基于边缘计算的无人机通感融合网络波束成形与资源优化

李斌[1]，彭思聪[1]，费泽松[2]
（1. 南京信息工程大学计算机学院，江苏 南京 210044； 2. 北京理工大学信息与电子学院，北京 100081）

摘 要：为了解决传统通信−感知融合网络模式对地面基础设施的依赖，针对复杂场景下通感融合网络系统功耗较大、信号阻塞、覆盖盲区等问题，提出了一种无人机搭载边缘计算服务器与雷达收发器辅助通感融合网络。首先，在满足用户传输功率、雷达估计信息率、任务卸载比例限制的条件下，通过联合优化无人机雷达波束成形、计算资源分配问题、任务卸载量划分、终端用户发射功率和无人机飞行轨迹，建立系统总能耗最小化问题；其次，将该非凸优化问题重新构建为一个马尔可夫决策过程，使用深度强化学习中的近端策略优化算法实现系统的优化决策。仿真结果表明，所提算法训练速度较快，能够在保证应用的感知与计算时延需求的同时有效降低系统能耗。

关键词：感知−通信−计算融合网络；无人机；深度强化学习；资源分配与优化
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023172

Beamforming and resource optimization in UAV integrated sensing and communication network with edge computing

LI Bin[1], PENG Sicong[1], FEI Zesong[2]
1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Abstract: To address the dependence of traditional integrated sensing and communication network mode on ground infrastructure, the unmanned aerial vehicle (UAV) with edge computing server and radar transceiver was proposed to solve the problems of high-power consumption, signal blocking, and coverage blind spots in complex scenarios. Firstly, under the conditions of satisfying the user's transmission power, radar estimation information rate and task offloading proportion limit, the system energy consumption was minimized by jointly optimizing UAV radar beamforming, computing resource allocation, task offloading, user transmission power, and UAV flight trajectory. Secondly, the non-convex optimization problem was reformulated as a Markov decision process, and the proximal policy optimization method based deep reinforcement learning was used to achieve the optimal solution. Simulation results show that the proposed algorithm has a faster training speed and can reduce the system energy consumption effectively while satisfying the sensing and computing delay requirements.

Keywords: integrated sensing-communication-computation network, UAV, deep reinforcement learning, resource allocation and optimization

收稿日期：2023−04−25；**修回日期：**2023−07−17
通信作者：费泽松，feizesong@139.com
基金项目：国家重点研发计划基金资助项目（No.2021YFB2900200）；国家自然科学基金资助项目（No.62101277）；江苏省自然科学基金资助项目（No.BK20200822）
Foundation Items: The National Key Research and Development Program of China (No.2021YFB2900200), The National Natural Science Foundation of China (No.62101277), The Natural Science Foundation of Jiangsu Province (No.BK20200822)

0　引言

下一代无线通信的网络节点被设想为超越单一通信维度，以一种综合的方式执行多种功能，如高精度、多目标环境感知和低时延计算[1]。由于无线感知在硬件设施和信号处理方面与无线通信技术有着惊人的相似之处，将无线通信与无线感知相结合，可为网络提供并发的通信和感知功能。为此，通信−感知一体化（ISAC, integrated sensing and communication）的研究备受关注，通过在感知和通信系统之间共享频谱资源和无线基础设施，可以实现资源的高效利用，同时保证感知与通信功能之间的互通、互惠、互利[2]。

随着无线设备类型逐渐异质化、应用形态日趋丰富化、网络数据越来越巨量化，将所有数据卸载到云端势必导致严重的网络拥堵和过高的服务时延。移动边缘计算（MEC, mobile edge computing）作为一种新兴范式，将云计算的功能扩展到网络边缘，实现业务的就近服务，已成为缓解核心网络拥堵、提高用户服务质量的备选方案[3]。在上述背景下，将 ISAC 网络架构与 MEC 架构有机结合，在网络节点实现感知与通信功能的同时，系统设备也实现数据边缘处理的过程，未来网络节点朝着感知−通信−计算融合（ISCC, integrated sensing, communication and computation）网络架构的方向发展[4-6]。

目前，有关 ISCC 的工作主要集中在地面网络。然而，地面网络存在许多固有的局限性，如地面周围障碍物和散射物造成的信号阻塞，可用基础设施有限导致信号覆盖不完整，从而导致服务性能严重下降，甚至无法使用[7]。为有效提高感知、通信和计算的服务质量，凭借无人机（UAV, unmanned aerial vehicle）的高机动性与灵活性，将其部署为空中移动基站和雷达感知器已被视为一种克服地面 ISCC 系统局限性的候补方案[8]。

1) 关于 UAV 辅助 MEC 的研究。为了应对复杂环境下固定基站存在的局限性，文献[9]提出了一种 UAV 中继辅助 MEC 网络方案，通过联合优化 UAV 波束成形、计算资源频率、飞行轨迹、发射功率和用户计算资源分配，以最小化系统能耗。文献[10]考虑了多 UAV 环境下用户安全通信与安全计算性能，以最大限度地提高系统的平均计算能力。文献[11]将 UAV 辅助 MEC 的问题分解为区域划分优化问题和轨迹优化问题，从而减少 UAV 传输能耗和悬停能耗之和。文献[12]考虑了多 UAV 场景下数据卸载策略以及任务时隙划分问题，其目的是最小化每个时隙的系统能耗。文献[13]在优化 UAV 飞行轨迹的同时，联合优化了 UAV 与用户之间上行链路和下行链路的通信资源，最大限度地为用户提供卸载机会。

2) 关于 ISCC 的研究。关于 MEC 辅助 ISAC 的研究着重于资源调度和波束成形，通过有效的资源调度和波束成形设计，可以优化无线资源、保障通信链路质量，并提高计算任务的处理效率。该方法有助于加快感知数据采集速度，增强通信链路的可靠性，以及降低计算任务的时延与能耗，进而改善网络的整体性能。文献[14]提出了一种智能反射面辅助 ISCC 网络的节能设计方案，采用迭代算法对计算资源和通信资源进行联合优化。综合考虑总体性能最大化和发射功率最小化多目标问题，文献[15]分别提出了 2 种联合波束成形算法。在此基础上，文献[16]提出一个多目标优化问题，联合优化通信资源和计算资源分配的同时，设计多终端下雷达波束成形，实现计算能耗最小化。文献[17]以系统最大吞吐量为目标，在满足异构资源需求的同时解决 ISCC 无线资源调度问题。尽管上述工作对 MEC 辅助 ISAC 网络进行了深刻的研究，但关于 UAV 辅助 ISCC 网络的资源调度和智能管理方面的研究鲜有关注。

本文研究 UAV 辅助 ISAC 网络波束成形和资源优化问题，UAV 作为空中平台，在执行感知与通信的同时，对感知任务进行进一步分析和计算处理。本文工作目标是在保证 UAV 感知、通信和计算服务质量的同时，通过联合优化 UAV 飞行轨迹、波束成形以及计算效率来最小化系统能耗。本文主要工作如下。

1) 将 UAV 引入 ISAC 网络中，其中，UAV 与多个地面用户通信，并在执行通信任务期间进行感知服务。此外，UAV 边缘服务器对感知任务进行计算处理。在该网络中，联合优化波束成形、用户与 UAV 计算资源分配、飞行轨迹、任务卸载量，建立系统能耗最小化优化问题。

2) 提出一种基于近端策略优化（PPO, proximal policy optimization）算法的深度强化学习（DRL, deep reinforcement learning）方法，在满足雷达信息估计率、计算卸载时延以及资源分配约束条件下，通过 DRL 训练框架，求得该优化问题的解。通过

实验仿真结果验证本文算法在动态环境下所实现的性能。

1 系统模型及问题描述

UAV 辅助 ISCC 网络模型如图 1 所示。该网络由一架有 M 根天线的 UAV 和 K 个单天线地面用户组成。其中，UAV 配有计算和存储资源以及雷达感知装置，在实现通信服务和目标感知的同时，为实时处理感知任务提供计算服务[18]。

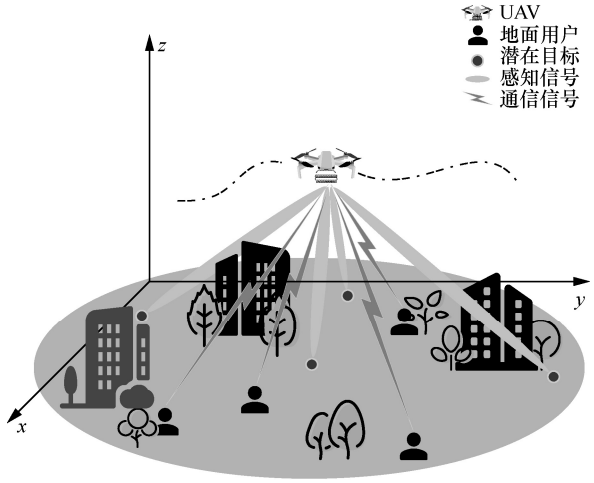

图 1 UAV 辅助 ISCC 网络模型

假设任务周期为 T，将其离散成 N 个足够短的时隙，每个时隙的持续时间为 $\delta_n=\dfrac{T}{N}$，使 UAV 与用户的相对位置在单位时隙内保持近似不变，而在相邻时隙内则有所不同。为了便于表述和分析，定义用户与时隙的集合分别为 $\forall k\in\mathcal{K}\triangleq\{1,\cdots,K\}$ 和 $\forall n\in\mathcal{N}\triangleq\{1,\cdots,N\}$。本文采用三维笛卡儿坐标系，其中，用户 k 的位置固定在 $(x_k,y_k,0)$，$\boldsymbol{q}_k=(x_k,y_k)$ 表示用户 k 的水平坐标，$(x[n],y[n],H)$ 表示 UAV 在第 n 个时隙的坐标，$\boldsymbol{q}[n]=(x[n],y[n])$ 表示 UAV 的水平坐标，H 表示 UAV 的固定飞行高度。相邻时隙内 UAV 的位移变化与飞行速度 $\boldsymbol{v}[n]$ 和加速度 $\boldsymbol{a}[n]$ 有关，因此 UAV 的位移变化应满足

$$\boldsymbol{q}[n+1]=\boldsymbol{q}[n]+\boldsymbol{v}[n]\delta_n+\frac{1}{2}\boldsymbol{a}[n]\delta_n^2 \tag{1}$$

$$\left\|\boldsymbol{q}[n+1]-\boldsymbol{q}[n]\right\|^2\leqslant v_{\max}\delta_n \tag{2}$$

其中，$v_{\max}$ 是 UAV 最大飞行速度。

1.1 通信模型

考虑到现实环境中发射信号会受到建筑、树木等障碍物的影响，用户 k 和 UAV 之间的信道模型遵循瑞利衰减[19]，可表示为

$$\boldsymbol{h}_k[n]=\sqrt{\beta_0 d_k^{-\alpha}[n]}\left(\sqrt{\frac{\kappa}{\kappa+1}}\bar{\boldsymbol{h}}_k[n]+\sqrt{\frac{1}{\kappa+1}}\tilde{\boldsymbol{h}}_k[n]\right) \tag{3}$$

其中，$d_k[n]=\sqrt{\left\|\boldsymbol{q}[n]-\boldsymbol{q}_k\right\|^2+H^2}$ 表示在时隙 n 用户 k 与 UAV 的距离，$\alpha\geqslant 2$ 表示路径损耗指数，β_0 表示参考距离 $d_0=1$ m 处的信道功率增益，κ 表示莱斯因子，$\bar{\boldsymbol{h}}_k[n]$ 表示视线线路（LoS, line-of-sight）信道分量，$\tilde{\boldsymbol{h}}_k[n]$ 表示非视距信道分量且服从均值为零、协方差为单位矩阵的复高斯分布，即 $\tilde{\boldsymbol{h}}_k[n]\sim\mathcal{CN}\left(0,\boldsymbol{I}_M\right)$。考虑 UAV 天线采用均匀线性阵列，则用户 k 到 UAV 的 LoS 分量为

$$\bar{\boldsymbol{h}}_k[n]=\boldsymbol{a}(\boldsymbol{q}[n],\boldsymbol{q}_k)=\left[1,\mathrm{e}^{\mathrm{j}2\pi\frac{d}{\lambda}\cos\theta(\boldsymbol{q}[n],\boldsymbol{q}_k)},\cdots,\mathrm{e}^{\mathrm{j}2\pi\frac{d}{\lambda}(M-1)\cos\theta(\boldsymbol{q}[n],\boldsymbol{q}_k)}\right]^{\mathrm{H}} \tag{4}$$

其中，$\boldsymbol{a}(\boldsymbol{q}[n],\boldsymbol{q}_k)$ 表示指向用户的转向矢量，λ 和 d 分别表示载波波长和相邻两根天线之间的距离，$\theta(\boldsymbol{q}[n],\boldsymbol{q}_k)$ 表示用户 k 和 UAV 之间的角度。

在时隙 n 中，UAV 接收信号 $\boldsymbol{x}[n]$ 包括用户传输信号 $\boldsymbol{x}_{\mathrm{off}}[n]$ 和雷达感知信号 $\boldsymbol{x}_{\mathrm{rad}}[n]$，即

$$\boldsymbol{x}[n]=\boldsymbol{x}_{\mathrm{off}}[n]+\boldsymbol{x}_{\mathrm{rad}}[n]+\boldsymbol{n}[n] \tag{5}$$

其中，$\boldsymbol{n}[n]\in\mathbb{C}^{M\times 1}$ 为独立同分布的高斯随机噪声，其均值为零，方差为 σ^2。

为了处理雷达感知信号 $\boldsymbol{x}_{\mathrm{rad}}[n]$，UAV 根据目标的先验知识生成一个对目标预测的雷达发射信号 $\boldsymbol{s}_{\mathrm{rad}}[n]$，UAV 信号接收机从接收的信号中减去 $\boldsymbol{s}_{\mathrm{rad}}[n]$，以减轻雷达信号引起的不必要的干扰，即目标被抑制的雷达返回信号 $\tilde{\boldsymbol{s}}_r[n]$ [20]。对雷达发射信号进行抑制后，UAV 接收到的雷达感知信号 $\boldsymbol{x}_{\mathrm{rad}}[n]$ 可表示为

$$\boldsymbol{x}_{\mathrm{rad}}[n]=\boldsymbol{H}_r[n]\boldsymbol{w}_r[n]\tilde{\boldsymbol{s}}_r[n] \tag{6}$$

其中，$\boldsymbol{w}_r[n]\in\mathbb{C}^{M\times 1}$ 表示雷达感知信号的波束成形矢量，$\boldsymbol{H}_r[n]\in\mathbb{C}^{M\times M}$ 表示雷达的目标响应矩阵。

对于用户传输信号 $\boldsymbol{x}_{\mathrm{off}}[n]$，在单位时隙内，用户 k 向 UAV 发送的传输信号 $\boldsymbol{x}_k[n]$ 可表示为

$$\boldsymbol{x}_k[n]=\sqrt{p_k[n]}\boldsymbol{h}_k[n]s_k[n] \tag{7}$$

其中，$s_k[n]$ 表示时隙 n 中用户 k 的传输信号，$p_k[n]$ 为用户 k 在时隙 n 的传输功率。因此，UAV 在时隙 n 接收到的用户传输信号 $\boldsymbol{x}_{\mathrm{off}}[n]$ 可表示为

$$\boldsymbol{x}_{\text{off}}[n]=\sum_{k=1}^{K}\boldsymbol{x}_k[n]=\sum_{k=1}^{K}\sqrt{p_k[n]}\boldsymbol{h}_k[n]s_k[n] \tag{8}$$

UAV 接收到用户传输信号后，采用波束成形矢量 $\boldsymbol{w}_k[n]$ 从接收信号中恢复用户 k 的信号。则恢复的用户 k 的信号 $\hat{\boldsymbol{x}}_k[n]$ 为

$$\begin{aligned}\hat{\boldsymbol{x}}_k[n]=&\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{x}[n]=\\&\sqrt{p_k[n]}\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{h}_k[n]s_k[n]+\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{H}_r[n]\boldsymbol{w}_r[n]\tilde{\boldsymbol{s}}_r[n]+\\&\sum_{j=1,j\neq k}^{K}\sqrt{p_j[n]}\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{h}_j[n]s_j[n]+\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{n}[n]\end{aligned} \tag{9}$$

根据文献[17]，$\tilde{\boldsymbol{s}}_r[n]$ 的方差为 $\eta^2B^2\sigma_r^2$，η 是雷达波形功率谱密度常数，σ_r^2 表示预测雷达回波的方差，B 表示 UAV 信道带宽。因此用户 k 在第 n 个时隙传输信号的信噪比为

$$\gamma_k[n]=\frac{g_k[n]}{n_k[n]} \tag{10}$$

其中，$g_k[n]$ 和 $n_k[n]$ 分别表示在第 n 个时隙用户 k 的信号功率和噪声功率，其可分别表示为

$$g_k[n]=p_k[n]\left|\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{h}_k[n]\right|^2 \tag{11}$$

$$\begin{aligned}n_k[n]=&\sum_{j=1,j\neq k}^{K}p_j[n]\left|\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{h}_j[n]\right|^2+\sigma^2\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{w}_k[n]+\\&\eta^2B^2\sigma_r^2\left|\boldsymbol{w}_k^{\mathrm{H}}[n]\boldsymbol{H}_r[n]\boldsymbol{w}_r[n]\right|^2\end{aligned} \tag{12}$$

故用户 k 在第 n 个时隙的卸载速率为

$$R_k[n]=B\mathrm{lb}\left(1+\gamma_k[n]\right) \tag{13}$$

1.2　感知模型

本文采用雷达信息估计率来衡量雷达的感知性能[21]。由于雷达照射在目标上的照度可视为目标被动地传递其参数信息。因此，可以将雷达估计信息率视为雷达与目标之间的互信息，即接收到的回波信号提供的关于目标参数的信息量。目标参数与接收回波之间的互信息越大，UAV 可以从目标处收集到的信息越多。

利用串行干扰消除法将通信信号从观测波形中去除，从而得到无通信干扰的雷达回波信号。因此，UAV 在第 n 个时隙接收到的被抑制雷达回波信号的信噪比 $\gamma_r[n]$ 可表示为

$$\gamma_r[n]=\frac{\eta^2B^2\sigma_r^2\left|\boldsymbol{c}[n]^{\mathrm{H}}\boldsymbol{H}_r[n]\boldsymbol{w}_r[n]\right|^2}{\sigma^2\boldsymbol{c}[n]^{\mathrm{H}}\boldsymbol{c}[n]} \tag{14}$$

其中，$\boldsymbol{c}[n]\in\mathbb{C}^{M\times1}$ 表示线性有限脉冲响应滤波器。因此，在第 n 个时隙雷达信息估计率 $R_{\text{rad}}[n]$ 为

$$R_{\text{rad}}[n]=\frac{\delta}{2\mu}\mathrm{lb}\left(1+2B\mu\gamma_r[n]\right) \tag{15}$$

其中，μ 为雷达脉冲时长，δ 为雷达占空比因子。

1.3　计算模型

在第 n 个时隙开始时，用户 k 生成一个任务 $\boldsymbol{\Omega}_k[n]=(L_k[n],C_k[n],T_k[n])$。$L_k[n]$ 为生成的计算任务数据量，$C_k[n]$ 为在用户 k 上处理每比特数据的周期数。为了简化分析，任务必须在一个时隙内完成。本文采取部分卸载策略，即根据卸载比例 $\rho_k[n]$ 将每个计算任务分成两部分，$(1-\rho_k[n])L_k[n]$ 的数据量在本地计算，剩余 $\rho_k[n]L_k[n]$ 数据卸载到 UAV 进行计算。因此，在第 n 个时隙内，用户 k 的本地计算时延为

$$t_k^{\text{loc}}[n]=\frac{(1-\rho_k[n])L_k[n]C_k[n]}{f_k[n]} \tag{16}$$

其中，$f_k[n]$ 表示用户 k 在第 n 个时隙的计算频率。用户 k 将任务卸载到 UAV 的传输时延为

$$t_k^{\text{tr}}[n]=\frac{\rho_k[n]L_k[n]}{R_k[n]} \tag{17}$$

UAV 执行用户 k 卸载的任务时所需要的计算时延可表示为

$$t_k^{\text{e}}[n]=\frac{\rho_k[n]L_k[n]C_k[n]}{f_k^{\text{e}}[n]} \tag{18}$$

其中，$f_k^{\text{e}}[n]$ 为 UAV 在第 n 个时隙的处理速率。

由于计算结果的数据量通常很小，相对于传输过程中的数据时延而言，计算结果的传输时延可以忽略不计。因此，为简化问题并提高系统的实时性，本文假设 UAV 返回结果的传输是即时完成的，以便更好地研究和优化 UAV 辅助 ISCC 网络的其他关键问题。综上，根据式(16)～式(18)，在每个时隙内执行用户 k 生成的任务所需要的最大时延为

$$t_k[n]=\max\left\{t_k^{\text{loc}}[n],t_k^{\text{e}}[n]+t_k^{\text{tr}}[n]\right\} \tag{19}$$

在每个时隙内，用户 k 的能耗 $E_k[n]$ 包括本地计算能耗 $E_k^{\text{loc}}[n]$ 和数据卸载能耗 $E_k^{\text{tr}}[n]$，即

$$E_k[n]=E_k^{\text{loc}}[n]+E_k^{\text{tr}}[n] \tag{20}$$

其中，用户 k 的本地计算能耗 $E_k^{\text{loc}}[n]$ 和数据卸载能耗 $E_k^{\text{tr}}[n]$ 可分别表示为

$$E_k^{\text{loc}}[n]=\phi_1f_k^2[n](1-\rho_k[n])L_k[n]C_k[n] \tag{21}$$

$$E_k^{\mathrm{tr}}[n]=\frac{\rho_k[n]L_k[n]}{R_k[n]}p_k[n] \tag{22}$$

同理，用户在每个时隙根据卸载比例将任务卸载至 UAV 上。UAV 对用户 k 卸载的任务进行的计算能耗 $E_k^{\mathrm{e}}[n]$ 可表示为

$$E_k^{\mathrm{e}}[n]=\phi_2\left(f_k^{\mathrm{e}}[n]\right)^2\rho_k[n]L_k[n]C_k[n] \tag{23}$$

其中，ϕ_1 和 ϕ_2 分别为用户和 UAV 有效电容系数。

在时隙 n 中，UAV 的飞行功率为

$$\begin{aligned}p_{\mathrm{fly}}[n]=&P_1\left(1+\frac{3\|\boldsymbol{v}[n]\|^2}{U_{\mathrm{tip}}^2}\right)+\\&P_2\left(\sqrt{1+\frac{\|\boldsymbol{v}[n]\|^4}{4v_0^2}}-\frac{\|\boldsymbol{v}[n]\|^2}{2v_0^2}\right)^{\frac{1}{2}}+\\&\frac{1}{2}d_0\rho_0 gA\|\boldsymbol{v}[n]\|^3\end{aligned} \tag{24}$$

其中，P_1 为 UAV 叶片旋转功率，P_2 为 UAV 悬停功率，U_{tip} 为叶片尖端速度，v_0 为 UAV 悬停平均转子速度，ρ_0、A、d_0 和 g 分别表示空气密度、转子盘面积、机身阻力比和转子稳定度。因此，UAV 的飞行能耗 $E_{\mathrm{fly}}[n]$可表示为

$$E_{\mathrm{fly}}[n]=p_{\mathrm{fly}}[n]\delta_n \tag{25}$$

根据式(23)和式(25)，UAV 在每个时隙内的能耗 $E_{\mathrm{U}}[n]$ 为

$$E_{\mathrm{U}}[n]=E_{\mathrm{fly}}[n]+\sum_{k=1}^{K}E_k^{\mathrm{e}}[n] \tag{26}$$

1.4 问题建立

本文通过联合优化任务卸载比例 $\boldsymbol{\rho}\triangleq\{\rho_k[n],\forall k\in\mathcal{K},n\in\mathcal{N}\}$、UAV 计算资源分配 $\boldsymbol{f}_e\triangleq\{f_k^{\mathrm{e}}[n],\forall k\in\mathcal{K},n\in\mathcal{N}\}$、UAV 飞行轨迹 $\boldsymbol{q}\triangleq\{\boldsymbol{q}[n],\forall n\in\mathcal{N}\}$、用户计算资源分配 $\boldsymbol{f}_K\triangleq\{f_k[n],\forall k\in\mathcal{K},n\in\mathcal{N}\}$ 和波束成形 $\boldsymbol{W}\triangleq\{\boldsymbol{w}_k[n],\forall k\in\mathcal{K},n\in\mathcal{N}\}$，旨在最小化整个周期 T 内的系统总能耗。故优化问题如式(27)所示

$$\begin{aligned}&\min_{\{\boldsymbol{q},\boldsymbol{W},\boldsymbol{\rho},\boldsymbol{f}_e,\boldsymbol{f}_l\}}\ \sum_{n=1}^{N}\left(\varpi_1E_{\mathrm{U}}[n]+\varpi_2\sum_{k=1}^{K}E_k[n]\right)\\&\text{s.t. C1: }0\leqslant\rho_k[n]\leqslant1,\forall k\in\mathcal{K},n\in\mathcal{N}\\&\quad\text{C2: }0\leqslant f_k^{\mathrm{e}}[n]\leqslant f_e^{\max},\forall k\in\mathcal{K},n\in\mathcal{N}\\&\quad\text{C3: }\sum_{k=1}^{K}f_k^{\mathrm{e}}[n]\leqslant f_e^{\max},\forall k\in\mathcal{K},n\in\mathcal{N}\\&\quad\text{C4: }0\leqslant f_k[n]\leqslant f_k^{\max},\forall k\in\mathcal{K},n\in\mathcal{N}\\&\quad\text{C5: }t_k[n]\leqslant T_k[n],\forall k\in\mathcal{K},n\in\mathcal{N}\\&\quad\text{C6: }R_{\mathrm{rad}}[n]\geqslant R_{\mathrm{rad}}^{\min},\forall n\in\mathcal{N}\\&\quad\text{C7: }0\leqslant p_k[n]\leqslant P_k^{\max},\forall k\in\mathcal{K},n\in\mathcal{N}\\&\quad\text{C8: }\|\boldsymbol{a}[n]\|\leqslant a_{\max},\forall n\in\mathcal{N}\\&\quad\text{C9: }\|\boldsymbol{v}[n]\|\leqslant v_{\max},\forall n\in\mathcal{N}\\&\quad\text{C10: }\|\boldsymbol{q}[n+1]-\boldsymbol{q}[n]\|^2\leqslant v_{\max}\delta_t,\forall n\in\mathcal{N}\end{aligned} \tag{27}$$

其中，ϖ_1 和 ϖ_2 为权重因子，$\varpi_1+\varpi_2=1$；$f_e^{\max}$ 和 $f_k^{\max}$ 分别为 UAV 和用户的最大计算频率资源；$R_{\mathrm{rad}}^{\min}$ 为最小雷达估计信息率；$P_k^{\max}$ 为用户最大传输功率；$a_{\max}$ 为 UAV 最大加速度；$v_{\max}$ 为 UAV 最大飞行速度。约束条件 C1 表示每个用户在时隙内的任务卸载比例，约束条件 C2 和 C3 分别限制 UAV 和用户的计算频率资源，约束条件 C4 表示 UAV 计算资源分配，约束条件 C5 为任务时延约束，约束条件 C6 为 UAV 雷达感知约束，约束条件 C7 限制了用户 k 的传输功率，约束条件 C8 ～C10 限制了 UAV 的飞行轨迹。

由于目标函数的非凸性、场景动态性和任务多样性，传统的离线优化方法难以对其求解[22-23]。为了实现在线实时决策，本文采用 DRL 方法求解该问题。DRL 是一种自适应机器学习方法，它能够与环境进行交互、学习，最终得到一个可以部署在用户上的策略模型，从而根据当前状态进行实时决策，进而得到问题的满意解。

2 优化问题求解

本节将式(27)表述成马尔可夫决策过程（MDP, Markov decision process）问题[21]，并通过 DRL 方法从训练环境中学习最优策略来解决 ISCC 中的能耗最小化问题。

2.1 MDP 模型

在本文场景中，UAV 不需要任何关于环境的先验信息，只能从环境状态中获取因果信息，因此本文模型中转移概率未知，可建模为无模型、无转移概率的马尔可夫决策过程[21]。在 MDP 中，智能体与动态环境不断交互以优化自身策略。例如，在时间步长 n，环境处于某一状态 s_n，智能体执行动作 a_n，环境以一定的概率转移到任一可行的后继状态 s_{n+1}中，智能体接收到奖励 r_n，随后 n 增加 1。智能体通过观察状态 s_{n+1}与奖励 r_n 来调整自身策略，从而使积累奖励最大化。在此过程中，状态空间、动作空间和奖励函数是 3 个关键要素。

1) 状态空间

为了综合考虑ISCC中设备任务与UAV资源之间的特性，本文定义在时间步长 n 的状态空间 s_n 可表示为

$$s_n = \{\boldsymbol{q}[n], \boldsymbol{v}[n], \boldsymbol{L}[n], \boldsymbol{C}[n]\} \tag{28}$$

其中，$\boldsymbol{q}[n]$ 表示当前UAV的水平坐标，$\boldsymbol{v}[n]$ 表示UAV的速度，$L[n]=\left[L_1[n],\cdots,L_k[n]\right]$ 和 $\boldsymbol{C}[n]=\left[C_1[n],\cdots,C_k[n]\right]$ 分别为用户任务数据量和任务所需的 CPU 资源。

2) 动作空间

智能体根据状态空间 s_n 输出动作 a_n，并将动作映射为任务卸载比、UAV 资源分配、UAV 飞行轨迹和波束成形的优化变量。因此，该动作可以表示为

$$a_n = \{\rho[n], f_e[n], \boldsymbol{w}[n], \boldsymbol{a}[n]\} \tag{29}$$

同时，为了最大限度地减少用户计算能量，本文根据动态电压频率调节技术，通过式(30)设置并估计 CPU 频率

$$f_k[n] = \min\{f_k^{\max}, \frac{1}{t_k[n]} L_k[n] C_k[n]\} \tag{30}$$

3) 奖励函数

智能体根据观察到的状态执行动作，并从环境中获得奖励，为了长期实现式(27)中的优化目标，并考虑约束条件的满足程度，本文设计了与系统能耗相似的奖励函数。奖励包含系统的能量消耗与违反时延约束和感知约束的惩罚，同时 UAV 的边界惩罚也考虑其中。因此本文设计的奖励函数 r_n 为

$$r_n = -\left[\varpi_1 E_{\mathrm{U}}[n] + \varpi_2 \sum_{k=1}^{K} E_k[n]\right] P_n^{\mathrm{rad}} P_n^{\mathrm{W}} P_n^{\mathrm{T}} \tag{31}$$

其中，感知约束惩罚 P_n^{rad} 和边界惩罚 P_n^{W} 为线性惩罚函数，时延约束惩罚 P_n^{T} 为指数惩罚函数，其可分别表示为

$$P_n^{\mathrm{rad}} = 1 + \frac{R_{\mathrm{rad}}^{\min}}{\bar{R}_{\mathrm{rad}} - R_{\mathrm{rad}}^{\min}} \tag{32}$$

$$P_n^{\mathrm{W}} = 1 + \frac{1}{v_{\max}} \left\| q[n] - \mathrm{clip}(q[n], 0, X) \right\| \tag{33}$$

$$P_n^{\mathrm{T}} = \frac{1}{K} \sum_{k \in K} P\left(t_k[n], T_k[n], T_k[n]\right) = \frac{1}{K} \sum_{k \in K} \left(2 - \exp\left(-\left\lceil \frac{t_k[n] - T_k[n]}{T_k[n]} \right\rceil^{+}\right)\right) \tag{34}$$

其中，$\mathrm{clip}(\cdot)$ 为梯度裁剪函数，$\bar{R}_{\mathrm{rad}}$ 为平均信息估计率，$\lceil\ \rceil^{+}$ 为向上取整。

2.2 基于 PPO 的 DRL 训练框架

由于上述状态空间和动作空间都是连续的，本文采用 PPO 算法实现系统能耗的最小化。该算法不仅考虑新动作策略，还兼顾旧动作策略，通过设置一个新的目标函数，将动作值稳定在近端区间，从而使新的动作策略可以参照旧策略进行更新，同时具有动态决策的优势，可以快速决定模型优化方向，进而在实现系统能耗最小化的同时提高算法效率。基于 PPO 的 DRL 训练框架如图 2 所示。

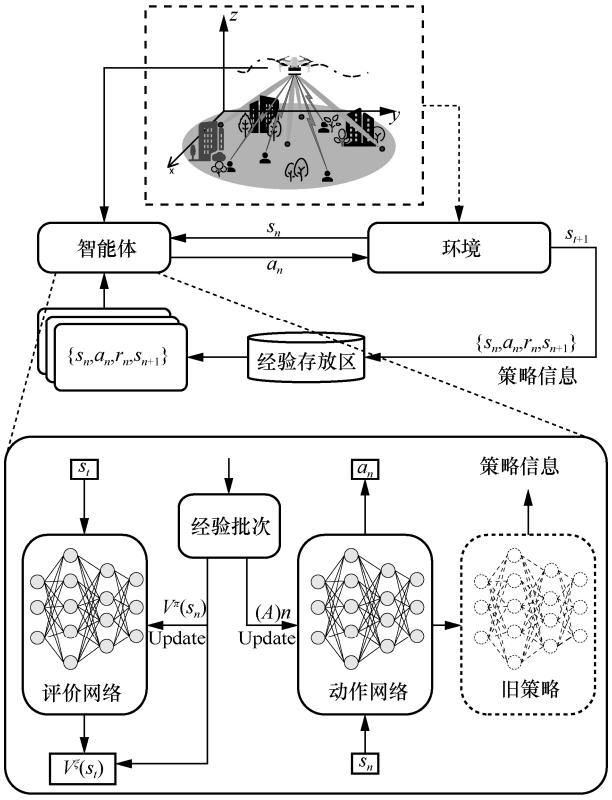

图 2　基于 PPO 的 DRL 训练框架

PPO 算法采用动作−评价（AC, actor-critic）结构，其中包含动作网络和评价网络，动作网络分为新旧 2 个部分，分别对应参数 θ 和 θ_{old}，评价网络参数为 ξ。动作网络根据状态 s_n 输出动作 a_n，并与环境交互；评价网络根据状态信息计算状态价值 $V^{\xi}(s_n)$，可表示为

$$V^{\xi}(s_n) = \mathbb{E}_{s_n, a_n}\left[\sum_{l=0}^{\infty} \gamma^l \mathcal{R}(a_{n+l} \mid s_{n+l})\right] \tag{35}$$

其中，γ 为折扣因子，$\mathbb{E}[\cdot]$ 表示期望值，$\mathcal{R}(\cdot)$ 表示关于状态和动作的奖励函数。为了评估动作 a_n 的性能，算法引入优势函数 $\hat{A}(s_n)$，即

$$\hat{A}(s_n)=\sum_{l=0}^{\infty}(\gamma\lambda)^l(r_n+\gamma V(s_{n+1})-V(s_n)) \tag{36}$$

为保障策略更新的稳定性，式(36)采用广义优势估计（GAE, general advantage estimation）的形式，其中，$0\leqslant\lambda\leqslant1$为 GAE 因子。随后，计算动作网络$\theta$和评价网络$\xi$的目标函数，其可分别表示为

$$L^{\text{actor}}(\theta)=\mathbb{E}_{\pi_\theta}\left\{\min\left[\frac{\pi_\theta(a_n\mid s_n)}{\pi_{\theta_{\text{old}}}(a_n\mid s_n)}\hat{A}(s_n),\ \text{clip}\left(\frac{\pi_\theta(a_n\mid s_n)}{\pi_{\theta_{\text{old}}}(a_n\mid s_n)},1-\varepsilon,1+\varepsilon\right)\hat{A}(s_n)\right]\right\} \tag{37}$$

$$L^{\text{critic}}(\xi)=\left[V^{\xi}(s_{n+1})-V^{\xi}(s_n)\right]^2 \tag{38}$$

其中，$\pi_\theta(\cdot)$和$\pi_{\theta_{\text{old}}}(\cdot)$分别表示新、旧策略函数，$\varepsilon$为截断参数。

为降低训练难度，在动作函数方面，本文引入具有剪切概率比的目标函数。ε用来决定新旧策略之间的差异。最后根据式(37)和式(38)计算目标网络的梯度，通过梯度下降法对参数θ和ξ进行更新，完成一轮迭代。算法 1 给出了基于 PPO 算法的 DRL 训练算法伪代码。

算法 1 基于 PPO 的 DRL 训练算法

输入 最大训练集l_{Me}，每一个训练集的长度l_{El}，学习率α，GAE 因子λ，截断参数ε，评价网络参数ξ

输出 动作网络参数θ

1) 初始化：评价网络参数ξ，动作网络参数θ
2) for $m=1,\cdots,l_{\text{Me}}$ do
3) 初始化：用户位置(x_k,y_k)，UAV 初始坐标$q[0]$，用户任务Ω_k，飞行高度H
4) for $n=1,\cdots,l_{\text{El}}$ do
5) 从环境中获取状态s_n
6) 智能体根据状态s_n做出决策π_θ，选择动作a_n
7) 根据动作a_n计算下一状态s_{n+1}
8) 根据式(31)计算奖励r_n
9) 存储经验(s_n,a_n,r_n,s_{n+1})
10) end for
11) for $n=1,\cdots,l_{\text{El}}$ do
12) 根据式(36)计算$\hat{A}(s_n)$
13) end for
14) 根据式(37)和式(38)更新动作网络参数θ和评价网络参数ξ
15) 更新$\theta_{\text{old}}\leftarrow\theta$
16) 清理经验数据
17) end for

2.3 计算复杂度分析

本文方案中，算法 1 的复杂度以一次迭代中乘法计算次数来衡量[24]。在 DRL 框架中，智能体首先将观测到的状态值发送至多层感知器（MLP, multi-layer perceptron），MLP 由一个输入层、一个输出层和若干个隐藏层组成。每一隐藏层的复杂度可表示为$O\left(U_{j-1}U_j+U_jU_{j+1}\right)$，其中$U_j$为第$j$层隐藏层神经元数量。由于输入层和输出层的乘法运算次数远少于隐藏层，可忽略其对复杂度的影响。因此，J层 MLP 的复杂度为$O\left(\sum_{j=2}^{J-1}\left(U_{j-1}U_j+U_jU_{j+1}\right)\right)$。本文采用的动作网络和评价网络均由一个 MLP 组成。结合上述分析，可以得到算法 1 的总复杂度为$O\left(l_{\text{Me}}l_{\text{El}}\left(\sum_{j=2}^{J-1}\left(U_{j-1}U_j+U_jU_{j+1}\right)\right)\right)$。

3 仿真结果与分析

3.1 参数设置

本节提供数据仿真以验证本文提出的基于 PPO 算法的 UAV 辅助 ISCC 网络对系统总能耗的影响，采用 Py-Torch 框架搭建仿真环境并分析所提方案的性能。考虑一个面积为 500 m × 500 m 的地面正方形区域，用户随机分布在该区域内，设置 UAV 飞行高度为 200 m。任务数据大小均匀分布在$[0.5\,\text{MB},L_{\max}]$，其中，$L_{\max}$默认为1.5 MB，单位比特平均计算次数$C_k[n]\in[500,1\,500]$cycles/bit，任务周期$T=200$ s，时隙持续时间$\delta_n=1$ s。若非特别说明，用户与 UAV 通信信道带宽设置为$B=10$ MHz，噪声功率σ^2和σ_r^2为−65 dBm，参考距离$d_0=1$ m 处信道功率增益β_0为−30 dB，莱斯因子$\kappa=4$，雷达波形功率谱密度常数η、雷达脉冲时长μ和雷达占空比因子δ分别为$\frac{2\pi}{\sqrt{12}}$、2×10^{-5} s和0.01。同时，本文设置最小雷达估计信息率$R_{\text{rad}}^{\min}=10^3$ dB，CPU 有效电容系数$\phi_1=\phi_2=10^{-27}$，用户最大传输功率$P_k^{\max}=0.5$ W，用户最大计算频率$f_k^{\max}=1$ GHz，

UAV 最大计算频率 $f_e^{\max}=10$ GHz。此外，在与无人机飞行相关的参数设置中，其最大飞行速度和加速度分别为 20 m/s 和 5 m/s^2，叶片旋转功率 $P_1=79.07$ W，悬停功率 $P_2=79.07$ W，叶片尖端速度 $U_{\text{tip}}=120$ m/s，悬停平均转子速度 $v_0=3.6$ m/s，转子盘面积 $A=0.503\,0$ m^2 和转子稳定度 $g=0.05$。PPO 算法相关的训练参数如表 1 所示。

表 1　　PPO 训练参数

参数	数值
最大训练集 l_{Me} /episode	300
每一个训练集的长度 l_{El} /step	200
学习率 α	0.0005
折扣因子 γ	0.98
截断参数 ε	0.2
GAE 因子 λ	0.95
隐藏层大小	64 和 128
优化器	Adam

为了验证基于 PPO 算法的性能，本文将其与以下基准算法进行比较。

1) 优势动作评论（A2C, advantage actor critic）算法。A2C 算法是一种将优势函数引入 AC 结构中的同策略算法，A2C 算法用优势函数代替评价网络中的原始回报，作为衡量所选动作与所有动作平均值好坏的指标。

2) 深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法。DDPG 算法为异策略 DRL 算法，该算法直接输出动作向量而不是概率分布，这需要一个较大的重放缓冲区来学习动作价值函数。

3.2　仿真评估

设置用户数量 $K=10$，UAV 天线数 $M=4$，PPO 算法的收敛性如图 3 所示，从图 3 可以看出，随着训练步数的增加，所提方案的奖励也逐渐上升，强化学习智能体可以显著提升每一训练步奖励值，这证实了 PPO 算法在计算卸载方面的有效性。使用 Py-Torch 收集训练 60 000 步的结果，每个结果为一个回合内的奖励值之和，随着训练步数的增加，智能体在通信、感知和计算方面的策略逐渐优化，收敛性曲线的振荡有明显渐弱的趋势，最终算法能获得较为稳定的奖励值。为了验证学习率对算法收敛性的影响，本文还比较了不同学习率下奖励值的收敛曲线。从图 3 可以看出，当学习率为 8×10^{-4} 时，奖励值曲线在 3 000 步左右实现收敛；当学习率为 8×10^{-5} 时，曲线在 10 000 步左右收敛；当学习率分别为 5×10^{-4} 和 2×10^{-4} 时，曲线收敛性介于两者之间。虽然不同的学习率对收敛性有一定影响，但当 4 条曲线均达到收敛后，可以看到所得到的奖励值相差不大且处于较为稳定的区间内，由此说明，学习率对本文 PPO 算法的收敛速度具有一定影响，但对于性能的影响较小。

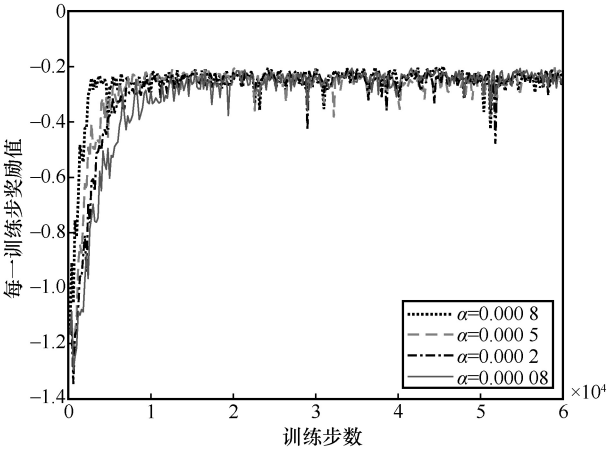

图 3　PPO 算法的收敛性

不同算法的奖励收敛性比较如图 4 所示。从图 4 可以看出，本文所提 PPO 算法比 A2C 算法和 DDPG 算法收敛更快，总体上获得了更高的奖励。另一方面，从图 4 中可以观察到，DDPG 算法在早期很难提高奖励，其训练过程相对于基于策略网络的 PPO 算法和 A2C 算法更加曲折。这是因为 DDPG 算法使用确定性动作输出而不是分布式的动作输出，这限制了其在动作空间的探索能力，导致其收敛困难且复杂。

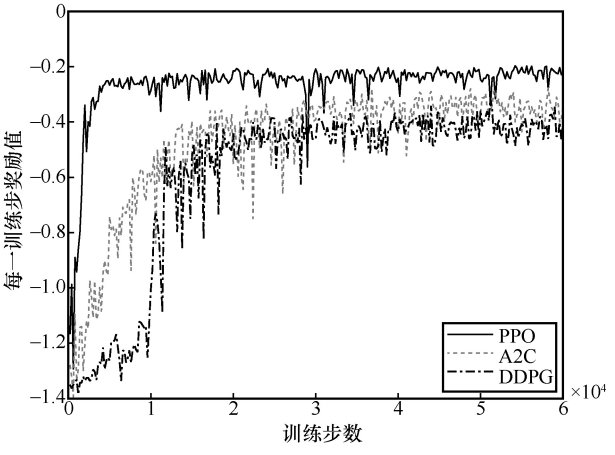

图 4　不同算法的奖励收敛性比较

图 5 比较了不同算法或卸载策略在不同用户数量下的系统加权能耗变化情况。整体而言，本文所提 PPO 算法在 ISCC 网络中表现最好，而 DDPG 算法在系统加权能耗方面与基于策略网络的算法差距较大，尤其当用户数量增加、趋于密集时，采用 DDPG 算法执行整个周期的 ISCC 任务所需能耗是基于 PPO 算法的近 2 倍。此外，与所提算法相比，无论是采用任务全部卸载策略还是用户随机卸载策略，产生的加权能耗都高于本文采用卸载因子 ρ 进行部分卸载决策的加权能耗，这证明在联合优化中考虑卸载因子 ρ 可以在减少系统能耗方面获得更好的性能。此外，从图 5 可以看出，相邻用户数量之间的加权能耗也有增大的趋势。这是因为当越来越多的用户接入网络时，用户之间的信号干扰增加，传输速率降低，进而提高传输成本，从而使用户卸载至 UAV 的任务量减少，本地计算任务量增加，用户需要更多的计算资源来处理任务，导致系统能耗不断增加。

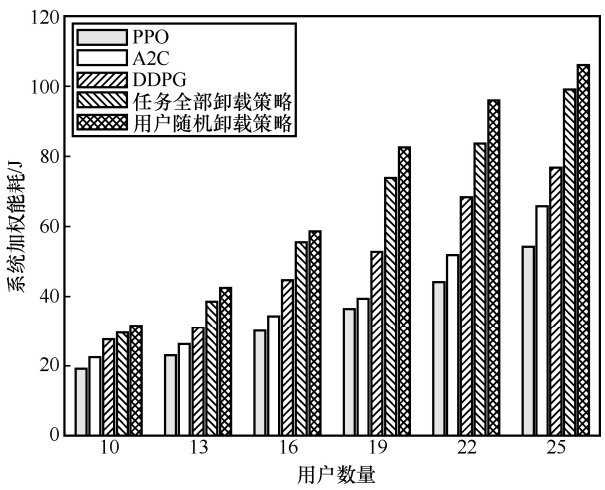

图 5 不同用户数量下的系统加权能耗变化情况

不同的用户数量下 UAV 飞行轨迹的变化情况如图 6 所示。从图 6 可以看出，UAV 能够选择用户较多的区域，并且能够根据用户的分布情况自适应地更新其位置。同时，这也意味着奖励可以引导 UAV 找到用户分布相对公平的区域，然后采取悬停或缓慢移动的策略以节省飞行能耗。相较于文献[8]中简单地控制方向和速度，本文 UAV 飞行轨迹相对平滑，适用于 UAV 的实际运动，这体现了本文算法在 UAV 飞行轨迹设计中的有效性。

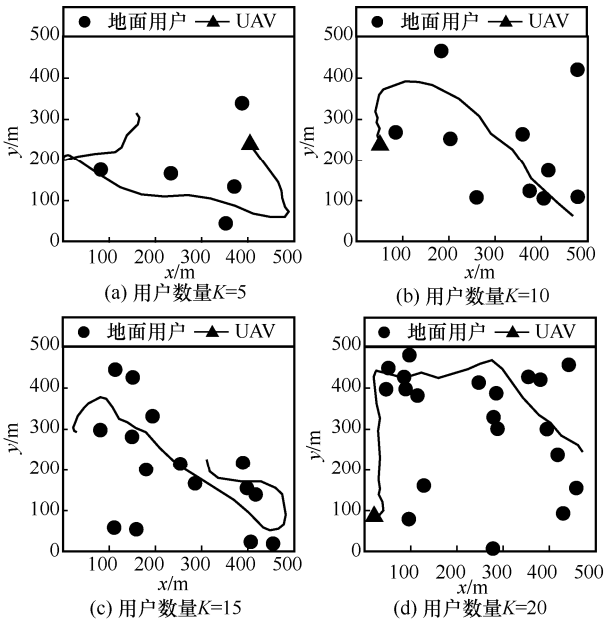

图 6 不同用户数量下 UAV 飞行轨迹的变化情况

4 结束语

本文研究了无人机辅助通感算融合网络波束成形和资源优化问题。为了最小化系统总能耗，本文联合优化波束成形、任务卸载比、用户和无人机计算资源分配和无人机飞行轨迹，并提出了一种基于 PPO 的资源分配与策略优化算法。仿真结果表明，训练得到的智能体能以较低的复杂度生成资源分配与策略优化。同时，与基准算法相比，本文算法能显著降低系统能耗。未来的工作将考虑用户移动环境下 ISCC 网络资源分配与优化决策问题。

参考文献：

[1] LETAIEF K B, SHI Y M, LU J M, et al. Edge artificial intelligence for 6G: vision, enabling technologies, and applications[J]. IEEE Journal on Selected Areas in Communications, 2021, 40(1): 5-36.

[2] ZHOU Y, LIU L, WANG L, et al. Service-aware 6G: an intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.

[3] LIU Y Q, PENG M G, SHOU G C, et al. Toward edge intelligence: multiaccess edge computing for 5G and Internet of things[J]. IEEE Internet of Things Journal, 2020, 7(8): 6722-6747.

[4] CHENG K J, FANG X M, WANG X B. Optimized resource allocation and time partitioning for integrated communication, sensing, and edge computing network[J]. Computer Communications, 2022, 194: 240-249.

[5] WANG Z L, MU X D, LIU Y W, et al. NOMA-aided joint communication, sensing, and multi-tier computing systems[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(3): 574-588.

[6] YU Z Y, HU X L, LIU C X, et al. Location sensing and beamforming design for IRS-enabled multi-user ISAC systems[J]. IEEE Transactions on Signal Processing, 2022, 70: 5178-5193.

[7] DO-DUY T, NGUYEN L D, DUONG T Q, et al. Joint optimisation of real-time deployment and resource allocation for UAV-aided disaster emergency communications[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(11): 3411-3424.

[8] 吴义豪, 齐彦丽, 周一青, 等. 通感算协同的无人机群轨迹规划与功率分配[J]. 西安电子科技大学学报, 2023, 50(3): 61-74.
WU Y H, QI Y L, ZHOU Y Q, et al. UAVs trajectory planning and power allocation based on the convergence of communication, sensing and computing[J]. Journal of Xidian University, 2023, 50(3): 61-74.

[9] LIU B Y, WAN Y Y, ZHOU F H, et al. Resource allocation and trajectory design for MISO UAV-assisted MEC networks[J]. IEEE Transactions on Vehicular Technology, 2022, 71(5): 4933-4948.

[10] LU W D, DING Y, GAO Y, et al. Secure NOMA-based UAV-MEC network towards a flying eavesdropper[J]. IEEE Transactions on Communications, 2022, 70(5): 3364-3376.

[11] WANG D, TIAN J, ZHANG H X, et al. Task offloading and trajectory scheduling for UAV-enabled MEC networks: an optimal transport theory perspective[J]. IEEE Wireless Communications Letters, 2022, 11(1): 150-154.

[12] GAN Y H, HE Y J. Trajectory optimization and computing offloading strategy in UAV-assisted MEC system[C]//Proceedings of 2021 Computing, Communications and IoT Applications (ComComAp). Piscataway: IEEE Press, 2021: 132-137.

[13] JEONG S, SIMEONE O, KANG J. Mobile edge computing via a UAV-mounted cloudlet: optimization of bit allocation and path planning[J]. IEEE Transactions on Vehicular Technology, 2018, 67(3): 2049-2063.

[14] HUANG N, WANG T S, WU Y, et al. Integrated sensing and communication assisted mobile edge computing: an energy-efficient design via intelligent reflecting surface[J]. IEEE Wireless Communications Letters, 2022, 11(10): 2085-2089.

[15] QI Q, CHEN X M, KHALILI A, et al. Integrating sensing, computing, and communication in 6G wireless networks: design and optimization[J]. IEEE Transactions on Communications, 2022, 70(9): 6212-6227.

[16] DING C F, WANG J B, ZHANG H, et al. Joint MIMO precoding and computation resource allocation for dual-function radar and communication systems with mobile edge computing[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(7): 2085-2102.

[17] ZHAO L D, WU D, ZHOU L, et al. Radio resource allocation for integrated sensing, communication, and computation networks[J]. IEEE Transactions on Wireless Communications, 2022, 21(10): 8675-8687.

[18] QI Y L, ZHOU Y Q, LIU Y F, et al. Traffic-aware task offloading based on convergence of communication and sensing in vehicular edge computing[J]. IEEE Internet of Things Journal, 2021, 8(24): 17762-17777.

[19] HUA M, WU Q Q, CHEN W, et al. Secure intelligent reflecting surface aided integrated sensing and communication[J]. IEEE Transactions on Wireless Communications, 2023, PP(99): 1.

[20] CHIRIYATH A R, PAUL B, JACYNA G M, et al. Inner bounds on performance of radar and communications co-existence[J]. IEEE Transactions on Signal Processing, 2015, 64(2): 464-474.

[21] PENG H X, SHEN X M. Multi-agent reinforcement learning based resource management in MEC-and UAV-assisted vehicular networks[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(1): 131-141.

[22] ZHAO N, YE Z Y, PEI Y Y, et al. Multi-agent deep reinforcement learning for task offloading in UAV-assisted mobile edge computing[J]. IEEE Transactions on Wireless Communications, 2022, 21(9): 6949-6960.

[23] LIANG J B, ZHANG H H, JIANG C, et al. Research progress of task offloading based on deep reinforcement learning in mobile edge computing[J]. Computer Science, 2021, 48(7): 316-323.

[24] ENGSTROM L, ILYAS A, SANTURKAR S, et al. Implementation matters in deep policy gradients: a case study on PPO and TRPO[J]. arXiv Preprint, arXiv: 2005.12729, 2020.

[作者简介]

李斌（1987− ），男，山东济宁人，博士，南京信息工程大学副教授、硕士生导师，主要研究方向为无人机通信、移动边缘计算等。

彭思聪（2000− ），男，江苏盐城人，南京信息工程大学硕士生，主要研究方向为移动边缘计算、通感算一体化等。

费泽松（1977− ），男，安徽合肥人，博士，北京理工大学教授、博士生导师，主要研究方向为无线通信、多媒体信号处理等。