基于GAT的分布式联合路由与频谱接入机制

周子铂[1,2]，任保全[1]，钟旭东[1]，刘琦[1]，秦蓁[1]
（1.军事科学院系统工程研究院，北京 100091；2.空军预警学院，湖北 武汉 430019）

摘　要：针对传统路由与频谱接入方法在动态拓扑下感知受限、决策耦合度高的问题，提出一种融合图注意力网络（GAT）与深度强化学习（DRL）的联合路由与频谱接入方法。首先将分布式路径建立过程建模为部分可观测马尔可夫过程（POMDP），利用DRL实现逐跳分布式决策；同时，通过GAT聚合局部感知信息，捕捉不规则拓扑结构与节点间干扰关系，以提升模型对复杂环境的适应能力。在模型训练阶段，采用优先经验回放机制提升样本利用效率。多种场景下的仿真实验验证了所提方法的有效性：在随机分布的拓扑结构中，所提方法可实现10%的数据率提升，同时降低频率切换次数以及路径建立跳数；在多数据流场景下，其性能与基线方法相当；在簇状拓扑结构中，切换次数与路径跳数分别降低约10%与13%。

关键词：图注意力网络；深度强化学习；路由；频谱接入
中图分类号：TN915
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2026003

GAT-based decision mechanism for decentralized joint routing and spectrum access

Zhou Zibo[1,2], Ren Baoquan[1], Zhong Xudong[1], Liu Qi[1], Qin Zhen[1]
1. Systems Engineering Institute, Academy of Military Science, Beijing 100091, China
2. Air Force Early Warning Academy, Wuhan 430019, China

Abstract: To address the limited situational awareness and high decision coupling of traditional routing and spectrum access methods in dynamic network topologies, a joint optimization framework that integrates graph attention networks (GAT) with deep reinforcement learning (DRL) was proposed. The distributed path establishment process was formulated as a partially observable Markov decision process (POMDP), enabling hop-by-hop decentralized decisions via DRL. GAT was implemented to aggregate local observations to capture irregular topologies and inter-node interference, improving adaptability to complex environments. During training, prioritized experience replay enhances sample efficiency. Extensive simulations under random, clustered, and multi-flow scenarios demonstrate the method's effectiveness: in random topologies, it achieves approximately 10% higher bottleneck throughput while reducing both channel switching frequency and path hop count. In clustered topologies, it reduces channel switches by about 10% and hop count by about 13%, and in multi-flow scenarios, its performance is comparable to baseline approaches.

Keywords: graph attention network, deep reinforcement learning, routing, spectrum access

0　引言

随着物联网（Internet of things, IoT）与车联网（vehicular Ad-hoc network, VANET）等新兴网络应用的快速发展，无线自组网凭借其支持灵活组网和对移动通信的良好支持，已成为支撑智能终端互联互通的重要使能技术[1]。在该类网络中，源节点与

收稿日期：2025–11–12；修回日期：2025–12–26
通信作者：任保全，renbq88@126.com
基金项目：中国博士后科学基金资助项目（No.2025M784510）
Foundation Item: China Postdoctoral Science Foundation (No.2025M784510)

目的节点之间往往需要通过多跳中继建立通信，而通信路径质量直接影响端到端吞吐率、时延与可靠性[2-3]。信道的时变特性以及频谱资源的稀缺性，使得构建高质量通信路径面临严峻挑战。传统通信网络采用分层优化架构，将路由选择、介质访问控制层与物理层资源管理割裂处理，难以实现跨层协同优化[4-5]。然而，无线网络工作在开放共享的电磁环境中，其信号质量受传播距离、遮挡、干扰等多重因素影响，仅依赖网络层跳数或链路质量作为路由度量，往往无法保障端到端性能。尽管已有研究考虑到信号质量对通信性能的影响，但较少充分挖掘物理层特性，如信干噪比等，在路由决策中的作用[6-7]。因此，建立节点间高质量的通信路径本质上是涉及网络层路由与物理层频谱资源管理的联合优化问题，其有效协同是提升通信网络整体效能的关键。

在频谱接入与路由优化方面，动态频谱接入（dynamic spectrum access, DSA）与智能策略路由（smart policy routing, SPR）已成为当前的研究热点[8-9]。文献[10]提出一种支持全频谱接入的基站天线设计方案，通过优化天线结构增强频谱接入能力。文献[11]针对认知车联网场景，设计了一种基于频谱稳定性评估的动态接入算法，提升了频谱利用的可靠性。文献[12]提出一种联合频谱感知与资源分配机制，兼顾频谱共享的安全性与效率。文献[13]则在毫微微蜂窝（femto-cell）网络中引入双边拍卖机制，实现频谱分配与功率控制的协同优化，有效提高了频谱利用率与系统效用。此外，博弈论[14]、凸优化[15-16]等经典方法也被广泛应用于频谱管理。然而，上述方法普遍依赖精确的系统模型和先验环境信息，在面对拓扑高度动态、干扰关系复杂、状态空间庞大的实际无线网络时，往往面临计算复杂度高、收敛速度慢、难以在线求解等挑战[17]。同时，多数研究聚焦于单一优化目标，缺乏对路由决策与频谱接入的联合建模。

在无线自组网中，从源节点到目的节点的路由决策与频谱管理问题可建模为典型的马尔可夫决策过程（Markov decision process, MDP），即序列决策问题。由于强化学习（reinforcement learning, RL）不需要精确环境建模，并能通过试错机制学习最优策略，其在动态环境下处理序列决策任务方面受到广泛关注[18-19]。深度强化学习（deep reinforcement learning, DRL）通过引入深度神经网络替代传统Q表，显著提升了在高维状态空间中的泛化能力与决策效率[20-21]。文献[22]提出一种基于DRL的动态频谱接入算法，验证了其在频谱资源分配中的有效性。文献[23]设计了一种融合频谱感知与非正交多址的大规模免授权随机接入方案，利用强化学习优化接入策略，提升了系统吞吐率与公平性。文献[24]将竞争双深度Q网络（dueling double DQN， D3QN）应用于动态频谱接入，显著提高了频谱分配准确性与信道容量。面向多业务非规则场景，文献[25]结合强化学习实现频谱与功率的联合分配。文献[26]则提出一种基于多智能体近端策略优化（multi-agent proximal policy optimization, MAPPO）的多信道动态频谱接入方法，有效提升了频谱利用效率。然而，现有DRL方法多采用全连接网络或卷积神经网络作为策略网络，其架构主要针对欧几里得空间，难以有效应对无线网络固有的非规则图结构。相比之下，图神经网络（graph neural network, GNN）凭借其对不规则图结构数据的强大特征提取能力及跨拓扑的泛化性能，为通信网络的优化提供了新的解决思路[27]。文献[28]将GNN应用于动态路由决策，通过显式建模网络拓扑关系，使智能体在训练中未见过的拓扑上仍具备良好泛化能力；文献[29]进一步提出一种基于图卷积网络（graph convolutional network, GCN）驱动的多任务DRL框架，实现了网络切片与路由的联合优化。

尽管上述研究取得了重要进展，但仍存在若干亟待解决的关键问题。首先，现有工作未能充分挖掘路由决策与频谱接入之间的强耦合关系；其次，各节点面临部分可观测马尔可夫决策过程（partially observable Markov decision process, POMDP）的挑战；最后，多数研究未充分考虑信道切换开销对系统性能的影响，导致路径建立过程中频繁切换信道，从而引入额外的时延与能量消耗。

针对上述挑战，本文提出一种融合图注意力网络（graph attention network, GAT）与DRL的联合路由与频谱接入方法，聚焦于无线自组织网络中高质量通信路径的构建，重点围绕节点局部感知能力增强、频谱资源高效管理与动态拓扑适应性等关键维度展开。本文主要贡献如下。

1) 采用基于DRL的逐跳联合决策框架实现路

由与频谱接入的分布式跨层优化，有效应对全局拓扑未知条件下的决策问题。

2) 引入GAT的多头注意力机制对局部感知信息进行聚合，提升模型对不规则拓扑结构与节点干扰关系的表征能力，增强决策的鲁棒性。

3) 在奖励函数中综合考虑吞吐率与信道切换次数，有效平衡通信性能与切换代价，降低路径质量波动。

1　系统模型与问题建模

1.1　系统模型

考虑一个部署在矩形区域$X \times Y$内的无线自组网，N个同质节点随机分布，第i个节点记为n_i，构成节点集合$\psi = \{n_1,\cdots,n_N\}$，如图1所示。节点$n_i$的位置表示为$p_i = (x_i,y_i)$，设所有节点静止处于相同高度。网络支持数据流集合$F$，每个数据流具有固定的源节点和目的节点。数据流$f \in F$发起通信请求时，需在该数据流的源节点与目的节点之间建立通信路径。

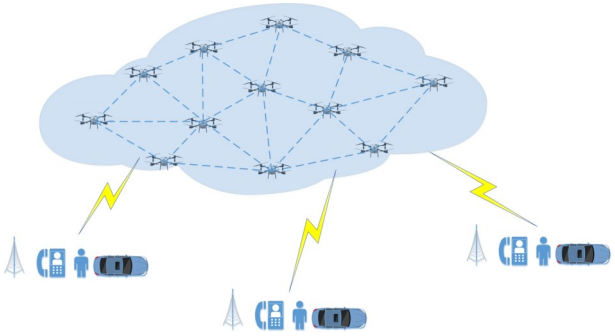


图1　系统模型

$$\mathbf{Route}_f = \left[n_{f,0},n_{f,1},\cdots,n_{f,H_f}\right] \tag{1}$$

其中，H_f为建立数据流f通信路径所需的最大跳数。路径建立成功后，$n_{f,0}$与n_{f,H_f}分别表示源节点与目的节点，其余节点为中继节点。

物理层总带宽W被等分为M个正交信道，并构成信道集合，各信道的带宽为$w = \dfrac{W}{M}$。

$$\boldsymbol{B} = \left[b_1,b_2,\cdots,b_M\right] \tag{2}$$

设节点发射功率p_t恒定，背景噪声功率为σ^2，信号传播过程中存在路径损耗。在节点n_i向节点n_j传输数据时，信道增益为$g_{i,j} \in \Re^+$，该增益数值取决于两节点之间的传播距离。

$$g_{i,j} = \left(\frac{\lambda}{4\pi d_{i,j}}\right)^2 \tag{3}$$

其中，$d_{i,j}$为两节点之间的距离，λ为信号传输的波长。

综合考虑信号传播过程中的衰减、噪声与干扰，接收端的信干噪比（signal-interference-noise-ratio, SINR）为

$$\mathrm{SINR}_{i,j} = \frac{g_{i,j}p_t}{\sum\limits_{k \neq i,j} g_{k,j}p_t + \sigma^2} \tag{4}$$

其中，$k \neq i$、j表示求和项中仅排除当前节点与候选节点，网络内其他节点均可对(i,j)节点对之间的信道质量造成影响。具体而言，干扰项包括数据流内部节点之间的干扰以及其他数据流的干扰。当多个数据流共存（$|F|>1$）时，对数据流f而言，其余数据流$F_{/f}$即被视为干扰源。

根据香农定理，两节点间的最大传输速率，即通信容量，由接收端的信干噪比决定。那么，两节点之间允许的最大传输速率为

$$C_{i,j}^f = w\,\mathrm{lb}\left(1 + \mathrm{SINR}_{i,j}\right) \tag{5}$$

1.2　问题建模

在多跳通信网络中，数据流f的端到端吞吐率受限于其路径上速率最低的链路，即瓶颈链路，其传输速率称为瓶颈吞吐率，即

$$C_f^{bn} = \min\left(C_{0,1}^f,C_{1,2}^f,\cdots,C_{H_f-1,H_f}^f\right) \tag{6}$$

本文的目标是通过联合优化路由选择与信道分配，最大化数据流中的瓶颈吞吐率，以提升系统整体通信质量。优化问题可建模为

$$\max\ C_f^{bn} \tag{7}$$

$$\text{s.t.}\quad n_{h_f} \in \boldsymbol{N}_{h_f-1},\quad \forall h_f = 1,\cdots,H_f \tag{7a}$$

$$n_{h_f} \neq n_{h'_f},\quad \forall h_f \neq h'_f \tag{7b}$$

$$b_{h_f} \in \boldsymbol{B},\quad \forall h_f = 1,\cdots,H_f \tag{7c}$$

$$H_f < H_{\max} \tag{7d}$$

其中，$\boldsymbol{N}_{h_f-1} = \left\{n_i | n_i \in \Psi, \left\| p_{h_f-1} - p_i \right\|_2 \leqslant R_s\right\}$为数据流$f$第$h_f-1$跳节点的邻域，$R_s$为节点的感知距离，$H_{\max}$为路径建立最大允许跳数。式(7a)约束下一跳节

点必须位于当前节点的感知范围内，体现了节点的局部感知能力；式(7b)防止同一节点被重复选择，避免出现环路；式(7c)规定选择的信道必须在信道集合内；式(7d)限制了路径建立的最大跳数。

由于节点间存在相互干扰，各数据流的传输速率不仅受到自身链路条件的影响，还与其他数据流之间存在强耦合关系。在优化整个网络时，需要在不同数据流之间进行协同与权衡。式(7)所述优化问题属于混合整数非线性规划（mixed-integer non-linear programming， MINLP）问题，且为典型的NP难问题。其复杂性源于：路由决策与信道分配之间存在强耦合关系，路径选择影响干扰分布，信道分配作用于链路质量；多数据流共存时，流间干扰导致目标函数高度非凸；节点仅能感知局部信息，无法获知全局拓扑与干扰状态，决策过程具有部分可观测性。为有效解决此类优化问题，本文提出一种基于图注意力网络的深度强化学习模型，以实现路径与频谱资源的联合优化管理。

2 分布式联合路由与频谱接入方法

2.1 联合决策模型

DRL主要包括决策代理，该决策代理通过与环境的交互，学习最优策略以解决面临的POMDP问题。环境提供任务的动态描述，决策代理通过试错–经验，学习执行任务的策略。所有可能的环境状态构成状态空间$\boldsymbol{S}$，决策代理在各状态下可采取的动作构成动作空间$\boldsymbol{A}$。在时间t，智能体获取感知信息，即环境状态s_t，依据策略$\pi:\boldsymbol{S}\times\boldsymbol{A}\to[0,1]$选择动作$a_t\in\boldsymbol{A}$，并根据奖励函数$\mathrm{R}:\boldsymbol{S}\times\boldsymbol{A}\to\Re^{+}$获得奖励$r_t$。决策代理的目标是最大化累积奖励。

$$R_t=\sum_{k=0}^{\infty}\gamma^k r_{t+k} \tag{8}$$

其中，折扣因子$\gamma\in(0,1]$反映策略对未来奖励的重视程度。决策代理通常为参数化的DNN，而学习获得的策略则体现在DNN的参数θ，训练过程中利用经验样本对参数进行持续的优化。在联合路由与频谱接入场景中，观测空间、动作空间与奖励函数定义如下。

2.1.1 观测空间

由于环境状态并非完全可观测，决策代理仅能感知环境的部分信息$\boldsymbol{o}\in\boldsymbol{O}$，即观测空间$\boldsymbol{O}$仅包含状态空间$\boldsymbol{S}$的部分信息，其余信息处于隐藏状态。受限于有限的感知范围，节点仅能收集其邻域内的信息，即与当前节点的距离小于感知距离R_s的节点信息。并且，邻域内的信息通过控制与非有效载荷链路（control and non-payload communications, CNPC）进行交换。具体而言，决策代理在时间t的观测信息如下。

1) 当前节点到邻域节点的距离d_f。

2) 邻域节点到目的节点的距离d_d。

3) 当前节点指向邻域节点的方向与指向目的节点方向之间的角度差φ。

4) 邻域节点接收端的SINR。

具体而言，在第h_f跳时，第h_f–1节点的观测向量定义为

$$\boldsymbol{o}_{h_f-1}=\mathop{\Vert}_{j\in N_{h_f-1}}\left[d_{f,j},d_{d,j},\varphi_j,\mathrm{SINR}_j\right] \tag{9}$$

其中，‖表示向量拼接。

考虑到上述数据之间的较大差异，在数据预处理阶段，分别根据网络部署区域的对角线长度与角度π对距离与角度进行标准化预处理，同时对获得的SINR数值进行归一化。

2.1.2 动作空间

动作空间$\boldsymbol{A}$是节点选择空间$\boldsymbol{A}_n$与信道空间$\boldsymbol{A}_{Fv}=\boldsymbol{B}$的笛卡儿积。

$$\boldsymbol{A}=\boldsymbol{A}_n\times\boldsymbol{A}_{Fv} \tag{10}$$

然而，节点的感知邻域随位置发生变化，因此节点选择空间为

$$\boldsymbol{A}_n=\left\{n_i|n_i\in\boldsymbol{N}_{h_f-1}\right\} \tag{11}$$

2.1.3 奖励函数

为避免传统时序差分（temporal-difference, TD）方法中价值估计偏差问题，本文采用蒙特卡罗（Monte-Carlo, MC）策略。因此，每一步的奖励r_t应准确反映该步决策的质量。此处，在路径建立成功后，将全局瓶颈吞吐率以及信道切换次数分解为逐跳奖励，为每个状态–动作对分配未来的累积奖励，以平衡通信性能与切换开销。在通信路径建立完成后，数据流f的通信速率需根据网络中所有节点空间位置与信道选择进行综合计算，并统一分配奖励。第t步的奖励定义为从该跳起至路径终点的最小链路速率与该跳之后所进行信道切换次数的加权组合。

$$r_t = (1-\beta)\min_{k=t,\cdots,H_f} C^f_{k,k+1} - \beta\upsilon_t \tag{12}$$

其中，υ_t为该跳之后直至路径建立成功还需要的信道切换次数，反映当前选择对整体路径信道切换次数的影响；$\beta \in (0,1)$为瓶颈吞吐率与信道切换次数之间的平衡系数。该设计使决策代理在追求高吞吐率的同时，避免频繁信道切换带来的性能损耗。

2.1.4　终止条件

路径建立过程在以下任一条件满足时终止。

1) 成功建立源节点与目的节点间的路径。

2) 当前节点为孤立节点，无可用下一跳。

3) 路径建立已达到预设最大跳数$H_{\max}$。

其中，条件1)为路径建立成功，将计算累积奖励并更新策略。条件2)与条件3)为路径建立失败，不进行奖励分配与经验存储，且该路径不产生实际通信流量。

值得注意的是，本文对路径建立失败的情况不施加显式惩罚与经验存储，主要基于以下考虑：失败路径无法形成有效通信路径，其性能在物理层面难以明确，强行设定负奖励可能引入与优化目标的偏差；采用蒙特卡罗奖励机制，仅对成功建立的路径进行性能评估，确保奖励信号与系统目标严格匹配。

2.2　图注意力网络

在无线自组网中，节点连接关系不规则且动态变化，其通信性能受邻近节点状态与干扰水平的显著影响。本节采用GAT作为DRL策略网络的核心组件，以实现对局部感知信息的动态聚合，辅助联合路由与频谱资源管理。当前节点仅能观测其直接邻域内的信息，同时每个邻域节点也可获取其邻域范围的感知信息。基于GAT的DRL决策模型如图2所示，展示了GAT对感知信息的处理流程，通过加权聚合这种邻域信息，在不违反物理感知约束的前提下，增强节点的感知能力。需要强调的是，下一跳节点的选择仍限于当前节点的感知范围。

GAT模块包含多个相同的图注意力层，各层通过注意力机制为不同邻居节点分配差异化权重，分别对邻域信息进行增强与汇聚。设当前节点在决策时刻t的观测向量集合为

$$\boldsymbol{O}(t) = [\boldsymbol{o}_0(t), \boldsymbol{o}_1(t), \cdots, \boldsymbol{o}_m(t)] \tag{13}$$

其中，m为当前节点感知范围内的节点数量，$\boldsymbol{o}_j(t) \in R^L\ (j = 1,\cdots,m)$为第$j$个邻域节点的感知信息，$\boldsymbol{o}_0(t)$为当前节点的感知信息。

首先，利用投影矩阵$\boldsymbol{U} \in R^{L' \times L}$对所有观测向量进行线性变换，然后通过注意力函数$q: R^{2L'} \to \Re^+$计算注意力系数。

$$e_{ij} = q\left[\boldsymbol{U}\boldsymbol{o}_i(t) \| \boldsymbol{U}\boldsymbol{o}_j(t)\right] \tag{14}$$

其中，$i,j \in [0,\cdots,m]$。

注意力计算利用LeakyReLU函数实现网络的非线性适应性。为便于不同邻居节点间的比较，使用Softmax函数对注意力系数进行归一化。

$$\alpha_{ij} = \frac{\mathrm{LeakyReLU}\left(q\left[\boldsymbol{U}\boldsymbol{o}_i(t) \| \boldsymbol{U}\boldsymbol{o}_j(t)\right]\right)}{\sum\limits_{i,j \in [0,\cdots,m]} \mathrm{LeakyReLU}\left(q\left[\boldsymbol{U}\boldsymbol{o}_i(t) \| \boldsymbol{U}\boldsymbol{o}_j(t)\right]\right)} \tag{15}$$

归一化后的注意力系数反映了节点n_j的感知信息对节点n_i的重要程度。随后，通过注意力系数的加权求和与非线性激活函数ρ得到输出。为提升注意力学习的稳定性，采用多头注意力机制。

$$\boldsymbol{o}'_i(t) = \Big\|_{k=1}^{K} \rho\left(\sum_{j \in [0,\cdots,m]} \alpha^k_{ij} \boldsymbol{U}_k \boldsymbol{o}_j(t)\right) \tag{16}$$

其中，K为注意力头数。

经过处理之后，第一层图注意力层输出新的观测集合。

$$\boldsymbol{O}'(t) = [\boldsymbol{o}'_0(t), \boldsymbol{o}'_1(t), \cdots, \boldsymbol{o}'_m(t)] \tag{17}$$

中间多层的处理方式与此类似，使得各节点的特征得到增强。

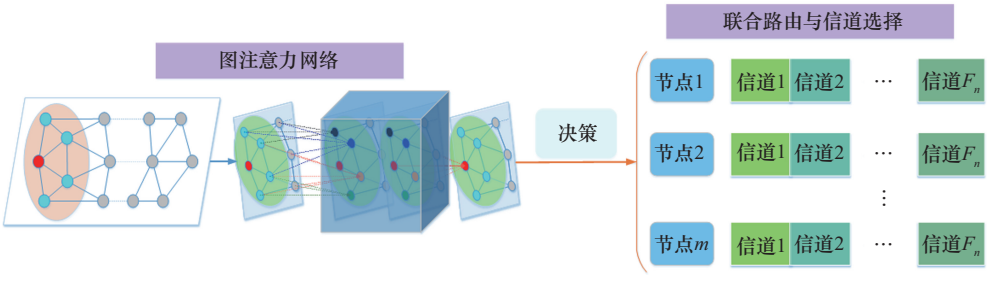


图2　基于GAT的DRL决策模型

$$\boldsymbol{O}''(t)=[\boldsymbol{o}_0''(t),\boldsymbol{o}_1''(t),\cdots,\boldsymbol{o}_m''(t)] \tag{18}$$

在最后一层，需要将所有节点的信息汇聚至当前节点，生成最终的观测向量，作为DRL网络的输入，用于Q值估计与动作选择。信息汇聚过程中，当前节点与各邻居节点之间的注意力系数为

$$\alpha'_{0j}=\frac{\mathrm{LeakyReLU}\left(q\left[\boldsymbol{U}\boldsymbol{o}_0''(t)\|\boldsymbol{U}\boldsymbol{o}_j''(t)\right]\right)}{\sum\limits_{j\in[1,\cdots,m]}\mathrm{LeakyReLU}\left(q\left[\boldsymbol{U}\boldsymbol{o}_0''(t)\|\boldsymbol{U}\boldsymbol{o}_j''(t)\right]\right)} \tag{19}$$

最终，GAT模块输出Q值估计向量为

$$Q(o_t,a;\theta)=\boldsymbol{o}_0^f(t)=\rho\left(\sum_{j\in[1,\cdots,m]}\alpha'_{0j}\boldsymbol{U}\boldsymbol{o}_j''(t)\right) \tag{20}$$

2.3　训练机制与策略优化

为提升样本利用与模型训练的效率，本节在DRL训练流程中引入优先经验回放（prioritized experience replay, PER）机制。与标准经验回放随机采样不同，PER机制优先选择估计误差较大的经验样本，因其蕴含更多可用于策略更新的信息。训练期间采用ε-贪婪策略与环境交互收集经验，该策略旨在平衡探索与利用。

$$a_t=\begin{cases}\mathrm{random}(\boldsymbol{A}), & \xi<\varepsilon\\ \underset{a}{\mathrm{argmax}}\left(Q(o_t,a;\theta)\right), & \xi>\varepsilon\end{cases} \tag{21}$$

其中，随机数$\xi\sim\mathrm{Uniform}(0,1)$，$\varepsilon\in[0,1]$为探索系数。训练初期，$\varepsilon$设为较高值，鼓励决策代理探索网络拓扑，广泛收集不同条件下的状态–动作–奖励数据。随着训练推进，ε逐步衰减，减少探索行为，转向利用已学策略。

因为此处深度强化学习采用MC策略，每个经验样本$\boldsymbol{E}_i$须为完整的连续观测序列，并且经验样本$\boldsymbol{E}_i$的存储需延迟至回合结束。对于经H_f跳完成的数据流，第t步的经验元组为$\langle o_t,a_t,r_{t+1}\rangle$，而完整经验样本$\boldsymbol{E}_i$为

$$\boldsymbol{E}_i=\left\{\left\langle\boldsymbol{o}_{i,0},a_{i,0},r_{i,1}\right\rangle,\cdots,\left\langle\boldsymbol{o}_{i,t},a_{i,t},r_{i,t+1}\right\rangle,\cdots,\left\langle\boldsymbol{o}_{i,H_f-1},a_{i,H_f-1},r_{i,H_f}\right\rangle,\delta_i,\tau_i\right\} \tag{22}$$

其中，δ_i与τ_i分别表示样本$\boldsymbol{E}_i$训练误差与回放概率。在经验存储阶段，将δ_i初始化为10^3，使得新获取的经验具备更高的回放概率。

在每一轮模型训练之后，δ_i进行更新。

$$\delta_i=\left|\mathrm{Loss}\left(\boldsymbol{E}_i,\theta\right)+\eta\right|^{\alpha}=\left|\mathrm{MSE}\left[Q\left(o_{i,t},a_{i,t},\theta\right),y_{i,\mathrm{tar}}\right]+\eta\right|^{\alpha} \tag{23}$$

其中，α为误差指数，$Q\left(o_i,a_i,\theta\right)$为样本$\boldsymbol{E}_i$经过DRL模型之后的$Q$值输出，$y_{i,\mathrm{tar}}=r_{i,t+1}$表示样本$\boldsymbol{E}_i$的训练目标，$\eta$为较小数，防止误差为0。

训练误差δ_i更新之后，同步对经验池内所有经验的回放概率进行更新。

$$\tau_i=\frac{\delta_i}{\sum\limits_i\delta_i} \tag{24}$$

在训练架构上，将蒙特卡罗奖励作为Q值学习目标，以降低注意力网络训练过程中的计算复杂度。同时，为缓解因缺少目标网络可能引发的训练不稳定问题，对梯度进行裁剪后再执行反向传播。算法流程如图3所示，当数据流请求生成时，源节点首先作为决策代理选择下一跳节点及其信道；一旦源节点与下一跳节点之间的链路建立完成，决策代理角色便转移至下一跳节点，使其成为新的决策代理，继续执行联合路由与信道分配决策。该过程迭代进行，直至目的节点加入路径。路径成功建立之后，将成功经验存储至经验池，然后根据优先经验回放机制对模型进行训练。

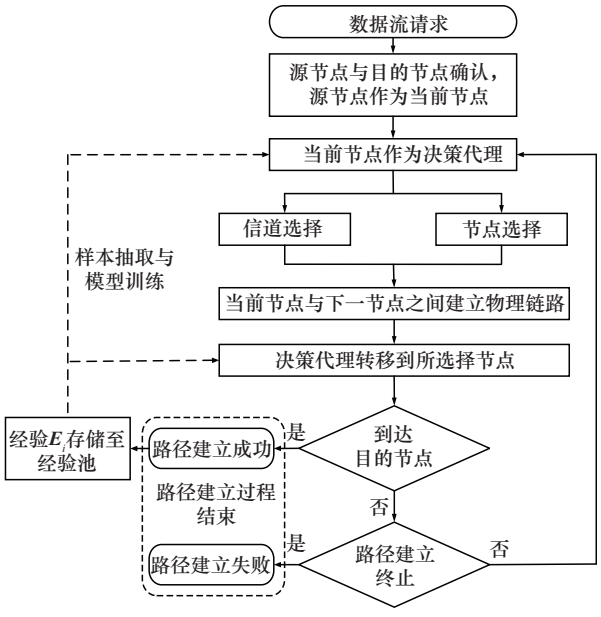


图3　算法流程

3　实验结果与分析

为全面评估基于GAT的DRL联合路由与频谱接入方法的性能，本文构建了仿真平台，并在多种网络场景下进行对比实验。实验环境设置如下：仿真区域为5 km × 5 km的矩形区域，按随机分布拓

扑结构分别部署20或35个节点，以及簇状拓扑结构部署30个节点；物理层中心频率为2.4 GHz，总带宽为50 MHz，划分为5个正交信道；节点发射功率固定，背景噪声功率$\sigma^2 = -130$ dBm/Hz，感知距离$R_s = 1.5$ km；通过实验调优确定平衡系数$\beta = 0.2$；经验回放池可容纳5 000条经验样本，每次训练按回放概率抽取128个样本。模型参数采用Adam优化器进行更新，学习率为1×10^{-4}。

为增强训练阶段模型的探索能力，单回合的最大跳数$H_{\max} = 20$。为避免测试阶段过度占据网络资源，单数据流最大允许跳数$H_{\max} = 10$。采用ε-贪婪策略训练30 000次迭代，每次迭代包含5个回合。在前90%的训练迭代中，探索率从1.0指数衰减至0.01，最后10%的迭代探索率固定为0，以稳定策略性能。网络拓扑在每次迭代开始时重新生成，每个回合随机选取新的源–目的节点对，且确保目的节点位于源节点的感知范围之外。在训练过程中，持续对模型进行测试，每轮测试包含100个测试回合，每回合使用不同的拓扑结构与源–目的节点。此外，测试阶段所用拓扑与训练阶段完全不重叠，以验证模型的泛化能力。GAT模块包含两层图注意力层，且各层均为4注意力头结构，输入观测经投影矩阵变换后为64维特征向量，该GAT模块仅聚合单跳范围内的邻居节点信息，以降低CNPC链路的开销。

为验证所提方法的有效性与先进性，选取以下3类典型神经网络作为DRL策略网络的对比基线：1)多层感知机（multi layer perceptron, MLP），传统全连接网络；2)Transformer-Encoder（以下简称Encoder），时序注意力网络，基于自注意力机制的时序模型；3)GCN，经典图神经网络，采用固定权重聚合邻域信息。

3.1　随机分布拓扑

本节实验考察网络在随机拓扑下的性能表现，网络由20个随机分布的节点构成，单数据流通信，相应的瓶颈吞吐率优化曲线如图4所示。可见，所提方法在随机分布的拓扑结构下收敛速度较快，且获得较高的瓶颈吞吐率6 Mbit/s，显著优于对比方法。GCN与Encoder性能相近，瓶颈吞吐率分别在5.5 Mbit/s与5.2 Mbit/s附近，而MLP则取得最低的瓶颈吞吐率，低于5 Mbit/s。该结果表明，GAT通过注意力机制自适应地关注高干扰或关键邻域节点信息，有效提升了状态表征能力，从而实现更优的联合决策效果。

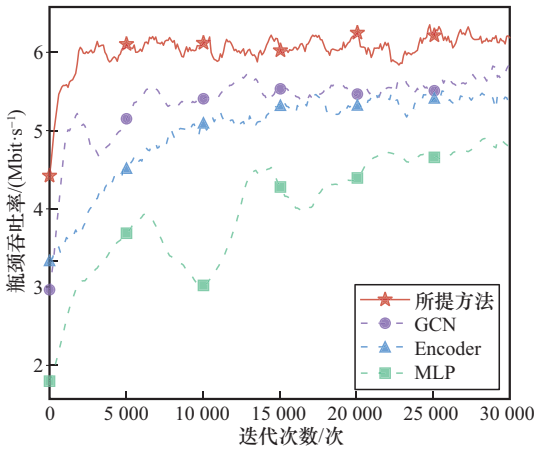


图4　随机分布拓扑下瓶颈吞吐率优化曲线

图5为随机分布拓扑下信道切换次数的优化曲线。在训练过程中，所提方法信道切换次数显著降低，平均低于4次，而MLP与GCN的切换次数维持在4.5次左右，Encoder的信道切换次数最高，接近5.5次。这表明GAT能够学习到更稳定的信道使用策略，减少信道频繁切换导致的通信质量波动。为对比各方法所建立路径的时延特性，图6展示了路径跳数的优化过程。所提方法收敛后平均跳数约为5.2跳，明显低于GCN（5.7跳）与MLP（6.8跳），远优于Encoder（7.0跳）。这说明GAT不仅关注链路质量，还能综合方向性与拓扑连通性，选择更短、更高效的路径。因此，在随机拓扑下，所提方法在瓶颈吞吐率、信道切换次数与路径跳数3项关键指标上均表现最优，验证了其在常规网络场景下的综合优化能力。

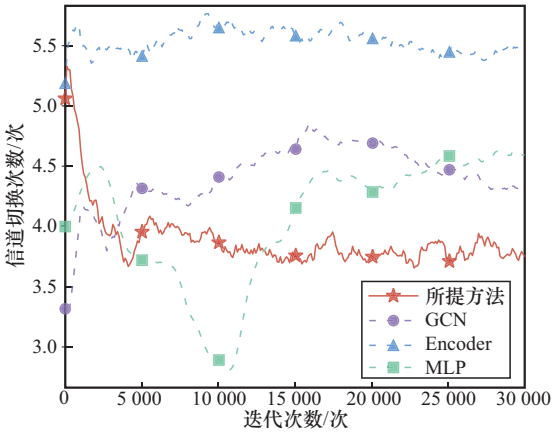


图5　随机分布拓扑下信道切换次数的优化曲线

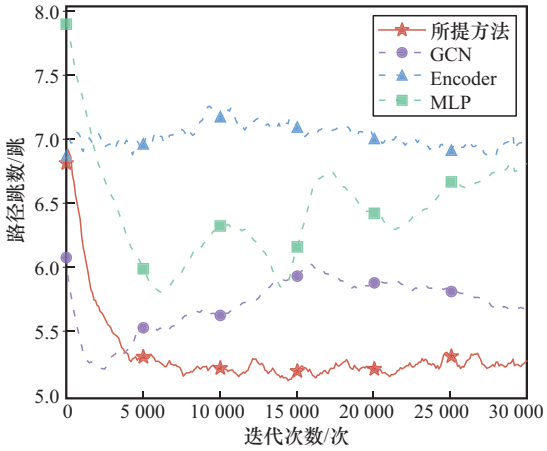


图 6　随机分布拓扑下路径跳数优化曲线

3.2　多数据流场景

为评估所提方法在高负载、强干扰环境下的鲁棒性，本节实验设置了 35 个随机分布的节点，并同时建立两条独立数据流。图 7 为多数据流瓶颈吞吐率之和的优化曲线，由于数据流之间的相互干扰，总瓶颈吞吐率低于单数据流场景。具体分析可知，所提方法实现的总瓶颈吞吐率相对较高，达到 4.6 Mbit/s，GCN 与 Encoder 的总瓶颈吞吐率分别在 4.5 Mbit/s 与 4.2 Mbit/s 附近，而 MLP 则取得最低的总瓶颈吞吐率，低于 4.0 Mbit/s。分析表明，GAT 的性能优势源于其多头注意力机制能够自适应地加权邻居节点的观测信息，有效区分干扰源与有用中继节点，从而在复杂干扰环境中作出更优的路由与频谱联合决策。相比之下，GCN 采用固定归一化邻接权重，在动态干扰场景下难以灵活调整信息聚合策略；Encoder 虽能提取局部特征，但未显式建模节点间关系；而 MLP 完全忽略网络拓扑，导致决策质量显著下降。

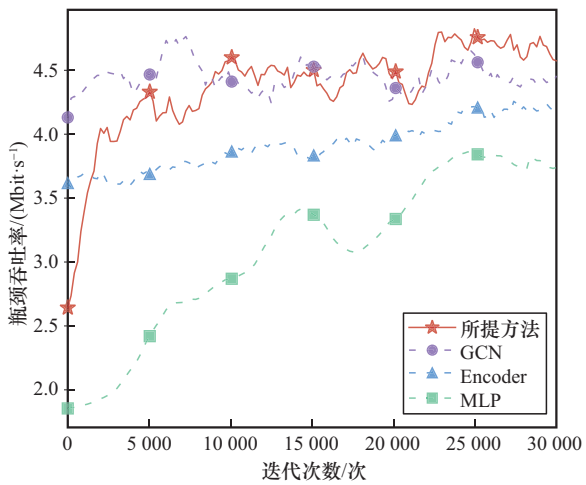


图 7　多数据流瓶颈吞吐率之和的优化曲线

此外，为进一步考察路径间的公平性，图 8 展示了多数据流之间最低瓶颈吞吐率的优化曲线。所提方法仍表现最优，最低瓶颈吞吐率接近 1.2 Mbit/s，表明其在提升系统总吞吐的同时兼顾了多数据流之间的资源分配公平性。GCN 次之，Encoder 与 MLP 因缺乏对跨流干扰的协调感知，易过度倾斜某一条流，导致另一条流性能严重受限，最低瓶颈吞吐率明显偏低。图 9 和图 10 分别为多数据流信道切换次数和路径跳数。两种图神经网络模型（GAT 与 GCN）在信道切换次数和路径跳数方面均优于 MLP 与 Encoder。因此，所提方法不仅有效控制了信道切换频率，还减少了路径跳数，展现出优异的资源协调性能。

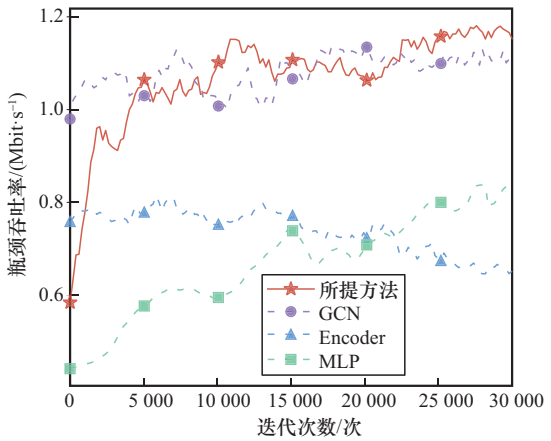


图 8　多数据流之间最低瓶颈吞吐率的优化曲线

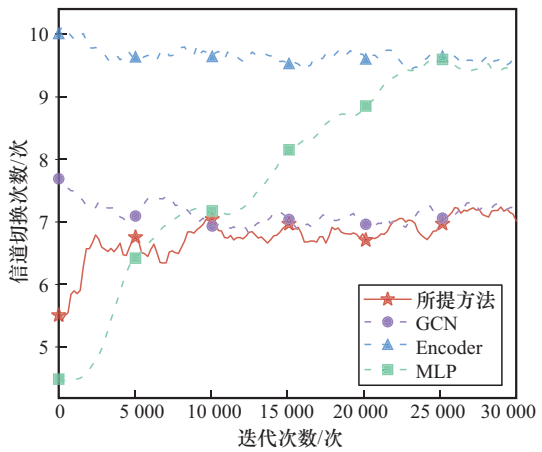


图 9　多数据流信道切换次数的优化曲线

3.3　簇状分布拓扑

为验证所提方法在非均匀、非规则拓扑下的适应能力，本节实验构建簇状网络结构：在实验区域内分布 6 个簇心，以 1.5 km 为半径生成互不重叠的

簇区域，每个簇内随机部署 5 个节点，共 30 个节点。此类拓扑模拟了城市热点区域、工业园区等实际场景。

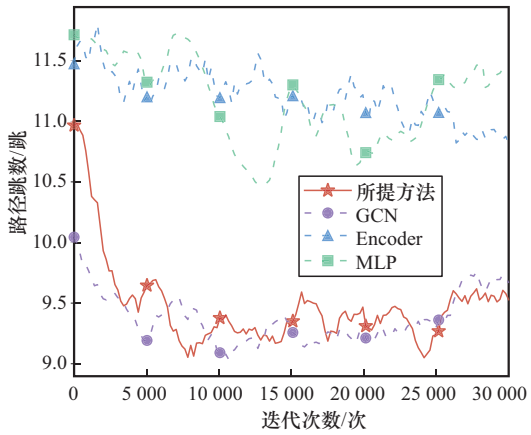


图 10　多数据流建立所需路径跳数的优化曲线

图 11 为簇状拓扑下单数据流的瓶颈吞吐率优化曲线，所提方法实现了较高的瓶颈吞吐率，达到 3.8 Mbit/s。对比 GCN 与 Encoder 的效果稍差，分别收敛在 3.6 Mbit/s 与 2.8 Mbit/s 附近，而 MLP 则取得最低的瓶颈链路质量，低于 2.5 Mbit/s。在簇状拓扑中，节点呈非均匀分布，簇内连接密集而跨簇链路稀疏，跨簇通信链路易受长距离衰减与遮挡影响，导致跨簇链路质量显著下降。在这类高度异构的环境中，GAT 凭借其注意力机制，能够自适应地识别并强化对高信道增益、低干扰链路的关注，有效抑制低质量邻居的负面影响，从而在路由与频谱联合决策中选择更可靠的跨簇路径。

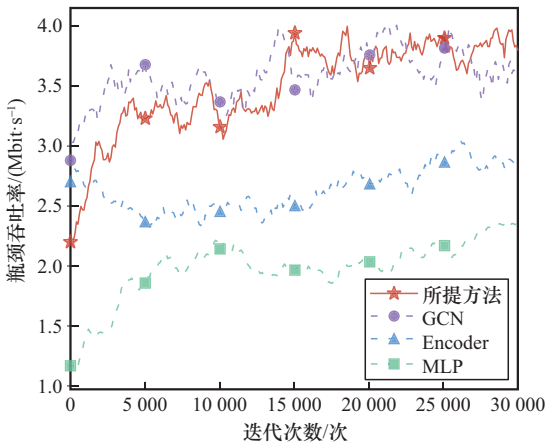


图 11　簇状拓扑下单数据流的瓶颈吞吐率优化曲线

图 12 和图 13 分别为簇状拓扑下信道切换次数和路径跳数优化曲线。经过训练，所提方法可以实现最低的信道切换次数与路径跳数，而 MLP 与 GCN 的优化结果则相对较高，并且 Encoder 的数值最高。结果表明，GAT 通过注意力机制有效识别了高质量、稳定的通信链路，在路径规划过程中倾向于选择信道状态良好且跳数较少的路由，从而避免不必要的频谱切换与冗余中继。因此，所提方法在簇状拓扑条件下仍能保持性能优势，验证了其对复杂、非规则网络结构的强适应性与泛化能力。

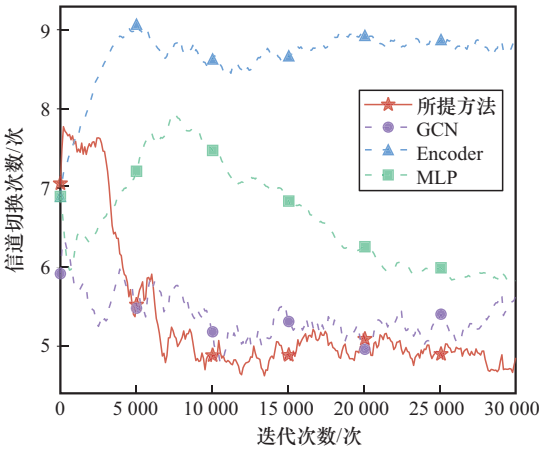


图 12　簇状拓扑下信道切换次数优化曲线

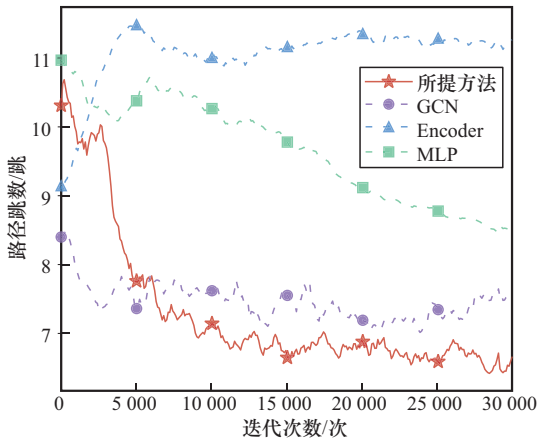


图 13　簇状拓扑下路径跳数优化曲线

4　结束语

面向无线自组网中高质量通信路径构建的挑战，本文提出了一种融合图注意力网络与深度强化学习的联合路由与频谱接入方法。将路径建立过程建模为部分可观测马尔可夫决策过程，并通过 DRL 实现分布式决策，克服了全局拓扑未知条件下的联合优化难题。在 DRL 框架内引入 GAT 对局部感知信息进行多头注意力聚合，有效捕捉非规则拓扑结构特征与节点间干扰关系，增强决策代理对

复杂网络环境的状态特征提取能力。在奖励函数设计中综合考虑链路瓶颈吞吐率与信道切换次数，实现了通信性能与切换开销的平衡，避免了频繁信道切换导致的性能波动。实验设置涵盖随机拓扑、多数据流共存及簇状非均匀分布等多种典型场景，且结果表明所提方法在瓶颈吞吐率、信道切换次数与路径跳数等关键指标上的先进性与鲁棒性。

参考文献：

[1] Fu B, Xiao Y, Deng H M, et al. A survey of cross-layer designs in wireless networks[J]. IEEE Communications Surveys & Tutorials, 2014, 16(1): 110-126.

[2] Popovski P, Stefanovic C, Nielsen J J, et al. Wireless access in ultra-reliable low-latency communication (URLLC)[J]. IEEE Transactions on Communications, 2019, 67(8): 5783-5801.

[3] Zhang L L, Zhang Y, Zheng J. Deep reinforcement learning based joint uplink and downlink resource allocation for URLLC[J]. IEEE Transactions on Vehicular Technology, 2025, 74(4): 6048-6063.

[4] Sivakumar R, Das B, Bharghavan V. Spine routing in ad hoc networks[J]. Cluster Computing, 1998, 1(2): 237-248.

[5] Cui W, Yu W. Scalable reinforcement learning for routing in ad-hoc networks based on physical-layer attributes[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 8118-8122.

[6] Le L, Hossain E. Cross-layer optimization frameworks for multihop wireless networks using cooperative diversity[J]. IEEE Transactions on Wireless Communications, 2008, 7(7): 2592-2602.

[7] Chen H, Baras J S. Distributed opportunistic scheduling for wireless ad-hoc networks with block-fading model[J]. IEEE Journal on Selected Areas in Communications, 2013, 31(11): 2324-2337.

[8] 周述淇, 陈发堂. 卫星互联网的高效频谱共享研究综述[J]. 光通信研究, doi: cnki.com.cn/Article/CJFDTotal-GTXY20250508002.
Zhou S Q, Chen F T. A survey on efficient spectrum sharing in satellite Internet[J]. Study on Optical Communications, doi: cnki.com.cn/Article/CJFDTotal-GTXY20250508002.

[9] 宋波, 叶伟, 孟祥辉. 基于多智能体强化学习的动态频谱分配方法综述[J]. 系统工程与电子技术, 2021, 43(11): 3338-3351.
Song B, Ye W, Meng X H. Review of multi-agent reinforcement learning based dynamic spectrum allocation method[J]. Systems Engineering and Electronics, 2021, 43(11): 3338-3351.

[10] 何业军, 黄伟, 陈亚玲, 等. 支持未来全频谱接入通信的基站天线研究综述[J]. 电波科学学报, 2023, 38(1): 15-26, 43.
He Y J, Huang W, Chen Y L, et al. Review of base station antennas for future all-spectrum-access communications[J]. Chinese Journal of Radio Science, 2023, 38(1): 15-26, 43.

[11] 马彬, 杨祖敏, 谢显中. 认知车联网中评估频谱稳定性的动态频谱接入算法[J]. 电子与信息学报, 2025, 47(5): 1474-1485.
Ma B, Yang Z M, Xie X Z. Dynamic spectrum access algorithm for evaluating spectrum stability in cognitive vehicular networks[J]. Journal of Electronics & Information Technology, 2025, 47(5): 1474-1485.

[12] 黄柳碧, 王威, 曹平, 等. 面向非可信节点的联合频谱感知与资源分配机制设计[J]. 物联网学报, doi: cnki.com.cn/Article/CJFDTotal-WLWX20250724002.
HUANG L B. WANG W, CAO P,et al. Joint spectrum sensing and resource allocation mechanism with untrusted sensing nodes[J]. Chinese Journal on Internet of Things, doi: cnki. com. cn/Article/CJFDTotal-WLWX20250724002.

[13] 李梓豪. Femto-cell网络频谱分配及能效算法研究[D]. 长沙: 中南大学, 2024.
Li Z H. Research on spectrum allocation and energy efficiency algorithms for Femto-cell networks[D]. Changsha: Central South University, 2024.

[14] 牛阳阳, 尉志青, 冯志勇. 基于博弈论的空时频谱共享: 动态接入与惩罚策略[J]. 通信学报, 2023, 44(12): 28-38.
Niu Y Y, Yu Z Q, Feng Z Y. Space-time spectrum sharing based on game theory: dynamic access and penalty strategy[J]. Journal on Communications, 2023, 44(12): 28-38.

[15] Pang X W, Tang J, Zhao N, et al. Energy-efficient design for mmWave-enabled NOMA-UAV networks[J]. Science China Information Sciences, 2021, 64(4): 140303.

[16] Du Y, Yang K, Wang K Z, et al. Joint resources and workflow scheduling in UAV-enabled wirelessly-powered MEC for IoT systems[J]. IEEE Transactions on Vehicular Technology, 2019, 68(10): 10187-10200.

[17] Mahfouzi R, Aminifar A, Samii S, et al. Stability-aware integrated routing and scheduling for control applications in Ethernet networks[C]// Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition. Piscataway:IEEE Press, 2018: 682-687.

[18] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J].arXiv Preprint, arXiv: 1312.5602, 2013. .

[19] Wang Z Y, Freitas N D, Lanctot M. Dueling network architectures for deep reinforcement learning[J]. arXiv Preprint, arXiv: 1511.06581 2015.

[20] Zhou X H, Xiong J, Zhao H T, et al. Joint UAV trajectory and communication design with heterogeneous multi-agent reinforcement learning[J]. Science China Information Sciences, 2024, 67(3): 132302.

[21] Zormati M A, Lakhlef H, Ouni S. Routing optimization based on distributed intelligent network softwarization for the Internet of things[C]// Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing.New York: ACM Press,2024: 1757-1764.

[22] 王子傲. 基于深度强化学习的动态频谱接入算法研究[D]. 北京: 北京邮电大学, 2024.
Wang Z A. Research on dynamic spectrum access algorithms based on deep reinforcement learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2024.

[23] 张晶, 马林, 高宏旭, 等. 基于频谱感知和非正交多址的大规模免授权随机接入方案[J]. 通信学报, 2025, 46(3): 151-163.
Zhang J, Ma L, Gao H X, et al. Spectrum sensing and non-orthogonal multiple access based massive grant-free random access scheme[J]. Journal on Communications, 2025, 46(3): 151-163.

[24] 于越, 陈玲玲, 刘文刚, 等. 基于D3QN的认知物联网动态频谱接入[J]. 信息技术与信息化, 2024(12): 69-72.
Yu Y, Chen L L, Liu W G, et al. Dynamic spectrum access of cognitive

Internet of Things based on D3QN[J]. Information Technology and Informatization, 2024(12): 69-72.

[25] 李鹏飞. 多业务非规则场景下频谱资源智能分配技术研究[D]. 西安: 西安电子科技大学, 2024.
Li P F. Research on intelligent allocation technology on spectrum resources in multi business irregular scenarios[D]. Xi'an: Xidian University, 2024.

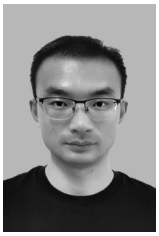
[26] 陈平平, 张旭, 谢肇鹏, 等. 基于多智能体近端策略优化的多信道动态频谱接入[J]. 电子学报, 2024, 52(6): 1824-1831.
Chen P P, Zhang X, Xie Z P, et al. Multi-channel dynamic spectrum access based on multi-agent proximal policy optimization[J]. Acta Electronica Sinica, 2024, 52(6): 1824-1831.

[27] Gu Y F, She C Y, Quan Z, et al. Graph neural networks for distributed power allocation in wireless networks: aggregation over-the-AI[J]. arXiv Preprint, arXiv:2207.08498, 2023.

[28] Almasan P, Suárez-Varela J, Rusek K, et al. Deep reinforcement learning meets graph neural networks: exploring a routing optimization use case[J]. Computer Communications, 2022, 196: 184-194.

[29] Dong T J, Zhuang Z R, Qi Q, et al. Intelligent joint network slicing and routing via GCN-powered multi-task deep reinforcement learning[J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(2): 1269-1286.

[作者简介]



周子铂（1994– ），男，江西九江人，军事科学院系统工程研究院博士生，主要研究方向为智能信号处理、无线资源配置与优化。

任保全（1974– ），男，陕西西安人，博士，军事科学院系统工程研究院研究员、博士生导师，主要研究方向为物联网、无线通信、移动通信网络技术等。

钟旭东（1991– ），男，湖南常德人，博士，军事科学院系统工程研究院副研究员，主要研究方向为智能信息网络技术、卫星通信、无线资源管理与优化。

刘琦（1993– ），女，吉林磐石人，军事科学院系统工程研究院博士生，主要研究方向为通信网络架构设计、智能化评估。

秦蓁（1995– ），女，山东济南人，军事科学院系统工程研究院在站博士后，主要研究方向为低空智能信息网络、边缘智能。