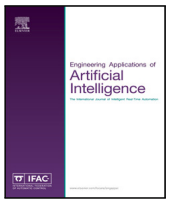
Engineering Applications of Artificial Intelligence 149 (2025) 110440

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

An adaptive traffic signal control scheme with Proximal Policy Optimization based on deep reinforcement learning for a single intersection

Lijuan Wang [a,b], Guoshan Zhang [a,*], Qiaoli Yang [b], Tianyang Han [c]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China
[b] School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, 730070, China
[c] Department of Civil Engineering, Graduate School of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Keywords:
Traffic signal control
Proximal policy optimization
Deep reinforcement learning

ABSTRACT

Adaptive traffic signal control (ATSC) is an important means to alleviate traffic congestion and improve the quality of road traffic. Although deep reinforcement learning (DRL) technology has shown great potential in solving traffic signal control problems, the state representation and reward design, as well as action interval time, still need to be further studied. The advantages of policy learning have not been fully applied in TSC. To address the aforementioned issues, we propose a DRL-based traffic signal control scheme with Poximal Policy Optimization (PPO-TSC). We use the waiting time of vehicles and the queue length of lanes represented the spatiotemporal characteristics of traffic flow to design the simplified traffic states feature vectors, and define the reward function that is consistent with the state. Additionally, we compare and analyze the performance indexes obtained by various methods using action intervals of 5s, 10s, and 15s. The algorithm is implemented based on the Actor-Critic architecture, using the advantage estimation and the clip mechanism to constrain the range of gradient updates. We validate the proposed scheme at a single intersection in Simulation of Urban MObility (SUMO) under two different traffic demand patterns of flat traffic and peak traffic. The experimental results show that the proposed method is significantly better than other compared methods. Specifically, PPO-TSC demonstrates a reduction of 24% in average travel time (ATT), a decrease of 45% in the average time loss (ATL), and an increase of 16% in average speed (AS) compared with the existing methods under peak traffic condition.

1. Introduction

In recent years, with the growth of vehicles and the acceleration of urbanization, urban traffic congestion has become a growing issue in our society. Due to the increasing amount of travel delay and fuel consumption, traffic congestion has caused huge economic loss. According to the report from INRIX company in 2018, traffic jams cost the United States nearly $87 billion, and almost cost each driver an average of $1300 for a year (Goetz, 2019). As a result, to deal with such costly traffic congestion, effective approach aiming at optimizing traffic signal regulation are crucial. Congestions are caused by various factors, such as traffic flow oversaturation, poor design of road infrastructures, and so on. Some factors require complex policies or long-term planning, and controlling road traffic through traffic signals is the economical and effective way and plays an important role in urban traffic management (Wei et al., 2019b).

In the past few decades, traditional traffic control methods (Sims and Dobinson, 1980; Hunt et al., 1982; Mirchandani and Head, 2001) have been widely applied and achieved good effects worldwide. With the continuous changes in traffic conditions, new traffic demands and new problems have emerged. In order to cope with complex and variant traffic problems, intelligent Traffic Signal Control (TSC) methods have also been studied to some extent (Ceylan and Bell, 2004; Qiao et al., 2010; García-Nieto et al., 2012). In addition, Some scholars have also conducted theoretical research on the phase setting of traffic signals (Yang et al., 2018; Jiang and Gao, 2020; Yang and Shi, 2021). Max-Pressure (MP) (Varaiya, 2013) theory in traffic engineering was used to adjust signal control to relieve traffic pressure at intersections, and MP-based TSC technologies was continuously improved to achieve better efficiency (Wang et al., 2022b; Faqir et al., 2023). Deep reinforcement learning (DRL) (Mnih et al., 2015), as an artificial intelligence technology, combines the strong perception of Deep Learning (DL) (LeCun et al., 2015) with the intelligent decision-making of Reinforcement Learning (RL) (Sutton and Barto, 2018), which provides a solution for the perception and decision-making problem of complex

* Corresponding author.
E-mail address: zhanggs@tju.edu.cn (G. Zhang).

https://doi.org/10.1016/j.engappai.2025.110440
Received 13 June 2023; Received in revised form 12 December 2024; Accepted 25 February 2025
Available online 11 March 2025

systems. In recent years, DRL has been studied and applied in many fields (Nguyen and La, 2019; Duan et al., 2020; Liang et al., 2019; Mei et al., 2023). In particular, DRL shows great potential to adaptively learn optimal policy by analyzing traffic signal timing and road traffic dynamics of the intersection (Rasheed et al., 2020). The TSC problem is a sequential decision-making problem, and can be solved by DRL techniques to regulate traffic flow, where the TSC system based on DRL technology is referred as a DRL-TSC agent (Wei et al., 2021).

Initially, most investigations about TSC employed the value-based DRL algorithms (Li et al., 2016; Genders and Razavi, 2016; Van der Pol and Oliehoek, 2016; Wei et al., 2018). Li et al. proposed a DRL approach to design signal timing plans, which can learn the Q-function from the sampled traffic state and the output corresponding to the performance of the traffic system (Li et al., 2016). Genders et al. designed the discrete traffic state encoding (DTSE), which is information dense and is utilized as input of a deep convolutional neural network (CNN) trained using Q-learning with experience replay (Genders and Razavi, 2016). To control coordinately traffic lights, Van et al. combined the popular deep Q-learning algorithm with a coordination algorithm, and concluded that the DRL-TSC approach reduced the travel time compared to earlier work on tabling RL methods (Van der Pol and Oliehoek, 2016). Wei et al. proposed a more effective DRL-TSC model and evaluated their method on a large-scale real traffic dataset gathered from surveillance cameras (Wei et al., 2018). Subsequently, many authors used DRL algorithms to investigate multi-intersection TSC problems and made some progress (Chu et al., 2020; Wang et al., 2021; Liu et al., 2021; Haddad et al., 2022).

On the other hand, scholars have applied policy-based DRL methods to solve TSC problems. Compared with value-based DRL methods, Proximal Policy Optimization (PPO) (Schulman et al., 2017), which is a policy gradient-based DRL algorithm aiming to improve Trust Region Policy Optimization (TRPO) (Schulman et al., 2015), is to directly parameterize the policy and updates the parameters of the action policy through gradient iteration. Shen et al. derived a fixed phase sequence signal scheme and proposed a variable successive phase duration approach based on PPO algorithm (Shen et al., 2020). Ma et al. defined DTSE state as the input of agent and adopted PPO to improve the convergence speed of the model (Ma et al., 2021). An et al. modified the advantage function of PPO algorithm and used the road snapshot as the state input to regulate traffic signals (An and Zhang, 2022). Huang et al. used the long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) network to handle input and output, and utilized PPO algorithm to adjust the signal phase (Huang and Qu, 2023).

In recent years, multi-intersection signal control technology has also received widespread attention. Fang et al. proposed a multi-objective RL model, which combined the multi-objective performance of traffic safety, efficiency, and network coordination (Fang et al., 2023). Liu et al. presented a collaborative MARL method to eliminate the non-stationary effects caused by the adjustment of multi-intersection control strategies (Liu et al., 2023). Mukhtar et al. designed a centralized collaborative graph network to achieve a signal-free corridor (Mukhtar et al., 2023). Wu et al. combined Nash equilibrium and RL game theory to propose a distributed agent-based DRL method, which alleviates traffic congestion by coordinating vehicle travel at road intersections (Wu et al., 2022). Although these studies have proposed methods from different perspectives to improve the efficiency of multi-intersection traffic, there are still some issues worth studying. Especially, it is important for the DRL agent to formulate an appropriate state and reward design because they influence the action and the policy optimization of DRL agent. The representation of the state in DRL models directly affects the computational complexity and the system performance, and high-resolution traffic states significantly raise the storage requirements (Liu et al., 2020).

Additionlly, we notice that DRL agents typically have a green phase interval time for training. Yen et al. (2020) set 5 s as the green phase interval, which is also the choice of most studies. But Huang and Qu (2023) set 30 s as the minimum phase interval. If the interval is too small, it may lead to switch phase frequently for dynamic traffic flow, which may cause green flicking. Excessive phase duration may lead to inflexible control of traffic flow and increase the waiting time of vehicles.

Table 1
Notations.

Notation	Meaning
TSC	Traffic Signal Control
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
MARL	Multi-agent Reinforcement Learning
DQN	Deep Q-network
D3QN	Double Dueling Deep Q-network
A2C	Advantage Actor-critic
PPO	Proximal Policy Optimization
DTSE	Discrete traffic state encoding
SUMO	Simulation of Urban MObility
AWT	Average waiting time
AQL	Average queue length
ATT	Average travel time
ATL	Average time loss
AS	Average speed

Inspired of these studies, in this paper, we investigate the problem of TSC based on PPO algorithm, encode the traffic state using the simplified feature value vector, and design a reward function to enable the agent to optimize traffic flow as much as possible. We utilize the Actor-Critic architecture to implement our scheme. Finally, we consider the impact of action time intervals on traffic efficiency. Our main contributions are:

- An adaptive traffic signal control scheme with Proximal Policy Optimization algorithm based on deep reinforcement learning for a single intersection is proposed, referred to as PPO-TSC.
- The scheme is implemented based on the Actor-Critic architecture and adopts the advantage estimation and the clip mechanism to constrain the range of gradient update. We define the traffic state with a simplified feature vector including waiting time of vehicles and queue length of lanes, and design the reward function consistent with the state.
- The proposed scheme is validated on a microscopic traffic simulation environment-SUMO (Simulation of Urban MObility) under two different traffic demand patterns of flat traffic and peak traffic. Experimental results show that our method achieves pretty good performance in three different action time intervals, i.e., 5 s, 10 s and 15 s.
- We analyze the effects of three different action time intervals on agent performance. The results indicate that PPO-TSC with appropriate action time interval is significantly superior to other methods.

The abbreviations and meanings of commonly used professional terms are listed in the notion table of Table 1.

The remainder of this paper is structured as follows. The literature is reviewed and relevant studies are discussed in Section 2. Section 3 provides DRL theoretical backgrounds and states the problem solved in this research work. Section 4 describes the proposed PPO-TSC including the state representation, action space, reward function and algorithm design. In Section 5, we conduct the experiments, as well as evaluating and discussing the results. Finally, we conclude our works in Section 6.

2. Related work

RL-based control approaches are one of the most active research directions in the field of the adaptive traffic signal control. Tabular RL methods, such as Sarsa and Q-learning, were applied to TSC in the early

Table 2
Overview of DRL approaches for TSC.

Literature	Algorithm	State	Action	Reward	Interval
Wang et al. (2022a)	D3QN	DTSE	4 phases	The change of the number of waiting vehicles	1 s
Liu et al. (2022)	DQN	DTSE	8 phases	The number of queued vehicles, the cumulated waiting times	5 s
Liu and Ding (2022)	DQN	DTSE	4 phases	The change of the number of waiting vehicles	6 s
Mohamad Alizadeh Shabestary and Abdulhai (2022)	DQN	DTSE	8 phases	The change of average cumulative delay of the intersection	1 s
Zheng et al. (2019)	DQN	The number of vehicles	4 phases	Queue length	1 s
Bouktif et al. (2023)	DDQN	Number of vehicles, or queue length, or waiting time	4 phases	Consistent with state definition	15 s
Wu and Hu (2023)	PPO	Vehicle number, and average vehicle speed	0/1 switch	Average queue length and throughput	10 s
Guo and Wang (2021)	PPO+MPC	DTSE	8 phases	The state of trams and the queue states	5 s
Li et al. (2021)	PPO	Queue length, and waiting time	4 phases	The change in the average waiting time	30 s

days (Thorpe and Anderson, 1996; Abdulhai et al., 2003). As the number of traffic states increases, the ability of Tabular RL to regulate traffic flow greatly decreases, and DRL using deep neural network is more suitable for solving complex traffic control problems (Yau et al., 2017). Different value-based DRL algorithms are used in traffic light control approaches to reduce the dimension curse and improve adaptability in large-scale traffic scenarios (Wei et al., 2019a; Chen et al., 2020; Huang et al., 2023). Research works on improving DRL-TSC generally focus on state representation, reward design, action definition, and the structure or learning mechanism of agents. DTSE, which discretizes the approaching lanes into various cells and records the presence and speed of the vehicles in each cell, is commonly used to describe the traffic data around crossings (Wang et al., 2022a; Liu et al., 2022). Liu et al. utilized CNN-based DQN algorithm with DTSE state and designed a distributed TSC scheme to learn from neighbors intersection without sharing experience data (Liu and Ding, 2022). Shabestary et al. developed a DQN-based TSC approach to receive un-preprocessed high-dimensional sensory information, using DTSE state as input and making a action decision per second (Mohamad Alizadeh Shabestary and Abdulhai, 2022). However, some researches indicate that the simplified traffic state also can make agent get good action policies, while ascending learning efficiency of the agent. Zheng et al. believed that too complicated state definition may dramatically slow down learning process without necessarily improving performance (Zheng et al., 2019). Bouktif et al. used the simplified state representation and defined both state and reward in a consistent manner to derive the best policies (Bouktif et al., 2023).

Compared with the widely used DQN algorithm, PPO can directly optimize policy parameters and simplify the implementation of the algorithm. Recently, some scholars have studied the application of PPO algorithm in the field of transportation (Wu and Hu, 2023; Guo and Wang, 2021; Li et al., 2021). Wu et al. proposed a phase-switching TSC scheme based on PPO with multi-head attention module to adapt different intersection topologies and flow distributions (Wu and Hu, 2023). Guo et al. combined model predictive control with PPO to achieve active signal priority control for trams (Guo and Wang, 2021). Li et al. compared the different DRL algorithms in intelligent traffic signal control at a single intersection, and concluded that PPO outperformed the others under unbalanced traffic flow scenarios (Li et al., 2021). Table 2 provides a summary of the recent relevant works.

Summarizing the related works in the literature, we notice that the influence of the difference between two encoding states, i.e., DTSE and the simplified feature vector state, on the performance of the agent has not been studied yet. Furthermore, different researchers adopted different action intervals as decision interval time, but the impact of action intervals on algorithm performance has not been given enough attention. In this paper, we adopt the simplified state representation as the input, design the corresponding reward function, and evaluate the performance of our method at three different action time intervals of 5 s, 10 s, and 15 s.

3. Preliminaries and problem statement

3.1. Background of reinforcement learning

RL is a machine learning method (Sutton and Barto, 2018) aimed at enabling agent to autonomously learn optimal policy by interacting with the environment through a trial-and-error manner. In the learning process, the agent observes the reactions of the environment by attempting different actions and adjusts its policy based on these reactions to achieve maximum long-term returns. The learning process of the agent is modeled as a Markov Decision Process (MDP) which is defined as a five-tuple $\langle S, A, P, \gamma, R\rangle$ as follows:

State Space S: S is a finite set of Markov states s_t, and s_t represents the state of time step t. The agent can depend on the current state s_t instead of the whole history to decide what to do next.

Action Space A: A is a set of legal actions a_t that can be taken by the agent. At time step t, the agent selects the optimal action from action space A following a policy π which maximizes the long-term expected return.

Transition Probablity P: For each triplet $(s_t, a_t, s_{t+1}) \in S \times A \times S$, $P(s_{t+1}|s_t, a_t)$ is the probility distribution of transition to state s_{t+1} after taking action a_t in state s_t.

Discounted Factor γ: γ is the discounted factor, which usually takes values in (0,1) and denotes the importance of future rewards versus immediate rewards.

Reward R: At time step t, the agent obtains a reward R_t by a reward function. R_t denotes the discounted accumulated return in time step t, namely Eq. (1):

$$R_t = \sum_{i=0}^{T-1} \gamma^i r_{t+i} \tag{1}$$

where r_t is the immediate reward at a time step t, T is a total number of time steps.

In value-based DRL methods, a typical solution to the DRL problem is to find a state–action value function (Q-function) that specifies the expected reward R_t for a given state–action pair under the specified policy:

$$Q^{\pi}(s, a) = E[R_t|s, a, \pi] \tag{2}$$

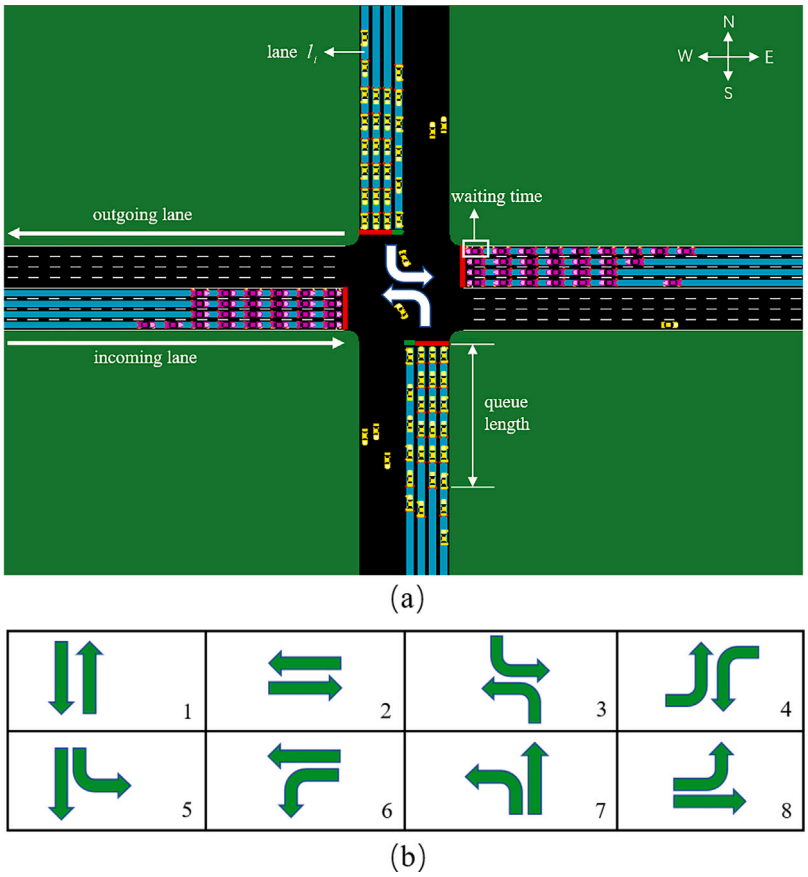


Fig. 1. Intersection model. (a) Traffic movement and descriptions. (b) Phase set.

An optimal Q-function $Q^*(s,a)$ maximizes the expected reward over all states. Having the optimal Q-function, the optimal policy $\pi^*(s)$ is usually determined using a greedy approach:

$$\pi^*(s) = argmax_a Q^*(s,a), \forall s \in S \tag{3}$$

In policy-based DRL methods, the policy $\pi(a_t|s_t;\theta)$ is a probability distribution parameterized by θ. Patameters θ are adjusted in order to maximize the expected reward as follows:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi(a_t|s_t;\theta)\hat{A}(s_t,a_t) \tag{4}$$

where $\hat{A}(s_t,a_t)$ is the advantage function, and α denotes the learning rate. We will specify it in detail in following section.

The DRL algorithm can use deep neural networks to approximate the Q function in value-based methods or the policy function in policy-based methods. Actor-critic DRL combines value-based method and policy-based method, where the actor control the action of agent and the critic evaluate the taken actions.

3.2. Problem statement

We define the optimal problem of traffic signal adaptive control, in which agent can ensure the smallest travel time for vehicles on the road from the origin position to the destination under given traffic conditions. The problem can be mathematically expressed as:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{K} T_i \\ &\text{subject to} \quad g^{min} \le g_k \le g^{max} \end{aligned} \tag{5}$$

where T_i is the total travel time of vehicles approaching the intersection, K is the total number of vehicles. g_k is the green time allocated to phase k, g_{min} and g_{max} are the minimum green phase time and the maximum green phase time, respectively.

4. Methodology

In this section, we firstly establish an environmental model for a four-way intersection. Then, we design the agent and define the state representation, action space, and the reward function. Finally, we present the PPO-TSC algorithm model.

4.1. Intersection environment model

The scenario of an intersection includes traffic flow, incoming and outgoing lanes, and phase setting. In order to make the model more universal, we used a regular intersection for modeling, as shown in Fig. 1. In Fig. 1(a), the intersection has four branches. There are four lanes in each entering edge and outgoing edge, respectively. In the north-south incoming lane, the leftmost lane allows left turns and straight going, the rightmost lane allows right turns and straight going, and the other lanes are straight going. The road design in the east–west direction is similar to the north-south direction. The traffic signal is used to determine whether traffic movement is allowed at a certain time. A green signal indicates the corresponding movement is allowed and a red signal indicates the movement is prohibited. Traffic signals are grouped into several conflict-free phases to regulate traffic movements at the same time. A signal phase is a combination of traffic movement signals that do not conflict, indicating the rights-of-way signal to vehicles by traffic lights. In Fig. 1(a), the phase of turning left from North and the South is activated, and white arrow lines indicate that the vehicles from the North and the South are allowed to turn left to their corresponding exiting lanes. The traffic signal at the intersection has eight phases, as shown in Fig. 1(b). There are no turning right movements in four directions because generally turning right movement is not limited in any signal phases. Noting that, when the number of lanes changes at an intersection, the number of control phases will not change, because the number of directions does not change.

Table 3
Traffic phases and traffic movements.

Action id	Phases	Traffic movements
a_1	NSG	Go straight from North and South
a_2	EWG	Go straight from East and West
a_3	NSGL	Turn left from North and South
a_4	EWGL	Turn left from East and West
a_5	NG	Go straight and turn left from North
a_6	EG	Go straight and turn left from East
a_7	SG	Go straight and turn left from South
a_8	WG	Go straight and turn left from West

4.2. Agent design

We define the basic ingredients, namely the state, the action and the reward, as well as employing PPO algorithm to model PPO-TSC.

• State

Considering the temporal and spatial characteristics of traffic scenarios, we define the following two indexes as traffic states: s_i^{QL} and s_i^{WT}, where s_i^{QL} represents the queue length of vehicles before the stop line on lane i, and s_i^{WT} represents the accumulated waiting time of the head vehicle before the stop line on lane i. The state s_t is defined as follows:

$$s_t = \{s_i^{QL}, s_i^{WT}\} \tag{6}$$

where $i = 1, 2, \ldots, n$ denotes the index of the incoming lanes at the intersection and n is the totol number of incoming lanes around the intersection.

• Action

At each time step, the agent selects an action phase from action set according to the policy and the current traffic condition to minimize the travel time of vehicles. Consider a signal timing plan containing 8 phases, the action space be expressed as follows:

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\} \tag{7}$$

The allowed traffic movements for each phase of the action set are shown in Table 3.

Where N, E, S, W represent the direction of North, East, South and West, respectively, G means green phase, and L means turning left.

• Reward

The reward provides a feedback of the performance of the previous actions to a DRL model. It is important to define the reward appropriately so as to effectively guide the learning process, which accordingly helps to take the best action policy. The higher the reward, the better the action that the agent chooses. In our system, the main goal is to minimize the travel time of the vehicles. To this end, it is necessary to minimize the delay of vehicles. Waiting time of vehicles and queue length of roads are indexes of delay caused by vehicles queuing at the intersection, which directly reflects traffic efficiency. In this paper, we define the reward as the weighted sum of the change in the average waiting time (AWT) and the average queue length (AQL) before and after the action is executed. In addition, the reward function remains consistent in the state representation. The reward function at time step t is defined as:

$$r_t = w_1 \times (WT_{t-\Delta t} - WT_t) + w_2 \times (QL_{t-\Delta t} - QL_t) \tag{8}$$

where w_1 and w_2 is the dynamically assigned weight coefficient, Δt is the action time interval of the model training, WT_t is the AWT of vehicles at time step t which is given by Eq. (9), and QL_t is the AQL on the lane at time step t which is given by Eq. (10).

$$WT_t = \frac{1}{N_t} \sum_{i_t}^{N_t} wt_{i_t} \tag{9}$$

where N_t denotes the total number of vehicles in the road network at time step t, and wt_{i_t} denotes the waiting time of the car i at time step t which is the accumulated stopping time from the time when the vehicle enters the road network to time step t.

$$QL_t = \frac{1}{M} \sum_{i}^{M} ql_{i_t} \tag{10}$$

where M denotes the total number of incoming lanes at the intersection, and ql_{i_t} denotes the queue length of the incoming lane i at time step t which is the total number of halting vehicles for the last time step on the given lane.

The reward R_t of each episode is the accumulated sum of the immidiat reward r_t, which is defined as follows:

$$R_t = \sum_{t=1}^{T} r_t \tag{11}$$

where T is the total training time steps of each iteration episode for the agent.

4.3. Algorithm model

The PPO algorithm theory is based on the theoretical foundation of TRPO. Both of them address the same issue of how to maximize policy improvement through existing data and control the step size of the policy. There are two different implementations of PPO that one is penalty-based PPO and the other is clip-based PPO. Penalty-based PPO adds KL divergence as a penalty function to the optimization function, And the clip-based PPO converts the optimization function into a clip function. Due to the excellent performance of the clip method in PPO, we develop a Traffic Signal Control model based on the clip-based PPO algorithm, named PPO-TSC, as shown in Fig. 2.

In our model, PPO based on Actor-Critic architecture includes two modules, in which Actor is responsible for outputting action policies based on input states, and Critic is responsible for evaluating action policies. The Actor module consists of two policy networks, which are respectively the current actor and the old actor. At time step t, receiving the state s_t, the current actor outputs a probability distribution $\pi_\theta(a_t|s_t) = P[A_t = a_t|S_t = s_t]$ and randomly samples an action a_t based on $\pi_\theta(a_t|s_t)$. The old actor has the identical state input with the current actor, and its output is denoted as $\pi_{\theta_{old}}(a_t|s_t)$. θ and θ_{old} are the parameters of the current actor and the old actor, respectively.

The critic uses the value network to evaluate the quality of the action selected by the current actor. In the critic module, the value network receives the state input s_t and outputs the predicted state value function V_φ, where φ is the model parameters of the critic.

The clip mechanism is introduced to constrain the gap between the old policy $\pi_{\theta_{old}}(a_t|s_t)$ and the current policy $\pi_\theta(a_t|s_t)$ to a range, beyond which a hyperparameter ϵ is used to limit. The objective function of the current policy network can be defined as

$$L_{clip}(\theta) = \hat{E}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{12}$$

where $\hat{E}_t$ denotes the empirical average over a finite batch of samples, ϵ is the clipped surrogate factor, $r_t(\theta)$ denotes the importance sampling ratio which is given by

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{13}$$

and $\hat{A}_t$ is an estimator of the advantage function at time step t. When $\hat{A}_t$ is greater than 0, it indicates that the current policy is superior to the old policy and the probability of the current action should be increased. Conversely, when $\hat{A}_t$ is less than 0, it indicates that the current action is not very good and the probability of the action should be reduced. Whether increasing or decreasing the probability of the current action, it needs to be limited to a range, namely $(1-\epsilon, 1+\epsilon)$. It can be denoted as $\hat{A}_t(s_t, a_t)$ which is written as:

$$\hat{A}_t(s_t, a_t) = r_t + \gamma V_\varphi(s_{t+1}) - V_\varphi(s_t) \tag{14}$$

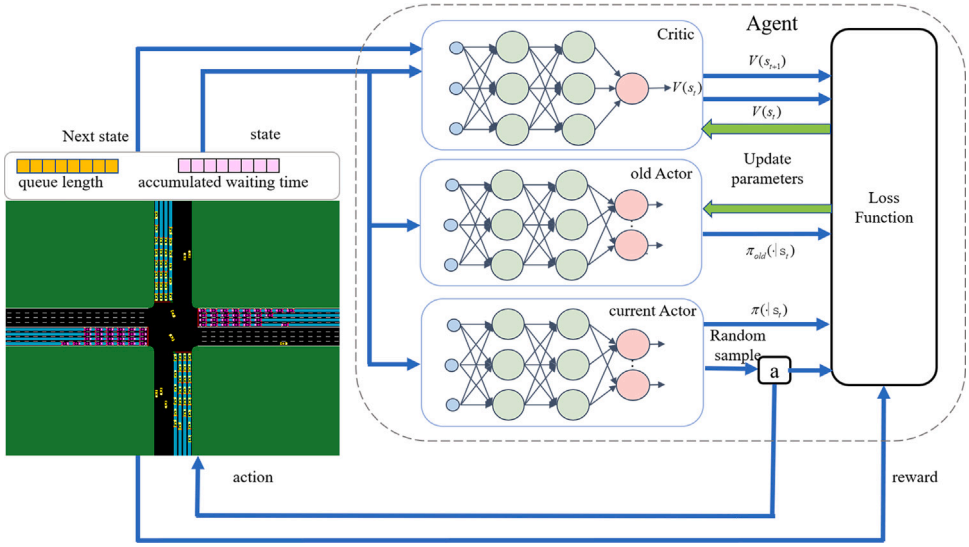


Fig. 2. Archetucture of PPO-TSC.

here r_t denotes the instant reward in time step t, and γ is a hyper-parameter that reduces variance by introducing deviation. The actor updates policy parameters θ by minimizing $L_{clip}(\theta)$ and copies the current parameters θ to θ_{old} for every update. The old policy network does not need to be trained, it periodically copies the parameters from the current policy network.

The goal of the critic network is to minimize the loss function which is given by

$$L_{critic}(\varphi) = (r_t + \gamma V_\varphi(s_{t+1}) - V_\varphi(s_t))^2 \tag{15}$$

where $r_t + \gamma V_\varphi(s_{t+1})$ is the discounted sum of rewards, $V_\varphi(s_t)$ is the predicted state value function, which will gradually approximate the true state value function by the optimization procedure. The algorithm structure can achieve policy evaluation and policy improvement. Due to simplify the dimensions of the state vector, this model adopts a fully connected artificial neural network to implement a three-layer network to avoid overfitting. Our PPO-TSC scheme is shown in Algorithm 1.

Algorithm 1: PPO-TSC Algorithm

Input: state feature vectors s_t, current action a_t
Output: the optimal action a_{t+1}
Initial episode numbers, action time interval ΔT, phase set A and traffic environment
Randomly initial actor network $\pi_\theta(a|s)$ and critic network $V_\varphi(s)$
Initial traffic environment parameters, traffic state s_0 and initial action a_0
for *episode*$= 1, 2, 3, \ldots$ **do**
  for *simulation step* $t <$ *duration every episode* **do**
    Run $\pi_{\theta_{old}}$ to sample trajectories $< s_t, a_t, r_t, s_{t+\Delta T} >$
    Compute $\hat{A}$ according to (14)
    Optimize π_θ according to (12) (13)
    Optimize V_φ according to (15)
    Update actor parameters θ by gradient ascend
    Update critic parameters φ by gradient descend
    $\theta_{old} \longleftarrow \theta$
    Obtain action $a_t \in \pi_\theta(\cdot|s_t)$
    if a_t=*current phase* **then**
      Excute action a_t for ΔT seconds.
    else
      Excute action yellow phase for T_y seconds, then excute action a_t for $\Delta T - T_y$ seconds.
    Obtain the average waiting time and the average queue length of vehicles in the intersection and transfer to a new traffic state $s_{t+\Delta T}$
    calculate R_t
    $t = t + \Delta T$

5. Experiments and analysis

5.1. Experimental settings

Our experiments are conducted on a Linux workstation with Intel Core i7-6700K CPU 4.00 GHz with 8 cores, NVIDIA GeForce GTX 1070, and 8 GB memory. Our traffic simulation platform is SUMO 1.12.0, which is an open source microscopic traffic simulation package-Simulation of Urban Mobility (SUMO). SUMO provides users Traffic Control Interface (TraCI) to control the traffic signal and obtain various traffic information. Meanwhile, we utilize Pytorch 1.8.1, one of the most popular deep learning frameworks, for our experiments.

The experiments take the intersection of Hongyan East Road and Xinwang South Road in Chaoyang District, Beijing, as the traffic simulation scene. The intersection model is shown in Fig. 3, where there are four direction edges, and each edge is 500 m long. There are 4 incoming lanes and 4 outgoing lanes in the north-south direction, and 3 incoming lanes and 3 outgoing lanes in the east–west direction. Our experiments select the traffic flow data from 13:00–15:00 and 18:00–20:00 on a certain workday in July, 2018 (Shen et al., 2020), corresponding to the medium flow and high flow of the simulation environment, namely flat traffic and peak traffic. The traffic demand distribution is shown in Fig. 4. Vehicle routes include going straight movements, and turning movements. Details about traffic flow configurations are shown in Table 4. The interval time ΔT is set as 10 s in the experiments, and the yellow light lasts for 3 s.

The total number of training iterations is 200, and each iteration process lasts for 7200 simulation steps, which implies that each experiment simulates 2 h in real world. PPO-TSC obtains the traffic information of the incoming lanes through D2 detectors in SUMO. To test the performance and adaptability of the proposed scheme, experiments are conducted with respect to both flat traffic and peak traffic. The improvement effect can be analyzed by the metrics of average waiting time (AWT) and average queue length (AQL) in the traffic simulation

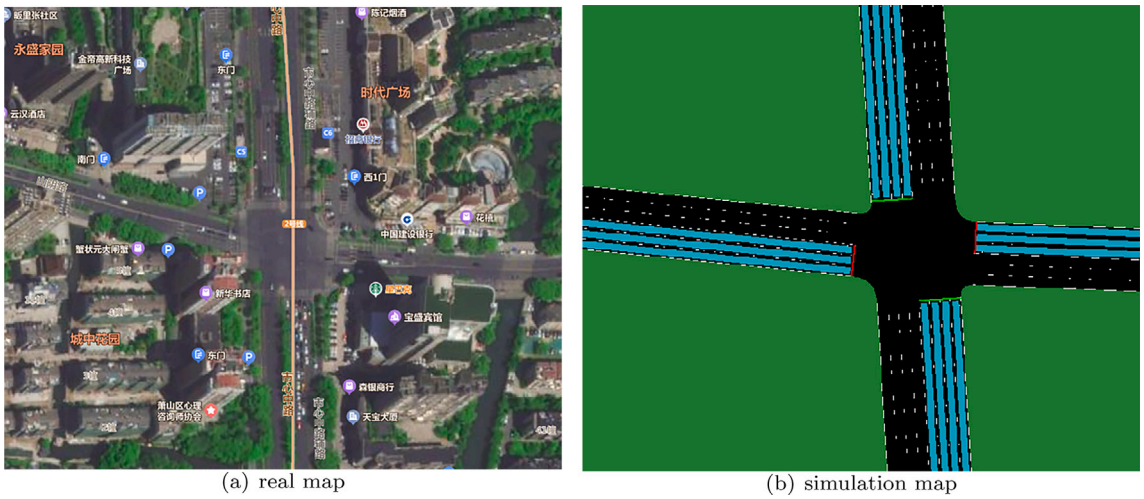


Fig. 3. Schematic diagram of intersection model. (a) Intersection scene in reality. (b) Intersection scene in SUMO.

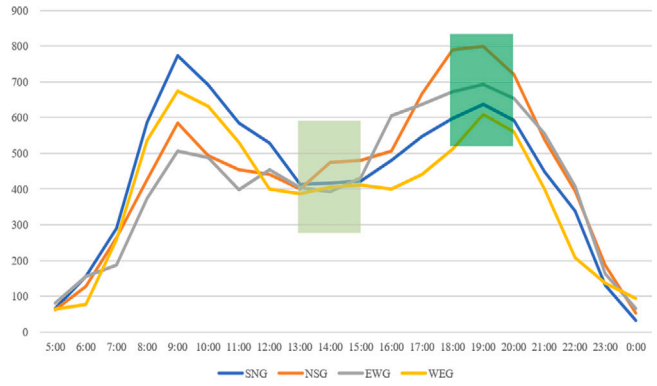


Fig. 4. The traffic demand in a workday, The light green and dark green respectively contain traffic data in medium traffic density (13:00–15:00) and high traffic density (18:00–20:00) scenarios.

Table 4
Parameter setting for traffic environment in SUMO.

Parameters	Value
Lane length	500 m
Vehicle length	5 m
Vehicle speed	13.89 m/s
Vehicle Acceleration	1 m/s^2
Vehicle deceleration	4.5 m/s^2
Minimum gap	2.5 m
Duration of yellow phase	3 s
Duration of green phase	10 s

environment. Table 4 shows the basic traffic parameters of vehicles and lanes.

Considering the stability of the training process, the initial parameters settings of the PPO-TSC algorithm are shown in Table 5.

5.2. Evaluation methods and metrics

To evaluate the effectiveness of our scheme, we compare the performance of the following methods:

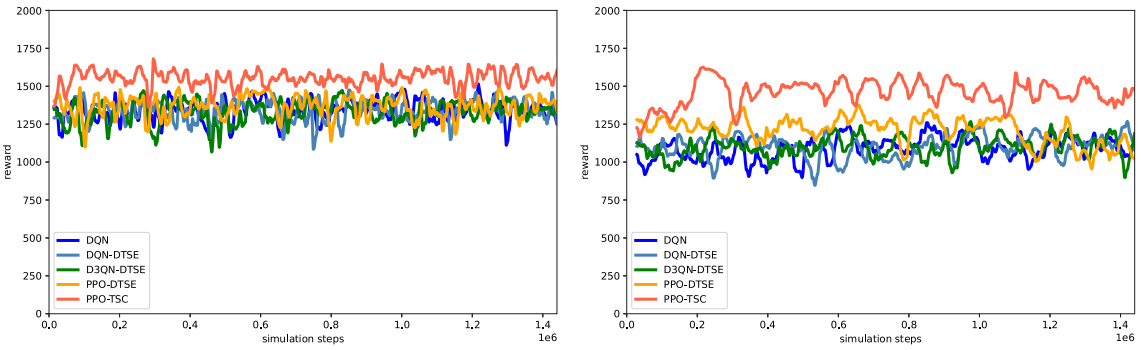
- DQN (Bouktif et al., 2023): This method adopts double Deep Q Network and prioritized experience replay for the agent architecture. It uses states and rewards with simlified and consistent definitions.

Table 5
Parameter setting for DRL algorithm.

Parameters	Value
Total episode	200
Duration every episode	7200
Discount factor γ	0.9
Actor learning rate	0.0001
Critic learning rate	0.001
Clip parameter ϵ	0.2
Grad update	0.5
Batch size	20
Optimizer	Adam

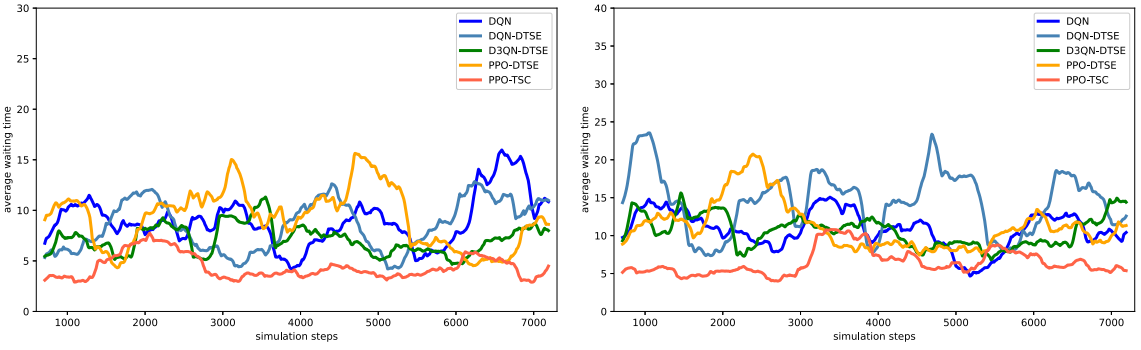
- DQN-DTSE (Liu et al., 2022): This method uses DTSE describing positions and speed of vehicles as state representation, and the weighted sum of some traffic factors as reward.
- D3QN-DTSE (Wang et al., 2022a): This method is a DRL approach of solving the traffic signal timing optimization problem using D3QN algorithm. It uses the position and speed of vehicles as the state and the difference of cumulative waiting time before and after action as reward.
- PPO-DTSE (Ma et al., 2021): This is a policy-based DRL method of solving TSC problem using PPO algorithm. It uses DTSE to define state representation, and the difference of waiting time before and after action with a penalty term as reward. Different from the value-based DRL methods, this method parameterizes and optimizes the policy directly.
- Max-PPO (An and Zhang, 2022): This method modifies the advantage function of PPO algorithm and uses the images as states.
- LSTM-PPO (Huang and Qu, 2023): This method combines the LSTM network with PPO algorithm to adjust the signal control strategy, with a 30 s action interval.
- ELM-MP (Faqir et al., 2023): This method integrates extreme learning machine (ELM) algorithm with MP technology to control traffic signal for a single intersection.
- PPO-TSC: This is the proposed method of this paper.

We evaluate the training of the proposed PPO-TSC and other algorithms under flat traffic and peak traffic, as shown in Fig. 5(a) and (b). Fig. 5(a) shows the training results of the algorithms during flat traffic hours. It can be seen that PPO-TSC still maintains an obvious advantage even if there is the small differences in various methods. In the case of peak traffic in Fig. 5(b), the DRL methods based on value functions, such as DQN, DQN-DTSE, D3QN-DTSE, etc., obtain similar rewards during the training process. However, policy gradient-based



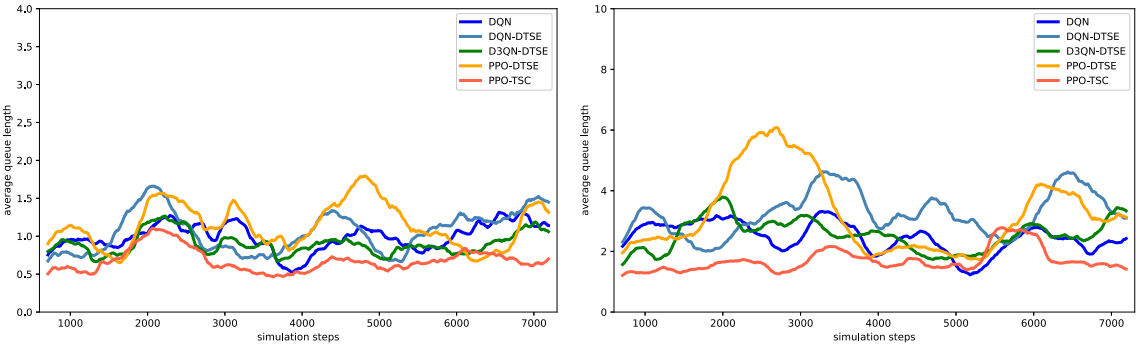
(a) flat traffic (b) peak traffic

Fig. 5. Evaluation of reward of different methods in flat traffic and peak traffic.



(a) flat traffic (b) peak traffic

Fig. 6. Evaluation of average waiting time in flat traffic and peak traffic.



(a) flat traffic (b) peak traffic

Fig. 7. Evaluation of average queue length in flat traffic and peak traffic.

DRL methods, such as PPO-TSC and PPO-DTSE, the rewards are higher than the former during the training process. Among them, PPO-TSC shows a notable superiority. In addition, During the experiment, we observe that the DRL method with DTSE state takes more than 2 times longer than the DRL method with features vector state for training, which means that the former has a slower learning speed.

In order to demonstrate the performance of the PPO-TSC algorithm, we conduct experiments to compare different DRL-TSC methods, and analyze two traffic metrics: AWT and AQL. AWT represents the time that vehicles wait on the road. AQL represents the road occupancy rate caused by vehicle stoppping and waiting. Figs. 6 and 7 demonstrate the variation curves of the three indexes in two traffic modes, where (a) is flat traffic and (b) is peak traffic.

The comparison of AWT of the proposed method with the baseline methods are shown in Fig. 6. When a vehicle travels from its origin to its destination, it is hoped that the waiting time can be as short as possible. From the figure, it can be seen that the orange curve representing PPO-TSC is the lowest in both traffic modes, which means that PPO-TSC obtains the minimum AWT. Moreover, Compared to the curves of other methods, the orange curve represented PPO-TSC changes relatively smoothly, which means that regardless of the traffic conditions, PPO-TSC can keep the traffic flow stable.

Queue length is another important index of traffic efficiency. Seen from Fig. 7(a) and (b), AQL of all methods are relatively small in flat traffic. It can infer that when there is not much traffic intensity at noon, there is rarely a queue of vehicles at the intersection. But during peak

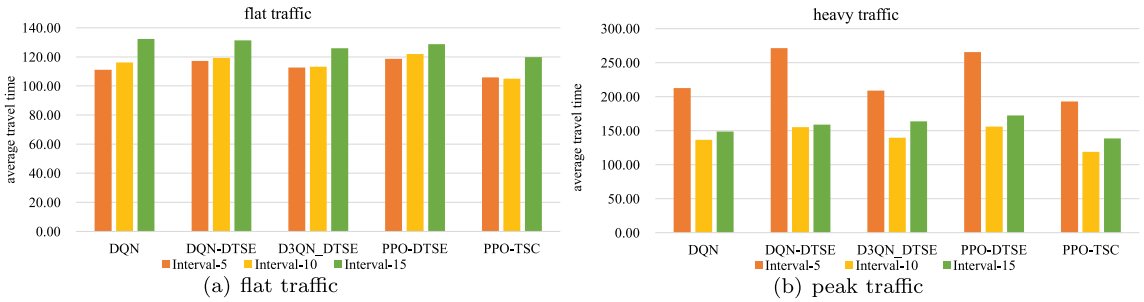


Fig. 8. Evaluation of average travel time in flat traffic and peak traffic with three action time intervals.

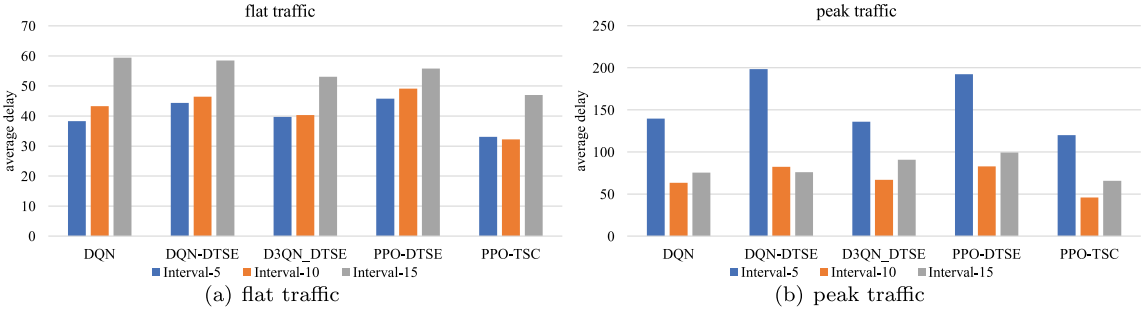


Fig. 9. Evaluation of average timeloss in flat traffic and peak traffic with three action time intervals.

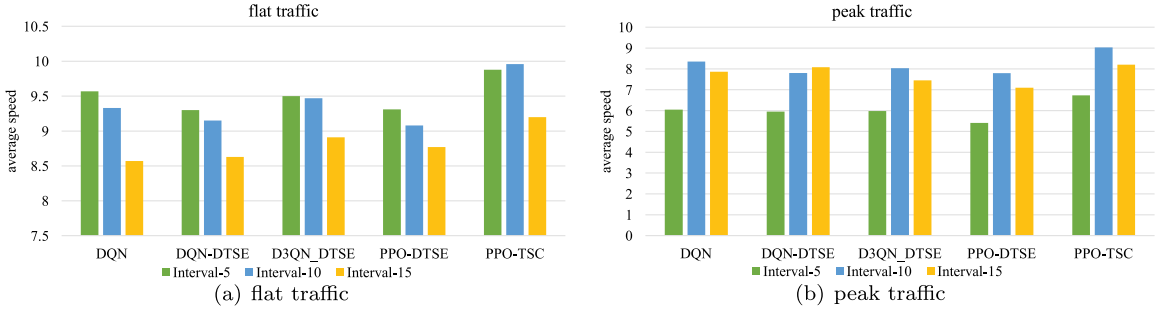


Fig. 10. Evaluation of average speed in flat traffic and peak traffic with three action time intervals.

hours in the evening, the curves exhibit significant fluctuations. AQL curves demonstrate that PPO-TSC exhibits the optimal results among all the methods.

5.3. Effects analysis of three different action time intervals on traffic

In existing DRL-TSC related studies, DRL agents typically have a green phase interval time Δt for training. Different studies selected different green phase action intervals. However, the influence of the size of the action interval on the performance of the algorithm is not considered. We analyze the effects of three action time intervals, namely 5 s, 10 s, and 15 s. Figs. 8, 9 and 10 intuitively show the performance of the DRL-TSC methods for the main traffic metrics, namely average travel time (ATT), average speed (AS), and average time loss (ATL) in sequence, under three different intervals, where (a) is flat traffic, and (b) is peak traffic. The corresponding data results are shown in Table 6. Fig. 8 shows that when the interval is 5 s, the travel time of all methods is not too much under flat traffic. But, during peak traffic periods, compared to the other two time intervals, the travel time of vehicles with 5 s interval significantly increases. Average timeloss in Fig. 9 also shows the similar trend as in Fig. 8.

We can infer that as more and more vehicles approach the intersection in each direction, the agent will increase the number of phase switches based on the changes in traffic dynamics on the roads, resulting in a rapid increase in the yellow light time and thus increasing the vehicle's timeloss time on the road. On the contrary, if the interval time is too long, it may occurs that the lanes where the queue vehicles has been cleared are still given green lights, but other lanes where

Table 6
PPO-TSC model training effect comparison with three action time intervals.

Method	ATT		ATL		AS			
		Flat	Peak	Flat	Peak	Flat	Peak
Interval-5	DQN	111.13	212.55	38.28	139.69	9.57	6.04
	DQN-DTSE	117.18	271.32	44.34	198.46	9.30	5.95
	D3QN-DTSE	112.56	208.91	39.72	136.04	9.50	5.98
	PPO-DTSE	118.64	265.43	45.78	192.55	9.31	5.41
	PPO-TSC	**105.85**	**192.94**	**33.01**	**120.05**	**9.88**	**6.73**
Interval-10	DQN	116.15	136.37	43.30	63.47	9.33	8.35
	DQN-DTSE	119.26	155.29	46.41	82.41	9.15	7.80
	D3QN-DTSE	113.15	139.78	40.29	66.98	9.47	8.03
	PPO-DTSE	121.99	155.83	49.13	82.94	9.08	7.79
	PPO-TSC	**105.06**	**118.82**	**32.22**	**45.94**	**9.96**	**9.03**
Interval-15	DQN	132.26	148.57	59.42	75.67	8.57	7.86
	DQN-DTSE	131.37	158.90	58.50	76.02	8.63	8.08
	D3QN-DTSE	125.91	163.68	53.06	90.80	8.91	7.45
	PPO-DTSE	128.64	172.31	55.79	99.42	8.77	7.10
	PPO-TSC	**119.87**	**138.69**	**47.01**	**65.80**	**9.20**	**8.20**

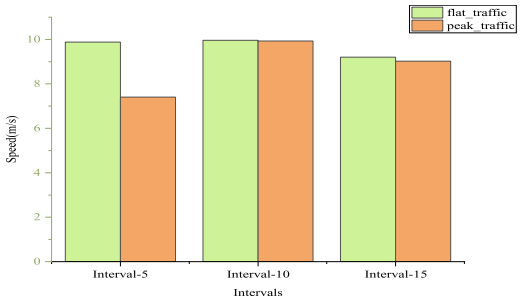


Fig. 11. Average speed of PPO-TSC in flat traffic and peak traffic with three action time intervals.

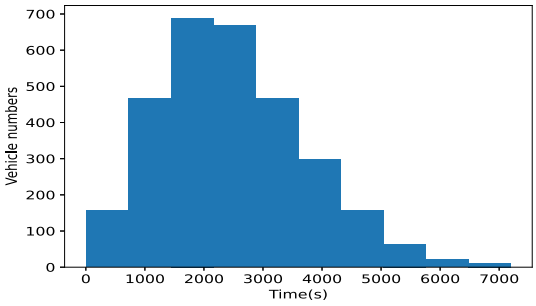


Fig. 12. Traffic distribution under non-stationary condition.

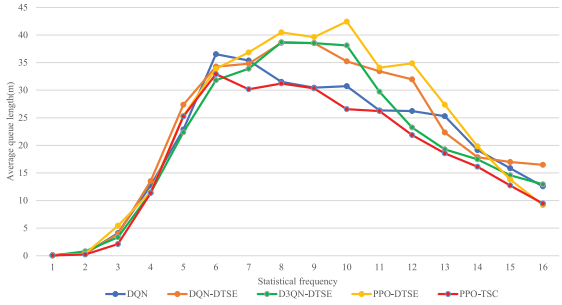


Fig. 13. Average queue length under non-stationary traffic.

vehicles are queuing and waiting are given red lights. Even if the period is short, it will increase the timeloss of queuing vehicles. As a result, a moderate training interval time is crucial for the effectiveness of algorithm performance. The experimental results indicate that when the interval is selected as 10 s, the approach performs the best control effect on road traffic flow. Fig. 10 demonstrates the average speed curves under both traffic patterns. In both flat and peak traffic, PPO-TSC maintains the maximum speed value in all methods.

Table 6 lists the results of five methods under three action time intervals. As shown in Table 6, PPO-TSC proposed in this paper performs optimally in various indexes during both flat and peak traffic. Under flat traffic conditions, compared to PPO-DTSE, ATT decreased by 14%, ATL decreased by 34%, and AS increased by 10%. Under peak flow conditions, the performance of PPO-TSC still has significant advantages. Compared to the PPO-DTSE, ATT has decreased by 24%, ATL has decreased by 45%, and AS has increased by 16%. According to the data in the table, it can be seen that even with different action time intervals, the effect of PPO-TSC on traffic flow is still significantly better than other methods. On the other hand, we compare the performance of flat traffic and peak traffic. From Table 6 and the above two figures, it can be seen that when the traffic volume is not high, traffic flow situation is improved, but the effect is not remarkable. During peak traffic hours when traffic density increases, PPO-TSC significantly improves various indexes.

We compare AS of PPO-TSC in flat traffic and peak traffic with three action time intervals, as shown in Fig. 11. Under flat traffic condition, due to the low density of vehicles, PPO-TSC with different time intervals get almost the same speed and the effect of interval time on speed is a little slight. AS of 10 s is 9.96 m/s, while The values of AS of 5 s and 15 s are 9.88 m/s and 9.20 m/s. Under peak traffic condition, AS using different action intervals shows significant difference. Furthermore, when the interval time is 10 s, AS of PPO-TSC is the highest. Specifically, AS of 10 s is 9.03 m/s, while AS of 6.73 m/s, and AS of 15 s is 8.20 m/s. PPO-TSC with 10 s action interval increases AS by 34% compared to 5 s and by 10% compared to 15 s.

We contrast the proposed method using 10 s action interval with the following methods, namely Max-PPO, LSTM-PPO, and ELM-MP, in the two traffic demand scenarios at the single intersection presented in this paper. ATT, ATL, and AS obtained by these methods are listed in Table 7. It can be seen that PPO-TSC achieves the highest AS, the lowest ATT and the least ATL among these methods under both flat and peak traffic. In terms of speed, under the peak traffic condition, AS obtained by these methods shows a decrease in different degrees. Specifically, Max-PPO decreases by 8%, LSTM-PPO decreases by 31%, ELM-MP decreases by 27%, and PPO-TSC only decreases by 10%. It indicates that PPO-TSC still achieves the best performance under peak traffic condition, and also demonstrates that PPO-TSC has good regulating ability for different traffic densities. Summarizing of the experimental results, in flat traffic, PPO-TSC has a decrease by 11% on ATT, a reduction by 28% on ATL, and an increase by 8% on AS compared with the better performing ELM-MP. In peak traffic, PPO-TSC has a decrease by 24% on ATT, a reduction by 45% on ATL, and an increase by 16% on AS compared with the better performing Max-PPO.

5.4. Performance under non-stationary conditions

Unlike stationary conditions, where traffic flow parameters such as speed, density, and volume remain relatively constant over time, non-stationary environments are characterized by abrupt traffic fluctuations. These changes are caused by some factors, including rush-hour congestion, road construction, traffic accidents, weather conditions, and so on. When the distribution of traffic changes, the smoothness of traffic on the road will change. To verify the performance of the proposed method in non-stationary traffic conditions, we simulated traffic dynamics within 2 h. The arrival of vehicles in all directions follows Poisson distribution, with vehicles going from sparse to dense and gradually dissipating, as shown in Fig. 12. We use AQL and AWT to evaluate four baseline methods and the proposed method. According to the changing traffic dynamics, we make statistics every 450 s. From Fig. 13, it can be seen that when the distribution of vehicles changes, AQL of vehicles on the road also changes in line with the trend of traffic distribution. Compared to other methods, PPO-TSC has the smallest AQL and is relatively stable. Fig. 14 shows AWT under non-stationary traffic conditions. When the traffic distribution fluctuates, AWT of vehicles is also affected. The results show that AWT of PPO-TSC is the smallest and the variation range is also the smallest, which indicates that our method can adapt well to the non-stationary phenomena in traffic.

5.5. Ablation analysis

To assess the influence of different factors on the results, we conduct the following ablation studies. We separately adopt waiting time and queue length as state characteristics and reward objectives, denoted as PPO-TSC(w) and PPO-TSC(q), respectively. They use 10 s as the action interval time. We validated three indexes including AS, ATT, and ATL under both flat and peak traffic conditions, and the results are shown in Table 8. This table shows the values of the performance

Table 7
Evaluation of the performance indexes of different methods in flat traffic and peak traffic.

Traffic condition		Max-PPO (An and Zhang, 2022)	LSTM-PPO (Huang and Qu, 2023)	ELM-MP (Faqir et al., 2023)	PPO-TSC (Ours)
Flat	ATT	133.75	123.17	117.50	**105.06**
	ATL	60.90	50.30	44.65	**32.22**
	AS	8.48	9.02	9.25	**9.96**
Peak	ATT	156.35	186.98	176.62	**118.92**
	ATL	83.46	114.08	103.73	**45.94**
	AS	7.81	6.90	7.29	**9.03**

Table 8
Performance indexes comparison of ablation analysis in flat traffic and peak traffic with 10 s.

Mode	Methods	Queue length	Waiting time	AS	ATT	ATL
Flat traffic	PPO-TSC(w)		✓	9.23	118.11	45.25
	PPO-TSC(q)	✓		9.32	119.44	46.57
	PPO-TSC	✓	✓	**9.96**	**105.06**	**32.33**
Peak traffic	PPO-TSC(w)		✓	8.21	149.77	76.89
	PPO-TSC(q)	✓		8.14	153.12	80.21
	PPO-TSC	✓	✓	**9.03**	**118.82**	**45.94**

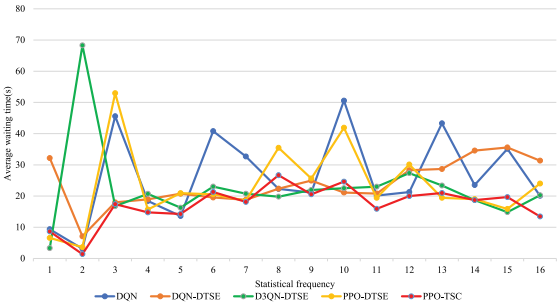


Fig. 14. Average waiting time under non-stationary traffic.

indexes of three algorithms with 10 s action interval in two traffic modes. When using a single waiting time or queue length as state representation and reward design, all indexes exhibit a decrease to different extent. Under flat traffic condition, PPO-TSC with 10 s action interval performs the best compared to the other two methods. Under peak traffic condition, due to the high density of vehicles, the difference between index values is more significant than under flat traffic. PPO-TSC still exhibits significantly better performance than PPO-TSC(w) and PPO-TSC(q) using a single traffic feature. We conclude that our method combines waiting time and queue length to design the state and reward, and can achieve the best control performance.

5.6. Convergence analysis

We analyze the convergence of the proposed method in the training process under different traffic conditions, as shown in Fig. 15. We conduct 200 epochs of training on the model and it tends to converge after approximately 20 epochs. Under flat traffic, the model can maintain a relatively stable state after convergence. Under peak traffic, although the model still exhibits some fluctuation after convergence, the overall trend converges to a stable range. In addition, as shown in Fig. 16, the loss curves of the actor and critic gradually decline and eventually tended to zero. This further confirms the convergence and stability of the model.

6. Conclusions

In this paper, we proposed an adaptive traffic signal control scheme with Proximal Policy Optimization based on deep reinforcement learning, namely PPO-TSC. This scheme uses the queue length of lanes and the waiting time of vehicles as simplified state feature vectors, and designs a reward function in a consistent manner. We compared and analyzed the influence of three action time intervals of 5 s, 10 s, and 15 s on traffic flow and designed a TSC agent based on the PPO algorithm. Compared with complex DTSE states, simplified state vectors make algorithm training more efficient, and selecting an appropriate action interval time (i.e 10 s) can enable the TSC agent to achieve the better performance. We construct a real-world single signalized intersection in SUMO and evaluate the performance of our method under both flat and peak traffic conditions. The experimental results show that the proposed method performs the best. Specifically, compared with the existing methods, our method reduces average travel time by 24%, decreases average time loss by 45%, and increases average speed by 16% under peak traffic condition. In the future, we will fully utilize the good scalability of PPO model and consider upgrading the proposed method to multi-agent collaborative control through distributed or centralized control scheme to further improve the traffic efficiency of the urban road network. Distributed control may lead to local optimization and computation intensive, while centralized control requires high computing power and may result in computational resource bottlenecks. Combining the advantages of both and adopting the architecture of hierarchical control is expected to solve the signal control problem of complex multi-intersections. For complex road networks, distributed decision-making and centralized learning may be used to extend the proposed algorithm in this paper to multi-intersection scenarios. In addition, adaptive learning techniques such as transfer learning or continuous learning enable models to adapt to dynamic and complex traffic environments without the need to retrain from scratch.

CRediT authorship contribution statement

Lijuan Wang: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Guoshan Zhang:** Writing – review & editing, Supervision, Resources. **Qiaoli Yang:** Writing – review & editing. **Tianyang Han:** Software.

Declaration of competing interest

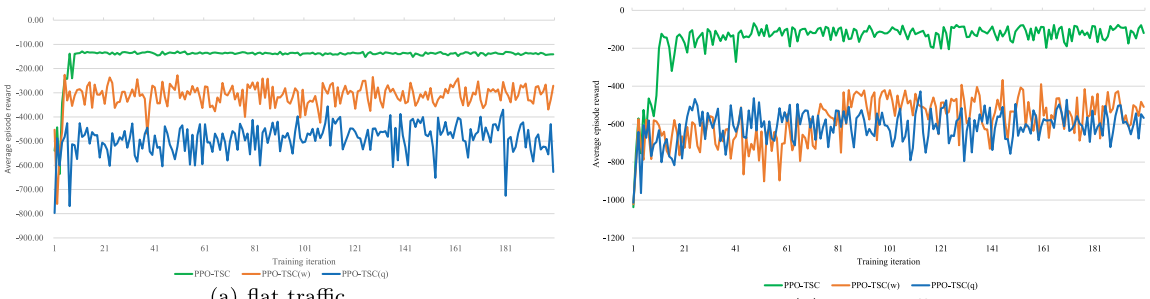
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by the National Natural Science Foundation of China under Grant No. 62073237 and No. 72171106; Key Program of Natural Science Foundation of Gansu Province of China under Grant No. 20JR5RA428.

Data availability

Data will be made available on request.



(a) flat traffic (b) peak traffic

Fig. 15. Training curves of PPO-TSC models under in flat traffic and peak traffic.

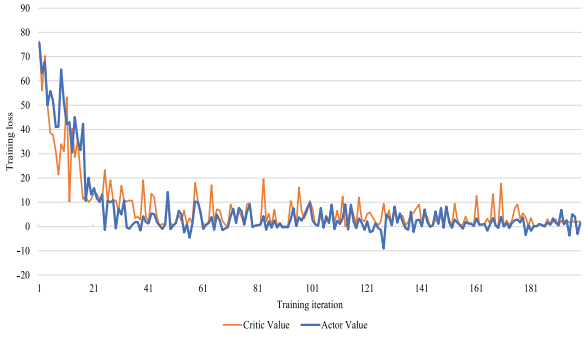


Fig. 16. Training loss curves of actor and critic.

References

Abdulhai, B., Pringle, R., Karakoulas, G.J., 2003. Reinforcement learning for true adaptive traffic signal control. J. Transp. Eng. 129 (3), 278–285. http://dx.doi.org/10.1061/(ASCE)0733-947X(2003)129:3(278).

An, Y., Zhang, J., 2022. Traffic signal control method based on modified proximal policy optimization. In: 2022 10th International Conference on Traffic and Logistic Engineering. ICTLE, IEEE, pp. 83–88. http://dx.doi.org/10.1109/ICTLE55577.2022.9901894.

Bouktif, S., Cheniki, A., Ouni, A., El-Sayed, H., 2023. Deep reinforcement learning for traffic signal control with consistent state and reward design approach. Knowl.-Based Syst. 267, 110440. http://dx.doi.org/10.1016/j.knosys.2023.110440.

Ceylan, H., Bell, M.G., 2004. Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing. Transp. Res. Part B: Methodol. 38 (4), 329–342. http://dx.doi.org/10.1016/S0191-2615(03)00015-8.

Chen, C., Wei, H., Xu, N., Zheng, G., Yang, M., Xiong, Y., Xu, K., Li, Z., 2020. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, pp. 3414–3421. http://dx.doi.org/10.1609/aaai.v34i04.5744.

Chu, T., Wang, J., Codecà, L., Li, Z., 2020. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans. Intell. Transp. Syst. 21 (3), 1086–1095. http://dx.doi.org/10.1109/TITS.2019.2901791.

Duan, J., Shi, D., Diao, R., Li, H., Wang, Z., Zhang, B., Bian, D., Yi, Z., 2020. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. IEEE Trans. Power Syst. 35 (1), 814–817. http://dx.doi.org/10.1109/TPWRS.2019.2941134.

Fang, J., You, Y., Xu, M., Wang, J., Cai, S., 2023. Multi-objective traffic signal control using network-wide agent coordinated reinforcement learning. Expert Syst. Appl. 229, 120535.

Faqir, N., Loqman, C., Boumhidi, J., 2023. Combined extreme learning machine and max pressure algorithms for traffic signal control. Intell. Syst. Appl. 19, 200255. http://dx.doi.org/10.1016/j.iswa.2023.200255.

García-Nieto, J., Alba, E., Carolina Olivera, A., 2012. Swarm intelligence for traffic light scheduling: Application to real urban areas. Eng. Appl. Artif. Intell. 25 (2), 274–283. http://dx.doi.org/10.1016/j.engappai.2011.04.011.

Genders, W., Razavi, S., 2016. Using a deep reinforcement learning agent for traffic signal control. arXiv preprint arXiv:1611.01142.

Goetz, A.R., 2019. Transport challenges in rapidly growing cities: is there a magic bullet? Transp. Rev. 39 (6), 701–705. http://dx.doi.org/10.1080/01441647.2019.1654201.

Guo, G., Wang, Y., 2021. An integrated MPC and deep reinforcement learning approach to trams-priority active signal control. Control Eng. Pract. 110, 104758. http://dx.doi.org/10.1016/j.conengprac.2021.104758.

Haddad, T.A., Hedjazi, D., Aouag, S., 2022. A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control. Eng. Appl. Artif. Intell. 114, 105019. http://dx.doi.org/10.1016/j.engappai.2022.105019.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Huang, H., Hu, Z., Wang, Y., Lu, Z., Wen, X., 2023. Intersec2vec-TSC: Intersection representation learning for large-scale traffic signal control. IEEE Trans. Intell. Transp. Syst. 1–13. http://dx.doi.org/10.1109/TITS.2023.3340153.

Huang, L., Qu, X., 2023. Improving traffic signal control operations using proximal policy optimization. IET Intell. Transp. Syst. 17 (3), 592–605. http://dx.doi.org/10.1049/itr2.12286.

Hunt, P., Robertson, D., Bretherton, R., Royle, M.C., 1982. The SCOOT on-line traffic signal optimisation technique. Traffic Eng. Control 23 (4).

Jiang, X., Gao, S., 2020. Signal control method and performance evaluation of an improved displaced left-turn intersection design in unsaturated traffic conditions. Transp. B: Transp. Dyn. 8 (1), 264–289. http://dx.doi.org/10.1080/21680566.2020.1764410.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Li, Y., He, J., Gao, Y., 2021. Intelligent traffic signal control with deep reinforcement learning at single intersection. In: 2021 7th International Conference on Computing and Artificial Intelligence. ICCAI '21, pp. 399–406. http://dx.doi.org/10.1145/3467707.3467767.

Li, L., Lv, Y., Wang, F.-Y., 2016. Traffic signal timing via deep reinforcement learning. IEEE/ CAA J. Autom. Sin. 3 (3), 247–254. http://dx.doi.org/10.1109/JAS.2016.7508798.

Liang, X., Du, X., Wang, G., Han, Z., 2019. A deep reinforcement learning network for traffic light cycle control. IEEE Trans. Veh. Technol. 68 (2), 1243–1253. http://dx.doi.org/10.1109/TVT.2018.2890726.

Liu, B., Ding, Z., 2022. A distributed deep reinforcement learning method for traffic light control. Neurocomputing 490, 390–399. http://dx.doi.org/10.1016/j.neucom.2021.11.106.

Liu, J., Qin, S., Su, M., Luo, Y., Wang, Y., Yang, S., 2023. Multiple intersections traffic signal control based on cooperative multi-agent reinforcement learning. Inform. Sci. 647, 119484.

Liu, F., Tang, R., Li, X., Zhang, W., Ye, Y., Chen, H., Guo, H., Zhang, Y., He, X., 2020. State representation modeling for deep reinforcement learning based recommendation. Knowl.-Based Syst. 205, 106170. http://dx.doi.org/10.1016/j.knosys.2020.106170.

Liu, S., Wu, G., Barth, M., 2022. A complete state transition-based traffic signal control using deep reinforcement learning. In: 2022 IEEE Conference on Technologies for Sustainability (SusTech). IEEE, pp. 100–107. http://dx.doi.org/10.1109/SusTech53338.2022.9794168.

Liu, J., Zhang, H., Fu, Z., Wang, Y., 2021. Learning scalable multi-agent coordination by spatial differentiation for traffic signal control. Eng. Appl. Artif. Intell. 100, 104165. http://dx.doi.org/10.1016/j.engappai.2021.104165.

Ma, Z., Cui, T., Deng, W., Jiang, F., Zhang, L., 2021. Adaptive optimization of traffic signal timing via deep reinforcement learning. J. Adv. Transp. 2021, 6616702. http://dx.doi.org/10.1155/2021/6616702.

Mei, P., Karimi, H.R., Xie, H., Chen, F., Huang, C., Yang, S., 2023. A deep reinforcement learning approach to energy management control with connected information for hybrid electric vehicles. Eng. Appl. Artif. Intell. 123, 106239. http://dx.doi.org/10.1016/j.engappai.2023.106239.

Mirchandani, P., Head, L., 2001. A real-time traffic signal control system: architecture, algorithms, and analysis. Transp. Res. Part C: Emerg. Technol. 9 (6), 415–432. http://dx.doi.org/10.1016/S0968-090X(00)00047-4.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529–533. http://dx.doi.org/10.1038/nature14236.

Mohamad Alizadeh Shabestary, S., Abdulhai, B., 2022. Adaptive traffic signal control with deep reinforcement learning and high dimensional sensory inputs: Case study and comprehensive sensitivity analyses. IEEE Trans. Intell. Transp. Syst. 23 (11), 20021–20035. http://dx.doi.org/10.1109/TITS.2022.3179893.

Mukhtar, H., Afzal, A., Alahmari, S., Yonbawi, S., 2023. CCGN: Centralized collaborative graphical transformer multi-agent reinforcement learning for multi-intersection signal free-corridor. Neural Netw. 166, 396–409.

Nguyen, H., La, H., 2019. Review of deep reinforcement learning for robot manipulation. In: 2019 Third IEEE International Conference on Robotic Computing. IRC, IEEE, pp. 590–595. http://dx.doi.org/10.1109/IRC.2019.00120.

Van der Pol, E., Oliehoek, F.A., 2016. Coordinated deep reinforcement learners for traffic light control. Proc. Learn. Inference Control. Multi- Agent Systems (At NIPS 2016) 8, 21–38.

Qiao, J., Yang, N., Gao, J., 2010. Two-stage fuzzy logic controller for signalized intersection. IEEE Trans. Syst. Man, Cybern.- Part A: Syst. Humans 41 (1), 178–184. http://dx.doi.org/10.1109/TSMCA.2010.2052606.

Rasheed, F., Yau, K.-L.A., Noor, R.M., Wu, C., Low, Y.-C., 2020. Deep reinforcement learning for traffic signal control: A review. IEEE Access 8, 208016–208044. http://dx.doi.org/10.1109/ACCESS.2020.3034141.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P., 2015. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. Vol. 37, PMLR, pp. 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Shen, S., Shen, G., Shen, Y., Liu, D., Yang, X., Kong, X., 2020. PGA: An efficient adaptive traffic signal timing optimization scheme using actor-critic reinforcement learning algorithm. KSII Trans. Internet Inf. Syst. (TIIS) 14 (11), 4268–4289. http://dx.doi.org/10.3837/tiis.2020.11.002.

Sims, A.G., Dobinson, K.W., 1980. The sydney coordinated adaptive traffic (SCAT) system philosophy and benefits. IEEE Trans. Veh. Technol. 29 (2), 130–137. http://dx.doi.org/10.1109/T-VT.1980.23833.

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction. MIT Press.

Thorpe, T.L., Anderson, C.W., 1996. Tra c light control using sarsa with three state representations. Tech. Rep. Citeseer.

Varaiya, P., 2013. Max pressure control of a network of signalized intersections. Transp. Res. Part C: Emerg. Technol. 36, 177–195. http://dx.doi.org/10.1016/j.trc.2013.08.014.

Wang, B., He, Z., Sheng, J., Chen, Y., 2022a. Deep reinforcement learning for traffic light timing optimization. Process. 10 (11), http://dx.doi.org/10.3390/pr10112458.

Wang, X., Ke, L., Qiao, Z., Chai, X., 2021. Large-scale traffic signal control using a novel multiagent reinforcement learning. IEEE Trans. Cybern. 51 (1), 174–187. http://dx.doi.org/10.1109/TCYB.2020.3015811.

Wang, X., Yin, Y., Feng, Y., Liu, H.X., 2022b. Learning the max pressure control for urban traffic networks considering the phase switching loss. Transp. Res. Part C: Emerg. Technol. 140, 103670. http://dx.doi.org/10.1016/j.trc.2022.103670.

Wei, H., Chen, C., Zheng, G., Wu, K., Xu, K., Gayah, V., Li, Z., 2019a. Presslight: Learning max pressure control for signalized intersections in arterial network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1290–1298.

Wei, H., Zheng, G., Gayah, V., Li, Z., 2019b. A survey on traffic signal control methods. arXiv preprint arXiv:1904.08117.

Wei, H., Zheng, G., Gayah, V., Li, Z., 2021. Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. ACM SIGKDD Explor. Newsl. 22 (2), 12–18. http://dx.doi.org/10.1145/3447556.3447565.

Wei, H., Zheng, G., Yao, H., Li, Z., 2018. Intellilight: A reinforcement learning approach for intelligent traffic light control. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, Association for Computing Machinery, pp. 2496–2505. http://dx.doi.org/10.1145/3219819.3220096.

Wu, Z., Hu, J., 2023. PhaseLight: An universal and practical traffic signal control algorithms based on reinforcement learning. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 4738–4743. http://dx.doi.org/10.1109/ITSC57777.2023.10422109.

Wu, Q., Wu, J., Shen, J., Du, B., Telikani, A., Fahmideh, M., Liang, C., 2022. Distributed agent-based deep reinforcement learning for large scale traffic signal control. Knowl.-Based Syst. 241, 108304.

Yang, Q., Shi, Z., 2021. The queue dynamics of protected/permissive left turns at pre-timed signalized intersections. Phys. A 562, 125406. http://dx.doi.org/10.1016/j.physa.2020.125406.

Yang, Q., Shi, Z., Yu, S., Zhou, J., 2018. Analytical evaluation of the use of left-turn phasing for single left-turn lane only. Transp. Res. Part B: Methodol. 111, 266–303. http://dx.doi.org/10.1016/j.trb.2018.03.013.

Yau, K.-L.A., Qadir, J., Khoo, H.L., Ling, M.H., Komisarczuk, P., 2017. A survey on reinforcement learning models and algorithms for traffic signal control. ACM Comput. Surv. http://dx.doi.org/10.1145/3068287.

Yen, C.-C., Ghosal, D., Zhang, M., Chuah, C.-N., 2020. A deep on-policy learning agent for traffic signal control of multiple intersections. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 1–6. http://dx.doi.org/10.1109/ITSC45102.2020.9294471.

Zheng, G., Zang, X., Xu, N., Wei, H., Yu, Z., Gayah, V., Xu, K., Li, Z., 2019. Diagnosing reinforcement learning for traffic signal control. arXiv preprint arXiv:1905.04716.