面向海上异构无线网络多路径协同拥塞控制

陈津，周轩，何如龙
（海军工程大学电子工程学院，湖北 武汉 430033）

摘 要：对海上异构无线网络多路径协同拥塞控制问题进行了研究，提出了多智能体强化学习与图神经网络（GNN，Graph Neural Network）增强的协同拥塞控制算法 MACC（Multi-Agent Reinforcement Learning for Cooperative Congestion Control），动态优化多路径传输参数以提高网络综合效用。该算法通过链路类型编码与交叉注意力融合实现异构状态统一表征，并耦合传输层、网络层与链路层多维反馈构建跨层耦合奖励，同时利用 GATv2 消息传递机制编码时变拓扑，自主调整各链路拥塞窗口与切换决策以适应动态拓扑和信道变化，从而缓解异构多链路间的非平稳性问题。算法采用集中式训练去中心化执行模式，通过单调混合网络将全局价值函数分解为各智能体的个体价值，适用于高维状态动作空间的多智能体协同决策。仿真结果表明，与 MPTCP、Aurora、QMIX 和 DDPG 相比，所提算法在吞吐量、端到端时延、丢包率及公平性方面表现更优。

关键词：海上异构无线网络；多路径协同拥塞控制；多智能体强化学习；图神经网络
中图分类号：TN915.03
文献标志码：A
doi：10.11959/j.issn.2096−3750.

Multipath Cooperative Congestion Control for Maritime Heterogeneous Wireless Networks

CHEN Jin, ZHOU Xuan, HE Rulong
College of Electronic Engineering, Naval University of Engineering, Wuhan 430033, China

Abstract: This paper investigates the cooperative congestion control problem for multipath in maritime heterogeneous wireless networks. We propose MACC (Multi-Agent Reinforcement Learning for Cooperative Congestion Control), a cooperative congestion control algorithm enhanced by multi-agent reinforcement learning and Graph Neural Networks (GNN), which dynamically optimizes the transmission parameters of multipath to improve the overall network utility. The algorithm achieves a unified representation of heterogeneous states through link-type encoding and cross-attention fusion, and constructs a cross-layer coupled reward by integrating multi-dimensional feedback from the transport, network, and link layers. Meanwhile, it leverages the GATv2 message-passing mechanism to encode time-varying topologies, autonomously adjusting the congestion window and switching decisions of each link to adapt to dynamic topology and channel variations, thereby alleviating the non-stationarity problem arising among heterogeneous links. Adopting the paradigm of centralized training with decentralized execution (CTDE), the algorithm factorizes the global value function into the individual values of per-link agents through a monotonic mixing network, making it well suited for multi-agent cooperative decision-making in high-dimensional state-action spaces. Simulation results demonstrate that, compared with MPTCP, Aurora, QMIX, and DDPG, the proposed algorithm achieves superior performance in terms of throughput, end-to-end delay, packet loss rate, and fairness.

收稿日期：XXXX−XX−XX；**修回日期：**XXXX−XX−XX
通信作者：何如龙，0909031043@nue.edu.cn
基金项目：国家自然科学基金（61871473），海军工程大学自主研发计划(202509)
Foundation Items: The National Natural Science Foundation of China (61871473), The Naval Engineering University’s Independent Research and Development Program Project (202509)

Key words: maritime heterogeneous wireless networks, multipath congestion control, multi-agent reinforcement learning, graph neural networks

0 引言

海事物联网（Maritime Internet of Things，MIoT）与智能航运的快速发展，对海洋信息基础设施提出了前所未有的通信需求[1]。海上通信环境的物理特性显著区别于陆地网络，船舶移动与海浪起伏导致网络拓扑持续变化，海上无线信道受雨衰、多径衰落与蒸发波导效应影响使得误码率较高，并且随着地理位置与气象条件动态变化，导致链路频繁中断[2-3]。严重的流量波动导致网络拥塞，进而引发网络丢包并降低网络可靠性。因此，如何有效管理与优化海上无线信道资源，避免拥塞发生，是当前海上异构无线网络（MHWNs，Marine Heterogeneous Wireless Networks）亟需解决的重要问题。

Li等人提出了一种基于Q学习的多路径拥塞控制方法SmartCC，采用异步强化学习框架学习一组拥塞控制规则，使发送端能够观测环境并执行动作，自适应地调整子流拥塞窗口以适配不同的网络状况，以应对异构网络中多通信路径的差异化特性[4]。在此基础上，Han等人开展了多路径传输控制协议（MPTCP，Multipath Transmission Control Protocol）增强研究，提出一种自适应控制框架OLSch，通过整合不同方案并发挥其各自优势，以实现持续稳定的高性能传输[5]。虽然TCP拥塞控制技术针对秒级时间尺度的流量波动效果明显，但是面对短时突发特点的亚秒级流量波动时，难以适用于实时通信、远程控制等延迟敏感业务[6-8]。

针对这一问题，已有研究通过融合软件定义网络（SDN，Software Defined Networking）的集中控制特性与命名数据网络[9]（NDN，Named Data Networking）的内容中心通信范式，设计了分层拥塞控制框架[10]，以适用于由时变拓扑协作多层网络构成的海上开放通信环境。在全局层面，SDN控制器采用纳什议价博弈[11]，依据接口质量与信息时效性约束，动态计算各接口的最优转发权重；在本地层面，中间转发节点建立双向拥塞检测机制，监测兴趣队列长度与数据包新鲜度（AoI，Age of Information）[12-13]。检测结果反馈至消费者端以实现实时流量调整，形成集感知、反馈与自适应于一体的闭环控制系统[14]。由于缺乏学习历史经验的能力，难以对拥塞控制策略进行精准调整，并将网络收敛至新的平衡。因此陆续有研究者提出了基于机器学习的拥塞控制算法以提升其泛化与自优化能力[15]。文献[16]提出了一种基于深度强化学习的拥塞控制框架Aurora，该框架通过PPO算法训练神经网络策略，能够依据网络统计信息动态映射发送速率，在单流时变链路中实现了较高的链路利用率与吞吐量。然而，该框架仅在单流环境中进行训练，缺乏对多流竞争场景下公平性的显式约束，难以适应多流共存及复杂不对称的网络拓扑结构。文献[17]针对Aurora在多流环境下的公平性缺陷，提出了奖励重塑、状态增强与丢包敏感度调谐三种后处理公平性策略，在保持聚合吞吐量不变的前提下有效改善了同构Aurora流间的带宽分配公平性。然而，这些策略主要基于流体级仿真器与固定架构背景流假设，在异构协议共存及动态流加入退出等复杂场景下仍面临公平性波动与策略退化风险。

尽管这种分层拥塞控制框架在自适应流量调控展现出潜力，但是本质上仍然遵循单智能体或集中式控制范式，难以有效应对海上异构无线网络中多路径协同拥塞控制[18]。更为关键的是，海上异构无线网络中各链路共享船舶端瓶颈缓冲区，一条链路的拥塞窗口调整将直接影响其他链路的可用带宽与队列状态，形成典型的多智能体竞争与协作并存环境[19]。现有方案缺乏对链路间非平稳性与信用分配问题的显式建模，当网络拓扑因船舶移动动态变化时，固定启发式规则或单智能体策略难以维持稳定的协同性能[20]。

近年来，集中式训练去中心化执行的多智能体深度强化学习（MARL, Multi-Agent Reinforcement Learning）在解决分布式决策与全局优化矛盾方面展现出独特优势[21]。以QMIX[22]和DDPG[23]为代表的拥塞控制算法，通过将每条异构链路建模为独立决策智能体，MARL能够在训练阶段利用全局信息学习协同策略，在执行阶段仅依赖局部观测实现去中心化拥塞控制，有效缓解多链路交互引发的环境非平稳性。然而，将MARL直接应用于海上异构

网络仍面临异构状态空间统一表征、跨层目标耦合与拓扑动态感知三项核心挑战[24]。

鉴于上述局限性，本文提出一种面向海上异构无线网络的多路径协同拥塞控制方法MACC（Multi-Agent Reinforcement Learning for Cooperative Congestion Control）。该方法基于集中式训练去中心化执行框架，在集中训练阶段利用全局状态信息学习协同策略，在去中心化执行阶段由各智能体仅依据局部观测动态调整拥塞窗口与切换决策，实现亚秒级的网络拥塞控制。在此框架下，为实现分布式设备间协同管理和网络效用最大化，设计了多智能体强化学习拥塞控制算法，通过单调混合网络将全局价值函数分解为各智能体的个体价值，保证各智能体局部贪婪决策与全局联合最优决策的一致性。仿真结果表明，与MPTCP、Aurora、QMIX和DDPG相比，所提算法在吞吐量、端到端时延、丢包率及公平性方面表现更优。

1　系统模型

本节建立海上异构网络拥塞控制模型。首先定义时变网络拓扑与异构链路特征，随后建立多链路传输的排队模型，最后将问题形式化为分散式部分可观测马尔可夫决策过程[25]（Dec-POMDP，Decentralized Partially Observable Markov Decision Process），并完整刻画状态空间、动作空间、状态转移函数和观测模型。

1.1　网络模型

对网络拓扑建模并表示为$G=(V,L)$，其中$V=\{v_1,v_2,v_3,.....v_n\}$表示路由节点集合，$L=\{l_1,l_2,l_3,.....l_m\}$表示链路集合。海上异构无线网络典型通信场景如图1所示，UAV、船只、USV节点平台借助无中心化扩展性强的自组网、高带宽低时延的5G专网、长距离低带宽的卫星通信、短距离高带宽的WIFI技术进行组网通信。依托SDN集中式架构，将海上USV、船只及空中UAV共同组建为数据平面。基站同时担负控制平面与数据平面的职责，负责网络拓扑管理和路由策略优化工作，并定期向数据平面下发流表规则。

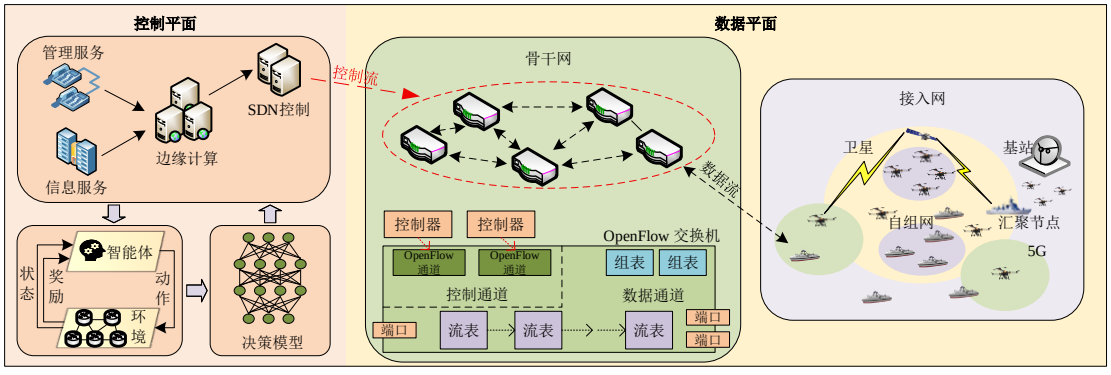

图1　海上异构无线网络典型通信场景

每条链路$l\in L$由四维向量$c_l=\{B_l,D_l,L_l,A_l\}$刻画，其中B_l表示链路带宽，D_l表示传输时延，L_l表示丢包率，$A_l=\dfrac{\mathrm{MTBF}}{\mathrm{MTBF}+\mathrm{MTTR}}$表示链路可用性，其中MTBF为平均无故障时间，MTTR为平均修复时间，直接影响状态转移模型中的链路失效概率。

每个节点$v\in V$配置有限缓冲容量的队列。针对链路l上的队列，本文采用M/D/1/K排队模型[26]，分组到达过程服从参数为$\lambda_{v,l}$的泊松分布，服务时间为固定值$\dfrac{1}{\mu_{v,l}}$，即服务率$\mu_{v,l}$为固定常数而非时变量，最大缓冲区容量为$K_{\max}$个分组。在决策时隙尺度上采用流体近似，以时隙内期望到达率刻画泊松到达的统计平均行为，队列占用率$q_{v,l}^t\in[0,1]$的演化过程定义为

$$q_{v,l}^{t+1}=\min\left\{1,\max\left\{0,q_{v,l}^t+\frac{\lambda_{v,l}^{in}(t)-\mu_{v,l}^{out}(t)}{K_{\max}}\cdot\Delta t\right\}\right\}\tag{1}$$

其中，$\lambda_{v,l}^{in}(t)$表示聚合到达率，$\mu_{v,l}^{out}(t)$为服务率，Δt为决策时隙长度。当队列处于满容量状态时，新到达的分组将被丢弃。

1.2　动态队列协同管控模型

为抵消流量波动，引入动态队列协同管控模型，如图2所示，每条链路具备不同数据传输能

力，用 C_l 表示链路 l 的传输能力。网络转发设备通过端口缓存队列来平滑输入与输出链路间的速率失配，当输入流量较大时，缓存队列暂存超额分组，并依据链路状态与输出端口负载实施动态调度，进而实现流量整形、抑制网络抖动并提升传输可靠性。假设网络中有 I 组流量入口出口对（IE，Ingress-Egress），每队 $i \in I$ 的流量需求定义为 d_i，各网络节点均配置尾部丢弃队列，记节点 v 在 t 时刻的队列容量表示为 $q_l^v(t)$，其最大容量表示为 $q_{\max}^v$。队列长度的增量 ΔQ 定义为

$$\Delta Q = \sum_{G_v^{out} \in G_{V'}^{out}} \sum_{i \in I} \left(\Delta \alpha_{G_v^{out}}^i \sum_{G_v^{in} \in G_{V'}^{in}} \beta_{G_v^{in}}^i a_{G_v^{in}}(t) - \Delta s_{G_v^{out}}^i(t) \right) \tag{2}$$

其中，G_v^{in} 和 G_v^{out} 分别表示网络节点 v 的输入端口和输出端口，$\alpha_{G_v^{out}}^i$ 表示输出端口所输出流量占总流量的比例，$\beta_{G_v^{in}}^i$ 表示源-目的对 i 中流量在端口 G_v^{in} 中总流量的比例，$a_{G_v^{in}}$ 表示输入端口的数据到达率，$s_{G_v^{out}}^i$ 表示网络节点 v 对第 i 组流量在端口 G_v^{out} 上的服务速率。

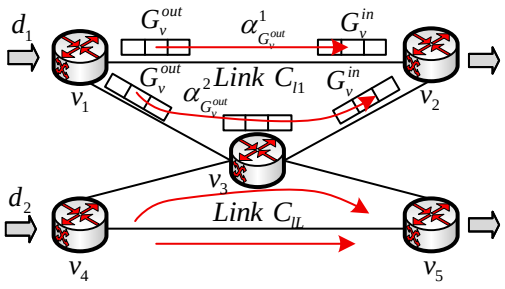

图2　动态队列协同管控模型

由于网络流量波动，导致输入端口瞬到达率可能短时突破设备峰值服务能力，进而引发拥塞丢包[27]。为描述流量瞬时波动，引入波动变量 d_i，并将 $\Delta d_i(t)$ 表示独立同分布的随机变量（如零均值正态分布）。到达速率 a 与服务速率 s 的波动满足如下性质

$$\begin{cases} a_{G_v^{in}}^i(t) = a_{G_v^{in}}^i + \Delta a_{G_v^{in}}^i(t) \\ s_{G_v^{out}(t)}^i = s_{G_v^{out}}^i + \Delta s_{G_v^{out}(t)}^i \\ \Delta a_{G_v^{in}}^i(t) = \sum_{i \in I} \Delta a_{G_v^{in}}^i(t) \\ \Delta s_{G_v^{out}(t)}^i = \sum_{i \in I} \Delta s_{G_v^{out}(t)}^i \\ a_{G_v^{in}}^i(t) \leqslant C_l \end{cases} \tag{3}$$

1.3　问题模型

将海上异构网络的多链路协同拥塞控制问题形式化为一个 Dec-POMDP 过程，其数学定义为

$$M = \langle N, S, \{A_i\}, P, \{O_i\}, \{\Omega_i\}, R, \gamma \rangle \tag{4}$$

其中 $N = \{1,2,.....,n\}$ 为智能体集合（每个智能体对应一条活跃链路），S 为全局状态空间，A_i 为智能体 i 的动作空间，$P: S \times A \times S \to [0,1]$ 为状态转移概率，O_i 为智能体 i 的观测空间，$\Omega_i: S \to O_i$ 为观测函数，$R: S \times A \to \mathbb{R}$ 为全局奖励函数，$\gamma \in [0,1)$ 为折扣因子。Dec-POMDP 是部分可观测随机博弈（POSG，Partially Observable Stochastic Game）在完全协作设置下的特例，所有智能体共享同一奖励函数。

全局状态 $s_t \in S$ 由链路状态、网络状态和环境状态三部分组成 $s_t = (s_{\text{link}}(t), s_{\text{net}}(t), s_{\text{env}}(t)) \in S$。链路状态包含每条链路的拥塞控制相关参数。对智能体 i 管理的链路 k，其状态分量为 $s_{i,k}^{\text{link}}(t) = (w_{i,k}(t), \text{RTT}_{i,k}(t), \theta_{i,k}(t), p_{i,k}(t), q_{i,k}(t))$。其中，$w_{i,k}(t)$ 为拥塞窗口大小，$\text{RTT}_{i,k}(t)$ 为往返时延，$\theta_{i,k}(t)$ 为吞吐量，$p_{i,k}(t)$ 为丢包率，$q_{i,k}(t)$ 为队列占用率。

网络状态 $s_{\text{net}}(t)$ 编码拓扑结构和流量需求为 $s_{\text{net}}(t) = (G(t), \{N_i^{\text{obs}}(t)\}_{i \in N}, \{\lambda_i(t)\}_{i \in N})$。其中，$G(t)$ 为时变网络拓扑图，$N_i^{\text{obs}}(t)$ 为智能体 i 的可见邻居集合，$\lambda_i(t)$ 为节点 v 的流量到达率。智能体通过网络状态，能够在拥塞控制决策中考虑拓扑上下文和流量需求变化。

环境状态 $s_{\text{env}}(t)$ 反映海洋信道的外部条件 $s_{\text{env}}(t) = (\text{SSC}(t), \text{v}_{\text{wind}}(t), \{\text{v}_v(t)\}_{i \in N})$。其中，$\text{SSC}(t) \in \{0,1,...,9\}$ 为道格拉斯海浪分级，$\text{v}_{\text{wind}}(t)$ 为风速，$\text{v}_v(t)$ 为节点移动速度。海浪分级直接影响海上信道衰落模型的参数设置，高 SSC 值对应更强的海面反射和散射损耗，从而改变链路的有效带宽和丢包率[28]。

考虑到由 N 条异构链路组成的并行传输系统，每条链路 $i \in N$ 对应一种网络模态（如卫星链路、5G 链路或自组网链路）。在每个时隙 t，链路 i 的拥塞控制智能体根据局部观测 $o_i(t)$ 选择动作 $a_i(t)$（如调整拥塞窗口或发送速率），其策略记为 $\pi_i(a_i|o_i)$。本文的优化目标是最大化全网效用函数，该函数综合反映了各链路的传输效用与时延代价、网络的全局公平性以及系统能耗，定义为

$$\max \sum_{i=1}^{N}\left[\alpha \cdot U(\theta_i) - \beta \cdot D(\theta_i)\right] + \gamma \cdot F(\theta) - \delta \cdot \sum_{i=1}^{N} E(\theta_i) \tag{5}$$
$$\text{s.t.} \sum_{i,j \in \mathrm{P}(\mathrm{i})} r_i(t) \leqslant C_j(t), \forall j \in L$$
$$\theta_i \in \Theta_i, d_i(t) \leqslant d_{\max}, \forall i \in N$$

其中，θ_i为链路i的拥塞控制参数向量，$U(\cdot)$为吞吐量效用函数，$D(\cdot)$为时延惩罚函数，$F(\cdot)$为跨链路公平性度量（Jain公平指数，公平指数定义为$F=\frac{(\sum_{i=1}^{n}x_i)^2}{n\sum_{i=1}^{n}x_i^2}$，其中$x_i$为第$i$条链路分配到的有效带宽，n为活跃链路数量，$F\in[\frac{1}{n},1]$，值越接近1表示分配越公平。），$E(\cdot)$为能耗函数；$\alpha,\beta,\gamma,\delta$为权重系数；$r_i(t)$为链路$i$的分配速率，$C_j(t)$为链路$j$的时变容量；$\Theta_i$为策略可行域，$d_i(t)$为端到端时延，$d_{\max}$为最大可接受时延。

2　算法设计

本文提出的MACC算法由四个核心组件构成，利用集中式训练去中心化执行（CTDE，Centralized Training with Decentralized Execution）框架实现可扩展的多智能体学习，使用异构状态统一编码器（UHSE，Unified Heterogeneous State Encoder）解决异构链路状态空间的表征不一致问题，通过跨层耦合奖励（CLCR，Cross-Layer Coupled Reward）机制将传输层、网络层和链路层的性能指标耦合为统一优化目标，依托图神经网络（GNN，Graph Neural Network）增强模块提供动态拓扑感知能力。四个组件的协同设计使得各链路智能体无需集中控制，即可在分布式执行中隐式逼近网络效用最大化。

2.1　多智能体协同架构

CTDE核心思想是在训练阶段利用全局状态信息学习协同策略，在执行阶段仅依赖局部观测进行去中心化决策[29]，有效缓解了由部分可观测性导致的环境非平稳性问题。集中训练阶段，中央协调器收集所有链路智能体的局部观测、动作和奖励信号，构建全局经验序列并存储于经验回放池；协调器基于全局状态训练混合网络，学习将全局价值函数分解为个体价值函数。去中心化执行阶段，各链路智能体仅基于局部观测和训练好的策略网络独立选择拥塞窗口调整动作，无需全局信息。图3展示了该架构的完整结构。

单调混合网络部署于集中式控制器，各链路智能体的个体Q网络部署于分布式网络设备。混合网络以全局状态S与各智能体的动作价值为输入，将其单调聚合为全局动作价值函数Q_{tot}，从而引导智能体的价值学习向全局最优收敛。各智能体的动作空间为离散集合，定义为拥塞窗口与链路切换决策的笛卡尔积。个体Q网络的输入是本地观测信息O_i，仅依靠本地感知数据对各候选离散动作进行价值评估。在策略执行过程中，智能体选择个体价值最大的动作执行，由混合网络的单调性保证该去中心化贪心决策与全局联合动作价值最大化一致，实现对网络状态的快速响应。

各链路智能体的个体Q网络参数记为θ，混合网络参数记为ω。全局动作价值函数由混合网络对n个智能体的个体动作价值进行单调聚合得到，其值分解表达式为

$$Q_{\text{tot}}(S,A) = f_\omega(S; Q_1(o_1,a_1), Q_2(o_2,a_2), ..., Q_n(o_n,a_n))$$
$$\text{s.t.} \frac{\partial Q_{\text{tot}}}{\partial Q_i} \geqslant 0, \forall i \in \{1,2,...,n\} \tag{6}$$

其中，S为全局状态，$A=(a_1,a_2,...,a_n)$为所有智能体的联合动作集合，o_i与a_i分别为智能体i的局部观测与动作，$Q_i(o_i,a_i)$为智能体i的个体动作价值函数，f_ω为单调混合网络。为保证个体贪心决策与全局价值最大化一致，混合网络须满足单调性约束$\frac{\partial Q_{\text{tot}}}{\partial Q_i}\geqslant 0$。

为训练上述值分解网络，引入时序差分误差（TD，Temporal Difference），以TD误差定义在线网络的回归损失函数为

$$L(\theta,\omega) = \mathbb{E}_{(s,a,r,s')\sim D}\left[\left(Q_{\text{tot}}(S,A;\theta,\omega) - y\right)^2\right] \tag{7}$$

其中，D为经验回放缓冲区，y为TD目标，表示为

$$y = r + \gamma \max_{a'} Q_{\text{tot}}(S',A';\theta^-,\omega^-) \tag{8}$$

其中，r表示当前状态下执行联合动作后获得的即时奖励，S'表示采取行动后转移到的下一状态，γ为折扣因子，θ^-和ω^-分别表示目标Q网络与目标混合网络的参数。得益于混合网络的单调性，式中对联合动作的最大化可等价分解为各智能体对各自离散动作的独立贪心选择，无需在执行阶段进行联合

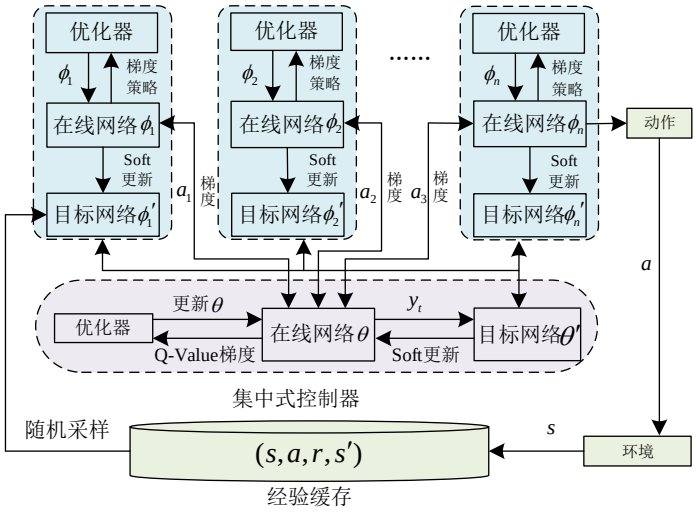

图3　集中式训练-去中心化执行模型

动作搜索。

为提升值分解网络训练的稳定性，引入经验回放机来削弱数据样本间的时间相关性。具体而言，智能体将与环境交互所产生的经验数据(s,a',r,s')存储至回放缓冲区D，参数更新时从D中随机采样一个小批量数据进行学习。这种先存储后重采样的方式有效缓解了因训练数据时间关联而引起的训练不稳定问题。

同时，为进一步提升训练稳定性，引入目标网络机制[30]。目标网络是一组与在线值分解网络结构相同，但参数更新滞后的独立网络。在更新参数时，利用目标网络计算目标值，以提供更稳定的学习目标，抑制参数更新带来的波动。算法执行期间，目标网络参数每隔若干步通过Soft更新从在线网络参数中缓慢同步，更新公式如下

$$\phi' = k\phi' + (1 - k)\phi \tag{12}$$

$$\theta' = k\theta' + (1 - k)\theta \tag{13}$$

2.2　异构状态统一表征

在海上异构网络中，若直接输入各链路状态，梯度更新将被大数值维度所主导，导致小数值维度（如MANET中吞吐量的微小变化）的有效信号被淹没。因此，需对异构原始特征进行统一映射，如图4所示，异构状态统一表征结构首先采用分层编码策略，通过独立的链路、网络、传输及环境编码器，分别对物理层特征（如SNR、BER）、队列状态、吞吐量、时延以及海况、节点移动性等异构输入进行非线性变换，提取出维度统一的隐含特征$h_i \in \mathbb{R}^{d_m}$。随后，这些多源隐含特征被输入至基于多头注意力的交叉注意力融合模块中，系统基于缩放点积注意力（Scaled Dot-Product Attention）并行计算不同状态特征间的相关性与权重分布，以深度捕捉各分层信息间的全局依赖关系。最后，交叉注意力融合后的综合语义表征通过映射函数f投影至特定的统一潜在空间，生成最终的状态统一表征向量$z \in \mathbb{R}^{d_z}$。通过这一端到端的处理流程，异构状态统一表征有效消除了网络协议栈不同层级及外部环境数据在语义与尺度上的异构性，为后续的智能决策与网络优化提供了紧凑且语义完备的全局状态输入。

对于每个模态$m \in \{\text{link,net,trans,env}\}$，编码器输出模态嵌入

$$h_m = f_\theta^{(m)}(x_m) \in \mathbb{R}^{d_m} \tag{14}$$

其中，$f_\theta^{(m)}$为对应模态的MLP编码器，x_m为该模态的原始输入特征。各编码器独立训练时采用层归一化（Layer Norm）和ReLU激活，确保不同模态的输出在数值范围上具有可比性。

多路径嵌入通过交叉注意力机制融合为统一状态表征。与直接拼接相比，交叉注意力允许模态间动态加权，即在拥塞严重时网络层和传输层特征应获得更高权重，在链路切换时链路层和环境特征更为关键。具体而言，以传输层嵌入为查询（Query），其余三个模态为键（Key）和值（Value），计算跨模态注意力权重

$$z_i = \sum_{m=1}^{M} \alpha_{i,m} \cdot h_m \tag{15}$$

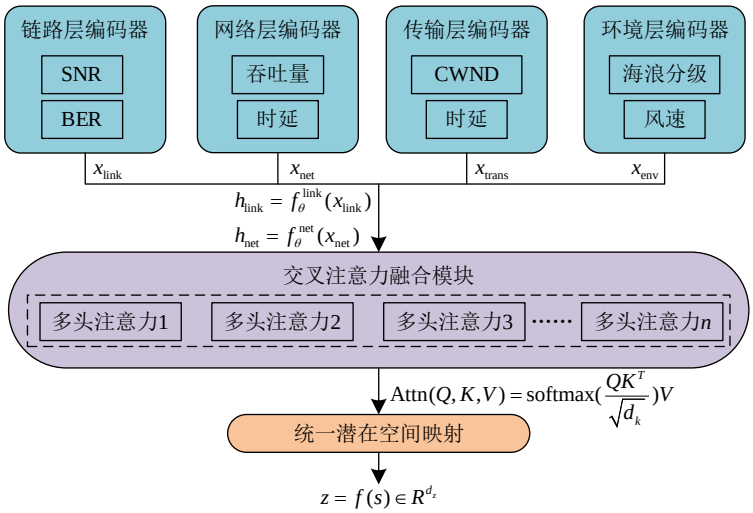

图4　异构状态统一表征结构

$$\alpha_{i,m} = \frac{\exp\left(\frac{h_{trans}^T W_Q^T W_K h_m}{\sqrt{d_k}}\right)}{\sum_{m'} \exp\left(\frac{h_{trans}^T W_Q^T W_K h_{m'}}{\sqrt{d_k}}\right)} \tag{16}$$

其中，M为模态数量，W_Q、W_K为可学习投影矩阵，d_k为注意力头维度。融合后的表征$z_i \in \mathbb{R}^{d_z}$通过残差连接和层归一化稳定训练，最终作为链路智能体i的策略网络输入。

2.3　跨层耦合奖励

传统基于强化学习的拥塞控制方法仅考虑传输层指标（吞吐量、时延、丢包率），忽略了网络层队列稳定性和链路层切换开销。因此，利用跨层耦合奖励机制[31]将OSI协议栈多个层次的性能信号耦合为综合奖励，使智能体在优化端到端性能的同时维持网络基础设施的稳定性。如图5所示，该机制首先针对传输层、网络层和链路层分别构建了具备物理意义的子奖励函数，其中传输层通过公式$r_{trans} = \alpha \cdot \log(\theta) - \beta \cdot D$实现吞吐量与时延的动态权衡，网络层以$r_{net} = -\gamma \cdot q_{occ}$评估队列稳定性，而链路层则利用$r_{link} = -\delta \cdot C_{ho} - \varepsilon \cdot N_{fail}$对切换开销与传输失败进行严格惩罚，且各层之间明确设定了耦合关联，强调状态间的相互制约。在提取各层级基础奖励的基础上，系统引入自适应权重网络，该网络以当前全局状态s_t为输入，通过非线性映射生成符合凸集约束的权重向量$\mathrm{w} = g_\phi(s_t) \in \Delta^3$，并利用Softmax确保各层权重满足$\omega_{trans} + \omega_{net} + \omega_{link} = 1$，从而实现对多目标的自适应动态侧重。最终，系统将上述加权子奖励与全局协同正则项相融合，计算总奖励$r_{total} = \omega_{trans} r_{trans} + \omega_{net} r_{net} + \omega_{link} r_{link} + \eta \cdot J(\{\theta_i\})$，其中最后一项通过超参数$\eta$控制全局策略熵或参数约束，旨在避免决策陷入局部最优，实现跨层耦合约束下的多目标协同与端到端全局优化。

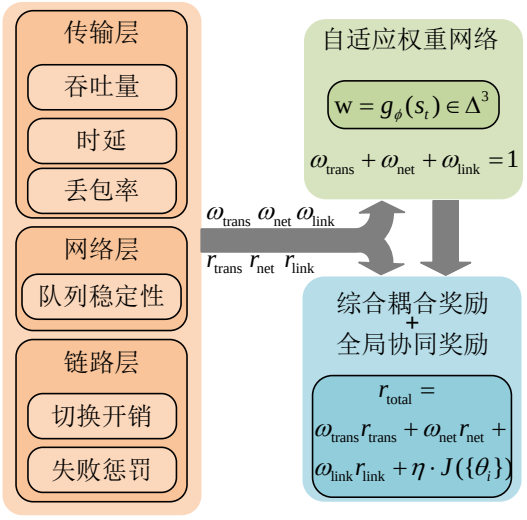

图5　跨层耦合奖励机制

传输层奖励r_{trans}基于吞吐量、时延、丢包率权衡设计，采用对数形式使奖励对吞吐量变化具有适度敏感性

$$r_{trans} = \alpha \cdot \log\left(1 + \frac{\theta}{\theta_{ref}}\right) - \beta \cdot \frac{D}{D_{ref}} - \zeta \cdot p_{loss} \tag{17}$$

其中，θ为吞吐量，D为平均端到端时延，p_{loss}为丢包率，θ_{ref}和D_{ref}为用于归一化的参考值，α、β、ζ为平衡系数。

网络层奖励 r_{net} 激励队列稳定性，防止缓冲区溢出导致的尾部丢包

$$r_{\text{net}} = -\gamma \cdot q_{occ} - \gamma' \cdot |\Delta q_{occ}| \tag{18}$$

其中，$q_{occ} \in [0,1]$ 为队列占用率，Δq_{occ} 为队列变化率，惩罚剧烈队列振荡，鼓励平滑的拥塞控制行为。

链路层奖励 r_{link} 惩罚不必要的链路切换和链路故障

$$r_{\text{link}} = -\delta \cdot C_{ho} \cdot \mathbb{I}_{\text{switch}} - \varepsilon \cdot N_{\text{fail}} \tag{19}$$

其中 C_{ho} 为切换开销常数，$\mathbb{I}_{\text{switch}}$ 为切换指示函数，N_{fail} 为监测窗口内的链路失败次数。

多目标奖励的简单加权求和存在两个固有缺陷，奖励信号在量纲和数值范围上的差异容易导致某一维度主导梯度更新，并且固定权重难以适应网络状态的动态变化[32]，例如轻载时应侧重吞吐量，拥塞时则应侧重时延与队列长度。针对这些问题，跨层耦合奖励机制采用自适应权重网络，以当前全局状态为输入，通过 Softmax 输出动态权重

$$\mathrm{w} = g_{\phi}(s_t) = \text{softmax}\left(\text{MLP}_{\phi}(s_t)\right), \mathrm{w} \in \Delta^3 \tag{20}$$

其中 Δ^3 为三维概率单纯形，g_{ϕ} 为轻量级权重网络。综合奖励为加权组合加上全局公平性激励

$$r_{\text{total}} = \omega_{\text{trans}} \cdot \bar{r}_{\text{trans}} + \omega_{\text{net}} \cdot \bar{r}_{\text{net}} + \omega_{\text{link}} \cdot \bar{r}_{\text{link}} + \eta \cdot J(\{\theta_i\}) \tag{21}$$

其中，$\bar{r}$ 为归一化后的层奖励（消除量纲差异），$J\left(\{\theta_i\}\right)$ 为各链路吞吐量的 Jain 公平性指数，η 为公平性权重。权重网络与策略网络联合训练，使奖励权重能够根据网络负载、海况和拓扑变化自适应调整。

2.4　图神经网络增强拓扑感知

海上网络的拓扑时变性和链路间耦合关系难以被独立 Q 网络捕获。本文将网络拓扑建模为时变带权图 $G = (V,L)$，其中边权重 $\omega_{ij}(l)$ 编码链路质量估计为

$$\omega_{ij}(l) = \frac{\text{SNR}_{ij}(l)}{\text{SNR}_{\max}} \cdot (1 - p_{\text{loss}}(l)) \cdot A_{ij}(t) \tag{22}$$

边权重综合了信噪比、丢包率和链路可用性三个指标，取值范围为 $[0,1]$。当链路质量恶化时，对应边权重降低，图神经网络（GNN，Graph Neural Networks）消息传递中该链路的邻居贡献减弱。

本文采用 GATv2（Graph Attention Network v2）[33] 实现动态注意力聚合。与图卷积网络（GCN，Graph Convolutional Networks）对所有邻居执行统一加权平均不同，GATv2 允许每条链路自适应地调整对邻居信息的关注权重，从而更灵活地刻画非均匀的拓扑依赖关系。节点 i 在第 l 层的消息传递过程为

$$e_{ij}^{(l)} = \text{LeakyReLU}\left(a^{(l)T}\left[W^{(l)}h_i^{(l)} \,//\, W^{(l)}h_j^{(l)}\right]\right) \tag{23}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k \in N_i} \exp(e_{ik}^{(l)})} \tag{24}$$

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right) \tag{25}$$

其中，$W^{(l)}$ 为第 l 层可学习权重矩阵，$a^{(l)}$ 为注意力参数向量，$//$ 表示向量拼接，N_i 为节点 i 的邻居集合，σ 为 ELU 激活函数。

GNN 输出的拓扑感知节点嵌入与异构状态统一表征输出的链路级状态表征融合，形成增强状态表示

$$\tilde{z}_i = \text{MLP}_{\text{fuse}}\left(\left[z_i \,//\, h_i^{(L)}\right]\right) \tag{26}$$

其中，z_i 为异构状态统一表征输出的统一链路状态表征（式 12），$h_i^{(L)}$ 为经过 L 层 GATv2 消息传递后的拓扑感知嵌入，$[\,\cdot\,//\,\cdot\,]$ 表示向量拼接，MLP_{fuse} 为融合网络。增强表征 $\tilde{z}_i$ 同时包含链路自身状态信息和拓扑上下文信息，使链路智能体能够感知邻居链路的拥塞状态并做出协同决策。

MACC 算法的完整流程如算法 1 所示。

3　仿真实验与性能评估

为验证 MACC 算法在海上异构网络中的有效性，本文基于 NS-3 仿真平台构建了高保真度的海上通信场景，从端到端性能、公平性、动态适应性三个维度开展系统性评估，并通过消融实验量化分析各核心组件的贡献。

3.1　实验设计

本文基于 NS-3 网络仿真器构建仿真实验平台。硬件平台为具有 24 核心 32 线程、256GB RAM 的 Intel i9-13900K CPU、NVIDIA RTX 4090 GPU 和 2TB 硬盘容量。其中 NS-3 离散事件仿真运行于 CPU 端，GPU 用于 MACC 多智能体强化学习算法的策略训练，仿真环境与训练框架之间通过 ns3-gym 接口完成网络状态观测与奖励信号的交互。为满足多智能体协同训练需求，本文对 ns3-gym 的 OpenGym 消息结构进行了扩展，使其支持多智能体观测与动作的同步交互，实现多智能体与仿真环

算法1　多智能体协同拥塞控制算法MACC

输入：链路智能体数量n，最大回合数$M_{\max}$，回合时隙数T，批量大小B，折扣因子γ，学习率，软更新系数k

输出：训练完成的各链路智能体个体Q网络参数

集中式训练阶段：

1　初始化个体Q网络参数θ、混合网络参数ω、UHSE编码器、GATv2模块与自适应权重网络参数；目标网络参数$\theta^{-} \leftarrow \theta$，$\omega^{-} \leftarrow \omega$；清空经验回放缓冲区$D$

2　for 回合 m = 1 to $M_{\max}$ do

3:　重置NS-3仿真环境，获取初始全局状态S_o与各智能体局部观测o_i

4:　for 时隙 t = 1 to T do

5:　构建时变带权图$G(t)$，更新边权重与邻域快照

6:　for 每个链路智能体 i = 1 to n do

7:　对链路、网络、传输、环境四模态观测编码，经交叉注意力融合得到统一状态表征z_i

8:　执行L层GATv2消息传递得到拓扑感知嵌入，与z_i融合得到增强表征$\tilde{z}_i$

9:　ε - *greedy*基于$Q_i(\tilde{z}_i,\cdot)$选择动作$a_{i,t}$

10:　end for

11:　执行联合动作a_t，NS-3环境演化一个决策时隙

12:　计算传输层、网络层、链路层子奖励，权重网络生成自适应权重，合成含公平性激励的总奖励r_t

13:　将转移样本$(S_t,o_t,a_t,r_t,S_{t+1})$存入回放缓冲区$D$

14:　if　D中样本量 ≥ B　then

15:　从D中随机采样小批量样本，计算TD目标并最小化TD损失，更新在线网络参数θ与ω

16:　按软更新规则同步目标网络参数θ^{-}与ω^{-}

17:　end if

18:　end for

19:　end for

去中心化执行阶段：

20:　各链路智能体加载训练完成的个体Q网络，仅依据局部观测独立执行步骤5~9选择动作，无需全局状态与混合网络

境之间的并发状态采集与联合动作下发。

步骤时间为1s，回合最大时间为50s（50步），训练批量大小为300回合，最大回合数20000回合，折扣因子0.95，目标网络软更新系数k=0.005，学习率为1×10^{-4}。在信道建模方面，海上场景以Two-Ray Ground模型为基础路径损耗模型，用于捕捉海面反射引起的衰减特性。针对接近视距边缘的中长距离链路，本文在双射线地面反射传播模型基础上通过SetNext级联自研的第三径镜面反射分量来刻画蒸发波导效应。小尺度衰落采用Nakagami-m分布，m取3以表征开阔海域以直射分量为主的信道特征，阴影衰落使用对数正态分布建模。仿真区域设定为1.5km×1.5km的开放海域，水面船只采用Gauss-Markov移动模型，以刻画船舶航向与航速的平滑时变特性，航速在3~10 m/s之间随机选取。无人机中继节点采用三维航点移动模型，巡航高度固定为150 m，飞行速度在15~25 m/s之间随机选取，在航点处的暂停时间服从[5, 20]秒的均匀分布，以适应空中中继快速部署与机动的业务需求。如图6所示，仿真拓扑包含50艘船只节点、6架无人机中继节点和4个浮标基站，共60个网络节点。

为全面评估算法在不同业务类型下的表现，实验采用三种流量模型混合注入，如表1所示。三类流量同时注入持续300秒的仿真时长，重复运行30次取统计平均值。

3.2　实验结果分析

实验选取MPTCP、Aurora、QMIX和DDPG这四种基准算法进行对比，对比算法所设置的参数在3.1节中有具体介绍。如图7所示，可以发现平均奖励随着学习过程的继续而增加。MACC在2500回合训练内平均奖励急剧增长，在10000回合后稳定于固定区间，且进入稳态后波动幅度明显小于Aurora与DDPG，表明集中式训练去中心化执行框架下的协同策略学习与跨层耦合奖励设计，使各链路智能体在动态海上环境中能够快速形成稳定有效的协同拥塞控制策略。Aurora凭借单智能体PPO在2500回合奖励快速攀升，但由于缺乏多智能体协同机制与公平性约束，其曲线在整个训练过程中剧烈震荡，始终无法收敛到稳定策略。QMIX的收敛曲线最为平滑，但其稳态奖励仅维持在较低水平，说明在缺少异构状态统一表征与拓扑感知增强的条件下，基于值分解网络的表达能力不足以充分挖掘多链路协同增益。DDPG在5000回合前长期陷入负奖

表1　仿真详细参数设置

流量模型	业务类型	单流速率	流数量	聚合速率	占比	特点
CBR	遥测信令	64kbps	625	40Mbps	40%	时延敏感、带宽需求低
Poisson突发	视频监控	每流均值在1~10 Mbps随机抽取，总体均值5 Mbps	7	35Mbps	35%	显著的带宽波动
On-Off	大型传输业务	忙时20 Mbps，占空比25%	5	25Mbps（均值）	25%	弹性带宽

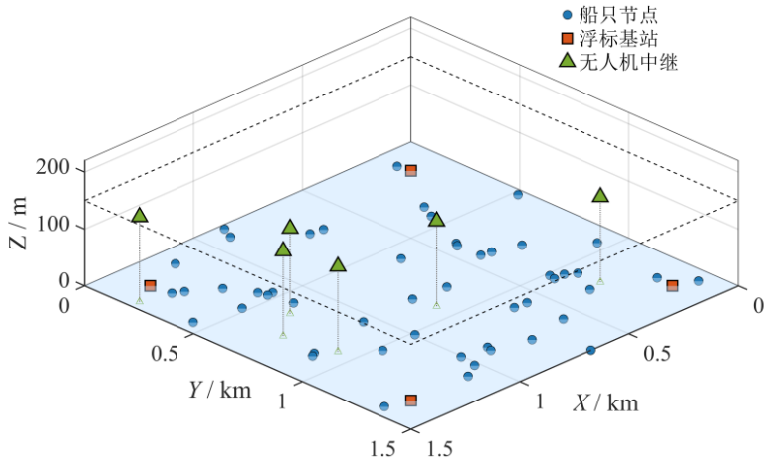

图6　仿真拓扑图

励区域，收敛速度最为缓慢且伴随大幅波动。作为非学习型基准，MPTCP 不具备训练过程。

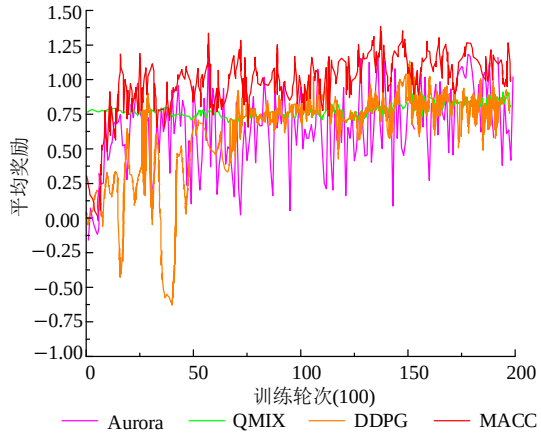

图7　各算法收敛性对比

各算法吞吐量的分布特征与统计稳定性如图 8 所示。MPTCP 的箱体与须线跨度最大，吞吐量散布于 69~95Mbps 区间并出现低值离群样本，Aurora 的分布范围次之，反映出传统多路径传输协议与基于强化学习的单路径拥塞控制算法在动态海上环境中易受链路质量扰动影响，吞吐性能波动剧烈且存在极端低值工况。QMIX 与 DDPG 的箱体宽度明显收窄，中位数分别提升至 91.1Mbps 与 93.5Mbps，但仍有少量离群样本出现，表明多智能体值分解与单智能体深度强化学习虽能改善传输稳定性，在抑制极端工况方面仍存在局限。相比之下，MACC 的箱体最为紧凑，吞吐量集中于 96.2~99.4Mbps 的四分位区间内，标准差仅为 2.6Mbps，较 MPTCP 降低约 61.8%，中位数与均值基本接近，且全部样本均未落入离群区域，表现出良好的鲁棒性。MACC 优势得益于多智能体框架通过集中式训练去中心化执行机制学习到的协同策略，能够在链路状态动态变化时及时调整各节点的传输决策，从而有效补偿拓扑与信道扰动带来的吞吐量波动。

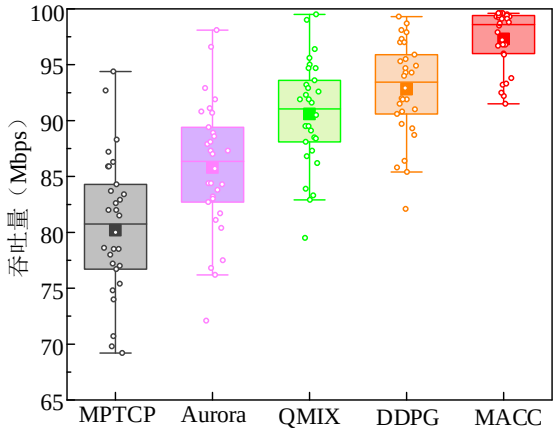

图8　各算法吞吐量对比

各算法在吞吐量-时延权衡空间中的分布特征及理想帕累托前沿如图 9 所示。五种算法的散点簇自左下方至右上方呈清晰的阶梯状分布，簇间分离程度与算法综合性能排序一致。MPTCP 的散点分布最为分散且整体偏向左下方，簇中心距理想帕累托前沿的垂直偏差达 70.6ms，反映出传统多路径传输协议在异构链路环境中难以同时兼顾吞吐与时延的固有局限。Aurora 与 DDPG 作为单智能体深度强化学习方法，其散点簇位于 MPTCP 与 MACC 之间，性能优于传统算法但距前沿仍有明显距离，表明单智能体决策在高维联合状态空间中易收敛至局部最优。QMIX 虽引入多智能体值分解机制，其散点簇与 MACC 仍存在可见差距，说明仅有集中式

协同训练框架而缺乏跨层耦合奖励与拓扑感知消息传递，难以充分挖掘多链路协同的潜在增益。相比之下，MACC的散点簇最为紧凑地集中于图的右上方区域，对应高吞吐与低时延的联合最优，簇中心（97.3Mbps，40.2ms）与理想帕累托前沿的垂直偏差仅为6.9ms。同时，理想前沿在吞吐量逼近链路容量重载区域时时延急剧增大，而MACC恰落于该拐点邻域，表明本文方法在传输效率与响应时延两个目标上实现了逼近理论下界的最佳权衡。注入的聚合负载约100Mbps由四类异构链路共同承载，端到端瓶颈容量高于任一单链路的标称容量，故该重载区域对应多链路协同承载工况，而非单一链路饱和。

图8与图9的统计结果来自同一批评估实验，其采用跨层耦合奖励训练策略至收敛，并在相同动态场景下对该收敛策略重复评估30次。两图分别展示该策略在不同指标维度上的表现，并非针对各指标分别优化所得的结果。

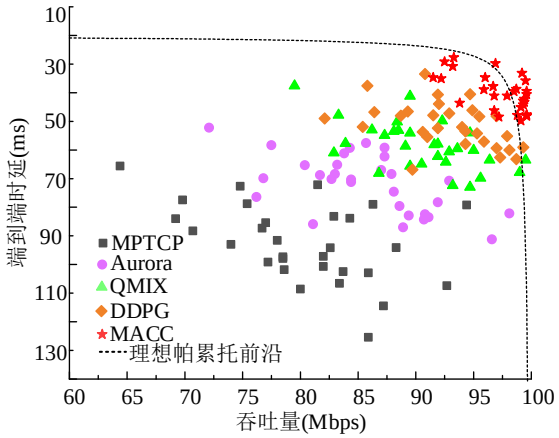

图9　各算法时延-吞吐量性能对比

如图10所示，四种学习型算法的Jain公平指数均随训练进程逐步提升并趋于收敛，但收敛速度与稳态水平差异显著。各学习算法均在离线场景库中预训练，图10横轴为训练仿真时间，评估阶段加载收敛后的策略运行300s，收敛时间均指训练收敛时间。MPTCP作为非学习型基准协议，其公平指数不随训练变化，以约0.874的水平参考线形式参与对比。仿真初期，智能体尚未充分学习链路特性，各学习型算法的公平指数均处于0.58~0.65的较低水平，暂时低于MPTCP基准。随着训练进行，完整MACC与四种消融变体在四项关键指标MACC的收敛速度显著领先，约20s达到0.95以上的公平水平，并在约33s时稳定于0.978，且全程波动幅度最小。Aurora与QMIX分别于约36s和48s收敛至0.891与0.927的稳态值。DDPG收敛最为缓慢，约88s达到稳态，但其终值0.935仍略高于QMIX。由此可见，MACC将收敛时间缩短至33s的同时将稳态公平指数提升至0.978，实现了收敛速度与收敛质量的双重优势，这得益于集中式训练去中心化执行框架下智能体间高效的协同策略学习与准确的信用分配。

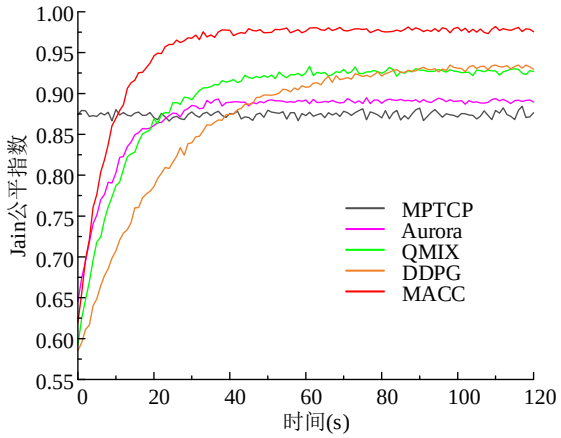

图10　各算法Jain公平指数收敛对比

3.3　消融实验

为量化MACC方法中各核心组件对整体性能的贡献，本文设计了四组消融实验进行验证。去除异构状态统一表征时，各链路原始观测仅经min-max归一化处理后直接拼接输入智能体网络与混合网络，不再经过统一表征空间的结构化映射。去除跨层耦合奖励时，奖励函数退化为传输层吞吐量效用、网络层队列稳定性与链路层切换开销三项以固定权重$(\frac{1}{3},\frac{1}{3},\frac{1}{3})$线性相加的形式，且不含公平激励项，权重不参与跨层耦合调节。去除图神经网络时，以MLP替代GATv2消息传递模块，智能体仅依据自身观测进行决策，不再聚合邻居节点的拥塞上下文。独立学习变体中，各链路智能体采用独立Q学习各自更新价值函数，不使用集中式混合网络与全局状态信息，以剥离CTDE框架的协同训练机制。除上述差异外，各变体的网络规模、超参数取值与训练评估流程均与完整MACC保持一致。

上的对比结果如图11所示。可以看出，完整

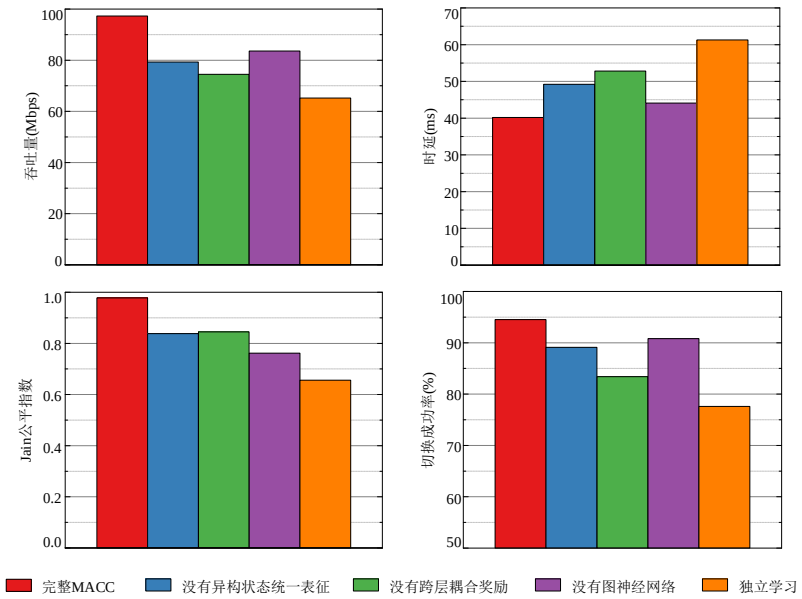

图 11　消融实验

MACC 在全部指标上均取得最优表现，而独立学习变体在各项指标上全面垫底，充分说明 CTDE 框架下中央协调器提供的全局信息交互与协同信用分配是系统性能的根基，完全去中心化的决策模式难以应对多链路交互引发的非平稳性问题。去除跨层耦合奖励后，吞吐量与切换成功率的下滑在三种变体中最为明显，这是因为缺失队列稳定性与切换开销的联合约束后，智能体的优化目标退化为单一的吞吐量最大化，频繁无效的链路切换直接侵蚀了有效传输能力。去除异构状态统一表征后，各项指标呈现均衡幅度的退化，表明量纲不同的原始观测会普遍削弱网络对各个体价值的比较与组合能力。去除图神经网络后，吞吐量与时延仍保持在各变体中的次优水平，但 Jain 公平指数的降幅在三种变体中最为突出，说明拓扑感知消息传递的核心作用在于为公平性提供高阶邻域信息，失去邻居拥塞上下文后，智能体倾向于各自抢占瓶颈资源，链路间的带宽分配趋于失衡。综上所述，四类组件缺一不可，只有协同组合才能在吞吐量、时延、公平性与切换稳定性之间取得最佳权衡。

3.4　超参数敏感性分析

本节重点关注相对变化趋势，缩减规模配置的状态维度与回合时长低于主实验，其收敛回合数与 3.2 节结果不可直接比较。

为在可控计算开销下完成多配置扫描，敏感性实验采用缩减规模评估配置。节点规模与链路容量按比例缩减，信道、移动性与流量模型设置与 3.1 节保持一致，重点考察各超参数扰动下性能的相对变化趋势。选取对训练动态影响最为显著的三个超参数，分别为折扣因子 γ、学习率与目标网络软更新系数 τ，在默认值附近设置 3 个水平（$\gamma\in\{0.90, 0.95, 0.99\}$，学习率$\in\{1\times10^{-4}, 3\times10^{-4}, 1\times10^{-3}\}$，$\tau\in\{0.001, 0.005, 0.02\}$），各配置独立训练后在留出场景集上评估，实验结果如表 2、图 12 所示。

可以看出，Jain 公平指数对三类超参数扰动均不敏感，全部配置下公平指数始终维持在 0.90 以上，最大偏差仅−5.5（γ=0.99 时），学习率与 τ 扰动下的偏差分别不超过 2.1 与 1.1。这是因为跨层耦合奖励中显式的公平性激励项与自适应权重网络的状态条件化加权机制，使链路间带宽分配在超参数扰动下仍能保持均衡。吞吐量方面，软更新系数 τ 与学习率在各自扫描区间内的最大相对偏差分别为−13.6%与−23.1%。γ=0.99 时吞吐量下降 38.4%，其原因在于过大的折扣因子使智能体过度侧重长期回报，对海上信道与拓扑的快时变拥塞响应缓慢。收敛性方面，各配置均能稳定收敛且未出现发散现象，收敛回合数随超参数扰动平缓变化。

4　结束语

本文针对海上异构网络多路径协同拥塞控制问题，提出了多智能体协同拥塞控制方法。首先构建基于 CTDE 的多智能体协同架构，用于缓解非平稳

表2　超参数敏感性实验结果

超参数	取值	吞吐量相对偏差(%)	Jain公平指数	收敛回合数
	0.90	+6.9	0.953	68
折扣因子γ	0.95（默认）	0	0.955	36
	0.99	-38.4	0.899	14
	1×10^{-4}	-6.5	0.951	12
学习率	3×10^{-4}（默认）	0	0.955	36
	1×10^{-3}	-23.1	0.934	30
	0.001	+10.6	0.954	42
软更新系数τ	0.005（默认）	0	0.955	36
	0.02	-13.6	0.945	18

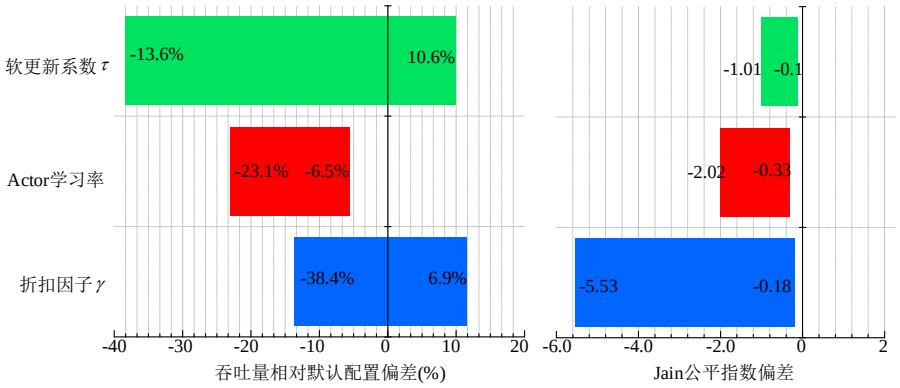

图12　超参数敏感性分析

性问题。其次设计异构状态统一表征，将带宽跨度三个数量级、时延差异百倍的链路状态映射至统一空间。然后提出跨层耦合奖励，耦合传输层吞吐量效用、网络层队列稳定性与链路层切换开销，使智能体隐式逼近网络效用最大化。最后引入GATv2消息传递实现GNN增强拓扑感知，使智能体自动编码邻居拥塞上下文，并在仿真实验中证明了本文所提出方法的有效性。未来工作将围绕真实海上环境、大规模节点网络以及在线学习与跨场景迁移学习场景开展进一步研究工作。

参考文献：

[1] Wang C, Peng Y, Wu J, et al. Deterministic scheduling and reliable routing for smart ocean services in maritime internet of things: A cross-layer approach[J]. IEEE Transactions on Services Computing, 2024, 17(6): 3387-3399.

[2] Dai M, Chang S, Wang Y, et al. Energy-efficient multi-access edge computing for heterogeneous satellite-maritime networks: A hybrid harvesting-and-offloading design[J]. IEEE Transactions on Mobile Computing, 2025.

[3] Xu C, Xiao H, Li Z, et al. Diversified service oriented multimedia intelligent transmission in 6G space-air-ground-ocean heterogeneous network[J]. IEEE Wireless Communications, 2025, 32(2): 130-138.

[4] Li W, Zhang H, Gao S, et al. SmartCC: A reinforcement learning approach for multipath TCP congestion control in heterogeneous networks[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(11): 2621-2633.

[5] Han J, et al. From an in-depth understanding of multipath TCP enhancement schemes to an adaptive control framework in wireless networks[J]. IEEE Transactions on Network and Service Management, 2025, 22(5): 4950-4963.

[6] Morawski M, Ignaciuk P. Choosing a proper control strategy for multipath transmission in industry 4.0 applications[J]. IEEE Transactions on Industrial Informatics, 2021, 18(6): 3609-3619.

[7] Lu J, Huang G, Hu D, et al. Multi-path data transmission system based on 5G communication technology[J]. Journal of ICT Standardization, 2024, 12(1): 71-94.

[8] Tian Z, Zhang Y, Guo P, et al. Heterogeneous radios meet intelligence: DRL-based delay minimization in multi-hop wireless networks[J]. IEEE Transactions on Network and Service Management, 2025.

[9] Wu F, et al. Named data networking enabled power saving mode design for WLAN[J]. IEEE Transactions on Vehicular Technology, 2020, 69(1): 901-913.

[10] Wang X, Di X, Zhu Y, et al. AoI-driven dynamic multipath congestion control for space – air – ground – sea integrated network[J].

IEEE Internet of Things Journal, 2026, 13(12): 26796-26810.

[11] Yang J, Wu J, Chiclana F, et al. An inter-subgroup compensation mechanism by Nash bargaining game for managing non-cooperative behavior in group decision making[J]. Computers & Industrial Engineering, 2025, 204: 111114.

[12] He X, Yang Y, Lee J, et al. Deep reinforcement learning based AoI minimization for NOMA-enabled integrated satellite-terrestrial networks[J]. IEEE Transactions on Vehicular Technology, 2025, 74 (2): 3567-3572.

[13] Han R, Wen Y, Bai L, et al. Age of information aware UAV deployment for intelligent transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(3): 2705-2715.

[14] Yang J, et al. Adaptive multi-source multi-path congestion control for named data networking[J]. IEEE/ACM Transactions on Networking, 2024, 32(6): 5049-5064.

[15] 陈锦前, 郭少勇, 刘畅, 等. 数据处理单元赋能的智算中心网络拥塞控制机制[J]. 通信学报, 2025, 46(2): 1-17.
CHEN J Q, GUO S Y, LIU C, et al. Congestion control mechanism for intelligent computing center networks empowered by data processing units[J]. Journal on Communications, 2025, 46(2): 1-17.

[16] Jay N, Rotman N, Godfrey B, et al. A deep reinforcement learning perspective on internet congestion control[C]//Proceedings of the 36th International Conference on Machine Learning. New York: PMLR, 2019: 3050-3059.

[17] Mbrice T, Liu Y. Fair-Aurora: Comparing fairness strategies for reinforcement learning-based congestion control in multi-flow environments[J]. arXiv preprint, arXiv:2605.19909, 2026.

[18] Gao K, et al. Freshness-aware age optimization for multipath TCP over software defined networks[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(6): 5589-5600.

[19] Zhang M, Tong W, Zhu G, et al. SQIX: QMIX algorithm activated by general softmax operator for cooperative multiagent reinforcement learning[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(11): 6550-6560.

[20] Guo W, Liu G, Zhou Z, et al. Enhancing the robustness of QMIX against state-adversarial attacks[J]. Neurocomputing, 2024, 572: 127191.

[21] Jia C, Liu W, He H, et al. Superior energy management for fuel cell vehicles guided by improved DDPG algorithm: Integrating driving intention speed prediction and health-aware control[J]. Applied Energy, 2025, 394: 126195.

[22] Kim J I, Kim S B. D-QMIX: Multi-step sequential forward dynamics modeling with global state and self-attention for sample-efficient multi-agent reinforcement learning[J]. Information Sciences, 2025: 122867.

[23] Liu Y C, Huang C Y. DDPG-based adaptive robust tracking control for aerial manipulators with decoupling approach[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 8258-8271.

[24] Xue J, He M, Chen J, et al. Improved DDPG based on enhancing decision evaluation for path planning in high-density environments [J]. Expert Systems with Applications, 2025, 279: 127378.

[25] Pey J J J, Samarakoon S M B P, Muthugala M A V J, et al. A decentralized partially observable Markov decision process for complete coverage onboard multiple shape changing reconfigurable robots[J]. Expert Systems with Applications, 2025, 271: 126565.

[26] Tang X, Li Q, Peng J, et al. Low-latency video streaming: Applications, challenges and trends[J]. IEEE Communications Surveys & Tutorials, 2026.

[27] Peroni L, Gorinsky S. An end-to-end pipeline perspective on video streaming in best-effort networks: A survey and tutorial[J]. ACM Computing Surveys, 2025, 57(12): 1-47.

[28] Liu R W, Liang M, Nie J, et al. Deep learning-powered vessel trajectory prediction for improving smart traffic services in maritime Internet of Things[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(5): 3080-3094.

[29] Amato C. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning[J]. arXiv preprint, arXiv:2409.03052, 2024.

[30] Zhang S, Sutton R S. A deeper look at experience replay[J]. arXiv preprint, arXiv:1712.01275, 2017.

[31] Ackermann J, Gabler V, Osa T, et al. Reducing overestimation bias in multi-agent domains using double centralized critics[J]. arXiv preprint, arXiv:1910.01465, 2019.

[32] Redondo J, Aslam N, Zhang J, et al. Optimising QoS in HD map updates: Cross-layer multi-agent with multi-task and mixed-dependence (MTMD)[J]. IEEE Transactions on Network and Service Management, 2026.

[33] Brody S, Alon U, Yahav E. How Attentive are Graph Attention Networks?[C]//International Conference on Learning Representations. Virtual: Open Review, 2022.