Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation

Yu Han, Meng Wang, Ludovic Leclercq

Communications in Transportation Research 3 (2023) 100104

Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Review

Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation

Yu Han [a,*], Meng Wang [b], Ludovic Leclercq [c]

[a] School of Transportation, Southeast University, Nanjing, 211189, China
[b] Faculty of Transport and Traffic Sciences, Technische Universität Dresden, Dresden, Saxony, 01067, Germany
[c] LICIT-ECO7, Université Gustave Eiffel, ENTPE, Lyon, F-69675, France

ARTICLE INFO

Keywords:
Reinforcement learning
Road traffic control
Learning cost
Transferability
Sim-to-real transfer

ABSTRACT

In recent years, the advancement of artificial intelligence techniques has led to significant interest in reinforcement learning (RL) within the traffic and transportation community. Dynamic traffic control has emerged as a prominent application field for RL in traffic systems. This paper presents a comprehensive survey of RL studies in dynamic traffic control, addressing the challenges associated with implementing RL-based traffic control strategies in practice, and identifying promising directions for future research. The first part of this paper provides a comprehensive overview of existing studies on RL-based traffic control strategies, encompassing their model designs, training algorithms, and evaluation methods. It is found that only a few studies have isolated the training and testing environments while evaluating their RL controllers. Subsequently, we examine the challenges involved in implementing existing RL-based traffic control strategies. We investigate the learning costs associated with online RL methods and the transferability of offline RL methods through simulation experiments. The simulation results reveal that online training methods with random exploration suffer from high exploration and learning costs. Additionally, the performance of offline RL methods is highly reliant on the accuracy of the training simulator. These limitations hinder the practical implementation of existing RL-based traffic control strategies. The final part of this paper summarizes and discusses a few existing efforts which attempt to overcome these challenges. This review highlights a rising volume of studies dedicated to mitigating the limitations of RL strategies, with the specific aim of enhancing their practical implementation in recent years.

1. Introduction

Dynamic traffic control is one of the primary research topics in intelligent transportation system. Road traffic systems can benefit from dynamic traffic control by mitigating the negative impacts of traffic congestion, e.g., reducing travel delays, alleviating pollutant emissions, or improving road traffic safety (Papageorgiou et al., 2003). Over the past decades, the majority of research in dynamic traffic control has predominantly concentrated on the development of model-based control strategies (Siri et al., 2021). Implementing model-based traffic control strategies in practice faces two significant challenges. Firstly, road traffic dynamics are inherently complex and typically described by nonlinear models. Consequently, the optimal control formulations derived from these nonlinear models are also nonlinear and non-convex. This nonlinearity poses difficulties in real-time implementation, particularly when dealing with large-scale optimization problems (Xi et al., 2013). Secondly, the performance of model-based traffic controllers is highly reliant on the accuracy of the underlying traffic flow models. Numerous studies have employed closed-form plant and prediction models to demonstrate their model-based control approaches (Carlson et al., 2010; Geroliminis et al., 2012; Hegyi et al., 2005). However, given the presence of various unpredictable factors of human behavior that may affect traffic dynamics, it becomes challenging to precisely predict the evolution of traffic processes using a deterministic traffic flow model. Studies have shown that a mismatch between the prediction model and the real traffic process can result in diminished performance for model-based controllers (Han et al., 2020).

To address the challenge of model mismatch, researchers have developed robust control approaches to optimize traffic control schemes by considering the worst-case prediction scenario, accounting for uncertain traffic conditions (Liu et al., 2021; Tettamanti et al., 2013). However, robust control methods also suffer from high computational

* Corresponding author.
E-mail address: yuhan@seu.edu.cn (Y. Han).

https://doi.org/10.1016/j.commtr.2023.100104
Received 24 August 2023; Received in revised form 24 September 2023; Accepted 25 September 2023
Available online 3 November 2023

burden, so their application scenarios remain in small-scale traffic networks. In recent years, reinforcement learning (RL), which has been successfully implemented in the areas of robotics and video games, has also gained substantial attention in traffic control area. RL offers several advantages that have the potential to overcome the limitations of model-based traffic control approaches. One notable advantage is that RL-based traffic control methods often do not rely on online traffic predictions and optimizations, resulting in reduced online computation time compared to model-based optimal control approaches. Moreover, by training an RL controller using real traffic data, the control policy can be optimized based on real traffic evaluation, potentially mitigating the model mismatch problem.

Existing studies on RL-based traffic control strategies have primarily focused on three aspects. Firstly, RL models have been developed to address various traffic control problems, such as traffic signal control, ramp metering, variable speed limits (VSLs), and vehicle motion control, e.g., Abdulhai et al. (2003), Aradi (2020), Belletti et al. (2017), and Li et al. (2017). Secondly, efforts have been made to enhance existing RL training algorithms, aiming for faster convergence and improved performance (El-Tantawy et al., 2013; Li et al., 2016). Lastly, novel RL algorithms have been devised to effectively train RL models for large-scale traffic control problems (Chen et al., 2020; Chu et al., 2019).

While a considerable number of RL-based traffic control strategies have been proposed in the literature, field implementation of such strategies is rather limited. There are several factors that limit the implementation potential of RL-based traffic control strategy. Firstly, during the training process, the exploration of random control actions may lead to additional delays or even unsafe traffic situations, which is unacceptable for traffic authorities. Secondly, given the limited training time available in real-world settings, there is no guarantee of achieving the desired level of improvement in traffic performance. Many existing studies have overlooked these problems by conducting the training of RL models solely in simulation environments, where random exploration is also acceptable, and training time can be flexible. Furthermore, when RL models are trained in simulators, there may be a mismatch between the training environment and the actual traffic process, potentially resulting in inferior control performance (Li et al., 2022).

In the literature, several studies have provided comprehensive overviews and analyses of RL-based traffic control approaches. For instance, the studies of Haydari and Yılmaz (2020) and Xiao et al. (2021) conducted comprehensive surveys on the broad application of deep RL in intelligent transportation systems and traffic engineering. These surveys covered a wide range of areas, including dynamic traffic control, routing optimization, autonomous driving, and energy management. On the other hand, Noaeen et al. (2022) and Wei et al. (2021) focused exclusively on literature reviews concerning RL applications for traffic signal control. The aforementioned studies primarily focus on the advancements and achievements in RL-based traffic control approaches, but they provide limited discussion on the challenges of implementing these strategies. Therefore, further exploration and discussion are needed to address the barriers that hinder the successful deployment of RL strategies in real-world traffic scenarios.

This paper distinguishes itself from other studies on the literature review of RL-based traffic control approaches in two key aspects. Firstly, it focuses specifically on RL-based studies in the field of dynamic traffic control, encompassing traffic signal control, traffic flow management on freeways, and microscopic traffic control utilizing intelligent vehicles as actuators. Secondly, the paper's objective extends beyond summarizing existing RL-based traffic control strategies to investigate their implementation potentials, considering them from a traffic engineering perspective. In addition to summarizing studies on model designs and algorithm architectures, the paper examines the practical implementation of RL-based traffic control strategies. Two possible methods of implementation are presented: online RL methods and offline RL methods. For online RL methods, the assumption is that RL models are directly trained in a real traffic environment, and the exploration and learning costs are explored through microscopic simulations. Offline RL methods, on the other hand, involve training RL models in a simulation environment and transferring the optimal control policies obtained to real traffic processes represented by different simulation models. Drawing from these investigations, the paper proposes several research directions that hold potential for overcoming the implementation challenges associated with RL-based traffic control strategies.

The remaining sections of this paper are organized as follows. Section 2 provides a comprehensive literature review of RL-based traffic control strategies. Section 3 explores the challenges associated with existing RL methods in terms of their implementation in real-world scenarios through microscopic traffic simulation experiments. Section 4 discusses potential approaches to address these challenges and enhance the field implementation of RL methods. Section 5 concludes the paper.

2. Literature review

In this section, we provide a comprehensive summary of existing research studies focusing on RL-based dynamic traffic control strategies. Section 2.1 presents an overview of RL fundamentals specifically applied to the domain of dynamic traffic control. Section 2.2 provides a comprehensive synthesis of agent formulations employed in various RL models. Finally, in Section 2.3, we discuss the learning algorithms that have been applied in previous studies on RL-based traffic control strategies.

2.1. Basics about RL in traffic control

RL concerns the problem of a learning agent that interacts with an environment to achieve a specific goal. The basic RL framework involves an agent, responsible for making decisions, and an environment, which encompasses everything the agent interacts with. Typically, the agent and the environment interact in discrete time steps. At each timestep k, the agent takes an action based on the state provided by the environment. In response, the environment assigns a reward to the agent and presents a new state, determined by a probability distribution. The reward function aims to quantify the advantage of taking a specific action in a given state. The agent's objective at time step k is to maximize the cumulative reward-to-go over a given time horizon, denoted as $G(k)$:

$$G(k) = \sum_{\tau=k}^{K_T} \gamma^{\tau-k} R(\tau) \tag{1}$$

where K_T denotes the time index when the state of the environment reaches the terminal state; $R(\tau)$ is the reward received at time τ; and $\gamma^{\tau-k}$ is the discount factor that defines the relative importance of the reward at time $\tau (0 \le \gamma \le 1)$.

The behavior of an agent, which involves mapping an observed state $s(k)$ of the environment to actions, is defined by a policy denoted as π. $\pi[a(k)|s(k)]$ represents the probability of taking action $a(k)$ upon observing state $s(k)$ at time k. To evaluate a policy, value functions V_π and Q_π are defined for a given state $s(k)$ and a given state-action pair $Q^*[s(k), a(k)] = \max_\pi Q^\pi[s(k), a(k)]$ as their expected cumulative reward-to-go.

$$V_\pi[s(k)] = E_\pi[G(k)|s(k)] \tag{2}$$

$$Q_\pi[s(k), a(k)] = E_\pi[G(k)|s(k), a(k)] \tag{3}$$

If the expected cumulative reward under policy π is greater than or equal to that under another policy π' for all states, π is considered to be better than or equal to π'. The policy that is better than or equal to all other policies is defined as the optimal policy, which is denoted as π^*. The optimal state value function and state-action value function are denoted as $V^*[s(k)]$ and $Q^*[s(k), a(k)]$, respectively, and $V^*[s(k)] = \max_\pi V^\pi[s(k)]$, $Q^*[s(k), a(k)] = \max_\pi Q^\pi[s(k), a(k)]$.

The presented basic RL framework can be applied to a traffic control system, where the traffic dynamics is the environment and the traffic controller is the agent. The agents are formulated based on the goal of the control system.

2.2. Agent formulations

In this section, we provide an overview of the agent formulations employed in existing RL models for freeway traffic control and urban traffic control, respectively. For large-scale urban traffic control problems, the coordination of each learning agent within RL strategies becomes essential to reduce learning complexity and enhance efficiency. The ways of coordinating the learning agents in existing RL-based strategies are summarized and discussed.

2.2.1. RL-based control approaches for freeway traffic

In the realm of RL-based freeway traffic control strategies, the agent formulations often closely resemble model-based controllers, which are derived from specific mechanisms described by physical traffic models. The model-based controllers employ the traffic state on the freeway as input, representing the state variables in the RL-based controllers. The decision variables in the model-based controllers directly correspond to the actions of the RL controllers, and the objective functions are aligned with the reward functions of the RL controllers. For instance, in a local ramp metering system, the main advantage lies in improving the throughput of the bottleneck by preventing capacity drop and spill back. As a result, the reward in RL-based ramp metering systems is often associated with the critical occupancy/density of the bottleneck, which leads to maximum throughput (Davarynejad et al., 2011; Schmidt-Dumont and Vuuren, 2015). In the case of coordinated ramp metering systems, the reward function in RL models should consider the balanced delay across each bottleneck (Belletti et al., 2017; Han et al., 2022b).

For a VSL control system, there are two commonly used mechanisms to improve traffic efficiency. When addressing a local bottleneck, it is commonly assumed that VSLs below the critical speed result in a fundamental diagram with reduced capacity. By implementing VSLs upstream of a bottle-neck, the mainstream arriving flow is permanently decreased to prevent bottleneck activation and the subsequent capacity drop. Consequently, RL models of VSLs developed based on this mechanism exhibit similar state representations to ramp metering models, with the reward function also being linked to the critical density of the bottleneck (Li et al., 2017; Wang et al., 2022b). Another application of VSLs in improving traffic efficiency is the mitigation of jam waves. As the congestion area of a jam wave propagates upstream, the RL state should encompass real-time information regarding the jam's location (Han et al., 2022a).

Besides improving traffic efficiency, ensuring traffic safety and sustainability are also crucial aspects of freeway traffic control. In the works of Li et al. (2020b) and Wu et al. (2020), reducing crash risks was considered as one of the primary objectives in their control design. To this end, the reward functions of their RL models are developed based on surrogate safety measures. The studies conducted by Li et al. (2021a) and Zhu and Ukkusuri (2014) have focused on reducing pollutant emissions as a control objective. As a result, the reward function of their RL models is designed based on emission models derived from the traffic state. Several recent studies have developed RL-based strategies for the microscopic control of connected and autonomous vehicles (CAVs) in merge situations, with the objective of enhancing the safety and efficiency of the merging zone (Hu et al., 2022; Nishitani et al., 2020; Wang and Chan, 2017). In these RL models, the state variables encompass the speeds and positions of both current and surrounding vehicles. The reward function takes into account safety performance represented by surrogate safety measures, as well as efficiency performance measures, such as the average velocity of vehicles in the merging zone.

2.2.2. RL-based control approaches for urban traffic signals

Given the scrutiny of state-of-the-art RL-based traffic signal control approaches in recent studies, e.g., Noaeen et al. (2022) and Wei et al. (2021), this section provides a concise summary of the RL agent formulations employed in urban traffic signal control. Existing RL-based strategies for urban traffic control encompass various scenarios, including single intersection control, multi-intersection coordination, and perimeter control.

The early works on RL-based traffic signal control date back to the early 2000s (Abdulhai et al., 2003; Thorpe and Anderson, 1996). In the context of single intersection signal control, the queue length of each lane and the current phase provide effective representations of the state variables (El-Tantawy et al., 2014; Li et al., 2016; Touhbi et al., 2017; Zhang et al., 2018). Some studies have explored the use of images to represent the state, extracting vehicle positions as an image input for convolutional neural networks (Genders and Razavi, 2018; Liang et al., 2019; Wei et al., 2018). The choice of action can involve setting the green duration of a phase, e.g., Casas (2017), or selecting which phase to activate, e.g., Zheng et al. (2019b), depending on the specific traffic signal settings. The rewards are commonly represented by the surrogate measures of total travel time for all the vehicles, such as the average queue length, average delay or throughput.

For RL-based multi intersection signal control, there exist two prevailing agent formulations, namely single global agent formulation and multi-agent formulation. A single global agent assimilates the state information from all intersections as input and learns to simultaneously determine the coordinated actions of all intersections (Nishi et al., 2018; Van der Pol and Oliehoek, 2016; Wiering, 2000). To address the issue of an expanding joint action space resulting from an increasing number of agents to model, some studies have proposed cooperative learning methods that combine the learning process of a centralized global agent with each local agent (Tan et al., 2019; Zhang et al., 2019). For multi-agent formulations, it is customary to enrich the ego agent's observation by incorporating details about the traffic conditions and past actions of neighboring agents (Chu et al., 2019; Wei et al., 2019; Xu et al., 2020). This integration enables the agent to optimize its policy by considering both its own state and the actions performed by neighboring agents.

Perimeter control utilizes traffic signals to regulate traffic flow upstream of a protected network in order to prevent over-saturation (Aboudolas and Geroliminis, 2013; Keyvan-Ekbatani et al., 2012). In the literature, RL-based perimeter control strategies have been investigated both in single-region networks (Ni and Cassidy, 2019; Su et al., 2023), and multi-region networks (Chen et al., 2022; Li and Hou, 2020; Zhou and Gayah, 2023). These strategies typically employ RL models with state variables that encompass regional accumulations and traffic demands. The implementation of the action can follow two approaches: discrete or continuous. Discrete perimeter control often employs bang-bang control, with perimeters in an all-open or all-closed state. In continuous control, the metering rate for each perimeter link is specified. To evaluate the effectiveness of these strategies, the reward functions commonly used are based on the trip completion rate of the protected network.

2.3. Learning algorithms

RL learning algorithms can generally be categorized into model-based methods and model-free methods. In road traffic systems, the dynamic evolution of traffic process is influenced by many stochastic human factors, such as varying driving skills and diverse reactions to disturbances. This complexity poses a challenge when attempting to estimate the state transition of an RL-based traffic control system using a probability model. Consequently, only a few works have investigated model-based RL methods in dynamic traffic control. In Wiering (2000), the

state transition probability of the signal control system was learned from observed experiences. This method has been further extended in subsequent studies to include coordination between agents (Kuyer et al., 2008), as well as multi-objective priority control for buses and emergency vehicles (Duan et al., 2010). The studies of Khamis et al. (2012) and Khamis and Gomaa (2012, 2014) proposed a Bayesian method to estimate the transition probability of non-stationary state in a multi-agent traffic signal control system. The recent study conducted by Kunjir et al. (2023) proposed a model-based offline RL method, enabling the RL agent to learn directly from offline data instead of actively interacting with the environment.

Most of existing RL-based traffic control systems use model-free training methods, which can be broadly divided into value-based methods and policy-based methods. For value-based RL methods, early attempts usually applied tabular approach such as Q-learning to update the RL policy (Arel et al., 2010; Li et al., 2017; Lu et al., 2008; Schmidt-Dumont and Vuuren, 2015; Touhbi et al., 2017). The Q-learning method estimates the optimal value function Q^* using temporal-difference learning. The Q-value, $Q_{(s,a)}$, stores the value of a state-action pair, and it is updated according to Eq. (4):

$$Q_{(s,a)} \leftarrow Q_{(s,a)} + \kappa_{(s,a)}\left[R + \gamma \max_{a'} Q_{(s',a')} - Q_{(s,a)}\right] \tag{4}$$

where R is the observed reward of the transition from the current state s to the new state s' under action a; a' denotes the action chosen at state s'; $\kappa(s,a)$ is the step-size parameter which controls how fast the Q-values are altered. While Q-learning is effective in handling local traffic control problems such as local ramp metering or individual intersection control, it is inefficient when the scale of the network gets larger because of the curse of dimensionality. Therefore, DRL approaches which use neural networks to approximate the value function have been developed (Gao et al., 2017; Genders and Razavi, 2016). The DQN family, comprising the original DQN, Double DQN, Duel DQN, and other variants, has been extensively investigated in dynamic traffic control strategies (Chen et al., 2020; Gong et al., 2019; Li et al., 2016; Mnih et al., 2015; Shabestary and Abdulhai, 2018; Wang et al., 2020; Wei et al., 2018; Xu et al., 2020). In DQN, the value function $Q_{s,a;\theta}$ is represented by a parameterized neural network. The parameters of the neural networks are updated by minimizing the following loss function as Eq. (5):

$$L = \left(R + \gamma \max_{a'} Q(s', a'; \theta') - Q(s, a; \theta)\right)^2 \tag{5}$$

where $R + \gamma \max_{a'} Q(s', a'; \theta')$ R is the estimated Q-value and $Q(s, a; \theta)$ is the target Q-value. θ' is updated by copying the values of θ. It is updated less frequently than $\boldsymbol{\theta}$ to improve the stability of the training.

Compared to value-based RL methods, policy-based RL methods can deal with continuous action space. The Policy Gradient (PG) algorithm is one of the earliest policy-based RL algorithms that have been applied to dynamic traffic control (Coşkun et al., 2018; Mousavi et al., 2017; Rizzo et al., 2019). The basic idea behind the PG algorithm is to compute the gradient of the expected cumulative reward with respect to the policy parameters, and use it to update the policy parameters in the direction of higher expected reward. The algorithm learns a parameterized policy that maps states to actions, and the policy is updated iteratively based on the observed rewards and actions.

$$\theta_{k+1} \leftarrow \theta_k + \alpha \nabla_\theta J(\pi_\theta) \tag{6}$$

where $J(\pi_\theta)$ is the expected cumulative reward under policy π parameterized by θ. α is the step-size parameter.

Actor-critic RL algorithms, which further improve the data sampling efficiency, have also been widely employed in dynamic traffic control (Aslani et al., 2017). In actor-critic algorithms, there are two main components: an actor network that learns the policy, and a critic network that learns the value function. The actor network decides which action to take in a given state, while the critic network evaluates the quality of the actor's actions.

Some actor-critic algorithms, such as Deep Deterministic Policy Gradient (DDPG) and Twin-Delayed Deep Deterministic Policy Gradient (TD3), learn deterministic policies that directly map states to actions. These algorithms have been extensively employed in RL-based traffic control systems (Casas, 2017; Li et al., 2021b; Lu et al., 2023; Tan et al., 2019; Wu et al., 2020; Zhou et al., 2019). On the other hand, some other RL traffic control systems utilized actor-critic algorithms that are capable of learning stochastic policies, which map states to action probability distribution, as the training methods. Examples of such systems include advantage actor-critic (A2C), and proximal policy optimization (PPO) (Chu et al., 2019; Lin et al., 2018; Kreidieh et al., 2018; Pandey et al., 2020; Peng et al., 2021), among others.

2.4. Evaluation method

Traffic control strategies are commonly evaluated in traffic simulations, with a focus on performance metrics related to traffic efficiency, safety, and sustainability. The evaluation of RL-based control strategies aligns with traditional model-based strategies. Traffic efficiency is typically measured by indicators such as total time spent, throughput, average travel delay, or average queue length within the network. Safety performance is often evaluated using surrogate safety measures like time-to-collision, crash potential index, and conflict index. Sustainability considerations involve assessing fuel consumption and pollutant emissions estimated from vehicle trajectories. Regarding the testing of traffic simulators, researchers have employed two main approaches. Some studies have utilized macroscopic simulations to replicate real-world traffic dynamics. However, an increasing number of studies opt for microscopic simulations, such as SUMO, VISSIM, and AIMSUN, which provide a more comprehensive representation of complex traffic behavior, including the stochastic nature of driving and routing decisions.

While many existing RL-based traffic control strategies have demonstrated superior performance compared to base-line control methods, such as model-based approaches, it is important to note that the comparisons made between them are not always fair. One notable issue in these comparisons is that the performance of RL controllers is often evaluated solely at the end of the training process, after the RL agents have been adequately trained. However, the learning cost incurred during the training process, particularly when exploring random actions, has often been overlooked. Given that many existing RL-based traffic control strategies employ random exploration, the associated cost of exploring such actions can be substantial. Furthermore, a significant number of studies fail to consider the mismatch between the training environment and the testing environment. Since there is an inherent discrepancy between a traffic simulator and real traffic processes, the performance of RL-based strategies may be limited by the accuracy of the simulators. The optimal control policy obtained from the training environment may be inferior when transferred to the real traffic environment.

In recent years, an increasing number of studies have taken into account the mismatch between the training environment and real traffic process when evaluating RL-based traffic control strategies, as summarized in Table 1. Some studies have employed macroscopic traffic flow models to establish the training environment and introduced perturbations to traffic demand and model parameters in the testing phase, thereby replicating this discrepancy (Pandey et al., 2020; Zhou and Gayah, 2021). Alternatively, certain researchers have employed one macroscopic traffic flow model for the training environment and a distinct model for the testing phase to replicate the simulation-to-reality transfer (Han et al., 2022a). Macroscopic traffic flow models provide significant advantages in terms of their tractability, interpretability, and ease of development. Nevertheless, this approach neglects the intricate

Table 1
Summary of RL-based traffic control strategies with separate training and testing environments.

Ref.	Research problem	Training environment (data)	Testing environment modification
Pandey et al. (2020)	Dynamic pricing of express lanes for network traffic optimization	**Macroscopic simulation** (CTM and a lane choice model)	Changes of parameters in the lane choice model
Zhou and Gayah (2021)	Perimeter control for two-region urban networks	**Macroscopic simulation** (MFD model)	Changes of parameters in the MFD model
Han et al. (2022a)	Resolve freeway jam waves using VSLs	**Macroscopic simulation** (CTM)	**Macroscopic simulation** METANET
Zhou and Gayah (2023)	Perimeter control for Multi-region urban networks	**Macroscopic simulation** (MFD model)	Changes of parameters in the MFD model
Aslani et al. (2018)	Signal control for single intersection	**Microscopic simulation** (AIMSUN)	Disturbances: incidents, detection errors, and pedestrians' jaywalking
Rodrigues and Azevedo (2019)	Signal control for single intersection	**Microscopic simulation** (AIMSUN)	Disturbances: demand surges, incidents, and detection errors
Tan et al. (2020)	Signal control for single intersection	**Microscopic simulation** (SUMO)	Disturbance: truck event
Wu et al. (2020)	Differential VSLs for freeway merge bottleneck	**Microscopic simulation** (SUMO)	Changes of model parameters in SUMO
Xie et al. (2022)	Dynamic route guidance for network traffic optimization	**Microscopic simulation** (VISSIM)	Changes of model parameters in VISSIM
Han et al. (2022b)	Local and coordinated ramp metering	**Macroscopic simulation** (METANET)	**Microscopic simulation** SUMO
Jang et al. (2019)	Merging control of CAVs	**Microscopic simulation** (SUMO)	**Scaled test bed**
Chalaki et al. (2020)	Merging control of CAVs	**Microscopic simulation** (SUMO)	**Scaled test bed**

behaviors of individual drivers and the associated heterogeneity due to their simplified representation of aggregated traffic dynamics, potentially leading to more favorable outcomes in training RL strategies.

Several researchers have employed microscopic traffic simulations to establish distinct training and testing environments. For instance, multiple studies formulated training scenarios involving ordinary traffic situations, while subjecting the strategies to a range of disruptions during testing, including truck events (Tan et al., 2020), jaywalking pedestrians (Aslani et al., 2018), and instances of detection failures (Rodrigues and Azevedo, 2019). In these investigations, despite the significant influence of these disturbances on the dynamic evolution of traffic processes, the underlying behavioral model of vehicles remained unchanged. Several studies have replicated the environmental mismatch phenomenon by introducing modifications to the underlying models within microscopic simulations (Han et al., 2022b; Wu et al., 2020; Xie et al., 2022). These adjustments to the underlying models are undertaken as a means to more effectively emulate the environmental mismatch, given that the simulation models can never perfectly reproduce the heterogeneous characteristics of drivers. Nevertheless, it should be noted that the application of microscopic simulation-based transferability tests in current research remains predominantly limited to small-scale traffic control scenarios, such as individual intersections or local freeway bottlenecks.

Instead of relying solely on traffic simulations, some recent studies have taken a different approach by testing control strategies for connected and autonomous vehicles in scaled test beds that replicate real-world traffic scenarios. For instance, Jang et al. (2019) proposed an RL-based strategy to optimize the trajectory of intelligent vehicles, aiming to enhance traffic operational efficiency in a roundabout scenario. The RL model was trained using the microscopic simulation tool SUMO and subsequently tested in a scaled test bed that closely resembled real-world traffic scenarios. Chalaki et al. (2020) further extended this approach by integrating an adversarial learner, aiming to improve the transferability of the control policy from the microscopic simulation to the scaled test bed. The study demonstrated superior performance compared to previous approaches. While a scaled test bed is a step closer to reality compared to simulations, it can only replicate light traffic streams, involving only a few vehicles, rather than capturing the complexities of traffic flow dynamics.

Table 1 provides a summary of the studies that have considered the mismatch between the training environment and the testing environment. Across these studies, while certain RL algorithms have exhibited greater robustness against the environmental mismatch, it is important to note that a general trend of performance degradation in simulation-to-reality transfer has been consistently observed.

3. Challenges for field implementation

There exist numerous challenges that must be addressed prior to the practical implementation of an RL-based traffic control strategy. In the literature, only a few studies have examined the feasibility of deploying their approaches in real-world scenarios. Based on the predominant sourcing of training data from either traffic simulators or real-world collections, we present two types of approaches for the practical implementation of RL-based traffic control strategies, namely online RL and offline RL. Figure 1(a) showcases the framework for online training, where the RL agent interacts directly with the real traffic environment to optimize the control policy. On the other hand, Fig. 1(b) illustrates the framework for offline RL training, where the agent is initially trained in a traffic simulation environment. Once a satisfactory level of training is achieved, the optimized control policy is then transferred to real-world traffic scenarios.

The online RL training approach entails two significant challenges. Firstly, the utilization of randomly explored control actions during the training process may lead to very poor traffic performance, e.g., high delays and unsafe traffic situations. Secondly, the training process relying on random exploration necessitates a substantial volume of training data, which may be impractical to collect due to the limited speed of data acquisition in the real world, influenced by physical time constraints and the inherent “slowness” of the traffic process.

The offline RL approach relies on the utilization of a traffic simulator due to the infeasibility of employing historical field data for training purposes. This is primarily attributed to the scarcity of effective training data collected from the field, as real-world traffic flows are regulated by a limited number of pre-defined control strategies. Additionally, many practical traffic control systems are not designed specifically to mitigate traffic congestion or enhancing traffic efficiency. For instance, many traffic signal control systems and speed control systems solely implement fixed signal timing plans and predetermined speed limit values. Consequently, the field data obtained from these control systems cannot be effectively employed for training an RL model.

However, the inherent mismatch between a traffic simulator and the real traffic process can impose constraints on the effectiveness of RL-based strategies due to the accuracy limitations of the simulators. While some studies argue that their RL-based control methods hold promising potential for field implementation, assuming continuous improvement in the accuracy of traffic simulators, it is important to note that the stochastic nature of traffic flow renders it challenging for traffic simulators to reach the level of descriptive accuracy found in other domains such as Newtonian physics. Thus, a major challenge of the offline RL approach revolves

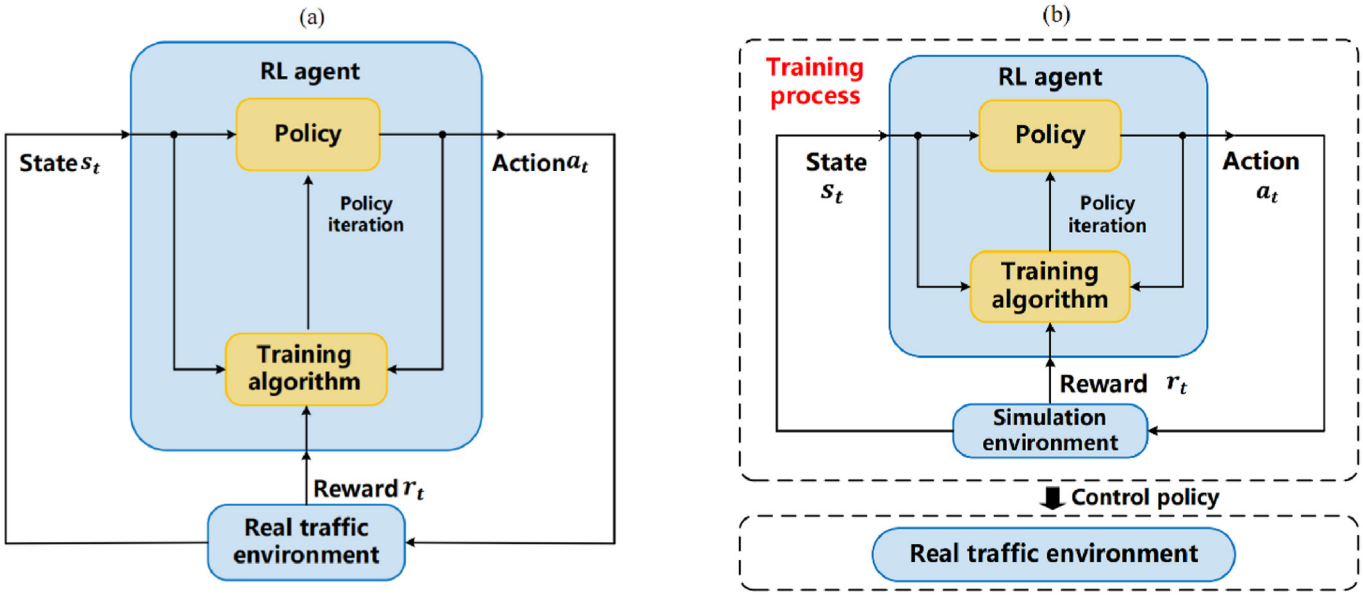


Fig. 1. Two types of approaches for implementing RL-based traffic control strategies: (a) Online training and (b) Offline training.

around achieving good transferability of the control policy from the simulation environment to the real-world traffic process.

There may be a third approach to implementing RL that integrates both offline RL (for pre-training) and online RL (for continual learning). However, challenges such as the transferability of policies across different environments in offline RL, as well as the high learning cost associated with random exploration in online RL, persist. Consequently, we will not exclusively discuss about this particular implementation method.

In the literature, while a handful of studies have touched upon the challenges mentioned in implementing their RL strategies (as observed, for instance, in Table 1), a systematic analysis that quantifies the learning cost and assesses the transferability of an RL traffic control strategy, along with comprehensive evaluations, remains notably absent. To address this gap and illuminate the limitations inherent in both online and offline RL implementations, we conduct a simulation test involving RL-based control strategies within a ramp metering scenario. In Section 3.1, we present the design of the simulation experiment. Section 3.2 evaluates the learning costs associated with the online RL method, while Section 3.3 focuses on the transferability of the offline RL method.

3.1. Simulation design

A ramp metering scenario is devised as the experimental setup, utilizing a real-life freeway stretch as the test bed, as depicted in Fig. 2. To simulate the traffic dynamics within the freeway stretch, we employ the open-source microscopic simulation software, SUMO. For our simulation experiment, we utilize the default car-following model in SUMO, known as the Krauss model. The parameter values adopted in this experiment are based on the work of Han et al. (2022b), in which the model was calibrated with real data. Specifically, we set the driver's desired (minimum) time headway to 1.1 s, while the driver imperfection (where 0 denotes perfect driving) is assigned a value of 0.4. The remaining parameters are set to their default values.

The traffic demand profile of the peak hour is depicted in Fig. 3(a). In this site, the merge area is an active bottleneck where congestion originated, as shown in the simulation result in Fig. 3(b). When congestion occurs, the downstream throughput is reduced as a result of capacity drop, and the upstream off-ramp is blocked, leading to more severe traffic congestion.

An RL-based ramp metering controller is designed to regulate the merge flow and optimize the traffic efficiency. The ramp control system operates based on a fixed cycle approach, where the cycle length is predetermined to be 20 s. The maximum queue length of the on-ramp is set to 120 vehicles. For the RL agent, the state, $s(k)$, is defined as

$$s(k) = [q_{\mathrm{m}}(k-1), \rho_B(k), v_B(k), n(k)] \tag{7}$$

where $q_{\mathrm{m}}(k-1)$ is the mainstream arriving flow of the bottleneck at time step $k-1$, measured at location 1 in Fig. 2. $\rho_B(k)$ and $v_B(k)$ are the density and speed of the bottleneck, measured at location 2. $n(k)$ is the queue length of the on-ramp. The action, $a(k)$, is defined as the green duration of cycle k. The reward, $R(k)$, is defined as

$$R(k) = -(\rho_B(k) - \rho_{cr})^2 \tag{8}$$

where ρ_{cr} is the critical density of the bottleneck. The model is trained using a state-of-the-art RL algorithm, namely TD3 (Fujimoto et al., 2018). The training process comprises 400 episodes, with each episode

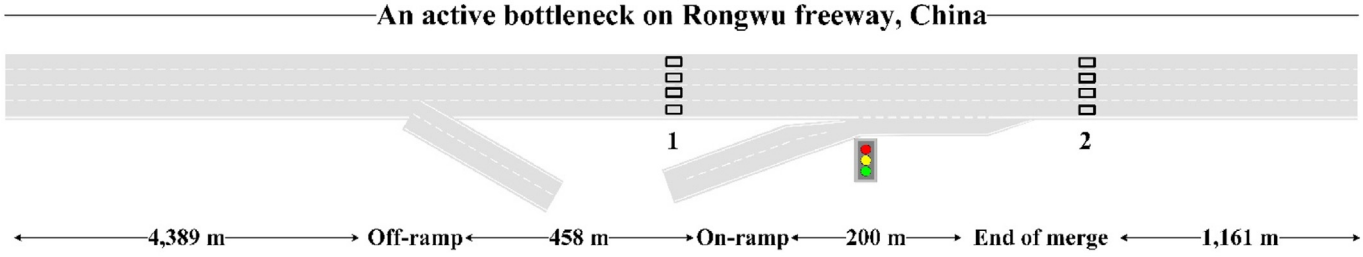


Fig. 2. A graphical representation of the freeway stretch.

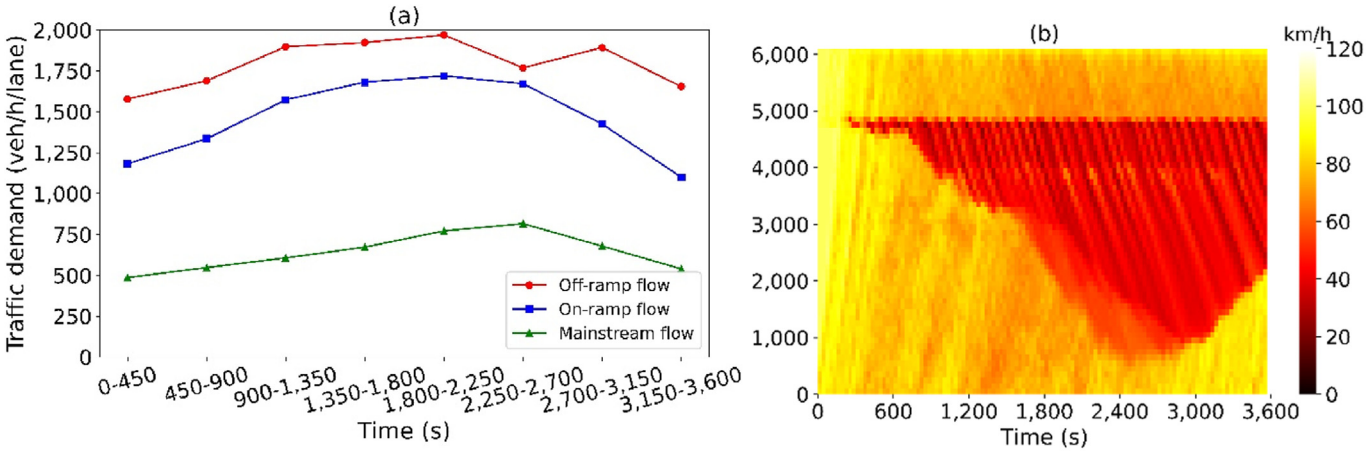


Fig. 3. (A) Traffic demand profile and (b) the speed contour plot of the simulation.

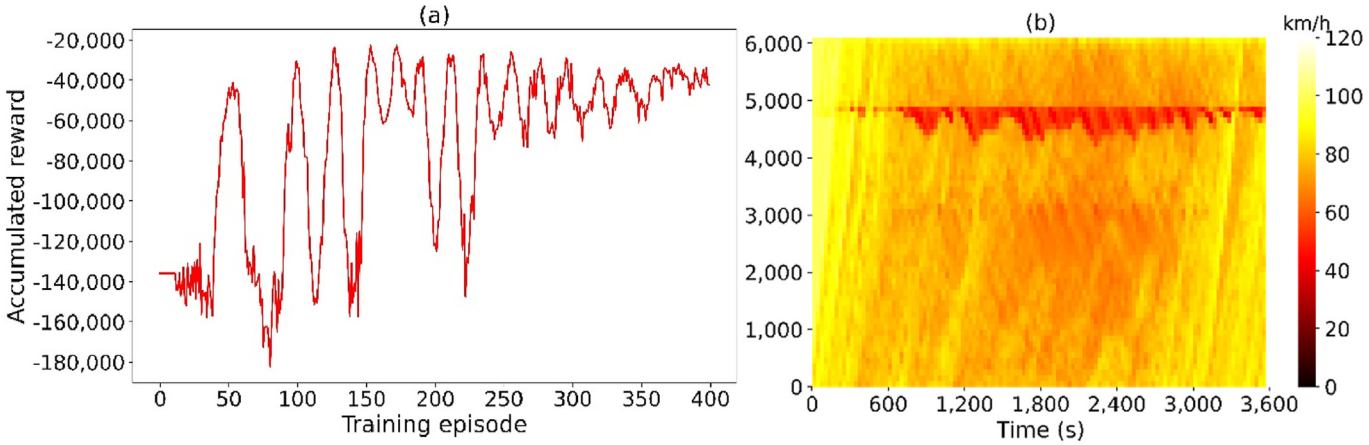


Fig. 4. (A) Performance curve of the training and (b) the speed contour plot with the optimal control policy.

representing a 1-h simulation. The performance curve of the training can be observed in Fig. 4(a). The performance of the RL controller is significantly improved after training. Figure 4(b) showcases the traffic performance achieved with the RL controller upon completion of the training. Notably, the congestion associated with the on-ramp bottleneck is substantially alleviated, resulting in a 13.7% reduction of the total travel time.

3.2. *Learning cost of online RL*

If the proposed RL-based ramp metering strategy is implemented in practice using the online RL method, additional costs will arise due to exploration and learning. To evaluate the learning cost associated with the proposed RL-based control strategy, we compare the performance of the RL controller with a baseline ramp metering controller throughout the entire training process. The feedback ramp metering strategy, ALINEA, which has been implemented in many freeways worldwide, is selected as the baseline control strategy. The comparison between ALINEA and the RL-based strategy is depicted in Fig. 5. Figure 5 illustrates the relative reduction in total travel time (TTT) achieved by the RL-based strategy in comparison to the baseline ALINEA for each training episode.

During the first 50 episodes, the RL controller exhibits notably inferior performance compared to ALINEA, resulting in a higher TTT ranging between 6% and 8%. This performance gap can be attributed to the RL agent's initial state of learning from scratch, as it lacks any prior knowledge of ramp metering. Between episodes 50 and 350, the RL controller's performance varies significantly due to the exploration and learning processes

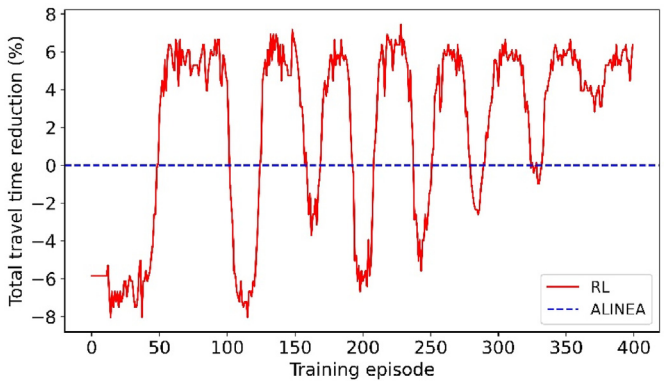


Fig. 5. Comparison between ALINEA and RL during the entire training process.

it undergoes. However, beyond the 350-episode mark, the RL controller consistently outperforms ALINEA. As a result, the RL-based control strategy incurs substantial learning costs within the initial 50episodes. Although the RL controller achieves an improved average TTT across the subsequent 300 episodes, there are still instances where its performance falls behind. Consequently, for the proposed RL strategy to confidently surpass ALINEA, approximately 350 rush hours of ramp control are necessary to train and enhance the control policy effectively.

Please be noted that in this simulation experiment, we only consider a simple local ramp metering scenario. For more complex traffic control

systems such as a coordinated ramp metering system or integrated traffic control for large-scale traffic networks, the learning cost is likely to increase significantly due to the much larger state-action space in the corresponding RL models. In addition, in the context of the local ramp metering scenario considered here, we do not account for any potential safety issues arising from randomly explored control actions. In the case of other control systems, such as a VSL control system, the learning cost pertaining to safety performance may also be considerably high. Furthermore, in practical applications, if the RL-based strategy exhibits inadequate performance, leading to a notable escalation in congestion, it is conceivable that traffic management authorities may opt to deactivate the system. This could potentially result in extended periods for both the learning process and the overall system enhancement. Consequently, finding effective approaches to reduce the learning cost of RL is crucial for facilitating the practical implementation of the online RL method.

3.3. *Transferability of offline RL*

For the offline RL method, the transferability of the control policy is one of the most critical challenges for field implementation. Transferability in the context of RL refers to the ability of an agent to apply the knowledge and skills learned in one task or environment to another related task or environment. While some studies have discussed the transferability of RL-based traffic control strategies, they have primarily focused on transferring knowledge within the same environment to different tasks. For instance, addressing the nonstationarity of traffic dynamics in RL-based traffic signal control strategies has been investigated in several studies Van der Pol and Oliehoek (2016), Yoon et al. (2021), Zang et al. (2020) and Zheng et al. (2019a), introduced meta RL approaches to enhance the agent's generalization capabilities for different phase structures of traffic signals. Ke et al. (2020) proposed a transfer learning approach to enhance the transferability of an RL-based VSL control strategy, specifically from normal merge traffic situations to rare scenarios such as adverse weather conditions. Regarding the control of CAVs, Kreidieh et al. (2018) presented an RL-based merging control strategy that considers the transferability from a synthetic merge bottleneck to a realistic counterpart.

This paper primarily discusses the transferability of RL in the context of environment mismatch. Although the aforementioned approaches aim to improve generalization across different tasks, they do not specifically address the issue of environment mismatch. In order to replicate the mismatch between the training environment and real-world traffic conditions, we employ various underlying models in the SUMO simulator for testing purposes. Specifically, we construct 15 different sets of testing environments, represented by different simulation models, parameters, and traffic demands. Please be noted that congestion and the associated capacity drop occurred in all sets of testing environments. The specific details of each environment can be found in Table 2. The degree of environment mismatch is quantified by measuring the difference in traffic state variables (the weighted average of speed and flow) using root mean square error, replicated in different environments. In essence, a larger extent of mismatch leads to more significant differences in simulated density and speed compared to those observed in the original training environment. The degrees of mismatch range from 10% to 30%. The testing environments constructed by the Krauss model with ordinary drivers and medium traffic demand is the same as the training environment, i.e., mismatch degree is 0.

This paper primarily explores the transferability of RL concerning the aspect of environment mismatch. While the aforementioned methodologies aim to enhance generalization across different tasks, they do not explicitly tackle the challenge of environment mismatch. To simulate the discrepancies between the training environment and real-world traffic conditions, we utilize diverse underlying models within the SUMO simulator for testing purposes. In particular, we configure 15 distinct sets of testing environments, each characterized by different simulation models, parameters, and traffic demands. The specific intricacies of each environment are outlined in Table 2. Notably, it is important to highlight that congestion and the associated capacity drop were observed across all testing environment sets.

The extent of environment mismatch is quantified by evaluating variations in traffic state variables (a weighted combination of speed and flow) through root mean square error calculation across different environments. Essentially, a higher level of mismatch corresponds to more pronounced disparities in simulated density and speed when compared to observations from the original training environment. The degrees of mismatch range from 10% to 30%. The testing environments generated using the Krauss model with ordinary drivers and medium traffic demand align precisely with the training environment, signifying a mismatch degree of 0.

The performance of the RL-based ramp control strategy across different testing environments is visually presented in Fig. 6. In scenarios where the training and testing environments are identical, the RL strategy achieves a 13.7% reduction in TTT. However, as the extent of environment mismatch increases, a noticeable decline in performance becomes evident. Notably, when the mismatch surpasses 20%, the TTT reduction attained by the ramp control strategy can be negative. Consequently, ensuring the precision of the training simulator holds paramount importance in realizing the efficacy of the RL-based control strategy. By way of comparison, the ALINEA strategy displays greater robustness compared to the RL approach. For mismatch levels below 20%, the average reduction in TTT by the RL strategy measures at 10.27%, whereas ALINEA achieves a TTT reduction of 6.22%. In contrast, when the degree of mismatch exceeds 20%, the RL strategy exhibits a negative average TTT reduction of 0.09%, while ALINEA still manages to achieve a positive gain, resulting in a TTS reduction of 1.43%.

Table 2
Degree of mismatch (left of |) and the performance (right of |) in different testing environments.

Simulation model	Low demand	Medium demand	High demand
Krauss (More timid drivers)	21.13 | 3.82	17.81 | 11.01	16.95 | 13.70
Krauss (Ordinary drivers)	24.86 | −1.95	0.00 | 13.71	16.03 | 12.89
Krauss (More aggressive drivers)	21.98 | −0.24	17.12 | 6.84	15.61 | 10.65
IDM	17.49 | 2.99	15.24 | 8.81	13.19 | 11.80
EIDM	28.25 | −4.45	26.28 | 1.42	22.08 | 0.89

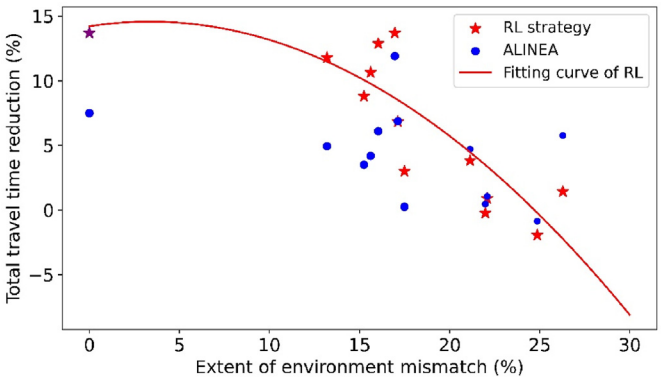


Fig. 6. Performance of the RL-based ramp control strategy and ALINEA in different testing environments.

4. Research directions to enhance RL methods for implementation

In recent years, researchers have increasingly recognized the challenges of implementing RL strategies in real-life traffic systems. They have put forward several methods with potential to address the aforementioned challenges, surpassing the mere suggestion of developing more accurate traffic simulators for training. This section provides a summary of these methods.

4.1. Integrating physical traffic flow models into the RL

Several studies have proposed new RL methods that integrate physical traffic flow models into conventional RL-based traffic control methods. Please be noted that those methods, referred to as physics-informed RL, are different from the aforementioned model-based RL methods. When referring to model-based RL, we specifically denote RL systems where state transitions are characterized by a probability model. In contrast, physics-informed RL emphasizes the utilization of physical traffic flow models to influence action selection during the training process.

Physics-informed RL methods have been demonstrated to improve the RL training performance in terms of both efficiency and safety (Lubars et al., 2021; Su et al., 2021). In Lu et al. (2014), an indirect RL-based ramp metering strategy for traffic flow management under incidents was proposed. The RL agent made action decisions by alternating between a *Q*-learning-based strategy and a model-based strategy. Simulation results indicated that the proposed method exhibited slightly better performance in reducing TTT than direct *Q*-learning. While the proposed method incorporates a traffic flow model into the training process, its objective was not specifically focused on reducing learning costs and enhancing transferability. Therefore, although the study demonstrated that the method outperformed the *Q*-learning method in reducing TTT after training, it did not provide a discussion on the associated learning cost during the training process.

Regarding improving the safety of action exploration during the training process, Bai et al. (2022) proposed a hybrid reinforcement learning (RL)-based eco-driving strategy for CAVs at signalized intersections. The strategy incorporated safety considerations by implementing model-based actions when RL-generated control actions violated pre-defined rules. A similar concept was also presented in a study by Lubars et al. (2021), where they introduced an RL-based ramp metering strategy. In their approach, the RL exploration of control actions was supervised by a model predictive controller to ensure safety. The integration of safety constraints in these methods opens up the possibility of training and testing them in real traffic environments.

The recent studies by Han et al. (2022b) proposed physics-informed RL strategies for local and coordinated ramp metering. In that method, the RL training process involves offline generation of control actions using a traffic prediction model, followed by online evaluation using online traffic data. That method offers several advantages compared to conventional RL-based ramp metering control strategies. Firstly, it eliminates the need for random exploration of control actions, resulting in substantially reduced learning costs during training. This enables its practical implementation in real traffic scenarios. Moreover, good transferability is demonstrated as knowledge is acquired by the RL agent not only from simulated traffic environments but also through practical evaluations and feedback.

The current studies on physics-informed RL methods primarily concentrate on small-scale traffic control problems. However, as the size of the network expands, the learning efficiency may not be guaranteed due to the exponential growth of the state-action space. Therefore, enhancing the scalability of physics-informed RL warrants further attention in future research.

4.2. Learning from demonstration

Learning from demonstration, also known as imitation learning, utilizes machine learning models to imitate behaviors of the experts or other models. In traffic control systems, model-based and rule-based strategies developed by researchers and engineers are considered the experts to imitate. For example, Li et al. (2020a) proposed a deep imitation learning method for traffic signal control. They treated signal control as a supervised learning problem, mapping the traffic state to control actions based on expert trajectories collected from the adaptive traffic signal control system known as SCATS. Although the method outperformed the fixed-time control scheme, it solely focused on imitating the expert system without incorporating reinforcement learning to further enhance the control performance.

Huo et al. (2020) integrated imitation learning and RL for cooperative traffic signal control. The learning model initially undergoes pre-training by imitating a rule-based signal control method, followed by interacting with the environment for continual learning. The method demonstrated faster convergence compared to other DRL methods that did not integrate imitation learning. In the studies conducted by Wang et al. (2022a) and Xiong et al. (2019), the adaptive traffic signal control system known as Self-Organizing Traffic Light (SOTL) was utilized as the expert system for imitation. Both studies demonstrated faster convergence using their respective methods. However, the discussion regarding the amount of data used for imitation learning was absent. In practical settings, the constrained speed of data collection, resulting from physical time limitations and the inherent slowness of the traffic process, presents a significant challenge in obtaining sufficient data for pre-training. This limitation may potentially affect the quality of the pre-training. The recent study by (Han et al., 2022a) proposed an RL-based controller for VSLs that incorporates imitation learning. The model initially learns from the expert VSL control system, SPECIALIST, which has been successfully implemented in practice. To demonstrate the effectiveness of the method, a reasonable amount of data, comparable in size to what can be collected from the field test experiment of SPECIALIST, was used in the pre-training stage. Simulation results indicate that the method significantly reduces learning costs compared to conventional RL methods without imitation learning.

In summary, learning from demonstration provides RL with a valuable advantage, granting it a head start by commencing from a reasonably competent performance instead of starting from scratch. This approach capitalizes on the utilization of existing knowledge and expertise, facilitating more efficient learning. However, a significant challenge arises concerning the availability of demonstration data, which may prove to be insufficient for achieving effective pre-training. Moreover, even with the integration of demonstration-based pre-training, the continual learning process still entails random explorations. These exploratory actions introduce uncertainty and potential inefficiencies, leading to higher learning costs. Hence, future research efforts should be directed towards minimizing these costs while simultaneously ensuring effective learning.

4.3. Meta-reinforcement learning

Meta-RL aims at acquiring general knowledge from a range of environments, enabling models to easily adapt to new tasks with minimal training data. This approach has also garnered interest in the field of traffic signal control. Zheng et al. (2019a) proposed an RL-based traffic signal control system that formulated action variables based on phase competition, resulting in a significant reduction of the state-action space. Building upon this work, the method proposed by Zang et al. (2020) further considered heterogeneous intersection scenarios with varying traffic intersection types and phase combinations. Zou and Qin (2020) extended those methods, introducing a Bayesian meta-RL method that learns a prior distribution as meta-knowledge from previously acquired tasks, rather than learning an initial point as meta-knowledge. The study

conducted by Huang et al. (2021) integrated model RL into the meta learning framework, further enhancing the data efficiency.

In the aforementioned methods, the aim of meta-learners is to determine an appropriate global initialization of parameters that can effectively adapt to the range of environments within the distribution. However, relying solely on a single global initialization may not always be enough for handling complex environments, particularly those characterized by varying traffic flow situations. To address this challenge, Kim et al. (2023) and Zhang et al. (2020) proposed meta-RL methods that learn multiple tasks for each intersection, encompassing different traffic regimes, including both under-saturated and over-saturated traffic situations. The meta-RL method proposed by Zhu et al. (2023) incorporates the observations and actions of neighboring agents, thereby further enhancing the stability of policy learning.

Previous studies have demonstrated the enhanced transferability of meta-RL signal control methods, attributed to their ability to learn general knowledge rather than being constrained by specific knowledge. Furthermore, training a meta-RL agent using data from various intersections enables the acquisition of generalized policies that are effective across diverse traffic scenarios, reducing the need for real-time data acquisition at each specific intersection. However, it is essential to consider certain factors when implementing the learned policies in real-life traffic signal configurations. Real-world traffic signals often operate based on fixed cyclic phases, which necessitate the adaptation of learned policies for successful deployment of a meta-RL agent within these constraints. Furthermore, it is worth noting that meta-RL methods for freeway traffic control remain an area yet to be thoroughly investigated.

4.4. Incorporating RL with other traffic control methods

Several studies have developed RL-based signal control methods that integrate with conventional traffic control methods (Zhang et al., 2022). In the PressLight method proposed by Wei et al. (2019), the state and reward of the RL model were designed based on the theoretically grounded max pressure algorithm. The PressLight method demonstrated superior performance compared to several baseline RL methods, as it directly optimized the total travel time through its reward design. Nevertheless, the authors also acknowledged the proposed method still relies on trial-and- error learning, which may result in potential risks and costs associated with deploying an online updated RL model in the real world.

The study conducted by Wang et al. (2022c) introduced an RL-based method aimed at optimizing a policy network with a pre-determined max pressure structure. The method focused on optimizing the position-weighted curves, which play a crucial role in calculating the pressure of movements, by treating them as the parameters to be optimized. The study demonstrated that the proposed method outperformed conventional max pressure methods. However, it is important to note that the method was not compared with other baseline RL methods to provide a comprehensive assessment of its advantages in reducing the learning cost.

Methods that combine RL with conventional traffic control approaches can avoid random exploration of control actions by following the logical structure embedded in these methods. Therefore, those methods offer the opportunity to leverage the knowledge and logical structure of conventional traffic control strategies, potentially reducing the learning cost. More experimental studies to demonstrate those advantages are needed in future research. Furthermore, it would be worthwhile to explore the integration of RL with conventional freeway traffic control approaches, such as feedback control, in future research.

4.5. Adversarial reinforcement learning

Adversarial reinforcement learning (ARL) has been proposed to enhance the robustness and transferability of RL-based approaches against the mismatch between training and testing environments (Pattanaik et al., 2017; Pinto et al., 2017). In an RL-based traffic signal control study conducted by Tan et al. (2020), the authors demonstrated that enhancing the robustness of the RL strategy's performance can be achieved by incorporating synthetic perturbations into the state space during training. Some studies have framed the policy learning in ARL as a zero-sum game, wherein the RL agent seeks the optimal policy while an adversary aims to find an optimal destabilization policy. For instance, Chalaki et al. (2020) conducted a study that introduced an adversarial learning method to optimize the driving behavior of CAVs at merging sections. This approach trained two agents against each other in a zero-sum game. The learning agent was trained to optimize merging behavior, while the adversarial agent was incentivized by the first agent's failure and aimed to minimize its reward by perturbing elements of the action and state space. Consequently, combining adversarial learning with RL offers a means to mitigate model mismatch between the training environment and the real environment. The effectiveness of the adversarial learning method was demonstrated through zero-shot policy transfer. In a recent study conducted by Han et al. (2023), a similar approach was employed for signal control within an urban network. The simulation results clearly illustrate that the incorporation of adversarial learning led to a substantial enhancement in the transferability of the RL-based control policy.

5. Conclusions

This paper presents a survey on the applications of RL in dynamic traffic control. Firstly, a comprehensive literature review on RL-based traffic control strategies is provided, including the agent design, training algorithms, and evaluation methods. Next, the challenges associated with implementing existing RL methods in real-world scenarios are discussed. Finally, potential approaches to address these challenges and enhance the practical implementation of RL methods are summarized. Our aim is to offer a thorough literature review and extensive discussion that inspires and encourages further advancements in RL-based applications for dynamic traffic control.

This review highlights the expanding body of literature focused on the examination of practical challenges inherent in RL-based traffic control strategies, with particular emphasis on the intricacies of simulation-to-reality transfer. Furthermore, an increasing number of studies are devoted to mitigating the limitations of RL strategies, demonstrating a strong focus on improving their practical feasibility in real-world contexts. Incorporating traffic flow models or other models derived from traffic domain knowledge into RL has been demonstrated to effectively reduce learning costs and enhance transferability, offering a promising solution to address the aforementioned challenges. Future research should continue to explore and propose versatile methods within this avenue. Nevertheless, it is important to note that the current methodology used to evaluate the transferability of RL strategies might be considered relatively simple. Consequently, there is a clear necessity to develop benchmarking methodologies that are more comprehensive, enabling effective assessment of the performance of RL-based traffic control strategies. In addition, the current research on RL strategy transferability primarily focuses on smaller-scale traffic control scenarios. To enhance the applicability of RL strategies in larger-scale traffic control, future studies should prioritize exploring methods capable of addressing the complexities of more intricate traffic situations.

Replication and data sharing

The code for replication can be found at https://github.com/YuHan-Research-Group-SEU/Deep-Reinforcement-Learning-Network-for-Highway-Ramp-Metering.

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.

Acknowledgement

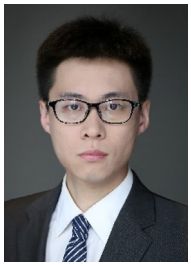
This research is supported by the National Natural Science Foundation of China (No. 52002065), the Natural Science Foundation of Jiangsu (No. BK20200378), and ZhiShan Scholar Program of Southeast University.

References

Abdulhai, B., Pringle, R., Karakoulas, G.J., 2003. Reinforcement learning for *True* adaptive traffic signal control. J. Transport. Eng. 129, 278–285.
Aboudolas, K., Geroliminis, N., 2013. Perimeter and boundary flow control in multi-reservoir heterogeneous networks. Transp. Res. Part B Methodol. 55, 265–281.
Aradi, S., 2020. Survey of deep reinforcement learning for motion planning of autonomous vehicles. IEEE Trans. Intell. Transport. Syst. 23, 740–759.
Arel, I., Liu, C., Urbanik, T., Kohls, A.G., 2010. Reinforcement learning-based multi-agent system for network traffic signal control. IET Intell. Transp. Syst. 4, 128–135.
Aslani, M., Mesgari, M.S., Wiering, M., 2017. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. Transport. Res. C Emerg. Technol. 85, 732–752.
Aslani, M., Seipel, S., Mesgari, M.S., Wiering, M., 2018. Traffic signal optimization through discrete and continuous reinforcement learning with robustness analysis in downtown Tehran. Adv. Eng. Inf. 38, 639–655.
Bai, Z., Hao, P., Shangguan, W., Cai, B., Barth, M.J., 2022. Hybrid reinforcement learning-based eco-driving strategy for connected and automated vehicles at signalized intersections. IEEE Trans. Intell. Transport. Syst. 23, 15850–15863.
Belletti, F., Haziza, D., Gomes, G., Bayen, A.M., 2017. Expert level control of ramp metering based on multi-task deep reinforcement learning. IEEE Trans. Intell. Transport. Syst. 19, 1198–1207.
Carlson, R.C., Papamichail, I., Papageorgiou, M., Messmer, A., 2010. Optimal motorway traffic flow control involving variable speed limits and ramp metering. Transport. Sci. 44, 238–253.
Casas, N., 2017. Deep deterministic policy gradient for urban traffic light control. arXiv: 1703.09035. https://arxiv.org/abs/1703.09035.pdf.
Chalaki, B., Beaver, L.E., Remer, B., Jang, K., Vinitsky, E., Bayen, A.M., et al., 2020. Zero-shot autonomous vehicle policy transfer: from simulation to real-world via adversarial learning. In: 2020 IEEE 16th International Conference on Control & Automation (ICCA). October 9-11, 2020, Singapore. IEEE, pp. 35–40.
Chen, C., Wei, H., Xu, N., Zheng, G., Yang, M., Xiong, Y., et al., 2020. Toward A thousand lights: decentralized deep reinforcement learning for large-scale traffic signal control. Proc. AAAI Conf. Artif. Intell. 34, 3414–3421.
Chen, C., Huang, Y.P., Lam, W.H.K., Pan, T.L., Hsu, S.C., Sumalee, A., et al., 2022. Data efficient reinforcement learning and adaptive optimal perimeter control of network traffic dynamics. Transport. Res. C Emerg. Technol. 142, 103759.
Chu, T., Wang, J., Codecà, L., Li, Z., 2019. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans. Intell. Transport. Syst. 21, 1086–1095.
Coşkun, M., Baggag, A., Chawla, S., 2018. Deep reinforcement learning for traffic light optimization. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). November 17-20, 2018, Singapore. IEEE, pp. 564–571.
Davarynejad, M., Hegyi, A., Vrancken, J., van den Berg, J., 2011. Motorway ramp-metering control with queuing consideration using Q-learning. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). October 5-7, 2011. IEEE, Washington, DC, USA, pp. 1652–1658.
Duan, H., Li, Z., Zhang, Y., 2010. Multiobjective reinforcement learning for traffic signal control using vehicular ad hoc network. EURASIP J. Appl. Signal Process. 2010, 7, 1–7.
El-Tantawy, S., Abdulhai, B., Abdelgawad, H., 2013. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto. IEEE Trans. Intell. Transport. Syst. 14, 1140–1150.
El-Tantawy, S., Abdulhai, B., Abdelgawad, H., 2014. Design of reinforcement learning parameters for seamless application of adaptive traffic signal control. J. Intell. Transp. Syst. 18, 227–245.
Fujimoto, S., Hoof, H., Meger, D., 2018, July. Addressing function approximation error in actor-critic methods. In: International conference on machine learning. PMLR, pp. 1587–1596.
Gao, J., Shen, Y., Liu, J., Ito, M., Shiratori, N., 2017. Adaptive traffic signal control: deep reinforcement learning algorithm with experience replay and target network. arXiv: 1705.02755. https://arxiv.org/abs/1705.02755.pdf.
Genders, W., Razavi, S., 2016. Using a deep reinforcement learning agent for traffic signal control. arXiv: 1611.01142. https://arxiv.org/abs/1611.01142.pdf.
Genders, W., Razavi, S., 2018. Evaluating reinforcement learning state representations for adaptive traffic signal control. Procedia Comput. Sci. 130, 26–33.
Geroliminis, N., Haddad, J., Ramezani, M., 2012. Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. IEEE Trans. Intell. Transport. Syst. 14, 348–359.
Gong, Y., Abdel-Aty, M., Cai, Q., Rahman, M.S., 2019. Decentralized network level adaptive signal control by multi-agent deep reinforcement learning. Transp. Res. Interdiscip. Perspect. 1, 100020.
Han, Y., Ramezani, M., Hegyi, A., Yuan, Y., Hoogendoorn, S., 2020. Hierarchical ramp metering in freeways: an aggregated modeling and control approach. Transport. Res. C Emerg. Technol. 110, 1–19.
Han, Y., Hegyi, A., Zhang, L., He, Z., Chung, E., Liu, P., 2022a. A new reinforcement learning-based variable speed limit control approach to improve traffic efficiency against freeway jam waves. Transport. Res. C Emerg. Technol. 144, 103900.
Han, Y., Wang, M., Li, L., Roncoli, C., Gao, J., Liu, P., 2022b. A physics-informed reinforcement learning-based strategy for local and coordinated ramp metering. Transport. Res. C Emerg. Technol. 137, 103584.
Han, G., Han, Y., Wang, H., Ruan, T., Li, C., 2023. Coordinated control of urban expressway integrating adjacent signalized intersections using adversarial network based reinforcement learning method. In: IEEE Trans Intell Transp Syst, pp. 1–15.
Haydari, A., Yılmaz, Y., 2020. Deep reinforcement learning for intelligent transportation systems: a survey. IEEE Trans. Intell. Transport. Syst. 23, 11–32.
Hegyi, A., De Schutter, B., Hellendoorn, H., 2005. Model predictive control for optimal coordination of ramp metering and variable speed limits. Transport. Res. C Emerg. Technol. 13, 185–209.
Hu, J., Li, X., Cen, Y., Xu, Q., Zhu, X., Hu, W., 2022. A roadside decision-making methodology based on deep reinforcement learning to simultaneously improve the safety and efficiency of merging zone. IEEE Trans. Intell. Transport. Syst. 23, 18620–18631.
Huang, X., Wu, D., Jenkin, M., Boulet, B., 2021. ModelLight: model-based meta-reinforcement learning for traffic signal control. arXiv: 2111.08067. https://arxiv.org/abs/2111.08067.pdf.
Huo, Y., Tao, Q., Hu, J., 2020. Cooperative control for multi-intersection traffic signal based on deep reinforcement learning and imitation learning. IEEE Access 8, 199573–199585.
Jang, K., Vinitsky, E., Chalaki, B., Remer, B., Beaver, L., Malikopoulos, A.A., et al., 2019. Simulation to scaled city: zero-shot policy transfer for traffic control via autonomous vehicles. In: Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems. April 16-18, 2019, Montreal, Quebec. ACM, Canada. New York, pp. 291–300.
Ke, Z., Li, Z., Cao, Z., Liu, P., 2020. Enhancing transferability of deep reinforcement learning-based variable speed limit control using transfer learning. IEEE Trans. Intell. Transport. Syst. 22, 4684–4695.
Keyvan-Ekbatani, M., Kouvelas, A., Papamichail, I., Papageorgiou, M., 2012. Exploiting the fundamental diagram of urban networks for feedback-based gating. Transp. Res. Part B Methodol. 46, 1393–1403.
Khamis, M.A., Gomaa, W., 2013. Enhanced multiagent multi-objective reinforcement learning for urban traffic light control. In: 2012 11th International Conference on Machine Learning and Applications. December 12-15, 2012. IEEE, Boca Raton, FL, USA, pp. 586–591.
Khamis, M.A., Gomaa, W., 2014. Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework. Eng. Appl. Artif. Intell. 29, 134–151.
Khamis, M.A., Gomaa, W., El-Shishiny, H., 2012. Multi-objective traffic light control system based on Bayesian probability interpretation. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems. September 16-19, 2012. IEEE, Anchorage, AK, USA, pp. 995–1000.
Kim, G., Kang, J., Sohn, K., 2023. A meta–reinforcement learning algorithm for traffic signal control to automatically switch different reward functions according to the saturation level of traffic flows. Comput. Aided Civil Eng. 38, 779–798.
Kreidieh, A.R., Wu, C., Bayen, A.M., 2018. Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). November 4-7, 2018. IEEE, Maui, HI, USA, pp. 1475–1480.
Kunjir, M., Chawla, S., Chandrasekar, S., Jay, D., Ravindran, B., 2023. Optimizing traffic control with model-based learning: a pessimistic approach to data-efficient policy inference. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. August 6-10, 2023. ACM, Long Beach, CA, USA. New York, pp. 1176–1187.
Kuyer, L., Whiteson, S., Bakker, B., Vlassis, N., 2008. Multiagent reinforcement learning for urban traffic control using coordination graphs. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, pp. 656–671.
Li, L., Lv, Y., Wang, F.Y., 2016. Traffic signal timing via deep reinforcement learning. IEEE/CAA J. Autom Sin. 3, 247–254.
Li, M., Cao, Z., Li, Z., 2021a. A reinforcement learning-based vehicle platoon control strategy for reducing energy consumption in traffic oscillations. IEEE Transact. Neural Networks Learn. Syst. 32, 5309–5322.
Li, X., Guo, Z., Dai, X., Lin, Y., Jin, J., Zhu, F., et al., 2020a. Deep imitation learning for traffic signal control and operations based on graph convolutional neural networks. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). September 20-23, 2020, Rhodes, Greece. IEEE, pp. 1–6.
Li, D., Hou, Z., 2020. Perimeter control of urban traffic networks based on model-free adaptive control. IEEE Trans. Intell. Transport. Syst. 22, 6460–6472.
Li, Z., Liu, P., Xu, C., Duan, H., Wang, W., 2017. Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks. IEEE Trans. Intell. Transport. Syst. 18, 3204–3217.
Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z., Zhou, B., 2022. MetaDrive: composing diverse driving scenarios for generalizable reinforcement learning. IEEE Trans. Pattern Anal. Mach. Intell. 45, 3461–3475.
Li, Z., Xu, C., Guo, Y., Liu, P., Pu, Z., 2020b. Reinforcement learning-based variable speed limits control to reduce crash risks near traffic oscillations on freeways. IEEE Intell. Transp. Syst. Mag. 13, 64–70.

Li, Z., Yu, H., Zhang, G., Dong, S., Xu, C.Z., 2021b. Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. Transport. Res. C Emerg. Technol. 125, 103059.
Liang, X., Du, X., Wang, G., Han, Z., 2019. A deep reinforcement learning network for traffic light cycle control. IEEE Trans. Veh. Technol. 68, 1243–1253.
Lin, Y., Dai, X., Li, L., Wang, F.Y., 2018. An efficient deep reinforcement learning model for urban traffic control. arXiv: 1808.01876. https://arxiv.org/abs/1808.01876.pdf.
Lu, S., Liu, X., Dai, S., 2008. Q-learning for adaptive traffic signal control based on delay minimization strategy. In: 2008 IEEE International Conference on Networking, Sensing and Control. April 6-8, 2008, Sanya, China. IEEE, pp. 687–691.
Liu, H., Claudel, C.G., Machemehl, R., Perrine, K.A., 2021. A robust traffic control model considering uncertainties in turning ratios. IEEE Trans. Intell. Transport. Syst. 23, 6539–6555.
Lu, C., Chen, H., Grant-Muller, S., 2014. Indirect reinforcement learning for incident-responsive ramp control. Procedia Soc Behav Sci 111, 1112–1122.
Lu, W., Yi, Z., Gu, Y., Rui, Y., Ran, B., 2023. TD3LVSL: a lane-level variable speed limit approach based on twin delayed deep deterministic policy gradient in a connected automated vehicle environment. Transport. Res. C Emerg. Technol. 153, 104221.
Lubars, J., Gupta, H., Chinchali, S., Li, L., Raja, A., Srikant, R., et al., 2021. Combining reinforcement learning with model predictive control for on-ramp merging. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). September 19-22, 2021, Indianapolis. IEEE, USA, pp. 942–947.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., et al., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.
Mousavi, S.S., Schukat, M., Howley, E., 2017. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. IET Intell. Transp. Syst. 11, 417–423.
Ni, W., Cassidy, M.J., 2019. Cordon control with spatially-varying metering rates: a Reinforcement Learning approach. Transport. Res. C Emerg. Technol. 98, 358–369.
Nishi, T., Otaki, K., Hayakawa, K., Yoshimura, T., 2018. Traffic signal control based on reinforcement learning with graph convolutional neural nets. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). November 4-7, 2018. IEEE, Maui, HI, USA, pp. 877–883.
Nishitani, I., Yang, H., Guo, R., Keshavamurthy, S., Oguchi, K., 2020. Deep merging: vehicle merging controller based on deep reinforcement learning with embedding network. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). May 31 - August 31, 2020, Paris, France. IEEE, pp. 216–221.
Noaeen, M., Naik, A., Goodman, L., Crebo, J., Abrar, T., Abad, Z.S.H., et al., 2022. Reinforcement learning in urban network traffic signal control: a systematic literature review. Expert Syst. Appl. 199, 116830.
Pandey, V., Wang, E., Boyles, S.D., 2020. Deep reinforcement learning algorithm for dynamic pricing of express lanes with multiple access locations. Transport. Res. C Emerg. Technol. 119, 102715.
Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., Wang, Y., 2003. Review of road traffic control strategies. Proc. IEEE 91, 2043–2067.
Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., Chowdhary, G., 2017. Robust deep reinforcement learning with adversarial attacks. arXiv: 1712.03632. https://arxiv.org/abs/1712.03632.pdf.
Peng, B., Keskin, M.F., Kulcsár, B., Wymeersch, H., 2021. Connected autonomous vehicles for improving mixed traffic efficiency in unsignalized intersections with deep reinforcement learning. Commun. Transp. Res. 1, 100017.
Pinto, L., Davidson, J., Sukthankar, R., Gupta, A., 2017. Robust adversarial reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70. ACM, Sydney, NSW, Australia. New York, pp. 2817–2826. August 6-11, 2017.
Rizzo, S.G., Vantini, G., Chawla, S., 2019. Time critic policy gradient methods for traffic signal control in complex and congested scenarios. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. August 4-8, 2019. ACM, Anchorage, AK, USA. New York, pp. 1654–1664.
Rodrigues, F., Azevedo, C.L., 2019. Towards robust deep reinforcement learning for traffic signal control: demand surges, incidents and sensor failures. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). October 27-30, 2019, Auckland, New Zealand. IEEE, pp. 3559–3566.
Schmidt-Dumont, T., Vuuren, J.V., 2015. Decentralised reinforcement learning for ramp metering and variable speed limits on highways. IEEE Trans. Intell. Transport. Syst. 14, 1.
Shabestary, S.M.A., Abdulhai, B., 2018. Deep learning *vs.* discrete reinforcement learning for adaptive traffic signal control. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). November 4-7, 2018. IEEE, Maui, HI, USA, pp. 286–293.
Siri, S., Pasquale, C., Sacone, S., Ferrara, A., 2021. Freeway traffic control: a survey. Automatica 130, 109655.
Su, Z.C., Chow, A.H.F., Zhong, R.X., 2021. Adaptive network traffic control with an integrated model-based and data-driven approach and a decentralised solution method. Transport. Res. C Emerg. Technol. 128, 103154.
Su, Z.C., Chow, A.H.F., Fang, C.L., Liang, E.M., Zhong, R.X., 2023. Hierarchical control for stochastic network traffic with reinforcement learning. Transp. Res. Part B Methodol. 167, 196–216.
Tan, T., Bao, F., Deng, Y., Jin, A., Dai, Q., Wang, J., 2019. Cooperative deep reinforcement learning for large-scale traffic grid signal control. IEEE Trans. Cybern. 50, 2687–2700.
Tan, K.L., Sharma, A., Sarkar, S., 2020. Robust deep reinforcement learning for traffic signal control. J. Big Data Anal. Transp. 2, 263–274.
Tettamanti, T., Luspay, T., Kulcsár, B., Péni, T., Varga, I., 2013. Robust control for urban road traffic networks. IEEE Trans. Intell. Transport. Syst. 15, 385–398.
Thorpe, T.L., Anderson, C., 1996. Traac Light Control Using SARSA with Three State Representations. Technical report, Citeseer.
Touhbi, S., Babram, M.A., Nguyen-Huu, T., Marilleau, N., Hbid, M.L., Cambier, C., et al., 2017. Adaptive traffic signal control: exploring reward definition for reinforcement learning. Procedia Comput. Sci. 109, 513–520.
Van der Pol, E., Oliehoek, F., 2016. Coordinated deep reinforcement learners for traffic light control. In: Proceedings of Learning, Inference and Control of Multi-Agent Systems (At NIPS 2016), vol. 8, pp. 21–38.
Wang, P., Chan, C.Y., 2017. Formulation of deep reinforcement learning architecture toward autonomous driving for on-ramp merge. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). October 16-19, 2017, Yokohama, Japan. IEEE, pp. 1–6.
Wang, M., Wu, L., Li, J., Wu, D., Ma, C., 2022a. Urban traffic signal control with reinforcement learning from demonstration data. In: 2022 International Joint Conference on Neural Networks (IJCNN). July 18-23, 2022, Padua, Italy. IEEE, pp. 1–8.
Wang, Y., Xu, T., Niu, X., Tan, C., Chen, E., Xiong, H., 2020. STMARL: a spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control. IEEE Trans. Mobile Comput. 21, 2228–2242.
Wang, C., Xu, Y., Zhang, J., Ran, B., 2022b. Integrated traffic control for freeway recurrent bottleneck based on deep reinforcement learning. IEEE Trans. Intell. Transport. Syst. 23, 15522–15535.
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. August 19-23, 2018. ACM, London, United Kingdom. New York, pp. 2496–2505.
Wang, X., Yin, Y., Feng, Y., Liu, H.X., 2022c. Learning the max pressure control for urban traffic networks considering the phase switching loss. Transport. Res. C Emerg. Technol. 140, 103670.
Wei, H., Chen, C., Zheng, G., Wu, K., Gayah, V., Xu, K., et al., 2019. PressLight: learning max pressure control to coordinate traffic signals in arterial network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. August 4-8, 2019. ACM, Anchorage, AK, USA. New York, pp. 1290–1298.
Wei, H., Zheng, G., Gayah, V., Li, Z., 2021. Recent advances in reinforcement learning for traffic signal control: a survey of models and evaluation. SIGKDD Explor Newsl. 22, 12–18.
Wiering, M., 2000. Multi-agent reinforcement leraning for traffic light control. In: Proceedings of the Seventeenth International Conference on Machine Learning. ACM, New York, pp. 1151–1158.
Wu, Y., Tan, H., Qin, L., Ran, B., 2020. Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm. Transport. Res. C Emerg. Technol. 117, 102649.
Xi, Y.G., Li, D.W., Lin, S., 2013. Model predictive control—status and challenges. Acta Autom. Sin. 39, 222–236.
Xiao, Y., Liu, J., Wu, J., Ansari, N., 2021. Leveraging deep reinforcement learning for traffic engineering: a survey. IEEE Commun. Surv. Tutor 23, 2064–2097.
Xie, J., Yang, Z., Lai, X., Liu, Y., Yang, X.B., Teng, T.H., et al., 2022. Deep reinforcement learning for dynamic incident-responsive traffic information dissemination. Transport. Res. Part E Logist Transp Rev 166, 102871.
Xiong, Y., Zheng, G., Xu, K., Li, Z., 2019. Learning traffic signal control from demonstrations. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. November 3-7, 2019. ACM, Beijing, China. New York, pp. 2289–2292.
Xu, M., Wu, J., Huang, L., Zhou, R., Wang, T., Hu, D., 2020. Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning. J. Intell. Transp. Syst. 24, 1–10.
Yoon, J., Ahn, K., Park, J., Yeo, H., 2021. Transferable traffic signal control: reinforcement learning with graph centric state representation. Transport. Res. C Emerg. Technol. 130, 103321.
Zang, X., Yao, H., Zheng, G., Xu, N., Xu, K., Li, Z., 2020. MetaLight: value-based meta-reinforcement learning for traffic signal control. Proc. AAAI Conf. Artif. Intell. 34, 1153–1160.
Zhang, R., Ishikawa, A., Wang, W., Striner, B., Tonguz, O., 2018. Using reinforcement learning with partial vehicle detection for intelligent traffic signal control. arXiv: 1807.01628. https://arxiv.org/abs/1807.01628.pdf.
Zhang, L., Wu, Q., Shen, J., Lü, L., Du, B., Wu, J., 2022. Expression might be enough: Representing pressure and demand for reinforcement learning based traffic signal control. In: International Conference on Machine Learning. PMLR, pp. 26645–26654.
Zhang, Z., Yang, J., Zha, H., 2019. Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization. arXiv: 1909.10651. https://arxiv.org/abs/1909.10651.pdf.
Zhang, H., Liu, C., Zhang, W., Zheng, G., Yu, Y., 2020. GeneraLight: improving environment generalization of traffic signal control via meta reinforcement learning. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. October 19-23, 2020, Virtual Event, Ireland. ACM, New York, pp. 1783–1792.
Zheng, G., Xiong, Y., Zang, X., Feng, J., Wei, H., Zhang, H., et al., 2019a. Learning phase competition for traffic signal control. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. November 3-7, 2019. ACM, Beijing, China. New York, pp. 1963–1972.
Zheng, G., Zang, X., Xu, N., Wei, H., Yu, Z., Gayah, V., et al., 2019b. Diagnosing reinforcement learning for traffic signal control. arXiv: 1905.04716. https://arxiv.org/abs/1905.04716.pdf.
Zhou, D., Gayah, V.V., 2021. Model-free perimeter metering control for two-region urban networks using deep reinforcement learning. Transport. Res. C Emerg. Technol. 124, 102949.

Zhou, D., Gayah, V.V., 2023. Scalable multi-region perimeter metering control for urban networks: a multi-agent deep reinforcement learning approach. Transport. Res. C Emerg. Technol. 148, 104033.
Zhu, F., Ukkusuri, S.V., 2014. Accounting for dynamic speed limit control in a stochastic traffic environment: a reinforcement learning approach. Transport. Res. C Emerg. Technol. 41, 30–47.
Zhou, M., Yu, Y., Qu, X., 2019. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: a reinforcement learning approach. IEEE Trans. Intell. Transport. Syst. 21, 433–443.
Zhu, L., Peng, P., Lu, Z., Tian, Y., 2023. MetaVIM: meta variationally intrinsic motivated reinforcement learning for decentralized traffic signal control. IEEE Trans. Knowl. Data Eng. 35, 11570–11584.
Zou, Y., Qin, Z., 2020. Bayesian meta-reinforcement learning for traffic signal control. arXiv: 2010.00163. https://arxiv.org/abs/2010.00163.pdf.



Yu Han received his Ph.D. degree from Delft University of Technology in 2017. He is currently an Associate Professor with the School of Transportation, Southeast University. His research interests involve employing interdisciplinary approaches to manage and control traffic flows, including techniques like model predictive control, reinforcement learning, and parallel learning that integrate traffic domain knowledge with data-driven learning models. He has authored numerous articles and has also served as a reviewer for prestigious journals within the fields of traffic and transportation, including *Transportation Research Part C* and *IEEE Transactions on Intelligent Transportation Systems*.

Meng Wang received his M.Sc. degree from Research Institute of Highway and his Ph.D. degree (with distinction) from TU Delft, in 2006 and 2014, respectively. He worked as a PostDoc Researcher (2014–2015) at the Faculty of Mechanical Engineering, TU Delft, and as an Assistant Professor (2015–2021) at the Department of Transport and Planning. Since 2021, he has been a Full Professor and Head of the Chair of Traffic Process Automation, "Friedrich List" Faculty of Transport and Traffic Sciences, TU Dresden. His main research interests are control design and impact assessment of Cooperative Intelligent Transportation Systems. He is an Associate Editor of *IEEE Transactions on Intelligent Transportation Systems* and *Transport-metrica B*.

Ludovic Leclercq is a Research Director at Université Gustave Eiffel, France. He is also a Full Professor (15% part-time) at TU Delft, the Netherlands, holding a chair in Transportation Systems Modeling in the Era of New Mobility. He has received his M.S. degree in civil engineering in 1998, his Ph.D. degree in 2002 and his habilitation thesis (HDR) in 2009. He is currently head of the LICIT-ECO7 laboratory, a joint research unit from Univ Eiffel and ENTPE. His research interests correspond to multiscale and multimodal dynamic traffic modeling and the related environmental externalities. Smart cities, mobility as a service, sustainable and reliable transportation systems are some of the applications that his researches are targeting. He is the current chair (2021–2024) of the committee "Traffic Flow Theory and Characteristics" ACP50 of the TRB. He is also a member of the international advisory committee of ISTTT. He is co-editor-in-chief of *Data Science in Transportation* and an associate editor for *Transportation Research Part C*, *Transportation Science*, and *Transport-metrica B*. He is also a member of many editorial boards including *Transportation Research Part B*. In 2015, he was awarded for 5 years an ERC consolidator grant in Social Science and Humanities. In 2020, he was awarded the "Grand Prix de l'Université de Lyon", a career award for his achievement in the transportation field.