计算机科学与探索
*Journal of Frontiers of Computer Science and Technology*
ISSN 1673-9418,CN 11-5602/TP

# 《计算机科学与探索》网络首发论文

题目：多智能体深度强化学习通信七维度综述
作者：陈荣敏，郭大波，吴宏坤，李程翔，胡海霄，刘李祥
网络首发日期：2025-10-13
引用格式：陈荣敏，郭大波，吴宏坤，李程翔，胡海霄，刘李祥．多智能体深度强化学习通信七维度综述[J/OL]．计算机科学与探索．
https://link.cnki.net/urlid/11.5602.tp.20251011.1450.004

计算机科学与探索
Journal of Frontiers of Computer Science and Technology　doi: 10.3778/j.issn.1673-9418.2506028

# 多智能体深度强化学习通信七维度综述

陈荣敏，郭大波+，吴宏坤，李程翔，胡海霄，刘李祥
三亚学院 信息与智能工程学院，海南 三亚 572000
+ 通信作者 E-mail: dabo_guo@sxu.edu.cn

**摘　要：**深度强化学习在解决单智能体机器学习中的各种顺序决策问题方面已取得显著成功，但在现实环境中任务往往涉及多个智能体，因此多智能体深度强化学习（MADRL）技术受到越来越多的关注。多智能体深度强化学习面临部分可观测性和环境非平稳性挑战，基于通信的多智能体深度强化学习（CB-MADRL）通过引入通信机制应对这些问题。聚焦 CB-MADRL 中的通信模块，针对现有综述在通信机制系统性分析上的不足，在该领域首次将通信时机进行分类讨论，并在通信对象和消息聚合维度设计新的分类方式，系统性提出了通信时机、通信对象、通信来源、消息聚合、通信作用、通信学习、通信约束的七个维度的分析框架，在此框架基础上对近些年有关 CB-MADRL 方法进行了分类梳理，设计统一的分类符号表示，并对各方法的通信七维度进行汇总。系统分析了通信七维度在现实应用中面临的多模态与异构消息处理、通信协议“黑箱化”以及现实通信约束耦合等关键挑战，并提出针对性的研究思路与可行方法。此外，深入探讨了通信维度间的相互作用与协同优化机制，提出了最优维度组合的探索路径，并展望了未来的研究方向。

**关键词：**多智能体系统；深度强化学习；多智能体通信；图神经网络；注意力机制
**文献标志码：**A　**中图分类号：**TP18

## A Review of Communication in Multi-Agent Deep Reinforcement Learning from Seven Dimensions

CHEN Rongmin, GUO Dabo+, WU Hongkun, LI Chengxiang, HU Haixiao, LIU Lixiang
School of Information and Intelligent Engineering, University of Sanya, Sanya, Hainan 572000, China

**Abstract:** Deep reinforcement learning has achieved remarkable success in solving various sequential decision-making problems in single-agent machine learning. However, in real-world environments, tasks often involve multiple agents, leading to growing interest in multi-agent deep reinforcement learning (MADRL) techniques. MADRL faces challenges such as partial observability and environmental non-stationarity. Communication-based multi-agent deep reinforcement learning (CB-MADRL) addresses these issues by introducing communication mechanisms. Focusing on the communication module in CB-MADRL, and addressing the lack of systematic analysis of communication mechanisms in existing surveys, this work, for the first time in the field, classifies communication timing, and proposes

**基金项目：**三亚学院人才引进项目(USYRC22-14)；三亚学院“产教融合”研究项目(USY23CJRH01、USY23CJRH02)；海南省高等学校教育教学改革研究项目(Hnjg2025ZC-89)；三亚学院硕士研究生一流课程。
This work was supported by the Sanya University Talent Introduction Project (USYRC22-14); the 2023 Integration of Industry and Education Research Projects of Sanya University (USY23CJRH01，USY23CJRH02)； the Higher Education and Teaching Reform Research Project of Hainan Province(Hnjg2025ZC-89) and Sanya University's top master's degree courses.

new classification methods for communication targets and message aggregation. A systematic seven-dimensional analytical framework is presented, encompassing communication timing, communication target, communication source, message aggregation, communication function, communication learning, and communication constraints. Based on this framework, recent CB-MADRL methods are categorized, a unified classification notation is designed, and the seven communication dimensions of each method are summarized. The paper systematically analyzes key challenges faced by the seven communication dimensions in real-world applications, including multi-modal and heterogeneous message processing, the “black-box” nature of communication protocols, and the coupling of real-world communication constraints, and proposes targeted research ideas and feasible approaches. Furthermore, it explores the interactions and synergistic optimization mechanisms among the communication dimensions, proposes a path for exploring optimal dimension combinations, and finally envisions future research directions.

**Key words:** Multi agent system; Deep reinforcement learning; Multi agent communication; Graph neural network; attention mechanism

随着人工智能技术的飞速发展，强化学习[1,2]（Reinforcement Learning, RL）已经成为机器学习领域中的一个重要分支。在强化学习中，智能体通过与环境的交互，以最大化累积奖励为目标函数，学习最优策略。强化学习在一系列决策问题的求解上达到或超过了人类水平[3]。深度强化学习（deep reinforcement learning, DRL）[4]通过在强化学习中引入深度神经网络作为函数逼近器，替代传统RL中的表格或线性模型，解决了维度灾难问题。

现如今，DRL已经被应用于机器人控制[5,6]，自动驾驶[7,8]，优化调度[9,10]和游戏[11,12]领域中，然而，在解决现实问题的时，研究者们往往高估了DRL的表现，低估了其工程实现难度[13]。特别是现实任务中，环境往往包括多个与环境互动并同时学习的智能体，这类问题通常被称作多智能体序列决策问题[14]。为解决这一类问题，多智能体深度强化学习[15](multi-agent DRL, MADRL)应运而生。

MADRL的目标是让多个智能体在共享的环境中通过不断与环境以及彼此之间进行交互，学习各自的最优策略，以实现个体目标或者共同的团队目标。在多智能体环境下，智能体只能访问其局部观测结果，而不是环境的完整状态，因此部分可观测（Partial Observability）成为MARL中的基本假设[16]。此外，智能体会将环境中的其他智能体也作为环境的一部分，由于其他智能体同样处于学习过程中，环境的状态转移函数不再是静态的而是动态变化的，从而打破了马尔科夫性假设，导致环境的非平稳性。

基于通信的多智能体深度强化学习（Communication Based MADRL, CB-MADRL）通过在MADRL中引入通信，以解决其部分可观测性和环境非平稳问题。在CB-MADRL中智能体可以将自己的观测、意图、经验等作为消息传输给环境中的其他智能体。智能体得到了对环境更全面的信息，这能使智能体做出更明智的决策，从而更好在的多智能体环境中完成目标任务[17]。

我们发现，已经有大量的工作聚焦于CB-MADRL，意旨通过通信交流共享信息来解决特定领域的任务，例如导航、交通和游戏[17]。在目前有关MARL综述中[14,15,18-20]，均有涉及有关智能体通信，但现有的专门关注CB-MADRL的综述文章还比较少。现有关注CB-MADRL文献综述中，文献[21]主要讨论不同性质多智能体系统下的CB-MADRL代表性方法。文献[22]关注现有CB-MADRL方法的性能分析。文献[3,17]开始关注CB-MADRL的通信维度的分类，是重要的先行工作。然而，文献[3]所涉及的CB-MADRL维度和方法范围较为有限；文献[17]对构建MARL信息交互的维度进行了较详细的分类，但仍存在不足：1)在消息聚合这一关键环节，缺乏对具体聚合方式的系统分类；2)缺乏对“通信时机”这一对系统效率至关重要的维度的讨论；3)仅在讨论通信类型时将智能体通

信对象划分为 3 类，未能充分挖掘通信对象维度的多样性与动态性；4)作为先行工作，其覆盖范围存在天然的时间局限性，未能纳入近两年发表的大量重要新进展，这些新工作往往在通信效率、鲁棒性、可解释性以及与新兴技术结合等方面取得了显著突破。表 1 比较了现有综述在通信维度覆盖方面的差异。

综上所述，现有综述尚未能提供一个全面、深入且结构化的框架来系统性解构和分析 CB-MADRL 中复杂的通信行为。通信机制的设计涉及多个相互关联的决策维度，缺乏一个统一的分析视角会阻碍对现有方法的深入理解、有效比较以及新方法的针对性创新。此外，现有 CB-MADRL 方法与综述普遍忽略了对通信多维度之间组合关系及其协同优化机制的系统探讨。

基于上述问题和空白，本文综述的贡献与必要性主要体现在以下六个方面：

1.构建了系统性分析框架：提出“通信七维度”分析框架包括通信时机、通信对象、通信来源、消息聚合、通信作用、通信学习、通信约束。首次将通信行为解构为七个核心决策维度，为理解、分析和设计 CB-MADRL 通信机制提供了统一的结构化理论工具。

2.弥补关键分类空白：首次在 CB-MADRL 领域对“通信时机”维度进行系统分类。在“通信对象”和“消息聚合”维度提出了新的、更细致的分类方式。

3.实现方法的统一符号表征：设计了统一的分类符号表示体系，对纳入综述的各类方法在七个维度上的特征进行标准化编码和汇总，极大地方便了方法的横向比较与研究脉络的梳理。

4.覆盖最新研究进展：系统梳理并纳入了最近的 CB-MADRL 通信研究成果，填补了相较于已有综述因发表时间较早而未能涵盖的研究空白，对近期涌现的关键方法进行了深入分析和归类，确保了对领域发展现状的全面性和时效性把握。

5.揭示通信维度间的耦合机制：剖析了通信七维度之间深刻的相互作用与依赖关系，首次系统论证了孤立优化单一维度的局限性，并前瞻性地提出了协同优化多维度通信框架的必要性。此外，还对未来探索最优维度组合指给出意见。

6.聚焦前沿挑战与方向：基于七维度框架，不仅回顾现有方法，更深入剖析了当前通信机制在应对现实复杂场景时面临的核心挑战，并探讨了解决思路和未来研究方向，为领域发展提供前瞻性地指引。

本综述整体组织结构如图 1 所示。

**表 1　涉及通信维度分类 CB-MADRL 综述**

Table 1　Review of CB-MADRL Reviews with Classification of Communication Dimensions

| 综述 | 通信时机 | 通信对象 | 通信来源 | 消息聚合 | 通信作用 | 通信学习 | 通信约束 |
|---|---|---|---|---|---|---|---|
| 综述[17] | | √ | √ | √ | √ | √ | √ |
| 综述[3] | | √ | √ | √ | √ | | √ |
| 本文 | √ | √ | √ | √ | √ | √ | √ |


**图 1　综述框架**

Fig.1　Review Framework

## 1 基础概述

### 1.1 问题建模

本文将具有通信的多智能体系统建模为部分可观测随机博弈（Partially Observable Stochastic Game, POSG）。POSG 可以由一个元组 $\langle \mathcal{N},\mathcal{S},\mathcal{A},\mathcal{P},\Omega,\mathcal{O},\mathcal{R},\gamma,\mathcal{M}\rangle$ 定义，其中 $\mathcal{N}=\{1,\cdots,n\}$ 是 $n$ 个智能体的集合，$\mathcal{S}$ 是全局状态的集合，$\mathcal{A}$ 是动作的集合，$\mathcal{P}:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ 是状态转移概率分布，$\Delta(\mathcal{S})$ 是 $\mathcal{S}$ 上的概率分布集合，$\Omega$ 是观测的集合，$\mathcal{O}:\mathcal{S}\times\mathcal{N}\to\Omega$ 是将全局状态和智能体观测进行映射的观测函数。$\mathcal{R}:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ 是奖励函数，$\gamma\in[0,1)$ 是折扣因子，$\mathcal{M}$ 是智能体可以通信的消息集合。在每个时间步 $t$，每个智能体 $i\in\mathcal{N}$ 从观测函数 $\mathcal{O}(s_t,i)$ 接收局部观测 $o_t^i\in\Omega$，其中 $s_t\in\mathcal{S}$ 是当前的全局状态。每个智能体遵循个体策略 $\pi^i(a_t^i\mid o_t^i,m_t^i)$，$m_t^i\in\mathcal{M}$ 是智能体 $i$ 在时间 $t$ 接收到的消息。智能体执行联合动作 $a_t=\langle a_t^1,\ldots,a_t^n\rangle$ 得到下一个状态 $s_{t+1}\sim\mathcal{P}(s_{t+1}\mid s_t,a_t)$ 并且获得奖励 $r=\langle r_t^1,\ldots,r_t^n\rangle$。

值得一提的是，在完全合作的任务中各智能体获得的奖励 $r_t^1=r_t^2=\ldots=r_t^n$，这时 POSG 退化为去中心化部分可观测马尔可夫决策过程（Decentralized Partially Observable Markov Decision Process,Dec-POMDP）。在完全合作任务中智能体的目标通常是寻找一个联合策略 $\pi=\langle\pi^1,\ldots,\pi^n\rangle$ 使期望折扣回报 $\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$ 最大化。如果在每个时间步观测集合 $\Omega$ 能唯一确定一个全局状态 $\mathcal{S}$，即智能体可以通过联合观测推断出真实全局状态，则 Dec-POMDP 退化为 Dec-MDP，即去中心化可观测马尔可夫决策过程。如果在每个时间步智能体局部观测 $o_t^i=s_t$，即每个智能体都能直接获取全局状态，则 Dec-MDP 退化为多智能体 MDP。如果 $n=1$，即智能体集合中只有一个智能体，多智能体 MDP 退化为个体 MDP，Dec-POMDP 则会退化为 POMDP，即部分可观测马尔可夫决策过程。

### 1.2 多智能体深度强化学习的范式

现有的 MADRL 方法针对学习-决策主要遵循三种范式：分布式训练分布式执行（decentralized training decentralized execution, DTDE）、集中式训练集中式执行（centralized training centralized execution, CTCE）和集中式训练分布式执行（centralized training decentralized execution, CTDE）。DTDE 采用独立学习模式，将每个智能体视为单智能体强化学习单元，如 IDQN[23]，IPPO[24]，虽在简单任务中有效，但因环境动态变化导致马尔可夫假设失效，难以应对更加复杂场景。CTCE 通过联合状态-动作空间训练规避环境非平稳性，但智能体数量增加时面临维度爆炸问题，严重制约其可扩展性。CTDE 作为主流范式，结合集中式训练克服非平稳性，同时保留分布式执行的灵活性，使智能体基于局部观测独立决策，兼具理论优势与实践可行性，成为当前复杂多智能体任务的首选方案。在遵循 CTDE 范式的多智能体深度强化学习方法中，基于值分解和基于演员-评论家的两类方法最具代表性。基于值分解的方法主要通过将全局的联合价值函数分解为各个智能体的局部价值函数来实现多智能体的协作。在集中式训练阶段，利用全局信息训练一个全局价值函数，并通过分解策略将其分解为智能体的局部价值函数，使得每个智能体在执行时只需依赖自身的局部观测即可选择动作。基于演员-评论家的方法通过引入策略网络和价值评估网络，将策略优化与价值评估分开，以实现多智能体的协作和复杂策略学习。在集中式训练阶段，评论家利用全局信息评估联合策略的价值，执行阶段每个智能体的演员网络仅依赖自身的局部观测执行动作。为了系统比较这两类方法的核心思想、代表方法及适用场景，在表 2 中进行了总结。

表 2 CTDE 范式下的 MADRL 分类

Table 2 Classification of MADRL under the CTDE Paradigm

| 方法类别 | 基于值分解 | 基于演员-评论家 |
|---|---|---|
| 核心思想 | 分解全局价值函数，局部执行 | 评论家网络集中式评估，演员网络局部执行 |
| 代表方法 | VDN[25],QMIX[26],QTRAN[27] | COMA[28],MADDPG[29],MAPPO[30] |
| 目标 | 避免直接处理指数级增长的联合动作空间 | 通过集中评论家解决信用分配问题，指导各演员优化策略，平衡探索与利用 |
| 适用场景 | 离散动作空间、合作任务 | 离散或连续动作空间，合作、竞争或合作-竞争混合任务 |
| 特点 | 计算效率高但分解限制表达能力 | 能有效处理复杂协作关系但对全局信息依赖强 |

### 1.3 多智能体通信

在有关 CB-MADRL 的论文中，我们注意到两个密切相关的研究领域分别是“新兴语言（emergent language）”和“通过通信学习任务（learning tasks with communication）”[15]。新兴语言[31-35]指的是多智能体系统在协作与交互过程中自发产生的通信协议或语言体系。在新兴语言的研究中，智能体通常没有预先定义的通信规则，而是通过共同完成任务的过程，自主学习出有效的信息表达方式。智能体学习到的语言可能包括符号化信号、序列编码、甚至接近自然语言的表达形式。新兴语言研究的目标是探索在没有人为干预的情况下，智能体如何发展出可以互相理解和高效协作的通信方式。通过通信学习任务[36-38]关注的是在多智能体系统中，智能体如何利用通信机制共同完成复杂任务。与新兴语言关注通信协议的自发产生不同，这一研究领域更侧重于通信机制对任务学习和完成的促进作用。通过通信学习任务的研究目标是设计和优化通信机制，使多智能体能够更高效地协作，从而提升整体任务表现。

在 MADRL 中有关于通信部分可以分为非显式通信（Non-Explicit Communication）和显式通信（Explicit Communication）。非显式通信的 MADRL 方法中智能体之间不直接传递结构化消息，而是通过环境状态、其他智能体的行为或共享的网络参数间接协调。MADDPG[28] 方法就是一个代表性工作，其并未设置显式通信模块，训练阶段中央评论家直接获得所有智能体的联合状态和动作，其梯度信号隐含智能体间的协作信息，而在执行阶段智能体依赖局部观测执行策略。显式通信方法则要求智能体通过明确的通信协议传递结构化消息，通常需要设计专用的通信信道和消息编码机制。智能体的行动策略受他们的观测和接收到的消息的共同影响，此类方法中消息对于智能体的决策至关重要的，在训练和执行阶段都是必不可少的。

本文将重点关注通过通信学习任务且显式通信的 MADRL 方法。

## 2 CB-MADRL 通信维度分类

多智能体通信机制的设计需兼顾动态环境适应性、信息交互效率与协作目标的一致性。为系统分析 CB-MADRL 方法，本节从通信行为的关键要素出发，提出通信七维度分析框架，涵盖通信时机、通信对象、通信来源、消息聚合、通信作用、通信学习、通信约束，如图 2 所示。

通信七维度框架中通信时机关注“何时通信”，探讨全时通信、自适应通信与事件触发通信的时序策略；通信对象研究“与谁通信”，涉及预定义拓扑与可学习拓扑的动态选择；通信来源分析“通信内容从何而来”，区分基于历史、当前信息的通信与基于未来信息预测的前瞻式通信；消息聚合解决“如何处理多源消息”，涵盖无显式聚合、神经网络聚合、注意力机制聚合等方法；通信作用分析“通信产生的影响”，包括对值函数、策略函数的单独或协同影响；通信学习探索“如何优化通信协议”，涉及可微分、强化学习、监督学习等训练范式；通信约束讨论“现实通信限制的应对”，如有限带宽、传输损耗等场景下的鲁棒性设计。这七维度构成了一个完整的 CB-MADRL 通信行为分析闭环。本章后续小节将依次深入探讨每个维度的具体内涵、现有研究的代表性方法及其分类，详见表 3 至表 10，并最终在表 11 和表 12 中提供统一的符号表

示和对现有方法的七维度总结。

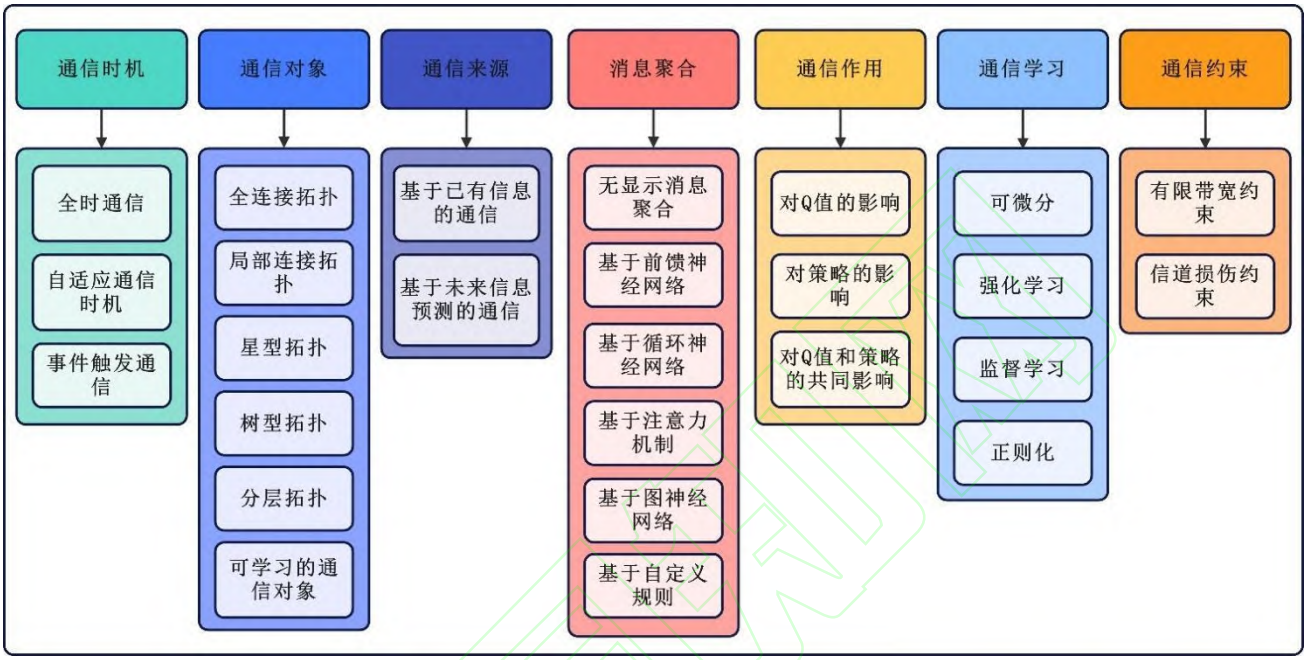


**图 2 CB-MADRL 通信维度分类**

Fig.2 Classification of Communication Dimensions in CB-MADRL

## 2.1 通信时机：效率与信息完备性的权衡

现有综述对通信机制的分析，多集中于通信对象、通信作用以及通信学习等维度，却普遍忽视了对“通信时机”这一关键决策维度的系统性梳理与分类。然而，通信时机的选择直接决定了多智能体系统的通信效率、资源消耗与协作效能，是平衡信息完备性与操作实时性的核心杠杆。忽略对这一维度的深入探讨，将难以全面理解与设计适用于现实场景的通信策略。本文首次在 CB-MADRL 综述框架中明确提出并构建了“通信时机”维度，将现有的通信时机策略归纳为全时通信、自适应通信时机与事件触发通信三类。这一分类不仅为剖析不同方法的通信效率提供了清晰的标尺，更重要的是，它揭示了通信机制设计中如何权衡通信效率与信息完备性这一矛盾中两方面的问题，即如何在尽可能保证关键信息的完备的前提下减少通信开销，减少冗余信息传输，从而应对部分可观测性问题中信道容量受限带来的挑战。通信时机分类及其代表性方法如表 3 所示。

**表 3 通信时机分类**

Table 3 Classification of Communication Timing

| 通信时机 | 相关方法 |
|---|---|
| 全时通信 | DIAL[36]、RIAL[36]、CommNet[37]、BiCNet[38]、ATOC[39]、DGN[40]、TarMAC[41]、MD-MADDPG[42]、SchedNet[43]、IMAC[44]、LSC[45]、Diff Discrete[46]、IS[47]、HAMMER[48]、GAXNet[49]、MAIC[50]、AMSAC[51]、AICNet[52]、TDU[53]、PMAC[54]、MAGI[55]、CroMAC[56]、PCGQ[57]、ExpoComm[58]、TMAC[59]、CoDe[60] |
| 自适应通信时机 | MADDPG-M[61]、IC3Net[62]、NDQ[63]、GA-Comm[64]、I2C[65]、TMC[66]、MACS[67]、PAGNet[68]、COCOM[69]、TGCNet[70]、AsynCoMARL[71] |
| 事件触发通信 | VBC[72]、ETCNet[73]、MBC[74] |

全时通信指环境中的智能体在每一个时间步都进行通信，即无论当前环境状态是否发生变化，在每个时间步，智能体都会与其他智能体完成信息交互，其本质是为了规避部分可观测性带来的协作障碍，通过持续信息交换确保智能体对全局状态的实时感知，从而可充分保证信息的完备性。早期的多智能体通信

研究通常采用全时通信的方式。如 DIAL[36]方法在每个时间步，智能体通过可微的通信通道生成连续向量消息，并利用跨智能体梯度反向传播优化通信策略。BiCNet[38]方法在每个时间步，通过双向递归神经网络构建的通信信道实现智能体间的信息交互并结合参数共享机制，使智能体能够动态协调动作以完成目标。全时通信保证了信息的最新性和同步性，有利于多智能体的高效协作，但其缺陷也显而易见，其仅适用于智能体数量少，如 2 至 5 个，任务对时延敏感的场景。当智能体规模扩大时，通信数据量呈指数增长，不仅导致带宽压力，如大规模的无人机智能体编队中将面临带宽崩溃问题，冗余信息还可能引入噪声干扰，例如在静态环境中反复传输不变状态会干扰策略学习。因此全时通信在大规模场景中面临根本性扩展瓶颈。

为突破全时通信的局限性，后续研究提出了自适应通信时机的通信，即智能体根据当前环境状态、任务需求或内部策略，动态地决定是否进行通信。核心是解决全时通信的资源浪费问题，其设计受信息论启发——根据通信带来的收益（如 Q 值提升）及成本（如带宽限制）动态调整通信时机。如 NDQ[63]方法通过引入信息论驱动的通信机制，结合值函数分解与变分推理，使智能体能够自主学习何时通信。I2C[65]方法根据任务场景需求和先验网络的动态决策机制，可使通信策略具有非周期性和高效性。IC3Net[62]、COCOM[69]等自适应通信时机的通信方法通过引入可学习的门控机制，由智能体自主决定在某个时间步是否进行通信。这些门控机制通常根据当前状态特征、环境信息或任务需求，动态调整通信行为。这类方法有效减少了通信冗余，显著提升了通信效率，因此在许多动态任务中展现出显著优势。得益于这种按需通信的特性，自适应通信机制特别适用于动态环境场景，例如自动驾驶中的车辆避障，这类亟需在信息时效性与通信开销之间取得精细平衡的场景。然而，这类方法也存在一定局限性： 首先，门控模块与主任务策略的联合优化过程容易导致训练不稳定，例如门控策略和动作策略的学习进度不匹配，导致通信决策失效或主任务性能下降；其次，它们通常依赖复杂的奖励函数设计，在奖励稀疏的场景，如长期探索任务中难以有效收敛。

基于事件触发的通信是一种稀疏化的通信策略，其核心思想是将通信与任务关键节点或固定规则绑定，智能体仅在特定事件发生或满足规则条件时触发通信。事件的定义可以多种多样，例如环境状态变化超过某一阈值、通信时间间隔超过设定值或决策置信度低等。如 VBC[72]方法采用基于置信度与方差信息的事件触发通信机制：当智能体对自身局部决策的置信度较低时，会主动发起通信请求，而接收方仅在其消息具有较高方差、即潜在信息价值较大时才进行回应。为进一步提升通信效率，VBC 在训练过程中对交换信息施加了基于方差的额外损失项，用于过滤冗余和低价值消息，仅保留信息量丰富的内容。通过事件触发通信的设计，VBC 在无需引入额外决策模块的情况下，即可动态调整通信模式，显著降低模型复杂度与不必要的通信开销，同时保证智能体在关键时刻能够获得有用信息，从而有效提升整体协作性能。ETCNet[73]方法通过引入事件触发的门控机制，将通信网络的有限带宽约束建模为数学上的惩罚阈值，并利用强化学习训练通信策略。该机制会在信息发生显著变化时才允许消息传递，从而动态决定是否进行通信。凭借“按需触发”逻辑，既避免了全时通信的冗余消耗，又通过事件条件的精准定义，确保关键信息不丢失，实现了“效率”与“完备性” 的动态平衡，在带宽受限场景下能够大幅降低通信频率，有效缓解因频繁通信带来的带宽占用和延迟问题，同时保持与全通信方式接近的协作性能。最新的相关研究 MBC[74]方法通过设置误差阈值和时间窗口的双条件事件触发机制，仅当估计偏差不可接受或长期未校正时发送消息，既保证协作效率，又最小化通信开销。适用于资源极端受限场景，如电池驱动的传感器网络或事件驱动型任务，如故障检测。基于事件触发的通信强调通信的高效性与任务相关性的同时，其能效实现面临的关键挑战在于事件定义的质量把控：过于保守的触发条件可能导致关键信息丢失，而过于敏感的阈值又可能使通信频率退化为频繁通信，事件定义依赖先验知识，因此泛化性差；可能因触发条件过严导致关键信息丢失，例如在森林火灾监测网络中，若烟雾浓度触发阈值设置过高，传感器节点可能在火灾初期因未达阈值而沉默，未能及时报告关键火情信息；其次，依赖先验知识定义“烟雾浓度阈值”难以适应不同季节、植被类型或气候条件下的火

灾特征变化。

## 2.2 通信对象：拓扑结构与协作模式的匹配

在 MADRL 中，通信对象的设计决定了智能体之间如何选择消息的发送与接收对象。我们可以使用图 $G=(V,E)$ 来表示多智能体系统中的通信拓扑结构，其中 $V$ 是智能体集合，$E$ 则代表通信信息传递。根据现有研究，通信拓扑结构可以分为预定义的通信对象和可学习的通信对象。

预定义的通信对象是指通信拓扑结构在算法运行前就已事先确定的通信机制。通常根据任务目标和方法设计思路的不同设计不同的拓扑，具体可分为如图 3 所示的以下五种主要主要类型：全连接拓扑、局部连接拓扑、星型拓扑、树型拓扑和分层拓扑。预定义的通信对象分类及其代表性方法如表 4 所示。

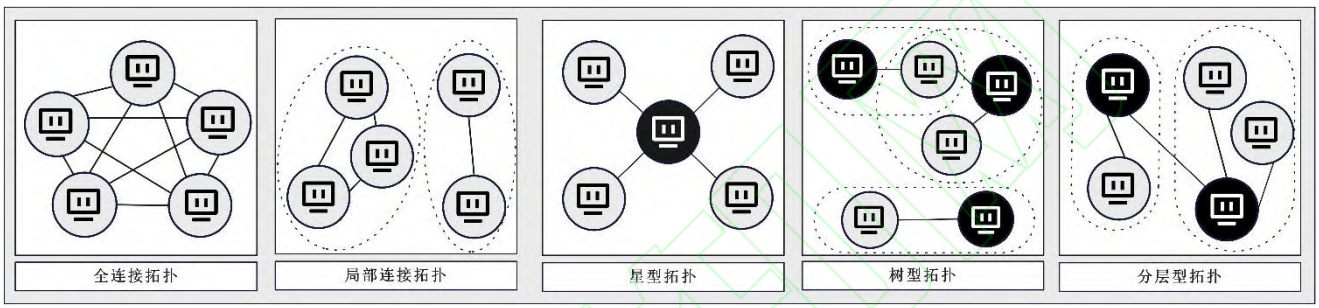


**图 3 预定义的通信对象拓扑分类**

Fig.3 Classification of Predefined Communication Object Topologies

全连接拓扑系统中，每个智能体均能与其他所有智能体直接进行信息交互。这种拓扑结构确保了任意两个智能体之间都存在通信链路，从而实现了最大化的信息交互，适用于分布式电力调度等需要全局协调的任务。相关的方法如 TarMAC[41]中每个智能体在发送消息时，会向所有其他智能体广播包含“签名”和“值”的消息，接收方会接收所有发送方的消息，并通过查询向量计算注意力权重。这种设计既保留了全连接拓扑的信息完整性，又通过注意力机制避免了广播通信的低效性。最新相关研究 CroMAC[56]方法中智能体接收所有其他智能体的消息，将其视为多视图以聚合全局信息，实现鲁棒协作。然而，全连接拓扑的高通信开销特性限制了其在大规模智能体系统中的应用，并且在智能体异质场景中，冗余连接导致无关信息干扰，如异构流水线上，执行精密装配的协作智能体会持续接收所有其他设的广播信息。这些与其核心装配任务无关的海量数据，不仅占用其有限带宽、增加延迟，还可能干扰接收关键的低延迟控制指令，最终影响装配精度和生产效率。

相比之下，局部连接拓扑通过限制通信范围，仅允许智能体与物理或逻辑意义上的邻近智能体进行信息交换。这种方法显著减少了通信负担，如 DGN[40]方法通过邻近节点定义实现局部高效交互，借助多层图卷积扩展合作范围，最终在动态场景中实现高效的策略学习。GAXNet[49]方法中每个无人机智能体基于局部相邻智能体的信息构建注意力图，并通过局部语义通信实现协作，但局部通信可能导致智能体获得全局信息不足，且无法很好的与未建立连接的智能体完整协同，如局部拥堵时无法获取全局绕行路径。

星型拓扑通过引入中心节点来实现信息的汇总与分发，智能体仅与中心节点进行通信，适用于需要集中控制的任务。如 MD-MADDPG[42]在智能体之间维护共享内存，学习从内存中有选择地存储和检索局部观测信息。AMSAC[51]通过建立消息共享网络，并在消息共享网络引入注意力机制，提取每个智能体的全局消息，降低冗余消息的影响，提高决策效率。当然，星型拓扑也并非完美，在星型拓扑中，中心节点的负载较高，容易成为系统瓶颈并且在中心节点失效时可能导致整个通信系统崩溃。

树型拓扑中智能体之间的通信关系呈树状结构组织。树型拓扑通常由一个“根节点”出发，逐层分布连接到若干“子节点”，形成分支结构。每个智能体只与其父节点和子节点进行通信，信息可以通过树状路径逐层传递和聚合。在 ATOC[39]方法中，通信发起者作为临时“根节点”，招募邻近智能体，形成局部通信组。合作者可进一步成为其他组的“分支节点”（如

被多个发起者招募时 ), 形成类似树的层级延伸。在这类方法中，各组既能通过共同节点共享信息，且在单个通信组失效时仅影响该组协作,其他组仍正常工作。但在 ATOC 中共享节点共享信息也存在弊端，即每个组的通信是依次进行的，对于实时系统来说，其较大的时间复杂度是无法接受的。

分层拓扑是一种将智能体按层级结构组织的架构模式，核心特征是通过功能分层实现全局协调与局部执行的分工。如 LSC[45]方法通过动态层级划分、层次化通信流程和自适应结构学习，构建了分层拓扑，既保留了分层结构在全局协调和效率上的优势，又通过强化学习实现了对任务的动态适配。这一设计使其在通信效率、可扩展性和复杂协作上超越传统非分层结构，成为 MARL 中分层拓扑的典型应用。

**表 4　预定义的通信对象分类**

Table 4　Classification of Predefined Communication Objects

| 拓扑类型 | 相关方法 |
|---|---|
| 全连接拓扑 | DIAL[36]、RIAL[36]、TarMAC[41]、VBC[72]、GA-Comm[64]、Diff Discrete[46]、IS[47]、ETCNet[73]、TMC[66]、CroMAC[56]、PAGNet[68]、CoDe[60]、COCOM[69]、MBC[74] |
| 局部连接拓扑 | BiCNet[38]、DGN[40]、SchedNet[43]、I2C[65]、GAXNet[49]、MAGI[55]、PCGQ[57]、ExpoComm[58]、TMAC[59]、AsynCoMARL[71] |
| 星型拓扑 | CommNet[37]、IC3Net[62]、MD-MADDPG[42]、IMAC[44]、HAMMER[48]、AMSAC[51]、AICNet[52]、TDU[53] |
| 树型拓扑 | ATOC[39] |
| 分层拓扑 | LSC[45] |

可学习的通信对象通过学习确定智能体之间的通信关系。这种动态通信机制可以根据任务需求自动调整通信拓扑,让智能体自主学习“谁对当前任务更重要”，摆脱了固定的人工设计拓扑结构的限制。相比于预定义结构，学习型通信对象更为灵活，能够适应复杂和动态的任务环境。根据其核心学习机制，可进一步细分为以下几类：

(1)基于注意力机制[75-78]：智能体利用注意力机制计算与其他智能体的关联权重，并选择权重最高的一个或几个作为通信对象。这种方式通常与消息聚合中的注意力机制紧密结合,代表方法如：MAIC[50]方法通过智能体学习一个有针对性的队友模型，使用类注意力机制来计算与其他所有智能体的通信权重，并通过优化通信权重的熵，修剪掉通信权重较小的无用通信链接，从而减少冗余信息 。CoDe[60]方法通过意图对齐的注意力机制，智能体动态选择与当前意图相关的交互对象。

(2)基于强化学习：将通信对象的选择视为一个独立的动作空间，或通过设计辅助奖励信号，利用强化学习策略直接优化“与谁通信”的决策。这类方法的核心思想是显式优化通信对长期收益的贡献。相关方法如，例如 I2C 方法利用多智能体强化学习中的联合行动-价值函数来推断一个智能体对另一个智能体的影响力，从而量化通信的必要性。这种影响被定义为一种因果效应，并用于为通信决策提供训练标签 。

(3)基于正则化：在通信协议的优化目标中引入信息论正则项，以鼓励生成稀疏、精炼的通信关系，间接地实现通信对象的筛选。相关研究如 NDQ[63]方法通过引入最大化互信息和最小化消息熵这两个信息论正则化器，让智能体能够在执行自身任务的同时，通过最小化通信来学习与谁通信。此外，MAIC[50]、PMAC[52]、MAGI[55]等方法也引入正则化为辅助约束，减少冗余链接，确保通信过程高效而简洁。

(4)基于图神经网络[79-82]：将智能体间的通信建模为动态图结构，并利用图神经网络对通信关系进行建模与优化。通过图结构学习，智能体不仅能够确定与哪些邻居通信，还能学习如何聚合和处理来自不同对象的消息。例如 AsynCoMARL[71]中使用基于智能体位置和行动时间的动态图结构，并通过图 Transformer 学习智能体与其邻居之间的权重来决定与谁通信以及如何处理消息。

可学习的通信对象分类及其代表性方法如表 5 所示。

可学习拓扑通信适用于动态分组任务,如救灾中机器人按需组队；局限性是学习复杂度随智能体数量激增，在智能体数量庞大的智能体系统中难以收敛，另外可能学到“伪相关性”，如两个智能体因偶然动作同步被误认为需持续通信。

预定义拓扑具备确定性与可预测性，通过固定规则或启发式方法建立智能体之间的连接，能够确保通信流稳定、简化协调过程，并降低系统复杂度，适用于任务流程稳定、角色明确、需求清晰的场景。例如分层、去中心化或中心化等静态拓扑结构，在稳定任务领域中可发挥较好效果，且设计、实施与维护相对简便。相比之下，可学习拓扑能够在运行时依据性能指标、工作负载变化或战略约束等动态调整智能体间的连接，更适应复杂多变的环境与突发情况，显著提升系统响应能力与灵活性。然而，其计算开销和收敛性挑战在大规模场景中尤为突出，并可能因偶然相关性而学习到低效或冗余的通信模式。

在此背景下，混合拓扑策略被提出作为折中方案。该类方法将预定义结构作为系统的基础通信框架，提供稳定和可预测的连接，以保证常规情况下的高效通信并降低计算负担；同时引入可学习机制，在局部范围内对连接进行动态调整，使系统能够根据任务的实时需求优化信息传递与协作效率。近年来的“物理/逻辑邻域基础+动态邻居筛选”方案便是代表性思路：其首先基于距离、观测范围或固定邻居列表建立局部连接，再在此受限集合内通过强化学习或注意力机制动态选择通信对象及聚合权重，从而兼顾任务相关性与通信效率。例如 MAGI[55]方法基于局部依赖假设限制智能体仅与邻域范围内的智能体交互，并在此基础上引入图注意力网络（Graph Attention network, GAT），通过动态学习注意力权重量化领域内智能体的重要性。

此外，在一些基于大语言模型的多智能体系统中，可先依据任务类型与智能体功能预定义层级式拓扑，明确角色与基本通信路径，再借助如 DyLAN [83]框架中的前向–后向团队优化步骤，结合执行过程中的性能反馈动态调整部分连接，实现按需的团队重构。这种设计在保持通信结构可控性与效率的同时，赋予系统更强的适应性与复杂任务处理能力，尤其适用于大规模协作型智能体系统的实际部署。

**表 5　可学习的通信对象分类**

Table 5　Classification of Learnable Communication Objects

| 学习机制 | 相关方法 |
|---|---|
| 基于注意力 | ATOC[39]、TarMAC[41]、GA-Comm[64]、MAIC[50]、PAGNet[68]、CoDem、TGCNet[70] |
| 基于强化学习 | MADDPG-M[61]、IMAC[44]、LSC[45]、I2C[65] |
| 基于正则化 | NDQ[63]、IMAC[44]、MAIC[50]、PMAC[52]、MAGI[55] |
| 基于图神经网络 | LSC[45]、GAXNet[49]、PMAC[52]、MAGI[55]、MACS[67]、AsynCoMARL[71] |

## 2.3　通信来源：信息时效性与预测风险的博弈

根据通信内容的来源与构成，现有研究中的通信内容可大致分为基于已有信息的通信与基于未来信息预测的通信两类。通信来源分类及其代表性方法如表 6 所示。

基于已有信息的通信是指智能体在当前时间步根据自身状态、观测、环境信息或历史交互数据进行通信。此类通信内容通常直接来源于智能体的局部信息或通过处理后的特征数据，旨在共享当前状态下的信息以实现更高效的协作。相关研究工作如 TDU[53]方法中，智能体依据当前观测值与历史信息生成原始通信信息，经自注意力机制的编码器和解码器筛选后传输高价值信息。TGCNet[70]方法通过动态有向图将局部观测和历史通信转化为智能的通信决策，实现了按需通信和精准协作。这种方法的深层动因是依赖历史/当前信息，本质是规避预测不确定性——已知信息的可靠性远高于预测结果，适合对决策安全性要求高的场景，如医疗机器人协作。这种方法的不足之处在于，缺乏对未来状态的推断，可能导致协作滞后，如自动驾驶中仅传递当前位置，未预测未来轨迹，易引发追尾。

基于未来信息预测的通信是指智能体通过预测未来的状态、意图或环境变化，生成用于通信的内容。这类通信方法试图在当前信息的基础上推断未来可能的变化，以更好地指导智能体的决策。基于未来信息的通信特别适用于环境变化快速或需要前瞻性协作的任务场景。其主要优点是由于实现了“预判式协作”，能够提升智能体在复杂动态环境中的适应性，促进智能体间的前瞻性协作，提升动态任务中的响应速度，如团队竞技游戏中的走位配合，减少未来可能的冲突或失误。相关研究工作如 ATOC[39]方法通过注意力机制预测未来通信需求，生成基于未来信息的通信内容，

有效减少了冗余通信。IS[47]方法通过共享“想象轨迹”协调行为，通过全局范围内的信息互通，让每个智能体能够基于其他所有智能体的意图调整自身策略。MACS[50]方法通过变分自编码器显式预测队友的未来观察和动作，生成隐藏特征，结合代理历史轨迹生成消息，并利用互信息约束使消息与队友未来 Q 值高度相关，经动态通信链路靶向传输，实现基于未来信息预测的高效协作通信。然而，这种基于未来信息预测的通信也带来了一些挑战，如需要额外的计算资源进行预测，不适合边缘设备，如低功耗传感器；且预测误差可能扩散，如错误预判队友动作导致通信误导，影响决策质量。

**表 6 通信来源分类**

Table 6 Classification of Communication Sources

| 通信来源 | 相关方法 |
|---|---|
| 基于已有信息的通信 | DIAL[36]、RIAL[36]、CommNet[37]、BiCNet[38]、DGN[40]、TarMAC[41]、MADDPG-M[61]、IC3Net[62]、MD-MADDPG[42]、SchedNet[43]、VBC[72]、NDQ[63]、IMAC[44]、GA-Comm[64]、LSC[45]、Diff Discrete[46]、I2C[65]、ETCNet[73]、TMC[66]、HAMMER[48]、GAXNet[49]、MAIC[50]、AMSAC[51]、AICNet[52]、TDU[53]、PMAC[54]、MAGI[55]、CroMAC[56]、PCGQ[57]、ExpoComm[58]、PAGNet[68]、TMAC[59]、CoDe[60]、COCOM[69]、TGCNet[70]、MBC[74]、AsynCoMARL[71] |
| 基于未来信息预测的通信 | ATOC[39]、IS[47]、MACS[67] |

## 2.4 消息聚合：信息融合效率与可解释性的矛盾

当智能体在多智能体系统中接收多个消息时，直接将所有消息作为输入会导致动作策略模型的输入维度爆炸，增加计算复杂度并可能引入冗余信息。消息聚合的目标是使智能体收到更有价值的消息。

文献[17]通过将智能体接收到的消息是否等值处理对消息聚合进行分类。在本节中，我们将重点关注消息聚合的方式，根据消息聚合处理方式的不同将 CB-MADRL 方法分为无显式消息聚合、基于前馈神经网络、基于循环神经网络、基于注意力机制、基于图神经网络和基于自定义规则的聚合。消息聚合方式分类及其代表性方法如表 7 所示。

无显式消息聚合是指在智能体间的信息交流过程中，未引入专门的消息融合机制，针对接收到的消息通常直接与本地状态拼接输入到价值或策略网络，如 NDQ[63]、ETCNet[73]等方法，或者是像 VBC[72]方法将局部 Q 值与多个智能体的有效消息进行元素求和。这种聚合方式主要是为了降低计算复杂度，适用于资源受限设备，如嵌入式系统中的微型机器人。主要局限在于接收方无法过滤噪声，并在消息异构场景（如视觉+文本信息）中性能骤降。

基于前馈神经网络[84-86]的消息聚合通常通过多层非线性变换，计算每条消息的重要性权重。相关研究如 IMAC[44]方法在其调度器中使用前馈神经网络学习各个消息的权重，并为每个智能体生成调度消息。HAMMER[48]方法中中央代理使用前馈神经网络将高维全局信息压缩为低维消息，避免通信冗余。

多智能体环境中，智能体的动作和观测具有时序连续性，循环神经网络（Recurrent Neural Network, RNN）[87]及其变体（如 LSTM、GRU）[88-90]的循环结构能够捕捉序列数据中的长期依赖关系。ATOC[39]方法中通信通道采用双向 LSTM 单元，通过门控机制选择性地保留或遗忘信息。GAXNet[49]中利用 GRU 编码器聚合当前权重与前一时隙相邻无人机的反向注意力权重生成语义消息。

基于注意力机制[75-78]的消息聚合方法通过学习不同消息的重要性分数，对收到的消息加权融合。这类方法能够根据当前环境和任务需求，动态调整对各智能体消息的关注度，从而提升关键信息的利用效率。相关研究如 AICNet[52]使用注意力模块生成动态权重，决定信息聚合的优先级。PAGNet[68]方法以智能体局部观测为查询，其他观测为键值，通过多头注意力计算动态权重，筛选关键信息并加权聚合，实现按需消息生成与协作。

基于图神经网络[79-82]的消息聚合方法，将多智能体系统建模为图结构，通过节点之间的邻接关系进行消息传递和融合。每个智能体可通过多层图神经网络聚合邻居的信息，实现多跳通信和全局信息整合。图神经网络结构包括图卷积网络（Graph Convolutional Network, GCN），图注意力网络等。典型的 GCN 方法有

CommNet[37]、IC3Net[62]、LSC[45]等。GAT 则通过引入可学习的注意力机制,根据节点特征动态计算每个邻居消息的权重,用以将接受到的消息聚合成更有价值的交互信息，相关方法如 MAGI[55]、MACS[67]等。基于注意力机制与图神经网络的消息聚合方式因能动态权衡消息重要性，如战场中优先关注敌方单位消息，且兼容大规模智能体系统，逐渐成为主流。其局限性是注意力权重的“黑箱性”导致难以解释,而图神经网络依赖邻接矩阵设计，在动态拓扑中需频繁重构，效率低下。

基于自定义规则的方法根据任务先验知识、领域经验或任务目标等，手动设定消息聚合方式。代表方法如：MADDPG-M[61]方法设计了基于竞争选择规则的自定义聚合方法。PMAC[54]方法根据智能体身份设计个性化聚合。ExpoComm[58]则是设计了混合的消息聚合方式，对于静态指数拓扑，使用注意力机制聚合消息，而对于单对等指数拓扑，使用 RNN 来聚合消息。

**表 7　消息聚合方式分类**

Table 7　Classification of Message Aggregation Methods

| 消息聚合 | 相关方法 |
|---|---|
| 无显式消息聚合 | DIAL[36]、RIAL[36]、VBC[72]、NDQ[63]、Diff Discrete[46]、ETCNet[73]、PCGQ[57] |
| 基于前馈神经网络 | IMAC[44]、HAMMER[48] |
| 基于循环神经网络 | BiCNet[38]、ATOC[39]、I2C[65]、GAXNet[49] |
| 基于注意力机制 | TarMAC[41]、IS[47]、MAIC[50]、AMSAC[51]、TDU[53]、PAGNet[68]、CoDe[60]、COCOM[69] |
| 基于图神经网络 | CommNet[37]、DGN[40]、IC3Net[62]、GA-Comm[64]、LSC[45]、AICNet[52]、MAGI[55]、MACS[67]、TMAC[59]、TGCNet[70]、MBC[74]、AsynCoMARL[71] |
| 基于自定义规则 | MADDPG-M[61]、MD-MADDPG[42]、SchedNet[43]、TMC[66]、PMAC[54]、CroMAC[56]、ExpoComm[58] |

## 2.5　通信作用：值函数与策略优化的路径选择

通信的核心作用体现在它如何影响智能体的学习与决策过程，进而影响系统的整体性能。在大多数现有文献中，消息被视为额外的观测值。智能体将消息作为策略函数、价值函数或两者的额外输入。根据通信在学习过程中所产生的影响，可以将其大致分为三类：对 Q 值的影响、对策略的影响以及对 Q 值和策略的共同影响。通信作用分类及其代表性方法如表 8 所示。

**表 8　通信作用分类**

Table8　Classification of Communication Functions

| 通信作用 | 相关方法 |
|---|---|
| 对 Q 值的影响 | DIAL[36]、RIAL[36]、DGN[40]、VBC[72]、NDQ[63]、LSC[45]、TMC[66]、MAIC[50]、MAGI[55]、CroMAC[56]、MACS[67]、PAGNet[68]、TMAC[59]、CoDe[60]、COCOM[69]、TGCNet[70] |
| 对策略的影响 | CommNet[37]、ATOC[39]、IC3Net[62]、MD-MADDPG[42]、SchedNet[43]、IMAC[44]、GA-Comm[64]、DiffDiscrete[46]、I2C[65]、IS[47]、ETCNet[73]、HAMMER[48]、GAXNet[49]、AMSAC[51]、AICNet[52]、TDU[53]、PMAC[54]、ExpoComm[58]、MBC[74]、AsynCoMARL[71] |
| Q 值和策略的共同影响 | BiCNet[38]、TarMAC[41]、MADDPG-M[61]、PCGQ[57] |

对 Q 值的影响主要体现在通过信息共享来改进 Q 值函数的估计。智能体通过通信获取其他智能体的观测或状态信息，从而弥补自身观测的局限性，优化联合 Q 值的学习。通信的目的是提升 Q 值函数的准确性，使各智能体的决策更加接近全局最优。此类方法适用于基于值函数的 MARL 方法，如 VDN[25]和 QMIX[26]。在这种机制下，通信内容通常包括与 Q 值相关的特征、状态或奖励信息，能够减少智能体间的信息不对称，提升联合 Q 值的估计精度，并加速学习过程。相关研究如 DGN[40]方法利用图神经网络对邻近智能体的 Q 值信息进行聚合,提升联合 Q 值计算的准确性。TMAC[59]方法中，通信通过自消息融合模块和消息处理器提取、整合多源信息，将处理后的特征输入 Q 网络，使 Q 值更准确反映动作价值。这种方法的局限性在于仅适用于离散动作空间,在连续动作场景，如机械臂协作中难以扩展。

对策略的影响则主要体现在通过信息共享来直接优化策略，避免值函数在多智能体交互中难以分解的问题。在策略优化过程中，智能体利用通信内容调整自身行为，从而实现更高效的协作或竞争。通信内

容通常涉及智能体的意图、动作建议或策略特征，有助于促进智能体间的协作，减少竞争或冲突，并提高策略的鲁棒性。相关研究如 IMAC[44]方法通过对策略的意图建模实现了智能体间的高效协作。IC3Ne[62]t 方法通过门控机制允许智能体共享策略相关信息，从而动态协调智能体行为。这种方法的局限性是策略过度依赖通信，可能导致 “盲从”，如某智能体错误消息引发群体决策失误。

通信对 Q 值和策略的共同影响是指通信机制同时作用于值函数的估计和策略的优化。此类方法试图通过通信实现 Q 值估计和策略生成的协同改进，以构建更高效的强化学习框架。这类方法兼顾估计精度与动作灵活性，适合复杂协作 - 竞争任务（如自动驾驶中的换道博弈）。在这一类的 CB-MADRL 方法中，一种是将接收到的消息作为演员和评论家模型的额外输入，如 BiCNet[38]方法，另一种则是将消息与局部观测相结合，生成新的内部状态然后与演员和评论家模型共享，如 TarMar[38]和 PCGQ[57]方法。这类方法的局限性在于双目标优化易导致训练不稳定，如 Q 值更新与策略梯度方向冲突。

## 2.6 通信学习：优化目标与场景约束的适配

通信协议的学习旨在动态调整智能体通信协议，使通信内容与任务目标高度一致，从而提升系统整体性能。现有研究主要探索了四种通信学习方法，包括可微分、强化学习、监督学习和正则化。通信学习分类及其代表性方法如表 9 所示。

可微分通信学习通过将通信模块设计为可微分的神经网络组件，使得通信生成与任务目标可以通过梯度反向传播进行端到端优化。这种方法将通信信息直接整合到策略学习或 Q 值学习过程中，使得通信学习与整体任务紧密结合，但该方法在处理离散通信信息时往往需要额外的技巧，如 GA-Comm[64]方法通过引入 Gumbel-Softmax[91]实现梯度的反向传播。而 Diff Discrete[46]方法提出了一种随机消息编解码程序，通过引入随机噪声扰动，将离散通信信道在数学上等效为含加性噪声的模拟信道，使得梯度能够通过离散信道进行反向传播。

强化学习方法使用环境奖励或自定义奖励来逐步更新通信策略或消息。智能体在环境中执行动作并根据奖励信号来评估动作的好坏，进而调整自己的策略，以最大化长期累积奖励。在通信协议学习中，智能体通过尝试不同的通信策略，根据获得的奖励来优化通信行为。如 ETCNet[73]方法中将带宽约束建模为带惩罚项的马尔可夫决策过程，通过优化门控网络的发送概率实现“按需通信”，平衡协作性能与带宽消耗。MADDPG-M[61]通过分层策略和设计内在奖励，使智能体自主学习信息共享，为高噪声多智能体系统提供有效解决方案。最新相关研究 PCGQ[57]方法利用 Q 值增益和奖励函数引导智能体动态优化通信决策。这种方法的局限性是当奖励稀疏时通信策略易陷入局部最优，如始终选择不通信以节省成本。

监督通信学习方法则依赖于明确的通信目标或标签来引导通信行为的生成，通过专家演示或预定义规则提供监督信号，从而在训练过程中直接优化通信内容的质量。由于监督信号明确，这种方法具有较高的稳定性和收敛速度，适用于具有明确任务先验的结构。I2C[65]方法中使用监督学习训练先验网络（Prior Network），通过因果效应标注通信，使智能体学会动态筛选通信对象。ATOC[39]方法中注意力单元被设计为一个二元分类器，通过计算通信前后智能体动作价值的差异，生成反映通信必要性的监督信号。

正则化通信学习通过在损失函数中加入正则化项（如信息熵约束或稀疏性约束）来抑制通信中的冗余信息，从而提高通信效率。该方法在优化任务目标的同时，致力于减少无用通信，适合资源受限或任务复杂的场景。IMAC[44]利用信息瓶颈理论，最大化消息对观测的压缩效率,即最小化互信息上界，提升通信性价比。MAGI[55]方法通过最大化消息与动作的互信息和约束消息与智能体特征的互信息，确保消息含动作选择必要信息并且剔除冗余噪声，从而平衡鲁棒性与表达性。

以上两种方法具有较强的稳定性，适合安全关键场景，如医疗机器人，但由于监督信号依赖先验，泛化到新任务时性能下降；正则项可能过度限制通信表达能力。

**表 9 通信学习分类**

Table 9 Classification of Communication Learning

| 通信学习 | 相关方法 |
|---|---|
| 可微分 | DIAL[36]、CommNet[37]、BiCNet[38]、DGN[40]、TarMAC[41]、MD-MADDPG[42]、VBC[72]、GA-Comm[64]、Diff Discrete[46]、IS[47]、TMC[66]、GAXNet[49]、AMSAC[51]、AICNet[52]、PMAC[54]、MAGI[55]、MACS[67]、PAGNet[68]、CoDe[60]、COCOM[69]、TGCNet[70] |
| 强化学习 | RIAL[36]、MADDPG-M[61]、IC3Net[62]、SchedNet[43]、LSC[45]、ETCNet[73]、HAMMER[48]、TDU[53]、CroMAC[56]、PCGQ[57]、TMAC[59]、TGCNet[70]、AsynCoMARL[71] |
| 监督学习 | ATOC[39]、I2C[65]、MACS[67]、ExpoComm[58]、COCOM[69]、MBC[74] |
| 正则化 | NDQ[63]、IMAC[44]、MAIC[50]、PMAC[54]、MAGI[55]、CoDe[60]、COCOM[69] |

## 2.7 通信约束：现实物理限制的妥协与突破

在 CB-MADRL 框架中，通信约束问题作为关键性挑战值得研究者特别关注。实际应用场景中的通信机制往往受制于物理环境和系统架构的多重限制，这些非理想通信条件会显著影响智能体间的协同效率，甚至导致整个系统的性能劣化。根据通信资源的可用性特征，我们可将通信约束问题划分为无约束通信和有约束通信。通信约束分类及其代表性方法如表 10 所示。

无约束通信作为理论研究的基准场景，该模式下假设存在理想的通信基础设施：智能体间可实现无延迟、无损耗的全维度信息交互。这种理想化假设虽然为算法设计提供了参考基准，但难以直接应用于实际任务中。相关方法如 CommNet[37]、TarMAC[41]、DGN[40]等。值得一提的是，类似于 ATOC[39]、IC3Net[62]、GA-Comm[64]等方法，这些研究中考虑通过预定义一些通信规则或可学习的方式来动态调整通信策略，使得智能体并不总是能完成通信，但这有别于我们后续所讨论的通信限制。这些研究并没有设定显式的通信限制，故也将上述方法归类到无约束通信中。

近些年来为了解决模拟环境中“理想信道”假设与现实的差距，通信约束被引入到 CB-MARDL 中，通过对通信约束的研究，促进多智能体系统在不可靠通信条件下仍能稳定协作。通信约束通常包含有限带宽约束、传输损失约束等。

有限带宽约束指通信过程受信道容量限制，基本思想是通过压缩消息或筛选关键信息，适配实际通信信道。如 MACS[67]在通信次数和消息大小上优化效率，以应对现实中带宽资源有限的挑战：通过图神经网络动态生成通信图，利用反事实假设剪枝冗余链路，并通过互信息约束生成高价值消息，实现按需通信，在有限带宽下减少无效传输，提升通信效率。而 IMAC[44]方法推导消息与观测的互信息上限，最小化消息熵，使智能体生成高信息密度的低维消息。COCOM[69]方法通过门控机制动态筛选必要消息、利用隐式共识减少通信依赖、以互信息优化消息内容避免冗余。这类方法的局限性是过度压缩可能丢失关键细节，如压缩后的图像消息无法区分障碍物类型。

传输损失约束指物理信道特性或环境干扰导致的通信质量下降，直接影响信息传输的完整性与时效性。基本思想是通过加密或鲁棒编码对抗噪声/干扰，适合实际无线通信场景。Diff Discrete[46]通过在发送消息前引入加密步骤，使信道噪声与消息内容无关，从而在梯度计算中忽略噪声影响，实现未知噪声下的无偏导数估计。CroMAC[56]方法通过多视图建模将消息视为状态的不同视角，利用 MVAE 和专家乘积聚合消息，结合区间边界传播，计算扰动边界，在潜在空间引入扰动优化 Q 值重叠，确保消息表示和决策在噪声干扰下的鲁棒性。安全约束指由恶意行为者或异常节点引发的通信可靠性风险。抗干扰机制增加计算开销，在实时性要求高的场景，如无人机避障，中可能失效。

值得注意的是，现实世界中的通信约束往往具有时空耦合效应，即不同类型的约束通常不是独立发生，而是伴随环境变化叠加出现。AsynCoMARL[71]方法利用动态图仅连接邻近活跃智能体并结合图 Transformer 筛选关键消息，应对有限带宽约束。同时通过定义独立时间尺度、随机行动间隔及掩码矩阵过滤非活跃节点，适应传输损失中的异步通信约束。

**表 10 通信约束分类**

Table 10 Classification of Communication Constraints

| 通信约束 | 方法 |
|---|---|
| 无约束通信 | CommNet[37]、BiCNet[38]、ATOC[39]、DGN[40]、TarMAC[41]、MADDPG-M[61]、IC3Net[62]、MD-MADDPG[42]、GA-Comm[64]、LSC[45]、I2C[65]、IS[47]、HAMMER[48]、GAXNet[49]、AMSAC[51]、AICNet[52]、TDU[53]、PMAC[54]、MAGI[55]、PCGQ[57]、ExpoComm[58]、PAGNet[68]、TMAC[59]、TGCNet[70] |
| 有限带宽约束 | DIAL[36]、RIAL[36]、SchedNet[43]、VBC[72]、NDQ[63]、IMAC[44]、ETCNet[73]、TMC[66]、MAIC[50]、MACS[67]、COCOM[69]、MBC[74]、AsynCoMARL[71] |
| 传输损失约束 | DIAL[36]、Diff Discrete[46]、CroMAC[56]、CoDe[60]、AsynCoMARL[71] |

## 3 讨论

### 3.1 总结与展望

本节的内容是对上述提出的CB-MADRL七个维度总结与未来展望。在表 11 给出了通信七维度分类符号表示，表 12 将本文涉及的 CB-MADRL 方法七维度进行了汇总。

在通信时机维度上，三类通信时机形成覆盖不同需求的应用场景：全时通信确保信息完备性但代价高昂，自适应通信时机在动态环境中寻求平衡，事件触发通信则为极端资源受限场景提供解决方案。如今在复杂的动态环境中的通信时机选择仍然有优化空间。未来可以设计使用混合通信时机的算法，如在常规状态下使用事件触发机制，遇突发状况自动切换为密集通信模式，以此适应更复杂的智能体环境。

在通信对象维度方面，得益于深度学习的发展，近期许多工作都集中在可学习的通信对象。大规模多智能体系统中，智能体因分组、权限、物理连接等形成受限的通信结构，导致直接通信不可行的或低效的。树型拓扑和分层拓扑这类结构性通信，将智能体划分为多个组，仅允许组内通信或通过共享节点，打破了组间隔离，适用于大规模多智能体系统。但无论是ATOC[39]还是 LSC[45]在组内共享信息时，高层向底层通信时往往采用广播的形式，这些共享信息往往缺乏针对性。结构性通信未来的一个重要方向是向目标智能体发送关键信息和意见。未来可以在该通信模块中引入如注意力机制[75-78]有关方法，来设计和优化针对组内不同智能体的动态个性化传输信息并学习更有通信意义的通信对象，使该类方法在存在有限带宽约束的环境下取得更好表现。

在消息来源维度的研究中，现有方法主要依赖智能体的历史观测、动作轨迹等生成消息。这类方法通过编码历史信息或传递当前动作概率分布，虽具有低维编码和计算高效的优势，却因局限于反映过去或当前状态，容易导致协作滞后。部分研究尝试融入意图或未来计划，虽能提升协作效率，却面临动态建模不足的挑战。基于模型的强化学习（Model-based RL）[92]可以帮助代理对未来情况做出更准确的预测，从而使智能体能够更确定地传达有关即将发生的变化的信息。此外，后续研究也还可以通过变分自编码器[93]或生成对抗网络[94]建模状态转移模型等，增强对未来环境的推演能力，使通信信息更具前瞻性和可靠性。

在消息聚合维度上，图神经网络和注意力机制能够充分挖掘多智能体间的结构信息与内容关系，在消息聚合中兼顾通信结构、信息重要性和模型扩展性，成为当前聚合方法中的主流选择。然而，现实任务的核心挑战在于处理多模态和异构通信消息。智能体需要整合来自视觉传感器、激光雷达、文本指令、音频信号、物理传感器读数等不同来源、不同类型的信息，这些信息在维度、语义和可靠性上存在巨大差异。如何有效整合这些信息，实现高效的消息聚合，是提升多智能体协作能力和智能水平的重要研究方向。未来可以在消息聚合维度中引入多模态模型，如 CLIP[95], Flamingo[96]等。多模态模型能有效编码视觉、语言、传感等异构信息为统一的语义表示进行通信，并在接收端解码还原，为高效的多模态信息融合通信提供新

范式。

在通信作用方面，以基于演员评论家模型为框架引入通信时，通信往往对策略或策略和 Q 值同时产生影响，而仅对 Q 值产生影响的方法通常基于各种值分解方法。在当前 CB-MADRL 方法中，通信消息的内容和机制设计多基于黑箱模型，缺乏对消息语义的明确建模与解释，也难以准确衡量每条消息对 Q 值或最终决策表现的具体贡献。这不仅限制了系统的透明性与调试效率，也制约了其在高安全要求场景，如智能医疗、自动驾驶中的应用。因此，未来通信机制的发展亟需借助可解释人工智能（Explainable Artificial Intelligence, XAI）[97]的理念与方法，对通信协议的结构、内容与决策路径进行可视化和归因分析，明确每条消息对全局 Q 值估计或个体策略优化的具体影响，从而增强通信协议的可信度和人类可控性。可进一步尝试通过设计带有因果解释能力的通信模块，在保证性能的前提下提升通信机制的可解释性。

与此同时，近年来兴起的大规模预训练模型，如 GPT[98]、LLaMA[99]等在自然语言、图像理解和跨模态生成方面表现卓越，也为 CB-MADRL 通信内容的构造提供了新思路。预训练模型可以作为通信生成模块的基础，让智能体之间实现自然语言或图文形式的信息交流。这种“类人通信”的机制不仅提升了通信内容的表达力，也为系统的可解释性与人机协同提供了可能。例如，在协作任务中，智能体间通过简洁的语义描述共享意图，有助于人类对通信过程的观察与干预。后续研究可进一步探讨如何将预训练模型与强化学习框架融合，使其具备任务适应性与泛化性，并通过联动 XAI 技术构建可解释、语义化的通信协议体系。

通信学习本质上是一个交互优化问题。发送方需要获知其消息是否被接收方有效利用，才能不断调整和优化自身的通信策略。可微分方法适合端到端优化和连续通信场景；强化通信学习能处理离散通信和动态调整任务；监督通信学习适用于有明确通信目标或先验知识的场景，如恶意消息检测、结构化通信；正则化通信学习则更适合在资源受限场景下限制冗余信息，如带宽受限的传感器网络或边缘计算场景。近期 CB-MADRL 方法尝试使用多种通信学习方式，以应对更复杂的任务环境。通信学习本质决定了通信协议的学习高度依赖于即时反馈。然而，在实际多智能体系统中，及时、准确地获取有效反馈存在诸多挑战：在复杂环境下，智能体间的通信行为往往难以与最终任务表现直接建立关联，导致反馈信号稀疏或延迟，并且在动态或非稳定环境中，过去的反馈可能不再适用于新的通信场景，如何实现反馈知识的泛化和迁移，也是通信学习面临的难题。在未来的研究中尝试通过自监督学习[100]和辅助任务设计[101]提供额外的学习信号，帮助智能体提升通信协议的学习效率和鲁棒性。如智能体可以基于当前观测和收到的消息，预测未来的环境状态或奖励，这样既能提升对消息作用的感知，也为消息发送方提供了间接反馈。此外，还可以引入预训练技术加速通信学习。一方面，可以在大规模多智能体交互数据集或模拟环境中预训练通用的消息编码器和解码器模型，学习基础通信模式和语义表示，后续通过微调快速适配新任务，大幅降低样本复杂度和训练时间。另一方面，可以借助探索提示工程技术，使智能体根据当前任务状态的文本描述（或嵌入表示），动态调整其通信策略，利用大型语言模型的上下文适应能力实现灵活通信。

在通信约束中，提到现实世界中的通信约束具有时空耦合效应，这种复合型、动态性的约束给现有的 CB-MADRL 算法带来了前所未有的挑战。目前大多数 CB-MADRL 算法只针对单一或静态的通信约束进行优化，缺乏对多种约束时序耦合、协同作用的建模与适应能力。这导致算法在面对现实复杂通信环境时，稳定性和适应性不足。后续工作可以给算法设计环境感知模块[102]实时监测信道状态、带宽变化等多种约束，并驱动智能体自适应地调整消息内容、压缩率、发送频率等通信参数，提高通信的鲁棒性和资源利用率，或者引入时序预测模型[103,104]，对未来的约束变化进

行提前感知和策略预调度，减轻突发性约束变化对系统性能的冲击。

面对极端通信约束，如极低带宽、高丢包率时，融合预训练模型所具备的语义理解与生成能力，正在成为提升通信鲁棒性的潜在突破口。一种可行思路是实现语义感知的极限压缩，具体而言，在带宽极度受限的情况下，智能体可以借助大语言模型对任务上下文和协作目标的理解能力，动态提取和传输最具信息价值的语义片段，而非依赖传统的数值压缩方法，从而实现超越比特层面的“语义级压缩”。这种方式能够在有限的传输资源下保留关键信息，保障多智能体协同的有效性。另一种方向是基于模型的容错推理机制，当通信遭遇严重的消息丢失或数据损坏时，智能体可以利用预训练模型所具备的世界知识和上下文理解能力，结合历史通信模式、环境常识以及当前的局部观测，对缺失的信息进行合理补全或推理，预测对方智能体的意图或可能传递的内容。这种推理能力显著增强了系统在恶劣通信环境下的稳定性与生存能力，使得智能体即使在通信受限的条件下，也能维持较强的协作水平。

### 3.2 通信维度间的耦合关系分析

本文提出的通信七维度框架为解构 CB-MADRL 的通信行为提供了结构化视角。然而，需要强调的是，这些维度在实践中并非孤立存在，而是存在深刻的相互作用与依赖关系。理解这些耦合关系对于设计高效、鲁棒、可扩展的通信机制至关重要。

通信时机的设计直接受限于现实通信资源条件。有限带宽或信道传输损失等约束，使得“全时通信”在实际环境中代价过高，甚至不可行。因此，“自适应通信”或“事件触发通信”成为主流选择，它们能在任务需求与资源消耗之间实现更合理的权衡。例如，在带宽受限场景中，事件驱动机制可仅在关键状态发生时通信，从而显著降低传输负担。此外，异步通信或动态带宽调度机制的引入，也对通信时机提出了更高的适配性要求。

通信对象的设计对消息聚合与通信学习机制构成直接约束。可学习的通信对象可根据任务动态调整通信邻居，从而降低冗余信息，提升通信价值密度，减少对复杂时机策略的依赖。而在拓扑结构确定的场景中，如局部连接或分层通信中，则需借助高效的聚合机制实现信息的整合与去噪，避免信息冗余和效能退化。

在此基础上，通信学习机制则需协同消息选择与聚合策略进行联合优化。例如在高通信负担的场景中，可通过正则化方法压缩冗余信息，在监督学习中引导关键信息生成。多维度间的这种交互协同，是实现高通信效率和模型性能的关键。

通信内容来源于历史信息或对未来的预测，会直接影响通信在学习中的作用。基于历史信息的通信既可以用来更新策略也可以用来优化 Q 值估计，帮助缓解部分可观测性带来的不确定性；而基于未来信息预测的通信，则为策略生成提供前瞻性的行为参考，对策略优化贡献更大。例如，通过想象轨迹或意图预测生成的消息，可帮助智能体协调未来行动方向，在合作-竞争混合场景中表现出更强的适应性。

与此同时，通信作用的定位也决定了消息生成与聚合策略的设计重点。通信仅用于优化策略时，则会关注信息表达的语义与行为关联；若用于优化 Q 值，则更需关注状态-动作之间的数值映射精度。

通信学习不仅是协议设计的实现手段，更是各维度协同的驱动器。例如，在带宽受限场景下，引入稀疏性正则项可引导通信策略向低频高质转化；在多模态输入环境下，监督学习可指导智能体从复杂感知数据中提取有用特征；在动态拓扑中，强化学习机制能根据通信反馈调整邻接策略，实现鲁棒信息流控制。因此，通信学习维度需与通信时机、对象、聚合等机制紧密结合，形成端到端的联合优化体系。

这些耦合关系揭示了 CB-MADRL 通信机制设计的复杂性和系统性挑战。孤立地优化单个维度往往难

以达到最优效果，甚至可能在其他维度引入瓶颈，例如，追求极致的消息压缩可能损害通信作用的有效性。因此，未来研究亟需采用更加协同的视角，探索联合优化多个维度的通信框架，理解不同维度组合在不同任务场景下的效能边界。

目前CB-MADRL维度组合优化仍以启发式设计和实验验证为主，但未来迫切需要发展系统的理论与经验方法来指导这一复杂过程。本文认为，可以通过参考神经架构搜索（Neural Architecture Search, NAS）[105]，提出一种通信架构搜索（Communication Architecture Search，CAS）方法：将维度选择视为超参数优化问题，利用贝叶斯优化、进化算法或可微分NAS，在多任务联合训练中探索Pareto最优的维度组合。如在无人机集群仿真中，通过CAS方法在密集障碍环境下，最优组合为事件触发+局部拓扑+图注意力聚合；而在开阔区域巡航时，自适应通信+星型拓扑+RNN聚合更高效。此外，还可以设计信息-价值-代价理论模型，建立量化分析框架，将通信行为建模为信息增益、任务价值提升、通信代价的函数，定义维度组合优化目标，指导后续维度组合优化。

### 3.3 未来研究方向

本文认为，CB-MADRL未来的核心方向有：

1. 场景驱动的通信机制设计：不同任务在智能体规模（小/大规模）、动作空间类型（离散/连续）、环境动态性（静态/高速变化）等方面存在显著差异，这些因素直接影响通信机制的适配性与性能。未来研究可面向不同应用场景建立任务特征与通信机制选择之间的映射关系，并探索基于场景特征的自适应通信机制生成方法，从而实现通信设计的针对性与普适性兼顾。

2.局限性的量化分析：现有研究多依赖实验验证通信机制的性能提升，但较少系统探讨其失效条件与适用边界。例如，全时通信虽能确保信息完备性，但在智能体规模扩张时会引发通信负载激增，其性能在智能体数量超过一定阈值后出现明显下降。未来可通过理论建模与系统性实验相结合，量化不同通信机制在智能体规模、带宽限制、噪声水平等关键因素下的性能拐点，为通信策略选择提供可预测的参考依据。

3.维度组合的系统化优化：当前CB-MADRL在多维度的组合优化上，缺乏系统性方法，未来可以通过引入CAS框架与建立信息-价值-代价理论模型，在多任务场景中探索最优组合并指导维度组合优化。

4.新兴技术的深度融合：积极探索大型预训练模型在语义通信、加速学习、极限压缩与容错方面的潜力，以及XAI在打开通信“黑箱”、实现归因驱动聚合和协议诊断方面的作用，是推动CB-MADRL通信迈向新阶段的关键动力。

总之，未来研究需更紧密结合现实任务的物理约束与协作需求，充分利用预训练模型和可解释AI等前沿技术，推动通信机制从“性能优化”向“可解释、鲁棒、语义化”的实用化方向发展。

## 4 结束语

多智能体深度强化学习正处于从理论探索向复杂现实场景渗透的关键阶段，而通信机制作为破解多智能体协作难题的核心枢纽，其中重要性不言而喻。本文系统地从通信时机、通信对象、通信来源、消息聚合、通信作用、通信学习与通信约束七个维度出发，对基于通信的多智能体深度强化学习方法进行了全面梳理与分析。首先，我们在现有文献的基础上，首次对通信时机进行了细致划分，并提出了针对通信对象和消息聚合的新分类方式；其次，通过设计统一的分类符号与标注体系，总结了各类方法的核心特点与应用场景；最后，在总结已有工作的基础上，针对多智能体通信面临现实问题于挑战，给出了研究思路与方向。本文提出的七维度分析框架，不仅为理解与设计CB-MADRL通信机制提供了系统化的理论工具，也为构建更高效、鲁棒且可解释的多智能体通信系统奠定了坚实基础，对未来研究具有重要的指导意义。随着多智能体系统在自动驾驶、智能制造、机器人群协同

等领域的广泛应用，希望后续研究者能够基于本文框架不断拓展理论边界，推动通信协议与学习算法的有机结合，为实现大规模、多场景下的智能体协作奠定坚实基础。

**表 11 通信七维度分类符号表示**

Table 11 Notation for Classification of Seven Communication Dimensions

| 维度 | 符号 |
|---|---|
| 通信时机 | $C_{es}$：全时通信、$C_{at}$：自适应通信时机、$C_{et}$：事件触发通信 |
| 通信对象 | $F_t$：全连接拓扑、$L_t$：LT 局部连接拓扑、$S_t$：星型拓扑、$T_t$：树型拓扑、$H_t$：分层拓扑、$L_{rl}$：基于强化学习、$L_{att}$：基于注意力机制、$L_{rg}$：基于正则化、$L_{gnn}$：基于图神经网络、 |
| 通信来源 | $E$：基于已有信息的通信、$F$：基于未来信息预测的通信 |
| 消息聚合 | $N$：无显式消息聚合、$F_{fnn}$：基于前馈神经网络、$R_{nn}$：基于循环神经网络、$A$：基于注意力机制、$G$：基于图神经网络、$R$：基于自定义规则的聚合 |
| 通信作用 | $Q$：对 Q 值的影响、$P$：对策略的影响、$PQ$：Q 值和策略的共同影响 |
| 通信学习 | $D$：可微分、$R_f$：强化学习、$S$：监督学习、$R_g$：正则化 |
| 通信约束 | $U$：无约束通信、$L_b$：有限带宽约束、$T_s$：传输损失约束 |

**表 12 CB-MADRL 方法七维度总览**

Table 12 Overview of Seven Dimensions in CB-MADRL Methods

| 方法名称 | 通信时机 | 通信对象 | 通信来源 | 消息聚合 | 通信作用 | 通信学习 | 通信约束 |
|---|---|---|---|---|---|---|---|
| DIAL[36] | $C_{es}$ | $F_t$ | $E$ | $N$ | $Q$ | $D$ | $L_b+T_s$ |
| RIAL[36] | $C_{es}$ | $F_t$ | $E$ | $N$ | $Q$ | $R_f$ | $L_b$ |
| CommNet[37] | $C_{es}$ | $S_t$ | $E$ | $G$ | $P$ | $D$ | $U$ |
| BiCNet[38] | $C_{es}$ | $L_t$ | $E$ | $R_{nn}$ | $PQ$ | $D$ | $U$ |
| ATOC[39] | $C_{es}$ | $T_t+L_{att}$ | $F$ | $R_{nn}$ | $P$ | $S$ | $U$ |
| DGN[40] | $C_{es}$ | $L_t$ | $E$ | $G$ | $Q$ | $D$ | $U$ |
| TarMAC[41] | $C_{es}$ | $F_t+$ | $E$ | $A$ | $PQ$ | $D$ | $U$ |
| MADDPG-M[61] | $C_{at}$ | $L_{rl}$ | $E$ | $R$ | $PQ$ | $R_f$ | $U$ |
| IC3Net[62] | $C_{at}$ | $S_t$ | $E$ | $G$ | $P$ | $R_f$ | $U$ |
| MD-MADDPG[42] | $C_{es}$ | $S_t$ | $E$ | $R$ | $P$ | $D$ | $U$ |
| SchedNet[43] | $C_{es}$ | $L_t$ | $E$ | $R$ | $P$ | $R_f$ | $L_b$ |
| VBC[72] | $C_{et}$ | $F_t$ | $E$ | $N$ | $Q$ | $D$ | $L_b$ |
| NDQ[63] | $C_{at}$ | $L_{rg}$ | $E$ | $N$ | $Q$ | $R_g$ | $L_b$ |
| IMAC[44] | $C_{es}$ | $S_t+L_{rl}+L_{rg}$ | $E$ | $F_{fnn}$ | $P$ | $R_g$ | $L_b$ |
| GA-Comm[64] | $C_{at}$ | $F_t+L_{att}$ | $E$ | $G$ | $P$ | $D$ | $U$ |
| LSC[45] | $C_{es}$ | $H_t+L_{rl}+L_{gnn}$ | $E$ | $G$ | $Q$ | $R_f$ | $U$ |
| Diff Discrete[46] | $C_{es}$ | $F_t$ | $E$ | $N$ | $P$ | $D$ | $T_s$ |
| I2C[65] | $C_{at}$ | $L_t+L_{rl}$ | $E$ | $R_{nn}$ | $P$ | $S$ | $U$ |
| IS[47] | $C_{es}$ | $F_t$ | $F$ | $A$ | $P$ | $D$ | $U$ |
| ETCNet[73] | $C_{et}$ | $F_t$ | $E$ | $N$ | $P$ | $R_f$ | $L_b$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TMC[66] | $C_{at}$ | $F_t$ | $E$ | $R$ | $Q$ | $D$ | $L_b$ |
| HAMMER[48] | $C_{es}$ | $S_t$ | $E$ | $F_{fnn}$ | $P$ | $R_f$ | $U$ |
| GAXNet[49] | $C_{es}$ | $L_t+L_{gnn}$ | $E$ | $R_{nn}$ | $P$ | $D$ | $U$ |
| MAIC[50] | $C_{es}$ | $L_{att}+L_{rg}$ | $E$ | $A$ | $Q$ | $R_g$ | $L_b$ |
| AMSAC[51] | $C_{es}$ | $S_t$ | $E$ | $A$ | $P$ | $D$ | $U$ |
| AICNet[52] | $C_{es}$ | $S_t$ | $E$ | $G$ | $P$ | $D$ | $U$ |
| TDU[53] | $C_{es}$ | $S_t$ | $E$ | $A$ | $P$ | $R_f$ | $U$ |
| PMAC[54] | $C_{es}$ | $L_{gnn}+L_{rg}$ | $E$ | $R$ | $P$ | $D+R_g$ | $U$ |
| MAGI[55] | $C_{es}$ | $L_t+L_{gnn}$ | $E$ | $G$ | $Q$ | $D+R_g$ | $U$ |
| CroMAC[56] | $C_{es}$ | $F_t$ | $E$ | $R$ | $Q$ | $R_f$ | $T_s$ |
| MACS[67] | $C_{at}$ | $L_{gnn}+L_{rg}$ | $F$ | $G$ | $Q$ | $D+S$ | $L_b$ |
| PCGQ[57] | $C_{es}$ | $L_t$ | $E$ | $N$ | $PQ$ | $R_f$ | $U$ |
| ExpoComm[58] | $C_{es}$ | $L_t$ | $E$ | $R$ | $P$ | $S$ | $U$ |
| PAGNet[68] | $C_{at}$ | $F_t+L_{att}$ | $E$ | $A$ | $Q$ | $D$ | $U$ |

| 方法名称 | 通信时机 | 通信对象 | 通信来源 | 消息聚合 | 通信作用 | 通信学习 | 通信约束 |
|---|---|---|---|---|---|---|---|
| TMAC[59] | $C_{es}$ | $L_t$ | $E$ | $G$ | $Q$ | $R_f$ | $U$ |
| CoDe[60] | $C_{es}$ | $F_t+L_{att}$ | $E$ | $A$ | $Q$ | $D+R_g$ | $T_s$ |
| COCOM[69] | $C_{at}$ | $F_t$ | $E$ | $A$ | $Q$ | $D+S+R_g$ | $L_b$ |
| TGCNet[70] | $C_{at}$ | $L_{att}$ | $E$ | $G$ | $Q$ | $D+R_f$ | $U$ |
| MBC[74] | $C_{et}$ | $F_t$ | $E$ | $G$ | $P$ | $S$ | $L_b$ |
| AsynCoMARL[71] | $C_{at}$ | $L_t+L_{gnn}$ | $E$ | $G$ | $P$ | $R_f$ | $L_b+T_s$ |

## 参考文献：


[1] LI Y. Deep reinforcement learning: An overview[J]. arXiv preprint arXiv:1701.07274, 2017.

[2] WANG X, WANG S, LIANG X, et al. Deep reinforcement learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(4): 5064-5078.

[3] 王戎骁, 冯旸赫, 张龙飞, 等. 多智能体强化学习中的信息交互方法[J]. 指挥与控制学报, 2024: 1-14.
WANG R X, FENG Y H, ZHANG L F, et al. A Review of Multi-agent Reinforcement Learning with Com- munication[J]. Journal of Command and Control , 2024: 1-14.

[4] FRANOIS-LAVET V, HENDERSON P, ISLAM R, et al. An Introduction to Deep Reinforcement Learning[J]. Foundations and Trends® in Machine Learning, 2018, 11(3-4): 219-354.

[5] ELUMALAI V K. A proximal policy optimization based deep reinforcement learning framework for tracking control of a flexible robotic manipulator[J]. Results in Engineering, 2025, 25: 104178.

[6] TANG C, ABBATEMATTEO B, HU J, et al. Deep Reinforcement Learning for Robotics: A Survey of Real-World Successes[J]. Proceedings of the AAAI Con- ference on Artificial Intelligence, 2025, 39(27): 28694-28698.

[7] JIN Y-L, JI Z-Y, ZENG D, et al. VWP: An efficient DRL-based autonomous driving model[J]. IEEE Tran- sactions on Multimedia, 2022, 26: 2096-2108.

[8] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: A survey[J]. IEEE transactions on intelligent transporta- tion systems, 2021, 23(6): 4909-4926.

[9] 冯斌, 胡轶婕, 黄刚, 等. 基于深度强化学习的新型电力系统调度优化方法综述[J]. 电力系统自动化, 2023, 47(17): 187-199.

FENG B, HU Y J, HUANG G, et al. Review on Optimization Methods for New Power System Dispatch Based on Deep Reinforcement Learning[J]. Automation of Electric Power Systems, 2023, 47(17): 187-199.

[10] 张延宇, 饶新朋, 周书奎, 等. 基于深度强化学习的电动汽车充电调度算法研究进展[J]. 电力系统保护与控制, 2022, 50(16): 179-187.
ZHANG Y Y, RAO X P, ZHOU S K, et al. Research progress of electric vehicle charging scheduling algorithms based on deep reinforcement learning[J]. Power System Protection and Control, 2022, 50(16): 179-187.

[11] SHAO K, TANG Z, ZHU Y, et al. A survey of deep reinforcement learning in video games[J]. arXiv preprint arXiv:1912.10944, 2019.

[12] ZHANG C, HE Q, YUAN Z, et al. Advancing DRL agents in commercial fighting games: Training, integration, and agent-human alignment[J]. arXiv preprint arXiv:2406.01103, 2024.

[13] IRPAN A. Deep reinforcement learning doesn't work yet[J]. Internet: https://www. alexirpan. com/2018/02/14/rl-hard. html, 2018.

[14] 丁世飞, 杜威, 张健, 等. 多智能体深度强化学习研究进展[J]. 计算机学报, 2024, 47(7): 1547-1567.
DING S F, DU W, ZHANG J, et al. Research Progress of Multi-Agent Deep Reinforcement Learning[J]. Chinese Journal of Computers, 2024, 47(7): 1547-1567.

[15] HERNANDEZ-LEAL P, KARTAL B, TAYLOR M E. A survey and critique of multiagent deep reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems, 2019, 33(6): 750-797.

[16] OLIEHOEK F A, AMATO C. A Concise Introduction to Decentralized POMDPs[J]. Springer Publishing Company, Incorporated, 2016.

[17] Zhu C, Dastani M, Wang S. A survey of multi-agent deep reinforcement learning with communication[J]. Autonomous Agents and Multi-Agent Systems, 2024, 38(1): 4.

[18] 赵立阳,常天庆, 褚凯轩, 等. 完全合作类多智能体深度强化学习综述[J]. 计算机工程与应用, 2023, 59(12): 14-27.
ZHAO L Y, CHANG T Q, CHU K X, et al. Survey of Fully Cooperative Multi-Agent Deep Reinforcement Learning[J]. Computer Engineering and Applications, 2023, 59(12): 14-27.

[19] 梁星星, 冯旸赫, 马扬,等. 多 Agent 深度强化学习综述[J]. 自动化学报, 2020, 46(12): 2537−2557.
LIANG X X, FENG Y H, MA Y, et al. Deep Multi-Agent Reinforcement Learning: A Survey[J]. Acta Automatica Sinica, 2020, 46(12): 2537−2557.

[20] 李明阳, 许可儿,宋志强, 等. 多智能体强化学习算法研究综述[J]. 计算机科学与探索, 2024, 18(8): 1979-1997.
LI Y M, XU K E, SONG Z Q, et al. Review of Research on Multi-agent Reinforcement Learning Algorithms[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(8): 1979-1997.

[21] 王涵,俞扬,姜远.基于通信的多智能体强化学习进展综述[J].中国科学：信息科学,2022,52(5):742-764
WANG H , YU Y, JIANG Y. Review of the progress of communication-based multi-agent reinforcement learning[J]. SCIENTIA SINICA Informationis, 2022, 52(5):742-764

[22] BEIKMOHAMMADI A. Learning to Communicate through Multi-Agent Reinforcement Learning (MARL): A Systematic Literature Review[J], Preprints, 2024.

[23] ARDI T, TAMBET M, DORIAN K, et al. Multiagent Cooperation and Competition with Deep Reinforcement Learning[J]. Plos One, 2017, 12(4): e0172395.

[24] DE WITT C S, GUPTA T, MAKOVIICHUK D, et al. Is independent learning all you need in the starcraft multi-agent challenge?[J]. arXiv preprint arXiv:2011.09533, 2020.

[25] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value- decomposition networks for cooperative multi-agent learning[J]. arXiv preprint arXiv:1706.05296, 2017.

[26] RASHID T, SAMVELYAN M, DE WITT C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(178): 1-51.

[27] SON K, KIM D, KANG W J, et al. Qtran: Learning to factorize with transformation for cooperative multi- agent reinforcement learning[C]. International con- ference on machine learning, 2019: 5887-5896.

[28] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.

[29] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[C]// Proceedings of the AAAI conference on artificial intelligence, 2018.

[30] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of ppo in cooperative multi-agent games[J]. Advances in neural information processing systems, 2022, 35: 24611-24624.

[31] LAZARIDOU A, BARONI M. Emergent multi-agent communication in the deep learning era[J]. arXiv preprint arXiv:2006.02419, 2020.

[32] CAO K, LAZARIDOU A, LANCTOT M, et al. Emergent communication through negotiation[J]. arXiv preprint arXiv:1804.03980, 2018.

[33] LOWE R, FOERSTER J, BOUREAU Y-L, et al. On the pitfalls of measuring emergent communication[J]. arXiv preprint arXiv:1903.05168, 2019.

[34] BULLARD K, KIELA D, MEIER F, et al. Quasi- equivalence discovery for zero-shot emergent commu- nication[J]. arXiv preprint arXiv:2103.08067, 2021.

[35] NOUKHOVITCH M, LACROIX T, LAZARIDOU A, et

al. Emergent communication under competition[J]. arXiv preprint arXiv:2101.10276, 2021.
[36] FOERSTER J, ASSAEL I A, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning[J]. Advances in neural infor- mation processing systems, 2016, 29.
[37] SUKHBAATAR S, FERGUS R. Learning multiagent communication with backpropagation[J]. Advances in neural information processing systems, 2016, 29.
[38] PENG P, WEN Y, YANG Y, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games[J]. arXiv preprint arXiv:1703.10069, 2017.
[39] JIANG J, LU Z. Learning attentional communication for multi-agent cooperation[J]. Advances in neural information processing systems, 2018, 31.
[40] JIANG J, DUN C, HUANG T, et al. Graph convolutional reinforcement learning[J]. arXiv preprint arXiv: 1810.09202, 2018.
[41] DAS A, GERVET T, ROMOFF J, et al. Tarmac: Targeted multi-agent communication[C]. International Conference on machine learning, 2019: 1538-1546.
[42] PESCE E, MONTANA G. Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication[J]. Machine Learning, 2020, 109(9): 1727-1747.
[43] KIM D, MOON S, HOSTALLERO D, et al. Learning to schedule communication in multi-agent reinforcement learning[J]. arXiv preprint arXiv:1902.01554, 2019.
[44] WANG R, HE X, YU R, et al. Learning efficient multi-agent communication: An information bottleneck approach[C]. International conference on machine learning, 2020: 9908-9918.
[45] SHENG J, WANG X, JIN B, et al. Learning structured communication for multi-agent reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems, 2022, 36(2): 50.
[46] FREED B, SARTORETTI G, HU J, et al. Communication learning via backpropagation in discrete channels with unknown noise[C]. Proceedings of the AAAI conference on artificial intelligence, 2020: 7160-7168.
[47] KIM W, PARK J, SUNG Y. Communication in multi-agent reinforcement learning: Intention sharing[C]. International conference on learning representations, 2020.
[48] GUPTA N, SRINIVASARAGHAVAN G, MOHALIK S, et al. Hammer: Multi-level coordination of reinforcement learning agents via learned messaging[J]. Neural Computing and Applications, 2023: 1-16.
[49] YUN W J, LIM B, JUNG S, et al. Attention-based reinforcement learning for real-time UAV semantic communication[C]. 2021 17th International Symposium on Wireless Communication Systems (ISWCS), 2021: 1-6.
[50] YUAN L, WANG J, ZHANG F, et al. Multi-agent incentive communication via decentralized teammate modeling[C]. Proceedings of the AAAI conference on artificial intelligence, 2022: 9466-9474.
[51] 臧嵘, 王莉, 史腾飞. 基于注意力消息共享的多智能体强化学习[J]. 计算机应用, 2022, 42(11): 3346-3353.
ZANG R, WANG L, SHI T F. Multi-agent reinforcement learning based on attentional message sharing[J]. Journal of Computer Applications, 2022, 42(11): 3346-3353.
[52] 马廷淮, 彭可兴, 周宏豪, 等. 具有实时注意力的多智能体强化学习通信模型[J]. 计算机仿真, 2023, 000(8): 445-450.
MA T H, PENG K X, ZHOU H H, et al. A Com- munication Model for Multi-Agent Reinforcement Learning with Time-Variant Attention[J]. Computer Simulation, 2023, 000(8): 445-450.
[53] ZHAO-RONG H, YU-HUA Q, GUO-QING L. Multi Agent Communication Based on Self Attention and Reinforcement Learning[J]. Journal of Chinese Com- puter Systems, 2023, 44(6).
[54] MENG X, TAN Y. Pmac: Personalized multi-agent communication[C]. Proceedings of the AAAI Con- ference on Artificial Intelligence, 2024: 17505-17513.
[55] DING S, DU W, DING L, et al. Learning efficient and robust multi-agent communication via graph information bottleneck[C]. Proceedings of the AAAI Con- ference on Artificial Intelligence, 2024: 17346-17353.
[56] YUAN L, JIANG T, LI L, et al. Robust cooperative multi-agent reinforcement learning via multi-view message certification[J]. Science China Information Sciences, 2024, 67(4): 142102.
[57] XU J, WEI W, ZHANG Y, et al. Partial Communication Model based on the Gain of Q-value in Multi-agent Reinforcement Learning[C]. 2024 14th Asian Control Conference (ASCC), 2024: 69-74.
[58] LI X, WANG X, BAI C, et al. Exponential Topology- enabled Scalable Communication in Multi-agent Re- inforcement Learning[J]. arXiv preprint arXiv:2502. 19717, 2025.
[59] LI X, XUE S, HE Z, et al. TMAC: a Transformer-based partially observable multi-agent communication method[J]. PeerJ Computer Science, 2025, 11: e2758.
[60] SONG S, LIN Y, HAN S, et al. CoDe: Communication Delay-Tolerant Multi-Agent Collaboration via Dual Alignment of Intent and Timeliness[J]. arXiv preprint arXiv:2501.05207, 2025.
[61] KILINC O, MONTANA G. Multi-agent deep reinforcement learning with extremely noisy observations[J]. arXiv preprint arXiv:1812.00922, 2018.
[62] SINGH A, JAIN T, SUKHBAATAR S. Learning when to communicate at scale in multiagent cooperative and competitive tasks[J]. arXiv preprint arXiv:1812.09755, 2018.
[63] WANG T, WANG J, ZHENG C, et al. Learning nearly decomposable value functions via communication minimization[J]. arXiv preprint arXiv:1910.05366, 2019.
[64] LIU Y, WANG W, HU Y, et al. Multi-agent game abstraction via graph attention neural network[C]. Proceedings of the AAAI conference on artificial intelligence, 2020:

7211-7218.
[65] DING Z, HUANG T, LU Z. Learning individually inferred communication for multi-agent cooperation[J]. Advances in neural information processing systems, 2020, 33: 22069-22079.
[66] ZHANG S Q, ZHANG Q, LIN J. Succinct and robust multi-agent communication with temporal message control[J]. Advances in neural information processing systems, 2020, 33: 17271-17282.
[67] JIANG R, ZHANG X, LIU Y, et al. Multi-agent cooperative strategy with explicit teammate modeling and targeted informative communication[J]. Neuro- computing, 2024, 586: 127638.
[68] ZHANG Z, CHENG B, WANG Z, et al. PAGNet: Pluggable Adaptive Generative Networks for Information Completion in Multi-Agent Communi- cation[J]. arXiv preprint arXiv:2502.03845, 2025.
[69] LI D, LOU N, XU Z, et al. Efficient Communication in Multi-Agent Reinforcement Learning with Implicit Consensus Generation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 23240- 23248.
[70] ZHANG Z, HE B, CHENG B, et al. Bridging training and execution via dynamic directed graph-based communication in cooperative multi-agent systems[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 23395-23403.
[71] DOLAN S, NAYAK S, ALOOR J J, et al. Asynchronous Cooperative Multi-Agent Reinforcement Learning with Limited Communication[J]. arXiv preprint arXiv:2502. 00558, 2025.
[72] ZHANG S Q, ZHANG Q, LIN J. Efficient com- munication in multi-agent reinforcement learning via variance based control[J]. Advances in neural infor- mation processing systems, 2019, 32.
[73] HU G, ZHU Y, ZHAO D, et al. Event-triggered communication network with limited-bandwidth con- straint for multi-agent reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(8): 3966-3978.
[74] HAN S,DASTANI M,WANG S.Model-based sparse communication in multi-agent reinforcement learning[C]. Seventeenth European Workshop on Reinforcement Learning, 2023.
[75] 夏庆锋, 许可儿, 李明阳, 等. 强化学习中的注意力机制研究综述[J]. 计算机科学与探索, 2024, 18(6): 1457-1475.
XIA Q F, XU K E, LI Y M, et al. A Review of Attention Mechanisms in Reinforcement Learning[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(6): 1457-1475.
[76] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.
[77] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
[78] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[79] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1): 4-24.
[80] CORSO G, STARK H, JEGELKA S, et al. Graph neural networks[J]. Nature Reviews Methods Primers, 2024, 4(1): 17.
[81] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.
[82] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
[83] LIU Z, ZHANG Y, LI P, et al. A dynamic llm-powered agent network for task-oriented agent collaboration[C]. First Conference on Language Modeling, 2024.
[84] MOLLER M.Efficient training of feed-forward neural networks, Neural Network Analysis[M].Florida: CRC Press, 2024: 136–173.
[85] BEBIS G, GEORGIOPOULOS M. Feed-forward neural networks[J]. Ieee Potentials, 1994, 13(4): 27-31.
[86] SAZLı M H. A brief review of feed-forward neural networks[J]. Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering, 2006, 50(01).
[87] CHO K, VAN MERRIëNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder- decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
[88] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[89] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
[90] SHERSTINSKY A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.
[91] JANG E, GU S, POOLE B. Categorical reparamete- rization with gumbel-softmax[J]. arXiv preprint arXiv:1611.01144, 2016.
[92] MOERLAND T M, BROEKENS J, PLAAT A, et al. Model-based reinforcement learning: A survey[J]. Foundations and Trends® in Machine Learning, 2023, 16(1): 1-118.
[93] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
[94] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
[95] RADFORD A, KIM J W, HALLACY C, et al. Learning

transferable visual models from natural language supervision[C]. International conference on machine learning, 2021: 8748-8763.
[96] ALAYRAC J-B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning[J]. Advances in neural information processing systems, 2022, 35: 23716-23736.
[97] GUNNING D, STEFIK M, CHOI J, et al. XAI—Explainable artificial intelligence[J]. Science robotics, 2019, 4(37): eaay7120.
[98] KASNECI E, SEßLER K, KüCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.
[99] CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models[J]. ACM transac- tions on intelligent systems and technology, 2024, 15(3): 1-45.
[100] LIU X, ZHANG F, HOU Z, et al. Self-supervised learning: Generative or contrastive[J]. IEEE transactions on knowledge and data engineering, 2021, 35(1): 857-876.
[101] LIN X, BAWEJA H, KANTOR G, et al. Adaptive auxiliary task weighting for reinforcement learning[J]. Advances in neural information processing systems, 2019, 32.
[102] KAZEROUNI A, HEYDARIAN A, SOLTANY M, et al. An intelligent modular real-time vision-based system for environment perception[J]. arXiv preprint arXiv:2303.16710, 2023.
[103] FAN J, ZHANG K, HUANG Y, et al. Parallel spatio- temporal attention-based TCN for multivariate time series prediction[J]. Neural Computing and Applications, 2023, 35(18): 13109-13118.
[104] ELSAYED S, THYSSENS D, RASHED A, et al. Do we really need deep learning models for time series forecasting?[J]. arXiv preprint arXiv:2101.02118, 2021.
[105] REN P, XIAO Y, CHANG X, et al. A comprehensive survey of neural architecture search: Challenges and solutions[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-34.

陈荣敏（2000—），男，浙江温州人，硕士研究生，CCF 学生会员，主要研究方向为深度强化学习、多智能体通信等。

**CHEN Rongmin**, born 2000, M.S. candidate, CCF student member. His research interests included deep reinforcement learning, Multi agent communication, etc.

郭大波（1963—），男，山西阳泉人，博士，教授，CCF 会员，主要研究方向为多智能体强化学习、自动驾驶等。

**GUO Dabo**, born in 1963, Ph.D., professor, CCF member. His research interests include Multi agent reinforcement learning, autonomous driving, etc.

吴宏坤（2001—），男，广东韶关人，硕士研究生，CCF 学生会员，主要研究方向为自动驾驶、深度强化学习等。

**WU Hongkun**, born 2001, M.S. candidate, CCF student member. His research interests included autonomous driving, deep reinforcement learning ,etc.

李程翔（2002—），男，浙江温州人，硕士研究生，主要研究方向为自动驾驶、深度强化学习等。

**LI Chengxiang**, born 2002, M.S. candidate. His research interests included autonomous driving, deep reinforcement learning ,etc.

胡海霄（1999—），男，内蒙古赤峰人，硕士研究生，CCF 学生会员，主要研究方向为自动驾驶、深度强化学习等。

**HU Haixiao**, born 1999, M.S. candidate. His research interests included autonomous driving, deep reinforcement learning ,etc.

刘李祥（1999—），男，四川成都人，硕士研究生，主要研究方向为自动驾驶、深度强化学习等。

**LIU Lixiang**, born 1999, M.S. candidate. His research interests included autonomous driving, deep reinforcement learning ,etc.