计算机科学
COMPUTER SCIENCE

基于值函数分解的多智能体深度强化学习方法研究综述

高玉钊, 聂一鸣

引用本文

高玉钊, 聂一鸣. 基于值函数分解的多智能体深度强化学习方法研究综述[J]. 计算机科学, 2024, 51(6A): 230300170-9.

GAO Yuzhao, NIE Yiming. Survey of Multi-agent Deep Reinforcement Learning Based on Value Function Factorization [J]. Computer Science, 2024, 51(6A): 230300170-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300170

基于值函数分解的多智能体深度强化学习方法研究综述

高玉钊　聂一鸣
军事科学院国防科技创新研究院　北京 100071
(958255669@qq.com)

摘　要　多智能体深度强化学习方法是深度强化学习方法在多智能体问题上的扩展,其中基于值函数分解的多智能体深度强化学习方法取得了较好的表现效果,是目前研究和应用的热点。文中介绍了基于值函数分解的多智能体深度强化学习方法的主要原理和框架;根据近期相关研究,总结出了提高混合网络拟合能力问题、提高收敛效果问题和提高算法可扩展性问题3个研究热点,从算法约束、环境复杂度、神经网络限制等方面分析了3个热点问题产生的原因;根据拟解决的问题和使用的方法对现有研究进行了分类梳理,总结了同类方法的共同点,分析了不同方法的优缺点;对基于值函数分解的多智能体深度强化学习方法在网络节点控制、无人编队控制两个热点领域的应用进行了阐述。

关键词:多智能体深度强化学习;值函数分解;拟合能力;收敛效果;可扩展性
中图分类号　TP181

Survey of Multi-agent Deep Reinforcement Learning Based on Value Function Factorization

GAO Yuzhao and NIE Yiming
National Innovation Institute of Defense Technology, Academic of Military Science, Beijing 100071, China

Abstract　The multi-agent deep reinforcement learning is an extension of the deep reinforcement learning method to the multi-agents problem, in which the multi-agents deep reinforcement learning based on the value function factorization has achieved better performance and is a hotspot for research and application at present. This paper introduces the main principles and framework of the multi-agents deep reinforcement learning based on the value function factorization. Based on the recent related research, three research hotspots are summarized: the problem of improving the fitting ability of mixing network, the problem of improving the convergence effect and the problem of improving the scalability of algorithms, and the reasons for the three hotspot problems are analyzed in terms of algorithm constraints, environmental complexity and neural network limitations. The existing research is classified according to the problems to be solved and the methods to be used, the common points of similar methods are summarized, and the advantages and disadvantages of different methods are analyzed; the application of multi-agent deep reinforcement learning method based on value function decomposition in two hot fields of network node control and unmanned formation control is expounded.

Keywords　Multi-agent deep reinforcement learning, Value function factorization, Fitting ability, Convergence effect, Scalability

1　引言

强化学习目前已经在电子竞技、工业控制等多个领域产生了广泛的应用,其在处理复杂问题、实时决策等方面有很大的优势。在实际的生产生活中,很多场景包含多个智能体,且需要其在分布式执行情况下合作完成任务,使用多智能体强化学习方法解决该类问题逐渐成为一个研究热点。现有的多智能体强化学习方法根据训练框架主要可分为独立训练独立执行方法和集中训练分布执行方法,如图1所示。

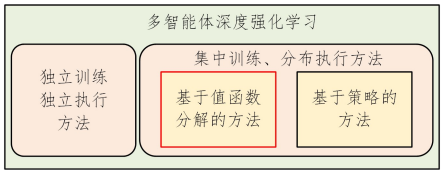

图1　多智能体深度强化学习方法分类

Fig. 1　Classification of multi-agent deep reinforcement learning

独立训练独立执行方法是使用单智能体强化学习方法处理多智能体问题[1]。每个智能体独立训练相当于把其他智能体看作环境的一部分。但实际上,其他智能体的策略会随着训练改变,相当于环境在改变,即环境是非平稳的,因此可能导致方法效果较差。

集中训练分布执行方法训练时,所有智能体共享同一个评价网络,根据评价网络的输出更新各自策略网络的参数[2]。执行时,智能体根据状态和自身的策略网络独立进行决策。这种方法可解决环境非平稳问题。然而由于使用全局奖励进行训练,每个智能体无法知道自己的动作对整体的贡献程度,即存在信用分配问题[3]。一些研究人员提出了基于值函数分解的多智能体深度强化学习方法解决此问题。该类方法假设全局的联合动作值与每个智能体动作值存在映射关系,通过人为设计或使用神经网络学习得到该函数。训练时使用全局奖励和联合动作值计算损失,通过梯度反向传播达到奖励贡献分配的效果。因此这类方法得到越来越多的关注和研究。

通信作者:聂一鸣(13370154812@189.cn)

许多研究人员对多智能体强化学习相关研究进行了综述和分析。Wong 等[4]总结了多智能体问题的 4 个主要挑战：计算复杂度、非平稳性、部分可观测和信用分配；从集中训练和分散执行、对手建模、沟通、高效协调和奖励塑造等领域介绍了现有多智能体深度强化学习方法的应对方案。Hao 等[5]将多智能体深度强化学习算法中探索方法困难的原因总结为状态维度指数增长、需要协同探索和全局与局部探索的平衡。从不确定导向和内在动机导向两个方面介绍了现有多智能体深度强化学习算法的探索方法。Du 等[6]对多智能体强化学习方法的研究和应用进行了综述，从算法的可扩展性和智能体意图等方面介绍了算法的改进和创新。Sun 等[7]对多智能体深度强化学习经典算法进行了分类阐述，并对算法的实际应用和现有测试平台进行了简要介绍。Yan 等[8]针对提高多智能体深度强化学习方法的可扩展性和可迁移性问题，对研究进行了分类梳理。Xiong 等[9]根据值函数分解的方式对基于值函数分解的多智能体深度强化学习经典算法进行了分类介绍。Li 等[10]介绍了多智能体深度强化学习在车载网络、物联网等新兴互联网领域的应用，重点对网络接入、计算迁移、分组路由等问题进行了分析。

本文主要关注基于值函数分解的多智能体深度强化学习方法，文章总体结构如图 2 所示。主要贡献如下：

(1)根据近期基于值函数分解的多智能体深度强化学习方法相关研究拟解决的问题，总结梳理了该类方法的热点研究问题，从方法原理、网络结构等方面分析了问题产生的原因。

(2)根据拟解决的热点研究问题和解决问题的方法对现有研究进行了分类介绍。

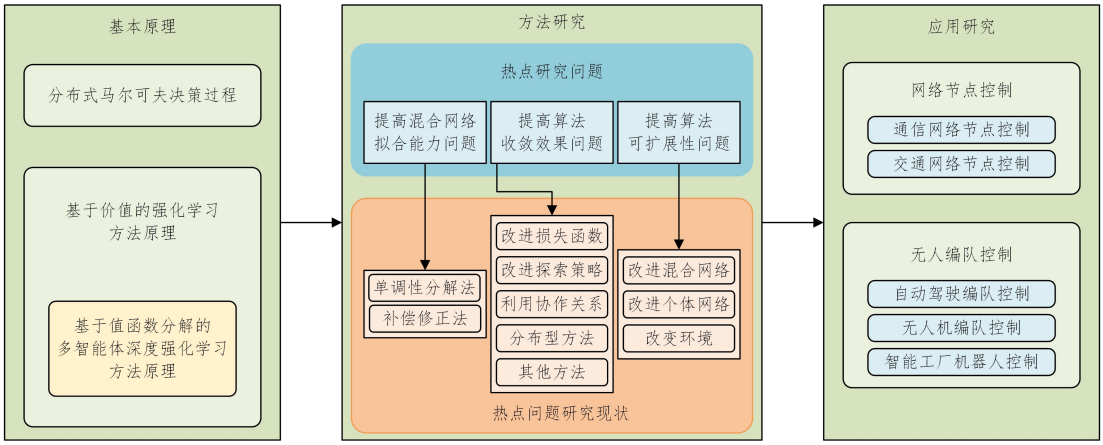

图 2 文章内容总体结构

Fig. 2 Overall structure of the main contents in this paper

2 基于值函数分解的多智能体深度强化学习方法相关理论

2.1 问题形式化

多智能体合作场景可使用分布式部分可观测马尔可夫过程描述。其由一个元组 $G=\langle S,U,P,r,Z,O,N,\gamma\rangle$ 组成。其中，$s\in S$ 代表环境状态；N 为环境中智能体的集合；$\boldsymbol{u}\equiv[u_i]_{i=1}^{n}\in U$ 代表智能体的联合动作；$P(s'|s,\boldsymbol{u}):S\times U\times S\to[0,1]$ 代表在 s 状态下，执行联合动作 $\boldsymbol{u}$ 后，状态变为 s' 的转移概率；$r(s,u):S\times U\to R$ 代表奖励函数，所有智能体共享该奖励；$\gamma\in[0,1)$ 是奖励折扣因子；$z\in Z$ 是每个智能体的局部观测值；$O(s,i):S\times N\to Z$ 为观测函数。为了弥补局部观测的不足，很多方法根据动作观测历史 $\tau^i\in T\equiv(Z\times U)^*$ 决策。

2.2 基于价值的强化学习方法原理

强化学习方法通常可以分为两类：基于值的方法和基于策略的方法。在强化学习中，智能体的目的是学得可获得最大环境奖励的策略。基于值的算法使智能体能够学习最优状态-动作值函数估计，也称为最优 Q 函数。最优 Q 函数如式(1)所示：

$$Q_\pi(s,u)=E_{s_{t+1}\sim\rho,u_{t>0}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,u_t,s_{t+1})\,|\,s_0=s,u_0=u\right] \tag{1}$$

通过采用所学习的最优 Q 函数的贪婪动作，可以间接地获得最优策略。当问题较为简单且状态动作为离散时，可使用表格表示 Q 函数。当状态为连续或复杂时，则使用神经网络来拟合 Q 函数。网络的输入为状态，输出为每个动作的评价值，经典方法有 DQN[11]和 DDQN[12]。在训练时，一般采用经验回放的方式，通过最小化 TD 误差，使网络逐渐逼近最优 Q 函数。损失函数如式(2)所示：

$$L(\theta)=\sum_{i=1}^{b}\left[(y_i^{\text{target}}-Q(s,u;\theta))^2\right] \tag{2}$$

其中，$y_i^{\text{target}}=r+\gamma\max_{u'}Q(s',\boldsymbol{u}';\theta^-)$，$\theta^-$ 为目标网络的参数，目标网络周期性与智能体网络同步。

2.3 基于值函数分解的多智能体深度强化学习方法原理

基于值函数分解的多智能体深度强化学习方法是基于价值的强化学习方法在多智能体合作问题中的扩展，目的是求得每个智能体的最优状态动作值函数。其使用集中训练分步执行的架构，如图 3 所示。

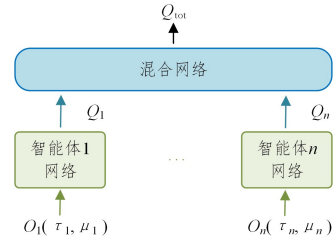

图 3 基于值函数分解的多智能体深度强化学习方法的基本框架

Fig. 3 Basic framework of multi-agent deep reinforcement learning based on value function factorisation

其将全局的联合动作值函数分解为个体动作值函数，如式(3)所示。F 称为值分解函数，可直接人为设置，也可通过

神经网络学习得到。该网络可称为混合网络。

$$Q_{\text{tot}}=F(Q_1,Q_1,\cdots,Q_n) \tag{3}$$

训练时其损失函数使用全局奖励和联合动作值计算，如式(4)所示。通过梯度的反向传播，每个智能体可根据全局奖励优化自身的网络，减轻信用分配问题。

$$L(\theta)=\sum_{i=1}^{b}[(y_i^{\text{target}}-Q_{\text{tot}}(s,\boldsymbol{u};\theta))^2] \tag{4}$$

其中，Q_{tot}为所有智能体的联合动作值，可由每个智能体的个体动作值计算得到。$y_i^{\text{target}}=r+\gamma\max_{\boldsymbol{u}'}Q_{\text{tot}}(s',\boldsymbol{u}';\theta^-)$，$\boldsymbol{u}'$为$s'$状态下的最优联合动作，由每个智能体的最优动作组成，可根据每个智能体的决策网络得到。这样选择最优联合动作将联合动作空间与智能体数量的关系转换为线性关系，解决了维度爆炸问题。

为保证上述两个优点，基于值函数分解的多智能体深度强化学习方法一般需满足两个条件。

(1)分解方式可进行梯度的反向传播，否则无法进行训练。

(2)个体的最优动作组合与最优联合动作一致，即满足IGM条件[13]，如式(5)所示。

$$\arg\max_{\boldsymbol{u}}Q_{\text{tot}}(\tau,u)=\begin{pmatrix}\arg\max_{u^1}Q_1(\tau^1,u^1)\\ \vdots\\ \arg\max_{u^n}Q_n(\tau^n,u^n)\end{pmatrix} \tag{5}$$

3 基于值函数分解的多智能体深度强化学习方法研究热点及现状

3.1 热点研究问题

分析VDN，QMIX，QTRAN和UNMAS等基于值函数分解的多智能体深度强化学习方法相关研究，可总结出以下3个研究热点。

(1)提高混合网络拟合能力问题。不同任务场景客观存在不同的最优值分解函数，一般使用混合网络通过训练拟合该函数。若混合网络的拟合能力较差，则无法得到准确的值分解函数，进而影响联合动作值的计算和算法的整体效果。由2.3节分析可知，值分解函数需满足IGM条件，因此需对混合网络施加约束，如单调性约束，这限制了混合网络的拟合能力，导致其在一些场景中效果较差[13]。在满足IGM条件的情况下，提高混合网络的拟合能力，使其适应更多的场景，是该领域的一大研究热点。

(2)提高算法收敛效果问题。该问题是强化学习方法普遍存在的问题。多智能体环境更为复杂。多智能体环境多为部分可观测环境[14]，智能体根据局部观测数据计算个体动作值，该值为基于全局状态的个体动作值的近似，这种近似导致算法更难收敛。多智能体环境联合动作空间大[15]、环境奖励稀疏[16]，智能体在探索过程中更难找到最优动作，使算法收敛到最优策略更为困难。许多研究从不同角度改进基于值函数分解的多智能体深度强化学习方法，以提高算法的收敛效果。

(3)提高算法可扩展性问题。可扩展性差主要是指当环境中智能体的规模改变后，算法的表现效果变差或无法使用。原因主要有下两点。

①目前大多数方法中，智能体的数量变化会导致观测数据维度的变化。智能体的网络一般为全连接神经网络，无法处理维度变化的观测数据[17]。当环境中智能体数量发生变化时，这类方法一般通过对数据补零使其维度固定。但这会导致方法可处理的最大观测数据维度有上限，限制了其可扩展性。

②在基于值函数分解的多智能体深度强化学习框架中，联合动作值根据每个智能体的动作值计算得到。智能体数量变化会导致个体动作值数量变化。当环境中智能体数量增多时，比较直接的改进方法是预先设置最大输入维度，通过补零来补齐数据，使得混合网络的输入维度不变。对于智能体减少的情况，比如被摧毁等，现有方法一般忽视该智能体的死亡，仍正常计算该智能体的个体动作值，以保证混合网络的输入维度不变，这种状态可以称为占用状态[18]。这种处理方式使现有算法能够在不改变任何网络结构的情况下在智能体数量减少的非定型场景中进行训练，简化了环境和实现方式。但这种处理方式可能导致联合动作值计算不准确；当混合网络使用神经网络作为函数逼近器时，将已经死亡的智能体个体动作值作为输入的一部分，使目标函数即值分解函数更难以逼近；同时，这种方式也浪费了计算资源；且因为需要预先设置最大输入维度，限制了方法的扩展能力。

提高算法的可扩展性是目前的一个研究热点问题。

3.2 热点问题研究现状

许多学者对基于值函数分解的多智能体深度强化学习方法进行了研究。本文对近期的相关研究进行了调研，主要根据不同研究拟解决的问题和使用的方法进行分类梳理，分类情况如图4所示。

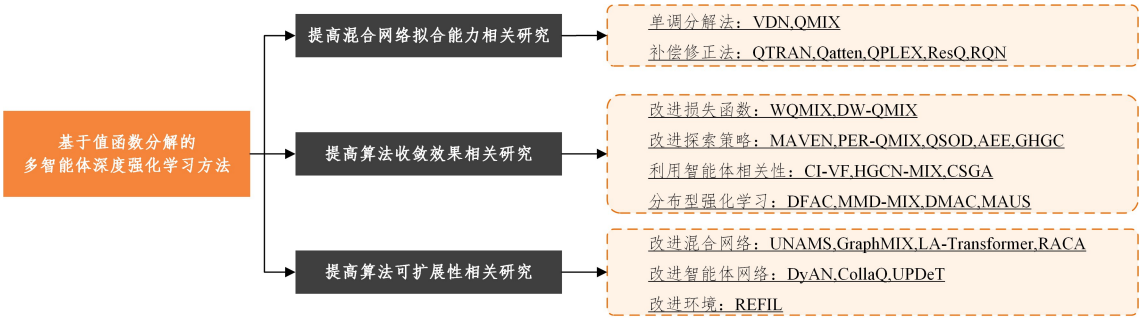

图4 基于值函数分解的多智能体深度强化学习方法的分类

Fig. 4 Classification of multi-agent deep reinforcement learning based on value function factorisation

3.2.1 提高混合网络拟合能力的相关研究

为保证个体的最优动作组合与最优联合动作一致，分解函数需满足IGM条件。许多研究中的分解函数满足该条件的充分不必要条件，但这样的约束过强，限制了分解函数的

拟合能力。后续研究主要通过不断放松分解函数的约束，使其逼近 IGM 条件的充分必要条件，来提高分解函数的拟合能力。根据分解函数的形式，主要可分为单调性分解法和补偿修正法两大类。

早期的研究主要使用单调性分解法。当联合动作值 Q_{tot} 与个体动作值 Q_i 为单调递增关系，即满足式(6)时，个体的最优动作组合为联合最优动作。

$$\frac{\partial Q_{tot}}{\partial Q_i}\geqslant 0,\forall i=1,\cdots,n \tag{6}$$

Peter 等[14]提出了 VDN 方法。其直接假设联合动作值是由每个智能体的动作值加和得到的，该分解方法满足式(4)的要求。这样的分解方法简洁明了，有易于扩展的优势，但满足的约束太严格，导致值分解函数的拟合能力较弱。Tabish 等[19]提出了 QMIX 方法，其使用混合网络拟合联合动作值函数，结构如图 3 所示。混合网络的参数 $W1$ 和 $W2$ 是由超参数网络根据全局状态 St 计算得到的，且经过绝对值激活函数处理，保证了参数的非负性，满足式(4)要求。该方法不仅利用了全局状态信息，也增加了值函数的拟合能力，在大部分情况下能取得较好的效果，但在非单调奖励环境下效果较差。

VDN 和 QMIX 方法满足的条件为 IGM 条件的充分非必要条件。一些研究人员对 IGM 条件进行了理论分析，提出了与 IGM 条件更一致的约束条件，并根据所提约束条件对值函数进行分解，提高其拟合能力。这类方法值分解函数表达式的主要形式如式(7)所示，通过加权或增加补偿函数，提高联合动作值函数的拟合能力。

$$Q_{tot}=\sum AQ_i+B \tag{7}$$

Kyunghwan 等[13]提出了一种仿射变换下 IGM 的充分必要条件，并根据该条件提出了 QTRAN 方法。在 VDN 分解方法的基础上，作者建立了联合动作值函数到个体值函数的映射。分解表达式如表 1 所列。其在 VDN 分解方法的基础上使用 $V_{tot}(\boldsymbol{\tau})$进行了修正，因此最终的混合网络拟合能力更强。Yang 等[20]提出了 Qatten 方法，对联合动作值与局部动作值的分解关系进行了理论分析，其分解方式与多头注意力计算公式类似，因此该方法使用基于多头注意力机制的混合网络来分解联合动作值。其通过引入非线性和全局信息，提升了值函数的拟合能力。同时该分解方式满足单调性要求。Wang 等[21]提出了 QPLEX 方法，目的是进一步增强分解方式与 IGM 条件的一致性。由 Dueling DQN[22]方法可知，Q 值可分解为 $Q=V+A$，状态值 V 与动作无关。因此，IGM 条件可转化为 Advantage-based IGM 条件。QPLEX 方法分解表达式如表 1 所列。其中，$\lambda_i(\boldsymbol{\tau},\boldsymbol{\mu})$为使用注意力机制计算的权重。该分解方式满足 Advantage-based IGM 条件，且也可看作在 VDN 分解方式上增加了补偿，使其拟合能力更强。Shen 等[23]为了增强其对非单调奖励的拟合能力，提出了 ResQ 方法。按照将奖励矩阵分解为单调矩阵加非单调矩阵的思想，将联合动作值函数分解为单调联合动作值函数和补偿值函数的加权和，分解表达式如表 1 所列。单调联合动作值函数 $Q_{tot}(\boldsymbol{\tau},\boldsymbol{\mu})$与 QMIX 分解方法相同。$Q_r(\boldsymbol{\tau},\boldsymbol{\mu})$为补偿函数，由观测动作历史、个体动作值和个体动作计算得到，经负绝对值函数处理，始终为负值，可不满足单调性约束。$\omega_r(\boldsymbol{\tau},\boldsymbol{\mu})$为掩码函数，根据联合动作确定是否进行修正。因此该分解方式同样满足 IGM 条件，且函数拟合能力优于 VDN 和 QMIX 等单调性分解方法。Pina 等[24]提出了一种残差动作值网络(RQN)计算个体动作值的因子，网络输入为每批数据中的平均个体动作值和最大个体动作值，输出因子与个体动作值加和计算联合动作值。增加该因子，引入了整体状态信息，可提高对较好状态下动作的探索和值分解函数的拟合能力。

表 1 补偿修正类方法及其值分解函数表达式

Table 1 Compensation correction class methods and their value decomposition function expressions

方法	值分解函数表达式
QTRAN	$\max\limits_{\boldsymbol{\mu}} Q_{tot}(\boldsymbol{\tau},\boldsymbol{\mu})=\sum\limits_{i=1}^{N}Q_i(\tau_i,\mu_i)+V_{tot}(\boldsymbol{\tau})$
Qatten	$Q_{tot}(\boldsymbol{s},\boldsymbol{\mu})=\sum\limits_{h}^{H}\sum\limits_{n}^{N}\lambda_{i,h}(s)Q_i(\tau_i,\mu_i)+c(s)$
QPLEX	$Q_{tot}(\boldsymbol{\tau},\boldsymbol{\mu})=\sum\limits_{i=1}^{n}Q_i(\boldsymbol{\tau},\mu_i)+\sum\limits_{i=1}^{n}(\lambda_i(\boldsymbol{\tau},\boldsymbol{\mu})-1)A_i(\boldsymbol{\tau},\mu_i)$
ResQ	$Q_{jt}(\boldsymbol{\tau},\boldsymbol{\mu})=Q_{tot}(\boldsymbol{\tau},\boldsymbol{\mu})+\omega_r(\boldsymbol{\tau},\boldsymbol{\mu})Q_r(\boldsymbol{\tau},\boldsymbol{\mu})$
RQN	$Q_{tot}(\boldsymbol{\tau},\boldsymbol{\mu})=\sum\limits_{i=1}^{N}Q_i(\tau_i,\mu_i)+\sum\limits_{i=1}^{N}\phi_i(\boldsymbol{\tau},\boldsymbol{\tau})$

3.2.2 提高算法收敛效果的相关研究

许多研究在使用 VDN 和 QMIX 值函数分解方法的基础上，通过改进损失函数、改进动作探索策略和利用智能体间的协作关系等方法，提高算法的收敛效果。

改进损失函数可直接影响算法收敛效果。Tabish 等[25]提出了 WQMIX 方法，根据不同的联合动作，对训练损失进行加权，提高算法在“非单调”环境下对最优动作的估计准确性。文中提出了两种加权函数，中心加权通过增大好的联合动作的权重，使算法对好的联合动作的估计更准确。乐观加权通过增加低估联合动作的权重，降低高估动作的权重，提高算法对所有联合动作值估计的准确性。WQMIX 方法中的加权值为固定的常数，为了使加权方式更精细，Du 等[26]在 WQMIX 基础上提出了动态加权方法 DW-QMIX，使用联合动作为最优联合动作的概率作为权重，通过贝叶斯神经网络计算。

动作探索策略对强化学习方法的收敛影响很大，提高探索效率可加速方法收敛。Mahajan 等[15]提出了 MAVEN 算法，在 QMIX 算法基础上，为每个智能体增加了一个隐策略来生成隐变量，由此可增大方法训练时的动作探索范围，提高找到全局最优解的可能性。Li 等[16]出了一种三级分层强化学习结构，将有效动作序列写成宏操作，降低动作空间的复杂性；引入 PER-QMIX 算法，提高重要经验的采样率，缓解奖励稀疏导致的影响。REHMAN 等[27]基于 QMIX 算法，提出了一种混合动作选择策略 QSOD，在贪婪策略基础上引入狼群优化动作选择策略，引导算法更快收敛。Hall 等[28]提出了一种自适应平均探索方法(AAE)，在训练时根据前一段时间的胜率决定贪婪选择动作的概率，以使算法收敛到高胜率策略。Jiang 等[29]出了一种分层组通信算法 GHGC，将对环境认知一致的智能体分为同一组，组内共享认知信息，组间通过通信共享部分知识。训练时先选择组内最优联合动作，再计算整体联合动作值。

多智能体环境中，智能体间合作程度高的策略一般优于合作少的策略。充分利用智能体间的协作关系可提高算法的收敛效果。Xiong 等[30]提出了一种基于个体之间相关性的方法：CI-VF。其在 QMIX 基础上增加了计算智能体相关性的结构，根据智能体的个体动作值计算相关性，将全体相关性与

混合网络输出相加，得到最终的联合动作值进行训练。Bai等[31]出了一种将超图卷积与值分解相结合的方法：HGCN-MIX。每个智能体根据局部观测计算与其他智能体的连接权重，使用该权重对个体动作值进行超图卷积，将处理后的个体动作值作为QMIX混合网络的输入得到联合动作值。Yun等[32]提出了一种分类状态图注意策略(CSGA)方法。其在QMIX框架下，将状态进行分类，根据观测状态计算自身与队友和敌人的注意力，得到对队友和敌人的动作值，进行加权求和后得到个体动作值。

VDN和QMIX等方法仅从频率主义角度对联合状态动作值的单个平均值进行建模，缺乏多个智能体的不确定性表示。这种忽略，导致算法容易陷入局部最优。Sun等[33]提出了DFAC框架。其根据IGM条件提出了Distributional IGM条件，并将VDN和QMIX改为值分布分解方法DDN和DMIX。其使用个体动作值的均值和分布计算联合动作值分布。Xu等[34]提出了MMD-MIX方法，将QMIX方法的混合网络改为了输出联合动作值的分布，并使用REM方法提高数据采样率，取得了较好的效果。Huang等[35]提出了一种分布式多智能体合作框架DMAC。在训练时，使用全局状态对每个智能体的动作值进行独立修正，使用修正后的个体动作值计算OR损失和平均值损失，依据该损失进行梯度下降，使得智能体的动作值输出分布逐渐趋近于目标分布。Yang等[36]提出了一种多智能体不确定性共享(MAUS)的新方法。基于QMIX结构，将智能体网络中的全连接层替换为贝叶斯网络来表达不确定性，使用TD误差和KL散度作为损失函数进行训练。

一些研究使用了强化学习中其他常用的改进收敛效果的方法。Liu等[37]提出了一种并行训练算法PS-QMIX，将参数服务器框架应用于并行训练QMIX智能体，加速数据收集和学习。参数服务器接收所有个体回传的梯度，更新网络权重并分发，每个个体使用新接收的网络参数与环境交互，产生数据本地存储，计算梯度并回传参数服务器。Wan等[38]提出了一种基于迁移学习的两阶段异构强化学习方法。第一阶段通过增加智能体网络输入层参数，固定隐藏层参数，重置输出层参数进行迁移学习，将在同构多智能体环境中学得的参数转移到异构环境中，加速方法收敛。第二阶段将同构的智能体分为一组，交替对不同的组进行训练。Xiong等[39]对环境奖励进行了改进，提出一种利用每个智能体个性特征的方法：PCQMIX。其设定每个智能体有3个个性特征(勇敢、恐惧和退缩)，根据环境状态和智能体的动作计算个性奖励，使用个性奖励和环境奖励共同进行训练。Wu等[40]提出了Sub-AVG方法，通过舍弃最近时刻高估误差最大的目标动作值，缓解动作值高估问题。

3.2.3 提高算法可扩展性的相关研究

在智能体规模变化的场景中，多智能体强化学习方法的效果变差。环境中智能体数量变化会导致观测数据维度变化和个体动作值的数量变化。智能体的网络为全连接神经网络，无法处理维度变化的观测数据，因此需要对方法进行改进以提高方法的可扩展性。改进主要可分3类：为对混合网络的改进、对智能体网络的改进及对训练环境的改进，如图6所示。

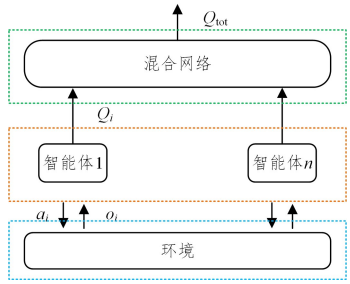

图6 提高可扩展性的主要改进点

Fig. 6 Major improvements to improve scalability

对混合网络改进，使其适应智能体数量变化。Chai等[41]提出了UNMAS方法，其采用自加权混合网络对各智能体的贡献进行评估，将得到的结果加权求和计算出联合动作值，可适应智能体的数量变化。为了保证分解满足IGM条件，自加权混合网络的权重参数由超网络计算得出，并经过绝对值处理，满足非负性。Naderializadeh等[42]提出了一种图混合网络(GraphMIX)进行值函数分解，图混合网络的参数由超参数网络计算得出，满足IGM条件，可一定程度适应智能体数量变化的情况。边的权重使用注意力机制计算，输入为RNN的隐含层特征向量。Zhou等[43]提出了LA-Transformer结构，可处理维度变化的观测数据，并将该结构应用到智能体网络和值混合网络中，以提高方法的可扩展性。Chen等[44]提出了一种关系感知信用分配(RACA)的多智能体强化学习算法。其使用注意力机制处理观测到的数量变化的智能体，使用基于图的关系编码来表示智能体之间的结构，将得到的编码向量作为智能体的个体动作值权重计算联合动作值。

对智能体网络进行改进，使其可处理维度变化观测数据。Wang等[17]设计了一种新的动态多智能体课程学习框架来解决大规模问题。其使用基于图神经网络的动态智能体数量网络(DyAN)来适应网络输入的动态大小变化，让智能体从一个小规模的多智能体场景开始学习，逐步增加场景中智能体的数量。Zhang等[45]针对VDN等方法在团队协作场景中表现不好的问题，提出了CollaQ方法。其将每个智能体的动作值分解为两个项：一个是仅依赖于智能体自身状态的自项，另一个是与当前智能体观察到的邻近智能体状态相关的交互项。使用注意力网络处理对邻近智能体的观测，可使用智能体数量变化场景。Hu等[46]针对动作与实体数量相关的场景下的扩展性和迁移问题，提出了UPDeT方法，每个智能体网络使用transformer处理观测数据，得到不同实体对应的编码向量，根据动作与实体的对应关系，使用相应的编码向量计算对应动作值，使用QMIX或VDN的方法计算联合动作值。对于星际争霸这类包含选择目标动作的场景，可实现输出动作与实体数量相对应。

智能体数量变化会影响环境的变化，这也是导致算法可扩展性差的一个原因。对环境改进，增加智能体探索得到的数据多样性，以适应智能体数量变化的场景。Iqbal等[47]为了得到适应智能体数量变化的控制策略，提出了REFIL方法。在多智能体场景中，有时智能体进行决策仅根据局部状态即可，较远的地方的状态对智能体决策的影响不大。作者从该思路出发，在QMIX基础上，在训练时随机将实体划分为两组，再根据智能体是否观察到该实体，取两者交集，将实体划分为2组。假设环境中仅有组内的实体，分别计算动作

值和损失，将该损失与原 QMIX 方法的加权和作为总损失进行训练。这种方法在训练中可学到适应不同智能体数量的策略。

4 基于值函数分解的多智能体深度强化学习方法应用

基于值函数分解的多智能体深度强化学习方法有较好的仿真效果及可局部观测下分布式执行等优点，适用于很多应用场景。目前的研究主要集中在对网络节点的控制和无人编队的控制两个领域。

4.1 网络节点控制

将网络节点视为单个智能体，整个网络即为多智能体环境。因此可使用基于值函数分解的多智能体深度强化学习方法控制节点解决问题。

在通信网中，可将通信节点或路由节点视为智能体，使用基于值函数分解的多智能体深度强化学习方法进行控制，以达到负载均衡和降低延迟等目的。Lei 等[48]为优化空闲信道之间的频谱利用率和负载平衡，提出了一种基于 QMIX 的智能动态频谱分配(DSA)方法。在训练阶段，用户将计算任务交给移动边缘计算服务器以获得最优分布式动态频谱分配策略，在执行阶段用户通过该策略在本地选择最优通道。Chen 等针对视频点播服务中的协同缓存问题，基于 VDN 提出了 VDDN 算法。其可应对大量视频和用户导致的动作空间复杂问题，提高收敛速度。Guo 等[49]为降低分布式网络通道访问中的延迟，基于 QMIX 提出了一种新的 MAC 协议：QLBT。其利用训练过程中所有智能体的整体信息，确保每个智能体都能根据自身的局部观察，独立地推断出最优的信道接入行为。Wang 等[50]针对具有多个发射机-接收机对的无线网络场景，提出了一种基于 QMIX 的分散多智能体功率控制(DECMAPC)算法来实现总速率最大化。该算法具有 QMIX 可处理局部观测信息的特点，对信道和环境相互作用引起的干扰变化具有鲁棒性。Han 等[51]针对耐延迟网络的路由问题，提出了一种基于 QMIX 的辅助路由算法，进行可靠的下一跳选择。Mseddi 等[52]针对网络中的请求用户和附近缓存设备之间的分布式配对问题，提出了一种基于 QMIX 算法的协作智能解决方案，用于蜂窝网络底层设备到设备用户之间的在线流量卸载。

在交通网中，可将路口节点视为智能体。多智能体深度强化学习方法有无模型、自适应强的优点[53]。可使用其进行交通信号灯控制，以达到减少堵车，提高通行效率等目的。Wang 等[54]为在带宽有限的情况下实现自适应交通灯信号控制(ATSC)，提出了一个通信高效的分散 ATSC 框架。其使用生成对抗网络进行交通数据恢复，利用交通统计数据恢复邻近的实时交通数据，使用值函数分解强化学习方法进行交通信号灯控制。Zhang 等[55]提出了一种基于合作博弈的多智能体强化学习方法(CG-MARL)，使用了基于值函数分解的方法框架，集中式训练，分布式执行，每个智能体根据局部观测进行决策。Chen 等[56]为满足大规模网络化交通信号控制中规模和协调的双重要求，提出了一种基于 Qmix 和 LSTM 通信模块的 C-Qmix 方法。在 QMIX 框架基础上，使用 LSTM 整合其他智能体的历史观察和行动，输出整合后的信息作为智能体的输入。

4.2 无人编队控制

基于值函数分解的多智能体深度强化方法可在局部观测分布式执行，这适用于很多通信环境差的多智能体场景，因此其在自动驾驶、无人机搜救、智能工厂等涉及无人编队控制的领域有很多应用研究。

在自动驾驶领域，Zhang 等[57]提出一种改进的电动汽车组队分布式模型预测控制(IDMPC)方法。IDMPC 中的对称权矩阵对最终的控制效果至关重要。基于多智能体强化学习中的 QMIX 算法，对 IDMPC 的权值进行优化，以协调地控制队列中跟随的车辆。Yuan 等[58]针对城市地区复杂的道路结构和移动受限的多车追击场景，定义了一个观测约束多车追击问题(OMVP)，提出了 T3OMVP 方法。其基于 QMIX 框架，在智能体网络中使用 transformer 编码器处理观测数据。Zhou 等[59]针对自动驾驶汽车载客和充电场景下的路径规划问题，提出了综合了值函数分解和 AC 算法的 GMIX 方法，使用基于图注意力网络的混合网络计算联合动作值，以更好地利用智能体之间的相互作用。

在无人机编队控制领域，Yin 等[60]针对多无人机运送救灾物资的任务分配问题，提出了一种基于 QMIX 的深度转移强化学习算法。通过使用在源任务组中训练好的参数，加速算法在新任务环境中的收敛速度，提高分配效率。Wang 等[61]针对使用无人机群对车联网提供边缘计算问题，提出了结合值函数分解方法的 AC-MIX 和 MA2DDPG 算法进行无人机移动控制和通信带宽分配。Ding 等[62]针对多无人机通信中继问题，提出了一种具有两种训练机制的多智能体强化学习算法：MAQMIX。每个 UAV 被分解为 3 个子代理，分别负责 UAV 轨迹设计、频率资源分配和下一跳选择，使用 QMIX 混合网络进行 UAV 内训练。使用下一跳选择和下一时刻无人机位置，计算传输用时，进行 UAV 间训练。

在智能工厂领域，Ritz 等[63]针对智能工厂中的智能体行动规范性问题，提出了将规范转换为奖励的方法，并使用 QMIX 等多个多智能体深度强化学习方法进行了验证。CHOI 等[64]针对物流仓库的自动运输场景，提出了一种基于 QMIX 的多 AGV 协同路径控制方法，通过动作屏蔽消除碰撞，增加智能体额外损失，提高智能体协作程度。

结束语 本文根据基于值函数分解的多智能体深度强化学习方法研究现状，总结分析了提高混合网络拟合能力、提高算法收敛效果、提高算法可扩展性 3 个热点研究问题；根据不同方法拟解决的问题和改进点，将方法进行了分类介绍；对基于值函数分解的多智能体深度强化学习方法在网络节点控制和无人编队控制领域的应用进行了综述。

由于基于值函数分解的多智能体深度强化学习方法的良好表现，将该类方法和思想引入其他多智能体深度强化学习领域也可取得较好的提升效果，如与基于策略的方法相结合，引入通信机制等。一些研究人员已经取得了一定成果[65-67]，后续可进一步关注。

参 考 文 献

[1] TAMPUU A, MATIISEN T, KODELJA D, et al. Multiagent cooperation and competition with deep reinforcement learning [J]. Plos One, 2017, 12(4): e0172395.

[2] LOWE R, WU Y, TAMAR A, et al. Multi-Agent Actor-Critic

for Mixed Cooperative-Competitive Environments[C]// Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017.

[3] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual Multi-Agent Policy Gradients[C]// The Thirty-second AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, Usa: AAAI Press, 2018: 2974-2982.

[4] WONG A, BÄCK T, KONONOVA A V, et al. Deep multiagent reinforcement learning: challenges and directions[J]. Artificial Intelligence Review, 2022, 56: 5023-5056.

[5] HAO J, YANG T, TANG H, et al. Exploration in Deep Reinforcement Learning: From Single-Agent to Multiagent Domain [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 1(1): 1-21.

[6] DU F, DING S F. A survey of multi-agent Reinforcement learning[J]. Computer Science, 2019, 46(8): 1-8.

[7] SUN Y, CAO L, CHEN X L, et al. Overview of multi-agent deep reinforcement learning[J]. Computer Engineering and Applications, 2020, 56(5): 13-24.

[8] YAN C, XIANG X J, XU X, et al. A Survey on the Scalability and Transferability of Multi-Agent Deep Reinforcement Learning[J]. Control and Decision, 2023, 37(12): 3083-3102.

[9] XIONG L Q, CAO L, LAI J, et al. Overview of Multi-agent Deep Reinforcement Learning Based on Value Factorization[J]. Computer Science, 2022, 49(9): 172-182

[10] LI T, ZHU K, LUONG N C, et al. Applications of Multi-Agent Reinforcement Learning in Future Internet: A Comprehensive Survey[J]. IEEE Communications Surveys & Amp; Tutorials, 2022, 24(2): 1240-1279.

[11] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with Deep Reinforcement Learning[EB/OL]. arXiv: 1312. 5602, 2013. https://ui. adsabs. harvard. edu/abs/2013arXiv1312. 5602M.

[12] HASSELT H V, GUEZ A, SILVER D. Deep Reinforcement Learning with Double Q-Learning[C]// Proceedings of the Thirtieth Aaai Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press, 2016: 2094-2100.

[13] SON K, KIM D, KANG W, et al. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement learning[C]// International Conference on Machine Learning. 2019.

[14] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-Decomposition Networks For Cooperative Multi-Agent Learning [EB/OL]. arXiv: 1706. 05296, 2017. https://ui. adsabs. harvard. edu/abs/2017arXiv170605296S.

[15] MAHAJAN A, RASHID T, SAMVELYAN M, et al. MAVEN: Multi-Agent Variational Exploration[C]// Advances in Neural Information Processing Systems 32 (NIPS 2019). California: Neural Information Processing Systems(NIPS), 2019.

[16] LI B. Hierarchical Architecture for Multi-Agent Reinforcement Learning in Intelligent Game[C]// 2022 International Joint Conference on Neural Networks(IJCNN). New York: IEEE, 2022.

[17] WANG W, YANG T, LIU Y, et al. From Few to More: Large-Scale Dynamic Multiagent Curriculum Learning[C]// Thirty-fourth Aaai Conference on Artificial Intelligence, the Thirty-second Innovative Applications of Artificial Intelligence Conference and the Tenth Aaai Symposium on Educational Advances in Artificial Intelligence. New York: Assoc Advancement Artificial Intelligence, 2020: 7293-7300.

[18] COHEN A, TENG E, BERGES V, et al. On the Use and Misuse of Absorbing States in Multi-agent Reinforcement Learning [EB/OL]. arXiv: 2111. 05992, 2021. https://ui. adsabs. harvard. edu/abs/2021arXiv211105992C. 10. 48550/arXiv. 2111. 05992.

[19] RASHID T, SAMVELYAN M, DE WITT C, et al. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning [J]. Journal of Machine Learning Research, 2020, 21.

[20] YANG Y, HAO J, LIAO B, et al. Qatten: A General Framework for Cooperative Multiagent Reinforcement Learning[EB/OL]. arXiv: 2002. 03939, 2020. https://ui-adsabs-harvard-edu-s. libyc. nudt. edu. cn: 443/abs/2020arXiv200203939Y.

[21] WANG J, REN Z, LIU T, et al. QPLEX: Duplex Dueling Multi-Agent Q-Learning[EB/OL]. arXiv: 2008. 01062, 2020. https://ui-adsabs-harvard-edu-s. libyc. nudt. edu. cn: 443/abs/2020arXiv200801062W.

[22] WANG Z, SCHAUL T, HESSEL M, et al. Dueling Network Architectures for Deep Reinforcement Learning[C]// International Conference on Machine Learning. 2016.

[23] SIQI S, MENGWEI Q, JUN L. ResQ: A Residual Q Function-based Approach for Multi-Agent Reinforcement Learning Value Factorization[C]// 36th Conference on Neural Information Processing Systems. New York: Curran Associates, 2022: 5471-5483.

[24] PINA R, DE SILVA V, HOOK J, et al. Residual Q-Networks for Value Function Factorizing in Multi-Agent Reinforcement Learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(2): 1534-1544.

[25] RASHID T, FARQUHAR G, PENG B, et al. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning[C]// Advances in Neural Information Processing Systems 33 (NEURIPS 2020). New York: Curran Associates, 2020: 10199-10210.

[26] DU W, DING S, GUO L, et al. Value function factorization with dynamic weighting for deep multi-agent reinforcement learning [J]. Information Sciences, 2022, 615: 191-208.

[27] REHMAN H M R U, ON B, NINGOMBAM D D, et al. QSOD: Hybrid Policy Gradient for Deep Multi-agent Reinforcement Learning[J]. Ieee Access, 2021, 9: 129728-129741.

[28] HALL G, HOLLADAY K. Adaptive Average Exploration in Multi-Agent Reinforcement Learning[C]// 2020 AIAA/IEEE 39th Digital Avionics Systems Conference(DASC) Proceedings. New York: IEEE, 2020.

[29] JIANG H, SHI D, XUE C, et al. GHGC: Goal-based Hierarchical Group Communication in Multi-Agent Reinforcement Learning[C]// 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). New York: IEEE, 2020: 3507-3514.

[30] XIONG L, CAO L, CHEN X, et al. A Value Factorization Method for MARL Based on Correlation between Individuals[J].

Mathematical Problems in Engineering, 2022, 2022: 1-8.

[31] BAI Y, GONG C, ZHANG B, et al. Cooperative Multi-Agent Reinforcement Learning with Hypergraph Convolution [C]// 2022 International Joint Conference on Neural Networks (IJCNHN). New York: IEEE, 2022.

[32] YUN W J, YI S, KIM J. Multi-Agent Deep Reinforcement Learning using Attentive Graph Neural Architectures for Real-Time Strategy Games [C]// 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). New York: IEEE, 2021: 2967-2972.

[33] SUN W, LEE C, LEE C. DFAC Framework: Factorizing the Value Function via Quantile Mixture for Multi-Agent Distributional Q-Learning [C]// International Conference on Machine Learning. 2021.

[34] XU Z, LI D, BAI Y, et al. MMD-MIX: Value Function Factorisation with Maximum Mean Discrepancy for Cooperative Multi-Agent Reinforcement Learning [C]// 2021 International Joint Conference on Neural Networks (IJCNHN). New York: IEEE, 2021.

[35] HUANG L, FU M, RAO A, et al. A Distributional Perspective on Multiagent Cooperation With Deep Reinforcement Learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 4246-4259.

[36] YANG G, CHEN H, ZHANG J, et al. Multi-Agent Uncertainty Sharing for Cooperative Multi-Agent Reinforcement Learning [C]// 2022 International Joint Conference on Neural Networks (IJCNN). New York: IEEE, 2022: 1-8.

[37] LIU X, LI X, LI Y, et al. PS-QMix: A Parallel Learning Framework for Q-Mix Using Parameter Server [C]// Advanced Data Mining and Applications (ADMA 2021). 2022: 341-352.

[38] WAN K, XU X, LI Y. Learning Distinct Strategies for Heterogeneous Cooperative Multi-agent Reinforcement Learning [C]// Artificial Neural Networks and Machine Learning (ICANN 2021). Switzerland: Springer International Publishing AG, 2021: 544-555.

[39] LIQIN X, LEI C, XILIANG C, et al. Character-Based Value Factorization For MADRL [J]. The Computer Journal, 2023, 66(11): 2782-2793.

[40] WU H, ZHANG J, WANG Z, et al. Sub-AVG: Overestimation reduction for cooperative multi-agent reinforcement learning [J]. Neurocomputing, 2022, 474: 94-106.

[41] CHAI J, LI W, ZHU Y, et al. UNMAS: Multiagent Reinforcement Learning for Unshaped Cooperative Scenarios [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(4): 2093-2104.

[42] NADERIALIZADEH N, HUNG F H, SOLEYMAN S, et al. Graph Convolutional Value Decomposition in Multi-Agent Reinforcement Learning [EB/OL]. 2020: arXiv: 2010. 04740. https://ui. adsabs. harvard. edu/abs/2020arXiv201004740N. 10. 48550/arXiv. 2010. 04740.

[43] ZHOU T, ZHANG F, SHAO K, et al. Cooperative Multi-Agent Transfer Learning with Level-Adaptive Credit Assignment [EB/OL]. arXiv: 2106. 00517, 2021. https://ui. adsabs. harvard. edu/abs/2021arXiv210600517Z. 10. 48550/arXiv. 2106. 00517.

[44] CHEN H, YANG G, ZHANG J, et al. RACA: Relation-Aware Credit Assignment for Ad-Hoc Cooperation in Multi-Agent Deep Reinforcement Learning [C]// 2022 International Joint Conference on Neural Networks (IJCNN). New York: IEEE, 2022.

[45] ZHANG T, XU H, WANG X, et al. Multi-Agent Collaboration via Reward Attribution Decomposition [EB/OL]. arXiv: 2010. 08531, 2020. https://ui. adsabs. harvard. edu/abs/2020arXiv201008531Z.

[46] HU S, ZHU F, CHANG X, et al. UPDeT: Universal Multi-agent Reinforcement Learning via Policy Decoupling with Transformers [EB/OL]. arXiv: 2101. 08001, 2021. https://ui. adsabs. harvard. edu/abs/2021arXiv210108001H. 10. 48550/arXiv. 2101. 08001.

[47] IQBAL S, DE WITT C, PENG B, et al. Randomized Entity-wise Factorization for Multi-Agent Reinforcement Learning [C]// International Conference on Machine Learning. San Diego: Jmlr-Journal Machine Learning Research, 2021.

[48] LEI C, ZHAO H, ZHOU L, et al. Intelligent Dynamic Spectrum Allocation in MEC-Enabled Cognitive Networks: A Multiagent Reinforcement Learning Approach [J]. Wireless Communications and Mobile Computing, 2022, 2022: 1-13.

[49] GUO Z, CHEN Z, LIU P, et al. Multi-Agent Reinforcement Learning-Based Distributed Channel Access for Next Generation Wireless Networks [J]. IEEE Journal on Selected Areas in Communications, 2022, 40(5): 1587-1599.

[50] WANG Z, ZONG J, ZHOU Y, et al. Decentralized Multi-Agent Power Control in Wireless Networks With Frequency Reuse [J]. IEEE Transactions on Communications, 2022, 70(3): 1666-1681.

[51] HAN C, YAO H, MAI T, et al. QMIX Aided Routing in Social-Based Delay-Tolerant Networks [J]. IEEE Transactions on Vehicular Technology, 2022, 71(2): 1952-1963.

[52] MSEDDI A, JAAFAR W, MOUSSAID A, et al. Collaborative D2D Pairing in Cache-Enabled Underlay Cellular Networks [C]// 2021 IEEE Global Communications Conference (globecom). New York: IEEE, 2021: 1-6.

[53] YU Z, NING N W, ZHENG Y L, et al. Survey of Intelligent Traffic Signal Control Strategies Driven by Deep reinforcement learning [J]. Computer Science, 2023, 50(4): 159-171.

[54] WANG Z, ZHU H, HE M, et al. GAN and Multi-Agent DRL Based Decentralized Traffic Light Signal Control [J]. IEEE Transactions on Vehicular Technology, 2022, 71(2): 1333-1348.

[55] ZHANG Z, QIAN J, FANG C, et al. Coordinated Control of Distributed Traffic Signal Based on Multiagent Cooperative Game [J]. Wireless Communications and Mobile Computing, 2021, 2021: 1-13.

[56] CHEN X, XIONG G, LV Y, et al. A Collaborative Communication-Qmix Approach for Large-scale Networked Traffic Signal Control [C]// 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). New York: IEEE, 2021: 3450-3455.

[57] ZHANG S, ZHUAN X. Distributed Model Predictive Control for Two-Dimensional Electric Vehicle Platoon Based on QMIX Algorithm [J]. Symmetry, 2022, 14(10): 2069.

[58] YUAN Z, WU T, WANG Q, et al. T3OMVP: A Transformer-

Based Time and Team Reinforcement Learning Scheme for Observation-Constrained Multi-Vehicle Pursuit in Urban Area[J]. Electronics，2022，11(9)：1339.

[59] ZHOU T，KRIS M L，CREIGHTON D，et al. GMIX：Graph-based spatial-temporal multi-agent reinforcement learning for dynamic electric vehicle dispatching system[J]. Transportation Research Part C：Emerging Technologies，2022，144：103886.

[60] YIN Y，GUO Y，SU Q，et al. Task Allocation of Multiple Unmanned Aerial Vehicles Based on Deep Transfer Reinforcement Learning[J]. Drones，2022，6(8)：215.

[61] WANG J，ZHANG X，HE X，et al. Bandwidth Allocation and Trajectory Control in UAV-Assisted IoV Edge Computing Using Multiagent Reinforcement Learning[J]. IEEE Transactions on Reliability，2023，72(2)：599-608.

[62] DING R，CHEN J，WU W，et al. Packet Routing in Dynamic Multi-Hop UAV Relay Network：A Multi-Agent Learning Approach[J]. IEEE Transactions on Vehicular Technology，2022，71(9)：10059-10072.

[63] RITZ F，PHAN T，MÜLLER R，et al. SAT-MARL：Specification Aware Training in Multi-Agent Reinforcement Learning [C] // Proceedings of the 13th International Conference on Agents and Artificial Intelligence：SCITEPRESS-Science and Technology Publications. 2021：28-37.

[64] CHOI H，KIM J，HAN Y，et al. MARL-Based Cooperative Multi-AGV Control in Warehouse Systems[J]. IEEE Access，2022，10：100478-100488.

[65] HANHAN Z，TIAN L，VANEET A. PAC：Assisted Value Factorisation with Counterfactual Predictions in Multi-Agent Reinforcement Learning[C] // 36th Conference on Neural Information Processing Systems. New York：Curran Associates，2022：15757-15769.

[66] WANG Y，HAN B，WANG T，et al. Off-Policy Multi-Agent Decomposed Policy Gradients[EB/OL]. arXiv：2007. 12322. https：//ui. adsabs. harvard. edu/abs/2020arXiv200712322W. 10. 48550/arXiv. 2007. 12322.

[67] WANG T，WANG J，ZHENG C，et al. Learning Nearly Decomposable Value Functions Via Communication Minimization[EB/OL]. arXiv：1910. 05366. https：//ui. adsabs. harvard. edu/abs/2019arXiv191005366W.

GAO Yuzhao，born in 1994，postgraduate. His main research interests include multi-agent deep reinforcement learning，UGV and task planning.

NIE Yiming，born in 1982，associate research fellow. His main research interests include intelligence unmanned systems and UGV.