基于深度强化迁移学习与干扰管理的无人机辅助无蜂窝网络部署策略

谢亚琴[1,2]，张念[1,2]，张余[1,2]，张治中[1,2]，李佳珉[3]，王东明[3]
（1.南京信息工程大学电子与信息工程学院，江苏 南京 210044；2.复杂环境智能保障技术教育部重点实验室，江苏 南京 210044；
3.东南大学信息科学与工程学院，江苏 南京 210096）

摘要：针对无蜂窝（Cell-Free）网络因依赖固定接入点（AP）而难以应对突发业务需求及无基础设施覆盖场景的问题，提出了一种基于深度强化迁移学习与干扰管理的动态无人机（UAV）辅助Cell-Free网络快速部署方法。该方法将UAV作为动态可扩展节点，通过深度强化学习驱动多UAV协同部署，兼顾补充覆盖与负载均衡；引入迁移学习机制，将源域习得策略快速适配至目标域，降低模型重训练复杂度；融入基于干扰管理的分布式频谱资源分配策略，实现相邻UAV间同频干扰的动态规避。仿真结果表明，与传统深度强化学习部署策略相比，所提方法收敛更快、拥塞缓解更优，用户等待概率小于2%，平均用户满意度提升44.3%，可突破固定AP的覆盖局限，为突发场景提供了一种高效、智能的解决方案。

关键词：无人机；深度强化学习；迁移学习；干扰管理
中图分类号：TN929.5
文献标志码：A
doi：10.11959/j.issn.1000–436x.TXXB260290

Deployment strategy for UAV-assisted cell-free networks based on deep reinforcement transfer learning and interference management

Xie Yaqin[1,2], Zhang Nian[1,2], Zhang Yu[1,2], Zhang Zhizhong[1,2], Li Jiamin[3], Wang Dongming[3]
1. School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. Key Laboratory of Intelligent Support Technology for Complex Environments, Ministry of Education, Nanjing 210044, China
3. School of Information Science and Engineering, Southeast University, Nanjing 210096, China

Abstract: To address the problem that cell-free networks relying on fixed access point (AP) was incapable of handling burst traffic demands and infrastructure-lacking scenarios, a rapid deployment method for dynamic unmanned aerial vehicle (UAV)-assisted cell-free networks based on deep reinforcement transfer learning and interference management (IM) was proposed. In the proposed method, UAV served as scalable nodes, and deep reinforcement learning (DRL) was employed to drive cooperative deployment, enabling both supplementary coverage and dynamic load balancing. A transfer learning mechanism was further introduced to adapt source-domain policies to target-domain dynamic scenarios, thereby substantially reducing the retraining overhead. Meanwhile, an IM-based distributed spectrum allocation strategy was incorporated to mitigate co-channel interference among densely deployed UAV. Simulation results demonstrate that, compared with conventional DRL-based strategies, the proposed method achieves faster convergence and better congestion relief, limits the user waiting probability to 2%, and improves the average user satisfaction by 44.3%. The proposed method overcomes the coverage limitations of fixed AP, providing an efficient and intelligent solution for bursty scenarios.

Key words: UAV, deep reinforcement learning, transfer learning, interference management

收稿日期：2026–05–21；**修回日期：**2026–07–25
通信作者：谢亚琴，xyq@nuist.edu.cn
基金项目：国家自然科学基金资助项目（No.62001238）；江苏省科技重大专项基金资助项目（No.BG2024002）
Foundation Items: The National Natural Science Foundation of China (No.62001238), Major Science and Technology Project of Jiangsu Province (No.BG2024002)

0 引言

随着5G及未来6G网络的迅速发展，数据流量和用户密度持续攀升，现有基站（base station, BS）在高负载及突发业务需求情况下逐渐暴露出覆盖不足和超负荷等问题[1-3]。特别是在应急救援、大型活动或灾后恢复等场景中，地面BS由于受基础设施损坏或网络拥塞的限制，难以及时为所有地面用户（ground user, GU）提供稳定的通信服务[1,4]。因此，亟须一种具备动态部署和高适应性的智能化网络补偿机制，以快速扩展通信覆盖范围，缓解BS压力。

针对上述问题，无蜂窝（Cell-Free）架构通过固定部署大量分布式接入点（access point, AP）并借助中央处理单元协同调度，可有效消除覆盖盲区并分担局部负载，被视为6G网络的关键使能技术之一。然而，Cell-Free架构的大规模部署依赖预设的光纤前传网络与固定AP，建设成本高昂，且难以应对灾区、临时集会等无基础设施覆盖或高度动态的突发场景。为弥补Cell-Free架构在上述突发场景中的不足，本文引入无人机（unmanned aerial vehicle, UAV）作为空中基站（UAV-BS），凭借其高机动性与按需部署能力，为GU提供灵活覆盖[5]。通过将UAV作为Cell-Free架构的动态扩展与补充节点，可在不依赖预设地面设施的前提下，快速响应突发业务需求，填补固定AP无法覆盖的盲区，从而形成地面固定AP构成基础覆盖层、UAV作为动态补充节点的协同网络架构。

UAV作为UAV-BS，可在短时间内进入地面网络的盲区或过载区域，提供临时通信支持，从而有效缓解地面BS压力并提高网络服务连续性[1-2,6-7]。例如，在物联网数据采集与移动边缘计算系统中，UAV可通过灵活的轨迹规划完成数据汇聚与传输任务，显著提升覆盖效率与系统吞吐量[1,4-5]。然而，在多用户、多BS、多UAV的复杂动态环境中，如何实现UAV的高效部署及任务调度仍是研究热点与挑战之一[1,5]。

Gao等[8]提出了一种多UAV-BS快速部署方法，专门针对灾害应急通信场景，通过协同部署多架UAV来快速恢复受损区域的通信网络，确保所有目标区域都得到有效覆盖。Ao等[9]研究了多UAV协同轨迹优化问题，最大化应急通信部署中的通信效率，同时维持UAV之间的连接性，这对于动态覆盖无线通信网络至关重要。尽管传统优化方法在UAV部署中取得了一定进展，通过几何优化来确定最佳悬停位置或使用启发式算法来规划飞行路径，但这些方法通常基于简化的信道模型和静态用户分布假设[10]。在实际动态且复杂的用户环境中，这些方法暴露出诸多局限性，包括计算复杂度高、可扩展性差，以及难以有效管理多UAV部署带来的同频干扰问题[11-13]。当多UAV在有限频谱资源下协同工作时，其三维部署特性会显著增加空对地和空对空之间的干扰链路，若缺乏精细的干扰管理机制，将严重影响通信效率和系统容量[14]。为此，Chen等[15]提出了一种基于图着色与禁忌搜索的干扰管理频谱资源分配（interference management-based spectrum resource allocation, IMSA）算法，通过为覆盖相同用户的相邻UAV分配不同频谱资源，有效避免了同频干扰，显著提升了系统容量。

为了应对传统方法的挑战，深度强化学习（deep reinforcement learning, DRL）为UAV网络中的智能决策提供了强大的框架[15-17]。DRL通过结合深度神经网络（deep neural network, DNN）的感知能力和强化学习（reinforcement learning, RL）的决策能力，使UAV能够从与环境的交互中学习最优策略，从而实现对复杂动态场景的自适应优化[15,18]。Chen等[15]提出了一种基于DRL近端策略优化的UAV-BS飞行优化方法，通过优化UAV战略定位来最大化救生潜力，特别适用于紧急情况下的快速网络部署。李斌等[19]提出了一种UAV搭载边缘计算服务器与雷达收发器辅助通信–感知融合网络，通过联合优化波束成形、计算资源分配、任务卸载与飞行轨迹，建立了系统能耗最小化模型，采用DRL实现动态环境下的优化决策，有效解决了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。Lagos等[17]利用DRL为UAV辅助蜂窝网络实现了公平覆盖，有效管理了部分可观测的移动用户、多跳回传连接以及同频干扰，并通过最小化UAV间距来优化覆盖效率。Qureshi等[12]提出了一种基于多智能体深度强化学习（multi-agent DRL, MADRL）的UAV-BS自主部署方案，不需要网络侧协助，在紧急场景中为GU设备提供无线覆盖，特别适用于用户集群分布未知的情况[12,20-21]。此外，Xu等[22]在卫星集群协调传

输研究中提出的多智能体 Double 和 Dueling 深度 Q 学习（multi-agent deep double dueling Q-learning, MAD3QL）方法，通过引入增量学习机制实现了对动态规模集群的高效适配。该方法虽面向卫星场景，但其核心的多智能体协调与增量学习思想对本文同样具有重要借鉴意义。这些 DRL 方法显著提升了 UAV 在通信覆盖和资源管理方面的性能。然而，DRL 方法也存在其固有的缺点，如训练过程对大量经验样本的依赖导致样本效率低下，在奖励稀疏或环境复杂的情况下收敛速度慢，并且在将一个场景下训练好的模型直接应用于另一个异构场景时，泛化能力往往不足，需要耗费大量时间和计算资源重新训练[10,12,23]。

为了克服 DRL 在泛化能力和训练效率上的不足，迁移学习（transfer learning, TL）提供了一个有效的解决方案[2,12]。TL 通过将在源任务或源域上预训练的模型知识迁移到目标任务或目标域，显著减少新场景的训练时间和数据需求。这种 TL 与 DRL 的结合，使 UAV 网络能够快速适应新的操作环境，同时保持对干扰和流量波动的鲁棒性[2]。Chu 等[24]提出了一种基于深度 Q 网络和 TL 的 UAV 动态服务调度算法（D3QL-TL），通过将预训练的 DRL 模型迁移至新任务，显著加速了模型收敛，并降低了系统成本。

基于上述分析，本文提出了一种基于深度强化迁移学习与 IMSA 的 UAV 辅助通信网络快速部署方法。该方法以最大化 GU 的业务需求为目标，在地面 BS 未覆盖区域动态部署 UAV，为 GU 提供临时通信服务，并缓解超负荷 BS 的负载。此外，本文还考虑了策略的泛化能力，将源任务训练获得的模型经验和策略迁移到不同目标任务中。本文的主要贡献如下。

（1）设计了一种基于迁移学习的多智能体 Double 和 Dueling 深度 Q 学习（multi-agent deep double dueling Q-learning with transfer learning, MAD3QL-TL）算法的跨场景快速部署策略。通过复用源域经验，显著增强 UAV 网络在突发业务场景及 BS 损毁下的泛化能力与自适应性能。

（2）提出了一种基于干扰管理的 MAD3QL-TL 算法。联合 IMSA 模块，根据 UAV 的空间分布动态分配频谱资源，从而抑制多 UAV 三维部署引发的同频干扰。

（3）在多用户、多 BS、多 UAV 的仿真环境中对所提方法进行验证。结果表明，该方法在网络覆盖率、资源利用率及收敛效率方面均显著优于传统 RL，实现了对复杂通信环境的高效自适应优化。

1　系统模型

考虑一个如图 1 所示的多 UAV 辅助 Cell-Free 网络通信场景。该场景以固定部署的分布式 AP 构成 Cell-Free 基础覆盖层，同时引入多架 UAV 作为动态补充节点，形成地面固定 AP 构成基础覆盖层、UAV 作为动态补充节点的协同网络架构。本文研究了以下 3 种动态场景：中等、高等突发需求以及 BS 损毁场景。在中等和高等突发需求下，与源域任务相比，区域内分别产生中等和高等突发业务量需求。在 BS 损毁场景下，由于部分 BS 损毁，源域内原有的业务量需求将得不到满足。在上述 3 种场景下，均可利用 UAV 快速部署至该区域，从而为业务量需求得不到满足的 GU 提供临时通信服务。

该场景进一步划分为源域与目标域任务场景。在源域中，UAV 通过 DRL 来学习不同任务类型（Type 1/2/3）下的网络部署与干扰管理策略；获得的经验与策略可迁移至目标域中不同的应急场景（如 BS 损毁、高负载突发等），实现快速适配与部署。假设 $L \times L$ 区域内存在 M 个 GU，其集合记为 $\mathcal{M}=\{1,2,\cdots,M\}$；第 m 个 GU 的三维坐标表示为 $p_m^{\mathrm{GU}}=\left(x_m^{\mathrm{GU}},y_m^{\mathrm{GU}},0\right),m\in\mathcal{M}$。考虑业务量需求由小到大的 3 种 GU 业务类型，即语音、图像和视频，第 m 个 GU 对应的业务量需求设为 T_m。此外，设区域内 BS 总数为 K，BS 集合记为 $\mathcal{K}=\{1,2,\cdots,K\}$；第 k 个 BS 的三维坐标为 $p_k^{\mathrm{BS}}=\left(x_k^{\mathrm{BS}},y_k^{\mathrm{BS}},h_k^{\mathrm{BS}}\right)$，覆盖半径为 r_k^{BS}，覆盖区域为 $\mathcal{R}_k^{\mathrm{BS}}$。

设区域内部署的 I 个 UAV 以固定高度 h 飞行，为 GU 提供服务，UAV 集合记为 $\mathcal{I}=\{1,2,\cdots,I\}$；第 i 个 UAV 的覆盖半径和覆盖区域分别表示为 $r_i^{\mathrm{U}}=\dfrac{h}{\tan\theta}$ 和 $\mathcal{R}_i^{\mathrm{U}}$；第 i 个 UAV 的三维坐标为 $p_i^{\mathrm{U}}=\left(x_i^{\mathrm{U}},y_i^{\mathrm{U}},h\right)$，其中，$(x_i^{\mathrm{U}},y_i^{\mathrm{U}})\in[0,L]$。

在无线通信系统中，信道模型是模拟真实无线通信场景的关键。针对空地通信场景，本文采用被广泛应用的概率路径损耗模型[25]，其中 UAV i 与

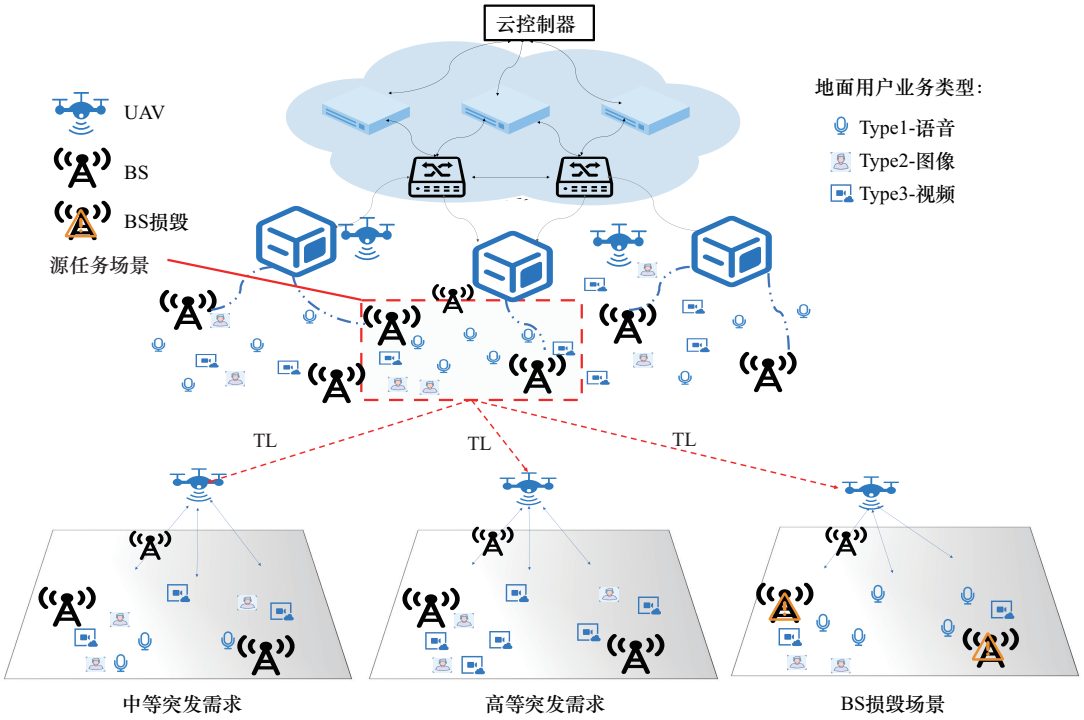


图1　多UAV辅助Cell-Free网络通信场景

GU m之间的视距（line of sight, LoS）通信概率可表示为：

$$P_{\mathrm{LoS}}\left(\theta_{i,m}\right)=\frac{1}{1+\alpha e^{-\beta\left(\theta_{i,m}-\alpha\right)}} \tag{1}$$

其中，α和β为环境参数，$\theta_{i,m}$为UAV i和GU m之间的仰角，UAV i和GU m之间的平均路径损耗$L_{i,m}$可表示为：

$$L_{i,m}=P_{\mathrm{LoS}}\left(\theta_{i,m}\right)L_{i,m}^{\mathrm{LoS}}+\left(1-P_{\mathrm{LoS}}\left(\theta_{i,m}\right)\right)L_{i,m}^{\mathrm{NLoS}} \tag{2}$$

LoS和非视距（non-line-of-sight, NLoS）下的路径损耗为：

$$\begin{cases} L_{i,m}^{\mathrm{LoS}}=20\lg\left(\dfrac{4\pi f d_{i,m}}{c}\right)+\eta_{\mathrm{LoS}} \\ L_{i,m}^{\mathrm{NLoS}}=20\lg\left(\dfrac{4\pi f d_{i,m}}{c}\right)+\eta_{\mathrm{NLoS}} \end{cases} \tag{3}$$

其中，f为空地信道的载波频率，c为光速。η_{LoS}和η_{NLoS}分别为LoS和NLoS下的平均附加路径损耗[26]，$d_{i,m}$为UAV i和GU m之间的欧氏距离。

若第m个GU接收到某UAV的信噪比高于信噪比阈值Θ_{th}，则认为该UAV可向GU m提供通信服务。UAV i与GU m之间的信噪比计算如下：

$$\Gamma_{i,m}=\frac{p_{i,m}L_{i,m}}{P_{\mathrm{gw}}+\sum\limits_{j\neq i}p_{j,m}L_{j,m}} \tag{4}$$

其中，P_{gw}为高斯白噪声功率，$p_{i,m}$为UAV i向GU m发射的功率。UAV i与GU m之间的吞吐量可表示为：

$$R_{i,m}=\frac{B}{N_i}\mathrm{lb}\left(1+\Gamma_{i,m}\right) \tag{5}$$

其中，B表示UAV的带宽，N_i表示接入UAV i的GU数量。

关于GU的服务质量（quality of service, QoS）要求，本文采用文献[27]中描述的满意度函数，由UAV i服务GU m的满意度可表示为：

$$f_{i,m}=\frac{1}{1+\exp\left[-\vartheta_m\left(R_{n,m}-R'_m+\dfrac{C}{\vartheta_m}\right)\right]} \tag{6}$$

其中，R'_m表示GU m的吞吐量需求，ϑ_m表示GU m的紧急程度，$\vartheta_m\in\{1,2,3\}$；C为常数，用于控制函数的形状，通常取$C\geqslant 7$。

根据Erlang C公式[28]，当BS k无空闲信道可用时，新接入的GU需等待（等待时间t_w满足$t_w>0$）的概率为：

$$W_k(t_w>0)=\frac{\dfrac{a_k^{l_k}}{l_k!}\dfrac{l_k}{l_k-a_k}}{\displaystyle\sum_{n=0}^{l_k-1}\frac{a_k^n}{n!}+\frac{a_k^{l_k}}{l_k!}\frac{l_k}{l_k-a_k}} \tag{7}$$

其中，a_k为BS k服务的GU总业务量，l_k为BS k的总信道数（文中假设所有BS的信道总数相等）。

2　问题描述

为应对局部区域的高业务负载需求，降低新接入GU的等待概率，本文在现有BS网络基础上引入UAV-BS构建动态辅助通信层，优先服务现有BS网络覆盖盲区的用户，并有效分流高业务负载BS的业务量。设UAV i为覆盖盲区内GU m提供服务的业务量为$T_{i,m}^{\mathrm{U}}$（$T_{i,m}^{\mathrm{U}}\geqslant 0$），则UAV i的总业务量可表示为：

$$T_i^{+}=\sum_{m\in\mathcal{M}}T_{i,m}^{\mathrm{U}} \tag{8}$$

同时，为缓解现有高负载BS的业务量压力，UAV还为这些BS提供流量卸载服务。设第k个BS服务的GU m业务量为$T_{k,m}^{\mathrm{BS}}(T_{k,m}^{\mathrm{BS}}\geqslant 0)$，BS最大业务量为$T_{\max}^{\mathrm{BS}}$。在业务高峰时段，总业务量$T_k^{\mathrm{BS}}=\sum_{m\in\mathcal{M}}T_{k,m}^{\mathrm{BS}}$可能超出容量限制。为保障QoS，需引入UAV进行负载分担，使第k个BS的业务量满足如下约束：

$$0\leqslant T_k^{\mathrm{BS}}\leqslant \epsilon T_{\max}^{\mathrm{BS}} \tag{9}$$

其中，ϵ为负载系数，且$\epsilon\in[0,1]$。在不失一般性的前提下，将GU业务量按降序排列。UAV优先承担业务量最高的GU，以快速缓解BS负载。设$\mathcal{M}_k^{\mathrm{off}}\subset\mathcal{M}$为从BS k转移至UAV i的用户集合，则UAV i从BS k承接的总卸载业务量为：

$$T_i^{-}=\sum_{m\in\mathcal{M}_k^{\mathrm{off}}}T_{k,m}^{\mathrm{BS}} \tag{10}$$

基于BS覆盖盲区的GU业务量T_i^{+}、BS服务范围内需要转移的GU业务量T_i^{-}和单个UAV可服务的最大业务容量$T_{\max}^{\mathrm{U}}$，UAV数量N_{U}可通过式(11)计算。

$$N_{\mathrm{U}}=\left\lceil\frac{T_i^{+}+T_i^{-}}{T_{\max}^{\mathrm{U}}}\right\rceil+o \tag{11}$$

其中，$\lceil\cdot\rceil$表示向上取整；o表示冗余因子，用于应对用户分布复杂性带来的不确定性。根据4.2节的仿真结果，下文这一冗余因子取值为$o=2$。

本文的目标是通过部署最少数量的UAV，同时满足两方面需求：一是优先服务现有BS覆盖盲区的GU；二是减轻现有BS的负载。具体而言，需在最大化所有UAV覆盖用户的总业务量的同时，高效平衡地面BS之间的负载。由于每架UAV独立执行任务，系统总业务量的最大化等价于最大化所有UAV个体服务量的加权和。该优化问题可建模为：

$$\begin{aligned}
&\max_{\{a_{i,t}\}_{i=1,t=0}^{I,T-1}}\ J(T;\{a_{i,t}\})=\\
&\sum_{i=1}^{I}\left\{\omega_1 T_i^{+}(T;\{a_{i,t}\})+\omega_2 T_i^{-}(T;\{a_{i,t}\})\right\}\\
&\text{s.t. C1:}\begin{cases}x_{i,t+1}=x_{i,t}+v_{i,t}\Delta t\mathbf{1}_{\{a_{i,t}=1\}}-v_{i,t}\Delta t\mathbf{1}_{\{a_{i,t}=3\}}\\ y_{i,t+1}=y_{i,t}+v_{i,t}\Delta t\mathbf{1}_{\{a_{i,t}=4\}}-v_{i,t}\Delta t\mathbf{1}_{\{a_{i,t}=2\}}\\ \left(x_{i,t}^{\mathrm{U}},y_{i,t}^{\mathrm{U}}\right)\in[0,L]^2,\forall i,t\end{cases}\\
&\quad\text{C2:}\omega_1+\omega_2=1,\omega_1,\omega_2\in(0,1)\\
&\quad\text{C3:}\Gamma_{i,m}\geqslant\Theta_{\mathrm{th}},\forall i\in\mathcal{I},\forall m\in\mathcal{M}\\
&\quad\text{C4:}W_k(t_w)\leqslant 2\%,\forall k\in\mathcal{K}\\
&\quad\text{C5:}T_i^{+}(T)+T_i^{-}(T)\leqslant T_{\max}^{\mathrm{U}},\forall i\in\mathcal{I}
\end{aligned} \tag{12}$$

其中，优化的决策变量$\{a_{i,t}\}$为考虑时间范围内所有UAV在每个时间步t的动作，用于引导UAV从初始位置移动至能最大化总体用户业务量的位置；$T_i^{+}(T)$和$T_i^{-}(T)$分别表示在动作序列$\{a_{i,t}\}_{t-1}^{T}$引导下，UAV i在终端时刻T累计服务的盲区用户业务量和卸载业务量；约束C1为UAV的实时位置及位置约束，其中，$v_{i,t}$表示UAV i在时间步t的运动速度，Δt表示相邻两个时间步的时间差；约束C2中的ω_1和ω_2为两个不同优化目标的权重；约束C3代表通信信噪比的阈值要求；约束C4是GU k的等待概率要求；约束C5为BS与UAV的资源容量限制。

3　基于干扰管理的MAD3QL-TL算法

为解决式(12)描述的大规模优化问题，本文将其转化为有限时域马尔可夫决策过程（Markov decision process, MDP）。基于MDP框架，UAV可在每个时间步t根据自身当前观测（即位置信息）动态选择最优动作，以最大化长期平均奖励。本节首先建立优化问题与MDP之间的转化关系，随后依次介绍MAD3QL-TL算法框架、迁移学习策略以及干扰管理频谱分配算法。

3.1 问题建模与MDP转化

式(12)所示的优化问题是一个有限时域内的终端性能指标最大化问题，其目标函数仅在部署过程结束时刻T被评估。直接求解该问题面临两个挑战：一是动作空间随UAV数量I和时域长度T呈指数增长；二是目标函数与动作序列之间的映射高度非线性且难以用解析式表达。

为此，本文将该优化问题转化为一个有限时域MDP。该MDP框架由五元组$(S, A, \boldsymbol{P}, r, \gamma)$定义，其中$S$为状态空间，$A$为动作空间，$\boldsymbol{P}$为状态转移概率矩阵，$r$为即时奖励函数，$\gamma$为折扣因子。通过DRL求解该MDP的最优策略，在方向上逼近原优化问题的最优解。

需要指出的是，尽管优化问题式(12)是一个全局合作性的目标函数（所有UAV的贡献加权求和），但该问题不能简单地通过单智能体RL解决。首先，每个UAV仅能获取其自身位置及覆盖范围内的用户信息，无法获知其他UAV的精确位置和全局用户分布。其次，在实际部署中，UAV之间无法依赖持续稳定的高速通信链路进行实时集中控制，各UAV必须基于本地信息独立作出部署决策。最后，单智能体方法的动作空间随UAV数量I呈指数增长，当I变化时需要重新训练整个网络。

因此，本文将上述问题建模为完全合作的多智能体部分可观测MDP，采用集中式训练和分布式执行的多智能体深度强化学习框架进行求解。在集中式训练阶段，各UAV可共享全局信息以加速收敛；在分布式执行阶段，各UAV仅依赖本地观测进行独立决策，从而满足实际部署需求。下面分别定义MDP的各要素。

（1）状态空间S

UAV i的状态空间S_i定义如下：首先，将所有UAV的可能目标区域离散为$N \times N$的等间隔网格，UAV的位置被映射至网格交点；同时，将时间范围T划分为若干时间步t，UAV在每个时间步t独立作出时序决策。据此，UAV i在时间步t的二维位置可被定义为其局部状态信息。

$$s_{i,t} = \left(x_{i,t}, y_{i,t}\right) \tag{13}$$

其中，$i \in \mathcal{I}, t \in [1,2,\cdots,\mathcal{T}]$。

（2）动作空间A

在任意网格交叉点处，单架UAV有5种水平运动可能：悬停、向右、向后、向左和向前，分别用0、1、2、3和4表示。UAV在移动时将跨越一个网格长度，因此其动作空间可表示为A，且$A = \{0,1,2,3,4\}$。UAV i在每个时间步t选择的动作值，将在整个时间范围内控制各UAV的动作，并相应改变它们的状态。

（3）状态转移概率矩阵$\boldsymbol{P}$

设UAV i在时间步t的状态$s_{i,t}$执行动作$a_{i,t}$转移到下一状态$s_{i,t+1}$的概率为$p(s_{i,t+1}|s_{i,t},a_{i,t})$，则UAV i的状态转移概率矩阵可表示为：

$$\boldsymbol{P}_i = \begin{pmatrix} p(s_t|s_t,a_{i,t}) & p(s_{t+1}|s_t,a_{i,t}) & \cdots & p(s_{t+N^2}|s_t,a_{i,t}) \\ p(s_t|s_{t+1},a_{i,t}) & p(s_{t+1}|s_{t+1},a_{i,t}) & \cdots & p(s_{t+N^2}|s_{t+1},a_{i,t}) \\ \vdots & \vdots & \ddots & \vdots \\ p(s_t|s_{t+N^2},a_{i,t}) & p(s_{t+1}|s_{t+N^2},a_{i,t}) & \cdots & p(s_{t+N^2}|s_{t+N^2},a_{i,t}) \end{pmatrix} \tag{14}$$

（4）即时奖励函数

在每个时间步t，UAV i根据当前状态$s_{i,t}$选择并执行动作$a_{i,t}$后，环境会反馈一个即时奖励$r_{i,t}$。为引导UAV在最大化终端业务量的同时满足安全约束，本文将即时奖励函数设计为每个时间步的增量贡献，UAV i在时间步t的即时奖励$r_{i,t}$定义为：

$$r_{i,t} = \begin{cases} \dfrac{\lambda_1 \Delta T_{i,t}^{+} + \lambda_2 \Delta T_{i,t}^{-}}{T_{\max}^{\mathrm{U}}}, & t = T \\ \dfrac{\lambda_1 \Delta T_{i,t}^{+} + \lambda_2 \Delta T_{i,t}^{-}}{T_{\max}^{\mathrm{U}}} - (p_{i,t}^{\mathrm{distance}} + p_{i,t}), & t < T \end{cases} \tag{15}$$

其中，λ_1和λ_2为奖励权重系数，用于调节MDP训练过程中即时奖励的敏感度，其值在表1中定义（本文通过网格搜索确定）。值得注意的是，这两个权重系数与优化问题式(12)中的权重系数ω_1和ω_2（定义终端性能评估的偏好）具有不同的功能。由于RL训练过程中即时奖励的稀疏性和延迟回报特性，使累计奖励的优化方向与终端目标函数的方向一致即可，两者数值不必相等。$\Delta T_{i,t}^{+} = T_{i,t}^{+} - T_{i,t-1}^{+}$和$\Delta T_{i,t}^{-} = T_{i,t}^{-} - T_{i,t-1}^{-}$分别为UAV i在时间步t新增的盲区覆盖业务量和负载卸载业务量；$T_{i,t}^{+}$和$T_{i,t}^{-}$为截止时间步t的累计服务量；$T_{\max}^{\mathrm{U}}$为单架UAV的最大

业务容量，用于归一化；$p_{i,t}$为UAV i在时间步t的边界惩罚，$p_{i,t}^{\text{distance}}$为UAV i在时间步t的距离惩罚函数（用于防止UAV之间发生碰撞），该函数可表示为：

$$p_{i,t}^{\text{distance}}=\begin{cases}\mu_1\sum\limits_{j=1,j\neq i}^{I}\left(1-\dfrac{d_{i,j}^t}{d_{\min}}\right), & d_{i,j}^t\leqslant d_{\min}\\ \mu_2\sum\limits_{j=1,j\neq i}^{I}\left(1-\dfrac{d_{i,j}^t-d_{\min}}{1.5r_i^{\mathrm{U}}-d_{\min}}\right), & d_{\min}<d_{i,j}^t\leqslant 1.5r_i^{\mathrm{U}}\\ 0, \quad \text{其他}\end{cases} \tag{16}$$

其中，μ_1和μ_2分别为碰撞权重系数与重叠权重系数，$d_{i,j}^t$为在时间步t时UAV i与UAV j之间的欧氏距离，$d_{\min}$为预设的安全距离阈值。当$d_{i,j}^t\leqslant d_{\min}$时，系统将施加与距离成反比的惩罚；当$d_{\min}<d_{i,j}^t\leqslant 1.5r_i^{\mathrm{U}}$时，施加较小的惩罚；其余情况则无惩罚。

将式(15)所示的即时奖励沿时间轴累加，UAV i的累计折扣奖励为：

$$\sum_{t=0}^{T}\gamma^t r_{i,t}=\frac{1}{T_{\max}^{\mathrm{U}}}\sum_{t=0}^{T}\gamma^t\left[\lambda_1\Delta T_{i,t}^{+}+\lambda_2\Delta T_{i,t}^{-}\right]=\frac{\gamma^t\left[\lambda_1 T_i^{+}(T)+\lambda_2 T_i^{-}(T)\right]}{T_{\max}^{\mathrm{U}}} \tag{17}$$

当折扣因子$\gamma=1$，惩罚项$p_{i,t}^{\text{distance}}=p_{i,t}=0$（即UAV始终满足安全约束）时，式(17)严格等于式(12)的归一化目标函数。当惩罚项不为零时，距离和边界惩罚作为软约束，引导UAV在提升服务性能的同时满足安全约束。当惩罚系数足够大时，违规行为的代价超过其可能带来的业务量增益，最优策略将趋于完全合规，MDP最优解逼近原问题最优解。因此，在RL框架下最大化累计折扣奖励，等价于求解原优化问题。

3.2　基于IMSA的MAD3QL-TL算法

本文提出的基于IMSA的MAD3QL-TL算法整体结构如图2所示。该算法由多智能体决策、TL和IMSA三部分组成。在训练阶段，各UAV基于本地观测输出动作；TL模块将源域预训练知识迁移至目标域，以加速收敛；同时，干扰管理器根据UAV的空间分布动态分配频谱资源，从而抑制同频干扰。

（1）MAD3QL算法

基于图2构建的模型，本节首先介绍MAD3QL算法。

在RL中，动作价值函数（即Q值）定义为在状态s执行动作a后能获得的期望累积折扣奖励。为提高学习过程的稳定性和效率，Dueling Deep Q Learning架构（如图3所示）将Q网络分解为两个并行计算流[29]，可同时估计状态价值函数$\mathcal{V}(s;\xi)$与优势价值函数$\mathcal{D}(s,a;\psi)$。其中，$\mathcal{V}(s;\xi)$表示处

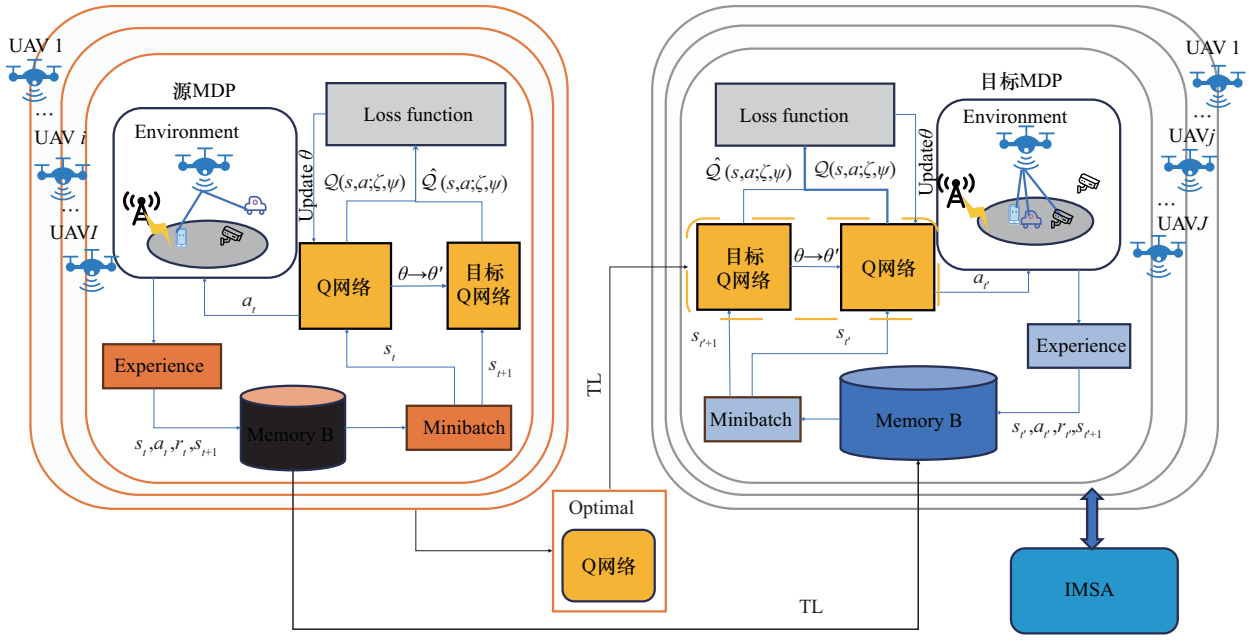


图2　基于IMSA的MAD3QL-TL算法架构

于状态 s 时能获得的期望回报，而 $\mathcal{D}(s,a;\psi)$ 表示在状态 s 下选择动作 a 相较于平均水平具有的优势。最终得到 Q 值为：

$$Q(s,a;\xi,\psi)=\mathcal{V}(s;\xi)+\left(\mathcal{D}(s,a;\psi)-\frac{1}{|\mathcal{A}|}\sum_{a'}\mathcal{D}(s,a';\xi)\right) \tag{18}$$

其中，ξ 和 ψ 分别为价值流与优势流的参数，$|\mathcal{A}|$ 为动作空间大小。

Double Deep Q Learning 机制[30]采用两个基于 Dueling 网络的估计器：一个是用于动作选择的 Q 网络 $Q(s,a;\xi,\psi)$；另一个是用于动作评估的目标 Q 网络 $\hat{Q}(s,a;\xi,\psi)$。因此，智能体 i 在时间步 t 的目标 Q 值 $G_{i,t}$ 可定义为：

$$G_{i,t}=r_{i,t}+\gamma\hat{Q}\left(s_{i,t+1},\arg\max_{a}Q(s_{i,t+1},a;\xi_t,\psi_t);\xi_t^-,\psi_t^-\right) \tag{19}$$

其中，$r_{i,t}$ 为智能体 i 在时间步 t 获得的即时奖励，$\gamma\in[0,1]$ 为折扣因子；ξ 和 ψ 为 Q 网络的参数，ξ^- 和 ψ^- 为目标 Q 网络的参数。

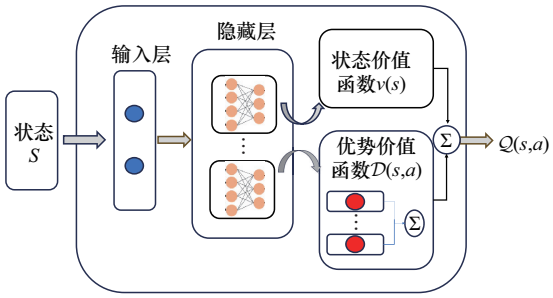


图 3　Dueling Deep Q Learning 架构

为了方便起见，设 ϕ_t 和 ϕ_t^- 分别表示时间步 t 时 Q 网络和目标 Q 网络的参数。Q 网络的训练目标为最小化其输出 Q 值与目标 Q 值之间的差异，因此可将智能体 i 在时间步 t 的损失函数 $\mathcal{L}_{i,t}(\phi_t)$ 定义为：

$$\mathcal{L}_{i,t}(\phi_t)=\mathbb{E}_{(s,a,r,s')}\left[\left(G_{i,t}-Q(s,a;\phi_t)\right)^2\right] \tag{20}$$

其中，(s,a,r,s') 表示经验池 $\boldsymbol{B}$ 中用于训练 Q 网络的数据点。

在此基础上，每个时间步 t 智能体 i 的成本函数 $\mathcal{J}_i(t)$ 可定义为：

$$\mathcal{J}_i(t)=\frac{1}{|\boldsymbol{B}|}\sum_{(s,a,r,s')\in\boldsymbol{B}}\mathcal{L}_{i,t}(\phi_t) \tag{21}$$

其中，$\mathcal{L}_{i,t}$ 为时间步 t 的损失函数，ϕ_t 为 Q 网络参数。然后，通过 Adam 优化器[31]更新 Q 网络的参数。

$$\phi_{t+1}=\phi_t-\alpha_t\mathrm{Adam}(\nabla_{\phi_t}\mathcal{J}_t(\phi_t)) \tag{22}$$

其中，α_t 表示时间步 t 的步长，用于决定 Q 网络参数更新的幅度；$\nabla_{\phi_t}(\cdot)$ 表示损失函数关于 Q 网络参数的梯度。

目标 Q 网络参数 ϕ_t^- 则采用软更新策略增强训练稳定性。

$$\phi_{t+1}^-=\tau\phi_t+(1-\tau)\phi_t^- \tag{23}$$

其中，$\tau\in(0,1)$ 为软更新速率，控制目标 Q 网络参数的更新幅度。

（2）TL 策略

当 GU 的业务量需求发生变化时，所需的 UAV 数量也会相应变化，DNN 可能需要从头重新训练，导致计算复杂度较高。为此，本文采用混合 TL 策略[24]，将源任务中积累的知识迁移至目标域，以提升目标任务的学习性能，具体执行流程如算法 1 所示。

设源域经验集合为：

$$E_S=\{e_k|e_k=(s_k,a_k,r_k,s_k')\} \tag{24}$$

其中，每个经验元组来自源 MDP，目标 UAV 的经验池为 $\mathcal{D}_T$。经验迁移过程可表示为：

$$\mathcal{D}_T\leftarrow\mathcal{D}_T\cup E_S \tag{25}$$

源域 UAV 的 Q 网络参数为 ϕ_s，通过策略迁移目标域 UAV 的 Q 网络参数 ϕ_T 初始化为：

$$\phi_T^{(0)}=\phi_s \tag{26}$$

该策略复用源 MDP 中积累的历史经验样本 E_S 参与目标 UAV 在目标 MDP 中的学习过程，同时，将源域 UAV 学习得到的策略参数 ϕ_s 迁移至目标 UAV 并将其作为策略网络初始参数 $\phi_T^{(0)}$。

（3）IMSA 算法

设所有 UAV 具有相同的覆盖半径 r，若 UAV i 和 UAV j 的覆盖圆在地面的投影存在重叠，即 $d_{i,j}<2r$，则二者不能分配相同频段；否则可复用同一频段。令 $\mathcal{V}=\{1,2,\cdots,N_v\}$ 表示由正整数构成的频段标签集合，每个标签代表一个正交的、互不干扰的通信频段。IMSA 算法的目标是为每个 UAV i 分配一个频段标签 $V_i\in\mathcal{V}$，使任意两个覆盖圆重叠的 UAV i 和 UAV j 满足 $V_i\neq V_j$，同时最小化使用的不同通信频段总数 N_v。IMSA 算法的执行步骤如算法 2 所示。

算法 1　MAD3QL-TL

输入　s_t

输出　a_t

1)建立经验池 $\boldsymbol{B}$，初始化初始探索率 ε

2)初始化 Q 网络参数 ϕ_i 和目标 Q 网络参数 $\phi_i^- = \phi_i$

3)根据式(26)迁移源域 Q 网络参数

4)根据式(25)迁移源域经验池

5)for step = 1 to T do

6)　初始化所有 UAV 的状态 $s_{i,0}, \forall i \in \mathcal{I}$

7)　使用 IMSA 算法（算法 2）更新 UAV 的频谱资源

8)　for all $i \in \mathcal{I}$ do

9)　　for n=1 to N do

10)　　根据贪婪策略选择动作 $a_{i,t}$

11)　　执行动作 $a_{i,t}$，获得奖励 $r_{i,t}$ 与下一时刻状态 $s_{i,t+1}$

12)　　保存经验 $(s_{i,t}, a_{i,t}, r_{i,t}, s_{i,t+1})$ 至经验池 $\boldsymbol{B}$

13)　　从经验池 $\boldsymbol{B}$ 中小批次采样

14)　　分别通过式(18)和式(19)计算 Q 值和目标 Q 值 $G_{i,t}$

15)　　根据式(22)更新 Q 网络参数

16)　　ε 值减小

17)　　end for

18)　　根据式(23)软更新目标 Q 网络参数

19)　end for

20)end for

算法 2　IMSA

输入　I，s_t，r

输出　$\{V_1, V_2, \cdots, V_I\}$，$N_v$

1)初始化 UAV 频谱资源分配 $\{V_1, V_2, \cdots, V_I\}$

2)for all $i \in \mathcal{I}$ do

3)　计算 UAV i 冲突度 g_i

4)end for

5)降序排列所有 UAV 的冲突度获得序列 ι

6)for $n \in \iota$ do

7)　$\mathcal{S}_i \leftarrow \{V_j | c_{i,j} = 1\}$

8)　$V_i \leftarrow \min\{v \geqslant 1 | v \in \mathcal{V}, v \notin \mathcal{S}_i\}$

9)end for

10)$N_v \leftarrow \max V_i$

11)return $\{V_1, V_2, \cdots, V_I\}, N_v$

首先，计算所有 UAV 之间的圆心距，标记冲突关系为：

$$c_{i,j} = \begin{cases} 1, d_{i,j} < 2r \\ 0, \ \text{其他} \end{cases} \tag{27}$$

并计算每个 UAV i 的冲突度 $g_i = \sum_{j \neq i} c_{i,j}$。然后，将 UAV 按冲突度降序排列，得到序列 ι。按照 ι 的顺序依次为每个 UAV 分配频段：对于 UAV i，设 $\mathcal{S}_i$ 为冲突邻居（即覆盖范围重叠的 UAV）占用的频段集合，则分配 $V_i = \min\{v \geqslant 1 | v \in \mathcal{V}, v \notin \mathcal{S}_i\}$。最后，所用频段总数为 $N_v = \max V_i$。

4　仿真分析

本文采用 Python 3.12 作为开发环境，并基于 PyTorch 框架构建 DRL 模型，以支持系统建模与优化。设定系统场景为 1 km×1 km 特定区域，将其划分为边长为 20 m 的单元格。部署 3 个 BS 提供通信服务，GU 围绕两个热点中心聚集。针对 UAV 辅助移动网络的信道建模[32]，关键环境参数设置如下：式(1)中的环境参数 $\alpha = 9.61, \beta = 0.16$；LoS 和 NLoS 条件下的平均附加路径损耗分别为 $\eta_{\text{LoS}} = 1$ dB，$\eta_{\text{NLoS}} = 20$ dB。其余关键系统参数如表 1 所示。

表 1　　系统参数设置

参数	取值
飞行高度 h/m	100
UAV 仰角 θ	51.34°
UAV 发射功率 p/W	1
载波频率 f/GHz	2
高斯白噪声功率 P_{gw}/W	3.98×10^{-15}
紧迫性 ϑ_m	[1-3]
吞吐量需求 R'_m	[0.1,1,2]
BS 的最大业务容量 $T_{\max}^{\text{BS}}$/erl	100
UAV 的最大业务容量 $T_{\max}^{\text{U}}$/erl	40
信噪比阈值 Θ_{th}/dB	1
负载系数 ϵ	0.8
优化目标权重系数 ω_1, ω_2	0.5, 0.5
奖励权重系数 λ_1, λ_2	0.732 2, 0.267 8
距离惩罚系数 μ_1, μ_2	0.9, 0.1

本文采用 DRL 算法进行 UAV 决策学习训练，相关的网络参数如表 2 所示。

表 2　　相关网络参数说明

参数	取值	说明
隐藏层结构	[256,128,64]	三层全连接层，神经元数分别为256，128和64
激活函数	LeakyReLU	非线性拟合
学习率	3×10^{-5}	Adam学习率
目标Q网络软更新系数	0.01	目标Q网络缓慢跟随Q网络
经验池容量	5×10^{6}	存储历史经验的最大容量
批量采样大小	128	每次更新的样本数量
折扣因子	0.98	平衡即时奖励与长期奖励
初始探索率	1.0	ε-贪心策略初始值
探索率衰减系数	0.998	每轮迭代的探索率衰减比例

4.1　敏感性分析

图4仿真了MAD3QL-TL算法的性能随权重大小的变化情况。由图4可知，随着ω_1的增大，UAV对覆盖盲区的关注程度提高，GU等待概率逐渐上升，而未覆盖用户数则呈现先下降后上升的趋势。这是因为当ω_1较大时，UAV更倾向于探索覆盖盲区内的GU，从而盲区用户数量减少。然而，当ω_1继续增大时，受用户分布复杂性的影响，UAV难以在探索盲区用户与卸载现有BS过载的业务量之间取得平衡，导致未覆盖用户数量和等待概率均随之增加。同时，当ω_1超过0.6后，用户的等待概率大于系统的等待概率阈值（2%）。因此，本文优化目标的权重取值为ω_1=ω_2=0.5。

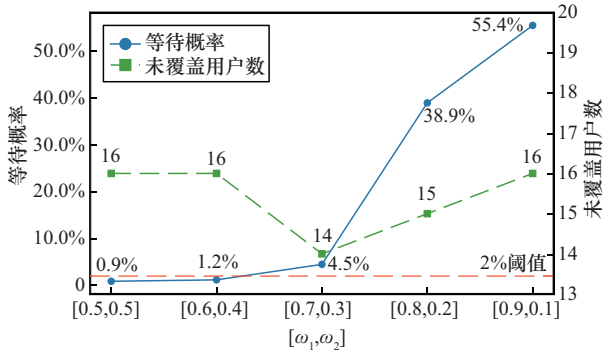


图4　敏感性分析

4.2　系统性能验证

图5验证了在UAV最大业务容量为40的条件下，不同UAV数量对系统性能的影响。在所有仿真实验中，MAD3QL-TL算法针对不同数量的UAV均迁移了源域（UAV最大业务容量为40、数量为8）的经验与策略。仿真结果分别从平均奖励、用户覆盖率、等待概率和用户满意度4个维度展示了不同算法的性能随UAV数量的变化情况。

由图5(a)可知，系统平均奖励均随着UAV数量的增多而减少。但是，MAD3QL-TL算法的平均奖励均高于MAD3QL、MADDPG和MADQL算法。

图5(b)和图5(c)分别展示了用户覆盖率与等待概率的变化情况。由图5(b)和图5(c)可知，当o = 2（对应UAV数量为8）时，用户的等待概率已降至2%以下，用户覆盖率达到92%，满足了用户的QoS要求；当继续增加UAV数量时，对系统的整体性能提升有限，考虑系统性能与资源利用效率，下文中的仿真中均选取o = 2。由图5(b)和图5(c)还可以看出，得益于源域经验与策略的迁移，当UAV数量较少时，尽管等待概率尚未完全满足QoS要求，但本文算法在这两项指标上仍优于其他对比算法。随着UAV数量的增加，用户覆盖率逐步提升，同时等待概率始终稳定在2%的阈值以下，有效缓解了网络拥塞。相比之下，对比算法的用户覆盖率明显偏低，且波动较大，部分场景下无法满足QoS要求。

图5(d)展示了用户满意度的变化趋势。可以看出，随着UAV数量的增加，用户满意度逐步提高，进一步验证了本文算法在提升用户体验方面的有效性。同时，MAD3QL-TL算法的用户满意度性能同样优于其他3种对比算法。此外，与对比算法相比，MAD3QL-TL算法的表现更为稳定，体现了TL机制在不同配置下的良好泛化能力。综合4项指标可知，MAD3QL-TL算法性能最优，其在保证稳定性的同时，实现了最高用户覆盖率、最低等待概率和最佳用户满意度。

图6验证了在UAV数量为8的情况下，UAV最大业务容量对系统性能的影响。在所有仿真实验

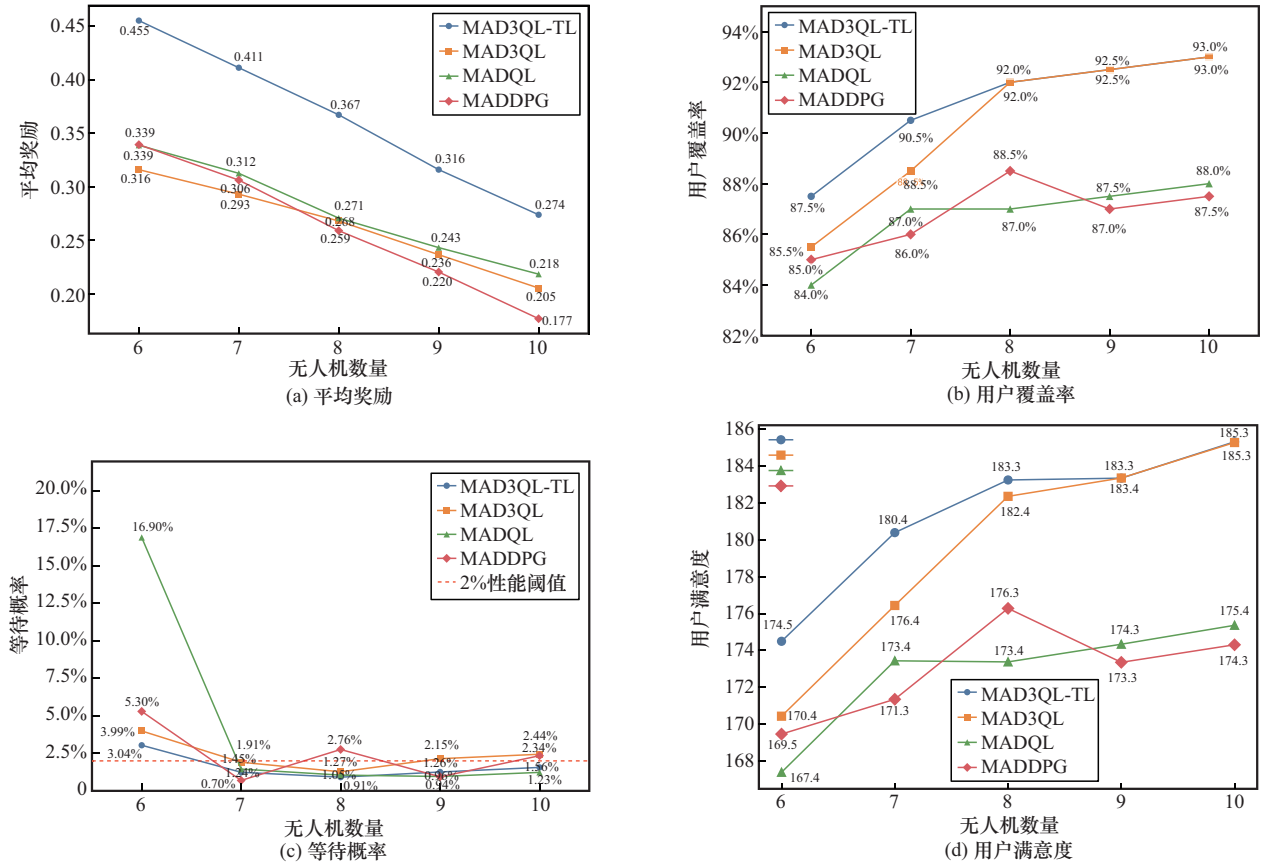


图5　不同UAV数量对系统性能的影响

中，MAD3QL-TL算法针对所有最大业务容量值均迁移了源域（UAV最大业务容量为40、数量为8）的经验与策略。

由图6(a)可知，随着UAV最大业务容量的增加，系统平均奖励均随之减少。在所有业务容量配置下，MAD3QL-TL算法的平均奖励均高于MAD3QL、MADDPG和MADQL算法。

图6(b)、图6(c)和图6(d)分别展示了用户覆盖率、等待概率和用户满意度的变化情况。在初始阶段（即UAV最大业务容量较低时），得益于MAD3QL-TL算法对源域经验与策略的TL，该算法始终保持较高的用户覆盖率，用户满意度也高于其他对比算法。然而，此时等待概率处于较高水平，尚未完全满足QoS要求。随着UAV最大业务容量的增加，UAV的业务承载能力逐步增强，本文算法在继续保持高用户覆盖率的同时，等待概率最终降至2%的阈值以下，满足了QoS要求。

值得注意的是，在整个变化过程中，用户覆盖率始终保持不变。这是因为本文优先保证BS覆盖盲区用户的QoS需求，在此基础上再为高业务负载的BS卸载业务量，从而确保了用户覆盖率指标的稳定性。

用户满意度的变化趋势与用户覆盖率呈明显正相关，这表明UAV提供的服务能够满足大部分用户的吞吐量需求，进一步验证了本文算法在提升用户体验方面的有效性。在对比算法方面，MAD3QL-TL算法在用户覆盖率、用户满意度和等待概率上均优于MADDPG、MADQL和DQL算法。同时，图6再次表明，MAD3QL-TL算法的表现比对比算法更加稳定，体现了TL机制在不同业务容量配置下的良好泛化能力。

4.3　算法收敛性分析

本节对比了不同算法的收敛性能，实验结果如图7所示。仿真结果表明，本文算法在学习效率方面具有显著优势。该算法在大约100次迭代后便呈现出明显的收敛趋势，而其他对比算法需要约1 500次迭代才能达到相近的性能水平。这种快速收敛特性主要归功于算法中引入的TL机制，通过复用源任务中已学习的经验和策略，有效避免了从零开始

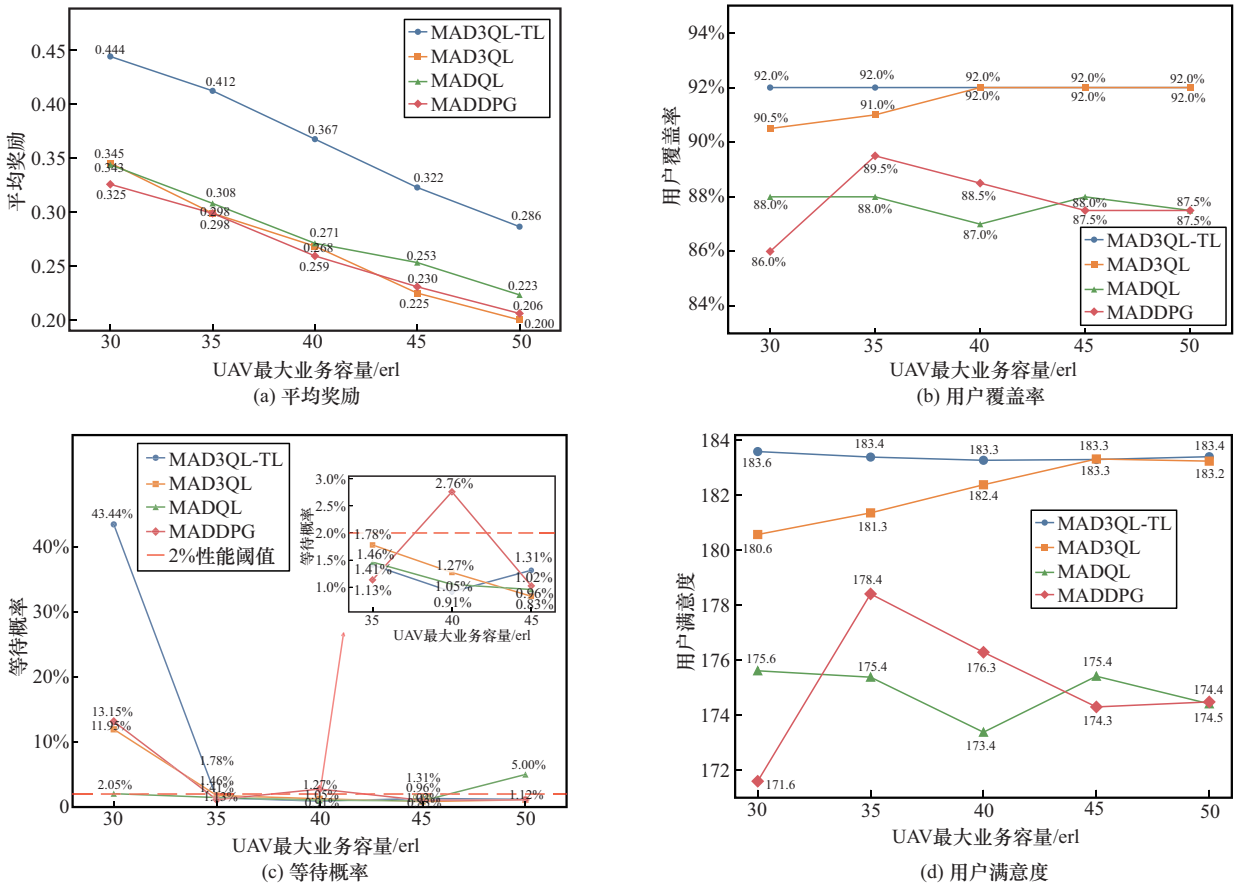


图 6　UAV 最大业务容量对系统性能的影响

探索的过程，从而显著加速了训练进程。从收敛曲线的最终性能还可以看出，MAD3QL-TL 算法在收敛速度和稳定性方面均优于对比算法，其稳定后的平均奖励值达到 0.38，相较于 MADDPG 和 MADQL 算法，分别提高了约 40% 和 90%。这一结果充分验证了该算法在解决复杂部署问题上的优越性。

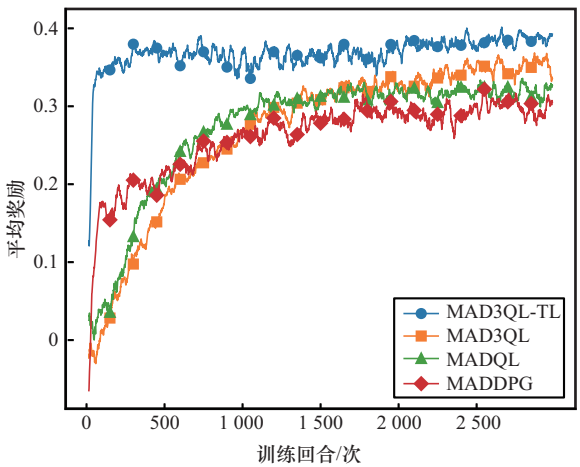


图 7　算法的收敛速度

4.4　网络部署结果分析

图 8 展示了 MAD3QL-TL 算法与其他对比算法的最终部署方案。在图 8(a)中，3 个加粗圆点表示地面 BS，虚线圆圈表示地面 BS 的覆盖范围，其内部的小圆点表示由该 BS 服务的 GU。加粗三角形表示 UAV，以其作为中心的虚线圆圈表示 UAV 的通信覆盖区域，其中，小三角形标示了位于 BS 覆盖盲区内由 UAV 提供服务的用户。特别值得注意的是，图 8 中以方形标记的部分 GU，原本处于地面 BS 的覆盖范围之内，但由于 BS 的负载限制，被智能转移至 UAV 提供服务。MAD3QL-TL 算法实现了 92% 的用户覆盖率，高于 MADQL（87%）和 MADDPG（88.5%）。BS 和 UAV 分别服务 55% 和 37% 的 GU，其中，由 BS 转移至 UAV 服务的 GU 约占用户总数的 9.5%。尽管本文算法与 MAD3QL 算法的最终部署策略无明显区别，但这恰恰反映了其高效的学习机制：通过成功迁移 MAD3QL 算法的经验与策略，本文算法能够以更快的速度收敛至其最优性能，在保证性能上限的同时，大幅降低了训练成本。

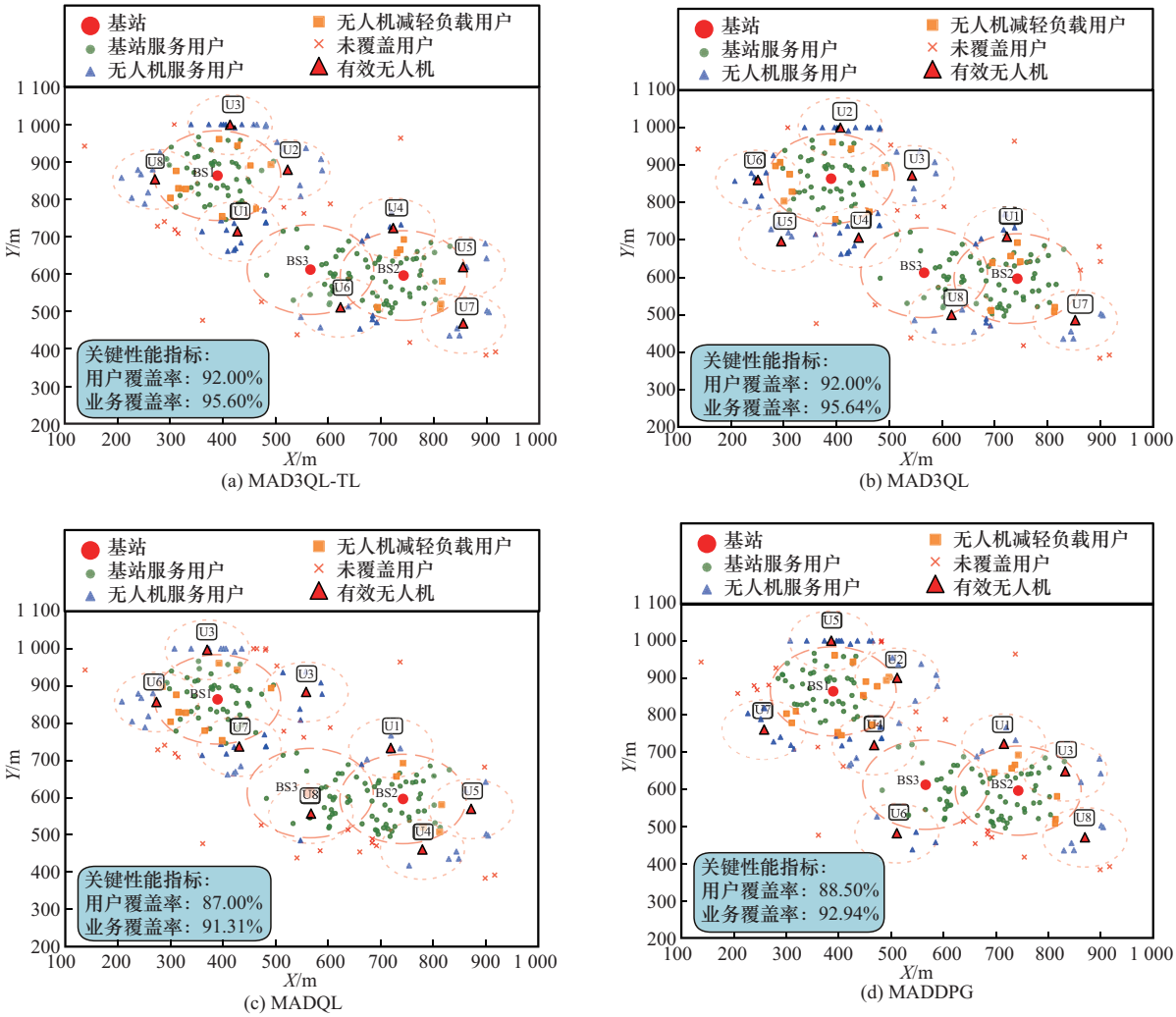


图8　UAV部署结果

表3给出几种算法的QoS对比结果，由表3可知，MAD3QL-TL算法在综合性能上最优。经过智能部署后，覆盖盲区用户数从初始的73大幅降至16，降幅达78.1%。MAD3QL-TL和MAD3QL算法均为16，优于MADQL（26）和MADDPG（23）。这表明本文算法能够更有效地为地面BS无法覆盖的用户提供通信服务，实现了“盲区优先覆盖”的设计目标。同时，负载转移机制使用户等待概率从79.50%大幅降至0.9%，显著缓解了网络拥塞。用户满意度从127.00提升至183.25，提升约44.3%。3项指标均显著优于其他对比算法，验证了本文算法在解决覆盖不足与网络拥塞问题上的有效性。上述结果充分证明了本文算法能够有效实现负载均衡，并体现了分层协同部署策略在复杂环境下的适应能力。

表3　　QoS对比

算法	覆盖盲区用户数		用户等待概率		用户满意度	
	优化前	优化后	优化前	优化后	优化前	优化后
MAD3QL-TL	73	16	79.5%	0.9%	127.00	183.25
MAD3QL		16		1.2%		182.36
MADQL		26		1.1%		173.37
MADDPG		23		2.6%		176.27

4.5　负载均衡效果评估

图9进一步揭示了本文算法在负载优化方面的成效。由图9(a)可知，经过MAD3QL-TL算法的智能调度，高负载BS（BS 1和BS 2）的业务量得到了显著削减，降幅分别达到37.2%和23.6%。由图9(b)可以看出，两个高负载BS（BS 1和BS 2）的等待概率均下降至2%以下，满足了QoS要求，有效缓解了网

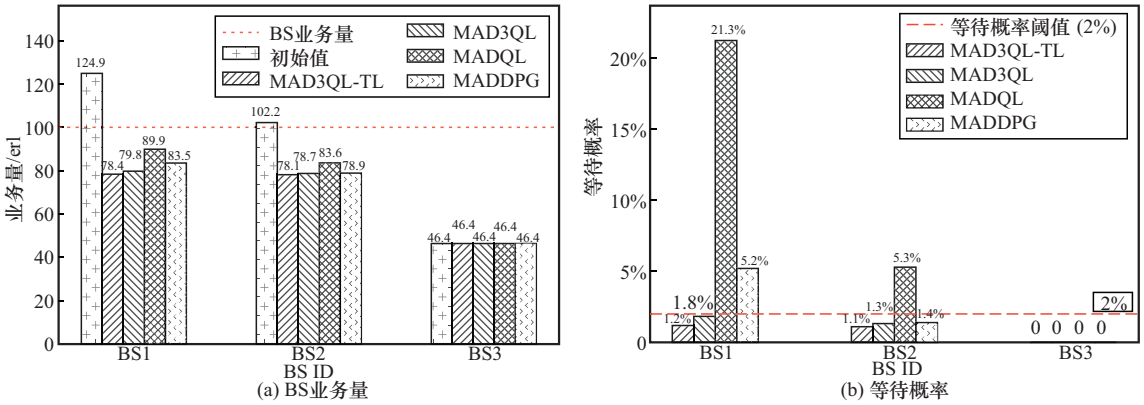


图9　BS 负载优化结果

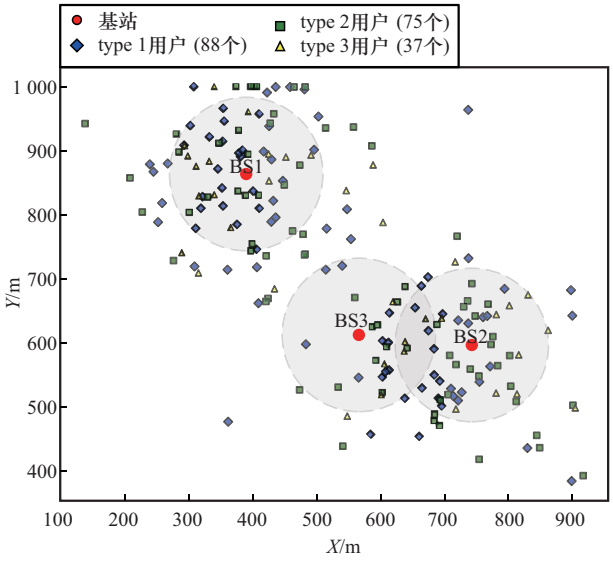
络拥塞状况。原本负载较低的BS 3则保持稳定运行状态，未受到调度策略的负面影响。由图9还可以看出，MAD3QL-TL和MAD3QL算法在负载均衡效果上显著优于MADQL和MADDPG，MADQL和MADDPG算法对高负载BS的业务量削减幅度较小，且等待概率仍维持在较高水平，未能完全满足QoS要求。本文算法通过差异化的智能调度策略，重点削减高负载BS的压力，同时避免对低负载BS进行不必要的调整，既保障了网络热点区域的QoS、提升了网络速率，又有效避免了资源浪费。

4.6　迁移学习性能验证

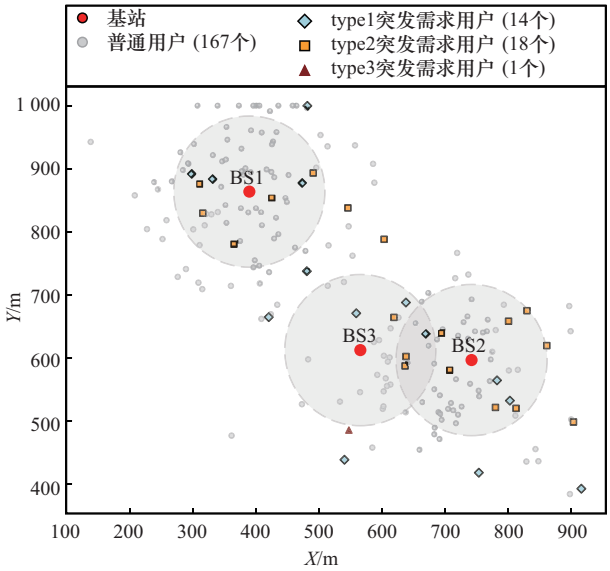
本节进一步评估了MAD3QL-TL算法在环境信息未知场景下基于用户业务需求的UAV部署方法的性能，同时给出了UAV在不同目标任务场景下（中等突发需求、高等突发需求和BS损毁场景）的学习效率与适应能力。图10(a)展示了基础场景下的用户分布情况，加粗圆点图标表示部署的3个BS，以其为中心的圆形区域表示其覆盖范围。菱形、方形和圆形分别表示3种业务量需求从小到大的任务类型，清晰显示了用户需求的空间异质性。图10(b)和图10(c)呈现了不同数量用户在突发业务场景下的用户分布，分别有33个用户和63个用户有突发业务需求。小圆点表示无突发业务需求的普通用户，菱形、方形和圆形分别表示3种类型的突发业务需求用户。图10(d)展示了在现有BS损坏的情况下，大量用户无法得到通信服务。

当用户业务模式发生变化（如突发业务需求或BS损坏）时，传统RL算法通常需要完全重新训练，整个过程不仅耗时，且收敛速度缓慢。图11验证了MAD3QL-TL算法在动态环境下的快速适应能力。在上述3种动态场景中，MAD3QL-TL算法均能有效复用源域中已习得的知识，仅需约100次迭代即可快速适应新环境。相比之下，传统算法重新训练所需的迭代次数远超这一水平。从训练效率来看，MAD3QL-TL算法相较于完全重新训练节省了约85%的学习时间，充分体现了TL机制在降低训练成本方面的显著优势。从收敛曲线的表现来看，MAD3QL-TL算法在初始阶段即获得正向平均奖励，且显著高于源域任务中从零开始训练的MAD3QL算法。这一结果反映出TL机制为智能体提供了更优的初始参数基础，使其能够在新任务中实现“热启动”，而非从零开始探索。

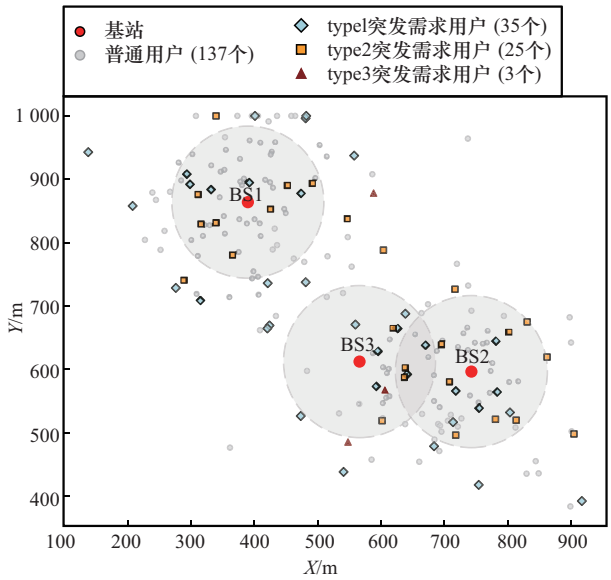
图12展示了在3种不同场景下，MAD3QL-TL算法的UAV部署效果。图12(a)展示了源域任务中MAD3QL算法训练后的结果，服务了92%的GU，BS和UAV分别服务了55.5%和36.5%的用户，其中，由BS转移至UAV服务的GU约占用户总数的9%。如图12(b)所示，当目标域与源域差异较小时，本文算法能够有效迁移源域的经验与策略，将多架UAV智能部署在热点区域，从而显著缓解网络拥塞。这表明在环境变化温和的情况下，TL机制可以充分发挥作用。如图12(c)所示，当突发业务需求增大、目标域与源域之间的差异扩大时，算法的迁移能力有所下降。具体而言，大多数UAV仍能部署在合适位置，但部分UAV未能有效学习源域经验，被部署在无效位置。这说明跨域差异过大对TL的有效性产生了一定影响。如图12(d)所示，在地面BS发生损坏的极端场景下，本文算法仍能快速部署UAV，为BS损坏范围内的GU提供临时通信服务，同时缓解正常BS范围内用户的网络拥塞。这验证了本文算法在应对突发基础设施故障时的鲁棒性和快速响应能力。



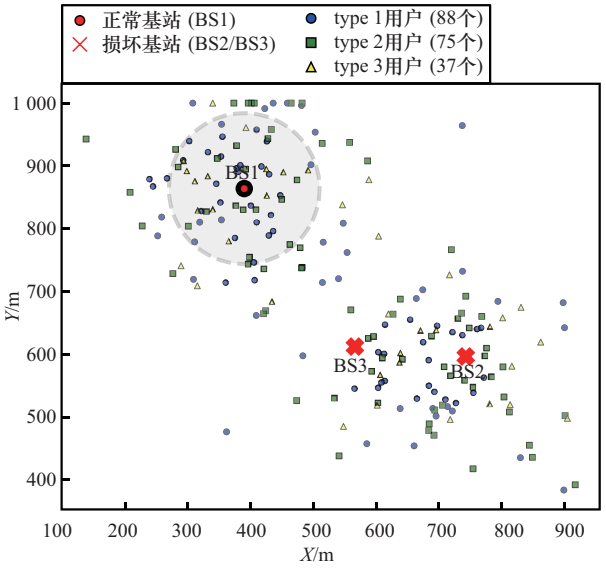
(a) 源MDP



(b) 目标MDP1（中等突发需求）



(c) 目标MDP2（高等突发需求）



(d) 目标MDP3（BS损毁场景）

图10　源域和目标域

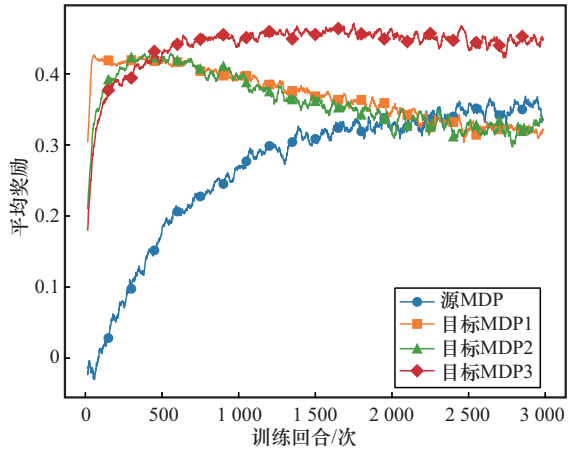
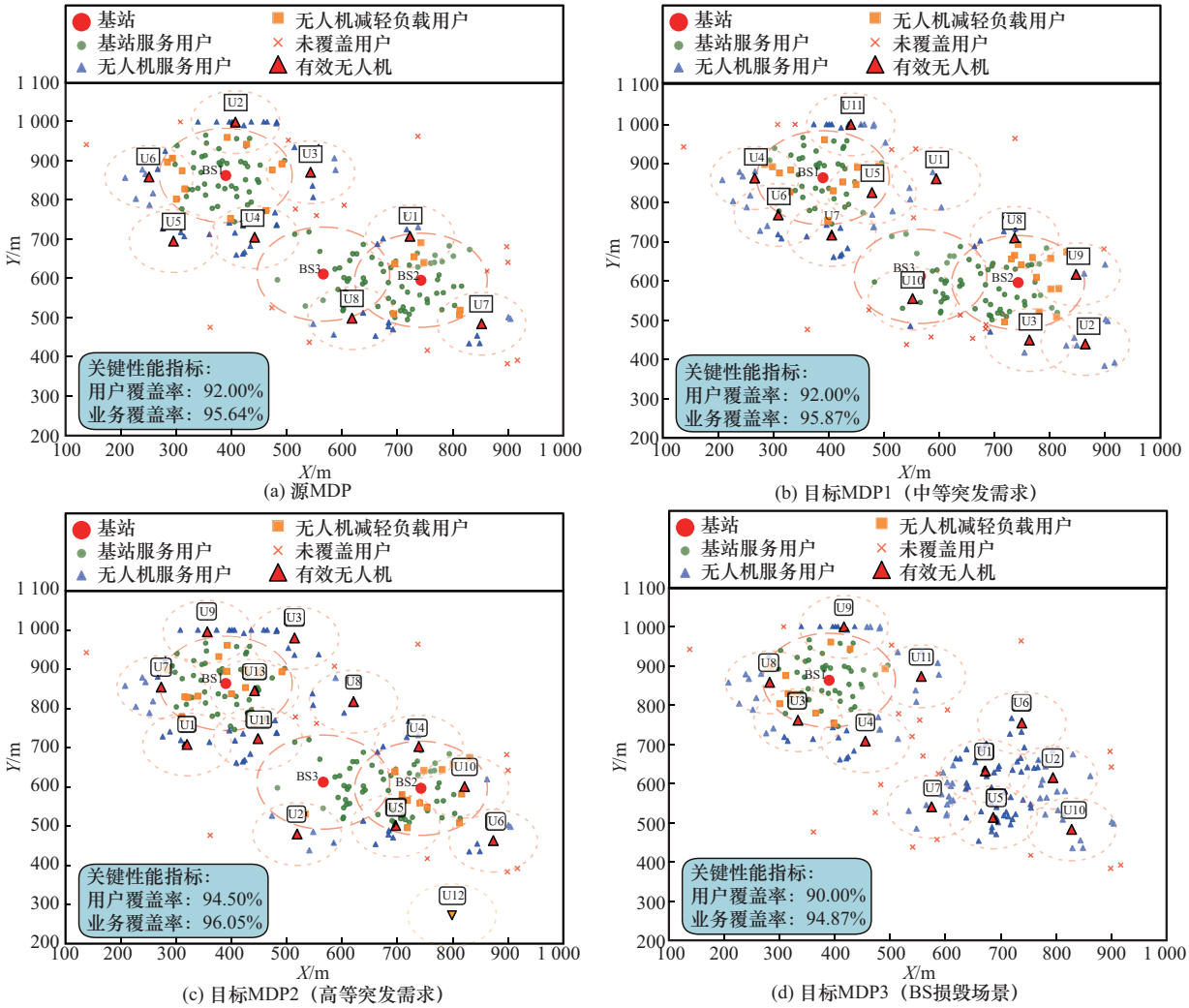


图11　基于TL的算法收敛速度

表4汇总了算法在不同子区域的性能表现，本文算法在3种不同场景下均取得了显著的优化效果。在目标域1（中等突发需求）中，经智能部署后，盲区用户得到了有效服务，网络拥塞得到了明显缓解。这表明当环境变化程度适中时，TL机制能够充分发挥作用。在目标域2（高等突发需求）中，尽管盲区用户获得了服务且过载BS业务量有所降低，但系统等待概率未能满足QoS要求。这说明当突发业务需求过大、目标域与源域差异超出一定范围时，算法的迁移能力受到一定限制，仍有待进一步优化。在目标域3（BS损坏场景）中，UAV



(a) 源MDP　(b) 目标MDP1（中等突发需求）

(c) 目标MDP2（高等突发需求）　(d) 目标MDP3（BS损毁场景）

图12　UAV部署结果

成功服务了大量因BS损坏而无法连接网络的用户，并有效缓解了正常BS范围内的网络拥塞。这验证了本文算法在应对突发基础设施故障时的鲁棒性和快速响应能力。在上述3种不同场景中，用户满意度均实现了大幅提升。上述结果表明，本文算法能够显著提升UAV在不同子区域间的学习效率与环境适应能力，同时降低计算复杂度和资源消耗，展现出良好的泛化性能和鲁棒性。

表4　不同目标任务场景下的性能验证

场景	覆盖盲区用户数		用户等待概率		用户满意度	
	优化前	优化后	优化前	优化后	优化前	优化后
源MDP	73	16	79.5%	0.9%	127.0	182.36
目标MDP1	73	16	80.9%	0.8%	127.0	182.55
目标MDP2	73	26	82.4%	42%	127.0	186.67
目标MDP3	145	23	100%	1.1%	55.0	179.17

5　结束语

针对Cell-Free网络难以应对GU需求突变及BS损毁等问题，本文提出了一种基于深度强化迁移学习与干扰管理的动态UAV辅助Cell-Free网络快速部署方法。MAD3QL解决了多UAV在高动态环境下的分布式协同部署决策问题，TL克服了传统DRL在场景变化时的重训练瓶颈，IMSA则从物理层消除了多UAV密集部署带来的同频干扰。三者的有机协同使UAV群能够在不同目标任务场景中“快速学会覆盖、同时避免相互干扰”。研究表明，本文算法有效克服了传统RL的训练延迟瓶颈，在不同突发场景下均能将等待概率控制在2%以内，平均用户满意度提升44.3%，且鲁棒性优于MAD3QL、MADDPG等算法。然而，当跨域差异过大时，复用经验可能误导网络陷入局部最优。未

来研究将聚焦于构建自适应TL驱动的智能决策架构，重点探索基于注意力机制的元迁移学习，以实现对负迁移的鲁棒免疫。

参考文献：

[1] Ning Z L, Hu H, Wang X J, et al. Mobile edge computing and machine learning in the Internet of unmanned aerial vehicles: a survey[J]. ACM Computing Surveys, 2024, 56(1): 1-31.

[2] Liu Y X, Mao W C, Li X L, et al. UAV-assisted integrated sensing and communication for emergency rescue activities based on transfer deep reinforcement learning[C]//Proceedings of the 30th Annual International Conference on Mobile Computing and Networking. New York: ACM Press, 2024: 2142-2147.

[3] Chen Q, Song X Q, Song T C, et al. Vehicular edge computing networks optimization via DRL-based communication resource allocation and load balancing[J]. IEEE Transactions on Mobile Computing, 2025, 24(9): 9222-9237.

[4] Yan H, Bao W D, Li X Q, et al. Fault-tolerant scheduling of heterogeneous UAVs for data collection of IoT applications[J]. IEEE Internet of Things Journal, 2024, 11(16): 26623-26644.

[5] Li J M, Pan Q J, Wan Z Q, et al. Low altitude 3-D coverage performance analysis of cell-free RAN for 6G systems[J]. IEEE Transactions on Vehicular Technology, 2023, 72(12): 16163-16176.

[6] Adnan M H, Zukarnain Z A, Amodu O A. Fundamental design aspects of UAV-enabled MEC systems: a review on models, challenges, and future opportunities[J]. Computer Science Review, 2024, 51: 100615.

[7] Li B W, Tripathi S, Hosain A K M S, et al. When learning meets dynamics: distributed user connectivity maximization in AAV-based communication networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2026, 12: 175-188.

[8] Gao R, Wang X. Rapid deployment method for multi-scene UAV base stations for disaster emergency communications[J]. Applied Sciences, 2023, 13(19): 10723.

[9] Ao T Y, Zhang K X, Shi H G, et al. Energy-efficient multi-UAVs cooperative trajectory optimization for communication coverage: an MADRL approach[J]. Remote Sensing, 2023, 15(2): 429.

[10] 张涛, 张迁, 朱颖雯, 等. 基于能量感知的智能反射面辅助无人机时效数据收集策略[J]. 电子与信息学报, 2025, 47(2): 427-438.
Zhang T, Zhang Q, Zhu Y W, et al. Energy aware reconfigurable intelligent surface assisted unmanned aerial vehicle age of information enabled data collection policies[J]. Journal of Electronics & Information Technology, 2025, 47(2): 427-438.

[11] Lyu J B, Chen X, Zhang J F, et al. Spatial deep learning for site-specific movement optimization of aerial base stations[J]. IEEE Transactions on Wireless Communications, 2024, 23(7): 7712-7727.

[12] Qureshi K I, Lu B X, Lu C, et al. Multi-agent DRL for air-to-ground communication planning in UAV-enabled IoT networks[J]. Sensors, 2024, 24(20): 6535.

[13] Zhang D, Wang M Y, Mango J, et al. A survey on applications of reinforcement learning in spatial resource allocation[J]. Computational Urban Science, 2024, 4: 14.

[14] 曾晓婉, 王海军, 黄蕾, 等. 无人机辅助D2D通信网络安全通信资源分配算法[J]. 通信学报, 2024, 45(2): 115-126.
Zeng X W, Wang H J, Huang L, et al. Resource allocation algorithm for secure communication in UAV-assisted D2D communication network[J]. Journal on Communications, 2024, 45(2): 115-126.

[15] Chen R R, Cheng W C, Ding Y, et al. QoS-guaranteed multi-UAV coverage scheme for IoT communications with interference management[J]. IEEE Internet of Things Journal, 2024, 11(3): 4116-4126.

[16] Cheng X, Jiang R, Sang H R, et al. Joint optimization of multi-UAV deployment and user association via deep reinforcement learning for long-term communication coverage[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 5503613.

[17] Lagos D M M, Azurdia-Meza C A, Ruiz-del-Solar J. Fair coverage for unmanned aerial vehicle-assisted cellular networks with deep reinforcement learning[J]. IEEE Internet of Things Journal, 2026, 13(3): 5202-5223.

[18] Ma H X, Yang G, Sun X X, et al. Improved DRL-based energy-efficient UAV control for maximum lifecycle[J]. Journal of the Franklin Institute, 2024, 361(6): 106718.

[19] 李斌, 彭思聪, 费泽松. 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[J]. 通信学报, 2023, 44(9): 228-237.
Li B, Peng S C, Fei Z S. Beamforming and resource optimization in UAV integrated sensing and communication network with edge computing[J]. Journal on Communications, 2023, 44(9): 228-237.

[20] Wang H Y, Li H, Wang X, et al. Resource allocation in UAV-D2D networks: a scalable heterogeneous multi-agent deep reinforcement learning approach[J]. Electronics, 2024, 13(22): 4401.

[21] Sharma R, Chopra S R, Gupta A, et al. Deployment of unmanned aerial vehicles in next-generation wireless communication network using multi-agent reinforcement learning[J]. IEEE Access, 2024, 12: 69517-69538.

[22] Xu L, Jiao J, Zhang X J, et al. Coded semantic-aware coordinated transmission in cohesive clustered satellite systems: an incremental MADRL approach[J]. IEEE Transactions on Wireless Communications, 2026, 25: 6362-6378.

[23] Omoniwa B, Galkin B, Dusparic I. Communication-enabled deep reinforcement learning to optimise energy-efficiency in UAV-assisted networks[J]. Vehicular Communications, 2023, 43: 100640.

[24] Chu N H, Hoang D T, Nguyen D N, et al. Joint speed control and energy replenishment optimization for UAV-assisted IoT data collection with deep reinforcement transfer learning[J]. IEEE Internet of Things Journal, 2023, 10(7): 5778-5793.

[25] Al-Hourani A, Kandeepan S, Lardner S. Optimal LAP altitude for maximum coverage[J]. IEEE Wireless Communications Letters, 2014, 3(6): 569-572.

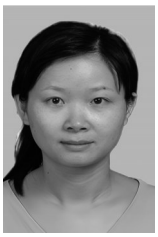
[26] Zhao H T, Wang H J, Wu W Y, et al. Deployment algorithms for UAV airborne networks toward on-demand coverage[J]. IEEE Journal on Selected Areas in Communications, 2018, 36(9): 2015-2031.

[27] He Y, Liu Y Q, Xie J. A distributed approach for user association and UAV deployment in QoE-aware multi-UAV networks[J]. IEEE Transactions on Network and Service Management, 2025, 22(5): 4394-4406.

[28] Chromy E, Misuth T, Kavacky M. Erlang C formula and its use in the call centers[J]. Advances in Electrical and Electronic Engineering,

2011, 9(1): 7–13.
[29] Wang Z Y, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on Machine Learning. New York: ACM Press, 2016: 1995-2003.
[30] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 2094-2100.
[31] Kingma D P, Ba J. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR). Vancouver: ICLR, 2015: 1-15.
[32] Zhang S, Ansari N. 3D drone base station placement and resource allocation with FSO-based backhaul in hotspots[J]. IEEE Transactions on Vehicular Technology, 2020, 69(3): 3322-3329.

[作者简介]



谢亚琴（1980– ），女，南京信息工程大学副教授，主要研究方向为室内定位、卫星导航、路径规划及人工智能。



张念（2001– ），男，南京信息工程大学硕士生，主要研究方向为无人机通信及资源分配。



张余（1987– ），男，南京信息工程大学讲师，主要研究方向为无蜂窝大规模MIMO通信技术、网络资源管理。



张治中（1972– ），男，南京信息工程大学教授，主要研究方向为5G/6G移动通信与信息处理等。



李佳珉（1983– ），男，东南大学教授，主要研究方向为6G无蜂窝智能无线接入网等。



王东明（1977– ），男，东南大学教授，主要研究方向为无线通信信号处理和无蜂窝大规模MIMO等。