results in the noise variance being set to $\sigma_l^{sat\,2} = \sigma_k{}^2 = 1, \forall l, \forall k$ [44]. For simplicity, we consider $\tau_{t_s}^{\text{tar}} = \tau^{\text{tar}} = 12$ dB, $\forall t_s$ and $\tau_{t_b}^{\text{rad}} = \tau^{\text{rad}} = 12$ dB, $\forall t_b$ [59]. The numerical results presented are the average outcomes from 100 distinct channel realizations. Unless stated otherwise, the parameter settings adhere to the aforementioned specifications.

B. Benchmark Schemes

For comparative analysis, we incorporate the following benchmark schemes.

1) **Centralized scheme**: In the centralized DRL framework, each iteration requires agents to share data with a central server for immediate control actions, facilitating the development of a unified model. This setup is implemented for both DDPG and DQN methods, denoted as C-DDPG and C-DQN, respectively, in the figures [60].
2) **Non-Federated Schemes:** This benchmarking scheme compares the proposed federated algorithms with DDPG and DQN in non-federated settings.
3) **Communication - only scheme**: This benchmark omits the sensing SINR requirements within a RIS-enhanced STIN, labeled as "w/o sensing" in the figures. This strategy assists in determining the effects of incorporating sensing capabilities on communication performance [61].
4) **Random RIS scheme**: In this scheme, while the satellite and BS utilize our proposed beamforming approach, the RIS adopts a random passive beamforming approach [62]–[65].
5) **No-RIS scheme**: This scheme examines the performance of our ISAC-STIN setup without RIS intervention in the terrestrial domain, marked as "w/o RIS" in the figures. underscores the significance of integrating RIS into our network design [62]–[65].

In Fig. 2, the cumulative network rate is presented for both DDPG and DQN algorithms implemented in a federated setting across MA systems, examining various aggregation intervals. The training process spanned 5000 episodes, with each episode consisting of a sequence of $T = 10$ time slots. The depicted curves highlight the impact of different server-agent aggregate frequencies. The aggregation frequencies considered for this simulation are $\rho = 50$ and $\rho = 500$. Specifically, it dictates the frequency of model update exchanges between the server and agents, affecting the convergence rate and effectiveness of the federated learning algorithms. Notably, the F-DDPG approach shows a more consistent and smooth path to convergence when compared to F-DQN, across all examined aggregation frequencies, ensuring steadier performance delivery. This phenomenon is likely due to the distinct exploration mechanisms employed by the two algorithms. On one hand, DQN relies on the ϵ-greedy method, introducing more randomness until convergence is attained. On the other hand, DDPG leverages a policy gradient framework that prioritizes updates along state-action paths associated with superior average rewards, which seems to benefit more directly from the aggregation process.

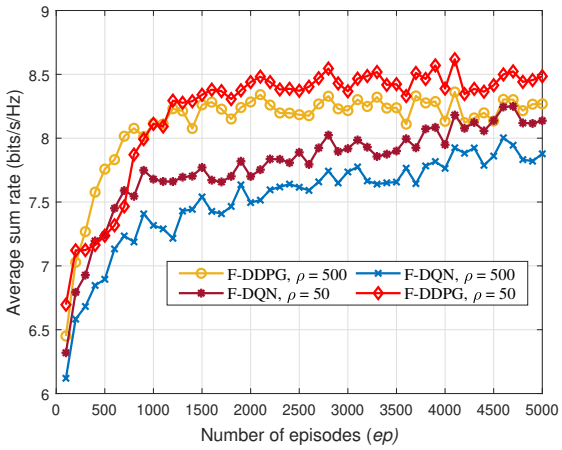

Fig. 2: Convergence behavior of the proposed federated algorithms at different aggregate frequencies.

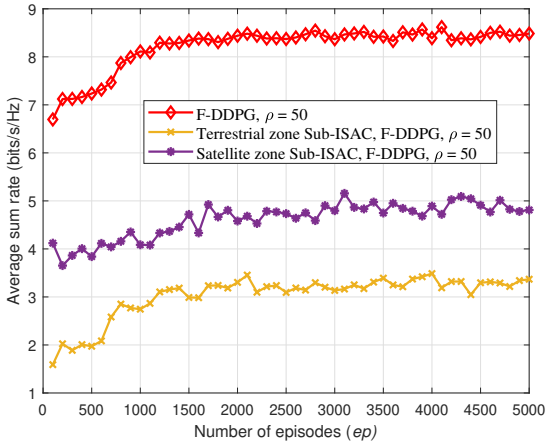

Fig. 3: Convergence behavior of the sub-ISAC systems.

In Fig. 3, we further explore the convergence characteristics by presenting the convergence plot of F-DDPG for the $\rho = 50$ case. Here, the average sum rate of the two sub-ISAC systems is shown separately: one for the terrestrial zone and the other for the satellite zone. This differentiation allows us to observe how each system works towards rate maximization individually and collectively, providing a comprehensive un-

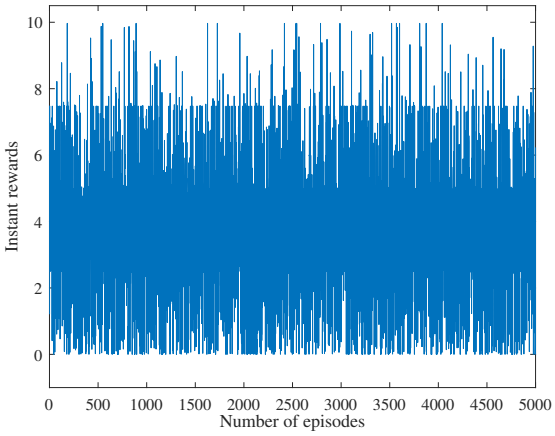

Fig. 4: Instant rewards versus number of episodes for the proposed federated algorithm.

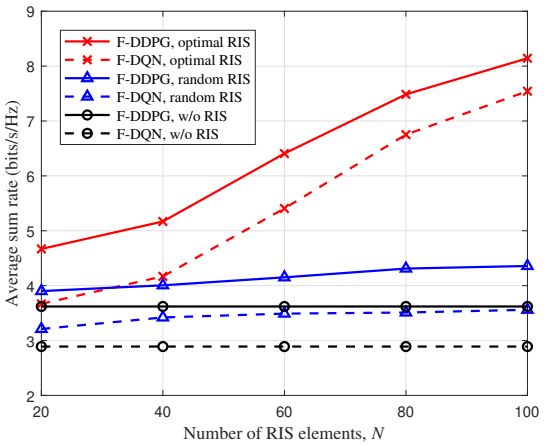

Fig. 6: Average sum rate versus N of the proposed federated algorithms for different benchmark schemes.

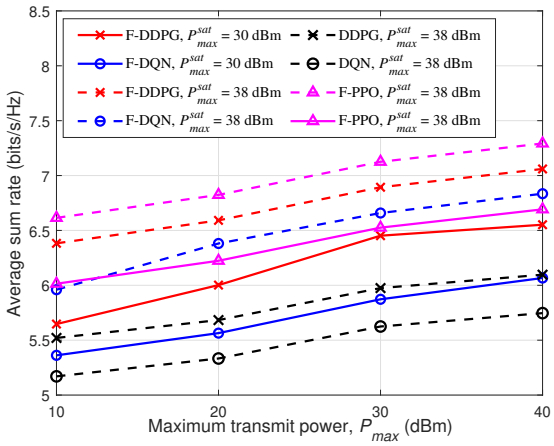

Fig. 5: Average sum rate versus P_{max} with different P_{max}^{sat}.

derstanding of the overall system performance. Additionally, in Fig. 4, the instant reward for the proposed F-DDPG system is included to more intuitively reflect learning fluctuations[6] in real-time scenarios. This inclusion showcases the dynamic adaptation of the learning algorithms throughout the training process. Overall, the instant reward in Fig. 4 is meant to display immediate fluctuations in step-wise rewards under initial settings, which may not capture the full efficacy of F-DDPG. In contrast, Fig. 3 is a cumulative measure, more accurately reflecting the algorithm's optimized performance trajectory over time.

In Fig. 5, the performance of both F-DDPG and F-DQN strategies is observed across varying levels of maximum transmission power at the BS. This exploration also considers the impact of altering maximum transmission power levels at the satellite alongside BS power adjustments on the overall system sum rate. The analysis reveals that the increasing transmission power, both at the BS and the satellite, leads to noticeable improvements in the operational metrics of both algorithms with F-DDPG demonstrating superior performance over F-DQN. Specifically, at a peak transmission power of 30 dBm for both the satellite and BS, F-DDPG and F-DQN attain system sum rates of 6.4[7] and 5.8, respectively. Furthermore, the comparison includes non-federated benchmark schemes DDPG and DQN [66], showing that our proposed federated schemes outperform these baselines. Additionally, we include federated proximal policy optimization (F-PPO) as a benchmark. While F-PPO shows slightly better performance in power versus average sum rate, it has several drawbacks: it requires fresh data[8] for each update and incurs higher computational overhead [67]. Thus, the federated schemes' superior performance can be attributed to their ability to leverage distributed data while maintaining global model consistency, resulting in better optimization of the sum rate.

Fig. 6 illustrates the impact of increasing the number of elements in an RIS on the system sum rate, indicating that a higher number of reflecting elements leads to improved system performance due to the additional spatial degrees of freedom. Furthermore, we compare the outcomes against scenarios without RIS (labeled as "w/o RIS") and those with randomly allocated RIS phase shifts ("random RIS"), for both the F-DDPG and F-DQN methods. Consequently, the optimal RIS configuration delivers a performance gain of about 54.2% over the random RIS setup and a notable 76.8% improvement against the w/o RIS scenario. This highlights the pivotal role that RIS technology plays in enhancing network efficiency, especially when moving from scenarios without RIS to those with optimally deployed RIS. The incorporation of RIS within the terrestrial segment notably amplifies performance in the zone, contributing to a substantial improvement in the performance metrics of the overall considered network.

Fig. 7 illustrates the trade-off between the communication performance and target detection capabilities. Specifically, Fig. 7 plots the system communication rate against the SINR

[6]The fluctuations in Fig. 4 represent expected variability in step-by-step rewards and do not detract from the reliability of our conclusions, as these are supported by the cumulative rewards and convergence shown in Fig. 3.

[7]The sum rate is expressed in bits/s/Hz. However, for brevity, the unit of measure is omitted in the text.

[8]Unlike DDPG and DQN, PPO requires fresh, on-policy data for each update, increasing data collection demands compared to the experience replay used in off-policy methods like DDPG and DQN.

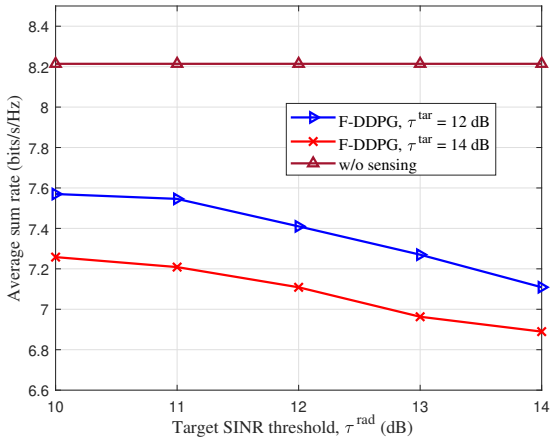

Fig. 7: Average sum rate versus τ^{rad} with different τ^{tar}.

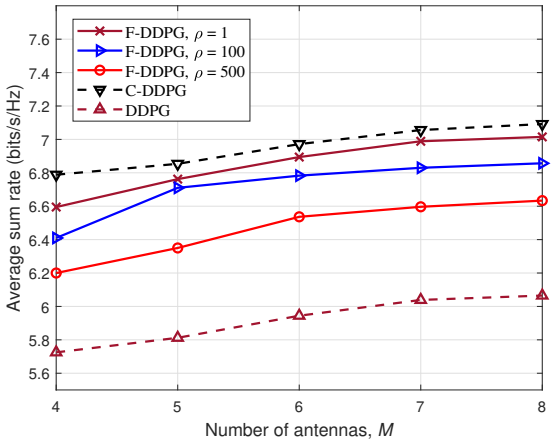

Fig. 8: Average sum rate versus M of the proposed federated algorithm at different aggregate frequencies and benchmark schemes.

threshold needed for targets' detection in both the satellite (τ^{tar}) and terrestrial domains (τ^{rad}). For a clearer comparison, a reference scenario focusing solely on communication excluding target sensing SINR requirements ("w/o sensing") is also analyzed. This comparison underscores the effect of incorporating radar sensing on the overall communication performance. As the SINR threshold for detection increases, the performance of our federated solution tends to decline. This decline occurs because, with higher SINR thresholds, more transmit power within the ISAC system must be allocated to meet the more demanding sensing requirements, consequently reducing the available power for communication. Therefore, this scenario underscores the inherent trade-off between communication and radar functionalities in ISAC systems.

In Fig. 8, we analyze the effect of antenna array sizes at both satellite and BS on the overall system performance, comparing outcomes from both the suggested federated approach and benchmark centralized scheme, as well as non-federated DDPG and DQN schemes. In the centralized DRL framework, each iteration requires agents to share data with a central server for immediate control actions, facilitating the development of

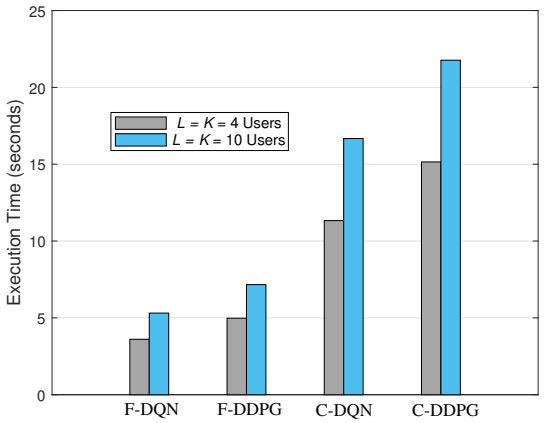

Fig. 9: Execution time comparison between the federated and centralized schemes for different number of users in the network.

a unified model. The results, as shown in Fig. 8, indicate that augmenting the number of antennas enhances spatial DoFs, thereby boosting the overall system performance. For this scenario, we set $\bar{N} = 4$ and evaluate the system performance with the varying number of antennas at the BS. When evaluating the federated method against the centralized model, their performances are nearly similar; however, the federated strategy significantly reduces the amount of data exchanged between agents and the central server by up to 0.002 times (for the case $\rho = 500$). Additionally, we compare the federated scheme with the non-federated DDPG to highlight the distinctions between federated/centralized schemes and the DDPG benchmark. The federated and centralized schemes benefit from global model updates and aggregated learning, leading to better performance in the sum rate optimization, while DDPG, being a localized learning method, lacks this holistic view, resulting in lower performance.

In Fig.9, the execution times for both federated and centralized configurations utilizing DDPG and DQN algorithms are examined for different number of users in the network. This evaluation focuses on the total execution time, which includes the local training period of each agent, the communication delay between the agents and the central server, and the aggregation time required by the server to aggregate the updates. These timings reflect the complete cycle duration for the operation of each model. In our federated models, we set the parameter ρ at 100. The findings demonstrate that federated setups outperform centralized ones in terms of execution speed. The analysis of Fig. 8 and Fig. 9 clearly shows that the federated algorithms not only align closely with the centralized model in performance while ensuring a reduced execution time, effectively balancing performance with efficiency.

Next, we detail the achieved beampattern gain for target functionality realized through Algorithm 1. By employing the optimized receive beamformer $\mathbf{u}_{t_b}^*$, which is normalized to ensure $\left\|\mathbf{u}_{t_b}^*\right\| = 1$, along with the optimized transmit signal $\mathbf{x}^{bs*}$, we define the beampattern directed towards the target as

$$p(\vartheta_{t_b}) = |\mathbf{u}_{t_b}^* \mathbf{a}_r(\vartheta_{t_b}) \mathbf{a}_t^H(\vartheta_{t_b}) \mathbf{x}^{bs*}|. \tag{32}$$

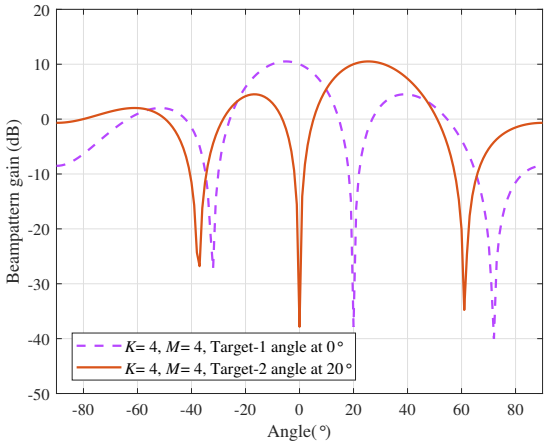

Fig. 10: Beampattern gain of terrestrial zone targets.

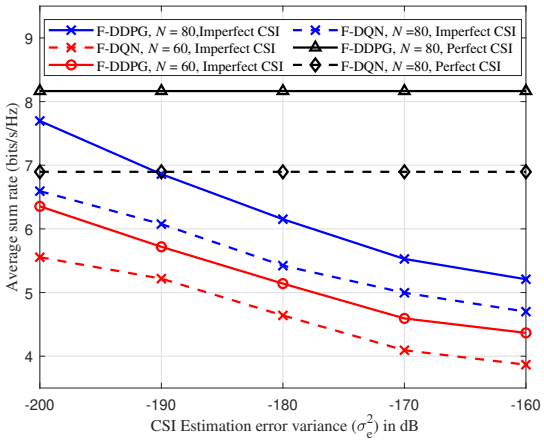

Fig. 12: Average sum rate versus channel estimation error variance (σ_e^2) with different N.

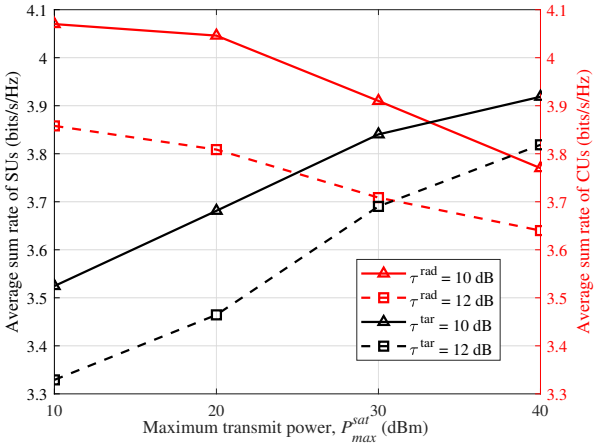

Fig. 11: Trade-off between average sum rate of SUs and CUs for varying P_{max}^{sat} and multiple targets' SINR thresholds.

Further, we illustrate the beampattern gains achieved for specific target locations in Fig. 10. We consider a case with $\mathcal{T}_b = 2$ targets positioned at angles 0° and 20°. Keeping other parameter settings constant, we address the challenge of maximizing the sum rate subject to radar SINR constraints for $\mathcal{T}_b$ targets in the terrestrial zone using Algorithm 1. The target beampattern gains showcased in Fig. 10 is the result of employing the optimized receive beamformer $\mathbf{u}_{t_b}^*$, as outlined in (32). The illustrations confirm that the main lobes are directed toward the intended targets, underscoring the capability of the proposed beamforming design to detect multiple targets effectively.

Fig. 11 illustrates the impact of varying the maximum transmission power at the satellite (P_{max}^{sat}) on the average sum rates for SUs and CUs within the network, maintaining a constant transmit power of $P_{max} = 40$ dBm at the BS. The analysis further carried out against different SINR threshold values required for target detection in satellite (τ^{tar}) and terrestrial (τ^{rad}) zones. The results highlight an improvement in the sum rate for SUs as the P_{max}^{sat} increases, as facilitated by the F-DDPG algorithm. On the contrary, this increase in P_{max}^{sat} introduces heightened interference in the terrestrial zone, diminishing the average sum rate for CUs. This delineates a trade-off, revealing the differential impacts of satellite transmission power on SUs and CUs at a consistent BS transmit power level. Additionally, the inclusion of SINR threshold effects on targets reveals a decline in the performance of our federated approach as the SINR threshold for detection ascends. This decline is due to the higher transmit power demands within the ISAC system to meet stricter sensing requirements, which in turn limits the power availability for communication tasks.

The analysis of the impact of CSI uncertainty on the performance of the STIN reveals significant insights, as depicted in Fig. 12. The study thoroughly examines system performance, highlighting the effects of different levels of CSI estimation inaccuracies (σ_e^2). The results show a noticeable decline in performance with increasing CSI estimation uncertainty. Additionally, the advantages of adding more RIS elements, such as greater channel diversity and improved system performance, outweigh the difficulties caused by CSI estimation uncertainties. This is clearly demonstrated in Fig. 12, where the average sum rate for the scenario with $N = 80$ RIS elements exceeds that of the scenario with $N = 60$ RIS elements for the proposed federated algorithms. The findings underscore the importance of considering CSI uncertainty in the STIN and validate the effectiveness of the proposed approach in addressing practical challenges like robustness.

Fig. 13 illustrates the performance comparison of the average sum rate as a function of the number of users. In this analysis, we set $L = K$ for simplicity. This figure demonstrates the scalability of our proposed federated learning scheme using DDPG in comparison to F-DQN and centralized benchmarks. The centralized schemes, which involve direct data exchange, set an upper bound on performance. In contrast, federated learning exchanges only weight updates, which impacts performance but reduces data communication overhead. As the number of users increases, the performance gap between centralized DDPG, centralized DQN, and their federated counterparts (F-DDPG and F-DQN) tends to lessen. The diminishing performance gap is attributed to the limited system resources, which constrain the achievable performance

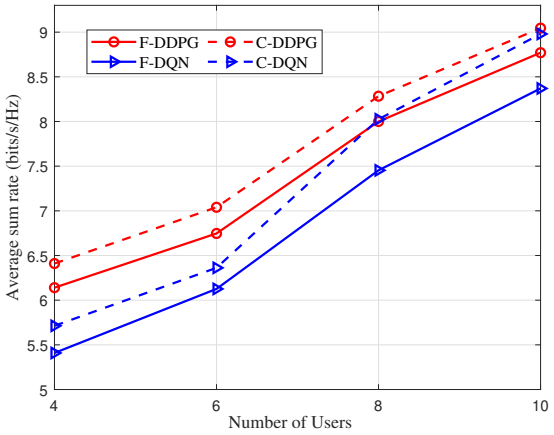

Fig. 13: Average sum rate versus L, K $(L = K)$ of the proposed federated algorithm in comparison with different benchmark schemes.

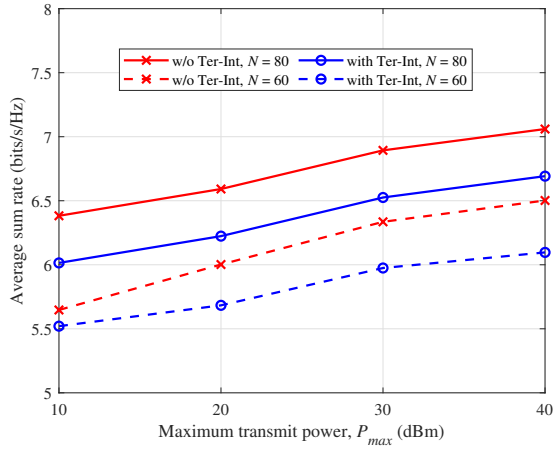

Fig. 14: Impact of terrestrial zone interference on average sum rate.

as the number of users increases. Despite these constraints, our proposed federated learning approach demonstrates robust scalability, maintaining competitive average sum rates and effectively managing the increasing number of users. This highlights the efficiency of federated learning in distributed environments and its potential to achieve high performance with reduced data exchange.

Fig. 14 illustrates the relationship between average rate and power for varying numbers of RIS elements, under conditions both with and without terrestrial interference, denoted as "with Ter-Int" and "w/o Ter-Int," respectively. The data indicate that terrestrial interference markedly diminishes the average rate, especially at elevated interference power levels and with a reduced number of RIS elements. This underscores the sensitivity of the proposed methods to interference assumptions, implying that actual system performance could fluctuate based on the extent of terrestrial interference. The augmentation of RIS elements highlights their efficacy in mitigating the adverse effects of terrestrial interference on the system.

VI. Conclusions and Future Work

In this study, we introduced a cutting-edge analytical framework to enhance spectral efficiency in STINs through the use of RIS within the ISAC framework. By leveraging federated learning, our method dynamically adapted to network changes, facilitating efficient resource management and ensuring compliance with beamforming designs, multiple target SINR thresholds, and RIS phase-shift requirements via an effective feedback loop. The proposed F-DDPG algorithm across MA systems outperformed existing models, including F-DQN, centralized, and traditional methods, by aligning closely with centralized model performance while significantly reducing execution times. Simulation results have demonstrated that the optimal RIS configurations led to a 54.2% performance increase over random setups and a 76.8% enhancement compared to scenarios without RIS, underscoring the significant impact of our federated learning approach in optimizing ISAC-enabled STINs, signaling a breakthrough in the optimization of spectral efficiency. Additionally, our study evaluated the impact of channel estimation errors and interference, confirming the robustness of our approach and its potential to optimize ISAC-enabled STINs.

References

[1] A. Yazar, S. Dogan-Tusha, and H. Arslan, "6G vision: An ultra-flexible perspective," *ITU J. Future Evolving Technol.*, vol. 1, no. 1, pp. 121–140, 2020.

[2] J. A. Zhang, M. L. Rahman, K. Wu, X. Huang, Y. J. Guo, S. Chen, and J. Yuan, "Enabling joint communication and radar sensing in mobile networks—a survey," *IEEE Commun. Surv. & Tut.*, vol. 24, no. 1, pp. 306–345, Firstquart. 2022.

[3] L. Yin, Z. Liu, M. R. Bhavani Shankar, M. Alaee-Kerahroodi, and B. Clerckx, "Integrated sensing and communications enabled low earth orbit satellite systems," *IEEE Netw.*, pp. 1–1, 2024.

[4] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband LEO satellite communications: Architectures and key technologies," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 55–61, Apr. 2019.

[5] I. del Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta Astronautica*, vol. 159, pp. 123–135, 2019.

[6] J. P. Choi and C. Joo, "Challenges for efficient and seamless space-terrestrial heterogeneous networks," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 156–162, May 2015.

[7] "(Release 15) study on new radio (NR) to support non-terrestrial networks," *3GPP Sophia Antipolis, France, Rep. TR38.811 V15.3.0. Release 15*, Jul. 2020.

[8] J. Peisa, P. Persson, S. Parkvall, E. Dahlman, A. Grovlen, C. Hoymann, and D. Gerstenberger, "5G evolution: 3GPP releases 16 & 17 overview," *Ericsson Technology Review*, vol. 2020, pp. 2–13, 03 2020.

[9] B. Aazhang *et.al*, *Key drivers and research challenges for 6G ubiquitous wireless intelligence (white paper)*, 09 2019.

[10] L. Kuang, C. Jiang, Y. Qian, and J. Lu, *Terrestrial-Satellite Communication Networks: Transceivers Design and Resource Allocation*. Springer, 2017.

[11] K. An, M. Lin, W.-P. Zhu, Y. Huang, and G. Zheng, "Outage performance of cognitive hybrid satellite–terrestrial networks with interference constraint," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9397–9404, Nov. 2016.

[12] X. Zhu, C. Jiang, L. Kuang, N. Ge, and J. Lu, "Non-orthogonal multiple access based integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2253–2267, Oct. 2017.

[13] L. Kuang, X. Chen, C. Jiang, H. Zhang, and S. Wu, "Radio resource management in future terrestrial-satellite communication networks," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 81–87, Oct. 2017.

[14] B. Li, Z. Fei, X. Xu, and Z. Chu, "Resource allocations for secure cognitive satellite-terrestrial networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 78–81, Feb. 2018.

[15] F. Guidolin, M. Nekovee, L. Badia, and M. Zorzi, "A cooperative scheduling algorithm for the coexistence of fixed satellite services and 5g cellular network," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 1322–1327.

[16] X. Zhu, C. Jiang, L. Kuang, N. Ge, and J. Lu, "Energy efficient resource allocation in cloud based integrated terrestrial-satellite networks," in *Proc. IEEE Int. Conf. Commun.*, May 2018, pp. 1–6.

[17] M. Lin *et.al*, "Joint beamforming and power control for device-to-device communications underlaying cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 138–150, Jan. 2016.

[18] M. Lin, L. Yang, W.-P. Zhu, and M. Li, "An open-loop adaptive space-time transmit scheme for correlated fading channels," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 2, pp. 147–158, Apr. 2008.

[19] M. A. Vazquez *et.al*, "Hybrid analog-digital transmit beamforming for spectrum sharing satellite-terrestrial systems," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2016, pp. 1–5.

[20] B. Li, Z. Fei, Z. Chu, F. Zhou, K.-K. Wong, and P. Xiao, "Robust chance-constrained secure transmission for cognitive satellite–terrestrial networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4208–4219, May. 2018.

[21] K. An, M. Lin, J. Ouyang, and W.-P. Zhu, "Secure transmission in cognitive satellite terrestrial networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 3025–3037, Nov. 2016.

[22] M. Lin, Z. Lin, W.-P. Zhu, and J.-B. Wang, "Joint beamforming for secure communication in cognitive satellite terrestrial networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 1017–1029, May 2018.

[23] 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. Commun.*, vol. 68, no. 6, pp. 3834–3862, Jun. 2020.

[24] S. Biswas *et.al*, "Design and analysis of FD MIMO cellular systems in coexistence with MIMO radar," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4727–4743, Jul. 2020.

[25] A. Kaushik, C. Masouros, and F. Liu, "Hardware efficient joint radar-communications with hybrid precoding and RF chain optimization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[26] F. Liu *et.al*, "Toward dual-functional radar-communication systems: Optimal waveform design," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4264–4279, Aug. 2018.

[27] K. Singh, S. Biswas, T. Ratnarajah, and F. A. Khan, "Transceiver design and power allocation for full-duplex MIMO communication systems with spectrum sharing radar," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 556–566, Sep. 2018.

[28] T. Tian, T. Zhang, L. Kong, and Y. Deng, "Transmit/receive beamforming for MIMO-OFDM based dual-function radar and communication," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4693–4708, May 2021.

[29] L. You, X. Qiang, C. G. Tsinos, F. Liu, W. Wang, X. Gao, and B. Ottersten, "Beam squint-aware integrated sensing and communications for hybrid massive MIMO LEO satellite systems," *IEEE J. Sel. Areas Communications*, vol. 40, no. 10, pp. 2994–3009, 2022.

[30] B. Zhao, M. Wang, Z. Xing, G. Ren, and J. Su, "Integrated sensing and communication aided dynamic resource allocation for random access in satellite terrestrial relay networks," *IEEE Commun. Lett.*, vol. 27, no. 2, pp. 661–665, Feb. 2023.

[31] S. Pala, M. Katwe, K. Singh, B. Clerckx, and C.-P. Li, "Spectral-efficient RIS-aided rsma URLLC: Toward mobile broadband reliable low latency communication (mBRLLC) system," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3507–3524, Apr. 2024.

[32] A. Fascista *et.al*, "RIS-aided joint localization and synchronization with a single-antenna receiver: Beamforming design and low-complexity estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1141–1156, 2022.

[33] S. P. Chepuri, N. Shlezinger, F. Liu, G. C. Alexandropoulos, S. Buzzi, and Y. C. Eldar, "Integrated sensing and communications with reconfigurable intelligent surfaces," *arXiv preprint arXiv:2211.01003*, 2022.

[34] R. P. Sankar, S. P. Chepuri, and Y. C. Eldar, "Beamforming in integrated sensing and communication systems with reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, 2023.

[35] R. Liu, M. Li, and A. L. Swindlehurst, "Joint beamforming and reflection design for RIS-assisted ISAC systems," in *Proc. IEEE Eur. Signal Process. Conf. (EUSIPCO)*, 2022, pp. 997–1001.

[36] M. Hua, Q. Wu, C. He, S. Ma, and W. Chen, "Joint active and passive beamforming design for IRS-aided radar-communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2278–2294, 2022.

[37] M. Wu, K. Guo, X. Li, A. Nauman, K. An, and J. Wang, "Optimization design in RIS-assisted integrated satellite-UAV-served 6G IoT: A deep reinforcement learning approach," *IEEE Internet Things Mag.*, vol. 7, no. 1, pp. 12–18, Jan. 2024.

[38] Z. Lin *et.al*, "Refracting RIS-aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3717–3724, Aug. 2022.

[39] F. Vatalaro, G. Corazza, C. Caini, and C. Ferrarelli, "Analysis of LEO, MEO, and GEO global mobile satellite systems in the presence of interference and fading," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 2, pp. 291–300, Feb. 1995.

[40] K. Bessho, K. Date, M. Hayashi, A. Ikeda, T. Imai, H. Inoue, Y. Kumagai, T. Miyakawa, H. Murata, T. Ohno *et al.*, "An introduction to himawari-8/9—japan's new-generation geostationary meteorological satellites," *J. Meteorol. Soc. Jpn. Ser. II*, vol. 94, no. 2, pp. 151–183, 2016.

[41] D. Christopoulos *et.al*, "Multicast multigroup precoding and user scheduling for frame-based satellite communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4695–4707, Sep. 2015.

[42] W. J. Szajnowski, "Estimators of log-normal distribution parameters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-13, no. 5, pp. 533–536, Sep. 1977.

[43] P. Stoica, J. Li, and Y. Xie, "On probing signal design for MIMO radar," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4151–4161, Aug. 2007.

[44] L. Yin *et.al*, "Rate-splitting multiple access for satellite-terrestrial integrated networks: Benefits of coordination and cooperation," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 317–332, Jan. 2023.

[45] Y. Xing and T. S. Rappaport, "Terahertz wireless communications: Co-sharing for terrestrial and satellite systems above 100 ghz," *IEEE Commun. Lett.*, vol. 25, no. 10, pp. 3156–3160, Oct. 2021.

[46] X. Zhu and C. Jiang, "Integrated satellite-terrestrial networks toward 6G: Architectures, applications, and challenges," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 437–461, 2021.

[47] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[48] X. Yin, Y. Zhu, and J. Hu, "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–36, 2021.

[49] W. Xu, J. An, C. Huang, L. Gan, and C. Yuen, "Deep reinforcement learning based on location-aware imitation environment for RIS-aided mmwave MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1493–1497, Jul. 2022.

[50] Z. Peng, Z. Zhang, L. Kong, C. Pan, L. Li, and J. Wang, "Deep reinforcement learning for RIS-aided multiuser full-duplex secure communications with hardware impairments," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21 121–21 135, Nov. 2022.

[51] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[52] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[53] Z. Liu, L. Yin, W. Shin, and B. Clerckx, "Max-min fair energy-efficient beam design for quantized isac leo satellite systems: A rate-splitting approach," *arXiv preprint arXiv:2402.09253*, 2024.

[54] R. Zhang, K. Xiong, Y. Lu, P. Fan, D. W. K. Ng, and K. B. Letaief, "Energy efficiency maximization in RIS-assisted SWIPT networks with RSMA: A PPO-based approach," *IEEE J. Sel. Areas in Commun.*, vol. 41, no. 5, pp. 1413–1430, May 2023.

[55] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[56] Q. V. Do, Q.-V. Pham, and W.-J. Hwang, "Deep reinforcement learning for energy-efficient federated learning in uav-enabled wireless powered networks," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 99–103, Jan. 2022.

[57] P. Tehrani, F. Restuccia, and M. Levorato, "Federated deep reinforcement learning for the distributed control of NextG wireless networks," in *2021 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 2021, pp. 248–253.

[58] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[59] Z. He, W. Xu, H. Shen, D. W. K. Ng, Y. C. Eldar, and X. You, "Full-duplex communication for ISAC: Joint beamforming and power optimization," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 9, pp. 2920–2936, Sep. 2023.

[60] F. D. Calabrese *et.al*, "Learning radio resource management in RANs: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138–145, 2018.
[61] D. Nguyen, L.-N. Tran, P. Pirinen, and M. Latva-aho, "On the spectral efficiency of full-duplex small cell wireless systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 4896–4910, Sep. 2014.
[62] L. Chai, L. Bai, T. Bai, J. Shi, and A. Nallanathan, "Secure RIS-aided MISO-NOMA system design in the presence of active eavesdropping," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 19 479–19 494, Nov. 2023.
[63] S. Pala, K. Singh, M. Katwe, and C.-P. Li, "Joint optimization of URLLC parameters and beamforming design for multi-RIS-aided MU-MISO URLLC system," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 148–152, Jan. 2023.
[64] S. Pala, O. Taghizadeh, M. Katwe, K. Singh, C.-P. Li, and A. Schmeink, "Secure RIS-assisted hybrid beamforming design with low-resolution phase shifters," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2024.
[65] P. Saikia, S. Pala, K. Singh, S. K. Singh, and W.-J. Huang, "Proximal policy optimization for RIS-assisted full duplex 6G-V2X communications," *IEEE Trans. Intell. Veh.*, pp. 1–16, 2023.
[66] S. Pala, M. Katwe, K. Singh, T. A. Tsiftsis, and C.-P. Li, "Robust design of RIS-aided full-duplex RSMA system for V2X communication: A DRL approach," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, Dec. 2023, pp. 2420–2425.
[67] Y. Gu, Y. Cheng, C. L. P. Chen, and X. Wang, "Proximal policy optimization with policy feedback," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 7, pp. 4600–4610, Jul. 2022.