文章编号：0253-374X(2024)07-1089-10 **DOI**：10.11908/j.issn.0253-374x.22441

基于多智能体深度强化学习的高速公路可变限速协同控制方法

余荣杰[1]，徐　灵[2]，章锐辞[1]
（1. 同济大学 道路与交通工程教育部重点实验室，上海 201804；2. 浙江杭绍甬高速公路有限公司，浙江 杭州 310000）

摘要：面向高速公路多路段可变限速协同控制需求，针对高维参数空间高效训练寻优难题，提出了应用多智能体深度确定性策略梯度(MADDPG)算法的高速公路可变限速协同控制方法。区别于既有研究的单个智能体深度确定性策略梯度(DDPG)算法，MADDPG将每个管控单元抽象为具备Actor-Critic强化学习架构的智能体，在算法训练过程中共享各智能体的状态、动作信息，使得各智能体具备推测其余智能体控制策略的能力，进而实现多路段协同控制。基于开源仿真软件SUMO，在高速公路典型拥堵场景对提出的控制方法开展管控效果验证。实验结果表明，提出的MADDPG算法降低了拥堵持续时间和路段运行速度标准差，分别减少69.23 %、47.96 %，可显著提高交通效率与安全。对比单智能体DDPG算法，MADDPG可节约50 %的训练时间并提高7.44 %的累计回报值，多智能体算法可提升协同控制策略的优化效率。进一步，为验证智能体间共享信息的必要性，将MADDPG与独立多智能体DDPG(IDDPG)算法进行对比：相较于IDDPG，MADDPG可使拥堵持续时间、速度标准差均值的改善提升11.65 %、19.00 %。

关键词：交通工程；可变限速协同控制；多智能体深度强化学习；交通拥堵；高速公路；交通效率；交通安全
中图分类号：U491.5　　**文献标志码**：A

Coordinated Variable Speed Limit Control for Freeway Based on Multi-Agent Deep Reinforcement Learning

YU Rongjie[1], *XU Ling*[2], *ZHANG Ruici*[1]
(1. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China; 2. Zhejiang Hangshaoyong Expressway Co., Ltd., Hangzhou 310000, China)

Abstract: In order to meet the needs of coordinated variable speed limit (VSL) control of multi-segment on freeways, and to solve the problem of efficient training optimization in high-dimensional parameter space, a multi-agent deep deterministic policy gradient (MADDPG) algorithm is proposed for freeway VSL control. Different from the existing research on the single agent Deep Deterministic Policy Gradient (DDPG) algorithm, MADDPG abstracts each control unit as an agent with Actor-Critic reinforcement learning architecture, and shares each agent in the algorithm training process. The state and action information of the agents enable each agent to have the ability to infer the control strategies of other agents, thereby realizing multi-segment coordinated control. Based on the open source simulation software SUMO, the effect of the control method proposed is verified in a typical freeway traffic jam scenario. The experimental results show that the MADDPG algorithm proposed reduces the traffic jam duration and the speed standard deviation by 69.23 % and 47.96 % respectively, which can significantly improve the traffic efficiency and safety. Compared with the single-agent DDPG algorithm, MADDPG can save 50 % of the training time and increase the cumulative return value by 7.44 %. The multi-agent algorithm can improve the optimization efficiency of the collaborative control strategy. Further, in order to verify the necessity of sharing information among agents, MADDPG is compared with the independent DDPG (IDDPG) algorithm: It is shown that MADDPG can improve the traffic jam duration and speed standard deviation by 11.65 %, 19.00 % respectively.

Keywords: traffic engineering; coordinated variable speed limit control; multi-agent deep reinforcement learning; traffic jam; freeway; traffic efficiency; traffic safety

交通事故导致的偶发交通拥堵是高速公路运营

收稿日期：2022-10-18
基金项目：浙江省交通运输厅科技计划项目(2021047)
第一作者：余荣杰，教授，博士生导师，工学博士，主要研究方向为道路交通事故风险辨识与主动管控。E-mail：yurongjie@tongji.edu.cn
通信作者：章锐辞，硕士生，主要研究方向为主动交通安全。E-mail：zhang_ruici@tongji.edu.cn

论文拓展介绍

长期面临的痛点问题[1],其造成的经济损失约为2.5亿元·d^{-1}[2]。目前,我国高速公路的交通运行安全管控仍以事故后的被动处置、救援为主,而欧美国家应用主动交通管理系统(active traffic management system, ATMS)[3]已显著降低事故发生概率和拥堵持续时间(分别降低约20 %[4]、12 %[5])。其中,可变限速(variable speed limit, VSL)控制是主动交通管理系统的核心管控手段。

可变限速控制以路侧限速信息发布设备为界划分管控单元,基于实时交通流状态对管控单元的限速值进行调节[6]。早期高速公路路侧限速信息发布设备布设间隔约为5 km,相邻路段交通流运行状态的相互影响程度较低,各单元的可变限速控制策略相对独立。近年来,在智慧高速持续建设的背景下,信息发布设备间隔缩减至平均2 km,密集处仅数百米,相邻路段的交通流运行状态强相关,促使可变限速控制策略由单一路段独立控制向多路段协同控制。

传统可变限速控制研究基于模型预测控制(model predictive control, MPC)框架,求解优化模型获得最优的控制策略[7],核心是对管控后交通运行态势的预测精度。既有研究多使用宏观交通流模型来模拟交通流演化规律,主要包括一阶交通流模型和二阶交通流模型两类,前者以元胞传输模型(cell transmission model, CTM)[8]、Lighthill-Whitham-Richards(LWR)模型[9]为代表,后者以METANET模型[10]为代表。由于宏观交通流模型本身未考虑可变限速控制对交通流运行的影响效应,预测偏差较大,大量研究尝试通过对交通流期望速度、平均速度的计算公式进行修正来提高交通流预测模型精度[11]。但交通流状态时变复杂[12],多路段环境下交通流运行态势预测误差大,无法满足MPC的高预测精度要求。

强化学习(reinforcement learning, RL)方法通过智能体与环境间的不断交互进行试错学习,自主探索最优控制策略,克服了传统MPC研究依赖交通流预测模型精度的局限[13],近年来广泛应用于可变限速控制策略的研发。基于RL的可变限速控制研究将可变限速控制过程视为马尔科夫决策过程(Markov decision process, MDP),以交通仿真环境作为交互对象,基于仿真模型推演各种管控策略下的交通运行态势,通过转移后的交通流状态计算回报值作为管控效果评价,最终得到最优管控策略。然而传统RL方法无法直接处理连续的交通流状态输入,需要对连续状态进行离散化处理,随后进行遍历式测试[12]。在多路段协同控制条件下,系统状态参数空间是由每个路段的交通流状态组成的高维度连续空间(假设有6个路段,3个交通流参数,每个参数划分10个状态,则需遍历$(10^3)^6=10^{18}$种离散参数组合),传统RL方法优化效率低。

基于深度强化学习(deep reinforcement learning, DRL)的可变限速控制克服了传统RL方法不适用于高维连续状态空间的缺陷。DRL将RL与深度学习结合,利用神经网络的函数逼近能力,拟合管控效果与交通流状态之间的复杂非线性关系。Wu等[14]基于深度确定性策略梯度算法(deep deterministic policy gradient, DDPG)在连续高维交通状态环境内高效优化差异化可变限速控制(differential variable speed limit, DVSL)策略,并通过高速公路瓶颈区的仿真实验验证了算法对于安全、效率和排放改善具有良好效果。Ke等[15]基于双层深度Q网络(double deep Q network, DDQN)算法进行可变限速控制策略优化,能有效缩短行程时间,提高交通效率,并考虑使用迁移学习提升新场景建模的训练效率。Roy等[16]采用深度Q网络(deep Q network, DQN)模型优化可变限速控制策略,以东京高速公路繁忙路段的仿真实验结果表明,其所提出的VSL能有效降低19 %的事故风险。然而,上述基于DRL的可变限速控制研究多采用单智能体算法,即在训练过程中使用一个智能体来控制所有路段,在训练过程中未考虑不同路段间的相互影响,无法实现多路段协同管控。

多智能体深度强化学习(multi agent deep reinforcement learning, MADRL)是DRL的拓展,主要可分为两类:①智能体间相互独立;②智能体间能进行交互。在第一类研究中各智能体只能观测到本地的信息,无法接收到其他智能体的状态及动作决策信息[17]。该类算法简单但无法解决多智能体协同问题。为了克服这一缺陷,第二类研究在训练过程中为每个智能体提供其余智能体的状态、动作信息,使得每个智能体在进行决策前都能推测其余智能体的策略。多智能体间的协同提升了算法收敛速度及表现效果,被广泛应用于机器人控制[18]、车辆轨迹控制[19]、无人机路径规划[20]等领域。在交通控制领域,MADRL目前主要应用于大规模路网的信号控制策略优化,并取得了良好的效果。Wu等[21]应用多智能体循环深度确定性策略梯度(multi-agent recurrent deep deterministic policy gradient, MARDDPG)算法,开展多交叉口的信号协同控制。Li等[22]提出了一种新的多智能体强化学习方法知识共享深度确定性策略梯度(knowledge sharing deep deterministic policy gradient, KSDDPG),通过增

强交通信号控制机之间的协作来实现最优控制，相较现有的DRL方法能显著提高大规模交通网络的信号控制效率。然而，目前MADRL算法还未被应用于高速公路可变限速控制。

面向高速公路多路段协同控制需求，针对高维参数空间高效训练寻优难题，本文提出了基于MADRL算法的高速公路可变限速协同控制方法。利用深度网络捕捉连续高维状态输入特征，通过训练过程中系统全局信息共享进行多路段联动，最终实现了可变限速协同控制策略的高效寻优，并基于仿真实验验证了该方法在典型拥堵管控场景下的效果及优越性。

1　基于多智能体深度强化学习的高速公路可变限速协同控制

1.1　马尔科夫决策过程推导

本文考虑的高速公路可变限速协同控制实验路段如图1所示。研究对象为高速公路主线，间隔2 km布设龙门架发布限速值信息。以龙门架为界，高速公路主线划分为若干长度为2 km的管控单元（称为路段），路段的实时交通流数据（流量、密度、速度等）可通过线圈等数据采集器获取。

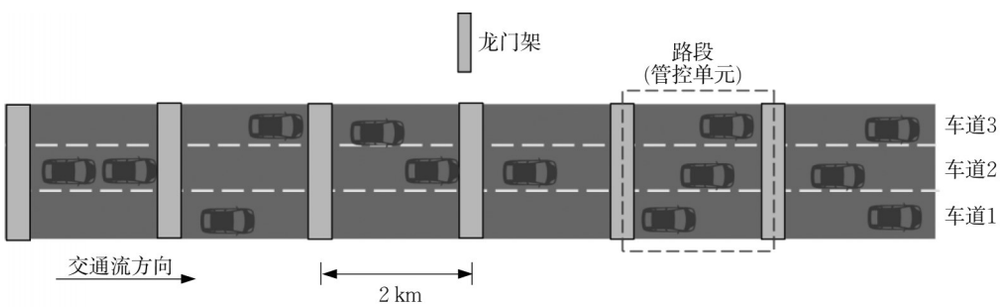


图1　高速公路可变限速协同控制实例

Fig. 1　Example of coordinated VSL control for freeways

为了将可变限速控制与MADRL关联起来，控制问题需要推导为通过与环境交互进行试错学习的MDP。MDP过程包括(S, A, P, R, γ)，S代表状态空间，A代表动作空间，P代表状态转移概率，R代表回报值分布，$\gamma\in[0,1]$表示定义即时奖励和历史奖励的相对重要性的折扣因子。在每一个离散的时间步长$t=1,2,3,\ldots$（每个时间步长之间间隔为5 min），智能体都会根据一些策略π来选择动作，环境相应的由状态s_t转移到状态s_{t+1}，智能体收到回报值r_t，智能体的训练目标是最大化累计回报值。考虑上述高速公路可变限速协同控制场景，智能体（Agent）、状态（State）、动作（Action）和回报值函数（Reward）设计如下所示：

① 智能体：将路段可变限速控制器视为智能体，每个路段的智能体可以对该路段区域设置不同限速随后通过路侧设备进行发布，智能体数量即为路段数量N。

② 状态：状态是实时交通流环境或交通流演变的体现。鉴于动态交通流的复杂性，很难精确表示交通流如何从一个状态转移到新状态。本文将交通流状态定义为路段速度均值、路段速度标准差、路段流量均值和路段流量标准差的集合，即每个路段的交通流状态均为4维空间，总状态空间的维度为$4N$。

③ 动作：动作即智能体施加给控制路段的限速值。结合实际工程应用情况，将限速值最小值设置为60 km·h^{-1}，最大值设置为100 km·h^{-1}，间隔为5 km·h^{-1}，即每个路段的离散限速值共有9种情况：60，65，70，75，80，85，90，95，100 km·h^{-1}。限速值每5 min调整一次，总限速动作空间的维度为$9N$。

④ 回报值：目标导向是深度强化学习的基础，深度强化学习通过学习选择动作，使累计回报值最大化。在本研究中，多路段可变限速协同控制的主要目标包括提升交通安全和交通效率两方面。

交通安全方面，选用对事故发生概率（称为事故风险）有显著影响的速度标准差和速度均值指标[23]，而既有研究表明理论上事故风险与速度均值负相关，与速度标准差正相关，因此交通安全回报值函数r_1设置为：$r_1=w_1V_{AS}-w_2V_{SS}$。其中，V_{AS}为路段速度均值，V_{SS}为路段速度标准差，单位均为km·h^{-1}，w_1、w_2分别为确定为1.5和2.5。

交通效率方面，既有研究[24]表明低速车辆是导致交通拥堵的重要影响因素，低速车辆的数量是交通拥堵严重程度的重要表征指标，当低速车辆总数较少时路段整体通畅，因此采用低速车辆总数作为

反应交通效率的指标。参考《中华人民共和国交通安全法》[25]相关条例,结合实证交通流数据分布特征,本文将低速车辆定义为速度低于50 km·h^{-1}(13.89 m·s^{-1})的车辆。交通效率回报值函数r_2设置为:$r_2=-V_{SV}$。其中V_{SV}为低速车辆总数。

最终综合交通安全和效率两方面,确定回报函数r如下:

$$r=W_1r_1+W_2r_2 \tag{1}$$

式中:W_1、W_2分别代表安全、效率两类奖励值的权重,其数值由高速公路管理部门结合实际工程经验确定为0.8、0.2。

1.2 多智能体深度确定性策略梯度算法

本文采用多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)算法[26]来对可变限速控制策略进行优化。MADDPG算法是DDPG算法在多智能体环境下的拓展,在训练多智能体协作能力的任务上有着较好的表现。本文基于MADDPG算法设计的N个路段的可变限速协同控制优化算法框架如图2所示。每个智能体都为Actor-Critic架构,其中Actor网络为策略生成网络,即基于实时交通流状态给出相应的限速动作,而Critic网络为策略评价网络,即基于实时交通流状态对当前的限速动作进行效果评价,在每个时间步长下,每个路段的Actor网络基于交通流状态$s_{t,i}$选择限速动作$a_{t,i}$并施加到环境中,每个路段环境在相应的限速动作的干预下会转变到新的状态$s_{t+1,i}$并且能计算相应的回报值$r_{t,i}$,$i\in[1,N]$以$\sum_i(s_{t,i},a_{t,i},r_{t,i},s_{t+1,i})$数据为基础对所有的Actor、Critic网络进行训练更新。

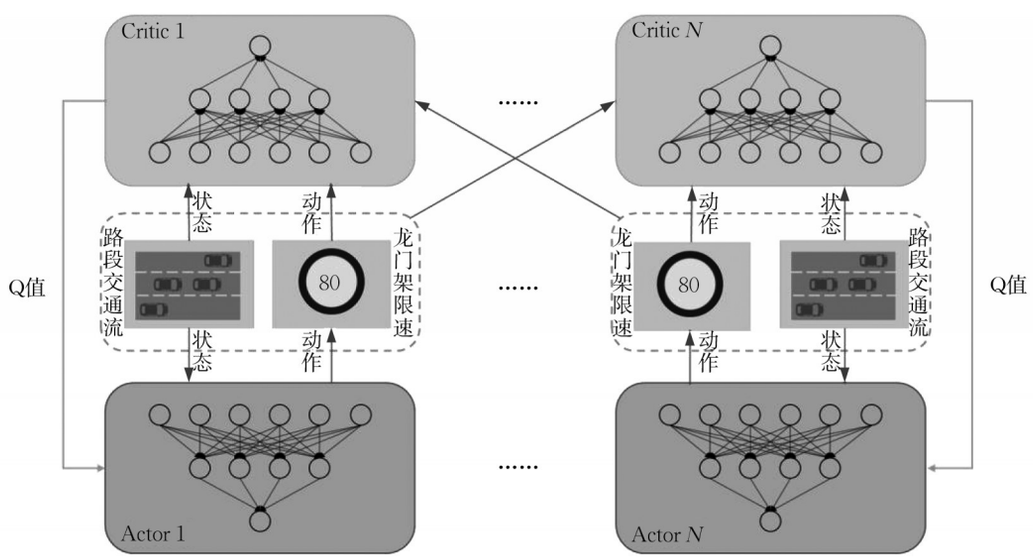


图2 基于MADDPG的高速公路可变限速协同控制优化算法框架

Fig. 2 Coordinated VSL control optimization framework for freeways based on MADDPG

为了实现多智能体间的协同,利用整个交通流环境的整体状态和动作(即所有路段的交通流状态及限速值)对每个智能体进行训练,对于每个时间步长t,每个智能体的Critic网络都能获得所有智能体的观测值和给出的动作,即所有路段的交通流状态及限速,并依据全局状态、动作信息输出$Q_i^{(\phi_i)}(s_{t,1},\ldots,s_{t,N};a_{t,1},\ldots,a_{t,N})$,作为对智能体$i$的策略评价值Q,其中$\phi_i$为第$i$个Critic网络的参数值。而Actor i通过其当前的管控策略$\pi_i^{(\theta_i)}$,根据观测到的单一路段的状态$s_{t,i}$给出相应路段的限速动作$a_{t,i}$,即$a_{t,i}=\pi_i^{(\theta_i)}(s_{t,i})$,因此当离线训练过程结束进行实际应用时不需要使用Critic网络,仅通过每个路段的Actor网络对该路段进行管控,实现了集中式训练和分布式应用。通过这种集中式训练的方法使得每个智能体具备推断其余智能体策略的能力,从而加速训练并获得最优的协同策略。由于每个智能体对应路段的空间位置不同,各智能体能学习到互异的分布式策略。最终训练得到的各智能体有不同的Actor、Critic网络参数。

传统的RL智能体会对状态、动作和回报值进行逐步采样,在利用这些数据更新参数后立即丢弃这些经验数据,这种方法导致样本之间具有强烈的时间相关性并且可能会将重要的数据快速遗忘,本文通过经验回放(experience replay)来解决上述问题。经验回放设

置会将训练过程中产生的数据进行存储记忆(replay memory)，存储上限为N_m，当数据量达到阈值N_{T}后，不断地从存储记忆中随机批量采样B个数据样本以更新所有智能体。为了探索更多潜在最优策略，在每次选择动作时都都加入随机高斯噪声N，随后通过$a_{t,i}=\pi_i^{(\theta_i)}(s_{t,i})+N_t$来选择具体动作。

为了使得训练过程更为稳定，引入目标网络(target networks)$\pi_i^{(\theta_i')}$和$Q_i^{(\phi_i')}$。每个Critic网络通过最小化损失值函数$L(\phi_i)$来更新网络参数，其中α为学习率，∇为偏导符号，$L(\phi_i)$如式(3)、(4)所示，其中$s^j=\{s_1^j,\ldots,s_N^j\}$为样本$j$对应的所有智能体状态的集合，$a^j=\{a_1^j,\ldots,a_N^j\}$为样本$j$对应的所有智能体动作的集合，$r^j=\{r_1^j,\ldots,r_N^j\}$为通过式(1)计算得到的各路段的回报值集合，$j'$代表训练过程中样本$j$的下一步长的更新样本。所有Critic网络的参数集合为$\phi=\{\phi_1,\cdots,\phi_N\}$。

$$\phi_i=\phi_i-\alpha\cdot\nabla_{\phi_i}L(\phi_i) \tag{2}$$

$$L(\phi_i)=\frac{1}{B}\sum_{j=1}^{B}[(Q_i^{\phi_i'}(s^j,a^j)-y^j)^2] \tag{3}$$

$$y^j=r^j+Q_i^{(\phi_i)}(s^j,a^j)|_{a_i^j=\pi_i^{(\theta_i')}(s_i^{j'})} \tag{4}$$

Actor网络的参数更新如式(5)所示，Actor网络更新的目标为最大化长期累计回报值，其损失函数如式(6)所示。所有Actor网络的参数集合为$\theta=\{\theta_1,\cdots,\theta_N\}\phi=\{\phi_1,\cdots,\phi_N\}$。

$$\theta_i=\theta_i-\alpha\cdot\nabla_{\theta_i}L(\theta_i) \tag{5}$$

$$L(\theta_i)=\frac{1}{B}\sum_{j=1}^{B}Q_i^{(\phi_i)}(s^j,a^j)|_{a_i^j=\pi_i^{(\theta_i')}(s_i^j)} \tag{6}$$

基于MADDPG的算法训练流程总结如下所示：

(1)确定智能体的个数N，初始化每个Actor和Critic网络分别为$\pi_i^{(\theta_i)}$和$Q_i^{(\phi_i)}$，初始化目标网络的参数$\phi_i'\leftarrow\phi_i$，$\theta_i'\leftarrow\theta_i$，$i=1,2,3,\cdots,N$，设置训练周期$M$、时间步长$T$、批次数据量$B$、经验回放的存储上限$N_m$及启动阈值$N_{\mathrm{T}}$，令$m=t=1$。

(2)进行第m个周期中第t个步长的训练。

(3)感知每个路段的当前交通流状态集合$s_t=\{s_{t,1},\ldots,s_{t,N}\}$，在随机高斯噪声下每个路段Actor网络的输出限速动作组成集合$a_t=\{a_{t,1},\ldots,a_{t,N}\}$，各路段在限速下转移到状态集合$s_{t+1}=\{s_{t+1,1},\ldots,s_{t+1,N}\}$，并基于式(1)得到各路段的回报值集合$r_t=\{r_{t,1},\ldots,r_{t,N}\}$。

(4)将数据(s_t,a_t,r_t,s_{t+1})进行存储记忆，当存储数据量大于N_{T}时随机抽取B个数据用于更新各路段智能体。

(5)基于式(2)—(6)对所有Actor网络和Critic网络进行更新。

(6)$t=t+1$，若$t<T$，转步骤(2)，否则转步骤(7)。

(7)$m=m+1$，利用$\phi_i'\leftarrow\tau\phi_i+(1-\tau)\phi_i'$、$\theta_i'\leftarrow\tau\theta_i+(1-\tau)\theta_i'$来更新目标网络，$\tau$为目标网络更新率。若$m<M$，则令$t=1$，转步骤(2)，否则结束算法，输出最终的目标网络参数作为智能体网络参数。

1.3　深度网络结构

各智能体的Actor、Critic网络结构如图3所示。Critic网络由5个全连接层组成，全连接层1和全连接层2的输出合并后输出至全连接层3。每一层线性层的输出参数在经过下一层线性层之前，使用激活函数leaky_relu处理，各层的具体参数如表1所示。Actor网络由4个线性层组成(表2)，每一层线性层的输出参数在经过下一层线性层之前，使用激活函数leaky_relu处理，全连接层4输出的值通过tanh激活函数保证其数值在区间[−1,1]内，并最终映射为离散的动作，即路段限速值。

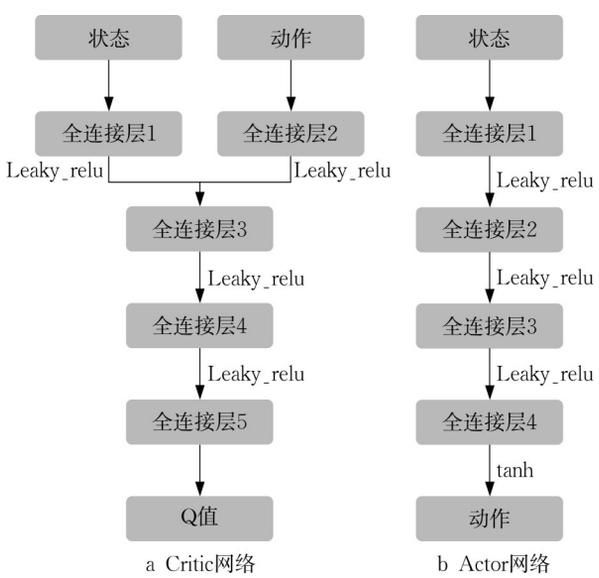


图3　深度网络结构

Fig. 3　Structure of deep neural networks

表1　Critic网络结构信息

Tab. 1　Architecture of Critic networks

全连接层	输入维度	输出维度
FC1	4N	196
FC2	2N	128
FC3	324	128
FC4	128	64
FC5	64	1

表2 Actor网络结构信息

Tab. 2 Architecture of Actor networks

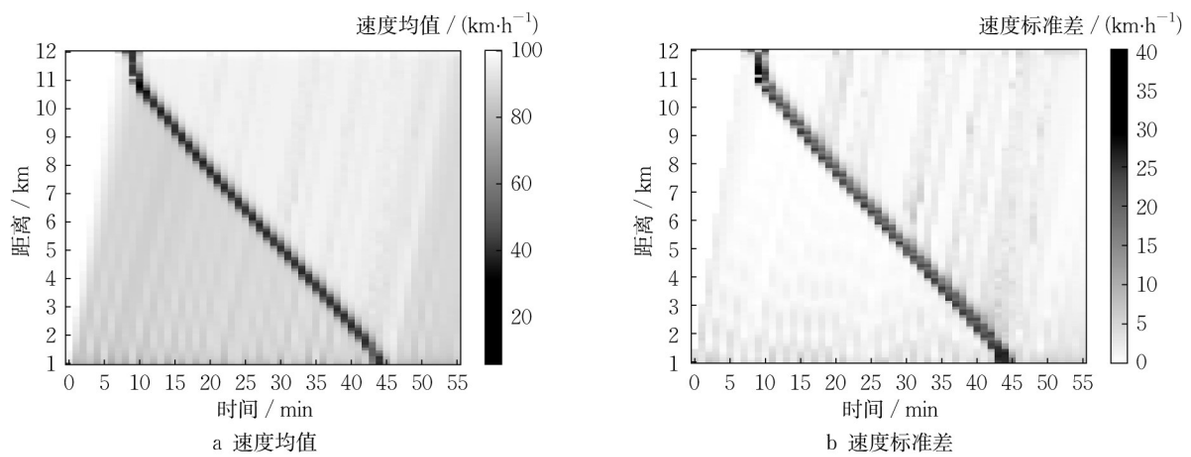
全连接层	输入维度	输出维度
FC1	4	256
FC2	256	128
FC3	128	64
FC4	64	1

2 仿真实验及结果

基于开源仿真软件SUMO搭建仿真模型模拟典型拥堵管控场景,以仿真环境作为MADDPG算法的交互环境对算法进行训练,验证了本文提出的MADDPG算法对于交通流安全、效率的改善效果。

2.1 仿真实验设计

由事故等异常事件导致的偶发性拥堵(后续简称为拥堵)是高速公路的典型管控场景之一,当发生拥堵时交通效率会大幅下降,同时也极易发生追尾事故,影响交通安全程度。本文参考经典文献[17]中的仿真设置,以拥堵场景作为仿真场景,仿真路段为12 km的单向3车道高速公路,以2 km为间隔划分为6个路段,默认限速为100 km·h^{-1},仿真时长为1 h,流量输入为5 400 veh·h^{-1}。为了产生从下游向上游传播的交通拥堵波(traffic jam wave),在5~10 min时令11.5~12 km处的车辆速度为30 km·h^{-1},在无管控的情况下,以1~12 km的路段为研究对象,0~55 min为研究时段,拥堵场景的速度均值和速度标准差的时空分布热力图如图4所示,单位均为km·h^{-1},可明显观察到拥堵波传播的过程。



a 速度均值 b 速度标准差

图4 无限速控制下拥堵场景交通流状态时空分布热力图

Fig. 4 Traffic state spatiotemporal heatmaps in traffic jam scenario without VSL control

将单次场景仿真作为可变限速控制策略优化算法的一个训练周期,针对单次仿真,0~5 min用于仿真预热,5~10 min用于产生交通拥堵波,10~15 min让其自然向上传播,从第15 min开始,每隔5 min每个路段对应的智能体都会根据前5 min的路段交通流状态选择路段限速并施加到仿真环境中,即限速的变化间隔为5 min,并基于算法训练流程不断优化控制策略,共训练100个周期。仿真过程中的跟驰模型和换道模型分别设置为智能驾驶人模型(intelligent driver model, IDM)和LC2013模型,并基于实证交通流数据进行标定,驾驶行为模型、仿真环境及MADDPG算法的相关参数的取值如表3所示。

表3 相关参数设置

Tab. 3 Setting of related parameters

参数	取值
期望加速度/(m·s^{-1})	1.4
期望减速度/(m·s^{-2})	2.0
期望车头时距/s	1.6
最小车头间距/m	2.4
单次仿真时长/s	3 600
训练周期M	100
限速动作更新频率/(min·次$^{-1}$)	5
控制步长上限T	10
智能体数量(路段数量)N	6
批次数据量B	64
经验回放存储上限N_m	100 000
经验回放启动阈值N_T	100
学习率α	0.000 1
目标网络更新率τ	0.01

在基于SUMO构建的拥堵场景中对算法进行训练和测试,可变限速控制智能体的设置通过Python完成,利用TraCI接口实现智能体和SUMO仿真环境的交互。控制逻辑如图5所示,每5 min利用TraCI接口实时感知交通运行状态作为输入,智能体网络基于状态输入自动输出各路段的最优限速值,并利用TraCI

接口将限速值施加到仿真环境中。

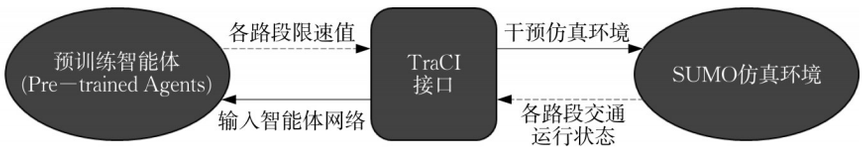


图5　SUMO平台可变限速控制实施逻辑

Fig. 5　VSL control logic in SUMO platform

从效率和安全两方面对算法的效果进行量化评估。考虑到拥堵场景管控的主要目标是使由低速车辆导致的交通拥堵波加速消散,结合拥堵场景仿真设计及低速车辆的判定阈值设计,本研究中的拥堵判定条件设定为:路段中存在某一长度超过100 m的区域的平均速度低于50 km·h^{-1}。将拥堵持续时间和低速车辆总数作为效率评价指标,将速度标准差均值作为安全评价指标。同时,以累计回报值作为效率和安全的综合评价指标。

2.2　仿真结果分析与评价

2.2.1　交通流运行改善效果

图6和图7对比了有无可变限速协同控制干预下的交通流状态时空图,可以明显观察到相较于无限速控制的情况,MADDPG能有效的加速交通拥堵波的消散。评价指标结果汇总如表4所示,括号内数值表明指标增长幅度。在交通效率方面,干预条件下,交通拥堵持续时间下降69.23 %,低速车辆总数减少35.91 %。在交通安全方面,干预条件下,速度标准差减少了47.96 %。在综合效率和安全方面,累计回报值提升了33.53 %。基于MADDPG的可变限速控制显著改善了交通效率与安全。

2.2.2　MADDPG与DDPG性能对比

为验证多智能体设置的优越性,将MADDPG与经典文献中的DDPG[14]算法进行对比。将智能体数量设置为1,即通过一个智能体收集6个路段的交通流状态,并一次性输出6个路段的限速值,其余参数(如状态、限速值范围、回报值设置等)保持不变,其框架示意图如图8所示。两种算法均在同一拥堵场景中进行训练和测试。

MADDPG和DDPG的训练过程对比如图9所示,MADDPG算法大约在第20个周期开始收敛,DDPG算法大约在第40个周期开始趋于收敛,相较于DDPG,

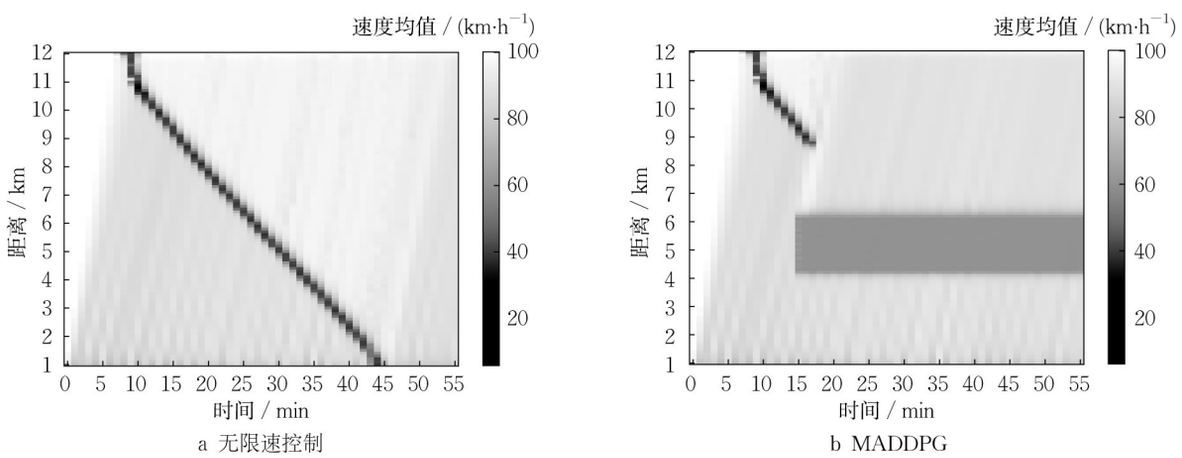


图6　测试场景速度均值时空分布热力图

Fig. 6　Average speed spatiotemporal heatmaps in testing scenario

表4　评价指标结果汇总

Tab. 4　Comparison of evaluation indicators

评价指标\\控制算法	无限速控制	MADDPG	IDDPG
拥堵持续时间/min	39	**12 (−69.23 %)**	17 (−56.67 %)
低速车辆总数	582	**373 (−35.91 %)**	433 (−25.60 %)
速度标准差均值/(km·h^{-1})	2.21	**1.15 (−47.96 %)**	1.57 (−28.96 %)
累计回报值	532.51	**711.06 (+33.53 %)**	630.60 (+18.42 %)

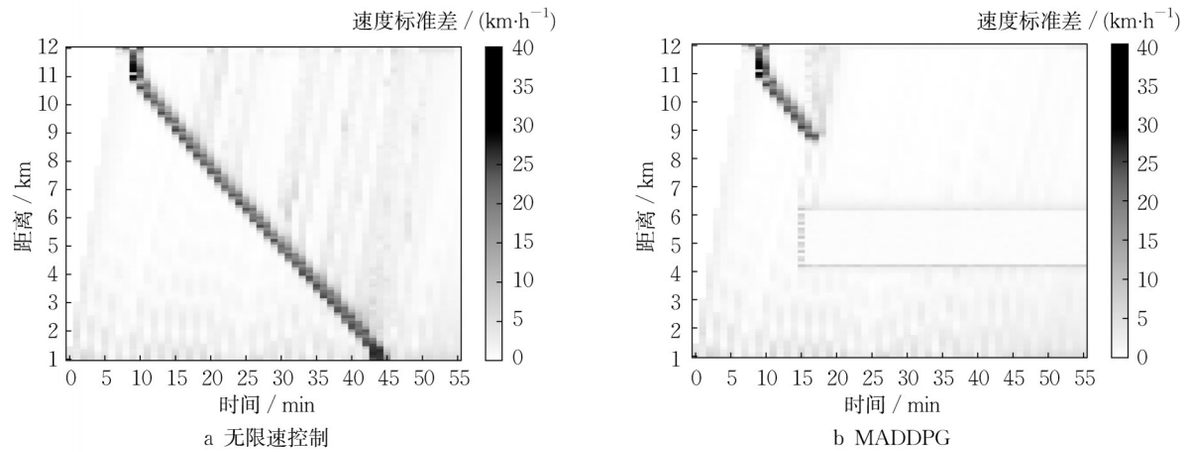


图7 测试场景速度标准差时空分布热力图

Fig. 7 Speed standard deviation spatiotemporal heatmaps in testing scenario

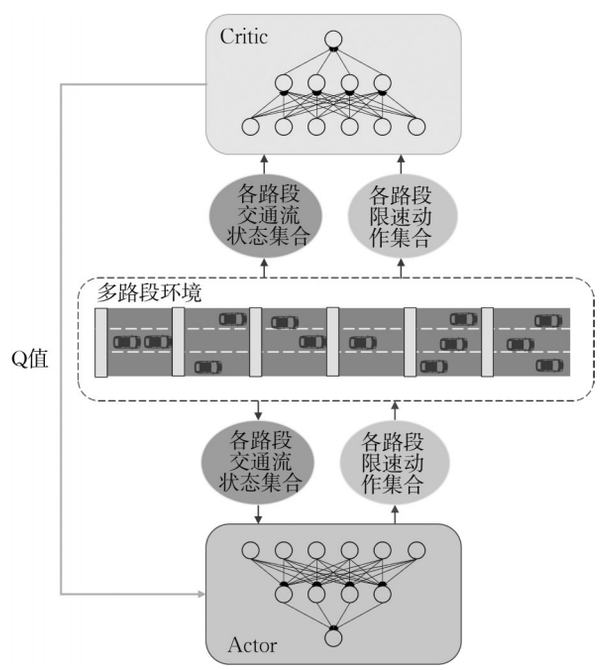


图8 基于DDPG的高速公路可变限速控制优化算法框架

Fig. 8 VSL control optimization framework for freeways based on DDPG

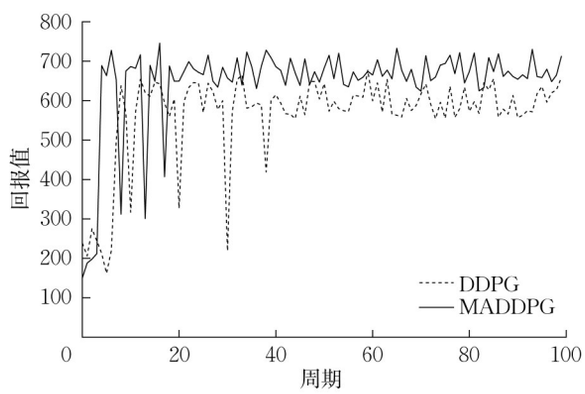


图9 MADDPG和DDPG训练过程对比

Fig. 9 Comparison of training progress of MADDPG and DDPG

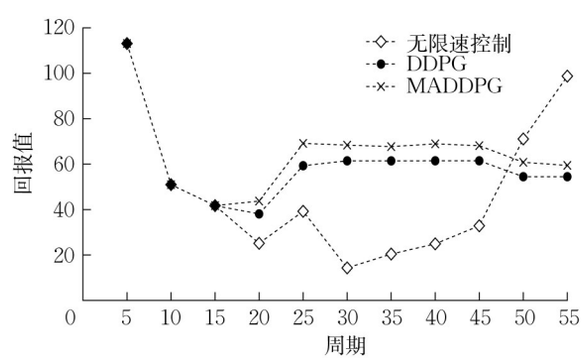


图10 MADDPG和DDPG在测试场景的回报值对比曲线

Fig. 10 Reward comparison of MADDPG and DDPG in testing scenario

MADDPG可以节约大约50 %的训练的时间,且算法收敛后MADDPG的回报值整体高于DDPG。

两种算法在拥堵场景测试时的回报值对比曲线如图10所示,每5 min基于该5 min内的交通流计算回报值函数,由于限速控制从第15 min开始,各类情形下的前15 min的回报值相等。从第20 min开始,相较于无限速控制,MADDPG和DDPG均能获得较高的回报值。两类控制下的累计回报值分别为711.06和658.15,相较于DDPG,MADDPG能提高7.44 %的累计回报值。

综上所述,MAADPG的性能优于DDPG。从训练效率来看,MADDPG算法的达到收敛所需的训练时间约为DDPG的50 %,从测试回报值曲线来看,两者均能改善拥堵场景的交通流运行,但MADDPG算法能获得更高的回报值、取得更好的改善效果,表明多智能体设置能有效提升算法性能。

2.2.3　智能体间信息共享的必要性验证

为验证多智能体利用全局信息训练管控算法的必要性，将MADDPG算法与同为多智能体深度强化学习算法的独立DDPG（independent DDPG，IDDPG）[27]算法进行比较。IDDPG与MADDPG相比，区别仅在于训练过程中利用各路段的局部信息训练相应的智能体，即各智能体之间无信息交互、相互独立，其余参数（如智能体数量、状态、限速值范围、回报值设置等）保持不变。其框架示意图如图11所示。两种算法均在同一拥堵场景中进行训练和测试。

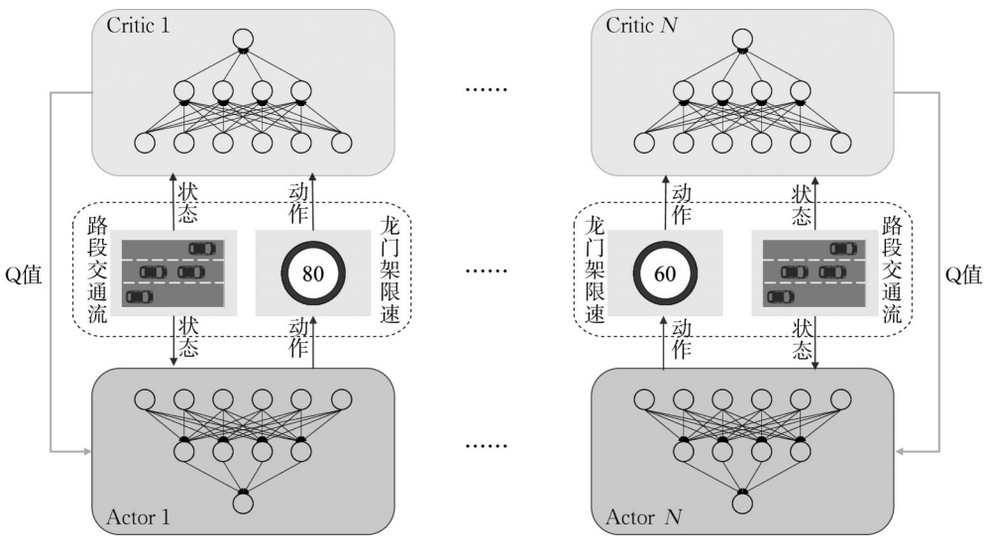


图11　基于IDDPG的高速公路可变限速控制优化算法框架

Fig. 11　VSL control optimization framework for freeways based on IDDPG

评价指标结果汇总如表4所示，在交通效率方面，IDDPG使拥堵持续时间下降了56.67 %，使低速车辆总数减少了25.60 %。在交通安全方面，IDDPG使速度标准差减少了28.96 %。在综合效率和安全方面，IDDPG使累计回报值提升了18.42 %。综上所述，IDDPG同样可改善交通效率与安全，但其指标改善效果在各方面均劣于MADDPG，表明了通过智能体间共享信息实现的多路段协同控制能进一步提高管控成效。

3　结 语

（1）本文面向高速公路多路段协同管控需求，提出了一种基于MADRL算法（MADDPG）的高速公路可变限速协同控制方法。利用深度网络提取高维连续交通流状态特征，将每个路段均视为一个智能体，进行了集中式训练和分布式应用：集中式训练过程中共享各智能体的交通流状态和限速动作信息，使得在训练过程中每一路段的智能体能推测其余路段智能体的策略，而分布式应用保证各智能体在实际应用过程中能仅基于本路段的交通流状态进行最优可变限速管控。

（2）基于SUMO软件搭建高速公路多路段环境，以典型拥堵管控场景仿真实验为例对所提出算法的效果进行验证。结果表明MADDPG使拥堵持续时间下降了69.23 %，使低速车辆总数减少了35.91 %，使速度标准差减少了47.96 %，使累计回报值提升了33.53 %，能显著提高交通效率与安全。

（3）与单智能体DRL算法（DDPG）相比，MADDPG使算法收敛的训练耗时缩短约50 %、累计回报值提升7.44 %，表明MADDPG在训练效率和交通流运行改善方面均优于DDPG，多智能体算法可提升协同控制策略的优化效率。

（4）与智能体间相互独立、无信息交互的MADRL算法（IDDPG）相比，MADDPG在交通流运行改善方面有显著优势：拥堵持续时间、低速车辆总数、速度标准差均值分别进一步降低11.65 %、10.31 %、19.00 %，累计回报值进一步提升15.11 %。表明利用全局信息训练算法实现多路段协同能有效提高交通流管控成效。

（5）本文的算法效果在典型拥堵场景仿真实验中得到了验证。在未来研究中，可以考虑将算法在更大规模的路网中进行测试分析，同时也可以考虑将方法拓展到多种管控策略的协同控制研究中，例如进行可变限速控制与匝道控制的协同研究等。

作者贡献声明：

余荣杰：提供研究思路、技术指导，完善、修订论文。

徐　灵：提供研究思路、技术指导及实验数据。

章锐辞：提供研究思路，设计并开展实验，撰写论文。

参考文献：

[1] KEJUN L, MEIPING Y, JIANLONG Z, *et al*. Model predictive control for variable speed limit in freeway work zone[C]//2008 27th Chinese Control Conference. Kunming: IEEE, 2008: 488-493.

[2] 包杰. 基于多源数据的城市路网交通事故风险研究[D]. 南京：东南大学，2019.
BAO Jie. Research on crash risk of urban road network based on multi-source data[D]. Nanjing: Southeast University. 2014.

[3] HARBORD B. M25 controlled motorway-results of the first two years [C]//9th International Conference on Road Transport Information and Control. [S.l.]: IET Digital Library, 1998: 149-154.

[4] MIRSHAHI M, OBENBERGER J, FUHS C A, *et al*. Active traffic management: The next step in congestion management[R]. [S.l.]: United States. Federal Highway Administration, 2007.

[5] HOOGENDOORN S P, DAAMEN W, HOOGENDOORN R G, *et al*. Assessment of dynamic speed limits on freeway A20 near Rotterdam, Netherlands[J]. Transportation Research Record, 2013, 2380(1): 61.

[6] 李志斌. 快速道路可变限速控制技术[D]. 南京：东南大学，2014.
LI Zhibin. Variable speed limit technique on expressways[D]. Nanjing: Southeast University, 2014.

[7] HAN Y, YU H, LI Z, *et al*. An optimal control-based vehicle speed guidance strategy to improve traffic safety and efficiency against freeway jam waves[J]. Accident Analysis & Prevention, 2021, 163: 106429.

[8] LI Z, ZHU X, LIU X, *et al*. Model-based predictive variable speed limit control on multi-lane freeways with a line of connected automated vehicles[C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). Edmonton: IEEE, 2019: 1989-1994.

[9] HAN Y, HEGYI A, YUAN Y, *et al*. Resolving freeway jam waves by discrete first-order model-based predictive control of variable speed limits [J]. Transportation Research Part C: Emerging Technologies, 2017, 77: 405.

[10] LU X Y, SHLADOVER S. MPC-based variable speed limit and its impact on traffic with V2I type ACC [C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Edmonton: IEEE, 2018: 3923-3928.

[11] YU R, ABDEL-ATY M. An optimal variable speed limits system to ameliorate traffic safety risk[J]. Transportation Research Part C: Emerging Technologies, 2014, 46: 235.

[12] WANG C, XU Y, ZHANG J, *et al*. Integrated traffic control for freeway recurrent bottleneck based on deep reinforcement learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9): 15522.

[13] LI Z, LIU P, XU C, *et al*. Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(11): 3204.

[14] WU Y, TAN H, QIN L, *et al*. Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm [J]. Transportation Research Part C: Emerging Technologies, 2020, 117: 102649.

[15] KE Z, LI Z, CAO Z, *et al*. Enhancing transferability of deep reinforcement learning-based variable speed limit control using transfer learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(7): 4684.

[16] ROY A, HOSSAIN M, MUROMACHI Y. A deep reinforcement learning-based intelligent intervention framework for real-time proactive road safety management [J]. Accident Analysis & Prevention, 2022, 165: 106512.

[17] 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.

[18] SARTORETTI G, KERR J, SHI Y, *et al*. Primal: Pathfinding via reinforcement and imitation multi-agent learning [J]. IEEE Robotics and Automation Letters, 2019, 4(3): 2378.

[19] GUILLEN-PEREZ A, CANO M D. Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrows intersections[J]. IEEE Transactions on Vehicular Technology, 2022, 71(7): 7033.

[20] QIE H, SHI D, SHEN T, *et al*. Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning[J]. IEEE Access, 2019, 7: 146264.

[21] 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.

[22] LI Z, YU H, ZHANG G, *et al*. Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning [J]. Transportation Research Part C: Emerging Technologies, 2021, 125: 103059.

[23] YU R, ABDEL-ATY M. Utilizing support vector machine in real-time crash risk evaluation[J]. Accident Analysis & Prevention, 2013, 51: 252.

[24] 周召敏. T-CPS下考虑低速车影响的交通拥堵特征分析及抑制策略研究[D]. 重庆：重庆大学，2020.
ZHOU Zhaomin. Research on low-speed-vehicles-based congestion characteristics and congestion control methods in T-CPS[D]. Chongqing: Chongqing University, 2020.

[25] 全国人大常委会. 中华人民共和国道路交通安全法[M]. 北京：全国人大常委会，2021.
The Standing Committee of the National People's Congress. Road traffic safety law of the People's Republic of China[M]. Beijing: The Standing Committee of the National People's Congress, 2021

[26] LOWE R, WU Y I, TAMAR A, *et al*. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). [S.l.]: Curran Associates Inc., 2017: 6379-6390.

[27] ZHANG Z, ZHENG J, ZOU C. Multi-empirical discriminant multi-agent reinforcement learning algorithm based on intra-group evolution[C]// 2019 2nd International Symposium on Big Data and Applied Statistics. [S.l.]: IOP Publishing, 2020: 012038-012053.