计算机科学
COMPUTER SCIENCE

深度强化学习驱动的智能交通信号控制策略综述

于泽，宁念文，郑燕柳，吕怡宁，刘富强，周毅

引用本文

于泽，宁念文，郑燕柳，吕怡宁，刘富强，周毅深度强化学习驱动的智能交通信号控制策略综述[J]．计算机科学，2023，50(4)：159-171.

YU Ze，NING Nianwen，ZHENG Yanliu，LYU Yining，LIU Fuqiang，ZHOU Yi. Review of Intelligent Traffic Signal Control Strategies Driven by Deep Reinforcement Learning [J]. Computer Science，2023，50(4)：159-171.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500261

深度强化学习驱动的智能交通信号控制策略综述

于 泽[1] 宁念文[1,4] 郑燕柳[2] 吕怡宁[1] 刘富强[3] 周 毅[1,4]
1 河南大学人工智能学院 郑州 450046
2 湖南大学信息科学与工程学院 长沙 410006
3 同济大学电子与信息工程学院 上海 201804
4 河南大学深圳研究院 广东 深圳 518000
(yuze@henu.edu.cn)

摘 要 随着城市人口快速增加,私家车数量呈指数级增长,使本已不堪重负的交通系统将承受更大的压力,交通拥堵问题愈加凸显。传统交通信号控制技术难以适应复杂多变的交通情况,数据驱动的方法为基于控制的系统带来了新方向。深度强化学习方法与交通控制系统的结合在自适应交通信号控制中扮演着重要角色。首先,文中综述了智能交通信号控制系统应用的最新进展,对智能交通信号控制方法进行了分类讨论,总结了这一领域的现有工作。其次,采用深度强化学习方法能够有效解决智能交通信号控制中状态信息获取不准确、控制算法鲁棒性差以及区域协调控制能力弱等问题,在此基础上,给出了智能交通信号控制的仿真平台和实验设置概述,并通过实例进行了分析和验证。最后,探讨了智能交通信号控制领域面临的挑战和有待解决的问题,并总结了未来的研究方向。

关键词:智能交通系统;深度强化学习;交通信号控制;多智能体
中图法分类号 TP181

Review of Intelligent Traffic Signal Control Strategies Driven by Deep Reinforcement Learning

YU Ze[1],NING Nianwen[1,4],ZHENG Yanliu[2],LYU Yining[1],LIU Fuqiang[3] and ZHOU Yi[1,4]
1 School of Artificial Intelligence,Henan University,Zhengzhou 450046,China
2 College of Computer Science and Electronic Engineering,Hunan University,Changsha 410006,China
3 College of Electronic and Information Engineering,Tongji University,Shanghai 201804,China
4 Shenzhen Research Institute of Henan University,Shenzhen,Guangdong 518000,China

Abstract With the rapid growth of urban populations,the number of private cars has grown exponentially,which makes overwhelming traffic congestion problem become more and more acute. The traditional traffic signal control technology is difficult to adapt to the complex and changeable traffic conditions,and the data-driven methods bring new research directions for the control-based system. The combination of deep reinforcement learning and traffic control systems plays an important role in adaptive traffic signal control. First,this paper reviews the latest progress in the application of intelligent traffic signal control systems,the methods of intelligent traffic signal control are classified and discussed,and the existing works in this field are summarized. The deep reinforcement learning method can effectively solve the problems of inaccurate state information acquisition,poor algorithm robust and weak regional coordination control ability in intelligent traffic signal control. Then,on the basis of the above,this paper gives an overview of the simulation platforms and experimental setup for intelligent traffic signal control,and analyzes and verifies it through examples. Finally,The challenges and unsolved problems in this field are discussed and future research directions are summarized.

Keywords Intelligent transportation system,Deep reinforcement learning,Traffic signal control,Multi-agent

到稿日期:2022-05-28 返修日期:2022-09-11
基金项目:国家自然科学基金(62176088);河南省科技攻关计划(222102210067,222102520028);深圳市中央引导地方科技发展专项(2021Szvup029)
This work was supported by the National Natural Science Foundation of China(62176088),Key Science and Technology Program of Henan Province,China(222102210067,222102520028) and Shenzhen Special Foundation of Central Government to Guide Local Science & Technology Development(2021Szvup029).
通信作者:宁念文(nnw@henu.edu.cn)

1 引言

近年来，由于人口数量迅速增长和城市化快速发展，城市交通需求不断增长。据公安部统计[1]，2021 年全国机动车保有量达 3.95 亿辆，其中汽车有 3.02 亿辆，已跃居世界第一。随着高层、超高层楼群大规模的兴建，超高层建筑具有容量大、人口集中等特点，给附近城市交通带来了巨大压力。滴滴和 Uber 等打车软件的普及化，加剧了城市道路的拥堵情况。调查结果显示，打车软件公司加入后，交通拥堵强度增加了近 1%，拥堵持续时间增加了 4.5%[2]。交通拥堵问题日益严重，缓解城市交通拥堵、提高出行效率已成为重要的民生问题。

物联网时代，智慧城市成为了一种新的城市发展理念，智能交通系统（Intelligent Transportation System，ITS）成为智慧城市建设的重要组成部分。《十四五规划》框定了以智能交通为首的十大数字化应用场景的具体范围，成为了构建数字社会、提升政府数字治理水平、营造良好数字生态的重要指引。智能交通系统通过人、车、路等要素的高效协同，提升了交通出行效率。其中，解决交通拥堵是智能交通系统的关键任务。《国家综合立体交通网规划纲要》指出：到 2035 年，基本建成便捷顺畅、经济高效、绿色集约、智能先进、安全可靠的现代化高质量国家综合立体交通网。道路基础设施扩建方案耗费大量人力物力，短时间内已不再适用。目前，缓解交通拥堵最直接有效的办法就是对交通信号进行智能控制与配时优化。在实际应用中，复杂的交通环境以及各个路口之间的耦合关系，使得如何高效协调多个交叉口的交通信号面临巨大挑战。

传统自适应交通信号控制采用三级控制方式（路口级、区域级和中心级）。路口级使用摄像头或传感器等设备收集车辆信息，通过通信网络传输给区域管理级。区域管理级进行调度优化，并形成调度指令传输给路口级。中心级为最高指挥级，在紧急情况下进行宏观调度。自适应交通信号控制很大程度上依赖当前交通状态，并没有考虑到未来的交通情况，因此很难达到全局最优状态。此外，自适应交通信号控制系统过分依赖计算机硬件设备，在计算复杂度过高时，系统可能无法得到全局最优交通信号控制策略。

日益严重的交通拥堵问题得到了越来越多研究人员的关注与思考，并产生了一系列重要且有效的解决方案和研究成果。如表 1 所列，按交通信号控制算法类型将智能交通信号控制研究分为 3 类，分别为统计学方法、模型驱动方法和强化学习方法。

（1）基于统计学的交通信号控制方法（定时控制方法和车辆驱动控制方法）

定时控制方法依靠固定时长法[3]或手动启发式算法[4]控制交通信号，缓解交通拥堵问题的效率低下。另外，由于人为进行预先设定的限制，无法适应动态不确定的交通场景。车辆驱动的控制方法，通常根据预先设定的交通规则和实时交通数据触发交通信号控制系统。其中，车辆驱动控制方法中常见的一种方法是通过设置阈值来判断交通信号是否需要改变。当队列长度超过该阈值时，则设置该车道交通信号为绿色信号。由于不同流量车道的最佳阈值是不同的，因此必须进行手动调优，这在实际应用中是很难实现的。

（2）基于模型的交通信号控制方法（如 Webster，GreenWave 和 Max-pressure 等）

基于模型的智能交通信号控制策略，例如 Webster[5]，GreenWave[6]和 Max-pressure[7]等，此类方法通常是在预设周期或基于固定周期相序[8]等假设下，对交通信号配时进行求解。Webster 是交通信号控制中广泛应用的方法，常用于孤立的交叉路口。该模型假设某一时段内的交通流是均匀的，构建一个封闭方程，求解该交叉口下的最优周期时长，并进行相位分割，使得通过该交叉口的所有车辆的行驶时间最短。GreenWave 是交通领域中实现协调的经典方法，其目的是优化偏移量，以减少特定方向上车辆停靠的次数。Max-pressure 旨在通过平衡相邻交叉口之间的车辆排队长度和最小化交叉口压力来降低过饱和风险。然而，这些方法中存在不合理的假设条件（例如统一到达率和无限制的车道容量），并不能够满足实际的交通需求。

在此基础上研究人员将神经网络[9]、粒子群算法[10]、遗传算法[11]等方法应用于自适应交通信号控制（Adaptive Traffic Signal Control，ATSC），以提高系统寻找全局最优策略的能力。但是，这些算法本身存在一些问题，例如粒子群算法的计算精度较低，系统不容易收敛，局部寻优能力较差，遗传算法收敛速度较慢。

表 1 智能交通信号控制方法分类及其优缺点

Table 1 Classification, advantages and disadvantages of intelligent traffic signal control methods

	方法	优点	缺点
统计学	定时控制方法	计算复杂度低、可实施性强	难以满足实时性
	车辆驱动控制方法	高鲁棒性	手动调优、难以适应真实场景
模型驱动	Webster，GreenWave，Max-pressure 等	模型简单、可构建非线性数据模型	模型固定、难以处理不确定性因素
数据驱动	强化学习方法	无模型、离线学习、计算复杂度低	维度爆炸问题
	深度强化学习方法	自适应性强	训练时间长、难以适应复杂交通场景

（3）基于强化学习的交通信号控制方法（强化学习方法和深度强化学习方法）

随着人工智能技术的兴起，互联网和无线通信的持续发展，数据获取变得更方便、更快捷。基于数据驱动的强化学习（Reinforcement Learning，RL）方法在智能交通信号控制中的应用愈加广泛。与传统模型驱动方法相比，强化学习方法不依赖启发式假设和启发式方程。多数基于强化学习的智能交通信号控制方法是针对单智能体展开研究的。单智能体强化学习观测交通状态信息，执行当前状态下的最优交通信号策略，但忽略了各个交叉路口之间的影响关系[12-14]。为了解决此类问题，研究人员提出了基于多智能体强化学习[15-16]方法，以解决多个交叉口以之间的协同问题。如何高效地协调多个

交叉口以提高出行效率，已成为当前应用研究的重点和难点。

现有的多智能体强化学习算法注重智能体之间的协同，根据不同的信息结构可分为集中式和分布式两种结构[17]。独立强化学习方法（Independent Reinforcement Learning，IRL）作为一种完全分布方法，直接对每个智能体执行强化学习算法且智能体之间不进行信息交换。该方法已被应用于解决多交叉口交通信号控制问题[18]。然而，环境状态在多智能体强化学习中是共享的，并且会随着每个智能体策略和状态的改变而改变[19]，这使得强化学习算法能够得到当前状态下单个智能体的最优策略，却很难得到该场景下的全局最优策略。文献[20]比较了独立强化学习和值分解网络（Value-Decomposition Networks，VDN）。作为一种集中式方法，VDN通过中央控制器，将各个智能体的值函数进行集成，得到一个联合动作值函数。QMIX[21]作为VDN的扩展，利用状态信息进行非线性集成，获得了比VDN更强的逼近能力。随机分解学习（Randomized Entity-wise Factorization Learning，REFL）通过考虑每个智能体的贡献对QMIX进行改进，并在其基础上改变收益函数，引入多头注意力机制[22]。QMIX和VDN都是典型的集中式多智能体强化学习算法。但是，集中式系统控制缺乏灵活性，系统进行大量数据计算时会降低系统的实时性。

传统强化学习系统的计算复杂度会随着状态空间的增加呈指数增长。交叉路口和智能交通信号控制系统数量的增加会导致系统计算复杂度迅速提高。为了解决这一问题，研究人员采用近似函数方法来降低状态空间维度。近似Q学习（Approximate Q-learning，AQL）是将Q函数近似为线性函数的一种典型方法，主要思想为使用一定数量的参数来近似Q值[23]，即计算权值向量与基于状态行为特征向量的内积。

深度学习与强化学习相结合的深度强化学习算法[24-26]的诞生，使得城市交通信号控制技术进一步迈入智能时代。深度强化学习算法采用神经网络代替近似函数，以解决状态空间维度迅速增加的问题，在实际应用中，解决单路口拥堵并不具有实际意义[27-29]。因此，将深度强化学习算法应用于多路口，去解决多智能体的协同控制问题十分具有前景。当前利用深度强化学习方法能够较好地解决单个交叉路口的拥堵问题，但在多个交叉路口的应用上还存在较多的问题。例如，各个交叉路口之间的耦合关系，智能交通信号控制系统的冗余问题以及交叉路口之间信息传递的安全性问题。

因此，智能交通信号控制系统配时优化研究所面临的挑战主要来源于以下几方面。

（1）状态信息获取不准确

由于进入交叉路口的交通流是时变且不确定的，每个交叉路口只能获取到局部的状态信息，而不能获取到完整的状态信息，这将导致状态观测信息获取不准确。因此，根据环境观测信息来判断自身状态存在偏差，这使得场景变得部分可观测。如何解决部分场景可观测下的智能交通信号配时问题，已成为首要挑战。

（2）控制算法鲁棒性差

在多交叉路口协调过程中，若其中一个交叉路口的观测环境是动态的、非平稳的，则将会引起整个系统难以收敛的问题。大多数智能交通信号控制算法通常在孤立的交叉口或在小型交通网络中进行实验。如何将单路口扩展到多路口，并在交通流时变且不确定的情况下保持系统良好的鲁棒性和稳定性，将是智能交通信号控制面临的核心挑战。

（3）区域协调控制能力弱

在拥有多个交叉路口的场景中，由于各个交叉路口的交通拥堵情况不同，路口之间会相互影响。每个交叉路口只执行当前观测状态下的最优策略，而缺少与其他交叉路口之间的信息交互，将导致在该场景下不能得到全局最优策略。如何在大规模场景下实现全局最优策略并保持较好的收敛性是一项非常重要的挑战。

综上所述，本文针对智能交通信号控制策略的研究现状及所面临的挑战进行了总结，主要综述了基于深度强化学习的智能交通信号控制方法。本文第2节对应用于智能交通信号控制的系统模型进行定义；第3节对目前所面临的挑战（状态信息获取不准确、控制算法鲁棒性差和区域协调控制能力弱）进行讨论，并对解决交通信号控制的典型方法以及面临的挑战进行了综述；第4节总结了用于智能交通信号控制的仿真平台以及相应数据集，可为相关后续研究提供借鉴和参考；第5节深入挖掘智能交通信号控制中潜在的研究方向；最后总结全文。

2　系统模型

智能交通信号控制具有无后效性，即只与现在和历史状态有关，与将来预测状态无关。因此，研究人员将智能交通信号控制问题定义为马尔可夫决策过程（Markov Decision Processes，MDP）[30]。由于深度强化学习模型更适合处理高维数据和连续状态问题，现有研究大多集中于深度强化学习的智能交通信号控制方法研究，其模型如图1所示。MDP作为强化学习的基本框架，实现智能体与环境之间的信息交互。将交叉路口当作智能体，利用智能体实现对交通信号的控制。将单路口下智能体控制过程定义为马尔可夫决策过程，使用$MDP=(S,A,P_{sa},\pi,\gamma)$来表示。智能交通信号控制系统通过智能体之间的相互协同和竞争，来得到此场景下全局最优交通信号策略组合。因此，交通信号控制过程可以定义为有限步内多智能体的马尔可夫决策过程。将此过程使用一个数组来表示，即$MDP=(S,A,N,P_{sa},\pi,\gamma)$。

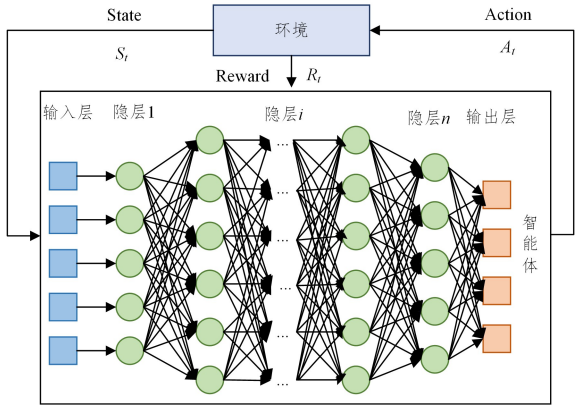

图1　深度强化学习系统框架

Fig. 1　Framework of deep reinforcement learning systems

2.1 系统状态观测空间 S

假设在该场景下存在 N 个交叉路口，每个交叉路口将部分状态信息作为其系统状态观测空间 $s \in S$。智能体根据系统状态观测信息执行相应的最优动作。本节综述了当前研究中常用的几种状态观测信息。

(1)车辆队列长度信息[12,31]。该信息表示在该条车道上等待的车辆数量。队列长度信息反映了所在道路上的拥堵程度，通常将不同车道上的队列信息组成一个状态向量作为状态观测信息。

(2)车辆的位置、速度信息[27,32-33]。首先将道路进行网格化离散处理，采用二进制矩阵来表示车辆的位置信息。该位置存在车辆表示为 1，否则为 0。另外，与车辆位置信息表示方法相似，同样使用矩阵来表示车辆速度信息。通常将速度小于 0.1m/s 的车辆视为等待状态，用 0 来表示该等待车辆，否则用具体的车辆速度来表示车辆的运动状态。这种状态形式化方法能够更加准确地表示真实交叉路口信息。

(3)交通信号相位信息[12-14,27-29]。使用矩阵表示当前交通信号所处的状态信息。交通信号相位状态信息的表示能有效减少在相位时间段内的冲突，从而提高交通安全性和通行效率。

除了上述总结的状态观测信息以外，还存在着其他的状态观测信息。例如，车道上车辆数目、行人状态信息、进出车道车辆数目以及相位持续时间等信息。状态观测信息的选择可以是上述总结中的一种，也可以是多种状态信息的随机组合。

2.2 动作空间 A

智能体接收到来自环境的状态观测信息，并根据该信息执行当前状态下的最优动作 A_m。在大多数的智能交通信号控制中，动作集合通常是离散的，可分为固定相序和可变相序。

(1)固定相序[12,32,34]

智能体的动作集合 $A_m = \{A_{m1}, A_{m2}\}$。智能体执行动作为 A_{m1} 时，保持当前相序，相序不发生改变。当执行动作为 A_{m2} 时，改变相序为下一种交通信号组合状态。在动作执行过程中将设置相序执行最长时间。若该相序在限制时长内动作状态没有发生改变，则智能体将自动切换到下一相序。

(2)可变相序[13,28,31]

智能体的动作集合 $A_m = \{A_{m1}, A_{m2}, \cdots, A_{mm}\}$。该集合表示智能交通信号系统所有可能执行的动作状态。智能体将执行该集合中概率最大的一个动作。虽然可变相序使得动作状态执行更加灵活，但是，由于下一阶段的不确定性会增加发生交通事故的概率，因此在实际的交通应用过程中通常不采用可变相序的智能体动作。

2.3 奖励函数 R

智能体执行相应动作后，环境反馈给智能体一个相应的值，这个值被称作奖励值。智能交通信号控制系统会根据该奖励值来进行网络参数更新。合适的奖励会对智能体的动作起到指导性作用。智能体通过选择合适的策略使得该交叉路口的拥堵程度最小。针对当前研究介绍几种常见的奖励函数：

(1)车辆的总等待时间[12,27,35]。该信息表示为车辆静止或速度小于某个阈值的时间。过长的车辆等待时间会严重影响交通效率，并且随着等待时间的增加，驾驶员负面情绪程度也会相应地增加。

(2)车辆队列总长度[12-14]。该信息表示该场景下所有车道上等待车辆的队列长度总和。队列长度的计算式如下：

$$L = \sum_{x=1}^{x} \sum_{y=1}^{y} D_{xy} \tag{1}$$

其中，D_{xy} 表示第 x 条街道上第 y 个车道上的车辆队列长度。

(3)系统延迟[12,28,36]。系统延迟表示车辆实际行驶速度与道路上允许的最高行驶速度之差和最高行驶速度的比值。系统延迟定义为：

$$f = 1 - \frac{v_q}{v_{\max}} \tag{2}$$

其中，v_q 表示车辆实际的行驶速度，$v_{\max}$ 表示该车道上车辆允许的最大行驶速度。

为了考虑实际交通场景以此获得更优异的动作状态。除上述关于一般车辆性质的奖励函数外，还将在奖励函数中考虑自行车、公交车和行人等因素[32]。不同的奖励对智能体的策略具有不同的影响，因此通常对不同的奖励赋予不同权重，并对奖励进行加权求和，表示为最终的奖励函数。

2.4 状态转移概率 P_{sa}

在 MDP 框架下，下一时刻的状态信息与之前多个时刻的状态信息有关。为简化系统模型，假设状态转化具有马尔可夫性，即下一时刻的状态信息只取决于前一时刻的状态信息。状态转移概率表示，在当前状态 S 下执行动作 A 后，转移到下一个状态 S' 的概率分布，记为 $P(S'|S,A)$。

2.5 策略 π 及折现因子 γ

在强化学习中，智能体通过在环境中的不断尝试采样来学习到一个最优策略 π。衡量一个策略的优劣取决于智能体执行该策略之后所得到的累计奖励值。为找到长期累积奖励，不仅要考虑当前时间步 t 的奖励，还需要考虑未来时间步的奖励。未来累计奖励 R_t 的计算式如下：

$$R_t = r_t + r_{t+1} + r_{t+2} + \cdots + r_n \tag{3}$$

其中，r_t 表示当前时间步 t 的奖励值。

在实际的交通场景中，环境状态在不断变化，导致下一状态也是随机的，因此无法保证下一次执行相同的动作。对未来探索得越多，会导致所产生的不确定性越多。为了避免这一问题，在实际应用过程中，引入折现因子 γ，使用 G_t 来替代未来累计奖励：

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^{n-t} r_n \tag{4}$$

其中，折现因子 γ 是介于[0,1]区间的常数，对于距离当前时间步 t 越远的奖励，其重要性就越小。

智能交通信号控制系统使用马尔可夫决策过程对系统进行建模。马尔可夫决策过程通过智能体与环境进行交互，在环境中进行不断的试错，来得到全局最优策略。

表 2 列出并阐述了针对智能交通信号控制系统的状态观测空间、动作状态空间以及奖励函数。

表 2　基于深度强化学习的智能交通信号系统模型

Table 2　Intelligent traffic signal system model based on deep reinforcement learning

文献	状态观测 S	动作状态 A	奖励函数 R
[7]	交通灯相位、出车道车辆数量、进车道车辆数量	可变相序	交叉路口压力差值
[13]	车辆数量、交通灯相位		平均队列长度
[27]	车辆位置、车辆速度、交通灯相位		平均等待时间
[28]	交通灯相位、检测器最后时间步占比		交通灯相位、系统延迟、平均等待时间
[31]	车辆队列长度		等待车辆队列长度、道路上移动的车辆数量
[33]	车辆位置、车辆速度		车辆累计延迟
[35]	第一辆车累计延时、进车道车辆数		车辆队列长度、车辆平均等待时间
[36]	路口图像的快照图		系统总延迟
[37]	路口图像的快照图		系统延迟、车辆等待时间、紧急停车、交通相位、瞬移惩罚
[38]	车辆位置、交通相位		车辆等待时间
[39]	交通灯相位、红灯/绿灯相位时长、左转车辆数量、剩余车辆数量		系统延迟
[12]	车辆队列长度、车辆数量、车辆位置、交通灯相位、车辆等待时间	固定相序	等待车辆总数、系统延迟、车辆等待时间、交通灯相位、固定步长内通过车辆数、车辆行驶时间
[14]	车辆数量、交通灯相位		车道队列长度总和
[29]	等待车辆数量、交通灯相位		队列长度总和
[32]	车辆位置、车辆速度、等待通过行人数量、交通灯相位、每条车道车队长度		车队排队长度、车辆总等待时间、系统延迟、交通灯相位、行人总等待时间、单位时间内通过车辆数
[34]	车辆数量、交通灯相位、当前相位持续时间、当前时刻、离路口最近车辆的距离		车道限速与实际车速差值
[40]	交通密度、平均速度、交通相位		单位时间内停在停车线前的车辆总数
[41]	交通相位、进入车道车辆数、接近车道的车辆分布		车辆队列长度
[42]	平均行驶时间、上游队列长度		所有上游车道队列长度总和以及车辆延迟、通过路口车辆数量、剩余车辆数量

3　智能交通信号控制面临的挑战及解决方法

针对目前智能交通信号控制系统所面临的挑战，研究人员采用不同的控制算法解决具体问题。本节将详细介绍基于深度强化学习的智能交通信号控制系统，并分析基础算法的优势以及局限性。针对状态信息获取不准确的问题，本节使用传统方法、机器学习方法和深度强化学习方法进行分析阐述。为了解决算法鲁棒性差的问题，通过单点交通控制系统-线性交通控制系统-区域交通控制系统的方法去分析算法的优势和不足。最后，对解决区域协调控制能力弱的问题采用传统方法和深度强化学习方法进行阐述。

典型的深度强化学习算法模型主要分为三大类，即基于值(Value-based)的方法、基于策略(Policy-based)的方法和同时基于值和策略的方法。基于值的典型算法有 Q-learning[43]，DQN[44]等；基于策略的典型算法有策略梯度[45](Policy Gradients，PG)等；同时基于值和策略的典型算法有 Actor-Critic[46]模型等。深度强化学习算法的对比如表 3 所列。

表 3　深度强化学习算法对比

Table 3　Comparison of deep reinforcement learning algorithms

算法类型	代表算法	算法机制	优势	不足	适用场景
基于值	Q-learning	使用 Q 表存储状态-动作对值，选择最大 Q 值对应动作	所需参数和数据量少	Q 表过大，查找和存储耗费大量计算资源	离散动作空间
	DQN	使用深度卷积神经网络逼近状态值函数	数据利用率高，减少更新方差	无法处理长时记忆问题，神经网络不易收敛，参数调整困难	离散动作空间
基于策略	PG	利用梯度下降法更新网络参数	可以在连续动作空间中选取动作	系统容易收敛到局部最优解	连续动作空间
基于值＋基于策略	AC	单步更新策略，Actor 根据值函数更新策略，Critic 根据动作计算值函数	更新速度快，有效减小方差	Critic 收敛困难，导致系统收敛困难	离散/连续动作空间
	A2C	在 AC 网络基础上，采用优势函数代替累计收益	有效减小方差	Critic 收敛困难，导致系统收敛困难	离散/连续动作空间
	DDPG	在 DQN 基础上，分别使用 2 个 Actor 和 2 个 Critic 网络训练当前网络和目标网络	样本利用率高，可在连续动作空间上有效学习策略	学习过程缓慢，Q 值估计过高，系统难以收敛	离散/连续动作空间

(1)基于值的方法(Q-learning 和 DQN)

Q-learning 是使用 Q 表存储每个动作—状态对的值。当状态空间和动作空间是高维或者连续时，Q 表会很大，导致查找和存储操作都需要消耗大量的时间和空间。DQN 的主要思想是使用深度神经网络来得到动作-状态 Q 值，并通过更新参数使 Q 函数逼近最优 Q 值。引入双网络结构以及经验回放单元结构，以有效地将深度学习和强化学习相结合。但是，DQN 算法只能处理短时记忆问题，并不能处理长时记忆问题。

(2)基于策略的方法(PG)

基于策略的强化学习方法的主要思想为：如果执行该动作使得最终奖励值变大，将增加该动作出现的概率，否则减少该动作出现的概率。策略梯度算法调整的是动作出现的

概率，并没有给出相应的动作-状态值。但策略梯度容易收敛到局部最优解，而非全局最优解。

(3)基于值和策略的方法(AC，A2C 和 DDPG)

AC 网络充分结合了策略梯度算法和 Q-learning 算法。策略梯度的更新是基于回合制的，并且需要大量的训练样本。Q-learning 输出为每个动作的状态值，要求动作状态必须是离散的。AC 网络将两者结合，在实现离线学习的同时，可以执行连续型动作状态。A2C 在 AC 网络的基础上，采用优势函数代替累计收益，能够有效减小方差。基于 AC 算法提出了深度确定性策略梯度算法(Deep Deterministic Policy Gradient，DDPG)[47]，该算法主要是为了实现连续动作空间下的深度强化学习。

3.1 状态信息获取不准确

智能交通信号控制系统能够观测到当前所有车辆的状态信息和当前自身的状态信息，根据观测状态信息去执行该状态下的最优策略。但是，由于智能体观测环境状态是时变的、非平稳的，智能体无法准确获取周围其他智能体的状态信息，只有通过间接的观察来感知状态，从而采取相应的动作。因此，可以将智能交通信号优化配时问题看作一个部分场景可观测的马尔可夫决策问题(Partially Observable Markov Decision Process，POMDP)[48]。智能体需要根据全域与部分区域的观测结果来推断当前状态的分布情况，这导致直接解决 POMDP 问题是十分困难的。

传统方法使用保持状态信息来估计状态的概率分布，一些基于记忆的方法使用递归 Q 学习方式来解决 POMDP 问题。一些其他的方法使用局部优化的方式来找到近似解[49]，但局部优化并不能保证全局最优。文献[34]提出了基于 DSRC(Dedicated Short-Range Communication)的部分车辆检测的智能交通信号控制系统。通过将部分车辆设置为不可观测车辆，将环境定义为部分场景可观测环境，并定义新的状态空间及奖励函数。由于 DSRC 能力的限制，在部分检测 ITS 中，只有部分车辆能够被检测到。因此，需要获得关于这些车辆更具体的信息，将状态空间设置为车辆数量、最近车辆数量、当前相位时间、当前相位和当前时间。奖励函数不再通过各个奖励值进行加权求和，而是将奖励函数设置为最大化交通信号 U 状态下的速度 $v_U(t)$，具体表达式如下：

$$r_t=-\frac{1}{v_{\max}}[v_{\max}-v_U(t)] \tag{5}$$

其中，$v_U(t)$ 表示最大化交通信号 U 状态下的速度，$v_{\max}$ 表示该车道上车辆允许的最大行驶速度，如果不同车辆的 $v_{\max}$ 不同，则取奖励函数为所有车辆的函数的算术平均值。该实验场景分别设置在 1×5 的线性路口和 4×4 的路网结构上。设置车辆检测率从 0～100%变化，并在稀疏、中等、密集车流场景下进行仿真实验。实验结果表明，该智能交通信号控制系统在解决部分场景可观测的问题上具有较好的性能。

机器学习的方法将智能交通信号控制问题表述为 POMDP 模型[50]。通过迁移学习[37]或将算法扩展到动态集群[51]等方法，来提高系统的观测性。文献[50]应用 PG 方法来保证在部分可观测环境下的局部收敛，但该方法难以达到全局收敛状态。文献[37]采用 DQN 算法，通过迁移学习缓解了部分可观测性问题，但是该研究模拟的交通环境过于简单。文献[51]将线性回归的 IQL 扩展到动态集群区域来提高可观测性。文献[52]通过共享邻居信息提高了独立 SARSA 学习算法的可观测性。

深度强化学习方法通常使用 LSTM 来解决部分场景可观测问题。文献[40]提出了深度循环 Q 网络(Deep Recurrent Q-network，DRQN)算法，该算法将 DQN 中的一个全连接层替换为了长短时记忆网络(Long-Short Term Memory，LSTM)[53]。LSTM 是一种特殊的循环神经网络(Recurrent Neural Network，RNN)，结构图如图 2 所示，其通过门结构来去除或增加记忆单元所储存的交通状态信息，选择性地保留对当前交通状态影响较大的交通信息。与此同时，丢弃那些对当前交通状态几乎不产生影响的交通信息，从而避免输入序列过长造成的信息缺失问题。文献[33]将 DRQN 与 RNN 相结合，在解决部分场景可观测问题的同时，缓解了标准强化学习中目标 Q 值过高和长期经验学习不足的问题。在此基础上，文献[32,35]分别在 MADDPG 和 A2C 的基础上加入 LSTM 结构，来解决部分场景可观测下的交通信号控制问题。通过消融实验结果可知，加入了 LSTM 结构的方法在解决部分场景可观测问题上取得了较好的效果。

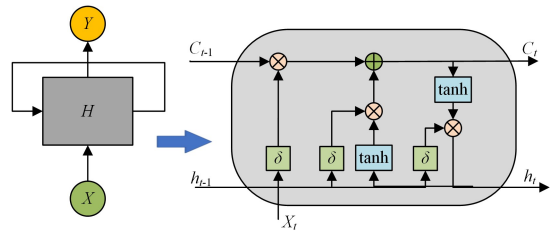

图 2 LSTM 结构图

Fig. 2 Framework of LSTM

综上，现有解决智能交通信号控制中部分场景可观测问题的传统方法提高了系统的观测能力，但并没有很好地解决部分场景可观测的问题。文献[34]虽然提出了一种基于 DSRC 的新型研究方法，并在部分场景可观测的情况下取得了较好的效果，但是，由于路口设备安装等问题，目前算法的实际实施还存在着较大问题。目前基于深度强化学习方法解决部分场景可观测问题采用的主要方法为，在网络结构中加入 RNN 或 LSTM 等网络结构。但是，LSTM 本身也具有参数较多、计算复杂度较高等缺点。

3.2 控制算法鲁棒性差

大多数深度强化学习算法应用于交通信号控制领域中时，实验场景大多集中在单个交叉路口或小型路网结构。深度神经网络的加入提高了强化学习的可扩展性，但训练一个集中式强化学习智能体对于大规模的 ATSC 是十分困难的。在实际应用过程中，集中处理这些状态信息会产生高时延和高故障率，这将导致交通流量信息丢失。在一个大规模城市中，有成千上万个车辆探测器和数量庞大的交通信号灯，由此产生了巨大的状态空间和动作空间。传统强化学习方法难以处理上述问题且算法可扩展性差。研究人员针对单点交通信号控制-线性交通信号控制-区域交通信号控制进行研究，解决了算法鲁棒性较差的问题。这种联合作用的思想已经在 TSC 中得到应用。

为解决可扩展性问题，文献[54]提出了 IQL 方法。该方法

中的每个智能体将其他智能体作为观测状态的一部分，独立执行当前状态下的最优策略。因此，该方法具有良好的可扩展性。然而，其他智能体的状态和策略在不断发生变化，使得环境变得部分可观测，导致该方法在使用过程中难以收敛。文献[55]在 IQL 中通过增加记忆回放单元，解决了 IQL 方法导致的 Q 值估计过高的问题。

文献[38]提出了一种网络级分散自适应信号控制算法，即 3DQN(Double Dueling Deep Q network)，解决了单点交通控制系统到线性交通控制系统鲁棒性问题。对于每个独立的交叉口，采用基于高分辨率实时交通数据的强化学习智能体控制交通信号。为了实现更好的网络性能，智能体通过共享交通信息和信号状态信息来相互协调。该网络结构采用两个卷积层和两个池化层将状态输入整合为一个向量。设置冻结期，将不会选择动作空间内的部分动作状态，以此提高系统的鲁棒性。该实验在一个基于真实场景(限制通行的高速公路(SR 417)和其他几条平行或相连的干线)进行实验。实验结果表明，该控制系统能够有效减少车辆行驶时间以及系统总延迟。

文献[35]在 IQL 的基础上提出了 IA2C 方法，旨在解决单点交通控制系统到区域交通控制系统的扩展性问题。由于 IA2C 方法是完全可扩展的，在部分场景可观测的情况下，采用两种方法来稳定 IA2C，以提高系统的鲁棒性及最优性。每个智能体能够获取其他智能体的最新策略，将相邻智能体的策略与智能体当前状态观测信息一同输入到深度神经网络(Deep Neural Network，DNN)中。另外，通过引入新的空间折扣因子来降低来自远处智能体信息的影响，定义新的奖励函数 $\tilde{r}_{t,i}$。

$$\tilde{r}_{t,i}=\sum_{d=0}^{D_I}\Big(\sum_{j\in\nu\mid d(i,j)=d}\alpha^d r_{t,j}\Big) \tag{6}$$

其中，引入空间折现因子 α 来调整智能体的全局奖励函数，d 为任意两个智能体之间连接它们的最小边数。与多个智能体之间共享相同的全局奖励相比，加入空间折扣因子的全局奖励权衡更为灵活。该实验在 5×5 的网格化结构和大型真实交通网络摩纳哥城市场景下进行仿真实验，实验结果表明该算法在队列长度、车辆等待时间等方面均优于非稳定下的 IA2C。

文献[31]提出了区域级到区域级的扩展方法，其主要思想是将大规模场景下的交通信号控制系统分解成多个相对较为容易的小型交通信号控制系统。通过 Per-Action DQN 算法和 Wolpertinger 算法分别在 2×3 的子网络系统上进行对比仿真，结果表明 Per-Action DQN 算法具有更好的性能指标。使用区域到区域的扩展方法，分别在 2 个子网络的 4×3 网格和具有 4 个子网络的 4×6 网格下进行仿真实验。每个子网络学习自己的策略并执行动作，然后使用集中式方法将不同区域内的智能体进行聚合，最终得到该区域的最优控制策略。

文献[38]提出的算法虽然在线性交通网络中表现出了良好的性能，但由于计算资源的限制，算法迁移性差，在应用于大型区域网络时还存在着一些问题。文献[35]有效地解决了从单点控制系统到区域控制系统算法可扩展性问题，但该算法难以确定合适的空间折现因子来减弱来自其他智能体的状态和奖励信息。文献[31]在完全可观测场景下进行是不切实际的，并且此模型采用集中式的控制方式，随着子网络个数的增多，状态空间的增大，集中式方法最终很难达到全局收敛状态，并没有完全解决区域到区域的可扩展性问题。

基于上述研究，针对智能交通控制系统鲁棒性差的问题，采用从单点交通信号控制-线性交通信号控制-区域交通信号控制进行研究。应用于解决可扩展性的主要方法采用独立学习的方法(IQL，IA2C 等)。此类方法是完全可扩展的，将其他智能体作为环境观测状态，并且不与其他智能体进行信息交互，但是此类方法在解决算法可扩展性差的同时，难以解决全局最优策略的问题。

3.3　区域协调控制能力弱

传统的方法在解决大规模场景下的全局最优问题时，通常将全局优化目标分解成多个子任务进行优化，并求取局部最优解。智能交通信号控制是不可微的，导致传统求解局部最优的算法无法使用，并且随着智能体数量增多、维度升高和局部最优解个数的增多，难以找出全局最优解。为解决这一问题，需要了解相邻智能体的状态信息和动作信息。

文献[56]采用 Q 函数作为信息传递。文献[45]采用 max-plus 方法进行信息传递。这些方法通过信息传递来获取相邻智能体的信息。但是，此类方法需要额外的计算来获得执行过程中的协调控制，并且需要启发式方程和假设来分解 Q 函数并制定消息传递策略。但是，这些信息可能会被算法本身所学习，近而影响其执行最优策略。文献[57]考虑使用协调图来学习主体之间的显式协调机制，但利用协调图方法得到的解是针对路网结构的。当这些特性发生变化时(例如，当增加新的交叉点，道路结构会发生改变)，必须重新创建当前的协调策略。集中式方法可以缓解这种情况，但会导致大规模的优化问题，尤其是在实际场景下，这些问题在计算上往往是不可行的。文献[41]提出，在没有明确协调策略的情况下，基于序列信息和迁移学习的强化学习方法也可以达到协调的效果，从而实现全局最优策略。由于上述方法都需要整个网络中智能体之间进行相互协调，因此计算代价很高。

尽管 DDPG 算法已经十分成熟，并且在交通场景中应用广泛[58]，但该算法自身也存在着一些不足之处。DDPG 以 AC 算法为基础，Critic 收敛困难将会导致系统收敛困难，并且 DDPG 算法对参数设置十分敏感，不合理的参数设置可能会导致系统不收敛。在实际应用场景下，由于各路口之间存在紧密的关系，解决单交叉路口的交通拥堵问题是不具有现实意义的。在 DDPG 的基础上，文献[32]提出了多智能体 DDPG(Multi-Agent Deep Deterministic Policy Gradient，MADDPG)，将 DDPG 算法扩展到多智能体。MADDPG 算法的结构如图 3 所示。

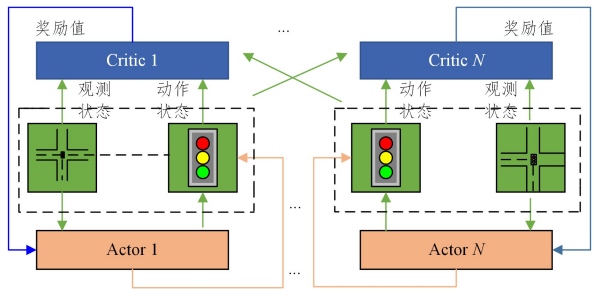

图 3　MADDPG 结构图

Fig. 3　Framework of MADDPG

各个智能体可获取该区域内其他智能体当前的状态信息以及所执行的策略。然而，将 MADDPG 应用于多智能体交通信号控制过程中存在不足之处。随着交通网络规模的增大，各智能体输入的状态空间和动作空间也会相应地增大，训练学习时间也会变长。此外，当智能体数量过多时，单个智能体需要考虑所有其他智能体的状态。然而，远离该交叉口的其他交叉口状态并没有对该交叉口的策略产生较大的影响，因此过多的无用信息会导致智能体的学习效率降低。最终，使该算法不能达到大规模场景下的全局收敛状态。

文献[59]提出了一种基于双估计器和上置信界(Upper Confidence Bound，UCB)策略的独立双 Q 学习方法——Co-DQL。该算法在保证探索性的同时，消除了传统 IQL 算法中 Q 值高估问题。利用平均场近似来进行智能体之间的交互，从而使智能体学习到更好的合作策略。为了提高学习过程的稳定性和鲁棒性，引入了一种新的奖励分配机制和局部状态共享方法。尤其在多智能体强化学习过程中可能会存在信用分配问题。因此，每个智能体往往不直接将全局奖励作为自己的奖励。具体奖励分配机制函数定义为：

$$\hat{r}_k = r_k + \beta * \sum_{i=N(k)} r_i \tag{7}$$

其中，β 作为策略因子来平衡合作程度，r_k 和 r_i 分别表示第 k 个智能体的奖励值和全局奖励值，$\hat{r}_k$ 为经过奖励分配机制反馈给第 k 个智能体的奖励值。以西安部分地区的路网作为真实路网进行实验。实验结果表明，在平均系统延迟、平均车辆速度和平均队列长度等指标方面，Co-DQL 均优于其他去中心化多智能体深度强化学习算法。

采用深度强化学习的方法解决区域协调能力弱的主要思想为，将其他路口的状态信息作为道路的辅助属性特征，与当前路口信息相融合作为模型的输入，并与其他智能体进行信息交互。集中式系统可以有效解决区域协调能力弱的问题，但随着状态信息的增加，集中式系统无法解决算法可扩展性差的问题。文献[59]提出采用新的奖励分配机制去解决区域协调能力弱的问题。由于交叉路口不同的路网结构，每个智能体的奖励将出现多次，这取决于相连接的相邻交叉路口的数量。例如，具有 5 条路的交叉口比具有 3 条路的交叉口接收更多的权重，可能会导致有偏执的最优解。

针对目前智能交通信号控制系统所面临的挑战，可采用 RNN，LSTM 或 GRU 等模型来获取时间相关性，通过分析前一时刻状态信息来解决信息获取不准确的问题。采用分布式算法，将其他路口的状态信息作为道路的辅助属性特征，与当前路口观测状态信息相融合作为模型的输入，与其他智能体进行信息交互，以解决控制算法鲁棒性差和区域协调控制能力弱的问题。表 4 总结了深度强化学习算法解决智能交通信号控制所面临的挑战。

表 4 智能交通信号控制系统解决方法

Table 4 Intelligent traffic signal control system solution

解决问题	文献	采用算法	解决方式
状态信息获取不准确	[36]	PG	传统方法
	[37]	迁移学习	
	[51]	扩展到动态集群	
	[32]	MADDPG+LSTM	深度强化学习方法
	[33]	DQN+LSTM+RNN	
	[34]	DQN+优先经验回放	
	[35]	A2C+LSTM	
	[40]	DQN+LSTM	
可扩展性(单点控制系统-线性控制系统)	[7]	DQN+max-pressure	深度强化学习方法
	[29]	DQN	
	[37]	DQN+max-plus	
	[38]	Double dueling DQN	
	[57]	DQN+迁移学习	
可扩展性(单点控制系统-区域控制系统)	[12]	DQN+相位门+记忆宫殿	
	[13]	分布式 Ape-X DQN+FRAP	
	[14]	DQN+定义新奖励函数	
	[35]	IA2C	
	[59]	独立 Double Q-learning	
可扩展性(区域控制系统-区域控制系统)	[31]	每动作 DQN+区域合作	
	[51]	近似 Q 学习+局部区域合作	
区域协调能力弱	[56]	使用 Q 函数传递信息	传统方法
	[57]	使用 max-plus 传递信息	
	[27]	IQL+获取其他智能体策略信息	深度强化学习方法
	[32]	DDPG+获取其他智能体状态策略信息	
	[35]	引入空间折现因子+获取其他智能体状态策略信息	
	[41]	DQN+获取相邻智能体状态信息+迁移学习	
	[59]	DQN+UCB+共享智能体之间的状态信息	

4 仿真平台及实验分析

用于智能交通信号控制的深度强化学习模型需要依靠大量的数据进行训练。其主要目的是从训练场景中学习到交通信号控制策略，并将学习到的策略应用于智能交通信号控制场景中。

本节将主要对实验所用的应用场景、仿真数据、仿真平台以及评价指标进行概述。

4.1　应用场景

智能交通信号控制系统已经在全球数百座城市中得到应用。本节将介绍几种应用于智能交通信号控制系统的交通网络结构。

(1)单路口交通网络

单路口交通系统[36,39,42]能够产生低维度的状态空间及动作空间，从而获得最优策略。

(2)多路口线性交通网络

多路口线性交通系统是由两个及以上交叉路口组成的线性封闭交通系统[37]。在该场景下，智能交通控制系统根据线性路口的车流程度着重研究主干道及小支路车道的情况。

(3)网格化交通网络

网格化交通网络是基于网格的网络结构交通系统。例如，3×3 路网[41]、5×5 路网[35]等规则网格化结构。区域范围内的交通信号控制，不但需要考虑单个交叉口的最优策略，而且需要各路口相互协调控制以得到全局最优策略。

(4)真实场景交通网络

现实世界的交通网络是基于城市布局网络的交通系统，因此需要考虑更多的交叉路口，例如美国佛罗里达州[38]、中国济南[12]和摩纳哥城市[35]。

研究人员通常在网格化交通网络或真实场景交通网络结构下使用合成数据或真实数据对训练完成的模型进行评估。在评估过程中，评估数据使用一个数组(O,t,D)表示。其中，O,D代表车辆的起点和终点，t代表车辆的行驶时间。评估指标主要包括平均奖励、车辆的平均等待时间、交叉路口平均延时、平均队列长度和车辆平均速度等。

4.2　仿真数据及评估指标

一些常见的研究不仅使用真实的路网结构，还会根据公开的交通数据情况设置相应的模拟参数，常见的数据集如下。1)济南 24 个路口的交通信息。该数据集包含超过 4.05 亿辆汽车，近 2 000 个摄像头进行为期 31 天的信息收集。数据集中每条记录包含时间、摄像头 ID 和车辆信息。通过摄像机位置分析记录信息得出车辆通过交叉路口的轨迹信息。2)美国佛罗里达州的 8 个路口场景下的交通数据，该场景为高速公路下的测试场景。由于在该场景内行人的活动效率极低，因此在该场景下行人的因素并没有被考虑其中。3)摩洛哥城市交通场景包括 30 个路口、11 个两相位交通信号灯、4 个三相位交通信号灯、10 个四相位交通信号灯、1 个五相位交通信号灯和 4 个六相位交通信号灯。

4.3　仿真平台

交通模拟器是用于验证和评估基于深度强化学习交通控制系统的仿真平台。在大多交通控制系统的研究中，通常使用基于微观方法的交通模拟器，包括 SUMO[60]，CityFlow[61]，VISSIM[62]和 Paramics[63]等。在已知的智能交通信号控制系统的研究中，大多采用 SUMO 仿真平台。SUMO 可以针对每辆车单独进行控制，因此更适合交通控制模型的开发；CityFlow 作为近几年新兴的交通仿真模拟器，与 SUMO 相比处理速度更快，可用于合成数据或真实数据进行多智能体交通信号仿真；VISSIM 是一个离散的、随机的仿真模拟器，其时间步长为 0.1 s。该系统中存在可用于检测数据和交通信号控制进行信息交互的接口；Paramics 具有较完善的路网结构和丰富的编程接口。目前，用于智能交通信号控制的交通模拟仿真器，通常采用泊松分布、正态分布等方法随机生成车辆。此类方法不能反映真实的交通情况，如早晚高峰、工作日和周末等时间的交通情况。基于真实路网构建网络结构将是后续研究工作的重点方向。表 5 总结了智能交通信号控制系统的仿真平台、路网结构、常用数据集以及评估指标。

表 5　智能交通信号控制系统仿真平台、路网结构及评估指标

Table 5　Simulation platforms, road network structures and evaluation indicators of intelligent traffic signal control system

文献	仿真平台	路网结构	使用数据集	评估指标
[12]	SUMO	单路口、济南(24 个路口)	济南(24 个路口)	平均奖励函数、平均队列长度、平均系统延迟、平均行驶时间
[14]		中国两个城市、美国洛杉矶、3×4 路网、4×4 路网、1×4 路网	中国两个城市数据集、美国洛杉矶	平均行驶时间
[27]		2×2 路网	无	平均奖励值、系统延迟、平均延迟、累计延迟
[31]		2×3 路口、4×6 路口、4×3 路口	无	平均等待时间、平均队列长度
[32]		1×2 路网、2×3 路网	无	平均奖励值、平均队列长度、平均系统延迟、平均行驶时间、平均行人数量
[33]		单路口	无	平均等待时间
[35]		5×5 路口、Monaco	Monaco 城市数据	平均奖励值、平均队列长度、平均系统延迟、平均车辆速度
[36]		单路口	无	平均奖励值、平均累计延迟、平均队列长度
[37]		1×2 路网、1×3 路网、2×2 路网	无	平均行驶时间
[51]		10×10 路网	无	交通密度分布
[40]		单路口	无	平均等待时间
[41]		单路口、1×10 路网、3×3 路网	无	平均行驶时间
[59]		4×4 路网、西安场景	西安城市数据	平均奖励值、平均系统延迟、平均车辆速度、平均队列长度、车辆到达率
[7]	CityFlow	6 路口、10 路口、20 路口、3×3 路网、济南青岛路、美国比弗大道、美国麦哈顿	济南青岛路、美国比弗大道、美国麦哈顿	平均行驶时间
[13]		3 叉单路口、4 叉单路口、5 叉单路口、3×4 路网、4×4 路网、1×5 路网、济南、杭州、Atlanta	济南、杭州、Atlanta(5 个路口)	平均行驶时间
[18]		3 个城市	3 个城市的真实数据	平均行驶时间

（续表）

文献	仿真平台	路网结构	使用数据集	评估指标
[38]	Aimsun Next	佛罗里达州场景	MVDS 和 ATSPM	平均行驶时间、平均系统延迟、车流量
[58]		单路口、2×3 路口、西班牙(43 个路口)	西班牙城市数据	平均奖励值
[42]	Vissim	单路口	无	平均队列长度
[39]		单路口	无	平均系统延迟、平均吞吐量

4.4 仿真实验分析

本文对文献[35]中的源代码进行复现，该文献采用 IA2C 的方法解决了算法鲁棒性差的问题。通过加入 LSTM 模型解决部分场景可观测问题，将相邻路口的状态信息作为道路的辅助属性特征，与当前路口信息相融合作为模型的输入，并引入空间折现因子来稳定 IA2C，以达到执行全局最优策略的目的。

定义状态观测信息为：

$$s_{t,i}=\{wait_t[l],wave_t[l]\} \tag{8}$$

其中，l 是交叉口 i 的进车道，$wait$ 表示第一辆车的累计延迟，而 $wave$ 代表距离目标交叉口 50 m 内每个进车道的车辆总数。

定义奖励函数为：

$$\gamma_{t,i}=-\sum(queue_{t+\Delta t}[l]+a*wait_{t+\Delta t}[l]) \tag{9}$$

其中，a 是一个折现系数，$queue$ 是每个进车道的队列长度。

在 SUMO 仿真模拟器中，对 5×5 网络结构中的 IA2C 算法和 MA2C 算法进行模型训练。两种算法的训练步长均设置为 1M。训练过程中，每辆车的运行路线在路网中随机生成。训练平均奖励值如图 4 所示。图 4 中，阴影部分表示其标准差，实线部分为拟合后的曲线。从图中可看出，MA2C 具有较好的学习能力和鲁棒性，训练曲线先稳定增长后趋于稳定。IA2C 和 MA2C 的平均奖励值为−1 139 和−749，MA2C 显然优于 IA2C。

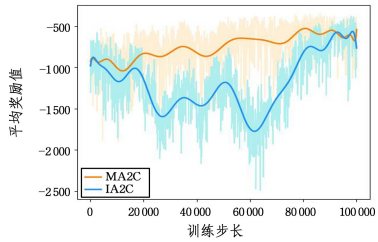

图 4 5×5 路网下的训练曲线图

Fig. 4 Reward during all training episodes in 5×5 network

将平均队列长度和平均系统延迟作为评估指标，对模型进行性能评估。图 5 给出了每个仿真步骤下的平均队列长度，IA2C 无法依靠可持续的策略来恢复拥塞的交通网络，而 MA2C 倾向于更稳定和可持续的策略，通过在网络早期负载较轻时支付较高的队列成本，来实现更低的拥塞程度和更快的恢复速度。

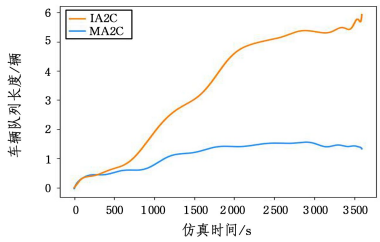

图 5 5×5 路网下的平均队列长度

Fig. 5 Average queue length in 5×5 network

图 6 给出了每个仿真步骤下的系统延迟。IA2C 算法无法从拥堵情况下快速恢复，导致延迟单调增加，而 MA2C 即使在峰值时也能将延迟维持在较低水平。

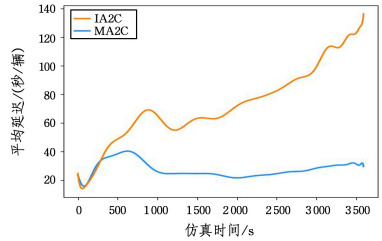

图 6 5×5 路网下的平均系统延迟

Fig. 6 Average system delay in 5×5 network

在 5×5 网格的基础上，本文针对真实数据集——摩洛哥城市交通场景进行了仿真实验。同样，设置算法训练步长均为 1 M，并增加了两种算法的对比实验，分别为在 IQL 基础上使用线性回归方法(IQLl)和在 IQL 基础上使用神经网络方法(IQLd)。训练平均奖励值如图 7 所示。从图中可以看出，IA2C 和 MA2C 算法系统得到较好的收敛效果，MA2C 显示出了更快、更稳定的收敛。

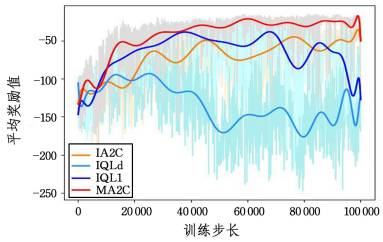

图 7 摩洛哥场景下的训练曲线图

Fig. 7 Training curve in Monaco traffic network

在摩洛哥的场景下，将平均队列长度作为评估指标，对模型进行性能评估。图 8 给出了每个仿真步骤下的平均队列长度。IA2C 和 MA2C 都能够减小峰值后的队列长度，并且 MA2C 实现了更好的恢复率。MA2C 能够通过协调共享相邻交叉口的信息，在交叉口之间更均匀地分配交通信息，从而实现更短的队列长度。

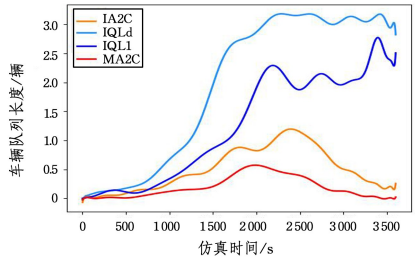

图 8 摩洛哥场景下的平均队列长度

Fig. 8 Average queue length in Monaco traffic network

5 研究展望

本文主要针对基于深度强化学习下的智能交通信号控制

策略进行综述和分析。根据状态信息获取不准确、控制算法鲁棒性差和区域协调控制能力弱的挑战来介绍当前该方向的研究现状。在实际应用智能交通信号控制系统过程中还存在着许多值得关注的问题。本节将讨论未来智能交通信号控制系统可能的发展方向。

5.1　基于图神经网络的智能交通信号控制

本文介绍了基于深度强化学习的智能交通信号控制策略。相邻路口之间的状态信息相互影响,但影响程度不同,因此不能将所有相邻交叉路口设置为同一等级。在此基础上可以引入图注意力网络来解决智能交通信号控制问题。在图神经网络中,使用边和节点构建图结构,将交叉路口表示为节点,两路口之间的道路表示为边。根据相邻节点对中心节点的影响为不同的边赋予不同的权重。使用图神经网络来捕获时空相关性,通过空间相关性协调各个路口之间的关系,并根据历史数据的时间相关性来解决部分可观测场景下的交通拥堵问题。实际场景下,路网结构是天然的图结构。同时,由于图神经网络优异的性能,将智能交通控制系统与图神经网络方法相结合,有望成为以后研究的重点。

5.2　多目标奖励函数优化

多目标决策是对多个相互矛盾的目标进行科学、合理的选优,并做出相应的决策。在当前方法中,奖励函数通常为几个不同指标的加权和,其中权重是根据经验设置的。在多目标强化学习领域中是以更可控的方式来找到这些权重。使用熵权法能够更有效地找到各奖励权重。熵权法的基本思路是,通过指标的重要程度确定客观的权重。使用极差标准化、均方差标准化等方式将多个指标进行归一化处理以求取信息熵值。因此,可以根据信息熵值求取各个指标对应的权重信息。

5.3　算法复杂度分析

由于大多数状态矩阵是道路信息,不包含车辆,因此在状态矩阵中总是用零表示。状态信息使用的位置矩阵和速度矩阵是稀疏的,这种存储信息的方法会增加系统的空间复杂度。利用矩阵稀疏性的方法可以有效缩短计算时间。空间稀疏卷积网络[64]能够较好解决矩阵稀疏问题。在传统的卷积神经网络中,任意输入神经元与任意输出神经元均存在连接。空间稀疏卷积网络则将部分存在状态信息的输入神经元与输出神经元相连接,这意味着需要存储的参数变少,计算复杂度与空间复杂度将会降低。

5.4　交通干扰对算法精准性的影响

在实际的深度强化学习算法部署过程中,交通信号控制系统必须能够抵御意外交通干扰,如恶劣天气条件、道路事故和施工等情况。当发生意外交通干扰时,可能会出现状态信息收集不准确的情况。为了避免此类情况,可采用 LED 灯进行辅助通信。LED 灯通信不会受到暴雨等恶劣天气的影响,并且具有高速率和高带宽等优点。此外,只有处在光线传播直线上的人才可能截获信息,因此采用 LED 灯通信安全性较高。同时,使用无人机辅助进行交通信息的收集,相比传统的信息收集方式,其具有检测范围广、检测视角灵活和检测效率高等优点。同时,在发生交通事故造成交通拥堵时,无人机可以更加迅速地到达事故现场。将无人机与传统信息收集方式相结合能够更准确且更及时地汇集交通信息。在面对交通干扰时,采集交通信息会相对准确。

5.5　安全性问题

深度强化学习算法应用到实际的交通信号控制系统中,还存在着许多实际的安全性问题。交通信号灯故障会增加交通事故发生的概率,对交通信号控制进行安全管理是十分必要的。将每个行动都与一个风险因素相关联,然后设计相应规则,从一组可行动作中排除高风险的行动。在初始阶段,通过仿真对不同动作的风险因素进行探索和验证,并在实际应用过程中随着时间推移进行改进,使学习成本最小化。这在未来的研究过程中可以提高智能交通信号控制的安全性。在多个交叉路口协调过程中,安全性主要包含两个方面:一方面为各路口之间信息传递的安全性;另一方面为智能交通信号控制系统稳定后,如何抵御外来信号攻击的安全性。

5.6　实际交通场景应用

现有的智能交通信号控制系统大多基于仿真或数据集进行训练和测试,而在实际场景中的应用是不充分的。在实际的应用场景中,可以将智能交通信号控制策略与交通流预测、路径优化[65]等进行结合。根据实时的交通控制策略,为每辆车提供最优路径策略,从而减少交通拥堵情况的发生。随着 6G 网络[66]的快速发展,未来的研究将更专注于如何将智能交通信号系统应用于实际场景中,为智能交通系统提供高质量的交通管理、交通服务和交通资源配置。将智能交通信号控制系统与其他方面相结合,能够加速智能交通信号控制系统在实际生活中的应用。

结束语　本文主要介绍了智能交通信号控制系统的发展现状,对智能交通信号控制系统当前存在的挑战(状态信息获取不准确、控制算法鲁棒性差和区域协调控制能力弱)进行了总结分析。全面地综述了当前主流的基于深度强化学习解决交通信号控制挑战的策略方法。目前深度强化学习算法取得了一定的效果,但是仍有一些问题需要解决。最后,本文提供了智能交通信号控制系统在图神经网络、交通信息收集、安全性问题、实际交通场景的应用等未来可能的发展方向。智能交通信号控制系统的应用不仅可以解决实际的交通拥堵问题,还有望为车辆的路径规划、交通预测等问题提供坚实的基础。

参 考 文 献

[1]　公安部交通管理局.《今年上半年新注册登记机动车 1 871 万辆》[EB/OL].(2022-01-11). https://www.mps.gov.cn/n2254314/n6409334/c8322353/content.html.

[2]　DIAO M,KONG H,ZHAO J,et al. Impacts of transportation networkcompanies on urban mobility[J]. Nature Sustainability,2021,4(6):494-500.

[3]　SUN H,CHEN C L,LIU Q,et al. Traffic Signal Control Method Based on Deep Reinforcement Learning[J]. Computer Science,2020,47(2):169-174.

[4]　VARAIYA P. The max-pressure controller for arbitrary networks of signalized intersections [M]. Springer: Advances in Dynamic Network Modeling in Complex Transportation Systems,2013:27-66.

[5]　ALI M E M,DURDU A,ÇELTEK S A,et al. An adaptive method for traffic signal control based on fuzzy logic with webster and modified webster formula using SUMO traffic simulator

[J]. IEEE Access, 2021, 9: 102985-102997.
[6] SHI Y, LI J, HAN Q, et al. A Coordination Algorithm for Signalized Multi-Intersection to Maximize Green Wave Band in V2X Network[J]. IEEE Access, 2020, 8(3): 213706-213717.
[7] WEI H, CHEN C, ZHENG G, et al. Presslight: Learning max pressure control to coordinate traffic signals in arterial network [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 1290-1298.
[8] WEI H, ZHENG G, GAYAH V, et al. A survey on traffic signal control methods[J]. arXiv: 1904. 08117, 2019.
[9] SRINIVASAN D, CHOY M C, CHEU R L. Neuralnetworks for real-time traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(3): 261-272.
[10] MANANDHAR B, JOSHI B. Adaptive traffic light control with statistical multiplexing technique and particle swarm optimization insmart cities[C]// 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS). IEEE, 2018: 210-217.
[11] SÁNCHEZ-MEDINA J J, GALÁN-MORENO M J, RUBIO-ROYO E. Traffic signal optimization in "La Almozara" district in saragossa under congestion conditions, using genetic algorithms, traffic microsimulation, and cluster computing[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 11(1): 132-141.
[12] WEI H, ZHENG G, YAO H, et al. Intellilight: A reinforcement learning approach for intelligent traffic light control[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2496-2505.
[13] ZHENG G, XIONG Y, ZANG X, et al. Learning phase competition for traffic signal control[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1963-1972.
[14] ZHENG G, ZANG X, XU N, et al. Diagnosing reinforcement learning for traffic signal control [J]. arXiv: 1905. 04716, 2019.
[15] GRONAUER S, DIEPOLDK. Multi-agent deep reinforcement learning: a survey [J]. Artificial Intelligence Review, 2022, 55(2): 895-943.
[16] TAMPUU A, MATIISEN T, KODELJA D, et al. Multiagent cooperation and competition with deep reinforcement learning [J]. PloS One, 2017, 12(4): e0172395. https://doi. org/10. 1371/journal. pone. 017 2395.
[17] ZHANG K, YANG Z, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[J]. Handbook of Reinforcement Learning and Control, 2021, 325(2): 321-384.
[18] XIONG Y, ZHENG G, XU K, et al. Learning traffic signal control from demonstrations[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 2289-2292.
[19] FOERSTER J, ASSAEL I A, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning [C]// Advances in Neural Information Processing Systems. 2016: 2145-2153.
[20] TAN M. Multi-agent reinforcement learning: Independent vs. cooperative agents[C]// Proceedings of the Tenth International Conference on Lachine learning. 1993: 330-337.
[21] RASHID T, SAMVELYAN M, SCHROEDER C, et al. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning[C]// International Conference on Machine Learning. PMLR, 2018: 4295-4304.
[22] IQBAL S, DE WITT CAS, PENG B, et al. Randomized Entity-wise Factorization for Multi-Agent Reinforcement Learning [C]// International Conference on Machine Learning. PMLR, 2021: 4596-4606.
[23] PANDEY D, PANDEY P. Approximate Q-learning: An introduction[C]// 2010 Second International Conference on Machine Learning and Computing. IEEE, 2010: 317-320.
[24] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: A brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.
[25] LEI L, TAN Y, ZHENG K, et al. Deep reinforcement learning for autonomous internet of things: Model, applications and challenges [J]. IEEE Communications Surveys & Tutorials, 2020, 22(3): 1722-1760.
[26] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[27] LIU M, DENG J, XU M, et al. Cooperative deep reinforcement learning for traffic signal control[C]// 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD). Halifax. 2017.
[28] SCHUTERA M, GOBY N, SMOLAREK S, et al. Distributed traffic light control at uncoupled intersections with real-world topology by deep reinforcement learning[C]// 32nd Conference on Neural Information Processing Systems, within Workshop on Machine Learning for Intelligent Transportation Systems. Canada, 2018: 1-9.
[29] LIU X Y, DING Z, BORST S, et al. Deep reinforcement learning for intelligent transportation systems [C]// 32nd Conference on Neural Information Processing Systems. Canada, 2018.
[30] PUTERMAN M L. Markov decision processes: discrete stochastic dynamic programming[M]. John Wiley & Sons, 2014.
[31] TAN T, BAO F, DENG Y, et al. Cooperative deep reinforcement learning for large-scale traffic grid signal control [J]. IEEE Transactions on Cybernetics, 2019, 50(6): 2687-2700.
[32] WU T, ZHOU P, LIU K, et al. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks[J]. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8243-8256.
[33] ZHAO T, WANG P, LI S. Traffic Signal Control with Deep Reinforcement Learning[C]// 2019 International Conference on Intelligent Computing, Automation and Systems(ICICAS). IEEE, 2019: 763-767.
[34] ZHANG R, ISHIKAWA A, WANG W, et al. Using reinforcement learning with partial vehicle detection for intelligent traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(1): 404-415.
[35] CHU T, WANG J, CODECÀ L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(3): 1086-1095.
[36] MOUSAVI S S, SCHUKAT M, HOWLEY E. Traffic light control using deep policy-gradient and value-function-based rein-

forcement learning[J]. IET Intelligent Transport Systems, 2017, 11(7): 417-423.
[37] VAN DER POL E, OLIEHOEK F A. Coordinated deep reinforcement learners for traffic light control[C]// Proceedings of Learning, Inference and Control of Multi-agent Systems(at NIPS 2016). 2016: 1-8.
[38] GONG Y, ABDEL-ATY M, CAI Q, et al. Decentralized network level adaptive signal control by multi-agent deep reinforcement learning[J]. Transportation Research Interdisciplinary Perspectives, 2019, 1: 100020.
[39] WAN C H, HWANG M C. Value-based deep reinforcement learning for adaptive isolated intersection signal control[J]. IET Intelligent Transport Systems, 2018, 12(9): 1005-1010.
[40] ZENG J, HU J, ZHANG Y. Adaptive traffic signal control with deep recurrent Q-learning[C]// 2018 IEEE Intelligent Vehicles Symposium(IV). IEEE, 2018: 1215-1220.
[41] WEI H, CHEN C, WU K, et al. Deep reinforcement learning for traffic signal control along arterials[C]// Proceedings of the 2019. DRL4KDD, 2019.
[42] TAN K L, PODDAR S, SARKAR S, et al. Deep reinforcement learning for adaptive traffic signal control[C]// Dynamic Systems and Control Conference. American Society of Mechanical Engineers, 2019.
[43] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine learning, 1992, 8(3): 279-292.
[44] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[45] PETERS J, SCHAAL S. Reinforcement learning of motor skills with policy gradients[J]. Neural Networks, 2008, 21(4): 682-697.
[46] KONDA V, TSITSIKLIS J. Actor-critic algorithms [C]// Advances in Neural Information Processing Systems. 1999: 1008-1014.
[47] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[C]// International Conference on Learning Representations. American, 2016.
[48] MONAHAN G E. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms [J]. Management Science, 1982, 28(1): 1-16.
[49] EREZ T, SMART W D. A scalable method for solving high-dimensional continuous POMDPs using local approximation[C]// Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence. California, 2010.
[50] RITCHER S. Traffic light scheduling using policy-gradient reinforcement learning[C]// The International Conference on Automated Planning and Scheduling. ICAPS, 2007.
[51] CHU T, QU S, WANG J. Large-scale traffic grid signal control with regional reinforcement learning[C]// 2016 American Control Conference(ACC). IEEE, 2016: 815-820.
[52] AZIZ H M A, ZHU F, UKKUSURI S V. Learning-based traffic signal controlalgorithms with neighborhood information sharing: An application for sustainable mobility[J]. Journal of Intelligent Transportation Systems, 2018, 22(1): 40-52.
[53] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[54] TAN M. Multi-agent reinforcement learning: Independent vs. cooperative agents[C]// Proceedings of the Tenth International Conference on Machine Learning. 1993: 330-337.
[55] FOERSTER J, NARDELLI N, FARQUHAR G, et al. Stabilising experience replay for deep multi-agent reinforcement learning[C]// International Conference on Machine Learning. PMLR, 2017: 1146-1155.
[56] GUESTRIN C, KOLLER D, PARR R. Multiagent planning with factored MDPs[C]// Advances in Neural Information Processing Systems. 2001, 14: 1523-1530.
[57] KOK J R, VLASSIS N. Collaborative multiagent reinforcement learning by payoff propagation[J]. Journal of Machine Learning Research, 2006, 7(1): 1789-1828.
[58] CASAS N. Deep deterministic policy gradient for urban traffic light control[J]. arXiv: 1703. 09035, 2017.
[59] WANG X, KE L, QIAO Z, et al. Large-scale traffic signal control using a novel multiagent reinforcement learning[J]. IEEE Transactions on Cybernetics, 2020, 51(1): 174-187.
[60] LOPEZ P A, BEHRISCH M, BIEKER-WALZ L, et al. Microscopic traffic simulation using sumo[C]// 2018 21st International Conference on Intelligent Transportation Systems(ITSC). IEEE, 2018: 2575-2582.
[61] ZHANG H, FENG S, LIU C, et al. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario[C]// The World Wide Web Conference. 2019: 3620-3624.
[62] FELLENDORF M, VORTISCH P. Microscopic traffic flow simulator VISSIM[M]// Fundamentals of Traffic Simulation. Springer, New York, NY, 2010: 63-93.
[63] CAMERON G D B, DUNCAN G I D. PARAMICS—Parallel microscopic simulation of road traffic[J]. The Journal of Supercomputing, 1996, 10(1): 25-53.
[64] GRAHAM B. Spatially-sparse convolutional neural networks [J]. arXiv: 1409. 6070, 2014.
[65] HUANG D, OU J, XIAO H X, et al. Collaborative optimization of traffic signal lights and vehicle fleet trajectory at intersection [J]. Journal of Chongqing University of Technology(Natural Science), 2022, 36(4): 84-93.
[66] ZHOU Y, LIU L, WANG L, et al. Service-aware 6G: An intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-256.

YU Ze, born in 1998, postgraduate. His main research interests include intelligent traffic and reinforcement learning.

NING Nianwen, born in 1991, Ph.D, lecturer. His main research interests include intelligent traffic and graph neural network.

（责任编辑：喻藜）