基于深度强化学习的移动边缘计算安全传输策略研究

王义君[1]，李嘉欣[1]，闫志颖[1]，吕婧莹[1]，钱志鸿[2]
（1.长春理工大学电子信息工程学院，吉林 长春 130022；2.吉林大学通信工程学院，吉林 长春 130012）

摘　要：在移动边缘计算中，任务卸载过程中会面临信息泄露和被窃听等安全问题。为了提高移动边缘计算系统的安全传输效率，提出了无人机辅助物理层安全传输策略。首先，构建了无人机（UAV）搭载的移动边缘计算系统，由I个用户设备、M架合法无人机（L-UAV）和N架窃听无人机（E-UAV）构成；其次，保证L-UAV在规定周期内完成卸载任务的同时，以通信系统安全传输效率最大化为目标，采用引入注意力机制的多智能体深度确定性策略梯度（A-MADDPG）算法进行问题求解与优化；最后，在保证卸载前提下实现用户的机密信息不被窃听者窃听和安全计算效率最大化，保障系统整体安全性。仿真结果表明，所提算法相较于其他基准算法展现了更佳性能，在安全传输效率方面表现优越。

关键词：移动边缘计算；物理层安全；深度强化学习；无人机辅助卸载
中图分类号：TN918.91
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025060

Research on secure transport strategy of mobile edge computing based on deep reinforcement learning

WANG Yijun[1], LI Jiaxin[1], YAN Zhiying[1], LYU Jingying[1], QIAN Zhihong[2]
1. College of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
2. College of Communication Engineering, Jilin University, Changchun 130012, China

Abstract: In mobile edge computing, the process of task unloading will face security problems such as information leakage and eavesdropping. To improve the unloading efficiency of mobile edge computing system, the physical layer security transmission strategy of mobile edge computing was proposed. Firstly, the mobile edge computing system based on unmanned aerial vehicle (UAV) was studied, which was composed of I user devices, M legal UAV (L-UAV) and N eavesdropping UAV (E-UAV). Secondly, while ensuring the unloading of L-UAV within a specified period, the multi-agent depth deterministic policy gradient (Attention-MADDPG) algorithm with the addition of attention mechanism was adopted to solve and optimize the problem with the aim of maximizing the safety unloading efficiency of the communication system. Finally, under the premise of ensuring uninstallation, the user's confidential information was not eavesdropped by the eavesdropper, and the secure computing efficiency was maximized to ensure the overall security of the system. Simulation results show that compared with other benchmark algorithms, the proposed algorithm has better performance in terms of secure transmission efficiency.

Keywords: mobile edge computing, physical layer security, deep reinforcement learning, UAV assisted offloading

收稿日期：2024–11–26；**修回日期：**2025–03–21
通信作者：王义君，wangyijun@cust.edu.cn
基金项目：国家自然科学基金资助项目（No.61540022）；吉林省科技厅重点研发基金资助项目（No.20230203091SF）
Foundation Items: The National Natural Science Foundation of China (No.61540022), The Jilin Province Science and Technology Department Key Research and Development Project (No.20230203091SF)

0　引言

随着各种智能设备的广泛普及，其产生的庞大任务数据为终端设备和云端服务器带来巨大计算压力。为了解决此问题，移动边缘计算（MEC, mobile edge computing）可以在网络边缘部署计算资源，在靠近数据源的地方进行数据分析和处理，能够有效地减少数据传输时延，提升实时响应能力，同时在一定程度上缓解带宽压力并提高数据隐私安全性。

现有的边缘基站大多为固定部署[1]，在一些偏远地区部署边缘基站成本较高。无人机因其灵活性可作为移动边缘基站应用于MEC系统中[2]。MEC系统采用无线网络传输数据，容易受到安全攻击，尤其是无人机的视距链路通信[3]在任务卸载时易被窃听从而产生信息泄露。基于以上问题，物理层安全技术[4]被认为是保护用户数据免受窃听的一种可行解决方案。

近年来，已有许多学者对移动边缘计算物理层安全通信进行研究。文献[5]研究了非正交多址下的无人机移动边缘计算网络，利用逐次凸逼近和块坐标下降法进行优化来提高系统安全性能。文献[6]研究了具有空中窃听者的边缘系统，将块坐标下降法与连续凸逼近（SCA, successive convex approximation）结合，实现了安全通信能效最大化。文献[7]研究了双无人机辅助MEC系统的安全问题，通过基于块坐标下降算法和惩罚块坐标下降算法解决时分多址和正交多址方案下的安全传输不确定性。

针对无人机通信的移动边缘计算物理层安全进行优化时，上述研究虽能有效提升系统安全性与能效，但存在局部最优的问题。深度强化学习（DRL, deep reinforcement learning）算法[8-9]能够自适应学习优化策略，具有更强的全局优化能力，从而实现更高效和更安全的通信。文献[10]研究了无人机辅助边缘计算，优化时延和能耗，提出基于深度强化学习的双深度Q网络（DDQN, double deep Q network）算法。文献[11]研究了多无人机辅助的移动边缘计算系统，提出了一种协作式智能体深度强化学习框架。多智能体深度强化学习算法作为一种新兴智能算法，应用在移动边缘计算系统中可以进行自主决策，从而寻找最优策略[12-13]。文献[14]研究了多无人机在目标区域上空飞行并支持地面上的用户设备，提出了基于多智能体深度强化学习的轨迹控制算法。

从已有的研究可以看出，多智能体深度强化学习算法应用在移动边缘计算系统中可以使该系统获得更高效的自主决策，以寻找最优策略。然而，多智能体在处理复杂的动态环境时，无法有效地聚焦于关键因素，主要包括2个方面：1) 环境动态性：MEC环境中的任务需求和资源状态变化迅速，智能体难以及时调整策略，同时，环境中的随机性和不可预测性增加了策略制定的难度；2) 多智能体协作：多个智能体在决策时缺乏有效协调，可能导致冲突或资源浪费，同时，智能体的目标可能不完全一致，增加了协作的复杂性。此外，现有研究仍未充分考虑无人机的能耗和安全性问题[15-16]。因此，针对多智能体深度强化学习算法进行改进，同时优化能耗与安全性是必要的。

本文搭建了存在空中窃听者的无人机辅助移动边缘计算系统，在物理层安全方面实现MEC系统下的安全传输，以最大化通信系统安全传输效率为目标，在保证卸载和能耗的前提下对多架无人机的卸载比例和传输功率进行优化。本文主要贡献如下。

1) 搭建了一个无人机辅助MEC的安全通信系统，该系统中用户设备（UD, user device）可以选择将计算任务进行本地计算或者部分卸载到合法无人机（L-UAV, legal UAV）上计算，窃听无人机（E-UAV, eavesdropping UAV）窃听用户卸载过程中的任务信息。为抵抗窃听者窃取信息，L-UAV向E-UAV发送干扰信号。

2) 提出一种基于注意力机制的多智能体深度确定性策略梯度（MADDPG）算法A-MADDPG，它是在MADDPG基础上加入注意力机制，增强了智能体之间的协作能力，从而做出更优的卸载决策，并对无人机位置以及传输功率进行优化，使通信系统安全传输效率最大化。

1　系统模型

1.1　网络模型

本文搭建了一个UAV辅助MEC的物理层安全系统架构，具有$I=\{1,2,\cdots,i\}$个UD、$M=\{1,2,\cdots,m\}$个L-UAV和$N=\{1,2,\cdots,n\}$个E-UAV。其中，L-UAV搭载了小型MEC服务器，用于辅助用户设备进行任务计算；E-UAV通过单一天线窃听用户设备发

送给L-UAV的任务信息，并由L-UAV向E-UAV发送干扰信号，以防止任务信息被窃听者窃取。UAV辅助MEC的物理层安全系统架构如图1所示。

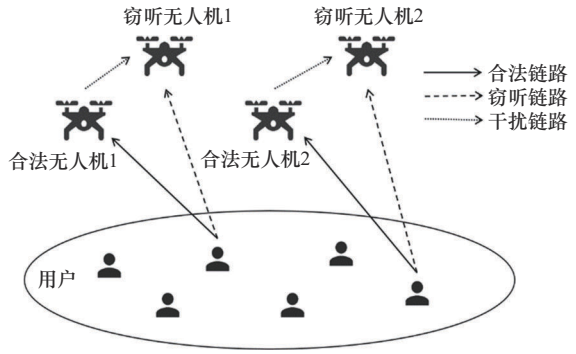

图1 UAV辅助MEC的物理层安全系统架构

在该系统架构中，每个用户设备配备单天线将计算任务进行部分卸载，需要向L-UAV进行任务数据传输，L-UAV接收用户的卸载信号进行计算并传回，同时向E-UAV发送干扰信号，以防任务信息被窃听者窃取。为了有效地抵御E-UAV窃听，L-UAV配备了双天线，分别用来接收用户的卸载信号和发送干扰信号给E-UAV。假设L-UAV通过多种技术手段，事先获取每一个UD和E-UAV的信道状态信息（CSI, channel state information），拥有全局完美CSI以及位置信息，这些信息为后续的信号处理与干扰策略提供了基础。具体而言，L-UAV利用机载光学相机和无线电频谱探测技术来全面获取这些关键信息。无线电频谱探测技术通过专门的硬件和软件设备，对无线电频谱进行扫描和分析，从而实时估算CSI。在干扰目标选择上，L-UAV优先选择信道增益较低的E-UAV进行干扰，若增益相近，则选择距离较近的E-UAV进行干扰。同时，E-UAV通过类似的技术，也能够获取所有用户设备的CSI和位置信息，但是E-UAV无法识别L-UAV的存在，这导致E-UAV会将接收到的干扰信号误认为是用户设备发送的任务数据，从而无法主动反制。L-UAV通过隐蔽自己的身份和伪装的干扰信号，避免被E-UAV发现。这样设定有效地确保了L-UAV的隐蔽性，使干扰策略的评估更加专注于L-UAV干扰的效果，确保安全分析不受外部因素影响，并且不影响测试结果的有效性。

整个系统在离散时间内运行，为了便于讨论，将整个通信周期T划分为t个时隙，L-UAV在一个时隙内只能服务于一个用户。

在三维笛卡儿坐标系下，L-UAV和E-UAV分别在固定的高度H_m和H_n飞行，这样设计可以有效避免碰撞风险。将其投影到xoy平面上，坐标分别为 $p_m(t)=\{x_m(t),y_m(t)\}^{\mathrm{T}}$ 和 $p_n(t)=\{x_n(t),y_n(t)\}^{\mathrm{T}}$，UD固定在坐标$w_i=\{x_i,y_i\}^{\mathrm{T}}$的位置上。

UD到L-UAV、UD到E-UAV和L-UAV到E-UAV的距离l分别表示为

$$l_{m,i}^2[t]=\left\|w_i-p_m(t)\right\|^2+H_m^2 \tag{1}$$

$$l_{n,i}^2[t]=\left\|w_i-p_n(t)\right\|^2+H_n^2 \tag{2}$$

$$l_{m,n}^2[t]=\left\|p_m(t)-p_n(t)\right\|^2+(H_m-H_n)^2 \tag{3}$$

1.2 通信模型

在该系统中，本文假设UAV与地面UD之间的通信受视距链路影响。系统通过动态链路选择机制，结合实时环境数据和链路条件，优先选择视距链路，忽略非视距链路。用户设备i到L-UAV之间的信道增益表示为

$$g_{mi}=\alpha_1 l_{m,i}^{-2}[t]=\alpha_1\left(\left\|w_i-p_m(t)\right\|^2+H_m^2\right)^{-1} \tag{4}$$

其中，α_1表示单位距离为1 m时UD到L-UAV的信道增益系数。

同样地，用户设备i到E-UAV和L-UAV到E-UAV之间的信道增益分别表示为

$$g_{ni}=\alpha_2 l_{n,i}^{-2}[t]=\alpha_2\left(\left\|w_i-p_n(t)\right\|^2+H_n^2\right)^{-1} \tag{5}$$

$$g_{mn}=\alpha_3 l_{m,n}^{-2}[t]=$$

$$\alpha_3\left(\left\|p_m(t)-p_n(t)\right\|^2+\left(H_m-H_n\right)^2\right)^{-1} \tag{6}$$

其中，α_2与α_3分别代表参考距离为1 m时UD到E-UAV和L-UAV到E-UAV的信道增益系数。

在时隙t，用户的发射功率$p_i[t]$不能超过峰值发射功率$P_{i\max}$，即$0\leqslant p_i[t]\leqslant P_{i\max}$。

L-UAV和E-UAV都知道UD的位置信息，L-UAV与用户设备属于合法通信网络，而E-UAV处于窃听网络中。当L-UAV向E-UAV发送干扰信号时，E-UAV会将其视为有用信号，因此在时隙t L-UAV和E-UAV的信噪比分别表示为

$$r_{m,i}[t]=\frac{p_i[t]\left|g_{mi}\right|^2}{p_m\left|g_m\right|^2+\delta_m^2},\forall m,i,t \tag{7}$$

$$r_{n,i}[t]=\frac{p_i[t]\left|g_{ni}\right|^2}{p_{mn}^{\mathrm{eva}}\left|g_{mn}\right|^2+\delta_n^2},\forall m,n,i,t \tag{8}$$

其中，δ_m^2 和 δ_n^2 分别为L-UAV与E-UAV下的高斯噪声，p_m 表示合法无人机的发射功率，g_m 表示自干扰信道增益，p_{mn}^{eva} 表示L-UAV向E-UAV发射的干扰功率。

因此，在时隙 t 用户设备 i 到L-UAV的任务卸载率和用户设备 i 到E-UAV的任务数据窃听率分别表示为

$$R_{m,i}[t]=\mathrm{lb}\,(1+r_{m,i}[t]) \tag{9}$$

$$R_{n,i}[t]=\mathrm{lb}\,(1+r_{n,i}[t]) \tag{10}$$

综上所述，用户设备 i 可实现的安全传输效率[17]可用 $R_{\mathrm{sec},i}[t]$ 表示，安全传输效率指在卸载数据时，从用户设备传输到边缘端的安全性和效率，确保数据在卸载过程中安全传输，可表示为

$$R_{\mathrm{sec},i}[t]=\left[R_{m,i}[t]-\max_{\forall n}R_{n,i}[t]\right]^+ \tag{11}$$

其中，$[x]^+\triangleq\max(x,0)$。

1.3　计算模型

用户设备任务执行方式采用部分卸载，将计算任务的一部分卸载到L-UAV进行计算，另一部分在本地执行计算。定义 λ_i 为用户将数据卸载到L-UAV的比例，$1-\lambda_i$ 为在本地计算的比例，应满足 $0\leqslant\lambda_i\leqslant1,\forall i$ 的约束，并假设用户设备 i 产生的任务数据大小为 D_i。由于L-UAV作为小型设备其电池容量有限，因此对能耗进行分析十分重要。本文中L-UAV的能耗主要分为4个部分。

1) L-UAV在飞行时的飞行能耗

L-UAV从一个位置飞行到一个新位置产生的能耗表示为

$$E_m^{\mathrm{fly}}(t)=0.5Gt_{\mathrm{fly}}v^2(t) \tag{12}$$

其中，G 表示无人机的飞行载荷；t_{fly} 是飞行时间；$v(t)\in\left[0,v_{\max}\right]$ 代表飞行速度，$v_{\max}$ 是飞行最大速度。

2) L-UAV接收用户任务数据时的通信能耗

用户设备 i 将数据任务卸载到L-UAV进行计算，卸载过程总时延由3个部分组成：从用户设备到L-UAV的上行链路传输时延、L-UAV的计算时延和结果传回时延。与整个通信周期的计算时延相比，结果传回时延非常短，因此忽略下行链路传输时延。则将计算任务卸载到L-UAV产生的通信传输时延为

$$t_i^{\mathrm{tra}}=\frac{\lambda_iD_i}{BR_{m,i}} \tag{13}$$

其中，B 表示传输带宽，λ_iD_i 表示数据卸载到合法无人机计算的大小。

定义 p_m^{tra} 为L-UAV的通信功率，则L-UAV在卸载任务过程中产生的通信能耗为

$$E_i^{\mathrm{tra}}=p_m^{\mathrm{tra}}t_i^{\mathrm{tra}}=\frac{p_m^{\mathrm{tra}}\lambda_iD_i}{BR_{m,i}} \tag{14}$$

3) L-UAV计算任务数据时的计算能耗

用户设备 i 将数据任务卸载到L-UAV进行计算，任务卸载到L-UAV上的计算执行时延为

$$t_i^{\mathrm{exe}}=\frac{\lambda_iD_iC_m}{f_{m,i}} \tag{15}$$

其中，C_m 代表L-UAV处理1 bit任务所需的CPU周期，$f_{m,i}$ 为L-UAV分配给用户设备 i 的计算频率。则L-UAV在时隙 t 内计算任务产生的能耗为

$$E_i^{\mathrm{exe}}=p_m^{\mathrm{exe}}t_i^{\mathrm{exe}}=\delta f_m^2\lambda_iD_i \tag{16}$$

其中，p_m^{exe} 为计算功率，$\delta=10^{-27}$ 为影响因子[18]。

4) L-UAV向E-UAV发射干扰信号时产生的干扰能耗

L-UAV需要在接收用户设备的任务数据过程中发送干扰信号，以避免E-UAV窃听用户信息，因此L-UAV在这个阶段中产生的干扰信号为

$$E_i^{\mathrm{eva}}=p_m^{\mathrm{eva}}t_i^{\mathrm{tra}}=\frac{p_m^{\mathrm{exe}}\lambda_iD_i}{BR_{m,i}} \tag{17}$$

根据以上能耗分析，L-UAV在卸载过程中产生的能耗总和为

$$E_{\mathrm{sum}}=\sum_{m=1}^{M}\left\{E_m^{\mathrm{fly}}+E_i^{\mathrm{tra}}+E_i^{\mathrm{exe}}+E_i^{\mathrm{eva}}\right\} \tag{18}$$

当用户设备选择将另一部分计算任务在本地执行时，在本地计算所产生的时延为

$$t_i^{\mathrm{loc}}=\frac{\left(1-\lambda_i\right)D_iC_i}{f_i} \tag{19}$$

其中，$\left(1-\lambda_i\right)D_i$ 表示数据在本地计算的大小，C_i

代表有用户设备处理1 bit任务所需的CPU周期，f_i代表用户设备i的计算频率。

1.4 问题建模

本文的优化目标是在保证L-UAV在规定周期内完成卸载的同时，使UAV辅助的MEC安全通信系统达到安全传输效率最大化，因此系统在保障数据安全传输的同时，还需高效利用资源，以实现长期的安全计算效率。安全计算效率[19]是衡量单位能耗下系统能够实现的安全数据传输能力，可以表示为

$$U_{\text{off}}=\sum_{m=1}^{M}\sum_{i=1}^{R}\sum_{t=0}^{T}\frac{R_{\text{sec},i}[t]}{E_m^{\text{fly}}(t)+E_i^{\text{tra}}+E_i^{\text{exe}}+E_i^{\text{eva}}} \tag{20}$$

本文假设UAV只能在给定范围内进行移动、L-UAV需要在整个周期结束前完成所有计算任务且UAV在卸载过程中产生的能耗不得超过电池总能量。这些假设为求解优化问题提供了明确的约束条件。在此基础上，结合深度强化学习方法，可以优化卸载过程中的安全传输效率，从而优化问题可以表示为

$$\begin{aligned}
\text{P:}\ &\max_{\{\lambda_i,p(t),f\}} U_{\text{off}}=\\
&\max_{\{\lambda_i,p(t),f\}}\sum_{m=1}^{M}\sum_{i=1}^{R}\sum_{t=0}^{T}\frac{R_{\text{sec},i}[t]}{E_m^{\text{fly}}(t)+E_i^{\text{tra}}+E_i^{\text{exe}}+E_i^{\text{eva}}}\\
\text{s.t.}\ &\text{C1: } 0\leqslant\lambda_i\leqslant 1,\forall i\\
&\text{C2: } f_i\leqslant F_i^{\max},\forall i\\
&\text{C3: } \sum_{i=1}^{R} f_{mi}\leqslant F_{mi}^{\max}\\
&\text{C4: } E_{\text{sum}}\leqslant E\\
&\text{C5: } \sum_{i=1}^{R}\max\left\{t_i^{\text{tra}}+t_i^{\text{exe}},t_i^{\text{loc}}\right\}\leqslant T\\
&\text{C6: } x(t)\in[0,X],y(t)\in[0,Y]
\end{aligned} \tag{21}$$

约束条件C1表示任务卸载到L-UAV的比例，约束条件C2和C3分别代表用户设备i和L-UAV的功率限制，约束条件C4表示L-UAV的能量不能超过电池总能量，约束条件C5表示L-UAV必须在周期T结束前完成计算任务，约束条件C6表示无人机只能在给定的区域内进行移动。

上述问题中的安全传输效率函数涉及安全传输率与总能耗，这与计算任务所需的任务数据大小和UAV电量有关。在优化问题中，任务卸载比和功率分配是主要的求解目标，以达到最大化安全传输效率。

2 基于深度强化学习的安全计算效率优化

为解决安全计算效率的优化问题，本文采用基于多智能体深度强化学习方法来最大化安全计算效率U_{off}。首先，本文给出了基于上述优化问题式(21)转化为马尔可夫博弈的模型，并给出了模型中的相关定义，然后提出了加入注意力机制的多智能体深度确定性策略梯度算法求解优化问题。

2.1 马尔可夫博弈建立

在本文模型中，UAV的状态会随时隙变化而更新，且下一个时隙的状态与当前时隙的状态相关，因此上述公式化列出的问题可以转换为马尔可夫决策过程。而多架UAV协同计算优化卸载任务，每架L-UAV被视为一个智能体，在多智能体环境中，每个智能体的决策受到所有智能体联合行动的影响。将马尔可夫决策过程视为多智能体扩展，用马尔可夫博弈[20]对多智能体强化学习进行建模。M个智能体的马尔可夫博弈可以用元组$(\mathcal{S},\mathcal{A},\mathcal{R},\gamma)$表示，元素分别为状态空间、动作空间、奖励空间和奖励折扣因子。在时隙t，智能体i观察到状态$s_i(t)$，并根据策略采取动作$a_i(t)$，从而产生一个新的状态$s_i(t+1)$。在上述模型中，无人机在飞行过程中不能直接相互通信，每架L-UAV只能获得包含自身信息的自身状态，由于环境是被完全观察到的，因此观察结果与状态是等价的。并能（L-UAV）都有自己的Actor-Critic网络，并能观察自己的状态，根据自己的策略采取行动，然后从环境中获得奖励，进入下一个状态，具体的状态空间、动作空间以及奖励值的定义如下。

1) 状态空间。在时隙t的任意一个智能体状态被定义为5个相对应的时变变量，其执行的状态由R个用户设备的位置信息$I(t)=\{I_1(t),I_2(t),\cdots,I_R(t)\}$、用户设备$i$生成的任务大小$D(t)=\{D_1(t),D_2(t),\cdots,D_R(t)\}$、L-UAV的位置$p_m(t)$、E-UAV的位置信息$p_n(t)$和UAV剩余的电量$E_r(t)$以及环境共同确定。因此，$t$时刻状态空间可以表示为

$$s(t)=\{I(t),p_m(t),p_n(t),D(t),E_r(t)\} \tag{22}$$

2) 动作空间。动作空间被考虑为L-UAV负责与环境交互并做出决策。其执行的动作包括任务卸载比$\lambda_i(t)$、L-UAV分配给用户设备i的计算频率$f_{m,i}(t)$、干扰功率$p_{mn}^{\text{eva}}(t)$、飞行角度$\theta(t)$和飞行速

度$v(t)$。飞行角度和飞行速度决定了任务卸载的执行效率和实时性，进而影响任务执行与环境交互的效果。因此，动作空间表示为

$$a(t)=\left\{\lambda_i(t),f_{m,i}(t),p_{mn}^{\mathrm{eva}}(t),v(t),\theta(t)\right\} \tag{23}$$

3) 奖励值。本文中奖励值是最大化安全计算效率，U_{off}代表L-UAV的整体奖励回报，使用$-R_t$表示条件限制惩罚。具体来说，当智能体执行的动作超出约束范围时，就会获得一个无穷大的负奖励。为了在尽可能小的发射和干扰功率下实现良好的安全计算效率，使用$-R_g$表示功率惩罚，则L-UAV的整体奖励回报为

$$R=U_{\mathrm{off}}-R_t-R_g \tag{24}$$

2.2　改进的MADDPG算法

考虑到多智能体环境中存在大量连续时变变量以及智能体间协作不稳定的情况，为了解决MADRL中的非平稳性问题和连续性动作，本文使用了MADDPG算法，该算法基于全局信息为每个智能体学习集中的Q函数。利用MADDPG算法处理全局信息，无人机可以基于目前的网络安全状态，动态调整资源分配和通信策略，应对网络安全威胁。如果被窃听者窃听，无人机群体可以通过改变飞行轨迹或发射协同电子干扰等集体反应策略来应对。本文所提出的A-MADDPG算法框架如图2所示。

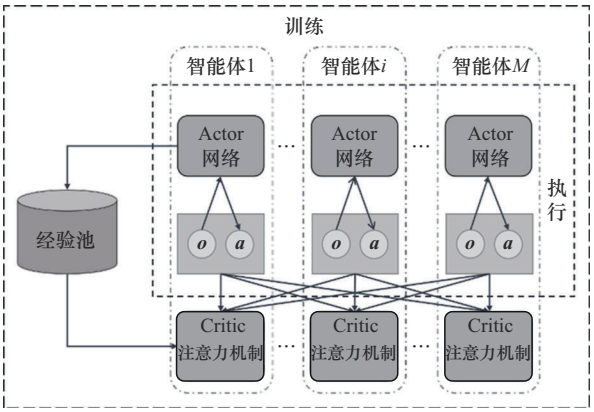

图2　A-MADDPG算法框架

每个智能体都有自己的Actor-Critic网络，其在每个时隙通过观察得到自身的状态，并根据自身的策略采取行动，然后从环境中获得奖励，进入下一个状态。MADDPG算法框架包含4个神经网络，Actor网络模拟策略函数$\mu\left(\boldsymbol{s}\middle|\theta^{\mu}\right)$，输入状态$\boldsymbol{s}$来确定动作$\boldsymbol{a}$。Critic网络模拟Q函数$Q\left(\boldsymbol{s},\boldsymbol{a}\middle|\theta^{Q}\right)$，用于评估Actor网络输出的动作。所谓Q函数，指的是某个状态$\boldsymbol{s}$下的某个动作$\boldsymbol{a}$对多个时隙后的多个动作产生的累积奖励期望，即

$$Q^{\mu}\left(s(t),\mu(t)\right)=\mathrm{E}\left[R\left(s(t),a(t)\right)+\gamma Q^{\mu}\left(s(t+1),\mu(t+1)\right)\right] \tag{25}$$

其中，γ是折扣因子，表示当前状态对未来状态作出的折扣贡献；当共有L个智能体时，令$\mu_\theta=\left\{\mu_{\theta_1},\cdots,\mu_{\theta_L}\right\}$代表策略，$\mu^*$代表最佳策略，$Q^*(\boldsymbol{s},\boldsymbol{a})=\max_{\mu^*}Q^{\mu}(\boldsymbol{s},\mu)$表示智能体选择最优策略$\mu^*=\left\{\mu_1^*,\mu_2^*,\cdots,\mu_N^*\right\}$时对应的累积奖励，$\theta=\left\{\theta_1,\cdots,\theta_L\right\}$是所有智能体的确定性策略的集合。每个智能体的策略梯度更新式为

$$\nabla_{\theta_i}J(\mu_{\theta_i})\approx\mathrm{E}_{\boldsymbol{o},\boldsymbol{a}}[\nabla_{\theta_i}\mu_{\theta_i}(a_i|s_i)\cdot\nabla_{a_i}Q_i^{\mu}(\boldsymbol{o},a_1,a_2,\cdots,a_L)|_{a_i=\mu_{\theta_i}(s_i)}] \tag{26}$$

其中，所有智能体的动作值函数输入为$\boldsymbol{a}=(a_1,\cdots,a_L)$以及观测信息$\boldsymbol{o}$，通常$\boldsymbol{o}$包括所有智能体的状态以及一些附加信息。经验回放池$D$是容量为$C$的有限经验池，用于存储所有智能体的经验。当经验回放池未满时，引入噪声N使每个智能体在选择动作时增加随机性，动作$a_i=\mu_{\theta_i}(o_{\mathrm{new}}^i)+N$。这种噪声通常使用OU噪声，不仅具有随机性，还具有一定的相关性，这对于连续动作空间中的探索来说是很有效的。当经验回放池容量已满时，最旧的样本将被丢弃，所有的智能体根据策略执行操作，并通过最小化损失函数来更新价值网络

$$L(\theta_i^Q)=E_{\boldsymbol{o},\boldsymbol{o}',\boldsymbol{a},r}[(Q_i^{\mu}(\boldsymbol{o},a_1,\cdots,a_L)-y)^2] \tag{27}$$

其中，

$$y=r_i+\varphi Q_i^{\mu'}(\boldsymbol{o}',a_1',\cdots,a_L')|_{a_i'=\mu'(o_i')} \tag{28}$$

在上述模型中，不同的L-UAV需要关注不同的E-UAV和地面用户的位置，这不可避免地会增加探索时长，而注意力机制可以减少探索空间，从而提高学习效率。因此，本文提出了一种带有注意机制的集中式训练Critic网络算法Attention-MADDPG，用来针对连续性动作。本文所提出的A-MADDPG算法的流程可以总结如下。

首先，将注意力网络添加到Critic网络执行之前，注意机制功能通过相似性函数$f\left(Q,K_i\right)$来计算给定的查询和每个键（Key）的相似性。$f\left(Q,K_i\right)$常见形式由点积$\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}_i$、2个向量的余弦相似度$\left(\boldsymbol{Q}\cdot\boldsymbol{K}_i\right)\cdot\left(\left\|\boldsymbol{Q}\right\|\cdot\left\|\boldsymbol{K}_i\right\|\right)^{-1}$和多层感知器（MLP, multi-layer perceptron）网络构成，并通过softmax函数得到归一化的注意力权值$\alpha_i=\mathrm{softmax}\left(f\left(\boldsymbol{Q},\boldsymbol{K}_i\right)\right)$。使用导出的注意力权重进行加权求和，表示为

$$\mathrm{Attention}\left(\boldsymbol{Q},\boldsymbol{K}\right)=\sum_i\alpha_i\mathrm{Value}_i \tag{29}$$

随后，将集中的Q值函数重写为

$$Q_i^{\mu}\left(\boldsymbol{o},\boldsymbol{a}\right)=H_i\left(J_i\left(s_i,a_i\right),k_i\right) \tag{30}$$

其中，$H_i(\cdot)$为第i个智能体的四层MLP，$J_i(\cdot)$为第i个智能体的单层MLP，k_i表示其他智能体的贡献。

最后，采用新的Q值函数来指导更新Actor网络的参数，观察环境反馈，逐步学习最优策略，从而在给定状态下选择的动作可以获得更高的长期累积奖励。

基于上述优化，A-MADDPG算法伪代码如算法1所示。

算法1 A-MADDPG算法伪代码

输入 Actor、Critic和注意力网络结构，批大小B，训练批次E，训练步长T，动作噪声N以及智能体数量M

输出 每个智能体Actor网络权值$\theta_i^{\mu'}$

初始化 价值网络Q和策略网络μ，使用随机参数θ^{μ}，θ^{Q}和θ^{A}；目标网络$\theta^{\mu'}\leftarrow\theta^{\mu}$，$\theta^{Q'}\leftarrow\theta^{Q}$，$\theta^{A'}\leftarrow\theta^{A}$；容量为$C$的经验回放池$D$；探索计数器Counter=0。

1) for each episode $e=1,2,\cdots,\mathcal{E}$ do
2) 获取初始观察值o_{init}，$o_{\mathrm{new}}\leftarrow o_{\mathrm{init}}$，同时将奖励值大小置于0；
3) for each step $t=1,2,\cdots\mathcal{T}$ do
4) if Counter<$\mathcal{C}$ then
5) 每个智能体i随机选择一个动作a_i；
6) Counter ← Counter + 1；
7) else
8) $a_i=\mu_{\theta_i}\left(o_{\mathrm{new}}^i\right)+N$；
9) end if
10) 执行动作$\boldsymbol{a}=[a_1,a_2,\cdots,a_M]$，并观察对应的奖励$R$和下一时隙的状态$o_{\mathrm{new}}$；
11) 将所有采集的样本$(\boldsymbol{o},\boldsymbol{a},R,o_{\mathrm{new}})$存储到经验回放池$D$中；
12) 从经验池D中随机抽取Z个批量样本$(o^e,a^e,R^e,o_{\mathrm{new}}^e)$训练策略和价值网络；
13) 令$y^e=r^e+\varphi\cdot Q_i^{\mu'}\left(o_{\mathrm{new}}^e,a_1',\cdots,a_M'\right)\big|_{a_i'=\mu_i'(o_i)}$；
14) 根据式(30)中的最小化损失函数来更新Critic和注意力网络；
15) 根据式(29)中的策略梯度来更新Actor网络；
16) 依次更新目标网络参数：$\theta_i^{\mu'}$，$\theta_i^{Q'}$，$\theta_i^{A'}$；
$\theta_u'\leftarrow\tau\theta_u+(1-\tau)\theta_u'$，$\theta_Q'\leftarrow\tau\theta_Q+(1-\tau)\theta_Q'$，$\theta_A'\leftarrow\tau\theta_A+(1-\tau)\theta_A'$
17) end for
18)end for

算法1的时间复杂度说明如下。第1)行是加载训练好的MADDPG模型中的网络参数，包括主策略网络和目标网络，在实际运行环境中仅执行一次。第2)行是获取环境的初始状态信息，这部分的时间复杂度为$O(1)$。第3)行到第14)行是智能体在每个时隙进行决策的部分。智能体通过神经网络（包含注意力机制）进行前向传播来获得当前时隙的动作决策。网络包含n_1个神经元的输入层、n_2个神经元的注意力层、n_3个隐藏层和n_4个神经元的输出层。单个样本的传播时间复杂度为$O(n_1n_2+n_1n_2+n_2n_3+n_3n_4)$。即算法的时间复杂度约为$O(n^2)$。第15)行和第16)行用于更新策略。更新涉及所有智能体的梯度计算，梯度计算和注意力机制的计算复杂度仍受网络规模影响，因此时间复杂度仍为$O(n^2)$。基于此，A-MADDPG算法的总时间复杂度为$O(n^2)$。

综上所述，物理层安全模型考虑安全传输策略时需要在多个参数之间进行动态调整，使目标优化问题呈现非凸性。为了提高系统的学习效率并减少空间探索的复杂性，本文在MADDPG中引入注意力机制，用于集中式训练Critic网络，该网络允许每个智能体在处理全局信息时更专注于与其利益相关的智能体。与直接将所有信息同时输入Critic网络的方法相比，本文方案可以更高效地提取和利用复杂多智能体系统中的关键信息，进而增强学习策略能力和决策效率，提高网络训练的精度和效果。

3　仿真分析

3.1　参数设置

本节将通过仿真实验对本文算法进行性能评估。为了公平性起见，本文算法和对比算法均在PyTorch 框架下进行验证和评估，并模拟了一个100 m×100 m的正方形区域，L-UAV、E-UAV以及用户设备UD分布在该范围内，并且所有L-UAV已经知晓用户设备和E-UAV的位置。考虑到实际情况，L-UAV的数量小于用户设备的数量，其中L-UAV的数量可以进行调整。本实验其他系统参数设置如表1所示。

表1　　系统参数

参数	数值
L-UAV飞行高度H_m/m	40
E-UAV飞行高度H_n/m	50
最大飞行速度v/(m·s^{-1})	5
UAV最大功率P_{max}/W	2.4
L-UAV计算频率f_m/GHz	2.5
episode	1 400
奖励折扣因子γ	0.95
经验回放池容量C	2 000
学习率α	0.01
批量大小k	64
探索策略θ	0.15
目标网络更新率τ	0.01
DNN优化器	Adam

3.2　实验结果和性能分析

图3为不同算法的奖励曲线。从图3可以看到，A-MADDPG的效果最好，与DDPG单智能体算法相比，A-MADDPG算法在性能上提升了83.3%左右，相较于单智能体，多智能体之间可以相互协作学习，共同提升性能；相较于MADDPG算法，A-MADDPG算法提升了约50%，因为添加了注意力机制的A-MADDPG算法能够高效处理信息并且增强智能体之间的协作能力，从而提高效率，由此证明了A-MADDPG算法的有效性。

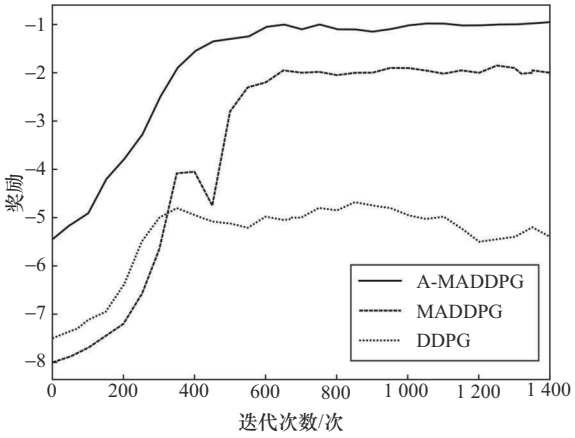

图3　不同算法的奖励曲线

图4为不同数量E-UAV下的安全计算效率。从图4可以看出，当E-UAV的数量增加时，安全计算效率越低，这是因为窃听无人机数量增多会导致更多的信号干扰，安全计算效率随之降低。设置L-UAV的数量为3架，当存在3架E-UAV时，安全计算效率约为2.0 Mbit/J，3架E-UAV与1架E-UAV的安全计算效率差值约为3.0 Mbit/J。这一结果表明，在E-UAV增多的情况下，系统的安全性受到更大的影响，因此需要采取更加强有力的措施来保护数据传输的安全性。

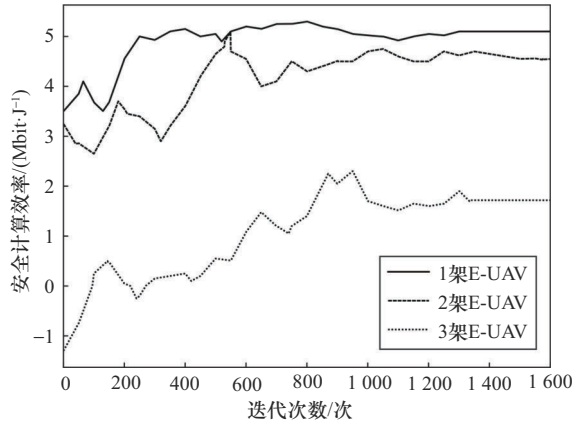

图4　不同数量E-UAV下的安全计算效率

图5为不同数量L-UAV的发射功率对比，此发射功率是1 600次迭代次数的平均值。从图5可以看出，2种算法下所有无人机的功率总和几乎相同，且A-MADDPG算法优于MADDPG算法，约为4.2%。

图6为不同C_m下用户任务数据大小与安全计算效率的关系。从图6中可知，当用户任务数据增加时，系统需要处理更多的计算任务，导致系统能

耗增加，无法保持高效的计算性能，进而降低了整体的安全计算效率。当C_m保持不变时，随着用户任务数据的增加，系统的安全计算效率随之降低；当用户任务数据保持不变时，C_m越大，安全计算效率越低。当C_m=800 cycle/bit时，处理2 Mbit的用户数据的安全计算效率为5.65 Mbit/J；当C_m=1 500 cycle/bit时，处理2 Mbit的用户数据的安全计算效率为5.23 Mbit/J。

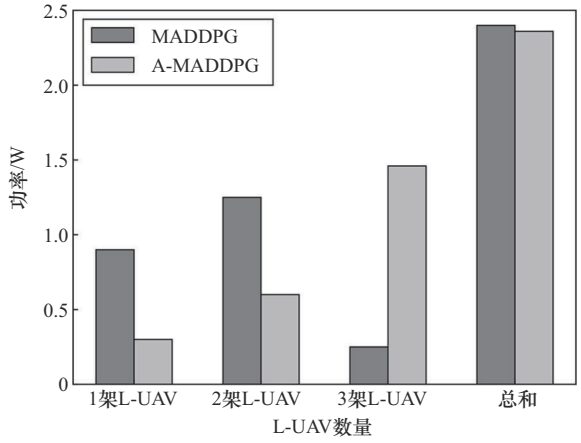

图5　不同数量L-UAV的发射功率对比

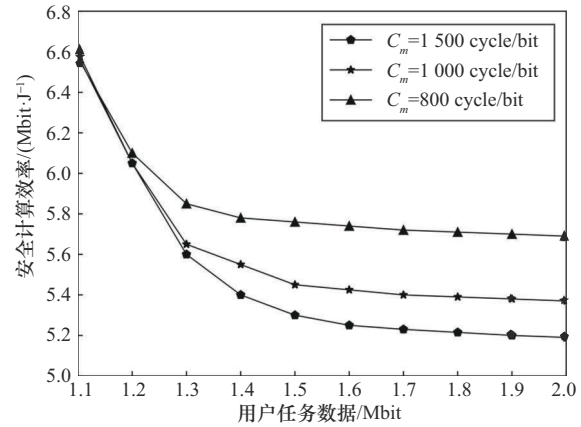

图6　不同C_m下用户任务数据大小与安全计算效率的关系

图7为不同算法下UAV最大功率与安全计算效率的关系。由图7可以看出，随着UAV最大功率的不断增加，安全计算效率也在增加，因为功率的持续性增加，UAV可以获得更多的能量，从而更好地支持复杂的计算任务，减轻任务卸载负担，同时无人机的干扰信号发射能力随之增强，从而提升通信链路质量，提高安全计算效率。

图8为不同C_m下用户任务数据大小与安全传输效率的关系。从图8中可知，当用户任务数据增加时，需要传输更多的计算任务，进而降低了整体的安全传输效率。当C_m保持不变时，随着用户任务数据的增加，系统的安全传输效率随之降低；当用户任务数据保持不变时，C_m越大，安全传输效率越低。

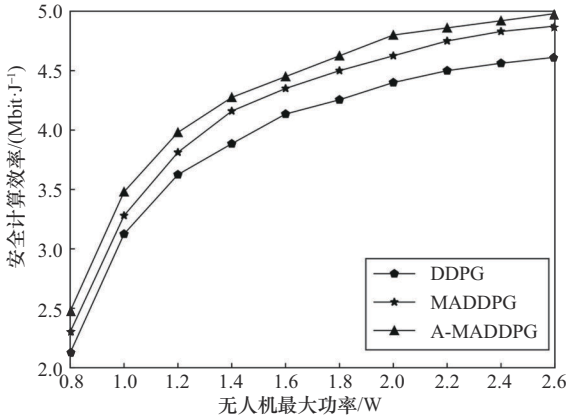

图7　不同算法下UAV最大功率与安全计算效率的关系

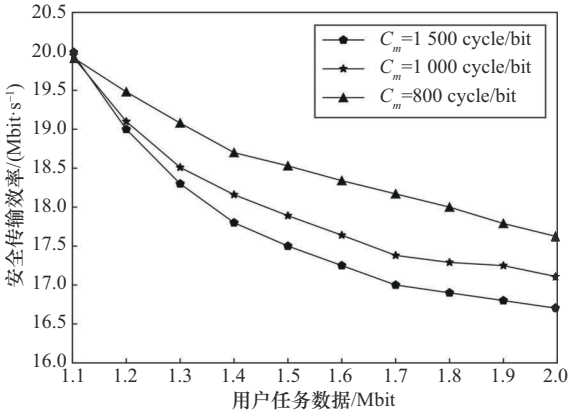

图8　不同C_m下用户任务数据大小与安全传输效率的关系

4　结束语

本文构建了一个存在任务数据被窃听情况下的无人机辅助MEC安全通信系统，在该系统下以最大化安全传输效率为目标，提出了在MADDPG算法的基础上加入注意力机制的A-MADDPG算法，提高了学习效率。通过采用A-MADDPG算法对L-UAV的任务卸载比例以及功率分配进行优化，证明了安全传输效率的提升，同时验证了本文提出的A-MADDPG算法相对于基准MADDPG算法与单智能体DDPG算法的有效性。

参考文献：

[1] 谢人超, 廉晓飞, 贾庆民, 等. 移动边缘计算卸载技术综述[J]. 通信学报, 2018, 39(11): 138-155.

XIE R C, LIAN X F, JIA Q M, et al. Survey on computation offloading in mobile edge computing[J]. Journal on Communications, 2018, 39 (11): 138-155.

[2] YANG L, YAO H P, WANG J J, et al. Multi-UAV-enabled load-balance mobile-edge computing for IoT networks[J]. IEEE Internet of Things Journal, 2020, 7(8): 6898-6908.

[3] UPADHYA A. On the reliability of interference limited unmanned aerial vehicles[J]. Wireless Personal Communications, 2023, 129(1): 119-131.

[4] GU X H, ZHANG G A, WANG M X, et al. UAV-aided energy-efficient edge computing networks: security offloading optimization[J]. IEEE Internet of Things Journal, 2022, 9(6): 4245-4258.

[5] 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.

[6] 余雪勇, 邱礼翔, 宋家宁, 等. 无人机辅助边缘计算中安全通信与能效优化策略[J]. 通信学报, 2023, 44(3): 45-54.
YU X Y, QIU L X, SONG J N, et al. Security communication and energy efficiency optimization strategy in UAV-aided edge computing[J]. Journal on Communications, 2023, 44(3): 45-54.

[7] XU Y, ZHANG T K, YANG D C, et al. Joint resource and trajectory optimization for security in UAV-assisted MEC systems[J]. IEEE Transactions on Communications, 2021, 69(1): 573-588.

[8] FATIMA N, SAXENA P, GIAMBENE G. Deep reinforcement learning based computation offloading for xURLLC services with UAV-assisted IoT-based multi-access edge computing system[J]. Wireless Networks, 2024, 30(9): 7275-7291.

[9] TAN L, KUANG Z F, GAO J, et al. Energy-efficient collaborative multi-access edge computing via deep reinforcement learning[J]. IEEE Transactions on Industrial Informatics, 2023, 19(6): 7689-7699.

[10] LIU X, CHAI Z Y, LI Y L, et al. Multi-objective deep reinforcement learning for computation offloading in UAV-assisted multi-access edge computing[J]. Information Sciences, 2023, 642: 119154.

[11] 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.

[12] SEID A M, LU J F, ABISHU H N, et al. Blockchain-enabled task offloading with energy harvesting in multi-UAV-assisted IoT networks: a multi-agent DRL approach[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(12): 3517-3532.

[13] DUAN W W, LI X M, HUANG Y, et al. Multi-agent-deep-reinforcement-learning-enabled offloading scheme for energy minimization in vehicle-to-everything communication systems[J]. Electronics, 2024, 13(3): 663.

[14] WANG L, WANG K Z, PAN C H, et al. Multi-agent deep reinforcement learning-based trajectory planning for multi-UAV assisted mobile edge computing[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(1): 73-84.

[15] YOO S, JEONG S, KANG J. Hybrid UAV-enabled secure offloading via deep reinforcement learning[J]. IEEE Wireless Communications Letters, 2023, 12(6): 972-976.

[16] LI H, MENG S M, SHANG J, et al. Value-based multi-agent deep reinforcement learning for collaborative computation offloading in Internet of things networks[J]. Wireless Networks, 2024, 30(8): 6915-6928.

[17] CHEN P P, LUO X S, GUO D K, et al. Secure task offloading for MEC-aided-UAV system[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(5): 3444-3457.

[18] CHEN P P, LUO L L, GUO D K, et al. Secure task offloading for rural area surveillance based on UAV-UGV collaborations[J]. IEEE Transactions on Vehicular Technology, 2024, 73(1): 923-937.

[19] WANG Y P, FANG W W, DING Y, et al. Computation offloading optimization for UAV-assisted mobile edge computing: a deep deterministic policy gradient approach[J]. Wireless Networks, 2021, 27(4): 2991-3006.

[20] LIN W S, MA H, LI L X, et al. Computing assistance from the sky: decentralized computation efficiency optimization for air-ground integrated MEC networks[J]. IEEE Wireless Communications Letters, 2022, 11(11): 2420-2424.

[作者简介]

王义君（1984– ），男，内蒙古通辽人，博士，长春理工大学副教授，主要研究方向为5G/6G移动通信、物联网及通感一体自主无人通信系统等。

李嘉欣（2000– ），女，吉林四平人，长春理工大学硕士生，主要研究方向为移动边缘计算安全传输技术。

闫志颖（1997– ），女，吉林白城人，长春理工大学硕士生，主要研究方向为移动边缘计算任务卸载策略。

吕婧莹（1999– ），女，吉林松原人，长春理工大学硕士生，主要研究方向为无线通信与数据传输。

钱志鸿（1957– ），男，吉林长春人，博士，吉林大学教授，主要研究方向为基于物联网、D2D、Wi-Fi、RFID等的无线网络与通信技术。