中国科学: 信息科学 2024 年 第 54 卷 第 6 期: 1474–1486
SCIENTIA SINICA Informationis
论文

# 基于 MATD3 的空地网络资源优化

秦鹏[1*], 王硕[1,2], 付民[1], 赵雄文[1]

1. 华北电力大学电气与电子工程学院, 北京 102206
2. 中国电信股份有限公司河北分公司, 石家庄 050036
* 通信作者. E-mail: qinpeng@ncepu.edu.cn

收稿日期: 2023–07–20; 修回日期: 2023–10–31; 接受日期: 2023–12–13; 网络出版日期: 2024–06–11

国家自然科学基金 (批准号: 62201212)、河北省自然科学基金 (批准号: F2022502017) 和中央高校基本科研业务费专项资金 (批准号: 2023JC003) 资助项目

**摘要** 移动边缘计算通过将计算任务卸载到无线网络边缘, 可有效减少任务延迟与终端能耗. 对于偏远地区分布的大量物联设备 (如风电、光伏等电力物联终端), 现有地面网络无法为其提供有效的网络服务. 因此, 本文重点研究空地一体化异构网络模型, 通过联合设计无人机轨迹、任务卸载与计算资源分配, 以最大限度地减少物联设备任务执行延迟与能耗. 针对目标函数的非凸性和网络动态造成的信息不确定性, 本文将问题建模为马尔可夫 (Markov) 决策过程, 并提出一种基于 MATD3 的 UAV 轨迹与网络资源协同优化算法. 实验结果表明, 与基准算法相比, 本文提出的方案在系统计算能耗和时延方面性能更优.

**关键词** 空地一体化异构网络, 卸载决策, 资源分配, UAV 轨迹优化, 多智能体深度强化学习

## 1 引言

随着移动通信技术以及物联网技术的迅速发展, 终端数量呈爆炸式增长. 与此同时, 各类新兴业务应用, 如虚拟现实 (virtual reality, VR)、增强现实 (augmented reality, AR) [1] 等不断涌现, 这对网络计算与服务性能提出了更高的要求 [2]. 移动边缘计算 (mobile edge computing, MEC) 作为一种新的范式, 通过在靠近用户的网络边缘侧部署计算与存储资源, 将终端任务卸载到边缘节点执行, 大大降低了任务时延并缓解了网络拥塞 [3~5]. 然而, 现有的边缘计算网络仍以地面为主, 对于偏远地区分布的大量物联设备 (如风电、光伏等电力物联终端), 地面设施无法为其提供有效的网络服务. 与地面网络不同, 空地网络可充分利用无人机 (unmanned aerial vehicle, UAV) 部署的灵活性与低成本实现广域无缝覆盖, 从而成为地面网络的重要补充. 近年来国内外涌现了大量关于空地异构网络的研究 [6~8], 包括资源管理 [9~11]、计算卸载 [12~14] 等多个方面. 如文献 [15] 提出了一种基于随机模拟的任务卸载和

**引用格式:** 秦鹏, 王硕, 付民, 等. 基于 MATD3 的空地网络资源优化. 中国科学: 信息科学, 2024, 54: 1474–1486, doi: 10.1360/SSI-2023-0223
Qin P, Wang S, Fu M, et al. Air-ground integrated network resource optimization based on MATD3 (in Chinese). Sci Sin Inform, 2024, 54: 1474–1486, doi: 10.1360/SSI-2023-0223

 

资源分配两阶段随机规划算法, 以最小化系统能耗. 但算法仅考虑了单个边缘服务器的情况, 限制了其在更复杂和规模较大的边缘计算网络中的适用性. 文献 [16] 研究了 UAV 网络中任务卸载和功率控制的联合优化问题, 在满足无人机能量约束的同时最小化平均服务时间. 然而, 该工作研究的是静态无人机网络系统. 实际上, 由于无人机需要飞往特定区域, 不同的飞行轨迹将提供差异化的通信质量, 从而导致不同的任务延迟与终端能耗. 文献 [17] 通过联合优化任务调度和多无人机部署以最小化系统任务时延和能耗的加权和. 文献 [18] 提出了一种获得最大系统效能的任务分配算法以及同时攻击目标的航迹规划算法. 然而, 在实际执行中还需考虑无人机上计算资源分配的问题. 因此, 有必要联合考虑 UAV 轨迹、任务卸载和计算资源管理的系统性问题. 此外, 无人机的动态性将引起网络状态的时变性与信道条件的不确定性, 使得传统通过频繁的信息交互获取网络全局信息 (global state information, GSI) 的方法不再可行. 由此带来高维空间问题, 需要寻求高效的优化方法以解决这一问题, 同时确保维持系统的性能和稳定性. 同时, 由于 UAV 能量、计算资源仍然有限, 当某一时段物联网设备产生任务量巨大而又对时延有一定要求时, 本地处理和卸载到 UAV 处理可能都不能满足. 因此需要在没有基站信号覆盖的偏远地区引入具有更大存储空间、覆盖面更广、计算能力更强的 HAPS (high altitude platform station), 以满足业务需求. UAV, HAPS 和终端设备之间的决策需要协同工作, 以实现任务卸载、计算资源分配和轨迹规划的一致性. 这需要设计协同决策算法, 以确保系统的协同工作能够高效实现. 幸运的是, 强化学习 (reinforcement learning, RL) 有望成为在信息不确定条件下解决上述问题的有效途径. 然而, RL 存在不能有效处理大状态空间和动作空间信息的问题. 深度强化学习 (deep reinforcement learning, DRL) 作为一种在复杂动态网络环境下优化决策的方法, 通过集成深度神经网络 (deep neural network, DNN) 的强大学习能力和 RL 的决策能力, 可通过与高维动态环境交互来获得最优决策. 因此, 为满足偏远地区设备的时延和能耗需求, 本文提出一种空地一体化异构网络, 其中无人机和 HAPS 能够协同处理物联设备卸载的计算任务. 通过联合优化无人机轨迹、任务卸载与计算资源分配, 以最大限度地减少物联设备任务执行延迟与能耗. 针对目标函数的非凸性和网络动态造成的信息不确定性 [16], 将问题建模为马尔可夫 (Markov) 决策过程, 并提出一种基于 MATD3 (multi agent twin delayed deep deterministic policy gradient) 的 UAV 轨迹与网络资源协同优化方法. 本文主要贡献如下:

(1) 首先, 设计了一个空地异构网络模型为偏远地区物联设备提供服务. 其中, UAV 与高空平台 (HAPS) 作为空基边缘服务器, 终端任务可以在本地执行, 也可以卸载到服务器端. 这种灵活性为偏远地区终端提供了更多的计算和通信资源, 同时优化了任务执行的方式. 然后, 将问题数学建模为无人机轨迹、任务卸载和计算资源分配的联合优化问题, 目的是最小化系统长期运行成本 (即任务时延与终端能耗之和).

(2) 由于上述问题涉及连续变量 (无人机轨迹、计算资源分配) 与二值变量 (任务卸载决策), 因此是一个混合整数非线性规划 (mixed integer nonlinear programming, MINLP) 问题. 同时, 考虑到网络环境动态造成的信息不确定性与维度空间爆炸难题, 使得求解十分困难. 因此, 将其建模为马尔可夫决策过程, 并设计提出基于多智能体深度强化学习的 MATD3 算法.

(3) 通过大量仿真实验与 4 种基线算法进行了对比. 结果表明, 本文所提算法在终端能耗与时延方面系统性能更优, 能够有效地降低系统运行成本, 减少终端能耗并优化任务执行的时延.

## 2 系统模型

如图 1 所示, 本文考虑由 1 个 HAPS 以及 $M$ 架 UAV 组成的空地一体化异构网络模型, 为偏远

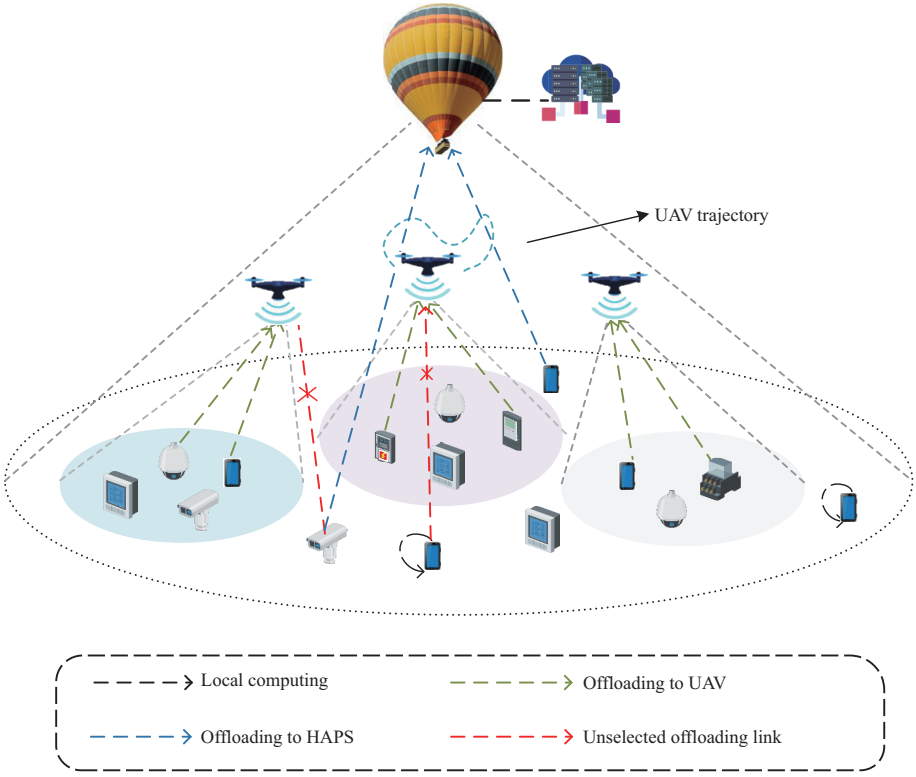


**图 1** (网络版彩图) 系统模型

**Figure 1** (Color online) System model

地区物联设备提供服务. 其中, 配备边缘服务器的 UAV 作为空中小基站只能提供有限的计算和通信资源, 其动态性导致在不同时间段和地点提供通信质量具备差异化, 因而需要优化飞行轨迹; 具备更强计算能力的 HAPS 作为空中宏基站以悬停模式部署在服务区域内, 用以确保广域无缝覆盖. 空基服务器集合表示为 $\mathcal{S}=\{S_0,S_1,\ldots,S_m\}$, 其中 $S_0$ 代表 HAPS. $K$ 个物联终端随机分布, 负责定期收集环境数据, 并生成计算密集型 (如电力设备视频监控类业务) 或时延敏感型任务 (如电力控制类业务), 相应的集合表示为 $\mathit{\Gamma}=\{U_1,U_2,\ldots,U_k,\ldots,U_K\}$. 本文采用时隙模型, 时间总长度包含 $T$ 个时隙, 集合表示为 $\mathcal{T}=\{1,2,\ldots,t,\ldots,T\}$. 物联设备生成的任务可以在本地执行, 也可以卸载到服务器端, 根据任务的性质和通信要求来灵活选择执行位置, 确保资源的有效利用.

定义地面终端的二维坐标为 $H_k(t)=[x_k(t),y_k(t)]$, 并用 $M_m(t)=[x_m(t),y_m(t),z_m(t)]$ 表示 UAV 的三维坐标, 其中, $x_m(t),y_m(t),z_m(t)$ 分别是 UAV 在 $t$ 时隙的 $X,Y,Z$ 的坐标. 令 $N_m(t)=[x_m(t),y_m(t)]$ 为 UAV 的二维坐标. 单个时隙内, UAV 在水平方向以角度 $\vartheta_m(t)$ 飞行距离 $l_m(t)$, 则其坐标更新公式为

$$\begin{aligned} x_m(t+1)&=x_m(t)+l_m(t)\cos(\vartheta_m(t)),\\ y_m(t+1)&=y_m(t)+l_m(t)\sin(\vartheta_m(t)). \end{aligned} \tag{1}$$

假定 UAV 具有最大俯仰角 $\varphi_m$, 则在 $t$ 时隙 UAV 的最大覆盖半径 $C_m(t)$ 为

$$C_m(t)=z_m(t)\tan(\varphi_m). \tag{2}$$

由于 UAV 水平飞行和垂直飞行速度有限, 因此, 其飞行距离受到以下约束:

$$\begin{aligned} &Z_{\min} \leqslant z_m(t) \leqslant Z_{\max}, \\ &\|N_m(t+1) - N_m(t)\| \leqslant L_{\max}^h, \\ &|z_m(t+1) - z_m(t)| \leqslant L_{\max}^v, \end{aligned} \tag{3}$$

其中, $Z_{\min}$ 和 $Z_{\max}$ 分别表示最小高度和最大高度; $L_{\max}^h$ 和 $L_{\max}^v$ 分别是 UAV 的最大水平飞行距离和垂直飞行距离. 此外, 为保证 UAV 在服务的矩形区域内移动, 须满足以下移动性约束:

$$\begin{aligned} &0 \leqslant x_k(t) \leqslant X_{\max}, \\ &0 \leqslant y_k(t) \leqslant Y_{\max}. \end{aligned} \tag{4}$$

为了避免任何两个 UAV 之间的碰撞, UAV 的距离应不小于最小距离 $D_{\min}$. 因此, 碰撞约束如下所示:

$$\|N_{m_1}(t) - N_{m_2}(t)\| \geqslant D_{\min}. \tag{5}$$

### 2.1 通信模型

对于 UAV/HAPS– 地通信, 我们重点关注任务卸载决策, 其时间尺度相比传统的资源调度时间要长得多 (通常为 1 ms). 一旦做出了任务卸载的决策, 该决策会持续到下一个操作时间点, 因此主要考虑大尺度衰落 [19]. 在任务卸载过程中, 终端设备 $U_k$ 与边缘服务器 $S_m$ 地对空通信链路的路径损耗定义为 [2]

$$L_{k,m,t} = 20\log_{10}\left(\frac{4\pi f_c\sqrt{d_{k,m,t}^2 + r_{k,m,t}^2}}{\mathrm{c}}\right) + P_{k,m,t}^{\mathrm{Los}}\eta_{k,m,t}^{\mathrm{Los}} + \left(1 - P_{k,m,t}^{\mathrm{Los}}\right)\eta_{k,m,t}^{\mathrm{NLos}}, \tag{6}$$

其中, $r_{k,m,t}$ 表示 $U_k$ 与 $S_m$ 之间的水平距离, $d_{k,m,t}$ 表示 $S_m$ 的飞行高度, $f_c$ 和 c 分别为载波频率以及光速, $\eta_{k,m,t}^{\mathrm{Los}}$ 和 $\eta_{k,m,t}^{\mathrm{NLos}}$ 表示视距链路以及非视距链路自由空间路径损耗之外产生的附加损耗. $U_k$ 与 $S_m$ 之间的视距通信概率 [20] 为

$$P_{k,m,t}^{\mathrm{Los}} = \frac{1}{1 + b_1\exp\{-b_2[\arctan(\frac{d_{k,m,t}}{r_{k,m,t}}) - b_1]\}}, \tag{7}$$

其中, $b_1, b_2$ 是由环境信息决定的变量. 在第 $t$ 时隙中, $U_k$ 与 $S_m$ 之间的通信链路的传输速率可以表示为

$$R_{k,m}(t) = B_{k,m}\log_2\left(1 + \frac{P^{\mathrm{TX}}10^{-\frac{L_{k,m,t}}{10}}}{\sigma^2}\right), \tag{8}$$

其中, $B_{k,m}$ 表示带宽, 这里每个信道为正交信道, 因而不考虑用户间干扰 [9]. $P^{\mathrm{TX}}$ 表示终端的发射功率, $\sigma^2$ 表示加性高斯 (Gauss) 白噪声功率.

### 2.2 计算模型

本小节介绍本地计算和边缘计算的计算模型. 三元组向量 $Z_k(t) = (d_k(t), c_k(t), \phi_k(t))$ 表示用户的计算任务, 其中 $d_k(t)$ 表示任务的数据量大小, $c_k(t)$ 表示处理任务所需要的 CPU 周期数, $\phi_k(t)$ 代表处理任务最大容忍时间. 当在第 $t$ 个时隙, 任务到达特定用户时, 系统必须确定任务是卸载到 MEC 服务器还是在本地计算. 定义 $a_{k,m}(t) \in \{0,1\}$ 为物联设备的卸载决策参数, $a_{k,m}(t) = 1$ 代表任务卸载到边缘服务器端执行, 否则 $a_{k,m}(t) = 0$.

(1) 本地计算模型: 物联设备本地计算的执行延迟和能耗可以表示为

$$\begin{aligned} T_k^L(t) &= \frac{c_k(t)}{f_k^L(t)}, \\ E_k^L(t) &= \mathcal{K}\big(f_k^L(t)\big)^2 c_k(t), \end{aligned} \tag{9}$$

其中, $f_k^L(t)$ 表示本地计算能力 (即分配的 CPU 资源), $\mathcal{K}$ 表示计算能力效率系数, 取决于芯片结构.

(2) 边缘计算模型: 对于边缘计算模型, $U_k$ 将其计算任务上传到 HAPS 或者 UAV 执行. 则任务传输时间和边缘计算时间分别通过下述公式计算:

$$\begin{aligned} T_{k,m}^{\text{trans}}(t) &= \frac{d_k(t)}{R_{k,m}(t)}, \\ T_{k,m}^{\text{exe}}(t) &= \frac{c_k(t)}{f_{k,m}^E(t)}, \end{aligned} \tag{10}$$

其中, $f_{k,m}^E(t)$ 表示 HAPS 或者 UAV 分配给 $U_k$ 的计算资源. 由于经 MEC 服务器处理后, 任务的数据量大小远小于任务处理之前的数据量, 并且下载速率远高于上传速率, 因此, 本文不考虑从 MEC 服务器返回到终端设备的时间. 则 $U_k$ 通过空基边缘计算的任务总时延为

$$T_{k,m}^E(t) = T_{k,m}^{\text{trans}}(t) + T_{k,m}^{\text{exe}}(t). \tag{11}$$

类似地, $U_k$ 卸载到 MEC 服务器的能耗可通过下式计算:

$$E_{k,m}^E(t) = T_{k,m}^{\text{trans}}(t) P^{\text{TX}}. \tag{12}$$

因此, $U_k$ 处理任务的总能耗表示为

$$E_k(t) = \left(1 - \sum_{m=0}^{M} a_{k,m}(t)\right) E_{k,m}^L(t) + \sum_{m=0}^{M} a_{k,m}(t) E_{k,m}^E(t). \tag{13}$$

$U_k$ 处理任务的总时延表示为

$$T_k(t) = \left(1 - \sum_{m=0}^{M} a_{k,m}(t)\right) T_k^L(t) + \sum_{m=0}^{M} a_{k,m}(t) T_{k,m}^E(t). \tag{14}$$

## 3 问题建模

本节通过联合优化 UAV 飞行轨迹、空地网络边缘服务器计算资源分配以及终端任务卸载决策, 提出任务执行时延与设备能耗加权成本最小化问题. 具体如下所示:

$$\min_{M,A,F} \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} [\omega_1 E_k(t) + \omega_2 T_k(t)] \tag{15}$$

$$\text{s.t.} \quad C_1: \quad a_{k,m}(t) \in \{0,1\}, \forall k \in \mathit{\Gamma}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{16}$$

$$C_2: \quad \sum_{m=0}^{M} a_{k,m}(t) \leqslant 1, \forall k \in \mathit{\Gamma}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{17}$$

$$C_3: \quad \sum_{i=1}^{I} a_{k,m}(t) f_{k,m}^{E}(t) \leqslant f_k^{\max}(t), \forall k \in \mathit{\Gamma}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{18}$$

$$C_4: \quad f_{k,m}^{E}(t) \geqslant 0, \forall k \in \mathit{\Gamma}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{19}$$

$$C_5: \quad Z_{\min} \leqslant z_m(t) \leqslant Z_{\max}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{20}$$

$$C_6: \quad |z_m(t+1) - z_m(t)| \leqslant L_{\max}^{v}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{21}$$

$$C_7: \quad \|N_m(t+1) - N_m(t)\| \leqslant L_{\max}^{h}, \forall m \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{22}$$

$$C_8: \quad \|N_{m_1}(t) - N_{m_2}(t)\| \geqslant D_{\min}, \forall m_1, m_2 \in \mathcal{S}, \forall t \in \mathcal{T}, \tag{23}$$

其中, $M = \{\vartheta_m(t), l_m(t)\}_{m\in\mathcal{S},\forall t\in\mathcal{T}}$ 代表 UAV 轨迹规划, 包含每个时隙 UAV 的飞行方向角与距离; $A = \{a_{k,m}(t)\}_{k\in\mathit{\Gamma},m\in\mathcal{S},\forall t\in\mathcal{T}}$ 代表物联网设备的卸载决策; $F = \{f_{k,m}^{E}(t)\}_{k\in m,\forall t\in\mathcal{T}}$ 代表 MEC 服务器的计算资源分配策略; $\omega_1$ 和 $\omega_2$ 分别代表能量消耗和执行时间的权重因子; $C_1$ 和 $C_2$ 表示物联设备的卸载决策约束; $C_3$ 和 $C_4$ 表示 MEC 服务器的计算资源约束; $C_5 \sim C_8$ 表示 UAV 飞行状态约束.

UAV 的飞行方向角和距离直接影响任务卸载和计算资源分配的决策. 边缘服务器的计算资源分配需要根据任务卸载决策来调整. 而卸载决策将直接影响设备的能耗和任务执行时延. 因此需要综合考虑各种约束条件和目标, 以找到最佳的解决方案.

## 4 问题求解

上述优化问题是一个包含二值变量和连续变量的 MINLP 问题, 通过传统的优化方法很难直接求解. 此外, UAV 轨迹的连续性与空地一体化网络环境的强动态性导致该优化问题具有庞大的动作维度空间. 更为困难的是, 出于对任务执行时延及信息传递开销的考虑, 获取全局状态信息是不现实的. MATD3 建立在深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 算法的基础上, 用于解决多智能体系统中的协同决策问题. 每个智能体都有两个神经网络, 分别是 Actor 网络和 Critic 网络. Actor 网络用于生成动作策略, 被训练来最大化累积奖励, 以指导智能体执行合适的动作. Critic 网络帮助 Actor 网络通过评估选择的动作是否有效来进行学习. MATD3 中的多个智能体彼此协同学习, 每个智能体都能观察整个系统的状态, 并根据系统奖励来调整其 Actor 和 Critic 网络. 这种合作学习使每个智能体能够在不同的状态下采取协同的动作, 以实现系统级的目标. 同时, 引入 “Twin Delayed” 机制, 它包括两组 Critic 网络, 每组有两个 Critic 网络, 以提高训练的稳定性和鲁棒性. 为此, 本节首先将上述问题建模为马尔可夫决策过程, 然后提出基于多智能体深度强化学习的 MATD3 算法对其进行求解.

### 4.1 基于马尔可夫决策过程的空地网络环境

我们考虑在空地异构网络中, 通过优化 UAV 的飞行轨迹、物联设备的卸载决策, 以及空基边缘服务器计算资源分配实现系统成本最小化. 该过程中, 终端设备、UAV 和 HAPS 的协同决策会对系统成本产生影响. 同时, 当前动作选择将触发系统从前一状态转移到下一状态. 因此, 本文将原始的 UAV 轨迹与网络资源协同优化问题建模为马尔可夫决策过程. 具体而言, 物联设备、UAV 与 HAPS 被视为 3 类智能体. 为了简化表达, 智能体的索引用 $z = \{1, \ldots, K, \ldots, K+M, K+M+1\}$ 表示, 总数为 $Z = K + M + 1$. 进而马尔可夫决策过程表示为 $\{S, A, R\}$, 其中 $S$ 表示状态空间, $A$ 表示动作空间, $R$ 为奖励函数. 在空地异构网络中, 状态空间、动作空间和奖励函数分别定义如下:

(1) 状态空间 (state and observation): 系统状态空间反应网络环境信息, 由物联设备与 UAV 位置坐标、任务信息以及设备到 UAV/HAPS 传输速率组成. 因此, $t$ 时隙, 状态空间表示为 $S(t)=\{\{H_k(t),Z_k(t)\}_{k\in\Gamma},\{M_m(t)\}_{m\in\mathcal{S}},\{R_{k,m}(t-1)\}_{k\in\Gamma,m\in\mathcal{S}}\}$.

(2) 动作空间 (action): 在空地网络中, 设备智能体需根据其本地观测结果选择边缘服务器进行任务卸载. 因此, $U_k$ 的动作空间 $A_k(t)$ 可以描述为一个有 $M$ 个分量的向量, $A_k(t)=\{a_{k,m}(t)\}$. UAV 智能体需确定飞行轨迹 (飞行方向角 $\vartheta_m(t)$、距离 $l_m(t)$) 及其计算资源分配 $f_{k,m}^E(t)$, 因此 $A_{K+m}(t)=\{\vartheta_m(t),l_m(t),f_{k,m}^E(t)\}$. 此外, HAPS 智能体的动作空间为 $A_{K+M+1}(t)=\{f_{k,m}^E(t)\}$, 指分配给各个物联终端任务的计算资源.

(3) 奖励函数 (reward): 一般来说, 网络即时奖励函数与目标函数有关. 在所考虑的优化问题中, 目标是最大限度地减少系统整体成本, 即网络能耗与任务执行时延的加权. 因此, 奖励的价值需要与目标函数的价值负相关. 此外, 为了通过合理的奖励函数设计避免智能体的决策违反计算资源限制、UAV 碰撞等约束, 还应对违反约束的智能体施加惩罚. 定义 $C(t)=\sum_{k=1}^{K}\omega_1E_k(t)+\omega_2T_k(t)$, 因此, 系统的奖励函数设置为

$$R(t)=R(S(t),A(t))=-C(t)-\nu_t\mathrm{Pen}_t, \tag{24}$$

其中, $\nu_t$ 是指示是否有智能体违反约束的二进制因子. 若 $\nu_t=1$, 代表有智能体违反计算资源或 UAV 飞行轨迹约束, 智能体受到惩罚 $\mathrm{Pen}_t$; 反之 $\nu_t=0$.

### 4.2 基于 MATD3 的空地网络资源分配算法

深度强化学习模型通过对智能体的训练, 可以捕获任务卸载、轨迹规划和资源分配的非线性关系, 学习到不同目标之间的权衡, 以找到最佳解决方案. 同时能够在较短的时间内生成决策, 满足实时性要求. 针对网络动态性导致的信息不确定性, 以及状态空间随网络规模指数级增长造成的维度空间爆炸难题, 本文提出基于多智能体深度强化学习 MATD3 的空地网络资源优化求解方法. 包括一个权重为 $\mu_n$ 的参与者 (Actor) 网络和两个权重为 $\theta_n^1$ 和 $\theta_n^2$ 的批评家 (Critic) 网络. 使用两个批评家网络, 每个智能体可以在一个批评家框架中处理 $Q$ 值的过拟合问题. 为进一步提升学习稳定性, 采用权重为 $\mu_n'$ 的 Actor 目标 (target Actor) 网络和权重为 $\theta_n^{i'}$ 的 Critic 目标 (target Critic) 网络. 此外, 考虑到网络环境的非平稳性, 为了保证收敛, 采用基于集中训练和分散执行的策略 [21].

以 UAV 为例, 为了稳定训练过程和提高样本效率, 每架无人机存储当前的经验 $s(t),s'(t),a(t),R(t)$. 对于每个 UAV, 从回放缓冲区随机抽样最小批量 (mini-batch) $\{s_j,s_j',a_j,r_j\}$. 然后, 将 $s_j$ 输入到 Critic 主体网络中生成策略, 每个 UAV 通过策略梯度更新 Actor 网络的权重 [22]:

$$\nabla J(\mu_n)=\frac{1}{M_b}\sum_{j=1}^{M_b}\nabla\mu_n\pi_n^{\mu}\left(s_n^j\right)Q_n^{\theta_1}\left(s_j,a_1^j,\ldots,a_N^j\right)\Big|a_n=\pi_n^{\mu}\left(s_n^j\right). \tag{25}$$

此外, 为了防止过度拟合, 将随机噪声添加到目标参与者网络中, 这可以实现更平滑的状态动作值估计. 修改的目标动作为

$$\widetilde{a_j}=\pi_n^{\mu'}\left(s_j'\right)+\tilde{\epsilon}, \tag{26}$$

进而可得目标值 $y_j$ 为

$$y_j=r_j+\delta_{i=1,2}^{\min}Q_n^{\theta_i}\left(s_j',a_j\right),\quad i=1,2. \tag{27}$$

然后, 基于策略 $\pi_n^{\mu}(s_j)$, 两个 Critic 网络将通过最小化损失函数 $L\left(\theta_n^i\right)$ 同时得到两个 $Q$ 值, 其定

**算法 1** 基于 MATD3 的空地网络资源分配算法

**输入:** Actor 网络和 Critic 网络, 经验回放池, 训练回合数.

初始化 Actor 网络参数以及 Critic 网络参数, 初始化经验回放池;
**for** episode = 1 to $E$ **do**
  重置空地网络环境;
  初始化 UAV 和设备位置, 设置动作探索噪声 $\rho(t)$;
  **for** time slot $t = 1$ to $T$ **do**
    **for** $z = 1$ to $Z$ **do**
      每个智能体 $z$ 根据当前的策略执行动作 $a_z(t) = \pi_z\left(o_z(t); {\theta_z}^{\pi}\right) + \rho(t)$;
      执行 $a(t) = \{a_1(t), \ldots, a_Z(t)\}$, 获得相应的奖励, 进入下一个状态 $S(t+1)$;
      **if** 经验回放池未满 **then**
        存储元组 $(S(t), A(t), R(t), S(t+1))$ 到经验回放池 $B$ 中;
      **else**
        用元组 $(S(t), A(t), R(t), S(t+1))$ 随机替换 $B$ 中一个元组;
        从经验回放池中随机选择大小为 $B_b$ 的小批量经验样本;
        根据最小化损失函数式 (28) 更新 Critic 网络权重 $\theta_n^{i'}$;
        **if** $t$ mod $d$ **then**
          根据式 (25) 更新 Actor 网络参数;
          根据式 (30) 更新 3 个目标网络参数;
        **end if**
      **end if**
    **end for**
  **end for**
**end for**

**输出:** 训练后的 Actor 网络模型.

义为

$$L\left(\theta_n^i\right) = \frac{1}{M_b}\sum_{j=1}^{M_b}\left[y_j - Q_n^{\theta_i}\left(s_j, a_j\right)\right]^2, \quad i = 1, 2. \tag{28}$$

每个 UAV 通过式 (29) 更新 3 个当前网络参数:

$$\begin{aligned} \mu_n &= \mu_n - \lambda \nabla_{\mu_n} J\left(\mu_n\right), \\ \theta_n^i &= \theta_n^i - \lambda \nabla_{\theta_n^i} L\left(\theta_n^i\right), \quad i = 1, 2, \end{aligned} \tag{29}$$

其中 $\lambda$ 表示学习率.

为了减少时间差异学习带来的误差, 每个 UAV 以低于评价评论家网络的频率更新评价行动者网络的权重. 在这里, 每个 UAV 选择每 $d$ 个时间步更新 3 个 target 网络参数, 如式 (30) 所示:

$$\begin{aligned} \mu_n' &= \tau\mu_n + (1-\tau)\,\mu_n', \\ \theta_n^{i'} &= \tau\theta_n^i + (1-\tau)\,\theta_n^{i'}, \quad i = 1, 2. \end{aligned} \tag{30}$$

基于 MATD3 的空地网络资源分配方法具体如算法 1 所示, 其他各类智能体训练过程类似, 此处不再赘述.

## 5 仿真与分析

本文进行了大量的仿真实验, 以验证所提方法的有效性, 仿真平台为 Python3.7, 采用 Tensorflow-

**表 1 仿真参数表**

**Table 1** Simulation parameter table

| Simulation parameter | Value |
|---|---|
| Number of IoT-devices | 50 |
| Number of UAV | 3 |
| Carrier frequency | 0.1 GHz |
| Noise power | −104 dBm |
| Environmental information $b_1, b_2$ | 4.88, 0.43 |
| Additional loss $\eta_{k,n,t}^{\mathrm{Los}}, \eta_{k,n,t}^{\mathrm{NLos}}$ | 0.1, 21 |
| Uplink transmission bandwidth | 10 MHz |
| Maximum transmission power of terminal | 0.5 W |
| Efficiency coefficient of terminal computing power | $5 \times 10^{-27}$ J/Hz$^3$/s |
| Computation capacity of UAV and HAPS | 30 GHz, 50 GHz |
| Computation capacity of UE | 2 GHz |
| Maximum and minimum flying altitude of UAV | 100 m, 50 m |
| Maximum horizontal distance and maximum vertical distance of UAV | 20 m, 12 m |
| Elevation angle of UAV | 42.44° |

2.0. 考虑 400 m × 400 m 区域, 物联设备在该区域内随机分布. 计算任务的输入数据大小在 [2.1, 5.4] Mbits 之间服从均匀分布, 所需 CPU 周期数在 $[10^8, 10^9]$ 范围内. 参考文献 [2, 8, 13, 23, 24] 中的设置, 具体仿真参数如表 1 所示. 本文的神经网络采用 ReLU 作为激活函数, 设置 3 个全连接的隐藏层, 每层神经元数分别为 256, 64 以及 16, 折扣系数 $\Gamma$ 为 0.99, 经验回放池 $B$ 为 100000, 最小批量 $B_b$ 为 100. 采用学习率为 0.001 的 ADAM 优化器更新所有智能体 Critic 网络和 Actor 网络. 为了评估所提算法的性能, 本文与 4 种基准方法进行比较.

(1) TD3-NT 算法: 采用 TD3 算法优化 MEC 服务器资源分配以及物联设备卸载决策. UAV 采用固定的飞行轨迹, 即以同心圆环绕地面物联网设备进行飞行.

(2) MADDPG 算法: 采用 MADDPG 算法优化.

(3) 单智能体深度确定性策略梯度 (DDPG) 算法 [25]: 每个智能体采用单智能体 DDPG 算法训练其决策, 该架构同样具有评估网络和目标网络的双网络结构. 然而, 训练阶段智能体之间无法交互信息, 只能基于本地信息更新神经网络.

(4) 双深度 Q 网络 (double deep Q network, DDQN) 算法 [26]: 使用 DDQN 方法. 因其动作空间是离散的, 为了处理连续作用空间, 需量化连续值, 并通过从有限离散值中近似取值来进行训练.

图 2 显示系统奖励随着 episode 的变化情况. 本文设置 500 轮 episode, 奖励函数设置为系统总能耗和时延的加权和, 其中 $\omega_1$ 和 $\omega_2$ 均设置为 1. 由该图可以观察到, 随着训练 episode 的增加, 系统奖励呈上升的趋势. 这表明随着训练轮数的增加, 系统总计算成本在不断减少, 算法通过不断与环境进行交互, 能够有效地学习到较优的资源分配策略以及任务卸载决策, 从而达到降低计算成本的目的. 当 episode 接近于 150 时, 本文所提算法趋近于收敛, 但由于空地一体化异构网络的动态性, 网络拓扑也随之变化, 导致算法收敛数值在一定区间内波动. 此外本文所提算法在收敛性能上优于 MADDPG 算法, 主要原因在于相比 MADDPG 方法, TD3 采用了两套 Critic 网络.

图 3 显示系统总成本与物联设备数量之间的关系. 随着物联设备的增加, 系统总计算成本呈上升

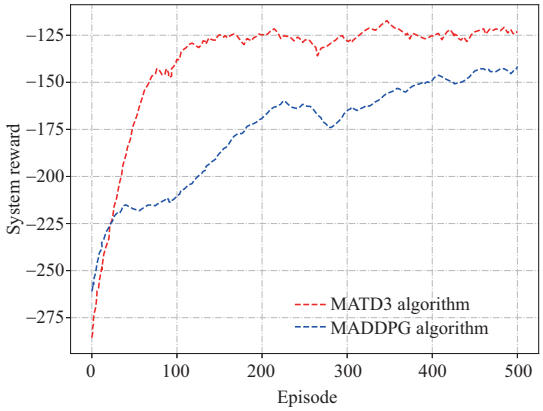


图 2 (网络版彩图) 收敛性能

**Figure 2** (Color online) Convergence performance

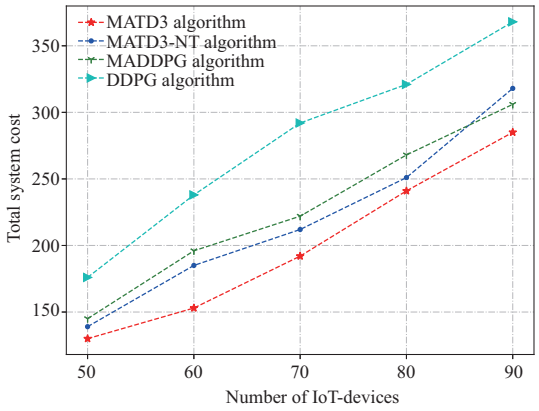


图 3 (网络版彩图) 终端数量影响

**Figure 3** (Color online) Impact of terminal's quantity

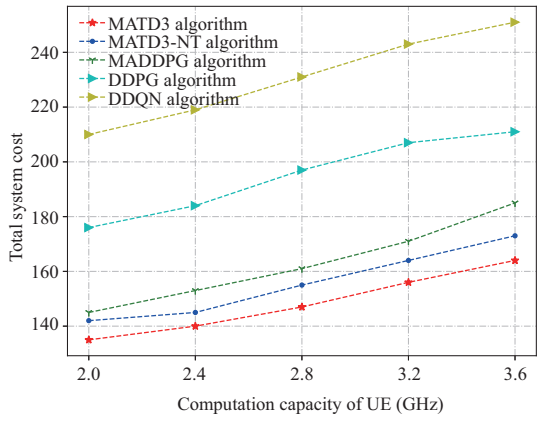


图 4 (网络版彩图) 终端设备计算能力影响

**Figure 4** (Color online) Impact of ternimal's computing capacity

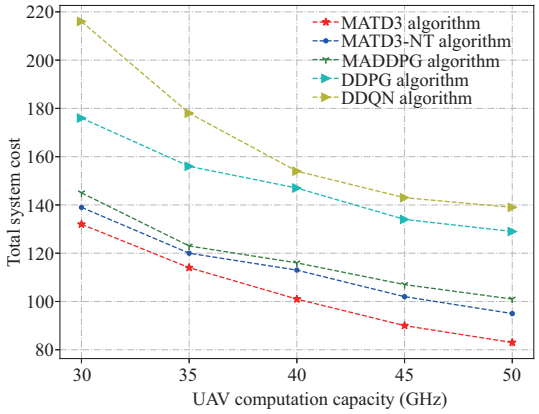


图 5 (网络版彩图) 无人机计算能力影响

**Figure 5** (Color online) Impact of UAV's computing capacity

的趋势. 原因在于, 随着终端数量增加, 由于 MEC 服务器所提供的计算资源有限, 这会导致设备来竞争计算资源. 但本文所提算法能够更为有效地对服务器资源进行分配, 并通过优化 UAV 轨迹为物联设备提供计算服务, 从而系统性能更佳. 此外, 与其他 3 种方法相比, 我们提出的方法成本更低且更稳定. 例如, 在终端数量为 60 时, 本文方法比 MATD3-NT 算法、MADDPG 算法、DDPG 算法分别降低了 15.4%, 18.7% 和 28.1% 的成本.

图 4 显示系统总成本受终端计算能力的影响. 随着设备计算能力的增加, 系统总成本呈上升的趋势. 原因在于, 当本地计算资源满足终端任务需求时, 任务更倾向于在本地执行. 尽管能耗随着用户计算能力的提高而增加, 但本文所提算法的能耗始终保持在较低水平, 且波动相对较小. 这是由于 MATD3 可以有效学习训练策略, 进而作出相应决策.

图 5 显示系统总成本受 UAV 计算能力的影响. 随着 UAV 计算能力的增加, 系统总成本呈下降趋势. 这是由于 UAV 可以将更多的计算资源分配给终端设备, 从而降低设备计算能耗和计算时延, 达到降低系统成本的目的. 但随着 UAV 计算能力的增加, 系统成本降低趋势逐步减缓, 这是由于有限

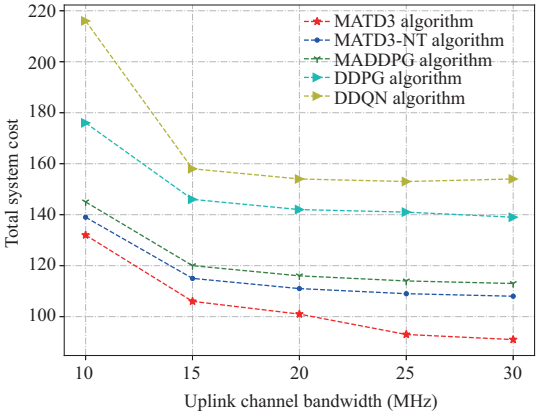


图 6 (网络版彩图) 传输带宽影响

**Figure 6** (Color online) Impact of transmission bandwidth

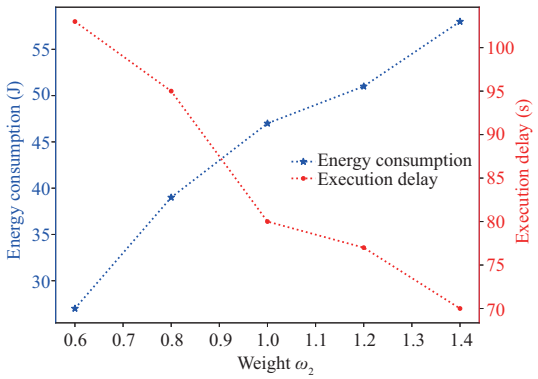


图 7 (网络版彩图) 权重因子影响

**Figure 7** (Color online) Influence of weight factor

的通信资源将制约系统计算成本的进一步降低. 由于 MATD3 算法可以根据其他智能体学习到更优的执行策略, 可以更好地降低系统成本. 本文算法由于联合优化 UAV 轨迹、MEC 服务器资源分配以及设备卸载决策, 因此在系统计算能耗和时延方面始终保持更优的性能. 例如, 当 UAV 计算能力为 50 GHz 时, 本文方法比 MATD3-NT 算法、MADDPG 算法、DDPG 算法、DDQN 算法分别降低了 12.8%, 19.8%, 24.8% 和 40.3% 的成本.

图 6 显示系统总成本受传输带宽的影响. 随着传输带宽的增加, 系统总成本呈下降的趋势. 随着上行传输带宽的增加, 可以提高物联设备上行传输速率, 进而降低传输延迟以及能量消耗. 从图 6 可以观测到本文所提算法可以更好地降低系统成本. 在 MATD3-NT 算法中, 由于 UAV 采用固定的飞行轨迹, 在系统成本方面相对较差. DDQN 算法随着状态空间以及动作空间的增大, 将更难学习到最佳策略, 导致其性能较差. 在传输带宽为 30 MHz 时, 本文方法比 MATD3-NT 算法、MADDPG 算法、DDPG 算法、DDQN 算法分别降低了 24.5%, 27.2%, 41.5% 和 46.8% 的成本.

图 7 显示系统总能耗和时延受权重因子的影响情况. 在权重因子 $\omega_1$ 为 1 的情况下, 权重因子 $\omega_2$ 由 0.6 增长到 1.6. 从图 7 可以观察到, 随着 $\omega_2$ 的增加, 系统能耗呈上升趋势, 而系统执行时间呈下降趋势. 这是由于当 $\omega_2$ 较小时, 算法训练智能体更加注重系统能耗, 当 $\omega_2$ 增大后, 系统训练智能体更加注重降低任务执行时间.

图 8 显示系统总成本受 UAV 数量的影响. 随着 UAV 数量的增加, 系统总成本呈下降的趋势. 反映了当较多 UAV 可用时, 系统变得更加高效和经济. 原因在于 UAV 数量增加, MEC 服务器的计算资源更充足, 计算时延降低, 同时设备更倾向于将任务卸载给边缘服务器, 降低了用户的计算能耗. 特别地, 所提算法在不同 UAV 数量下都能保持较高的性能水平, 表明该算法在系统的可扩展性方面表现出色.

## 6 总结

本文研究了空地网络无人机轨迹与资源分配的联合优化问题, 目标是最小化系统总成本. 针对目标函数的非凸性, 以及网络动态性导致的信息不确定与维度空间爆炸问题, 我们通过将该问题建模为马尔可夫决策过程, 研究提出一种基于多智能体深度强化学习的 MATD3 算法. 大量的实验结果表明,

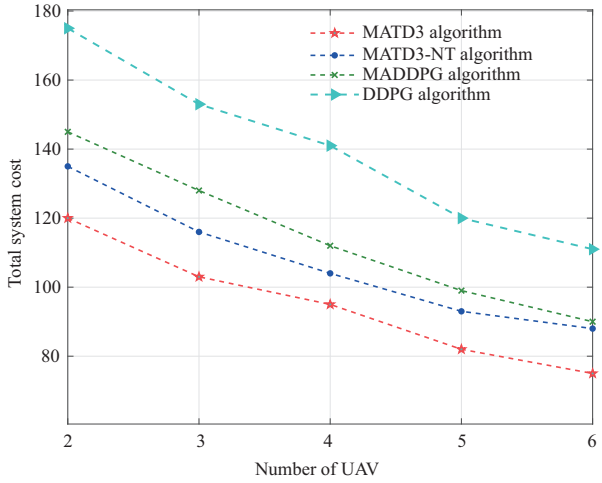


**图 8** (网络版彩图) UAV 数量影响

**Figure 8** (Color online) Influence of UAV's quantity

与基准算法相比, 本文所提的方法在系统计算能耗和时延加权的总成本方面性能更优.

## 参考文献


1 Yu Y. Mobile edge computing towards 5G: vision, recent progress, and open challenges. China Commun, 2016, 13: 89–99

2 Qin P, Fu Y, Zhao X W, et al. Optimal task offloading and resource allocation for C-NOMA heterogeneous air-ground integrated power Internet of Things networks. IEEE Trans Wirel Commun, 2022, 21: 9276–9292

3 Qin P, Wang M, Zhao X W, et al. Content service oriented resource allocation for space-air-ground integrated 6G networks: a three-sided cyclic matching approach. IEEE Internet Things J, 2022, 10: 828–839

4 Raza S, Wang S, Ahmed M, et al. Task offloading and resource allocation for IoV using 5G NR-V2X communication. IEEE Internet Things J, 2021, 9: 10397–10410

5 Qin P, He H T, Zhao X W, et al. Efficient resource allocation with context-awareness for parked car road side unit-based Internet of vehicles. J Commun, 2022, 43: 113–125 [秦鹏, 和昊婷, 赵雄文, 等. 基于停放车辆路边单元环境感知的车联网资源高效分配. 通信学报, 2022, 43: 113–125]

6 Li M, Cheng N, Gao J, et al. Energy-efficient UAV-assisted mobile edge computing: resource allocation and trajectory optimization. IEEE Trans Veh Technol, 2020, 69: 3424–3438

7 Wang Y, Ru Z, Wang K, et al. Joint deployment and task scheduling optimization for large-scale mobile users in multi-UAV-enabled mobile edge computing. IEEE Trans Cybern, 2020, 50: 3984–3997

8 Xu Y, Zhang T, Yang D, et al. Joint resource and trajectory optimization for security in UAV-assisted MEC systems. IEEE Trans Commun, 2021, 69: 573–588

9 Yu Z, Gong Y, Gong S, et al. Joint task offloading and resource allocation in UAV-enabled mobile edge computing. IEEE Internet Things J, 2020, 7: 3147–3159

10 Liu Y, Xie S, Zhang Y. Cooperative offloading and resource management for UAV-enabled mobile edge computing in power IoT system. IEEE Trans Veh Technol, 2020, 69: 12229–12239

11 Ji J, Zhu K, Yi C, et al. Energy consumption minimization in UAV-assisted mobile-edge computing systems: joint resource allocation and trajectory design. IEEE Internet Things J, 2021, 8: 8570–8584

12 Zhang J, Zhou L, Tang Q, et al. Stochastic computation offloading and trajectory scheduling for UAV-assisted mobile edge computing. IEEE Internet Things J, 2019, 6: 3688–3699

13 Sun C, Ni W, Wang X. Joint computation offloading and trajectory planning for UAV-assisted edge computing. IEEE Trans Wirel Commun, 2021, 20: 5343–5358

14 Zhan C, Hu H, Liu Z, et al. Multi-UAV-enabled mobile-edge computing for time-constrained IoT applications. IEEE

Internet Things J, 2021, 8: 15553–15567
15 Yao Z X, Xia S C, Li Y. Task offloading and resource allocation in an uncertain network. Sci Sin Inform, 2022, 52: 1349–1361 [姚枝秀, 夏士超, 李云. 不确定网络环境下的任务卸载和资源分配算法. 中国科学: 信息科学, 2022, 52: 1349–1361]
16 Yao J, Ansari N. QoS-aware machine learning task offloading and power control in Internet of Drones. IEEE Internet Things J, 2023, 10: 6100–6110
17 Zhang T K, Chen C B, Xu Y, et al. Joint task scheduling and multi-UAV deployment for aerial computing in emergency communication networks. Sci China Inf Sci, 2023, 66: 192303
18 Yan F, Zhu X P, Zhou Z, et al. Real-time task allocation for a heterogeneous multi-UAV simultaneous attack. Sci Sin Inform, 2019, 49: 555–569 [严飞, 祝小平, 周洲, 等. 考虑同时攻击约束的多异构无人机实时任务分配. 中国科学: 信息科学, 2019, 49: 555–569]
19 Cheng X, Lyu F, Quan W, et al. Space/aerial-assisted computing offloading for IoT applications: a learning-based approach. IEEE J Sel Areas Commun, 2019, 37: 1117–1129
20 Peng H X, Shen X M. Multi-agent reinforcement learning based resource management in MEC- and UAV-assisted vehicular networks. IEEE J Sel Areas Commun, 2021, 39: 131–141
21 Zhou S Y, Cheng Y F, Lei X, et al. Resource allocation in UAV-assisted networks: a clustering-aided reinforcement learning approach. IEEE Trans Veh Technol, 2022, 71: 12088–12103
22 Ding F, Xu L, Meng D D, et al. Gradient estimation algorithms for the parameter identification of bilinear systems using the auxiliary model. J Comput Appl Math, 2020, 369: 11257
23 Alzenad M, El-Keyi A, Lagum F, et al. 3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage. IEEE Wirel Commun Lett, 2017, 6: 434–437
24 Qin P, Fu Y, Tang G M, et al. Learning based energy efficient task offloading for vehicular collaborative edge computing. IEEE Trans Veh Technol, 2022, 71: 8398–8413
25 Qiu C, Hu Y, Chen Y, et al. Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications. IEEE Internet Things J, 2019, 6: 8577–8588
26 Tang F X, Hofner H, Kato N, et al. A deep reinforcement learning-based dynamic traffic offloading in space-air-ground integrated networks (SAGIN). IEEE J Sel Areas Commun, 2021, 40: 276–289

# Air-ground integrated network resource optimization based on MATD3

Peng QIN[1*], Shuo WANG[1,2], Min FU[1] & Xiongwen ZHAO[1]

1. *School of Electrical and Electronic Engineering, North China Electric Power University, Beijing* 102206*, China*;
2. *Hebei Telecom Co., Ltd., Shijiazhuang* 050036*, China*
* Corresponding author. E-mail: qinpeng@ncepu.edu.cn

**Abstract** Mobile edge computing effectively reduces service latency and terminal energy consumption by offloading tasks to the edge of the wireless networks. For tremendous IoT devices distributed in remote areas (e.g., wind power, photovoltaic, and other power IoT terminals), existing terrestrial communication networks are not able to provide effective computing services. Therefore, in this paper, we comprehensively consider an air-ground integrated heterogeneous network model to minimize the sum of execution delay and energy consumption of IoT devices by jointly designing UAV trajectory, task offloading, and computing resource allocation. Regarding the non-convexity of the objective function and the information uncertainty caused by network dynamics, the problem is modeled as a Markov decision process, and we propose a joint UAV trajectory and network resource optimization algorithm based on MATD3. Experimental results show that compared with other baselines, the proposed scheme has superior performance in system computing energy consumption and delay.

**Keywords** air-ground integrated heterogeneous network, offloading decision-making, resource allocation, UAV trajectory optimization, multi-agent deep reinforcement learning (MADRL)