UAV-enabled Integrated Sensing, Communication and Control: A Constrained RL Approach

Qingliang Li, *Graduate Student Member, IEEE,* Bin Li*, *Senior Member, IEEE,* Yue Rong, *Senior Member, IEEE,* Zhen-Qing He, *Member, IEEE,* and Zhu Han, *Fellow, IEEE,*

***Abstract*—In this paper, we propose a time-division integrated sensing, communication and control (ISCC) scheme designed to dynamically enhance communication and sensing capabilities on the UAV platform. The UAV is dispatched to track a randomly moving target for capturing and transmitting sensing data to the base station via wireless communication. The goal is to leverage the ISCC framework for maximizing the cumulative sensing mutual information while guaranteeing successful data transmission by optimizing the allocation of the communication and sensing time slots together with the UAV's control scheme. The formulated problem cannot be straightforwardly solved by off-the-shelf optimization algorithms due to the time-varying environment. To tackle this challenge, a constrained soft actor-critic (C-SAC) algorithm is developed, which dynamically switches between maximizing rewards and minimizing constraint violations to ensure robust performance in changing environments while maintaining the simplicity and efficiency of unconstrained policy optimization. Simulation results demonstrate that the proposed C-SAC algorithm outperforms dual-variable-based methods in handling the constrained problems, while extensive Monte Carlo tests confirm the robustness of the ISCC policy trained by the proposed algorithm, which adapts to varying target speeds and achieves higher cumulative mutual information compared to the point-mass UAV models.**

***Index Terms*—Integrated sensing, communication and control (ISCC), unmanned aerial vehicle (UAV), constrained reinforcement learning (CRL), trajectory planning.**

I. Introduction

Integrated sensing and communication (ISAC) is regarded as one of the six ITU use scenarios in 6G wireless networks [1]–[3], which not only provides high-throughput and low-latency communication services, but also affords high-precision sensing capabilities. As a flexible and cost-effective aerial platform, unmanned aerial vehicles (UAVs) have been widely used in various fields such as reconnaissance, disaster rescue operations [4], traffic monitoring [5], agriculture, forestry and animal husbandry operations, logistics and transportation [6]. Recently, UAV-enabled ISAC has attracted more attention, prompting a surge in academic research endeavors.

To fully exploit the potentials of UAV and ISAC, the optimal design of UAV-ISAC networks has been intensively studied, including UAV trajectory planning and resource allocation. [7] focused on minimizing the power consumption of a UAV-enabled ISAC system through the joint design of trajectory and resource allocation. In [8], the UAV utilized ISAC signals to provide communication services to ground users and locate unknown-positioned targets, optimizing its 2D flight path and hover points to balance the communication performance with the target positioning accuracy. However, it is noted that the UAV in [7], [8] is restricted to providing communication and sensing services solely during its hovering phases. In [9], a time-division multiple access (TDMA) based periodic ISAC mechanism was proposed. Simulation results underscore the importance of UAV trajectory design in effectively balancing the communication and sensing performance. Furthermore, in [10], a comprehensive framework was developed to encompass the joint optimization of UAV trajectory or static deployment and ISAC beamforming, addressing constraints such as sensing beam gain, transmission power, and flight limitations, with the aim of maximizing the communication users' average weighted sum-rate throughput.

The existing studies in the literature discussed above primarily utilize numerical optimization techniques to achieve suboptimal solutions with prior knowledge of the environment. However, the limited computational resources present a challenge in providing real-time online optimization support for UAVs operating in dynamically changing environments. Deep reinforcement learning (DRL) has attracted considerable attention as a method for tackling the sequential decision-making challenges of UAV-enabled ISAC systems in time-varying environments [11]. Recently, several studies have investigated the utilization of DRL to improve the capabilities of UAVs in terms of trajectory planning and resource allocation in unfamiliar environments. For instance, the study by [12] specifically addressed the optimization of service area selection and power allocation for UAVs during emergency communication scenarios. [13] investigated the enhancement of obstacle perception accuracy through the utilization of ISAC for designing

Manuscript received Jun. 07, 2025; revised Sept. 10, 2025; and accepted Sept. 30, 2025. This work was supported by the National Natural Science Foundation of China under Grant U24B20156 and the National Defense Basic Scientific Research Program of China under Grant JCKY2021204B051, and it is partially supported by NSF ECCS-2302469, CMMI-2222810, Toyota. Amazon and Japan Science and Technology Agency (JST) Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) JPMJAP2326. (*Corresponding author: Bin Li.)

Q. Li, B. Li and Z. He are with the School of Aeronautics and Astronautics, Sichuan University, Chengdu, 610065, China, e-mail:liqingliang@std.uestc.edu.cn, bin.li@scu.edu.cn; zhenqinghe@scu.edu.cn.

Y. Rong is with School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, GPO Box U1987, Perth, WA 6845, Australia, e-mail: y.rong@curtin.edu.au

Z. Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701, e-mail: hanzhu22@gmail.com.

UAV flight paths that circumvent obstacles. Furthermore, [14] delved into the application of DRL for the purpose of planning UAV 3D trajectories to improve user access rates, fairness, and energy efficiency. Moreover, [15] utilized DRL to tackle a challenge related to joint user association, UAV trajectory planning, and power allocation to enhance spectral efficiency. In addition, [16] investigated the application of multi-agent DRL algorithms in trajectory planning and resource allocation for multi-UAV systems.

However, it is worth noting that the aforementioned studies treat UAVs as point-masses and only take into account their kinematic paths. Given the significant difference between the oversimplified point-mass model and the actual UAV system, the planned UAV trajectory may not be fully trackable in practice, thus diminishing the feasibility of planned trajectories [17]. To improve the feasibility of planned UAV trajectory, [17] and [18] suggested using deep learning for dynamic trajectory planning of fixed-wing UAVs. For rotary-wing UAVs, [19] and [20] proposed breaking down the nonlinear dynamic trajectory planning into sub-problems solvable with convex optimization. However, [17]–[20] focused on UAV trajectory planning in scenarios involving obstacle avoidance. In the communications community, the trajectory planning that integrates UAV dynamics for UAV-enabled communication or sensing has not yet received sufficient research attention. Our prior works [21], [22] demonstrated that neglecting the UAV dynamics in the trajectory design can lead to degradation in the quality of service (QoS) of communication. Compared with the DRL schemes [12]–[16], [23] that treat the UAV as a mass point, the complexity of the problem is considerably increased when the dynamics of the UAV are incorporated. For instance, DRL frequently demonstrates limited efficacy in addressing the safety constraints and communication and sensing (C&S) constraints associated with UAV operations. This limitation suggests a shift in research focus towards safe reinforcement learning (SRL) [24] or constrained reinforcement learning (CRL) [25] methodologies.

This shift is exemplified by recent works: the authors in [26] proposed a SRL-based joint optimization framework for terahertz (THz)-band UAV-assisted communication networks. By co-optimizing the UAV's trajectory planning and dynamic channel allocation strategy, the study maximized network energy efficiency. Furthermore, [27] investigated UAV trajectory design and power allocation based on causal channel state information (CSI), while utilizing the CRL method to optimize the system's total transmission rate under average rate constraints for users. However, it is important to note that both studies [26] and [27] employed the Lagrangian multiplier method to manage constraints. Although this approach provides mathematical simplicity in formulation, its practical application exhibits significant limitations [24], [25]: (i) The convergence performance of Lagrangian multipliers is highly sensitive to initial parameter settings. (ii) The method exhibits poor adaptability in dynamic time-varying environments. (iii) Performance degradation becomes particularly evident when addressing sparse constraints (e.g., event-triggered safety conditions) or long-term cumulative constraints (e.g., temporal average requirements for latency or throughput).

Motivated by above discussion in this paper, our focus lies in the ongoing monitoring of randomly moving ground targets within a cellular-connected UAV-ISAC system. Operating under energy constraint, the UAV is required to sense the target efficiently and transmit sensing data to the base station (BS) in real-time. Our objective is to maximize the mutual information (MI) obtained from sensing the target, achieved through the integrated sensing, communication and control (ISCC) of the UAV to optimize its real-time trajectory and the allocation of communication and sensing time slots. In contrast to our prior studies [21], [22], this paper investigates a non-cooperative and non-adversarial (NCNA) sensing target with unpredictable movement patterns, which restricts the system to acquiring only causal CSI of the sensing channel. This characteristic renders conventional non-causal CSI-based approaches [7]–[10], [21], [22] ineffective for addressing the problem investigated in this study. Furthermore, the continuous control of UAVs becomes particularly challenging when simultaneously considering multiple constraints including communication quality, safety requirements, and energy consumption. To the best of our knowledge, no prior research has been conducted on this specific issue. In order to tackle this challenge, we customize the reward functions and the constrained soft actor-critic (C-SAC) algorithm, enabling the UAV to learn how to adjust its policy for monitoring randomly moving target.

In summary, the primary contributions of this paper can be outlined as follows:

- We propose a time-division ISCC (TD-ISCC) frame structure that can dynamically adjust the time slot allocation to maximize the ISAC rate, which represents the successful transmission of both the sensing and video data received by the UAV to the BS, reflecting the capacity of the ISAC system. Furthermore, we derive a rigorous closed-form solution for the allocation of communication and sensing time slots within this framework.
- We formulate the ISCC problem by integrating UAV dynamics, energy limitation, and safety prerequisites, resulting in a multi-constraint optimization problem. By customizing the reward and cost functions for each constraint, we transform this problem into a constrained Markov decision process (CMDP). Subsequently, for the single-step decision-making of UAVs, we construct the single-step reward and cost functions by incorporating the rewards and costs associated with all constraints, thereby aligning with the objectives of the ISCC problem. This approach simplifies the decision-making process.
- The C-SAC algorithm is proposed to address the ISCC problem within a UAV-based TD-ISCC system operating in changing environments. The algorithm leverages an unconstrained SAC policy to maximize rewards, thereby updating the network parameters. However, when a constraint violation is detected, C-SAC switches to a corrective mechanism: it employs the unconstrained policy to minimize the associated cost, updates the network parameters accordingly, and temporarily adjusts the policy in the regression direction of the violated constraint. Notably, the implementation of C-SAC maintains the simplicity

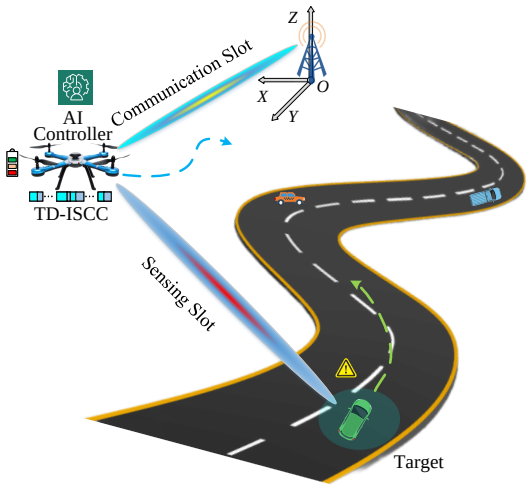


Fig. 1. An AI-controlled UAV using the TD-ISCC technique to sense moving targets.

and computational efficiency of traditional unconstrained policy optimization algorithms.

- Simulation results demonstrate that, compared to methods involving dual variables, C-SAC exhibits higher effectiveness in handling constrained problems. In addition, an extensive Monte Carlo tests demonstrate the robustness of the ISCC policy trained by the proposed algorithm. This policy can adapt to various target speeds and produces a greater cumulative MI compared to methods that consider the UAV as a point-mass model.

II. System Model and Problem Formulation

In Fig. 1, an AI-controlled quad-rotor UAV equipped with optical cameras and the TD-ISCC system is shown to conduct continuous surveillance over a specific area. The monitored target is subject to real-time traffic conditions on the ground, such as the presence of other vehicles and road bends, resulting in a time-varying speed. In this task, the UAV effectively gathers precise information about the target using both radar signals and video data through flexible flight maneuvers. The data acquired from the target is then transmitted to the BS in real-time, enabling timely and accurate monitoring and surveillance capabilities.

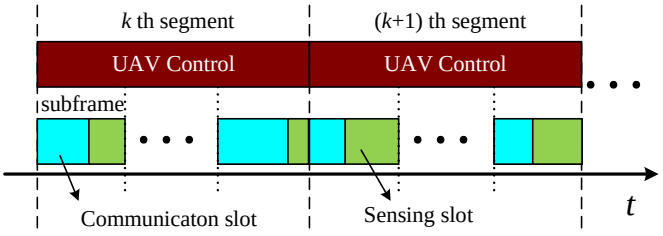


Fig. 2. The UAV control signal and the flexible sensing and communication frame structure of the proposed TD-ISCC system.

Fig. 2 depicts the UAV dynamic control signal segments and the flexible sensing and communication frame structure of the proposed TD-ISCC system. During the flight, the motion control commands of the UAV are divided into multiple segments. Each segment has the same duration and is further divided into multiple subframes, with each subframe containing both a communication slot and a sensing slot. The time proportion of the sensing slot in the subframe, denoted by ρ, can be adjusted flexibly according to the specific requirements of the surveillance mission. This flexible structure allows the UAV to efficiently balance its communication and sensing tasks while adapting to varying mission conditions and priorities.

We set the duration of each segment to dt, and the duration of each subframe can be designed to be 1 ms or even shorter. As a result, ρ can be approximated as a time-continuous variable, denoted as $\rho(t)$, representing its value at time t. In this scenario, the UAV and the target positions at time t are denoted by $\boldsymbol{p}(t) = [x(t), y(t), z(t)]$ and $\boldsymbol{p}_T(t) = [x_T(t), y_T(t), 0]$, respectively. In addition, the location of the BS is represented as $\boldsymbol{p}_B = [x_B, y_B, z_B]$.

Next, we provide a detailed description of the subsystem models and the problem formulation.

A. UAV Sensing Metric

The signal-to-noise ratio (SNR) of the radar signal reflected back from the monitored target to the UAV can be expressed as follows:

$$\Gamma_r(t \mid \boldsymbol{p}_T(t)) = \frac{P_t \mathbb{E}\left[|h_r(t \mid \boldsymbol{p}_T(t))|^2\right]}{\sigma_n^2}, \tag{1}$$

where P_t is the transmission power, σ_n^2 is the variance of the additive white Gaussian noise (AWGN). The non-cooperative channel gain can be obtained by [28]

$$\mathbb{E}\left[|h_r(t \mid \boldsymbol{p}_T(t))|^2\right] = \frac{G^2 \lambda^2 \sigma_{\text{cross}}}{(4\pi)^3 \|\boldsymbol{p}(t) - \boldsymbol{p}_T(t)\|^4}, \tag{2}$$

where G is the antenna gain of the UAV (assume that the transmitting and receiving antennas have the same gain), λ is the carrier wavelength, σ_{cross} is the target cross section. We investigate a general scenario where the UAV is assigned to sense an NCNA moving target. In this situation, the UAV can only obtain the target's current position and velocity information through perception technologies and is unable to predict the target's future trajectory. The motion equation of the target can be expressed as

$$\dot{\boldsymbol{p}}_T(t) = \boldsymbol{v}_T(t), \tag{3}$$

where $\boldsymbol{v}_T(t) = [v_{Tx}(t), v_{Ty}(t), 0]$ is the velocity of target. Under the aforementioned conditions, the radar MI estimation rate can be derived, which quantifies the system's capability to extract target-related information from echo signals. Consequently, we employ the radar MI estimation rate as a key metric to evaluate the UAV's sensing performance. For a given target position $\boldsymbol{p}_T(t)$, the UAV's sensing capability can be expressed as [29]–[31]

$$R_r\left(t \mid \boldsymbol{p}_T(t)\right) = B\rho(t) \log_2\left(1 + \frac{\lambda_1}{\|\boldsymbol{p}(t) - \boldsymbol{p}_T(t)\|^4}\right). \tag{4}$$

Here B is the ISAC system bandwidth and

$$\lambda_1 = \frac{P_t G^2 \lambda^2 \sigma_{\text{cross}}}{(4\pi)^3 \sigma_n^2}$$

represents the SNR at a reference distance of 1 meter.

B. UAV Sensing Information Transmission

Generally speaking, the transceiver antenna of the BS is positioned at a high location, enabling the establishment of a strong line-of-sight (LoS) channel with the UAV. In addition, it is assumed that the Doppler effect resulting from the UAV's flying motion is perfectly compensated at both the BS [32], [33] and the UAV sensing receiver [9], [34], [35]. This compensation ensures that any frequency shifts caused by the UAV's movement are effectively mitigated, leading to more reliable and accurate communication and sensing performance. Therefore, the SNR of the communication signal received by the BS from the UAV can be calculated as

$$\Gamma_c(t) = \frac{P_t \mathbb{E}\left[|h_c(t)|^2\right]}{\sigma_n^2}, \tag{5}$$

where $h_c(t)$ is the communication channel gain, which follows the free-space path loss model and the channel power gain from the UAV to the BS can be expressed as [28]

$$\mathbb{E}\left[|h_c(t)|^2\right] = \frac{GG_B\lambda^2}{(4\pi)^2\|\boldsymbol{p}(t)-\boldsymbol{p}_B\|^2}, \tag{6}$$

where G_B is the receiving antenna gain of the BS. Let

$$\lambda_2 = \frac{P_t G G_B \lambda^2}{(4\pi)^2\sigma_n^2}, \tag{7}$$

and then the data rate of the detection information transmitted by the UAV to the BS can be written as

$$R_c(t) = B\left(1-\rho(t)\right)\log_2\left(1+\frac{\lambda_2}{\|\boldsymbol{p}(t)-\boldsymbol{p}_B\|^2}\right). \tag{8}$$

We adopt the ISAC rate $\min\{R_c(t),\ R_r\left(t \mid \boldsymbol{p}_T(t)\right)+R_v\}$ to represent the capacity of the ISAC system, which characterizes the successful transmission of both the sensing and video data received by the UAV to the BS. To guarantee the successful transmission of the sensing and camera data collected by the UAV to the BS receiver, it is imperative that the communication rate of the UAV exceeds the combined rates of the video data and the MI from the sensing data. The constraint on the communication rate is as follows [36], [37]

$$R_c(t) \geq R_r\left(t \mid \boldsymbol{p}_T(t)\right) + R_v, \tag{9}$$

where R_v is the video data rate and is assumed to be stable at specific resolutions. Through analysis, we can derive a closed-form analytic constraint for $\rho(t)$ in (9), which can be expressed as

$$0 \leq \rho(t) \leq \frac{\log_2\left(1+\Gamma_c(t)\right) - R_v/B}{\log_2\left(1+\Gamma_c(t)\right)+\log_2\left(1+\Gamma_r\left(t \mid \boldsymbol{p}_T(t)\right)\right)}. \tag{10}$$

C. Energy Consumption Model

We consider a battery-powered electric quadcopter UAV, where the total energy available to perform the task is denoted as $E_{\max}$, due to the limitation of the onboard energy capacity. The remaining energy at time t can be expressed as

$$E_r(t) = E_{\max} - \int_0^t \left(P(\tau)+P_t+P_o\right)\,d\tau, \tag{11}$$

TABLE I
NOTATIONS AND TERMINOLOGIES

Notation	Terminology	Unit
C_t	Thrust coefficient	$\mathrm{N/(rad/s)^2}$
C_m	Torque coefficient	$\mathrm{N\cdot m/(rad/s)^2}$
C_{dx}	Drag coefficient of x-axis	$\mathrm{N/(m/s)^2}$
C_{dy}	Drag coefficient of y-axis	$\mathrm{N/(m/s)^2}$
C_{dz}	Drag coefficient of z-axis	$\mathrm{N/(m/s)^2}$
C_{mx}	Damping torque coefficient of x-axis	$\mathrm{N\cdot m/(rad/s)^2}$
C_{my}	Damping torque coefficient of y-axis	$\mathrm{N\cdot m/(rad/s)^2}$
C_{mz}	Damping torque coefficient of z-axis	$\mathrm{N\cdot m/(rad/s)^2}$
g	Acceleration of gravity	$\mathrm{m/s^2}$
I_x	Rotational inertia of x-axis	$\mathrm{kg\cdot m^2}$
I_y	Rotational inertia of y-axis	$\mathrm{kg\cdot m^2}$
I_z	Rotational inertia of z-axis	$\mathrm{kg\cdot m^2}$
I_m	Motor propeller inertia	$\mathrm{kg\cdot m^2}$
L	Fuselage length	m
m	Aircraft mass	kg

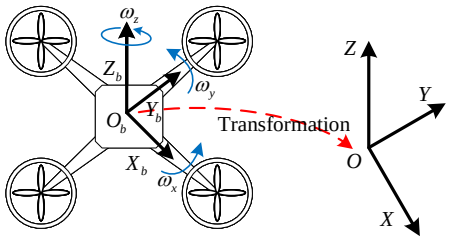


Fig. 3. Six DoF quadcopter coordinate transformation.

where P_o is an average power constant, encompassing the power consumption of optical cameras and other necessary onboard equipment of the UAV. As derived in [21], the propulsion power consumption of the UAV's motors can be expressed as

$$P(t) = \sum_{i=1}^{4} P_i(t), \tag{12}$$

where $P_i(t) = c_4\omega_i^4(t)+c_3\omega_i^3(t)+c_2\omega_i^2(t)+c_1\omega_i(t)+c_0$ represents the power consumption of each motor $i=\{1,2,3,4\}$. Here, c_0, c_1, c_2, c_3, and c_4 are constant parameters related to the motor [21], while $\omega_i(t)$ denotes the angular velocity of each motor $i \in \{1,2,3,4\}$ as illustrated in Fig. 3.

D. Quadcopter Rigid-Body Dynamic Model

As shown in Fig. 3, the UAV is equipped with four motors that enable it to perform six degree-of-freedom (6-DoF) movement. For simplicity, some of the terms and symbols related to UAV are listed in Table I. We denote the velocity vector of the UAV as $\boldsymbol{v}(t) = [v_x(t), v_y(t), v_z(t)]$, which is defined as

$$\boldsymbol{v}(t) = \dot{\boldsymbol{p}}(t). \tag{13}$$

In this paper, $\dot{\boldsymbol{p}}$ represents the derivative of $\boldsymbol{p}$ with respect to time t, following the conventional notation for time derivatives. The derivatives of other variables are defined similarly. We denote the UAV control variables as $\boldsymbol{u}(t) = [u_1(t), u_2(t), u_3(t), u_4(t)]$, where $u_1(t)$ represents the acceleration induced by the motors' tension, while $u_2(t)$, $u_3(t)$, and $u_4(t)$ correspond to angular accelerations around the X_b, Y_b, and Z_b axes, respectively. As illustrated in Fig. 3, the acceleration of the UAV is transformed from the body coordinate

system $O_b - X_bY_bZ_b$ to the Earth coordinate system $O-XYZ$ through a coordinate transformation. Consequently, the UAV acceleration can be obtained as [38]

$$\begin{aligned}\dot{v}_x(t) =&2u_1(t)\left[q_1(t)q_3(t)+q_0(t)q_2(t)\right]-D_x(t)/m,\\ \dot{v}_y(t) =&2u_1(t)\left[q_2(t)q_3(t)-q_0(t)q_1(t)\right]-D_y(t)/m,\\ \dot{v}_z(t) =&u_1(t)\left[q_0(t)^2-q_1(t)^2-q_2(t)^2+q_3(t)^2\right]\\ &-D_z(t)/m-g.\end{aligned} \tag{14}$$

The air drag of the UAV is proportional to the square of the flight speed and can be expressed as follows: $D_x(t) = C_{dx}|v_x(t)|v_x(t)$, $D_y(t) = C_{dy}|v_y(t)|v_y(t)$, and $D_z(t) = C_{dz}|v_z(t)|v_z(t)$. The quaternion $\boldsymbol{q}(t) = [q_0(t), q_1(t), q_2(t), q_3(t)]^\top$ is used to describe the body attitude of UAV, where '$[\cdot]^\top$' stands for the transpose operation. It is updated according to the following equation [39]:

$$\dot{\boldsymbol{q}}(t) = \frac{1}{2}\boldsymbol{\Omega}_q(t)\boldsymbol{q}(t), \tag{15}$$

where

$$\boldsymbol{\Omega}_q(t) = \begin{bmatrix} 0 & -\omega_x(t) & -\omega_y(t) & -\omega_z(t) \\ \omega_x(t) & 0 & \omega_z(t) & -\omega_y(t) \\ \omega_y(t) & -\omega_z(t) & 0 & \omega_x(t) \\ \omega_z(t) & \omega_y(t) & -\omega_x(t) & 0 \end{bmatrix}$$

and $\boldsymbol{\omega}(t) = [\omega_x(t), \omega_y(t), \omega_z(t)]$ represent the angular velocities around the UAV frame's respective axes. The angular velocities are controlled by the UAV's control signals $\boldsymbol{u}(t)$, which are given by [40]

$$\begin{aligned}\dot{\omega}_x(t) =&u_2(t)+\left[(I_y-I_z)\omega_y(t)\omega_z(t)-I_m\Omega(t)\omega_y(t)\right.\\ &\left.-D_{mx}\right]/I_x,\\ \dot{\omega}_y(t) =&u_3(t)+\left[(I_z-I_x)\omega_x(t)\omega_z(t)+I_m\Omega(t)\omega_x(t)\right.\\ &\left.-D_{my}\right]/I_y,\\ \dot{\omega}_z(t) =&u_4(t)+\left[(I_x-I_y)\omega_x(t)\omega_y(t)-D_{mz}\right]/I_z,\end{aligned} \tag{16}$$

where $\Omega(t) = \omega_1(t) - \omega_2(t) + \omega_3(t) - \omega_4(t)$. Due to the symmetry of the UAV body, the moment of inertia I_x around the X_b-axis is equal to the moment of inertia I_y around the Y_b-axis. The damping torque of the UAV body is proportional to the square of the angular velocities and can be expressed as follows: $D_{mx} = C_{mx}|\omega_x(t)|\omega_x(t)$, $D_{my} = C_{my}|\omega_y(t)|\omega_y(t)$, and $D_{mz} = C_{mz}|\omega_z(t)|\omega_z(t)$.

Intuitively, (14)-(16) capture the UAVs rigid-body dynamics in a physically meaningful way. (14) governs the translational acceleration along the three axes of the Earth frame, where the thrust generated by the motors is projected into the global coordinates through quaternion-based rotation, and the aerodynamic drag and gravity are also considered. (15) describes the time evolution of the UAVs attitude in terms of quaternion dynamics, ensuring a smooth representation of 3D orientation. (16) characterizes the rotational dynamics, showing how the control torques around the roll, pitch, and yaw axes $[u_2(t), u_3(t), u_4(t)]$ interact with gyroscopic effects and body damping to update the angular velocities. Taken together, these equations indicate that the UAVs next position and attitude are determined by the interplay of thrust, torque, aerodynamic drag, gravity, and inertia, which makes the model significantly more realistic than the commonly used point-mass model.

E. Problem Formulation

Based on the discussion of the models above, this section formulates the ISCC problem for the UAV TD-ISCC system. In this problem, the UAV starts from a point and continuously tracks and monitors a target while sending real-time sensing data to the BS. In addition, the UAV must reach a predetermined end point before depleting its onboard energy. The primary objective is to maximize the ISAC rate concerning the target, which is achieved by jointly optimizing the UAV's motor controls, communication and sensing ratio, and the time required to complete the mission. The goal is to achieve an efficient and effective surveillance task while optimizing the trade-off between energy consumption, sensing data quality, and the mission completion time. Since the video rate in the ISAC rate is a constant, our actual optimization objective shifts to maximizing the total radar MI estimation. Therefore, the ISCC problem for the UAV TD-ISCC system is formulated as

$$\begin{aligned}\mathcal{P}: \quad &\max_{\boldsymbol{u}(t),\rho(t),T} \quad Q_r(T) = \mathbb{E}\left[\int_0^T R_r(\tau \mid \boldsymbol{p}_T(\tau))\mathrm{d}\tau\right]\\ s.t. \quad &C_0: \ (13),(14),(15),(16),\\ &C_1: \ U_{1l} \le u_1(t) \le U_{1u},\\ &\qquad |u_i(t)| \le U_{iu}, \ i=2,3,4, \ t\in[0,T],\\ &C_2: \ T>0,\\ &C_3: \ (10),\\ &C_4: \ z(t) \ge h_{min}, \ t\in[0,T],\\ &C_5: \ E_r(T) \ge 0,\\ &C_6: \ \|\boldsymbol{p}(T)-\boldsymbol{p}_F\| \le \xi_d,\end{aligned}$$

where ξ_d is the critical distance.

Herein, C_0 include the dynamic equations of UAV. C_1 is introduced as a constraint to limit the force and torque produced by the UAV motors, where U_{1l} and U_{iu} $i = 1,2,3,4$ are boundary values. C_2 ensures that T satisfies real-world physical constraints. Constraint C_3 represents the communication requirements, guaranteeing the successful transmission of video data and radio sensing to the BS. For the safety of the UAV, its flight altitude is limited by constraint C_4, where h_{min} is the minimum allowable altitude. A critical factor for the mission success is reaching the end point before depletion of the onboard energy. C_5, the residual energy constraint, ensures sufficient energy residue at the final time. Constraint C_6 defines the permissible end point position for the UAV, where $\boldsymbol{p}_F = [x_F, y_F, z_F]$ is the terminal position.

In problem $\mathcal{P}$, the UAV cannot obtain the complete causal CSI of the sensing channels in advance, which makes traditional numerical optimization methods ineffective for obtaining the optimal solution. While online real-time computation approaches are theoretically feasible, they encounter fundamental trade-offs between optimizing the objective function, satisfying terminal constraints, and ensuring global energy efficiency. Although DRL techniques demonstrate promising capabilities in addressing sequential decision-making challenges induced by causal CSI, the inherent strong coupling between control variables and system states within the UAV dynamical model fundamentally constrains the ability to rigorously enforce state

constraints through purely hard-constrained neural network architectures, thus introducing substantial technical obstacles to effective problem resolution. To address these challenges, we propose an innovative and computationally efficient C-SAC algorithm.

III. CRL Based UAV Trajectory Design and Time Slot Allocation

Problem $\mathcal{P}$ is a sequential decision problem, which can be reformulated as a CMDP and then solved by SRL/CRL [24], [25]. In this section, the UAV is treated as an agent and trained with the C-SAC algorithm based on CRL to meet the objective and constraints in problem $\mathcal{P}$. At observation time t, the environment state, agent action, reward and cost are denoted as s_t, a_t, r_t and c_t, respectively. The state after taking action a_t is denoted as s_{t+1}. The $done$ variable is a binary flag that indicates whether the episode has ended. If the episode has ended, $done$ is 1, otherwise it is 0. The corresponding experience tuple $\langle s_t, a_t, r_t, c_t, s_{t+1}, done\rangle$ is stored in a experience replay buffer $\mathcal{B}$ for the training of the network. In the following section, we elaborate on the problem preprocessing, the definitions of system states, actions, rewards, and costs, as well as the composition of the C-SAC algorithm.

A. Preliminary Treatment of $\mathcal{P}$

In this subsection, we derive the optimal solution of $\rho(t)$ in $\mathcal{P}$ using an analytical method. Through analysis, we can derive a closed analytic expression of $\rho(t)$ in C_3 as Theorem 1, which simplifies the task of solving problem $\mathcal{P}$.

Theorem 1. *The objective function monotonically increases with $\rho(t)$. When $\rho(t)$ reaches its maximum value, the objective function becomes optimal. The optimal value of $\rho(t)$ is obtained as follows:*

$$\rho^*(t) = \frac{\log_2\left(1+\Gamma_c(t)\right) - R_v/B}{\log_2\left(1+\Gamma_c(t)\right) + \log_2\left(1+\Gamma_r\left(t \mid \boldsymbol{p}_T(t)\right)\right)} \quad s.t. \quad R_c(t) \geq R_v.$$

Proof. See Appendix A. □

B. Environment State Space

In problem $\mathcal{P}$, elements in the environment include the UAV, the BS and the moving target. For the UAV, the state variables that have an impact on the control decision include the current position coordinate $\boldsymbol{p}(t)$, the speed vector $\boldsymbol{v}(t)$, the current body rotation angular speed $\boldsymbol{\omega}(t)$, the quaternion used to describe the body attitude $\boldsymbol{q}(t)$, the remaining energy $E_r(t)$, and the MI throughput $Q_r(t)$, totaling 15 state variables. The BS's location $\boldsymbol{p}_B$ also has an impact on the UAV's trajectory planning. Both the target's present position $\boldsymbol{p}_T(t)$ and its moving speed $\boldsymbol{v}_T(t)$ are state variables that have an impact on the UAV control decision. In addition, for the UAV trajectory planning, it is necessary to have terminal position information $\boldsymbol{p}_F$.

Therefore, there are 26 states in the environment that have an impact on the UAV control decisions. Analysis reveals that the control decision solely pertains to the UAV's position in relation to both the base station and the target. Three states can be dropped when the relative position is utilized to describe the environmental states, which is advantageous for DRL. The position of the UAV relative to the BS, target, and the endpoint can be expressed as

$$\boldsymbol{p}_{UB}(t) = \boldsymbol{p}(t) - \boldsymbol{p}_B, \tag{17}$$

$$\boldsymbol{p}_{UT}(t) = \boldsymbol{p}(t) - \boldsymbol{p}_T(t), \tag{18}$$

$$\boldsymbol{p}_{UF}(t) = \boldsymbol{p}(t) - \boldsymbol{p}_F. \tag{19}$$

Therefore, the environment state space can be expressed as

$$s_t = \left\{ \begin{array}{l} \boldsymbol{p}_{UB}(t), \boldsymbol{v}(t), \boldsymbol{\omega}(t), \boldsymbol{q}(t), E_r(t), \\ Q_r(t), \boldsymbol{p}_{UF}(t), \boldsymbol{p}_{UT}(t), \boldsymbol{v}_T(t) \end{array} \right\}. \tag{20}$$

To ensure that the observed values of all environmental state variables fall within the range $[-1, 1]$, the observed values must be normalized [41]. Normalization aids in removing cross-dimensional influences between various state variables.

C. UAV Action Space

The action space of the UAV, denoted as $a_t = \{a_1(t), a_2(t), a_3(t), a_4(t)\}$, encompasses four dimensions, corresponding to the UAV's four control parameters. Given that the UAV's control variables adhere to the saturation control constraint C_1, the action components are constrained within the interval $[-1, 1]$, signifying that $a_i(t) \in [-1, 1]$ for $i = \{1, 2, 3, 4\}$. These action components $a(t)$ are responsible for determining the UAV's control inputs, and they can be mathematically expressed as follows:

$$\begin{aligned} u_1(t) &= U_{1l} + \left(U_{1u} - U_{1l}\right)\left(1 + a_1(t)\right)/2, \\ u_2(t) &= U_{2u} a_2(t), \\ u_3(t) &= U_{3u} a_3(t), \\ u_4(t) &= U_{4u} a_4(t). \end{aligned} \tag{21}$$

The UAV motors' angular velocities are controlled by input control variables, and its analytical expression can be derived from (21) as follows

$$\begin{aligned} \omega_1(t) &= 0.5\left(\tau_1(t) + \tau_4(t) - 2\tau_3(t)\right)^{0.5}, \\ \omega_2(t) &= 0.5\left(\tau_1(t) - \tau_4(t) + 2\tau_2(t)\right)^{0.5}, \\ \omega_3(t) &= 0.5\left(\tau_1(t) + \tau_4(t) + 2\tau_3(t)\right)^{0.5}, \\ \omega_4(t) &= 0.5\left(\tau_1(t) - \tau_4(t) - 2\tau_2(t)\right)^{0.5}, \end{aligned} \tag{22}$$

where $\tau_1(t) = mu_1(t)/C_t$, $\tau_2(t) = u_2(t)I_x/(C_t L)$, $\tau_3(t) = u_3(t)I_y/(C_t L)$, and $\tau_4(t) = u_4(t)I_z/C_m$.

D. Hybrid Reward and Cost Function Design

This study carefully designs the reward and cost functions to guide the UAV in maximizing the cumulative MI, subject to state constraints C_3 to C_6. Specifically, the objective function and constraints C_3 and C_4 are active throughout the entire UAV flight, and can be modeled as immediate rewards and costs at each step. In contrast, constraints C_5 and C_6 serve as terminal state constraints, typically modeled as sparse terminal rewards and costs.

1) The UAV motion space and the end of the episode settings: Setting an effective motion space for the UAV can avoid invalid exploration of the UAV, such as the UAV constantly lowering its altitude, the UAV continuously flying in the direction that violates the mission goal, etc. The motion space of the UAV is set to be larger than the space required for the normal flight of the UAV in the mission scenario. The UAV motion space is set as $\Omega_{3d} = \{[x_{min}, x_{max}]; [y_{min}, y_{max}]; [z_{min}, z_{max}]\}$, where x_{min}, y_{min} and z_{min} are the lower boundaries of the space, and x_{max}, y_{max}, and z_{max} are the space upper boundary values. If the UAV flies out of the motion space, the current episode is terminated and a cost is given

$$c_{3d}(t) = \|\boldsymbol{p}_{UF}(t)\|, \qquad \text{if } \boldsymbol{p}(t) \notin \Omega_{3d}, \text{else } 0. \tag{23}$$

When the UAV flies out of the motion space ($\boldsymbol{p}(t) \notin \Omega_{3d}$) or the given on-board energy runs out (constraint C_5), the episode ends, and the end flag is set as $done = 1$, otherwise $done = 0$.

2) Terminal reward: The training goal of this problem is that the UAV flies to the destination (constraint C_6) before the energy is exhausted, and the MI about the target is maximized during the flight. Therefore, the design of the terminal reward is related to the distance between the end point and the UAV, the residual energy, and the cumulative MI. The terminal reward and cost are obtained when the episode ends and they are designed to be

$$\begin{cases} r_d = \frac{\tilde{r}_d}{1+\varepsilon_d \|\boldsymbol{p}_{UF}(T)\|} + Q_r(T), \text{if } \|\boldsymbol{p}_{UF}(T)\| \le \xi_d, \text{else } 0, \\ c_d = \|\boldsymbol{p}_{UF}(T)\|, \qquad\qquad \text{if } \|\boldsymbol{p}_{UF}(T)\| > \xi_d, \text{else } 0, \end{cases} \tag{24}$$

where $\tilde{r}_d$ is the maximum terminal distance reward, ε_d is a positive constant parameter, and the cost c_d represents the degree of violation of constraint C_6. Specifically, in (24), $\tilde{r}_d$ is set to be on the same order of magnitude as $Q_r(T)$, ensuring that the agent balances attention between the terminal distance and cumulative MI. The parameter ε_d controls how the actual terminal distance $\|\boldsymbol{p}_{UF}(T)\|$ discounts the reward. An excessively large ε_d makes $\frac{\tilde{r}_d}{1+\varepsilon_d \|\boldsymbol{p}_{UF}(T)\|}$ too small in early training, leading the agent to ignore the terminal constraint, while an overly small value weakens its effect, hindering convergence. Empirically, $1 \le \varepsilon_d \xi_d \le 9$ gives satisfactory results.

3) Sensing reward: According to Theorem 1, transmission of the video information should be ensured during the UAV's flight, and as much sensing data as feasible should be collected. Therefor, the sensing reward is designed as

$$\begin{cases} r_s(t) = \varepsilon_s R_r(t), & \text{if } R_c(t) \ge R_v, \text{else } 0, \\ c_s(t) = R_c(t) - R_v, & \text{if } R_c(t) < R_v, \text{else } 0, \end{cases} \tag{25}$$

where ε_s is a positive constant parameter, and the cost $c_s(t)$ represents the degree of violation of constraint C_3.

4) Safe flight reward: To encourage continuous flying, the UAV is penalized for violating the safe flight height constraint and rewarded for keeping the safe flight. The safe flight reward is designed as

$$\begin{cases} r_h(t) = \tilde{r}_h, & \text{if } z(t) \ge h_{\min}, \text{else } 0, \\ c_h(t) = \varepsilon_h \left(h_{\min} - z(t)\right), & \text{if } z(t) < h_{\min}, \text{else } 0, \end{cases} \tag{26}$$

where $\tilde{r}_h$ is the safe flight reward, ε_h is the height penalty weight, and the cost $c_h(t)$ represents the degree of violation of constraint C_4.

5) Flight guidance reward: According to the expert knowledge, the UAV should fly to the target when the energy is sufficient, which reduces the sensing distance and enhances the sensing ability. When the UAV's energy level is low, it should fly towards its destination to ensure that it reaches the end point before running out of the onboard energy. Based on the above considerations, the flight guidance reward is set as

$$r_g(t) = \begin{cases} \max\left(\|\boldsymbol{p}_{UT}(t-1)\| - \|\boldsymbol{p}_{UT}(t)\|, 0\right), E_r(t) > \xi_e, \\ \max\left(\|\boldsymbol{p}_{UF}(t-1)\| - \|\boldsymbol{p}_{UF}(t)\|, 0\right), E_r(t) \le \xi_e, \end{cases} \tag{27}$$

where ξ_e is the critical energy value. According to (27), when the residual energy of the UAV is greater than the critical energy, the UAV can be rewarded for approaching the sensing target, and it is not penalized for moving away from the sensing target. This design is beneficial to avoid unreasonable ξ_e settings when the UAV needs to fly to the destination in advance without being penalized. The same is true for the guiding reward of the end point. When the value of ξ_e is unreasonable, the UAV uses more energy to sense the target without penalty.

6) Step reward: In summary, the step reward and cost can be formulated as

$$r_t = r_s(t) + r_h(t) + r_g(t) + done \cdot r_d, \tag{28}$$

$$c_t = c_s(t) + c_h(t) + c_{3d}(t) + done \cdot c_d. \tag{29}$$

For the process rewards $r_s(t)$, $r_h(t)$, and $r_g(t)$ in (28), the hyperparameters are tuned to normalize their values to a similar dimensionless range, ensuring that the agent pays balanced attention to different states. In contrast, the cost function is activated when a constraint is violated, and its value only needs to reflect the severity of the violation. For constraints that are more likely to be violated, the penalty can be amplified by increasing the corresponding hyperparameter, such as ε_h in (26).

7) Training objective: In CRL, the objective of the agent is to train a control policy π, which maximizes the cumulative discounted reward while adhering to constraints on the cumulative discounted cost. This objective can be expressed as follows

$$\begin{aligned} \mathcal{P}1: \quad & \max_{\pi} \quad \sum_{t=0}^{\infty} \gamma^t r_t \\ s.t. \quad & \sum_{t=0}^{\infty} \gamma^t c_t \le 0, \end{aligned}$$

where γ is the discount factor.

E. Preliminaries of SAC

SAC is an off-policy algorithm within the actor-critic (AC) framework [42]. The SAC framework, illustrated in Fig. 4, employs five neural networks: two parallel critic networks, e.g., $Q(s_t, a_t|\theta_{q_i})$, $i = \{1, 2\}$ with corresponding target networks, e.g., $Q(s_t, a_t|\theta_{\tilde{q}_i})$, $i = \{1, 2\}$, and one actor network, e.g., $\pi(s_t|\theta_\pi)$, where θ_{q_i}, $\theta_{\tilde{q}_i}$, and θ_π denote the trainable parameters of the respective networks. Specifically, the minimum Q-value from the two critic networks evaluates the actor's policy $\pi(s_t|\theta_\pi)$ (representing the action probability distribution), while the target networks provide stable Q-value estimates for

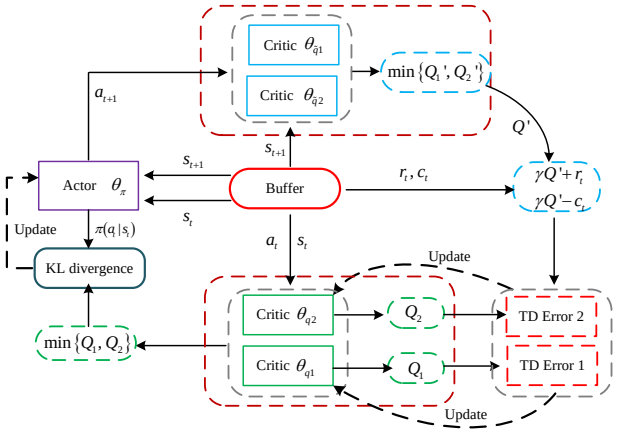


Fig. 4. The C-SAC algorithm architecture.

temporal difference learning. The real action can be obtained by sampling from the probability distribution of actions $a_t \sim \pi\left(s_t \mid \theta_\pi\right)$.

To address the issue of overestimation in the critic network, two critic target networks are employed for estimation. The estimation result is determined by selecting the one with the minimum valuation. The loss functions of the two critic networks (TD Error 1 and TD Error 2 in Fig. 4) can be represented as

$$\mathcal{L}\left(\theta_{q_i}\right)=\left(y_t-Q\left(s_t, a_t \mid \theta_{q_i}\right)\right)^2,\ i=\{1,2\}, \tag{30}$$

where

$$y_t=r_t+\gamma(1-done)[\min_{i=1,2} Q(s_{t+1}, \hat{a}_{t+1} \mid \theta_{\tilde{q}_i}) - \alpha \ln \pi(s_{t+1} \mid \theta_\pi)] \tag{31}$$

is the soft target Q value with the discount factor $0<\gamma<1$ and α is the temperature coefficient. The predicted action in (31) is given by

$$\hat{a}_{t+1} \sim \pi\left(s_{t+1} \mid \theta_\pi\right). \tag{32}$$

In addition, the policy network is optimized by reducing the Kullback-Leibler (KL) divergence between the action distribution and the soft Q-value s distribution. Therefore, the policy loss is defined as follows:

$$\mathcal{L}\left(\theta_\pi\right)=\alpha \ln \pi(s_t \mid \theta_\pi)-\min_{i=1,2} Q(s_t, a_t \mid \theta_{q_i}). \tag{33}$$

F. C-SAC Algorithm

To address the constraints in problem $\mathcal{P}1$, we propose a C-SAC algorithm that minimizes costs when constraints are violated and maximizes cumulative rewards when constraints are satisfied. Unlike the methods in [43], [44] that use separate cost and reward Q-functions, we employ a single Q-function to estimate the joint Q-value of costs and rewards, distinguishing between rewards and costs based on the sign of the Q-value (positive for rewards, negative for costs). In the implementation, when a constraint violation is detected, the negative cost value is used as a penalty to replace the reward term in (31), thereby minimizing costs during constraint violations and maximizing returns when constraints are satisfied. This unified Q-function design simplifies the algorithm structure and enhances the synergistic efficiency between constraint handling and reward optimization.

The loss functions of the two cost critic networks can be represented as

$$\mathcal{L}_c\left(\theta_{q_i}\right)=\left(y_t^c-Q_{rc}\left(s_t, a_t \mid \theta_{q_i}\right)\right)^2,\ i=\{1,2\}, \tag{34}$$

where

$$y_t^c=r_t^c+\gamma(1-done)[\min_{i=1,2} Q_{rc}(s_{t+1}, \hat{a}_{t+1} \mid \theta_{\tilde{q}_i}) - \alpha \ln \pi(s_{t+1} \mid \theta_\pi)], \tag{35}$$

and

$$r_t^c=\begin{cases} -c_t, & c_t>0, \\ r_t, & \text{otherwise.} \end{cases} \tag{36}$$

This constitutes a special case of the primal-dual method, with further details provided in Appendix B. Similar to (33), the loss function of the policy network can be expressed as

$$\mathcal{L}_c\left(\theta_\pi\right)=\alpha \ln \pi(s_t \mid \theta_\pi)-\min_{i=1,2} Q_{rc}(s_t, a_t \mid \theta_{q_i}). \tag{37}$$

To train the neural networks, N_b tuples are randomly sampled from the buffer $\mathcal{B}$ for each step of training. Specifically, the critic networks can be trained by minimizing the mean-squared soft Bellman error loss function which is derived as

$$\theta_{q_i}=\arg\min_{\theta_{q_i}} \frac{1}{N_b}\sum_{n=1}^{N_b}\left(y_n^c-Q_{rc}\left(s_n, a_n \mid \theta_{q_i}\right)\right)^2,\ i=\{1,2\}. \tag{38}$$

In addition, the actor network is trained by minimizing the policy loss

$$\theta_\pi=\arg\min_{\theta_\pi} \frac{1}{N_b}\sum_{n=1}^{N_b}\left[\alpha \ln \pi(s_n \mid \theta_\pi)-\min_{i=1,2} Q_{rc}(s_n, a_n \mid \theta_{q_i})\right]. \tag{39}$$

The temperature coefficient α is updated by minimizing the loss

$$\alpha=\arg\min_{\alpha}-\frac{1}{N_b}\sum_{n=1}^{N_b}\left[\alpha\left(\ln \pi\left(s_n \mid \theta_\pi\right)-\mathcal{H}\right)\right], \tag{40}$$

where $\mathcal{H}$ is equal to the negative value of the action dimension. The target networks are updated using soft update policy, which can be represented as

$$\theta_{\tilde{q}_i}=\tau\theta_{q_i}+(1-\tau)\theta_{\tilde{q}_i},\ i=\{1,2\}, \tag{41}$$

where $\tau \ll 1$ is the soft update coefficient.

For convenience, the C-SAC algorithm is summarized in Algorithm 1.

G. Complexity Analysis

Let N_π and N_Q denote the parameter sizes of the policy and critic networks, respectively, and let B be the batch size. The C-SAC algorithm has the same complexity as SAC, i.e., $O(B(N_\pi+2N_Q))$. The conservative augmented Lagrangian (CAL) [43] method requires additional networks to estimate both cost and reward, leading to a complexity order of $O(B(N_\pi+2N_Q+2KN_Q))$, where K is the number of constraints. The constrained variational policy optimization (CVPO) [44] approach further introduces an E-step involving

Algorithm 1 C-SAC Algorithm

Initialization: Randomly generating parameters of actor network θ_π, and critic networks θ_{q1}, θ_{q2}. Initialize target networks $\theta_{\tilde{q}_1} \leftarrow \theta_{q_1}, \quad \theta_{\tilde{q}_2} \leftarrow \theta_{q_2}$.

for $episode = 1, 2, \cdots, episode_{max}$ **do**
  Initialize the training environment and obtain the state s_0;
  for $t = 0, 1, \cdots$ **do**
    Sample action a_t from the policy distribution $\pi(s_t \mid \theta_\pi)$, UAV execution action a_k, and obtain the next environment state s_{t+1}, reward r_t, cost c_t, and the flag $done$;
    Store $(s_t, a_t, r_t, c_t, s_{t+1}, done)$ in replay buffer $\mathcal{B}$;
    if $done == 1$ **then**
      Break;
    end if
  end for
  for $i = 1, 2, \cdots, N_u$ **do**
    Randomly sample a batch of transitions $(s_t, a_t, r_t, c_t, s_{t+1}, done)$ from $\mathcal{B}$;
    Update critics as (38);
    Update policy network as (39);
    Update temperature coefficient as (40) ;
    Update target networks as (41);
  end for
end for

action sampling and dual-variable optimization, along with an M-step for supervised policy updates, leading to $O(B((1+M)N_\pi + 2(1+E+K)N_Q))$, where E and M denote the numbers of sampled actions in the E-step and M-step, respectively. Therefore, C-SAC achieves a lower computational complexity compared with the advanced CAL and CVPO schemes. Compared with the point-mass model, the rigid-body UAV model linearly increases computational complexity due to more input neurons, but enables trajectory optimization that strictly satisfies UAV dynamics.

IV. Simulation Results

In this section, we assess the effectiveness of the algorithm proposed in this paper through a series of simulation tests. The simulation environment is configured with the following specifications: Windows 10 operating system, Python 3.8 programming language, PyCharm 2023 as the compiler platform, and the open-source machine learning toolkit PyTorch 1.9.0 with CUDA 11.1 support.

A. Simulation Settings

In the simulations, we set the BS as the coordinate origin, with the receiving antenna positioned at a height of 30 meters. The UAV initiates its mission from the starting point $[500, 2600, 100]$ m and is tasked with sensing a moving target, ultimately reaching the endpoint $[0, 2600, 100]$ m. The UAV operates with a maximum energy capacity of 30 kJ. For the TD-ISCC system, the transmission power consumption at the UAV is held constant at $P_t = 1$ W. The UAV's antenna features a gain of $G = 17$ dBi, while the BS antenna has a gain of

TABLE II
KEY UAV PARAMETERS FOR SIMULATION

m	3	c_0	3.6×10^{-3}	C_t	4.848×10^{-5}
g	9.8	c_1	7.5×10^{-4}	C_m	8.891×10^{-7}
L	0.3	c_2	8.5938×10^{-6}	C_{mx}	0.016
C_{dx}	0.11	c_3	8.8949×10^{-7}	C_{my}	0.1
C_{dy}	0.11	c_4	5.1287×10^{-10}	C_{mz}	0.1
C_{dz}	0.2	h_{min}	60	I_x	4.29×10^{-2}
U_{1l}	6	I_m	8.02×10^{-4}	I_y	4.29×10^{-2}
U_{1u}	26	U_{2u}	0.7	I_z	7.703×10^{-2}
U_{3u}	0.7	U_{4u}	0.05	-	-

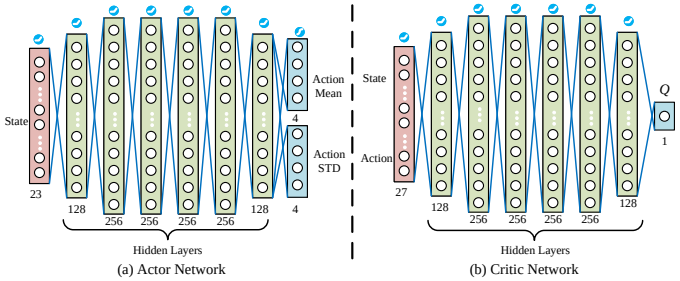


Fig. 5. The AC network architectures.

$G_B = 20$ dBi. The carrier frequency is set at 28 GHz, with a channel bandwidth of 10 MHz [31]. The UAV's camera has a data rate of $R_v = 80$ Mbps and consumes power of $P_o = 9$ W. The reference SNR values are determined as $\lambda_1 = 98.6$ dB and $\lambda_2 = 109.6$ dB. Key parameters related to the UAV are summarized in Table II [21].

The UAV operates within a 3D space defined as $\Omega_{3d} = \{[-300, 600]; [2500, 3100]; [30, 200]\}$ m. The key parameters of the reward and cost functions are set as $\tilde{r}_d = 500$, $\varepsilon_d = 0.3$, $\xi_d = 20$ m, $\varepsilon_s = 1$, $\tilde{r}_h = 1$, $\varepsilon_h = 10$, $\xi_e = 15$ kJ. The DRL algorithm has various hyperparameters, which are detailed in Table III. Fig. 5 illustrates the AC network architecture, including the input/output dimensions, number of neurons, hidden layers, and activation functions. The actor maps the environment state to the UAV action mean and standard deviation (STD) with $\tanh$ at the output layer and ReLU elsewhere, while the critic shares the same hidden-layer architecture and estimates $Q(s_t, a_t)$ from both the state s_t and action a_t. To evaluate performance, rewards between adjacent continuous states are estimated by uniform sampling, with the sampling interval, referred to as the control segment, set to one second.

B. Algorithms Comparison

To verify the effectiveness of the proposed algorithm, this section compares the training results of the C-SAC algorithm

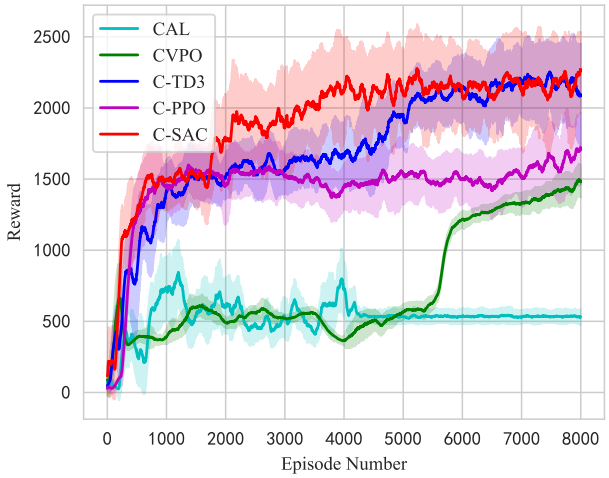
TABLE III
PARAMETER SETTING OF THE DRL ALGORITHM

Parameter	Symbol	Value
Actor learning rate	r_a	10^{-3}
Critic learning rate	r_c	10^{-3}
Discount factor	γ	0.99
Soft update coefficient	τ	0.005
Buffer	$\mathcal{B}$	2^{20}
Batch size	B	256
Update count	N_u	30

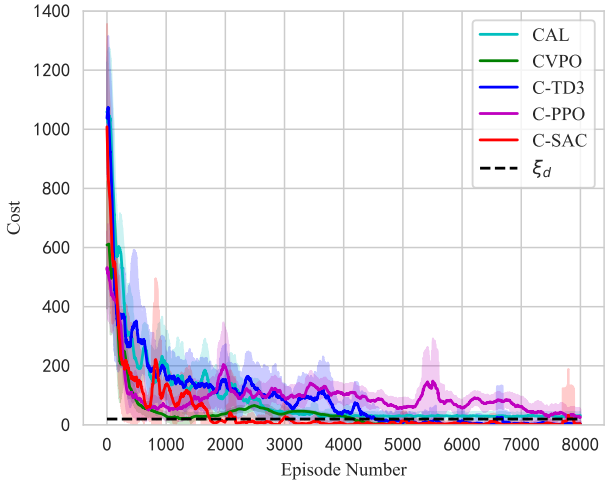
with the CAL [43], CVPO [44], constrained twin delayed deep deterministic policy gradient (C-TD3), and constrained proximal policy optimization (C-PPO) methods, where C-TD3 and C-PPO are standard TD3 [41] and PPO [45] algorithms adopting our recommended hybrid reward-cost (36) as reward. Furthermore, we include a non-RL numerical optimization (NO) [21], [22] baseline based on gradient descent for comparison. Fig. 6 illustrates the trends of the rewards and costs obtained by these algorithms during training with episodes. In the figure, the shaded areas show statistical results from a window of 101 episodes centered on each training round (including the current round, 50 preceding and 50 subsequent episodes). The solid lines represent the average reward, and the bounds of the shaded area are determined by the mean $\pm$ STD within the same window.

Fig. 6 illustrates the training performance of different algorithms, where ξ_d represents the distance cost allowed by constraint C_6. In terms of the cost convergence, both the CAL and CVPO algorithms exhibit satisfactory performance, demonstrating robust convergence characteristics. However, their effectiveness in reward training remains suboptimal, indicating a notable limitation in balancing constraint satisfaction and reward maximization. In contrast, the proposed method achieves a superior trade-off, effectively minimizing costs while simultaneously enhancing reward acquisition. Specifically, the analysis reveals that C-PPO, as an on-policy approach, suffers from significant limitations in both reward and cost optimization due to its inherent low data utilization efficiency. Meanwhile, C-TD3, as a deterministic policy algorithm, exhibits limited exploration capabilities during the training phase, which leads to a slower convergence rate compared to the more exploratory C-SAC approach. In addition to convergence, training stability can also be observed from the learning curves. As shown in Fig. 6, the proposed method exhibits smoother reward and cost trajectories, whereas the proposed method shows larger variance during iterations due to its high level of exploration. This stability is further reflected in the consistency of results across multiple runs, highlighting the robustness of the proposed method compared with baseline algorithms.

To validate the algorithm's efficacy in sensing time-varying motion targets and ensure its robustness, we conduct 1000 Monte Carlo simulations using the well-trained actor network. The simulation results are presented in Table IV. In Table IV, "Cumulative MI" denotes the successful reception of UAV sensing data ($Q_r(T)$) by the BS, while "Distance" represents the distance of the UAV from the destination at the time of energy depletion. The terms "Reward" and "Cost" refer to the cumulative reward and cost values provided as feedback from the environment, respectively. In addition, "Commun. Con." and "Height Con." indicate the probabilities of violating communication constraints (C_3) and altitude constraints (C_4) in the test results, respectively. In the presented table, the data are expressed in the format $a \pm b$, where a and b represent the arithmetic mean and STD of the results, respectively. From the simulation results, it can be observed that C-SAC achieved the maximum amount of sensing data, the shortest distance to the destination, the highest reward value, and did not violate the communication and altitude constraints throughout the process.



(a) Comparison of different algorithms on reward convergence.



(b) Comparison of different algorithms on cost convergence.

Fig. 6. Comparison of different algorithms.

In contrast, CVPO achieved the minimum cost. However, the NO scheme fails to deliver satisfactory performance under complex UAV dynamics and time-varying target trajectories. Overall, the simulation results underscore that the proposed C-SAC algorithm effectively accelerates training convergence and enhances the overall training performance.

C. Comparison of UAV Models: Rigid-body vs. point-mass

While UAVs are commonly abstracted as point-mass models in trajectory optimization studies for communication and sensing performance analysis [12]–[16], [23], a gap persists between this simplified representation and the actual 6-DoF rigid-body dynamics governing real-world UAV operations. This modeling gap significantly undermines the translational validity of neural network-based controllers trained with point-mass representations when implemented on physical UAV platforms. To rigorously assess this performance discrepancy, we conduct systematic Monte Carlo simulations that compare our proposed rigid-body dynamics framework with the conventional point-mass model-based approaches.

TABLE IV
MONTE CARLO SIMULATION RESULTS: PERFORMANCE OF DIFFERENT ALGORITHMS

Algorithm	Cumulative MI (Mbits)	Distance (m)	Reward	Cost	Commun. Con.	Altitude Con.
CAL	226.04 ± 62.31	17.51 ± 9.65	1304.89 ± 239.12	6.84 ± 13.83	0	0
CVPO	198.44 ± 46.70	8.89 ± 3.92	1483.83 ± 132.25	$\mathbf{0.02} \pm 0.67$	0	0
C-PPO	382.57 ± 95.08	30.62 ± 19.45	1674.06 ± 291.46	26.05 ± 23.86	0	0
C-TD3	456.20 ± 88.52	13.11 ± 3.61	2229.89 ± 255.41	0.86 ± 4.27	0	2.92×10^{-5}
C-SAC	$\mathbf{471.46} \pm 95.38$	$\mathbf{7.20} \pm 6.40$	$\mathbf{2298.64} \pm 276.75$	1.61 ± 7.04	0	0
NO	365.27 ± 11.07	31.33 ± 25.51	-	-	0	0

TABLE V
MONTE CARLO SIMULATION RESULTS: RIGID-BODY VS. POINT-MASS

Model		Cumulative MI (Mbits)	Distance (m)	Reward	Cost	Commun. Con.	Altitude Con.
Point-Mass	Expect	692.67 ± 123.89	1.90 ± 0.93	3174.68 ± 293.94	0.01 ± 0.21	0	7.68×10^{-5}
	Real	419.06 ± 31.98	856.67 ± 67.37	778.49 ± 47.96	1847.13 ± 344.51	0	0.32
Rigid-Body		471.46 ± 95.38	7.20 ± 6.40	2298.64 ± 276.75	1.61 ± 7.04	0	0

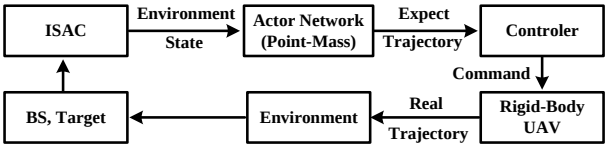


Fig. 7. Flowchart of the actor network implementation based on the point-mass model.

To systematically evaluate the impact of different UAV models on performance, this study employed the C-SAC algorithm to train both the UAV point-mass model and the rigid-body model. As illustrated in Fig. 7, the trajectory generated by the point-mass model was input into the real-world rigid-body model, and the UAV controller executed the corresponding control commands based on this trajectory to obtain the actual operational results of the point-mass model in a real-world environment. To compare the differences between the models in sensing time-varying moving targets, this study conducted 1,000 Monte Carlo simulations using a well-trained actor network. The statistical analysis of the simulation results is summarized in Table V.

As shown in Table V, the simulation results obtained from the point-mass ideal model are significantly superior to those obtained from training with the rigid-body model. This advantage primarily stems from the structural simplicity of the point-mass model, which facilitates easier training and enables more efficient optimization of the objective function, ultimately leading to enhanced simulation outcomes. However, when the control commands generated by the point-mass model are applied to the rigid-body model, the actual performance of the point-mass model is markedly inferior to the solution provided by our proposed method. Specifically, the point-mass model fails to reach the destination before depleting its energy and significantly increases the likelihood of violating altitude constraints. This phenomenon indicates that for underactuated systems such as UAVs, neural networks trained on the simplified point-mass model exhibit considerable discrepancies in practical applications, which can severely compromise the safety and reliability of UAV operations.

Fig. 8 illustrates the relationship between cumulative MI and average target speed when sensing targets with varying motion speeds under two model schemes. An analysis of the regression curves (RCs) indicates that the point-mass model theoretically achieves higher MI than the rigid-body model. However, its actual performance is contrary to this expectation. This discrepancy is primarily attributed to the point-mass model's inability to adequately account for the dynamic characteristics of UAVs, resulting in low energy efficiency during actual flights. Specifically, UAVs consume more energy when attempting to follow trajectories that are inconsistent with their dynamic properties, thereby diminishing the overall QoS in sensing. In addition, it is evident that the average speed of the target, whether excessively high or low, results in a reduction of the cumulative MI. This phenomenon can be attributed to the existence of an optimal flight speed for the UAV. If the target's movement speed significantly exceeds or falls below this optimal speed, the UAV will expend more energy while tracking the target, leading to a decrease in the total perceived MI. In contrast, the actual results of the point-mass model indicate that the cumulative MI for sensing is less influenced by the target's speed of motion. This phenomenon occurs primarily because the point-mass model allocates more time to close-range sensing when it is unable to return to the destination promptly, thereby partially mitigating the effects of speed variations on sensing performance.

In Fig. 9, we observe the average trend of the ISAC rate over 1,000 tests. This figure illustrates that during the initial phase, as the UAV maneuvers to track the target, it enhances the sensing channel, resulting in an increase in the amount of perceived information. However, in the later stage, as the UAV progresses toward its destination, the increasing distance between the UAV and the target diminishes the UAV's ability to perceive information. This decline subsequently leads to a decrease in the ISAC rate. In the point-mass model, the UAV is not constrained by dynamic limitations, allowing them to swiftly approach targets while maintaining relatively optimal sensing channels. Furthermore, the point-mass model overlooks the energy consumption associated with UAV rotation, resulting in lower energy expenditure compared to the rigid-body model and, consequently, longer sensing durations. However, in practical applications, the performance of the point-mass model often deviates significantly from theoretical expectations. The trajectories planned by the point-mass model do not align with the actual dynamic characteristics of UAVs, leading to significantly lower ISAC rates and sensing times compared to

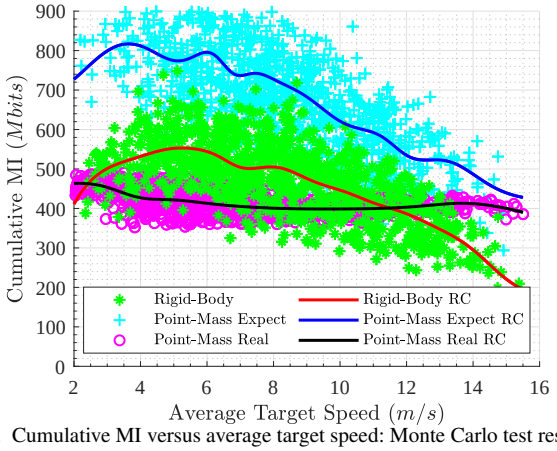


Fig. 8. Cumulative MI versus average target speed: Monte Carlo test results.

those achieved with the rigid-body model.

These results provide robust evidence of the effectiveness of the proposed method compared to the point-mass model in tracking targets with time-varying speeds.

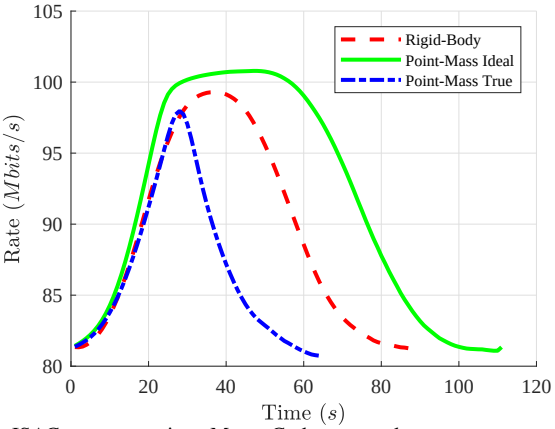


Fig. 9. ISAC rate versus time: Monte Carlo test results.

D. A Trajectory Planning Case

In Fig. 10, the real-time speed of the target is displayed, while Fig. 11 showcases the planned trajectory of the UAV for target sensing. Furthermore, Fig. 12 exhibits the UAV's video rate and real-time ISAC rate, while Fig. 13 illustrates the change trend of the UAV's control law.

In Fig. 11, the lines represent different trajectory categories, and the changing color of the lines corresponds to the change of time, as indicated by the color bar on the right side of the trajectory chart. By analyzing Figs. 11 and 12, several key observations can be made:

1) Initially, the UAV moves toward the target, which gradually enhances the sensing channel and improves the sensing rate, leading to a steady increase in the ISAC rate. Meanwhile, to ensure reliable communication transmission, the proportion of time slots allocated to sensing is reduced.
2) When the UAVs energy consumption reaches a certain threshold, its trajectory shifts toward the flight destination. This movement causes the UAV to move away from the target, weakening the sensing channel and reducing the sensing rate, which in turn leads to a decline in the ISAC rate. However, during this process, the proportion of sensing time slots increases, allowing for the acquisition of more sensing MI.
3) Eventually, the UAV successfully reaches its destination before depleting all its energy.

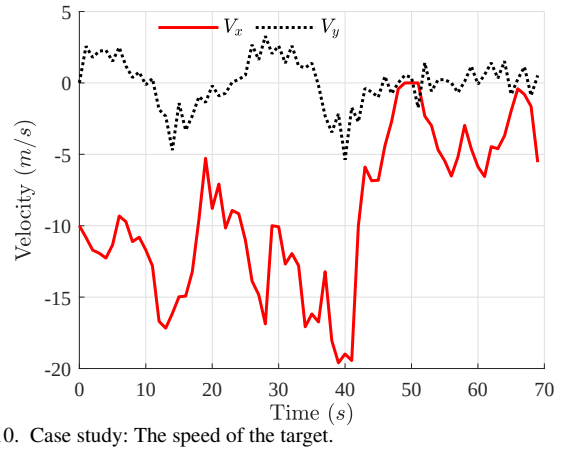


Fig. 10. Case study: The speed of the target.

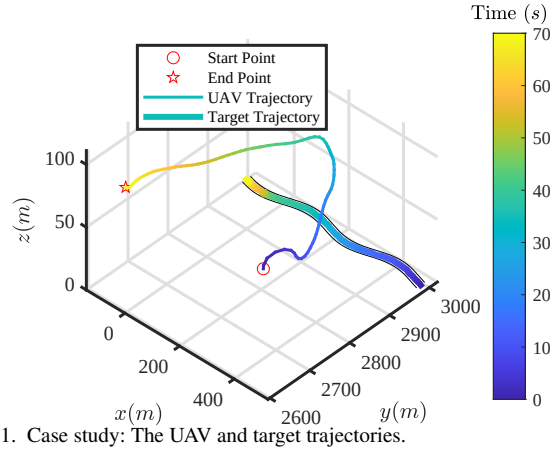


Fig. 11. Case study: The UAV and target trajectories.

In addition, in Fig. 13, the dashed lines in each subfigure represent the boundaries of the control variables. The control parameters of the UAV exhibit continuous changes without any violent oscillations. This observation indicates that the proposed control law schemes have the potential for real-world applications in UAVs, offering stable and efficient control during flight.

To verify the feasibility of deploying the proposed algorithm on embedded platforms, we tested the inference resource consumption of the actor network on an Intel Core i7-10700K (3.8 GHz) CPU. The actor model size is 1.02 MB, and the peak CPU memory usage during inference is 1.98 MB. In 100 random inference tests, the maximum delay is 1 ms, and the average delay is 0.45 ms, indicating that the proposed method satisfies the real-time requirements of UAV onboard systems.

V. Conclusion

In this paper, the C-SAC algorithm has been introduced to tackle the ISCC problem in the UAV TD-ISCC network, focusing on the task of monitoring a mobile target with varying velocity over time. We have proposed a TD-ISCC frame structure that adjusts its time slot allocation dynamically. This adaptive allocation approach enables the system to efficiently allocate communication and sensing slots, leading to enhanced system

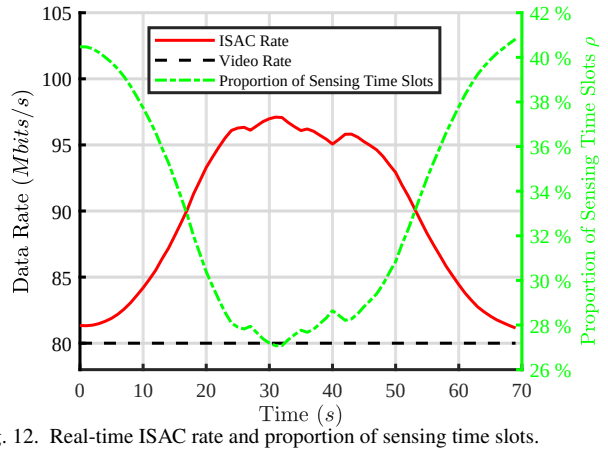


Fig. 12. Real-time ISAC rate and proportion of sensing time slots.

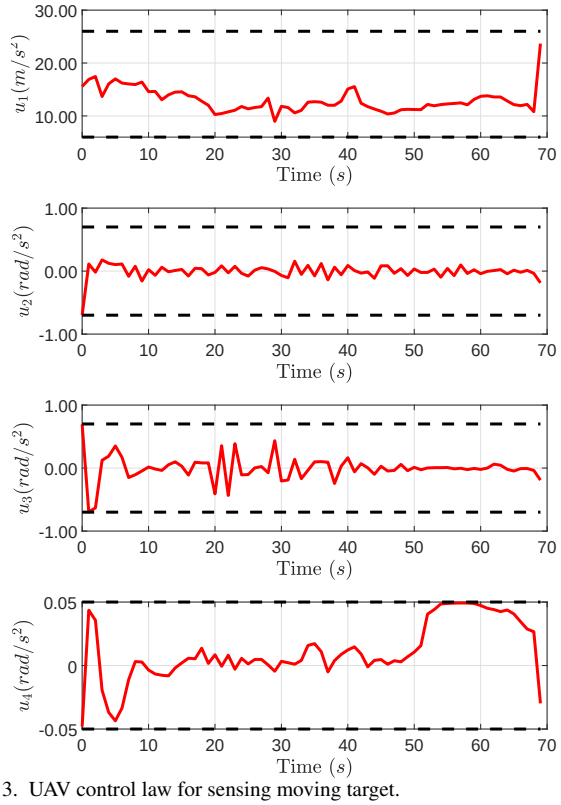


Fig. 13. UAV control law for sensing moving target.

capacity. By customizing the reward and cost functions for each constraint, we transformed the multi-constraint ISCC problem into a CMDP, making it suitable for resolution using CRL. A novel C-SAC algorithm has been developed that adaptively switches between reward maximization and constraint-driven corrective mechanisms. This dual-mode optimization preserves the computational efficiency of unconstrained policy gradient methods while ensuring strict adherence to constraints in dynamic environments. Extensive simulations validate that the proposed C-SAC algorithm outperforms dual-variable-based methods in constraint compliance efficiency and demonstrates superior robustness through Monte Carlo tests. The derived ISCC policy achieves a higher cumulative MI than point-mass UAV models while maintaining adaptability to varying target speeds.

Appendix A
Proof of Theorem 1

Let the objective function be denoted as $J(t) = \int_0^t R_r\left(\tau \mid \boldsymbol{p}_T(\tau)\right) d\tau$. Then, the partial derivative of $J(t)$ with respect to ρ is given by:

$$\frac{\partial J(t)}{\partial \rho} = \int_0^t B \log_2\left(1 + \frac{\lambda_1}{\|\boldsymbol{p}(\tau) - \boldsymbol{p}_T(\tau)\|^4}\right) d\tau > 0. \quad (42)$$

Hence, it follows that the objective function $J(t)$ monotonically increases with $\rho(t)$. Considering the feasible domain of $\rho(t)$ as defined in (10), we can express the optimal $\rho(t)$ as:

$$\rho^*(t) = \frac{\log_2\left(1+\Gamma_c(t)\right) - R_v/B}{\log_2\left(1+\Gamma_c(t)\right) + \log_2\left(1+\Gamma_r\left(t \mid \boldsymbol{p}_T(t)\right)\right)}, \quad (43)$$

subject to the constraint $R_c(t) \geq R_v$.

Appendix B
Relationship to the Primal-Dual Method

According to (35), the Q-function can be expressed as

$$\begin{aligned} Q_{rc}(s_0, a_0 \mid \theta_{q_i}) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t^c\right] \\ &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] - \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c_t\right] \\ &= Q_r(s_0, a_0 \mid \theta_{q_i}) - Q_c(s_0, a_0 \mid \theta_{q_i}), \end{aligned} \quad (44)$$

where $Q_r(s_0, a_0 \mid \theta_{q_i})$ and $Q_c(s_0, a_0 \mid \theta_{q_i})$ denote the reward and cost Q-functions under the action policy π, respectively. This formulation reveals that when $Q_c(\cdot) = 0$, the joint Q-function $Q_{rc}(\cdot)$ reduces to the reward Q-function $Q_r(\cdot)$; conversely, when $Q_r(\cdot) = 0$, we have $Q_{rc}(\cdot) = -Q_c(\cdot)$. Therefore, the sign of $Q_{rc}(\cdot)$ implicitly indicates whether the function reflects accumulated rewards or accumulated costs under the given policy π. In general, $Q_{rc}(\cdot)$ serves as a unified Q-function that balances both reward and cost, analogous to a primal-dual formulation. Specifically, in the standard primal-dual framework used to solve CMDP problems, the objective is typically formulated as

$$\min_{\lambda_L \geq 0} \max_{\pi} Q_r(s_0, a_0 \mid \theta_{q_i}) - \lambda_L Q_c(s_0, a_0 \mid \theta_{q_i}), \quad (45)$$

where λ_L is the Lagrange multiplier associated with the cost constraint. In contrast, our proposed method directly optimizes the joint Q-function

$$\max_{\pi} Q_{rc}(s_0, a_0 \mid \theta_{q_i}), \quad (46)$$

which can be viewed as a special case of the primal-dual approach with $\lambda_L = 1$. Furthermore, when the cost Q-function $Q_c(\cdot)$ converges to 0, our method becomes equivalent to the primal-dual formulation with $\lambda_L = 0$. Consequently, the proposed method not only maintains the optimality characteristics of the primal-dual approach but also provides a more concise and interpretable formulation by unifying rewards and costs within a single Q-function.

REFERENCES

[1] 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. Surveys Tuts.*, vol. 24, no. 1, pp. 306–345, 1st Quart. 2022.

[2] 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.

[3] Z. Fei, X. Wang, G. N. Wu, J. Huang, and J. A. Zhang, "Air-ground integrated sensing and communications: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 61, no. 5, pp. 55–61, May 2023.

[4] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," *IEEE Pervasive Comput.*, vol. 16, no. 1, pp. 24–32, Jan. 2017.

[5] K. Kanistras, G. Martins, M. J. Rutherford, and K. P. Valavanis, "A survey of unmanned aerial vehicles (UAVs) for traffic monitoring," in *Proc. IEEE ICUAS*, Atlanta, GA, May 2013, pp. 221–234.

[6] J. Zou, C. Wang, Y. Liu, Z. Zou, and S. Sun, "Vision-assisted 3-D predictive beamforming for green UAV-to-vehicle communications," *IEEE Trans. Green Commun. Network.*, vol. 7, no. 1, pp. 434–443, Mar. 2023.

[7] A. Khalili, A. Rezaei, D. Xu, and R. Schober, "Energy-aware resource allocation and trajectory design for UAV-enabled ISAC," in *Proc. IEEE GLOBECOM*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 4193–4198.

[8] X. Jing, F. Liu, C. Masouros, and Y. Zeng, "ISAC from the sky: UAV trajectory design for joint communication and target localization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 12 857–12 872, Oct. 2024.

[9] K. Meng, Q. Wu, S. Ma, W. Chen, K. Wang, and J. Li, "Throughput maximization for UAV-enabled integrated periodic sensing and communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 671–687, Aug. 2023.

[10] 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.

[11] H. Salem, H. Sadia, M. M. Quamar, A. Magad, M. Elrashidy, N. Saeed, and M. Masood, "Data-driven integrated sensing and communication: Recent advances, challenges, and future prospects," *ICT Express*, vol. 11, no. 4, pp. 790–808, Aug. 2025.

[12] C. Wang, D. Deng, L. Xu, and W. Wang, "Resource scheduling based on deep reinforcement learning in UAV assisted emergency communication networks," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3834–3848, Jun. 2022.

[13] C. Wang, Z. Wei, W. Jiang, H. Jiang, and Z. Feng, "Cooperative sensing enhanced UAV path-following and obstacle avoidance with variable formation," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 7501–7516, Jun. 2024.

[14] H. He, W. Yuan, S. Chen, X. Jiang, F. Yang, and J. Yang, "Deep reinforcement learning based distributed 3D UAV trajectory design," *IEEE Trans. Commun.*, vol. 72, no. 6, pp. 3736–3751, Jun. 2024.

[15] Y. Qin, Z. Zhang, X. Li, W. Huangfu, and H. Zhang, "Deep reinforcement learning based resource aladdress and trajectory planning in integrated sensing and communications UAV network," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8158–8169, Nov. 2023.

[16] Q. Gao, R. Zhong, H. Shin, and Y. Liu, "MARL based UAVs trajectory and beamforming optimization for ISAC system," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 40 492–40 505, Dec. 2024.

[17] Y. Liu, H. Wang, J. Fan, J. Wu, and T. Wu, "Control-oriented UAV highly feasible trajectory planning: A deep learning method," *Aerospace Science and Technology*, vol. 110, p. 106435, Dec. 2021.

[18] Y. Wang, H. Wang, Y. Liu, J. Wu, and Y. Lun, "6-DOF UAV path planning and tracking control for obstacle avoidance: A deep learning-based integrated approach," *Aerospace Science and Technology*, vol. 151, p. 109320, Aug. 2024.

[19] H. Sandhu, P. P. Pradhan, K. Rajawat, and M. Kothari, "Minimum time trajectory optimization for a 6-DoF quadrotor UAV using successive convexification," in *Proc. AIAA SCITECH Forum*, Orlando, FL, Jan. 2024.

[20] Z. Shen, G. Zhou, H. Huang, C. Huang, Y. Wang, and F.-Y. Wang, "Convex optimization-based trajectory planning for quadrotors landing on aerial vehicle carriers," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 138–150, Jan. 2024.

[21] B. Li, Q. Li, Y. Zeng, Y. Rong, and R. Zhang, "3D trajectory optimization for energy-efficient UAV communication: A control design perspective," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4579–4593, Jun. 2022.

[22] Q. Li, B. Li, Z. He, Y. Rong, and Z. Han, "Joint design of communication sensing and control with a UAV platform," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 19 231–19 244, Dec. 2024.

[23] 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.

[24] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 11 216–11 235, Dec. 2024.

[25] D. Yu, H. Ma, S. Li, and J. Chen, "Reachability constrained reinforcement learning," in *Proc. ICML*, MD, Jul. 2022, pp. 25 636–25 655.

[26] A. Termehchi, A. Syed, W. S. Kennedy, and M. Erol-Kantarci, "Distributed safe multi-agent reinforcement learning: Joint design of THz-enabled UAV trajectory and channel allocation," *IEEE Trans. Veh. Technol.*, vol. 73, no. 10, pp. 14 172–14 186, Oct. 2024.

[27] H. Yu and H.-C. Yang, "Causal CSI-based trajectory design and power allocation for UAV-enabled wireless networks under average rate constraints: A constrained reinforcement learning approach," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 61, no. 1, pp. 554–567, Feb. 2025.

[28] M. A. Richards, J. Scheer, W. A. Holm, and W. L. Melvin, *Principles of modern radar*. Citeseer, 2010.

[29] M. Thomas and A. T. Joy, *Elements of information theory*. Wiley-Interscience, 2006.

[30] Q. Zhang, X. Wang, Z. Li, and Z. Wei, "Design and performance evaluation of joint sensing and communication integrated system for 5G mmWave enabled CAVs," *IEEE J. Sel. Top. Sign. Proces.*, vol. 15, no. 6, pp. 1500–1514, Sept. 2021.

[31] Q. Zhang, H. Sun, X. Gao, X. Wang, and Z. Feng, "Time-division ISAC enabled connected automated vehicles cooperation algorithm design and performance evaluation," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2206–2218, Mar. 2022.

[32] E. S. Kang, H. Hwang, and D. S. Han, "A fine carrier recovery algorithm robust to Doppler shift for OFDM systems," *IEEE Trans. Consum. Electr.*, vol. 56, no. 3, pp. 1218–1222, Aug. 2010.

[33] Q. Wu, L. Liu, and R. Zhang, "Fundamental trade-offs in communication and trajectory design for UAV-enabled wireless network," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 36–44, Feb. 2019.

[34] M. Xing, X. Jiang, R. Wu, F. Zhou, and Z. Bao, "Motion compensation for UAV SAR based on raw radar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2870–2883, Apr. 2009.

[35] M. Pieraccini, L. Miccinesi, and N. Rojhani, "A Doppler range compensation for step-frequency continuous-wave radar for detecting small UAV," *Sensors*, vol. 19, no. 6, p. 1331, Mar. 2019.

[36] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Aug. 2013.

[37] D. Wen, P. Liu, G. Zhu, Y. Shi, J. Xu, Y. C. Eldar, and S. Cui, "Task-oriented sensing, computation, and communication integration for multi-device edge AI," in *Proc. IEEE ICC*, Rome, Italy, May 2023, pp. 3608–3613.

[38] R. Mahony, V. Kumar, and P. Corke, "Multirotor aerial vehicles: Modeling, estimation, and control of quadrotor," *IEEE Robot. Autom. Mag.*, vol. 19, no. 3, pp. 20–32, Aug. 2012.

[39] J.-Y. Wen and K. Kreutz-Delgado, "The attitude control problem," *IEEE Trans. Autom. Control*, vol. 36, no. 10, pp. 1148–1162, Oct. 1991.

[40] H. Liu, X. Wang, and Y. Zhong, "Quaternion-based robust attitude control for uncertain robotic quadrotors," *IEEE Trans. Ind. Informat.*, vol. 11, no. 2, pp. 406–415, Feb. 2015.

[41] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. ICML*, vol. 80, Stockholm, Sweden, Jul. 2018, pp. 1587–1596.

[42] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," *ArXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1812.05905

[43] Z. Wu, B. Tang, Q. Lin, C. Yu, S. Mao, Q. Xie, X. Wang, and D. Wang, "Off-policy primal-dual safe reinforcement learning," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2401.14758

[44] Z. Liu, Z. Cen, V. Isenbaev, W. Liu, S. Wu, B. Li, and D. Zhao, "Constrained variational policy optimization for safe reinforcement learning," in *Proc. ICML*, MD, Jul. 2022, pp. 13 644–13 668.

[45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *ArXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

Qingliang Li (Graduate Student Member, IEEE) received the B.E. degree in flight vehicle control and information engineering and M.E. degree in electronic information from Sichuan University, China, in 2020 and 2023, respectively. He is currently pursuing his Ph.D. degree with the National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China (UESTC), China. His research interests include sensing, communication and control co-design, movable antenna, and UAV communication.

Bin Li (M'18-SM'18) received the Bachelor degree in automation and the Master degree in control science and engineering from Harbin Institute of Technology, China, in 2005 and 2008, respectively, and Ph.D. degrees in mathematics and statistics from Curtin University, Australia, in 2011. From 2012-2014, he was a Research Associate with the School of Electrical, Electronic and Computer Engineering, the University of Western Australia, Australia. From 2014-2017, he was a Research Fellow with the Department of Mathematics and Statistics, Curtin University, Australia. Currently, he is a Professor with the School of Aeronautics and Astronautics, Sichuan University, China. His research interests include stochastic model predictive control, optimal control, optimization, signal processing, and wireless communications.

YUE RONG (Senior Member, IEEE) received the Ph.D. degree (summa cum laude) in electrical engineering from Darmstadt University of Technology, Darmstadt, Germany, in 2005.

He was a Postdoctoral Researcher with the Department of Electrical Engineering, University of California at Riverside, Riverside, CA, USA, from February 2006 to November 2007. Since December 2007, he has been with Curtin University, Bentley, WA, Australia, where he is currently a Professor. His research interests include signal processing for communications, underwater acoustic communications, underwater optical wireless communications, machine learning, speech recognition, and biomedical engineering. He has published over 230 journal and conference papers in these areas.

Prof. Rong was a Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2020 to 2024. He was an Editor of the IEEE WIRELESS COMMUNICATIONS LETTERS from 2012 to 2014 and a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Theories and Methods for Advanced Wireless Relays. He was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2014 to 2018.

Zhen-Qing He (Member, IEEE) received the Ph.D. degree in communication and information system from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2017. From 2015 to 2016, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA. He was a Post-Doctoral Researcher from 2018 to 2020 and was an Associate Research Fellow from 2021 to 2022 with the National Key Laboratory of Science and Technology on Communications, UESTC, respectively. He is currently an associate professor with the School of Aeronautics and Astronautics, Sichuan University, Chengdu, China. His main research interests include broad range of signal processing, wireless communications, and machine learning. He was a recipient of the IEEE Communications Society Heinrich Hertz Prize Paper Award in 2022 and was a co-recipient of the Best Paper Award of the International Conference on Wireless Communications and Signal Processing (WCSP) in 2022.

Zhu Han (S01M04-SM09-F14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. Dr. Hans main research targets on the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing, data science, smart grid, carbon neutralization, security and privacy. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011,ăthe EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, IEEE Vehicular Technology Society 2022 Best Land Transportation Paper Award, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018 and ACM Distinguished Speaker from 2022 to 2025, AAAS fellow since 2019, and ACM Fellow since 2024. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award (an IEEE Field Award), for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."