DOI:10.19651/j.cnki.emt.2519061

基于时空特征融合的交通信号控制与仿真分析*

刘振航 黄德启 黄德意 黄海峰
(新疆大学电气工程学院 乌鲁木齐 830047)

摘 要:针对现有方法因忽略历史交通信息导致时空特征感知不足的问题,提出一种融合深度强化学习与时空特征建模的交叉口信号控制方法。该方法使用 D3QN-LSTM 混合网络架构,通过离散交通状态编码将多时段交通信息表征为高维矩阵,采用卷积神经网络提取空间特征,结合长短时记忆网络捕捉时序依赖关系,并设计基于奖励反馈的动态探索机制优化策略训练过程。基于 SUMO 仿真平台进行实验,结果表明:相较于固定时长控制及传统强化学习方法,所提方法在早高峰流量情况下平均排队长度指标上分别降低 49.95%、35.04%和 16.72%,累积等待时间减少 63.03%、35.55%和 20.15%,有效验证了时空特征建模与动态探索策略的优越性。为评估算法鲁棒性,进一步开展平峰期交通流实验,结果表明:所提算法在平均排队长度与累积等待时间指标上依然保持显著优势,证明该方法对不同交通场景具有强适应性和良好的泛化能力。

关键词:交通信号控制;深度强化学习;时空特征建模;SUMO;智能交通系统
中图分类号:TN911.4;TP183 **文献标识码**:A **国家标准学科分类代码**:580.20

Traffic signal control and simulation analysis based on spatio-temporal feature fusion

Liu Zhenhang Huang Deqi Huang Deyi Huang Haifeng
(School of Electrical Engineering, Xinjiang University, Urumqi 830047, China)

Abstract: To address the insufficient spatio-temporal feature perception caused by neglecting historical traffic information in existing methods, this study proposes an intersection signal control method integrating deep reinforcement learning with spatio-temporal feature modeling. The approach employs a hybrid D3QN-LSTM network architecture, which encodes multi-period traffic information into high-dimensional matrices through discrete traffic state representation. A convolutional neural network extracts spatial features, while a long short-term memory network captures temporal dependencies. A reward-feedback-driven dynamic exploration mechanism is further designed to optimize policy training. Experiments conducted on the SUMO simulation platform demonstrate that during morning peak traffic, the proposed method reduces average queue length by 49.95%, 35.04% and 16.72%, and decreases cumulative waiting time by 63.03%, 35.55% and 20.15% compared to fixed-timing control, conventional reinforcement learning methods and D3QN, respectively, validating the superiority of spatio-temporal feature modeling and dynamic exploration strategies. To assess algorithmic robustness, off-peak traffic flow experiments further confirm that the proposed method maintains significant advantages in both average queue length and cumulative waiting time metrics, demonstrating strong adaptability and generalizability across varying traffic load conditions.

Keywords: traffic signal control; deep reinforcement learning; spatio-temporal feature modeling; SUMO; intelligent transportation system

0 引 言

交通拥堵已经成为了许多大城市面临的严重问题之一,解决这些问题的途径一般有两种:扩大现有城市道路基础设施的规模以及交通信号配时优化,但城市土地需求扩张与供给限制的矛盾性导致扩大城市基础设施规模的成本高昂[1-2]。现有的基础设施无法改变,因此,优化交通信号配时被广泛视为成本最低且见效最快的解决方案[3]。

收稿日期:2025-06-09
* 基金项目:新疆维吾尔自治区自然科学基金(2022D01C430)、国家自然科学基金(51468062)项目资助

随着智能交通系统的快速发展，交通信号控制研究经历了从固定配时、感应控制到自适应控制的发展历程。传统固定配时方法依赖历史数据，难以应对交通流的动态变化[4]；感应控制虽能根据实时检测数据调整相位，但其信息采集与决策执行过程存在显著的时滞性[5]；相比之下，自适应交通信号控制(adaptive traffic signal control，ATSC)[6]能根据实时交通状况动态优化信号配时策略，提高车辆通行效率、减少交通事故、降低道路拥堵。深度强化学习(deep renforcement learning，DRL)[7]的发展为自适应交通信号控制提供了技术支撑，其通过智能体与环境的持续交互学习实现策略优化。现有研究主要通过改进智能体网络架构[8-10]、设计不同的奖励机制[11-12]、以及引入优先级经验回放[13-14]等手段提升控制性能，但在时空特征建模方面仍存在明显局限：一方面，现有方法[15]多采用瞬时交通状态作为输入，忽视了历史交通模式的时序关联性；另一方面，传统状态表征方法对高维交通信息的处理效率不足，导致模型难以捕捉复杂路况的空间相关性。此外，探索策略的平衡与利用[16]也制约了算法在动态环境中的收敛速度与稳定性。

针对交通流时空依赖性建模问题，研究者们已提出多种解决方案。其中，以时空图卷积网络[17](spatio-temporal graph convolutional network，STGCN)、Graph WaveNet[18]为代表的方法，通过显式定义路网拓扑结构，利用图卷积操作捕捉空间邻域车辆的交互与依赖关系，并结合循环神经网络(recurrent neural network，RNN)或时间卷积网络(temporal convolutional network，TCN)捕获动态交通流的时序特征。基于 Transformer 的时空编码方法[19]则以自注意力机制替代循环神经网络结构，并行处理长序列数据，有效缓解 RNN 在长序列建模中易出现的梯度消失或爆炸问题。在面对因忽略历史信息导致的感知不足的问题时，Tan 等[20]将过去若干时间步的交通状态直接拼接或堆叠为高维张量输入模型，Jia 等[21]在 RNN 基础上引入注意力机制以增强对历史信息的捕捉能力。Liu 等[22]结合卷积神经网络(convolutional neural networks，CNN)的空间特征提取能力与 RNN 的时序建模优势构建混合网络架构，共同处理时空信息，深度挖掘历史交通流数据中蕴含的周期性、趋势性等关键时序模式，从而提升模型对动态交通环境演变的感知能力。

基于上述混合架构的技术路线，本文提出了一种改进的决斗双深度 Q 网络(dueling double deep Q-network，D3QN)算法以解决交通流时空特征感知能力低的问题。首先，将多维度历史交通信息表示为矩阵；其次，设计 CNN-LSTM 混合网络架构，通过卷积层提取空间特征，结合长短时记忆网络捕获交通流的时序特征，实现交通流时空特征融合；最后，使用 D3QN 训练智能体，采用基于奖励反馈的自适应探索策略加快智能体收敛速度，并在 SUMO 仿真平台上对所提出算法进行验证。

1　相关原理与问题定义

1.1　深度强化学习相关原理

深度强化学习作为强化学习的高级范式，通过深度神经网络对强化学习中的状态-价值函数进行高阶抽象表达，突破了传统 RL 算法在高维状态空间下的维度灾难的问题。如图 1 所示，通过构建深度 Q 网络(deep Q-network，DQN)实现状态-动作价值函数的非线性逼近，将交通信号控制转化为可微分优化问题，提升了复杂交通场景的建模能力。

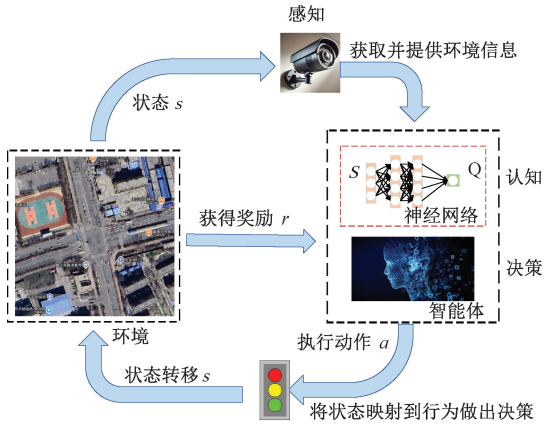


图 1　深度强化学习框架

Fig. 1　Deep reinforcement learning framework

DQN 算法是基于强化学习中 Q-learning 算法的一种改进算法，利用深度学习中的神经网络来记录每个状态下的动作值。网络的输入是状态信息，输出是每个动作的价值。从策略优化的角度来看，智能体的最终目标是求解最优策略 π^*，以最大化状态-动作价值函数。该算法价值函数定义为，从当前状态-动作对出发，遵循策略 π 所能获得的期望累积回报，如式(1)所示。

$$Q_\pi(s,a)=E[r^t+\gamma r^{t+1}+\gamma^2 r^{t+2}+\cdots+\gamma^T r^T] \tag{1}$$

式中：$\gamma\subset[0,1)$ 表示折扣因子，用以减少未来奖励的重要性，从而平衡当前奖励与未来奖励的关系，T 为终止时刻。

通过贝尔曼最优方程将其简化为如式(2)所示。

$$Q_{\pi^*}(s,a)=E\left[r_t+\gamma\max_{a'}Q_{\pi^*}(s',a')\right] \tag{2}$$

式中：π^* 是通过递归方法求得的最佳策略。

1.2　交通信号控制问题定义

1)交通环境描述

本文构建的交通控制模型以典型的四向十字交叉口为研究对象，其路口结构如图 2 所示。每条道路的长度为 360 m，每个方向有 4 个入口车道，其中右侧车道用于右转和直行，中间两条车道用于直行，左侧车道用于左转。图中的 E、W、S、N 分别代表东、西、南、北 4 个方向。交叉口由独立的智能体控制，采用东西直行、东西左转、南北直行和南北左转，四相位信号配时方案，并按顺序对这些信号配时进行编号处理。系统仅收集进口道检测器的数据，出口车道的车辆因已完成通行过程，其状态信息不被纳入状态空

间 $\boldsymbol{S}$ 中。

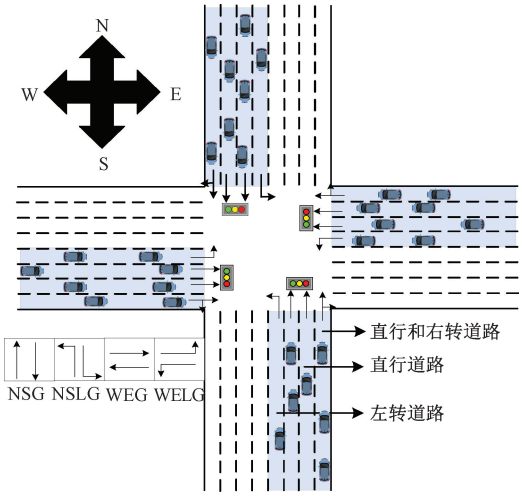


图 2 交通环境

Fig. 2 Traffic environment

2)状态定义

深度强化学习中的状态是用来描述环境的当前特征的,一般作为神经网络的输入,为智能的决策提供依据。本文基于离散交通状态编码(DTSE)来表示交通状态信息,如图 3 所示,将入口车道划分为非均匀元胞表示速度、位置信息,并且由于交通数据具有时间序列特征,需要在状态表示中添加时间维度。因此将状态空间表示为由历史环境信息、车辆的位置、速度堆叠以及与信号灯的相位拼接的形式表示,记作 $\boldsymbol{S}=[\boldsymbol{R},\boldsymbol{D},\boldsymbol{V},\boldsymbol{P}]$,其中 $\boldsymbol{R}$ 表示历史交通状。$\boldsymbol{D}$ 表示车辆位置矩阵用于表示车辆是否位于每个车道的网格中,1 表示有车,0 表示无车。$\boldsymbol{V}$ 表示每条车道单元格内的归一化车速矩阵。$\boldsymbol{P}$ 表示相位矩阵,采用 one-hot 编码方式,例如,$\boldsymbol{P}=\{1,0,0,0\}$ 表示南北方向直行和右转的绿灯信号处于激活状态。

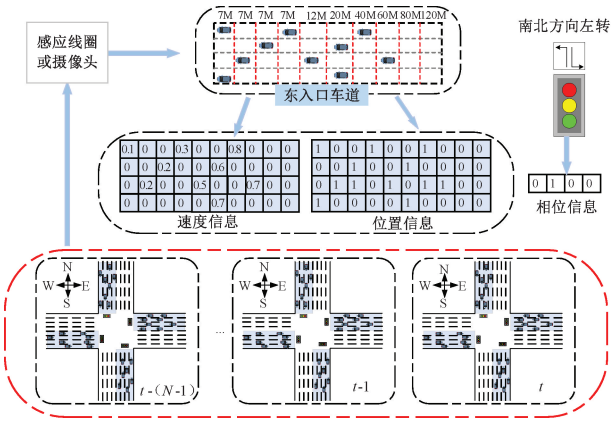


图 3 离散交通状态编码

Fig. 3 Discrete traffic state encoding

3)动作定义

智能体在感知当前环境状态后,利用神经网络评估所有潜在动作的价值,所有可行的动作组成了动作空间。动作空间定义了智能体可选择的交通信号相位,如图 4 所示,共包含 4 个离散动作,记作 $\boldsymbol{A}=\{\boldsymbol{NS},\boldsymbol{NSL},\boldsymbol{EW},\boldsymbol{EWL}\}$,分别对应不同方向的绿灯通行权限。动作的执行遵循安全过渡原则,智能体基于包含历史信息的交通状态(车辆位置、速度和相位的时序数据),通过 ε−贪婪策略选择动作。若新动作与上一动作不同(如从南北直行切换至东西直行),系统会先激活对应动作的黄灯相位,确保车辆安全减速。过渡完成后,才会执行新动作对应的绿灯相位。每个绿灯相位最短持续时间为 10 s,黄灯相位最短持续时间为 4 s。

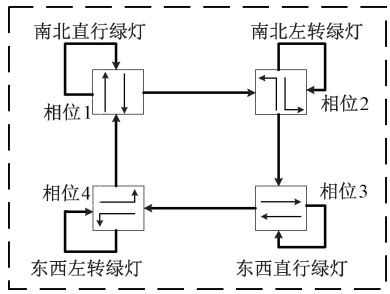


图 4 相位切换方案

Fig. 4 Phase switching scheme

信号相位如表 1 所示,G 为车道的通行信号,车辆可以不停车直接通过十字路口。r 为车道的停车信号。y 为车道的过渡信号,提醒车辆减速让行或停车等待。

表 1 交通信号相位

Table 1 Comparison of human evaluation scores

相位	当前相位	过渡相位
1	GGGGrrrrrrGGGGrrrrrr	yyyyrrrrrryyyyrrrrrr
2	rrrrGrrrrrrrrrGrrrrr	rrrryrrrrrrrrryrrrrr
3	rrrrrGGGGrrrrrrGGGGr	rrrrryyyyrrrrrryyyyr
4	rrrrrrrrrGrrrrrrrrG	rrrrrrrrryrrrrrrrry

4)奖励定义

智能体的核心目标是最大化其累积奖励。奖励函数定义了智能体在不同状态下采取某一动作所获得的即时反馈。为了有效地引导智能体优化交通信号控制策略,平衡车辆延误时间与排队长度等关键交通指标,本文将奖励函数设置为相邻时刻等待时间和排队长度差值的加权和的形式。

设 D_t 表示决策点 t 时刻交叉口所有入路车道的累积延误时间总和,D_{t-1} 表示决策点 $t-1$ 时刻的对应值,则相邻决策点之间的延误时间差定义如式(3)所示。

$$\Delta D = D_t - D_{t-1} \tag{3}$$

设 Q_t 表示决策点 t 时刻交叉口所有入路车道的排队长度总和,Q_{t-1} 表示决策点 $t-1$ 时刻的对应值,则相邻决策点之间的排队长度差定义如式(4)所示。

$$\Delta Q = Q_t - Q_{t-1} \tag{4}$$

最终的奖励函数如式(5)所示。

$$R_t = \alpha \times \Delta D + \beta \times \Delta Q \tag{5}$$

当智能体观察到 ΔD 和 ΔQ 持续减少时，环境反馈正向的激励信号，说明智能体正不断调整并优化其决策策略，逐步逼近累积奖励最大化的目标，直至策略收敛。

2　改进的目标网络

2.1　D3QN-LSTM 网络模型

Hochreit 等提出的长短期记忆网络被广泛用于解决时间序列问题[23-24]，如图 5 所示，本文将该网络引入 Dueling DQN 网络模型中，增强神经网络对包含历史交通状态信息的感知和表达能力。该模型中的状态由历史交通状态信息、车辆位置信息、归一化的车辆速度信息和交通灯的相位信息 4 个部分组成。首先，智能体将表征当前环境的状态特征和 R 个时间步的历史序列数据输入到神经网络模型中，如式(6)所示。

$$X_{t'} = \mathrm{Conv}(R, H, W, C) \tag{6}$$

式中：$t = t, t-1, \cdots, t-(N-1)$，$R$ 为 N 个相邻的历史状态，H 和 W 为车道数和道路元胞数，C 为速度和位置双通道。

在深度网络中，定义了两个卷积层、两个 ReLU 非线性激活层、一个最大池化层和展平层，包含序列数据信息输入经过深度网络如式(7)所示。

$$\begin{cases} X_{t''}^{\mathrm{CNN}} = \mathrm{ReLU}(X_{t'}) \\ X_{t'}^{\mathrm{Maxpolling,CNN}} = \mathrm{Maxpooling}(X_{t''}^{\mathrm{CNN}}) \\ X_{t''}^{\mathrm{CNN}} = \mathrm{ReLU}(X_{t''}) \\ X_{t''}^{\mathrm{flatten,CNN}} = \mathrm{flatten}(X_{t''}^{\mathrm{CNN}}) \end{cases} \tag{7}$$

然后将 $X_{t''}^{\mathrm{CNN}}$ 输入到 LSTM 网络中，产生最后一个隐藏输出 h_t，如式(8)所示。

$$\begin{cases} i_t = \sigma(X_{t''}^{\mathrm{flatten,CNN}} \boldsymbol{U}^i + h_{t-1} \boldsymbol{W}^i + b^i) \\ f_t = \sigma(X_{t''}^{\mathrm{flatten,CNN}} \boldsymbol{U}^f + h_{t-1} \boldsymbol{W}^f + b^f) \\ o_t = \sigma(X_{t''}^{\mathrm{flatten,CNN}} \boldsymbol{U}^o + h_{t-1} \boldsymbol{W}^o + b^o) \\ \tilde{C}_t = \tanh(X_{t''}^{\mathrm{flatten,CNN}} \boldsymbol{U}^g + h_{t-1} \boldsymbol{W}^g + b^g) \\ C_t = \sigma(f_t \odot C_{t-1} + i_t \odot C_t) \\ h_t = \mathrm{ReLU}(C_t) \odot o_t \end{cases} \tag{8}$$

式中：σ 是 sigmoid 激活函数；$\boldsymbol{U}^i$、$\boldsymbol{W}^i$ 和 b^i 分别是输入门的权重矩阵和偏置；$\boldsymbol{U}^f$、$\boldsymbol{W}^f$、b^f 是遗忘门的权重矩阵和偏置；$\boldsymbol{U}^0$、$\boldsymbol{W}^0$、b^0 是输出门的权重和偏置；$\boldsymbol{U}^g$、$\boldsymbol{W}^g$、b^g 是输入调制门的权重和偏置，$\odot$ 表示逐元素乘法，C_t 为细胞单元的状态，使得传输的信息不会发生梯度消失或爆炸的情况。

经过激活和展平操作后与相位信息合并，其过程可以表示为：

$$x_t^{a,\mathrm{LSTM}} = \mathrm{ReLU}(h_t \boldsymbol{W}^d + b) \tag{9}$$

$$X^{\mathrm{flatten,LSTM}} = \mathrm{flatten}(X^{a,\mathrm{LSTM}}) \tag{10}$$

$$X^{\mathrm{concat}} = \mathrm{concat}(\boldsymbol{P}, X^{FC,LSTM}) \tag{11}$$

式中：$\boldsymbol{P}$ 为相位矩阵。

最后把合并层提取的包含历史交通数据的交通状态信息输入到强化学习模型中。使用 Dueling DQN 的方法解耦状态价值函数与动作优势函数，通过共享前馈神经网络的全连接层参数 θ，将 Q 值分解为两个并行的子网络分支：状态值函数 $V(s;\theta)$ 和动作优势函数 $A(s,a;\theta)$。其中，状态值函数 $V(s;\theta)$ 用于表征环境状态 $\boldsymbol{s}$ 的全局潜在价值，独立于动作选择；而动作优势函数 $A(s,a;\theta)$ 则量化了在状态 $\boldsymbol{s}$ 下执行特定动作 $\boldsymbol{a}$ 相对于平均策略的局部优势。为了避免模型参数冗余并增强其可解释性，通过去均值化操作对优势函数进行正则化处理。最终，Q 值通过线性组合形式的输出如式(12)所示。

$$Q(s,a;\theta) = V(s,\theta) + \left(A(s,a;\theta) - \frac{1}{|A|}\sum_{a' \in A} A(s,a',\theta)\right) \tag{12}$$

式中：$|A|$ 表示动作空间。

2.2　训练过程

DQN 算法在面对最大化偏差问题时，容易导致 Q 值高估的问题。为了有效地减少 Q 值高估，采用 Double DQN 的方法训练模型参数，通过解耦动作选择与价值评估的过程，将动作选择和价值评估两个环节分别分配给不同的网络。

首先，使用参数为 θ 的主网络来确定下一状态下的最优动作，其过程如式(13)所示。

$$a^* = \mathrm{argmax}_a Q(s_{t+1}, a; \theta) \tag{13}$$

式中：S_{t+1} 为执行动作 a 后环境的新状态，a^* 为网络中 Q 值最大的动作。

随后，将该动作输入到参数为 θ' 的目标网络中使用式(14)进行价值评估。

$$Q_{\mathrm{Target}} = r_t + \gamma \cdot Q(s_{t+1}, a^*; \theta'_t) \tag{14}$$

式中：r_t 为 t 时刻的真实奖励，γ 为表示未来动作对当前状态影响的折扣因子。

通过这种解耦机制，有效抑制了传统 DQN 中由于最大化操作所引起的正偏差累积，进而避免了 Q 值的高估现象。

最后，分别采用主网络和目标网络的 Q 值作为真实值和预测值，采用梯度下降的方法更新网络参数，损失函数使用均方误差(mean square error，MSE)计算如式(15)所示。

$$l(\theta) = E(r_t + \gamma \max Q_\pi(s', a'; \theta') - Q_\pi(s, a; \theta))^2 \tag{15}$$

为了保证训练过程中 Q 值的稳定性，主网络参数采用实时更新的形式，而目标网络参数使用式(16)每 10 个动作更新一次。

$$\theta' = \omega \cdot \theta' + (1-\omega) \cdot \theta \tag{16}$$

2.3　算法伪代码

1) 初始化主网络参数 θ 和目标网络参数 θ′，创建经验

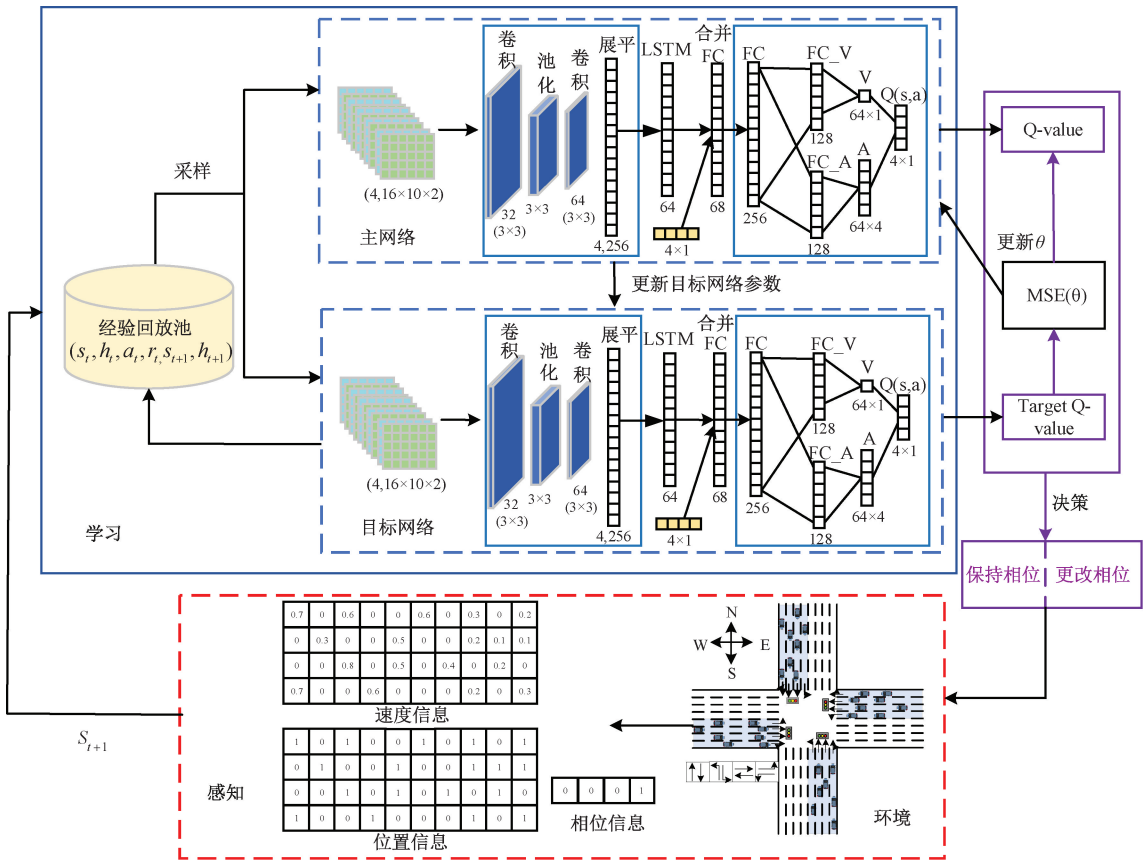


图 5 D3QN-LSTM 模型

Fig. 5 D3QN-LSTM model

回放池 D,设定采样批量大小 B,目标网络更新频率 F,初始化探索率 ε

2) for episode=1 到最大训练轮数 do:

初始化环境状态 s,重置 LSTM 隐藏状态

3) for t = 1 to Max-Steps do:

4) if 随机数 < ε then:

随机选择动作 $\boldsymbol{a}$

else:

通过主网络选择 Q 值最大动作 $a=\text{argmax}\ Q(s;\theta)$

5) 获得奖励 r,状态转移 s_{t+1},h_{t+1},终止标志 done

6) 更新状态序列,保留最近 R 个历史状态

7) 将转移样本 $(s_t,h_t,a_t,r_t,s_{t+1},h_{t+1})$ 存入经验池 D

8) 采样计数器 $C=C+1$

9) 更新当前状态 $s=s_{t+1}$

10) if $C \geqslant B$ then:

从经验池 D 随机采样批量样本

计算目标 Q 值:

$y=r+\gamma * Q'(s',\text{argmax}\ Q(s_{t+1};\theta);\theta')$

11) 通过均方误差计算损失:

$L=\text{MSE}(Q(s;\theta),y)$

12) 梯度下降更新主网络参数 θ

13) 更新目标网络参数 θ'

14) end if

15) end for

16) 衰减探索率 ε

17) end for

3 改进的探索策略

为了在训练过程中平衡智能体的探索与利用,本文提出一种基于奖励反馈的自适应 ε-greedy 策略实现探索与利用的动态平衡。该策略定义动作选择概率如式(17)所示。

$$a_t=\begin{cases}\underset{a}{\text{argmax}}Q(s_t,a),1-\varepsilon\\ \text{其他},\varepsilon\end{cases} \tag{17}$$

与传统的线性衰减机制不同,基于奖励反馈的自适应探索通过滑动窗口奖励评估实现 ε 值的动态调整。初始阶段充分探索状态空间,随着训练进程,根据最近 W 轮的奖励表现使用式(18)来自适应调整衰减速率:

$$\varepsilon_{t+1}=\max\left\{\begin{cases}0.85\varepsilon_t, R_t>\dfrac{1}{W}\displaystyle\sum_{i=t-W}^{t-1}R_i\\ 0.95\varepsilon_t, \text{其他}\end{cases},\varepsilon_{\min}=0.01\right\} \tag{18}$$

式中：R_t 表示第 t 合的累积奖励。

该机制具有双重调控特性：当近期奖励超过窗口均值时，加速探索衰减以强化策略利用，当表现低于历史水平时，减缓衰减速率以维持必要探索。这种动态平衡机制使智能体能够根据环境反馈自主调节行为模式，在训练初期保持 $\varepsilon>0.5$ 实现广泛状态探索，随着策略网络参数收敛逐步降低至 ε_{min}，最终实现以 Q 值最大化为目标的最优策略。

4　实验结果与分析

4.1　实验准备与评价指标

本文以乌鲁木齐市喀什东路与西平路交叉口为仿真对象。交通数据采集于 2025 年 3 月 25 日早高峰时段(08:30～10:00)，使用海康机器人 MV-CU060-10GCD 工业相机，在交叉口周边视野开阔的安全位置录制了 1.5 h 视频，清晰覆盖东、西、南、北 4 个入口车道的车辆到达、排队及通行过程。通过人工逐帧回放视频，统计各入口车道通过车辆数，汇总得到各驶入口流量如表 2 所示。

表 2　驶入路口车流量

Table 2　The traffic flow entering the intersection

驶入路口	车流量/辆
北入口	296
东入口	675
南入口	362
西入口	659

通过视频观察到交通高峰时段车辆呈先上升后下降的趋势，车辆开始时稳步上升，逐渐进入高峰时段；高峰期过后，车辆数量逐渐减少。基于此特征，采用威布尔分布对车辆到达时间间隔进行建模如式(19)所示。

$$f(x;\lambda;k)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}e^{-\left(\frac{x}{\lambda}\right)^{k}}, & x\geqslant 0\\ 0, & x<0\end{cases} \tag{19}$$

式中：x 为随机变量，k 为形状参数，为 λ 尺度参数。在实验中，形状参数 $k=2$，尺度参数 $\lambda=1\ 992$。车流量分布直方图如图 6 所示。

总仿真时间为 5 400 s，通过观察视频中排队长度、车辆速度范围、车辆的行驶轨迹等，确定辅助仿真参数。车辆的最大速度为 60 km/h，车辆长度为 4 m，最小间隔 2 m，统计车辆在路口时选择直行、左转、右转的比例，采用直行 70%、左转 15%、右转 15%的转向概率。将人工统计得到的车流量数据以及观察得到的车辆转向概率、车辆参数等作为 SUMO 交通需求的主要输入，交通信号状态则依据该路口实际运行方案在 SUMO 中进行设置。

评价指标为累计奖励(cumulative reword，CR)、平均排队长度(average queue length，AQL)、累计等待时间(average waiting time，CWT)。累计奖励反映了智能体在交通信号控制任务中的具体表现，在训练过程中累计奖励越大表明算法性能越强。平均排队长度和累计等待时间则反映了交叉口的拥堵程度和道路的吞吐量，其值越小，则代表车辆通行效率越高。

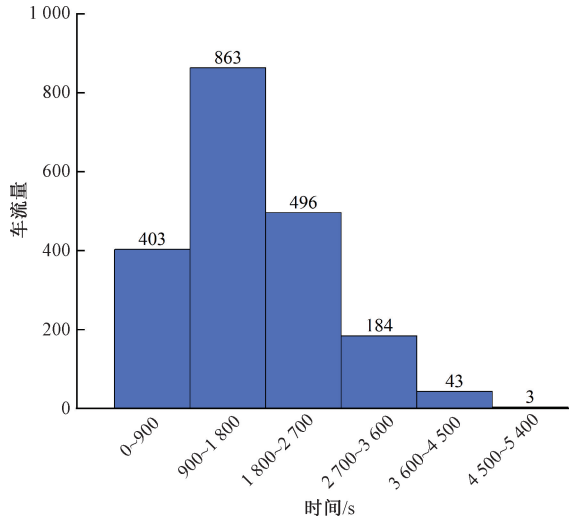


图 6　交通流量直方分布图

Fig. 6　Histogram of traffic flow distribution

4.2　仿真平台及参数设置

为验证所提方法的有效性，基于 SUMO 交通仿真平台开展基于真实交通数据的仿真实验。SUMO 可以根据真实道路得交通数据进行环境和参数设置，通过 SUMO 的控制接口 TraCI 与 Python 编程进行交互，控制仿真道路中的信号灯、车辆、交通流，实验的超参数如表 3 所示。

表 3　超参数

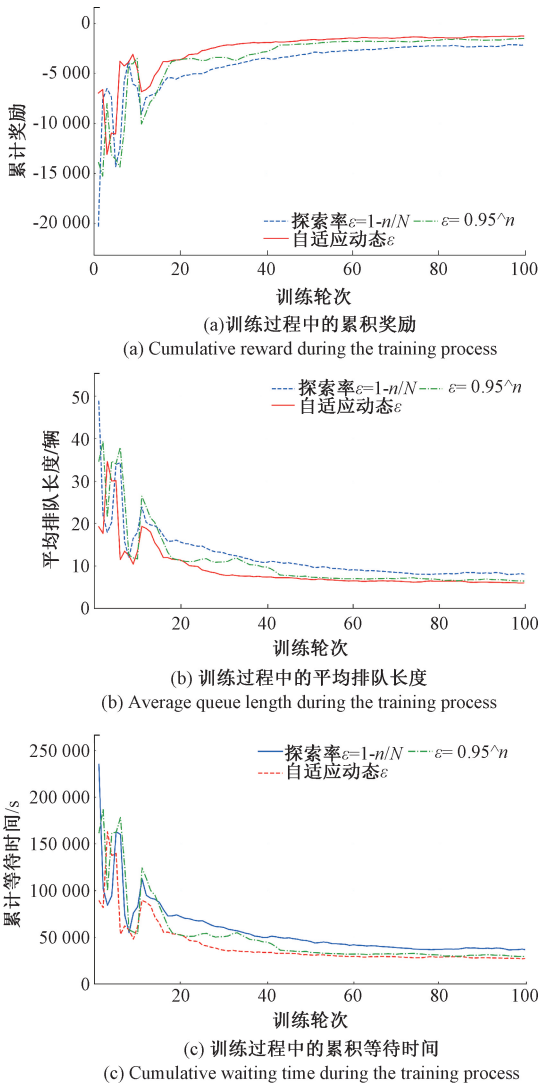
Table 3　Hyperparameter

超参数	值
训练轮数	100
每轮训练步数	5 400
批量大小	64
学习率 α	0.001
时间序列 R	4
折扣因子 γ	0.75
目标网络更新周期	10
α	0.1
β	0.9
经验池大小	50 000
LSTM 隐层单元	64

4.3　实验与分析

为验证改进探索策略的有效性，本文采用探索率线性衰减和幂函数衰减作为对比方法与本文提出的基于奖励反馈的自适应衰减探索策略进行比较，3 种方法均采用本文提出的算法模型。

在模型训练的初期,由于智能体没有先验知识,在动作选择上有很强的随机性。随着智能体对环境的逐步理解,智能体通过经验回放机制,更倾向于利用已知信息来进行决策,获得更多的奖励,最终逐步收敛。如图 7 (a)～(c)所示,在训练过程中,智能体使用基于奖励反馈的自适应衰减策略进行探索时,智能体获得的累计奖励最多,算法收敛速度更快,累计奖励较幂函数衰减优化 14.6%,较线性衰减优化 40%,算法收敛速度更快,平均排队长度比幂函数衰减法优化 6.9%,比线性衰减法优化 32.8%,累计等待时间较幂函数衰减优化 6.4%,较线性衰减优化 25.2%。



(a)训练过程中的累积奖励
(a) Cumulative reward during the training process

(b) 训练过程中的平均排队长度
(b) Average queue length during the training process

(c) 训练过程中的累积等待时间
(c) Cumulative waiting time during the training process

图 7 训练过程中各评价指标

Fig. 7 Each evaluation index during the training process

3 种探索策略的测试结果如图 8 所示,以平均排队长度为指标进行分析,在测试过程中,基于奖励反馈的自适应探索策略在减少平均排队长度方面明显优于其他两种探索方法,幂函数衰减次之,线性衰减策略表现最差,表明基于奖励反馈的自适应衰减探索策略具有较强的交通分流能力,能够将车辆的平均排队长度保持在相对较低的水平。

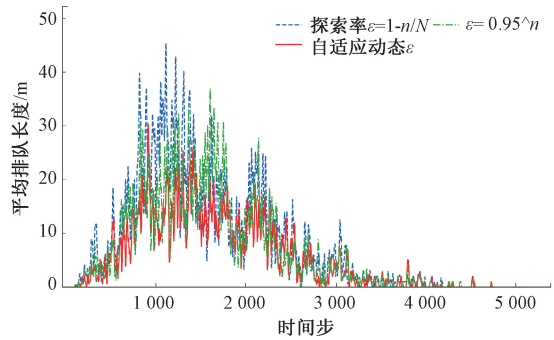


图 8 3 种算法控制效果对比

Fig. 8 Comparison of the control effects of three algorithms

为验证本文提出的算法的有效性,选择与以下 3 种基准方法进行比较。

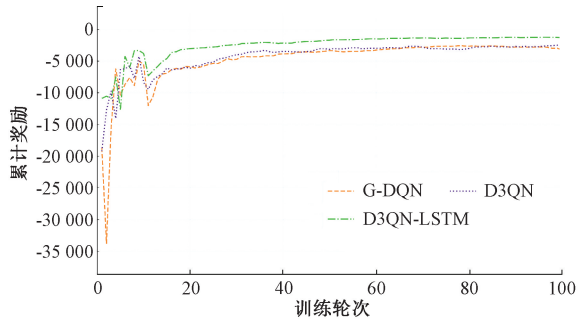
固定时间控制(fixed time control,FTC):FTC 是目前常见的交通信号控制方法,交通信号的切换周期通常依据历史数据或预测的交通需求进行规划。

G-DQN:继 Cao 等[25]的研究之后,本文实现了一个使用卷积神经网络提取特征信息的 DQN 模型进行比较,利用卷积神经网络,从当前路口的图像表示中提取交通信息,然后将提取的信息映射到 Q 值,与传统 DQN 相比,该方法运用了更复杂的深度神经网络模型。

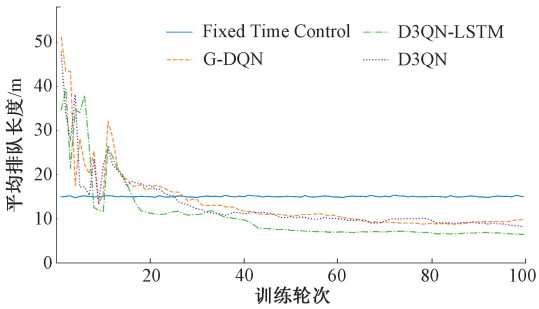
D3QN:与本文中概述的方法类似,参数如状态空间、动作空间、奖励函数和训练算法与本文中提出的方法一致。关键的区别在于,该基准算法中没有引入 LSTM 网络。

在收敛性与稳定性评估方面,图 9 (a)～(c)展示了早高峰流量下各算法 AQL、CWT 和 CR 随训练轮次的变化情况,其中 x 轴为训练轮次,y 轴为每次训练的累计奖励、平均排队长度和累计等待时间。FTC 由于采用固定相位配时方案,其 CWT 和 AQL 始终保持在一个较高的值,而 G-DQN、D3QN 以及 D3QN-LSTM 的 AQL 和 CWT 均随训练进程逐渐下降,其中,D3QN-LSTM 的收敛速度优于 3 种基准方法,且稳定性表现最佳。从算法机制来看,G-DQN 算法引入卷积神经网络与经验回放机制,虽缓解了特征维度过高的问题,但采用 DQN 训练智能体,易陷入局部最优或 Q 值高估问题。D3QN 作为 DQN 的改进方法,通过目标网络异步更新参数,并将 Q 值函数分解为状态值函数与优势函数,优化了不同动作对状态价值的估计,提升了模型的学习效率;但其因忽略历史数据中包含的交通信息导致模型感知能力不足。相比之下,D3QN-LSTM 方法融合了卷积神经网络结构与 Dueling DQN 的价值函数分解策略以及历史交通信息的交通状态表达,通过基于奖励反馈的自适应衰减动态调整智能体探索速率,从而实现最优控制策略。

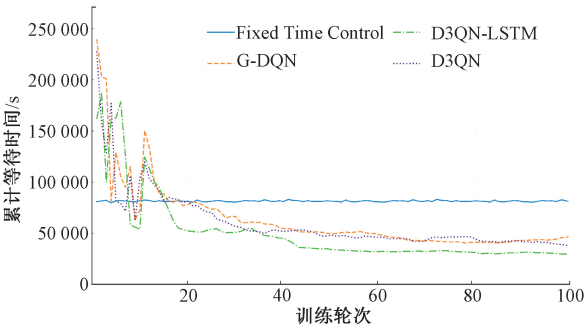
将 D3QN-LSTM 方法与 3 种基准方法使用不同随机种子进行仿真测试,以平均排队长度 (AQL) 和累计等待



(a) 算法获得的累计奖励对比
(a) Comparison of cumulative rewards obtained by the algorithms



(b) 4种算法平均排队长度对比
(b) Comparison of average queue lengths among the four algorithms



(c) 4种算法累积等待时间对比
(c) Comparison of cumulative waiting times among the four algorithms

图 9　早高峰流量下训练过程中各评价指标

Fig. 9　Evaluation indicators during the training process under morning peak traffic flow

时间（CWT）为评价指标。如图 10 所示，x 轴为不同算法，左 y 轴为各算法的车辆平均排队长度，右 y 轴为累计等待时间。该图展示了 1 992 辆驶入和离开交叉口的车辆经 20 次仿真测试计算所得的 AQL 和 CWT 结果。结果表明，D3QN-LSTM 方法在两项指标上均显著优于固定时长控制、G-DQN 和 D3QN 3 种基准方法：其中 AQL 分别降低 49.95%、35.04%和 16.72%，CWT 分别降低 63.03%、35.55%和 20.15%。实验结果证明本文提出的 D3QN-LSTM 模型的控制效果显著优于其他模型，表明交通流时空依赖性建模有助于信号控制策略的选择，并证明了基于 D3QN-LSTM 的强化学习算法使模型具备自学习和自适应的智能控制能力，进而可以优化控制策略，实现交叉口通行效率的提升。同时验证了排队长度与等待时间的正相关关系，即队列长度越短，车辆等待时间相应越短。

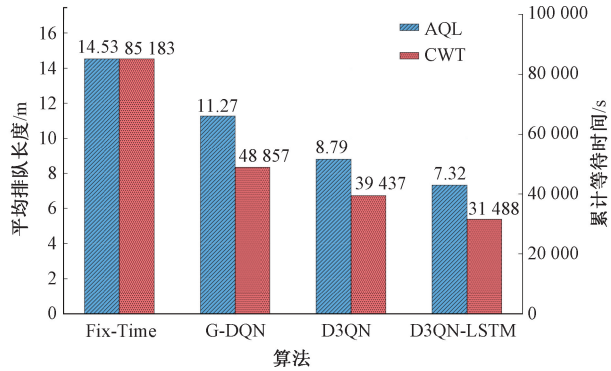


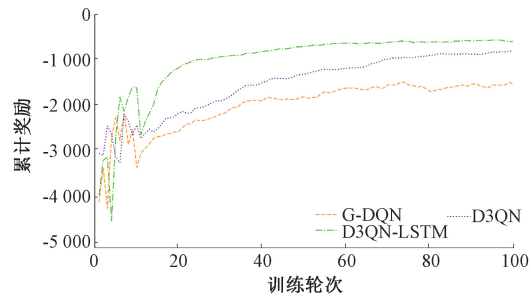
图 10　4 种算法 20 轮测试效果对比

Fig. 10　Comparison of the test effects of four algorithms in 20 rounds

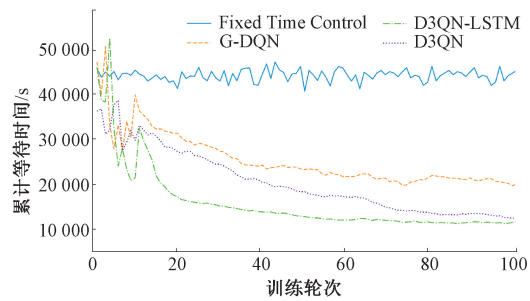
为进一步验证所提算法在不同交通负荷下的鲁棒性与泛化能力，本研究构建了包含 1 000 辆车辆的合成平峰期交通流数据集进行对比实验，保持早高峰场景的环境配置，仅替换交通需求数据。在相同评价指标下，对 D3QN-LSTM 及 3 种基准方法进行 100 轮训练与 20 轮测试。

100 轮训练结果如图 11 所示，即使在平峰期流量场景下，D3QN-LSTM 算法在累计奖励、平均排队长度（AQL）、累计等待时间（CWT）等关键指标上，仍优于 3 种基准方法。

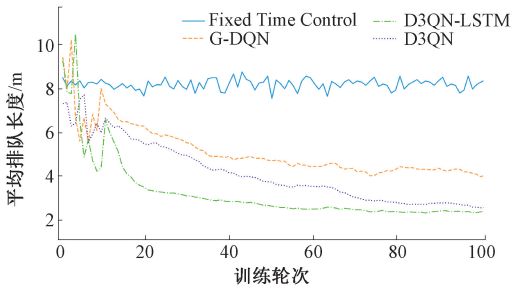
20 轮独立测试结果如表 4 所示，尽管评价指标因车流量降低而普遍下降，但 D3QN-LSTM 的控制效果仍表现出优秀的控制性能，与高峰场景下观察到的性能提升趋势一致，表明了 D3QN-LSTM 的控制优势并非局限于早高峰场



(a) 算法获得的累计奖励对比
(a) Comparison of cumulative rewards obtained by the algorithms



(b) 4种算法累积等待时间对比
(b) Comparison of cumulative waiting times among the four algorithms



(c) 4种算法平均排队长度对比

(c) Comparison of average queue lengths among the four algorithms

图 11 平峰流量下训练过程中各评价指标

Fig. 11 Evaluation indicators during the training process under off-peak traffic flow

表 4 平峰流量下不同算法控制效果对比

Table 4 Comparison of control effects of different algorithms under off-peak traffic flow

控制方法	累计奖励	AQL/m	CWT/s
FTC	—	9.15	4 939.85
DQN	−1 538.91	7.32	39 488.00
D3QN	−809.99	4.21	22 748.00
D3QN-LSTM	**−604.10**	**3.06**	**16 486.50**

景,其能够根据实时交通流动态学习并生成最优信号控制策略,从而在不同交通流量条件下仍能提升交叉口通行效率。

5 结 论

本文提出融合 LSTM 与 D3QN 的交通信号控制方法来解决当前交通流时空特征感知能力低的问题,通过离散交通状态编码实现高维交通信息表征,结合 CNN-LSTM 网络提取时空特征,采用 D3QN 训练智能体,设计基于奖励反馈的动态探索策略优化训练过程,加快收敛速度。使用基于真实交通环境的合成数据集验证了算法的有效性和不同交通流量下的鲁棒性与泛化能力。实验结果表明,相较于 D3QN 方法,该方法在单路口场景下平均排队长度降低 16.72%,累积等待时间减少 20.15%,验证了时空建模对提升动态交通流特征的感知能力的有效性。本文后续将重点解决以下问题:在大规模路网中验证算法的可扩展性;探索多智能体协同控制机制。

参考文献

[1] WU Q, SHEN J, YONG B B, et al. Smart fog based workflow for traffic control networks[J]. Future Generation Computer Systems, 2019, 97: 825-835.

[2] BAO Z K, NG S T, YU G, et al. The effect of the built environment on spatial-temporal pattern of traffic congestion in a satellite city in emerging economies[J]. Developments in the Built Environment, 2023, 14: 100173.

[3] KUMAR N, RAHMAN S S, DHAKAD N. Fuzzy inference enabled deep reinforcement learning-based traffic light control for intelligent transportation system[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(8): 4919-4928.

[4] WANG F, LAI G. Fixed-time control design for nonlinear uncertain systems via adaptive method[J]. Systems & Control Letters, 2020, 140: 104704.

[5] EOM M, KIM B I. The traffic signal control problem for intersections: A review[J]. European Transport Research Review, 2020, 12: 1-20.

[6] 牟海维,戚先锋,刘彦昌,等.单交叉口多目标联合优化的信号配时研究[J].电子测量与仪器学报,2020,34(9):62-68.

MOU H W, QI X F, LIU Y CH, et al. Research on signal timing for multi-objective joint optimization at a single intersection[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(9): 62-68.

[7] LADOSZ P, WENG L, KIM M, et al. Exploration in deep reinforcement learning: A survey[J]. Information Fusion, 2022, 85: 1-22.

[8] WANG T, CAO J H, HUSSAIN A. Adaptive traffic signal control for large-scale scenario with cooperative group-based multi-agent reinforcement learning[J]. Transportation Research Part C: Emerging Technologies, 2021, 125: 103046.

[9] CAI C, WEI M. Adaptive urban traffic signal control based on enhanced deep reinforcement learning[J]. Scientific Reports, 2024, 14(1): 14116.

[10] SU H R, ZHONG Y F, CHOW J Y J, et al. EMVLight: A multi-agent reinforcement learning framework for an emergency vehicle decentralized routing and traffic signal control system[J]. Transportation Research Part C: Emerging Technologies, 2023, 146: 103955.

[11] WANG B, HE Z K, SHEN J F, et al. Deep reinforcement learning for traffic light timing optimization[J]. Processes, 2022, 10(11): 2458.

[12] 刘志,曹诗鹏,沈阳,等.基于改进深度强化学习方法的单交叉口信号控制[J].计算机科学,2020,47(12):226-232.

LIU ZH, CAO SH P, SHEN Y, et al. Signal control of a single intersection based on an improved deep reinforcement learning method[J]. Computer Science, 2020, 47(12): 226-232.

[13] BOUKTIF S, CHENIKI A, OUNI A, et al. Deep reinforcement learning for traffic signal control with

consistent state and reward design approach [J]. Knowledge-Based Systems, 2023, 267: 110440.

[14] 金志琦，张正华，姜邦宇，等. 基于改进 D3QN 的单点交叉口信号控制研究[J]. 无线电工程，2025，55(1):28-35.
JIN ZH Q, ZHANG ZH H, JIANG B Y, et al. Research on signal control of a single-point intersection based on the improved D3QN[J]. Radio Engineering, 2025, 55(1): 28-35.

[15] 郭梦杰，任安虎. 基于深度强化学习的单路口信号控制算法[J]. 电子测量技术，2019，42(24):49-52.
GUO M J, REN AN H. Single intersection signal control algorithm based on deep reinforcement learning [J]. Electronic Measurement Technology, 2019, 42(24): 49-52.

[16] ZHENG Y L, LUO J, GAO H, et al. Pri-DDQN: Learning adaptive traffic signal control strategy through a hybrid agent [J]. Complex & Intelligent Systems, 2025, 11(1): 47.

[17] YU B, YIN H, ZHU Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting [J]. ArXiv preprint arXiv: 1709.04875, 2017.

[18] WU Z, PAN S, LONG G, et al. Graph wavenet for deep spatial-temporal graph modeling [J]. ArXiv preprint arXiv:1906.00121, 2019.

[19] 鲁思源，沈琴琴，包银鑫，等. 基于时空感知 Transformer 的交通流预测模型[J]. 电子测量技术，2024，47(10):85-92.
LU S Y, SHEN Q Q, BAO Y X, et al. Traffic flow prediction model based on spatio-temporal aware Transformer [J]. Electronic Measurement Technology, 2024, 47(10):85-92.

[20] TAN X, ZHOU Y, ZHAO L, et al. Short-term traffic flow prediction based on SAE and its parallel training [J]. Applied Intelligence, 2024, 54(4): 3650-3664.

[21] JIA P, CHEN H, ZHANG L, et al. Attention-LSTM based prediction model for aircraft 4-D trajectory[J]. Scientific Reports, 2022, 12(1): 15533.

[22] LIU L, FENG J, LI J, et al. Multi-layer CNN-LSTM network with self-attention mechanism for robust estimation of nonlinear uncertain systems [J]. Frontiers in Neuroscience, 2024, 18: 1379495.

[23] HUANG L B, QU X H. Improving traffic signal control operations using proximal policy optimization [J]. IET Intelligent Transport Systems, 2023, 17(3): 592-605.

[24] 尹刚，朱淼，全鹏程，等. 基于 PID 搜索优化的 CNN-LSTM-Attention 铝电解槽电解温度预测方法研究[J]. 仪器仪表学报，2025，46(1):324-337.
YIN G, ZHU M, QUAN P CH, et al. Research on CNN-LSTM-Attention based aluminum electrolytic cell temperature prediction method optimized by PID search[J]. Chinese Journal of Scientific Instrument, 2025, 46(1): 324-337.

[25] CAO K R, WANG L W, ZHANG S, et al. Optimization control of adaptive traffic signal with deep reinforcement learning [J]. Electronics, 2024, 13(1): 198.

作者简介

刘振航，硕士研究生，主要研究方向为强化学习与智能交通系统。

E-mail:18438625725@163.com

黄德启(通信作者)，博士，副教授，主要研究方向为模式识别与智能系统。

E-mail:dqhuang@126.com