IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 7, 2021, accepted October 22, 2021, date of publication October 26, 2021, date of current version November 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123273

Traffic Signal Control Under Mixed Traffic With Connected and Automated Vehicles: A Transfer-Based Deep Reinforcement Learning Approach

LI SONG AND WEI FAN
USDOT Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE), Department of Civil and Environmental Engineering, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Corresponding author: Wei Fan (wfan7@uncc.edu)

This work was supported in part by the United States Department of Transportation and University Transportation Center through the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) at The University of North Carolina at Charlotte under Grant 69A3551747133.

ABSTRACT Backgrounds: The traffic signal control (TSC) system could be more intelligently controlled by deep reinforcement learning (DRL) and information provided by connected and automated vehicles (CAVs). However, the direct training procedure of the DRL is time-consuming and hard to converge. Methods: This study improves the training efficiency of the deep Q network (DQN) by transferring the well-trained action policy of a previous DQN model into a target model under similar traffic scenarios. Different reward parameters, exploration rates, and action step lengths are tested. The performance of the transfer-based DQN-TSC is analyzed by considering different traffic demands and market penetration rates (MPRs) of CAVs. The information level requirements of the DQN-TSC are also investigated. Results: Compared to directly trained DQN, transfer-based models could improve both the training efficiency and model performance. In high traffic scenarios with a 100% MPR of CAVs, the total waiting time, CO2 emission, and fuel consumption in the transfer-based TSC decrease about 38%, 34%, and 34% compared to pre-timed signal schemes. Also, the transfer-based TSC system requires more than 20% to 40% MPRs of CAVs under different traffic demands to perform better than pre-timed signal schemes. Conclusions: The proposed model could improve both the traffic performance of the TSC system and the training efficiency of the DQN model. The insights of this study should be helpful to planners and engineers in designing intelligent signal intersections and providing guidance for engineering applications of the DQN TSC systems.

INDEX TERMS Deep reinforcement learning, traffic signal control, transfer learning, mixed traffic, connected and automated vehicles.

I. INTRODUCTION

With the rapid development of learning-based artificial intelligence technologies, combining the management of transportation systems with reinforcement learning (RL) technologies provides a new potential solution to improve the efficiency, safety, and sustainability of intelligent transportation systems. Also, the emerging development of the vehicle to infrastructure (V2I) communication technology enables connected vehicles (CVs) or connected and autonomous vehicles (CAVs) to transmit real-time information on vehicles to the traffic signal control (TSC) system. All these technologies make it feasible to control an intelligent TSC system by RL technologies.

Several studies optimized TSC strategies by assuming a 100% market penetration rate (MPR) of CAVs so that the TSC system could obtain information from all vehicles [1], [2]. MPRs of CAVs could determine the information level of vehicles that can be obtained by the TSC system. There is still a long transition time to achieve high MPRs of CAVs [3]. Hence, it remains an open question about how many MPRs of CAVs are sufficient to train a relatively good RL-controlled

The associate editor coordinating the review of this manuscript and approving it for publication was Michail Makridis.

TSC system. To investigate the validity and information-level requirement of the RL-controlled TSC system, it is important to study the impacts of different MPRs of CAVs on the RL-based TSC system.

Moreover, recent studies combined RL with the deep learning methods to approximate highly nonlinear functions from complex datasets, and this deep reinforcement learning (DRL) framework could provide a better TSC performance compared to RL-only methods [4]. However, the DRL-controlled TSC system still has many shortcomings. First, the training procedure of the DRL-controlled TSC system takes a long time to converge [5]. Secondly, the DRL training procedure requires lots of samples. Thirdly, the traffic flow at the intersection rapidly changes across time and space. It is extremely hard to train a model that could accommodate several traffic scenarios for real-world applications. When implementing this method in city networks with many intersections, training a specific DRL model for each intersection is extremely cumbersome. Thus, reusing or adjusting previous models under similar traffic scenarios provides a potential and feasible solution for engineering applications.

Currently, transfer learning enables the reuse of previously trained action policy developed from a similar task to initialize the learning of a target task, and it is expected to improve the training efficiency, sample efficiency, and training performance [5], [6]. It is expected that a model trained with a higher information level could also obtain a better solution compared to models trained with only partial information [7]. Several studies found that the proposed intersection control system performs better than traditional signal schemes after certain MRPs of CAVs are reached [8]–[10]. When transferring a prior model with a higher information level into a scenario with a lower information level, the transfer-based DRL TSC is expected to outperform the directly trained model as it could take a better action, which is given by the prior policy of the pre-trained model. Also, the training efficiency is expected to be improved by following and transferring prior policy as a start point in the target task. As few studies and research efforts have been made on transfer-based DRL TSC systems, it is meaningful to reuse pre-trained models and test the model performances in scenarios with similar traffic demands and/or MPRs of CAVs.

This paper aims to explore the performance and the transferability of the transfer-based DRL technologies in TSC systems and to bridge the research gap in terms of training efficiency and validity of the transfer-based DRL model. The transfer-based DRL TSC is tested at an isolated intersection with different traffic demands and MPRs of CAVs. The rest of the paper is organized as follows: Section 2 summarizes the state-of-art DRL TSC and studies on the impact of mixed traffic in the intersection control system. Section 3 describes the methodology of the transfer-based DRL TSC system. Section 4 introduces the simulation scenarios and model settings. Section 5 presents the results and findings. Finally, the article is summarized in Section 6.

II. LITERATURE REVIEW

A. REVIEW OF REINFORCEMENT LEARNING FOR TRAFFIC SIGNAL CONTROL

Optimizing traffic signal control with a reinforcement learning method has received great attention in previous studies [2], [11], [12]. The TSC agent is trained to learn an optimum policy for developing the signal phase or timeplan based on the information gathered from the traffic environment. With regards to the number of RL agents, these studies could be classified into centralized TSC with a single agent RL (for an isolated intersection or the entire intersection network) or decentralized TSC with multi-agent RL (for a network of intersections). The state of vehicles (numbers, locations, speeds, or other traffic performance criteria) is usually presented by image-like representation format (i.e., discrete traffic state encoding) or feature-based state vectors [2]. The actions are commonly defined as binary action sets (whether or not to prolong the green time) or multi-phase sets (usually four or eight green phases). Due to the large scale of the state and action representation, many recent TSC studies employed deep learning (neural networks) to approximate Q-values, which are returns for taking an action A at a state S [7]. Based on the target estimated by the deep learning, the deep reinforcement learning (DRL) could be classified into value-based (estimating Q value), policy-based (estimating action policy probability), and state-value-based method (estimating both Q value and action policy probability, such as actor-critic (A2C) framework).

Table 1 summarizes several deep reinforcement learning studies for traffic signal control systems. One of the earliest neural-network-based RL models for TSC was proposed in [13]. However, it was different from the typical deep Q network (DQN) algorithm as the lack of experience replay and the target network. After that, Genders and Razavi [14] implemented a convolutional neural network (CNN) to approximate the Q values for a single intersection with four green phases. The simulation in SUMO showed a better result compared to that using a single-layer neural network Q-learning approach. Wei *et al.* [15] introduced a DQN-based TSC, called IntelliLight, and utilized CNN to extract traffic features from real-world camera data collected in China. The IntelliLight is also selected as a benchmark in [5]. This research introduced a transfer learning framework with source task selection and batch learning. Results based on the real-world data from China indicated a quicker model convergence and better traffic performance compared to non-transfer models. Shi and Chen [4] also utilized transfer learning to speed up the training procedure of multi-agent DRL TSC with long short-term memory (LSTM) layers (a type of recurrent neural network, RNN) for Q-value approximation. The results on 2-by-2 grids of intersections indicated lower average delay compared to Q-learning and fixed-time signal under both low and high traffic demands. Moreover, Zhang *et al.* [7] trained a DQN for TSC with partial detection of vehicles. Results indicated that the DQN controlled TSC

TABLE 1. Summary of deep reinforcement learning studies for traffic signal control.

Paper	Scenario	Approach	Simulator	Result comparison
[4]	2 by 2 grid of intersections	Transfer DQN-RNN, multi-agent,	USTCMTS	Fixed-time signal, Q-learning
[5]	Single intersection, real data	Targeted Transfer DQN, CNN and LSTM, single-agent	SUMO	IntelliLight
[7]	Single intersection	DQN, single-agent	SUMO	Fixed-time signal
[11]	Single intersection	Double dueling DQN-CNN, single-agent	SUMO	Fixed-time signal, actuated signal, DQN
[14]	Single intersection	DQN-CNN, single-agent	SUMO	Q-learning with neural network
[15]	Single intersection, real data	DQN-CNN (IntelliLight), single-agent	SUMO	Fixed-time signal, Self-organizing TSC, DQN
[16]	Single intersection	Modified DQN, single-agent	VISSIM	Fixed-time signal, standard DQN
[17]	Single intersection	Asynchronous n-step Q-learning, single-agent	SUMO	actuated signal, random control
[18]	5 by 5 grid of Monaco	A2C-RNN, Multi-agent	SUMO	Q-learning, DQN, A2C

TABLE 2. Studies of the intersection control considering different MPRs of CVs/CAVs.

Paper	Object and method	Main result
[8]	Coordinate CVs at adjacent signalized intersections	Increase the MPRs of CVs would decrease fuel consumption. Fuel consumption will increase if the CV is following an HDV.
[9]	Eco-CACC (CAV) system that computes fuel-optimum trajectory	Eco-CACC system produces vehicle fuel savings up to 40% at a 100% MPR of CAVs. Lower MPRs of CAVs increase fuel consumption on multi-lane roads, and the system decreases the fuel consumption only after a 30% MPR of CAVs.
[10]	Optimizing speed of CAVs at isolated intersection	Benefits grow with the MPRs of CAVs until they level off at about 40% MPR.
[20]	Optimization of departure sequence and trajectory of CVs/AVs by maximizing the speed entering the intersection	This algorithm performs better than the actuated signal control after a 50% MPR of CVs. Even a 50% information level for CVs could significantly decrease the delay and stops.
[21]	A first-come-first-serve reservation for CAVs at an intersection	The proposed control system outperforms traffic signals after a 75% MPR of CAVs.
[22]	Cumulative travel-time responsive (CTR) real-time intersection control	The CTR algorithm improves the system performance after a 30% MPR of CVs. CTR algorithm outperformed the actuated controls after a 70% MPR of CVs.

could efficiently reduce the average waiting time even with a low detection rate.

B. REVIEW OF STUDIES ON MIXED TRAFFIC AT THE INTERSECTION

As shown in Table 2, it is noted that most simulation-based studies indicated a positive effect of the mixed traffic flow of human driving vehicles (HDVs) co-existing with CVs/AVs/CAVs at the intersection. Shladover *et al.* [19] found that a 40% MPR of the Cooperative Adaptive Cruise Control (CACC) vehicle is a critical threshold to achieve a 10% improvement of the highway capacity based on field experiment data. Yang *et al.* [20] also found a 50% information level for CVs could significantly decrease the delay and stops by maximizing the speed entering the intersection. However, several studies also found that the intersection performance only improved after certain MPRs of CVs/CAVs are reached [10], [21], [22]. Moreover, several studies found that the models trained with only partial information and the interaction between CAVs and HDVs could result in a negative impact on the intersection system performance [8], [9]. For TSC with DRL technologies, the MPR of the CAVs determines the information levels of the training inputs for the DRL system. According to the above studies, it is important to study the impacts of different MRPs of CAVs as they have a high potential to impact the performance, validity, and transferability of the transfer-based DQN TSC system.

In summary, the DRL-controlled TSC systems could have a better performance than traditional signal schemes. However, the training procedure of the DRL-controlled TSC system takes a long time and requires a lot of samples to accommodate different traffic conditions. Hence, reusing or modifying previous models under similar traffic scenarios provides a potential solution to improve the training efficiency and performance. Moreover, to bridge the research gap in terms of training efficiency and validity of the transfer-based DRL model, the impacts of different MPRs of CAVs on the DRL-TSC system still need to be further investigated.

III. MATH

In this study, the traffic light at an isolated intersection is controlled by a single agent DRL that interacts with the simulation environment. With the V2I technology, the TSC agent could choose an action a_t based on the state s_t and reward r_t transmitted by the CAVs in the timestep t. In this case, the information levels of vehicles within the simulation system are determined by the MPRs of CAVs. The Deep Q Network (DQN), which is a benchmark DRL method, is implemented in this research to train the TSC system. The action set A_t includes green phases for traffic movements. The state s_t is the traffic volume/state in each inlet segment of the intersection and is transmitted by the V2I technology of CAVs. The detailed definitions and settings for the action and

state are given in the empirical settings part. The framework of the transfer-based DRL TSC system is shown in Fig. 1.

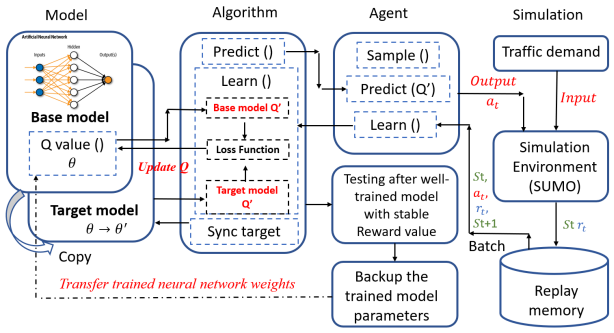


FIGURE 1. Framework of the transfer-based deep Q learning method.

The reward r_t denotes the feedback after the agent chosen an action a_t. Several traffic performance criteria are utilized as the reward in TSC systems, such as the queue length, throughput, and total waiting time [2], [12]. The total waiting time is the sum of the time for vehicles when the vehicle speed is less than 0.1 m/s. In comparison to the queue length and the throughput, the total waiting time considers both traffic volume and stopping time. Hence, the total waiting time is selected to describe the reward in this paper. Also, according to [12], a hyperparameter is added in the reward function to improve the training efficiency. The reward function is defined as:

$$r_t = \delta \times twt_{t-1} - twt_t \tag{1}$$

where twt_t denotes the total waiting time at the time step t, and δ ($\delta \leq 1$) could increase the magnitude of the reward value and is supposed to improve the training efficiency. When $\delta = 1$, the reward function changes to a commonly used reward function. The positive reward r_t denotes a better performance as the current action decreases the twt_t.

For value-based RL, Q learning is a benchmark model-free reinforcement learning technology [2]. The Q value denotes all rewards the agent could obtain when taking an action a_t in state s_t, and it could be approximated by selecting the action a_{t+1} that obtains the maximum Q value $Q^{'}$:

$$\begin{aligned} Q(s_t, a_t) &= r_{t+1} + \gamma \cdot r_{t+2} + \ldots + \gamma^{y-1} \cdot r_{t+y} \\ &\approx r_{t+1} + \gamma \cdot \max_{A} Q'(s_{t+1}, a_{t+1}) \end{aligned} \tag{2}$$

where $Q'(s_{t+1}, a_{t+1})$ is the Q value for taking an action a_{t+1} in the state s_{t+1}. γ is the discount rate that adds a penalization of the future reward compared to the immediate reward r_{t+1}. γ is set as 0.25 according to the test results in [12].

With the help of deep learning, a deep neural network is implemented to approximate the Q value. Experience replay is implemented to store and extract a batch of samples from the reply memory database. The random selection of the samples could mitigate correlations in samples and improve the utilization rate of the samples. As shown in Fig. 1, the deep Q network (DQN) contains two neural networks to improve the stability of the training results. The Q' value is the predicted value from the Neural Networks (NN) based on a given input sample. In the DQN framework, two NNs will be implemented for Q value prediction (with one from the base NN model and the other from the target NN model). Then, the difference between those two Q' values will compose a loss function which will be used to update the weights of the NN. After one episode updating of the base NN, the weights of the base NN are copied/updated to the target NN (i.e., synchronizing process).

The loss function $L(w)$ is denoted as the simple square error between Q' predicted from the base NN and the target NN network.

$$L(w) = E\left[\left(r + \gamma \max_{a'} Q_t\left(s', a', w\right) - Q_t(s, a, w)\right)^2\right] \tag{3}$$

To minimize the loss function $L(w)$, the Adaptive Moment Estimation (Adam) (i.e., a stochastic gradient descent method) is implemented. The weights w in the neural network are updated with the learning rate α as follows,

$$\frac{\partial L(w)}{\partial w} = E\left[\left(r + \gamma \max_{a'} Q\left(s', a'\right) - Q(s, a)\right) \frac{\partial Q(s, a, w)}{\partial w}\right] \tag{4}$$

$$w_{t+1} = w_t - \alpha \frac{\partial L_t}{\partial w_t} \tag{5}$$

Moreover, the epsilon-greedy method is used to explore possible actions at the beginning of the training stages. The agent would randomly choose an action with a probability of ϵ_h. Otherwise, the agent chooses the action a_{t+1} that obtains the maximum $Q^{'}$ value predicted from the training neural network,

$$\epsilon_h = 1 - \frac{h}{H} \tag{6}$$

where h is the current episode number. H is the total number of simulation episodes.

For two similar traffic scenarios, the trained policy for the action selection in one model is supposed to be useful and could be treated as an initial policy for another model [5], [6]. The transfer learning enables the reuse of a previously trained model between similar tasks. As the training procedure of the DRL is cumbersome and time-consuming, it is expected that transfer learning could improve the training efficiency and performance (when transferring models with higher information levels of the vehicles). In this paper, the neural network weights w in a prior task are transferred into a target task in scenarios with similar traffic demands or traffic information levels (determined by the market penetration rates of CAVs). The detailed algorithms of the DQN with the experience replay and transferred procedure are shown in Table 3.

In this paper, the total waiting time, CO_2 emission, and fuel consumption are utilized to investigate the traffic performance of the intersection system. All these three criteria are retrieved from the SUMO software. The waiting time of a vehicle/lane is calculated by accumulating the time when the

TABLE 3. Algorithms of deep Q network with the experience replay and transferred procedure.

Initialize experience replay memory D
Initialize the agent to interact with the environment
Get the current episode number h and the total number of simulation episodes H
If transferred from previous model
  Synchronize base and target neural network weights w and w' from previous model
Else
  Randomly initialize base and target neural network weights w and w'
End if
While cumulative reward value not converged **do**
  /*Sample phase
  Choose an action from states using policy ϵ_h-greedy(Q)
  If probability $\epsilon_h \leq 1 - \frac{h}{H}$
    Select a random action a_t
  Else
    Select a $a_t = \underset{a'}{argmax}\, Q_t(s_t, a, w)$
  End If
  Agent takes action a_t, observe reward r_t, and next state s_{t+1}
  Store sample (s_t, a_t, r_t, s_{t+1}) in the experience replay memory D
  If enough experiences in D **then**
    /*Learn phase
    Sample a random batch of N samples from D
    For every sample (s_t, a_t, r_t, s_{t+1}) in N **do**
      Set $\hat{Q}_t = \begin{cases} r_t & \text{If episode terminated at time step t} \\ r_{t+1} + \gamma \cdot \underset{A}{max} Q'(s_{t+1}, a_{t+1}) & \text{Otherwise} \end{cases}$
      Calculate the loss function value $L(w)$
      Update weight w using the SGD algorithm by minimizing $L(w)$
      Every C steps, copy weights from base NN to target NN
    End For
  End If
End While
Backup the trained NN weights w and w'
Test the performance of the TSC system

vehicle speed decreases to a value below 0.1m/s. Also, the waiting time would be reset to 0 after the vehicle moves. The emission and fuel consumption models of the gasoline-driven passenger car (Euro norm 4) are developed and calculated by the HBEFA3 (version 3.1.). The details of the calculation procedure and emission/fuel consumption factors could be referred to the HBEFA3 [23].

IV. EMPIRICAL SETTINGS

A. SIMULATION SCENARIOS

A typical four-way intersection with four lanes per approach is selected for the simulation. As shown in Fig. 2, the vehicle-based state array (the number of vehicles in each grid/segment) is determined by the discrete traffic state encoding (DTSE) method and is set as the input for the DQN model. Eight green phases are set as possible actions for the intersection. A 4-s yellow and all red-time is added if the

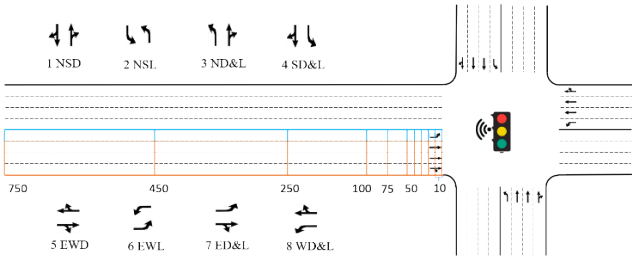


FIGURE 2. Discrete traffic state encoding of the vehicle-based state array and available traffic signal actions of a four-way intersection.

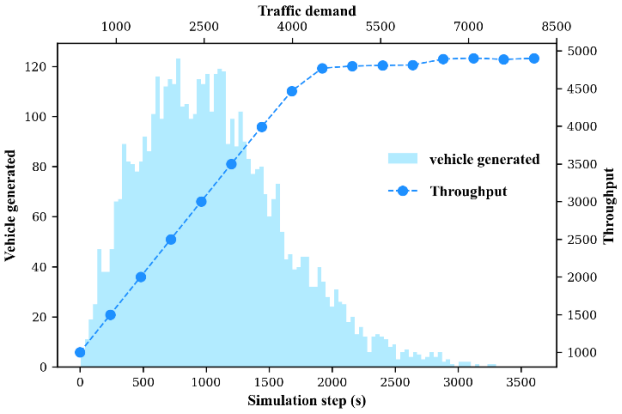


FIGURE 3. Traffic generated per simulation step and throughputs (vph) under different traffic demands.

traffic light changes its phase. The speed limit is set at 35 mph (i.e., 15.6 m/s). The peak-hour traffic, which is the main reason for the congestion at the intersection, is generated according to a Weibull distribution with a shape equal to 2. Meanwhile, the random seed, which equals the episode value, is utilized to generate heterogeneous traffic for each training episode. The saturated traffic demands of the intersection are determined by the simulated maximum throughput under a pre-timed signal scheme (100-s cycle length, two 30-s phases for the direct and right traffic, two14-s phases for the left-turn traffic). As shown in Fig. 3, with the increase of the traffic demand, the maximum throughput of the intersection increases to 4800 vehicles/hour in the simulation, and this saturated traffic is set as the high traffic demand scenario. The low, medium, and medium-high traffic demands are set at 20%, 40%, and 60% of the high traffic demand, respectively. The detailed traffic demand for each movement is presented in Table 4.

All simulation scenarios are processed in the Simulation of Urban MObility (SUMO) by the TraCI-Python interface. Each training episode of the simulation is set as 3600-s with a 0.1-s time step to accommodate the distribution of the peak hour traffic volume. The Intelligent Driving Model (IDM) is implemented for human driving vehicles (HDVs) according to [24]. The IDM has a simple model structure, accident-free logic, and continuous acceleration control function that can be used to describe the longitudinal movements of the

TABLE 4. Different traffic demand scenarios.

Traffic demand (veh/hr)	Low	Medium	Medium -High	High
Left traffic per approach	72	108	144	180
Through traffic per approach	336	504	672	840
Right traffic per approach	72	108	144	180
Total traffic per approach	480	720	960	1200
Total Throughput	1920	2880	3840	4800

HDVs/AVs [25]. A current study by Adil [26] also indicated that the IDM could have a better speed and acceleration control accuracy compared to the Krauss model or Wiedmann 99 model (a default car-following model adopted in VISSIM software). The Cooperative Adaptive Cruise Control (CACC) system is utilized for CAV simulation according to previous research [27], [28]. The default lane change model "LC2013" in SUMO is employed for all vehicles. Both HDVs and CAVs are assumed to have the same ability for acceleration (2 m/s^2) and deceleration (-4 m/s^2). The desired headways for HDVs and CAVs are 1.6-s and 0.7-s, respectively. To model heterogeneous driving behaviors of the human drivers, the maximum speed for the HDV follows a normal distribution $N(1.2, 0.1)$ with respect to the speed limits. Also, other parameters for CACC controlled CAVs are set according to previous research [28]–[31].

In this study, all vehicles are set as CAVs at first. A direct training procedure with 800 episodes is employed under the low traffic demand scenario. Then, the trained model is transferred to the next scenario with a higher traffic demand (from low to medium, medium to medium-high, and medium-high to high). After that, this paper tests the validity of the transfer-based DQN signal system by considering different information levels of the vehicles. For scenarios with the same traffic demand, the MPR of CAVs decreases from 100% to 20% by 20% per step. The trained model with higher MPRs of CAVs will be transferred into the subsequent scenario with lower MPRs of CAVs.

B. MODEL SETTINGS

A medium-sized fully connected neural network (4 hidden layers with 400 neurons per layer) is implemented in this study. This size of the NN is recommended as it could obtain good training results and save a lot of training time according to [12]. The NN is built in TensorFlow 2.0. The Rectified Linear Unit (ReLU) activation function is implemented for all hidden layers and the Liner activation function is used for the output layer. The Adam optimization algorithm is implemented for training NN models. The discount factor of the Q-learning equation is set at 0.25. The training iterations of the neural network weights will execute 800 times with a 0.01 learning rate, and each iteration will retrieve 100 samples according to the memory replay [12]. The total number of the training times for neural networks in one simulation episode is determined by the action step length and total simulation steps. The test results will be output after the convergence of the cumulative reward values. The following parts test the reward function parameter and the action step length.

As introduced in [12], the revision of the parameter ($\gamma = 0.9$) in the reward function could increase the magnitude of the reward value and improve the training efficiency. As shown in Fig. 4, this paper compares the results between the general reward parameter ($\gamma = 1$) and the revised reward parameter ($\gamma = 0.9$) for the transfer-based DQN procedure under the scenario of medium traffic and a 100% MPR of CAVs. The reward curves indicate that the proposed reward parameter ($\gamma = 0.9$) could not always improve the training efficiency and could result in more variations in action choices. Hence, the general reward function ($\gamma = 1$) is utilized in this paper.

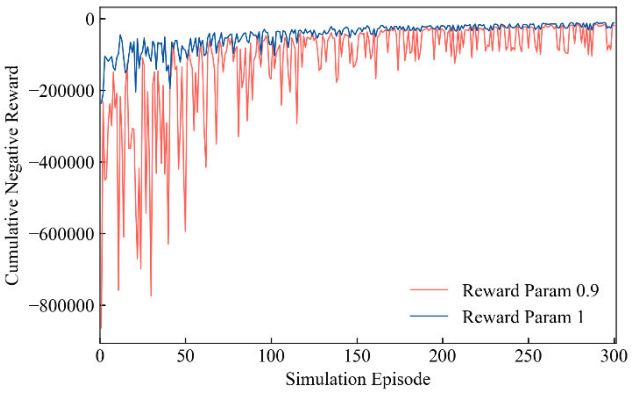


FIGURE 4. Reward curves for different reward function parameters in transfer-based models.

The ε-greedy exploration rate introduced in (6) is utilized to strike a balance between the exploration and exploitation of the actions. In general, the DQN training procedure is expected to explore more possible actions at the beginning and then exploit more when the action policy is well trained. As the transfer-based learning procedure could obtain prior action policy from previous scenarios, the training procedure might obtain the converged value without exploring all possible actions. To confirm this assumption, different ε-greedy exploration rates are tested, and the results are illustrated in Fig. 5. It is found that without full exploration (ε changes from 1 to 0), the transfer-based models could also obtain a similar stable reward, which indicates the validity of transferring models from similar scenarios.

When the current action a_t is different from the previous action a_{t-1}, a phase that includes 3-s yellow and 1-s all-red time is added. If the agent selects the same action (green phase) and that green phase exceeds a maximum cumulative green time (60 s in this paper), the agent would stop the current phase and change to the next green phase. As shown in Table 5, the model performances with different action step lengths (green time durations) are tested under a low traffic demand scenario. The start value of the action step length is set as the minimum green time (5s). It is noted that a

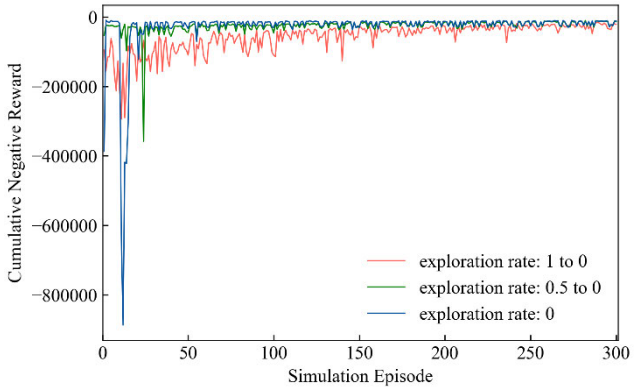


FIGURE 5. Reward curves under different ε-greedy exploration rate boundary in transfer-based models.

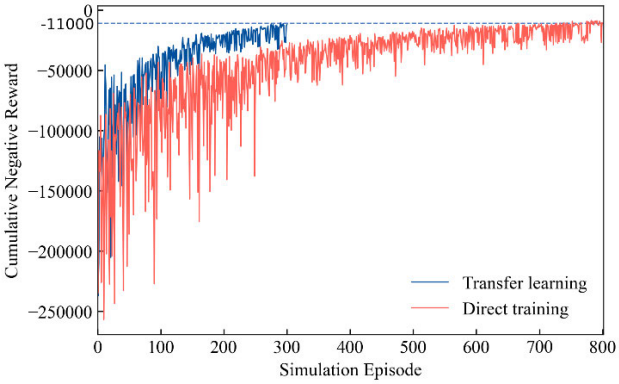


FIGURE 6. Comparison between the reward curves of direct and transfer-based leaning approaches.

TABLE 5. Green time duration per action for the DQN signal controller.

Green time (s)	Total waiting time (s)	Total CO_2 (kg)	Total fuel (L)
5	23191.3	318.8	137
10	24942.2	323.7	139.2
15	32772.6	344.9	148.3
20	39815.3	363.2	156

significant increase in the total waiting time, CO_2 emission, and fuel consumption are observed after 10-s of the green time. Also, the frequent change of the green phase would add more red/yellow time to the total time, and this would result in more green time loss. Hence, this paper sets 10-s green time for each action and 60-s for the maximum green time duration.

V. RESULTS AND DISCUSSIONS

A. COMPARISON BETWEEN DIRECT AND TRANSFER-BASED LEARNING

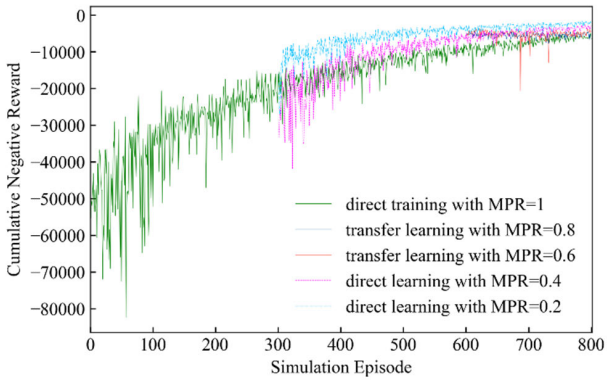
To test the efficiency of the transfer-based DQN approach, a comparison between direct training and transfer-based training with full exploration (ε-greedy from 1 to 0) is made under a scenario with medium traffic and a 100% MPR of CAVs. The cumulative negative reward curves in Fig. 6 demonstrate that the transfer-based method could get the stable maximum value with fewer training episodes compared to the direct training procedure. This result further proves that the prior action policy (neural network weights) provided by the previous model could be utilized in target models under similar traffic scenarios and promote the training efficiency with fewer adjustments of the pre-trained model. It is noted that the direct training procedure for an intersection with different traffic demands is very time-consuming. For example, in Fig. 6, the direct training and the transfer-based training take about 54.2 hours and 20.1 hours, respectively, in a computer with GTX-1050 GPU (for neural network training) and i5-7300 2.5GHz CPU. The significant decrease in the training time gives a possible engineering application of the transfer-based DQN TSC system at intersections with similar traffic demands.

B. IMPACTS UNDER DIFFERENT TRAFFIC DEMANDS AND MPRS OF CAVs

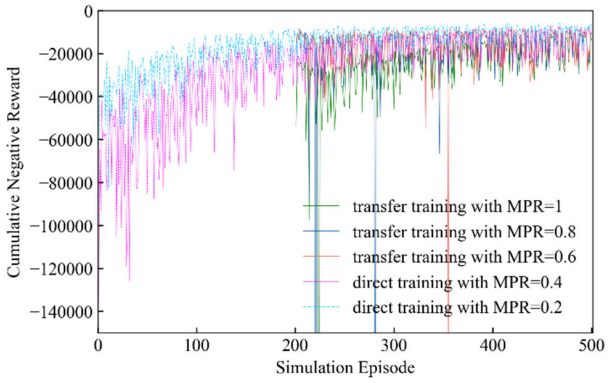
With the V2I communication technology, the TSC system could obtain state information (i.e., traffic volume, speed, waiting time, etc.) from the CAVs approaching the intersection. However, it is expected to have a long transition period during which human driving vehicles and intelligent vehicles will coexist [3]. This paper also tests the impacts of information levels of the mixed traffic on the transfer-based DQN TSC system.

Fig. 7 presents cumulative negative reward curves for scenarios with different traffic demands and MPRs of CAVs. First, with a 100% MPR of CAVs, the prior-trained NN weights of the trained DQN model are transferred from scenarios with high traffic demands to scenarios with low traffic demands (i.e., from low to medium, medium to medium-high, and medium-high to high). After that, for scenarios with the same traffic demands, the impacts of information levels of the vehicles on the DQN TSC system are investigated by transferring models with high MPRs of CAVs into models with low MPRs of CAVs (decreasing from 100% to 20% by 20% per step). For example, the direct training model for the scenario with a 100% MPR of CAVs is transferred to the scenario with an 80% MPR of CAVs. It is also noted that direct training procedures are utilized in some low MPRs scenarios to obtain more stable reward values at the end of the training procedure.

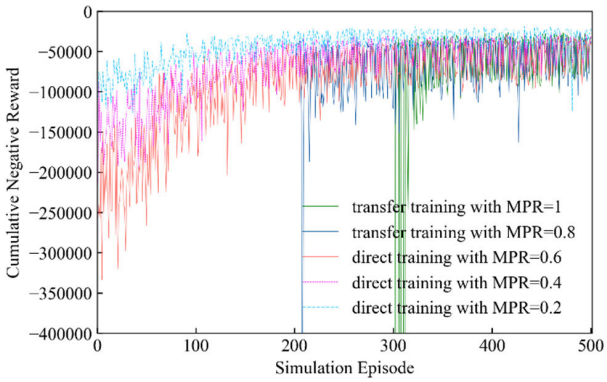
An interesting finding is that the reward values of the transfer-based model overlap with the reward values in models with higher MPRs of CAVs (i.e., higher information level). For example, in Figure 7 (b), the transfer-based curve with 60% and 80% MPRs of CAVs overlaps with the directly trained curve with 20% and 40% MPRs of CAVs. This overlapping is scarcely observed in directly trained models as the TSC system is trained based on partial traffic information. For scenarios with lower MPRs of CAVs, the cumulative



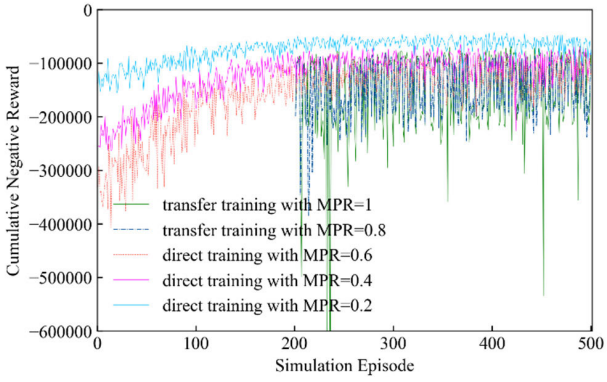
(a) Rewards under low traffic demand



(b) Rewards under medium traffic demand



(c) Rewards under medium-high traffic demand



(d) Rewards under high traffic demand

FIGURE 7. Training rewards for scenarios with different traffic demands and MPRs of CAVs.

TABLE 6. Performance comparison between transfer-based models and directly trained models.

Models	Transfer	Direct	Transfer	Direct	Transfer	Direct
Traffic demands (vph)	2880	2880	3840	3840	4800	4800
MPRs of CAV (%)	0.2	0.2	0.6	0.6	0.6	0.6
Total waiting time (s)	12432 3	16516 3	30723 4	31723 3	63943 0	67095 7
Total CO2 (kg)	653	745	1258	1311	2297	2520
Total Fuel (L)	281	320	541	564	988	1083

negative rewards values are also lower due to the missing of some rewards from HDVs (TSC could not obtain state values from HDVs). In this case, the final stable reward value would not overlap with others in directly trained models. The overlapping indicates that transfer-based models could obtain larger reward values than the directly trained models. The traffic light controller could not select a better choice if the system only gets limited/biased information on the vehicle states and system rewards. Table 6 also indicates that, in the same scenarios, the transfer-based model could improve the traffic performance compared to the directly trained models.

Table 7 to Table 9 present the test performance (total waiting time, total CO_2 emission, and total fuel consumption) of the proposed DQN TSC under different traffic demands and MPRs of the CAVs. Compared to the scenario with fixed signal schemes, a decrease in the total waiting time, CO_2 emission, and fuel consumption could be observed in scenarios with more than a 40% MPR of CAVs. Meanwhile, a decrease in indicator values (i.e., better system performance) could be observed with the increase of MPRs of CAVs. The DQN controlled signal system could get worse traffic performance with a 20% MPR of CAVs under low-, medium-, and high-traffic scenarios. Also, the performance indicator values decrease significantly when the MPRs of the CAVs increase from 20% to 40%. These results indicate that the proposed transfer-based DQN signal controller needs a certain information level of the vehicles, and the critical value

TABLE 7. Total waiting time for scenarios with different MPRs of CAVs and traffic demands.

Total waiting time (s)	Traffic demand			
	Low	Medium	Medium-high	High
Fixed signal	48009.7	129770.2	424350.3	816397.9
20% MPR	57641 (-0.201)	165163.1 (-0.273)	388409.3 (0.085)	882875 (-0.081)
40% MPR	35624 (0.258)	93565 (0.279)	341656.4 (0.195)	729716.3 (0.106)
60% MPR	29635.2 (0.383)	83949.1 (0.353)	317232.5 (0.252)	670957.4 (0.178)
80% MPR	29912.3 (0.377)	67767.1 (0.478)	251353.4 (0.408)	604543.4 (0.259)
100% MPR	24942.2 (0.48)	54472.9 (0.58)	218563.7 (0.485)	508737.1 (0.377)

*note: numbers in parentheses denote the change rate compare to fixed signal schemes

TABLE 8. Total CO_2 emission for scenarios with different MPRs of CAVs and traffic demands.

Total CO_2 (kg)	Traffic demand			
	Low	Medium	Medium-high	High
Fixed signal	363.1	684.2	1647.7	3086.1
20% MPR	369.3 (-0.017)	744.9 (-0.089)	1615.4 (0.02)	3089.3 (-0.001)
40% MPR	324.9 (0.105)	582.6 (0.148)	1419.2 (0.139)	2650.1 (0.141)
60% MPR	318.9 (0.122)	573.7 (0.162)	1311.3 (0.204)	2519.5 (0.184)
80% MPR	328.2 (0.096)	554.1 (0.19)	1152.2 (0.301)	2206.9 (0.285)
100% MPR	323.7 (0.109)	537.2 (0.215)	1105.2 (0.329)	2023.8 (0.344)

*note: numbers in parentheses denote the change rate compare to fixed signal schemes

TABLE 9. Total fuel consumption for scenarios with different MPRs of CAVs and traffic demands.

Total fuel (L)	Traffic demand			
	Low	Medium	Medium-high	High
Fixed signal	156.1	294.1	708.3	1326.6
20% MPR	158.8 (-0.017)	320.2 (-0.089)	694.4 (0.02)	1328 (-0.001)
40% MPR	139.6 (0.106)	250.4 (0.149)	610.1 (0.139)	1139.2 (0.141)
60% MPR	137.1 (0.122)	246.6 (0.162)	563.7 (0.204)	1083.1 (0.184)
80% MPR	141.1 (0.096)	238.2 (0.19)	495.3 (0.301)	948.7 (0.285)
100% MPR	139.2 (0.108)	230.9 (0.215)	475.1 (0.329)	870 (0.344)

*note: numbers in parentheses denote the change rate compare to fixed signal schemes

of the information level is between 20% to 40% according to different traffic demands. Moreover, with a 100% MPR of CAVs in a medium traffic demand scenario, the DQN TSC system indicates a decrease of 58% of the total waiting time, which is the best performance in total waiting time. For scenarios with high traffic demand, fixed signal schemes indicate significant congestion as all performance values almost doubled compared to medium-high traffic scenarios. However, for DQN TSC with a 100% MPR of CAVs, the total waiting time, CO_2 emission, and fuel consumption still decrease about 38%, 34%, and 34%, respectively.

VI. CONCLUSION

This paper presents a transfer-based DQN traffic light control system to improve the training efficiency of the deep reinforcement learning procedure. Different model settings (reward parameter, exploration rate, and action step length) are tested and discussed. Different traffic demands are determined according to the simulated maximum throughput of the intersection. The impacts of traffic demands and information levels of the vehicles on the transfer-based model are investigated. The trained DQN models are first transferred from scenarios with low traffic demands into scenarios with higher traffic demands (from low to medium, medium to medium-high, and medium-high to high). For scenarios with the same traffic demand, models are then transferred from high MPRs of CAVs scenario into low MPRs of CAVs scenario (decrease from 100% to 20% by 20% per step).

The result comparison between the transfer-based training procedure and direct training procedure indicates that the prior action policy of the DQN TSC model could be utilized in models with similar traffic demands or information levels of vehicles. The training efficiency is improved significantly in transfer-based models. Also, this paper tests the validity of the transfer-based DQN method by considering different information levels of vehicles. In this paper, the information level is determined by the MPR of CAVs and is transmitted to the TSC system. With the increase of MPRs of CAVs, a decrease in the total waiting time, CO_2 emission, and fuel consumption could be observed in transferred-based DQN TSC systems. Compared with pre-time signal schemes, the transferred-based DQN TSC systems perform better when the MPRs of CAVs are more than 20% under the medium-high traffic scenario and more than 40% under low, medium, and high traffic scenarios. Moreover, the transfer-based models could choose actions given by previous models with higher information level. Hence, the transfer-based model could choose actions with better performance than the model directly trained by the same information level.

The good performances in efficiency, validity, and transferability of the transfer-based DQN TSC method indicate a possible engineering application of this method in scenarios with similar traffic demands or information levels. With the rapid development of vehicles with V2I communication technologies, the information level requirement (between 20% and 40%) for this transfer-based DQN TSC system is expected to be met in the near future. These findings should be valuable to transportation researchers, decision-makers, and traffic engineers to improve intersection efficiency, design intelligent intersections, promote the technologies of V2I, and implement DQN-controlled traffic signals. Please note that adjacent intersections are more commonly seen in corridors or road networks. The coordination between adjacent intersection signals requires a multiagent DRL framework which is different from the single DQN framework used for the isolated intersection. Future research efforts could be focused on modeling multiagent signal controllers for adjacent intersections or intersection networks. Also, the size of the neural networks of the DQN model deserves further investigation.

REFERENCES

[1] Q. Guo, L. Li, and X. J. Ban, "Urban traffic signal control with connected and automated vehicles: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 101, pp. 313–334, Apr. 2019, doi: 10.1016/j.trc.2019.01.026.

[2] A. Haydari and Y. Yilmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 22, 2020, doi: 10.1109/TITS.2020.3008612.

[3] G. Sharon and P. Stone, "A protocol for mixed autonomous and human-operated vehicles at intersections," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2017, pp. 151–167, doi: 10.1007/978-3-319-71682-4_10.

[4] S. Shi and F. Chen, "Deep recurrent Q-learning method for area traffic coordination control," *J. Adv. Math. Comput. Sci.*, vol. 27, no. 3, pp. 1–11, May 2018, doi: 10.9734/JAMCS/2018/41281.

[5] N. Xu, G. Zheng, K. Xu, Y. Zhu, and Z. Li, "Targeted knowledge transfer for learning traffic signal plans," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, vol. 11440, Apr. 2019, pp. 175–187, doi: 10.1007/978-3-030-16145-3_14.

[6] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 1, 2021, doi: 10.1109/TITS.2021.3054625.

[7] R. Zhang, A. Ishikawa, W. Wang, B. Striner, and O. K. Tonguz, "Using reinforcement learning with partial vehicle detection for intelligent traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 404–415, Jan. 2021, doi: 10.1109/TITS.2019.2958859.

[8] Z. Du, B. HomChaudhuri, and P. Pisu, "Coordination strategy for vehicles passing multiple signalized intersections: A connected vehicle penetration rate study," in *Proc. Amer. Control Conf. (ACC)*, May 2017, pp. 4952–4957, doi: 10.23919/ACC.2017.7963722.

[9] H. Yang, H. Rakha, and M. V. Ala, "Eco-cooperative adaptive cruise control at signalized intersections considering queue effects," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1575–1585, Jun. 2017, doi: 10.1109/TITS.2016.2613740.

[10] H. Jiang, J. Hu, S. An, M. Wang, and B. Park, "Eco approaching at an isolated signalized intersection under partially connected and automated vehicles environment," *Transp. Res. C, Emerg. Techn.*, vol. 79, pp. 290–307, Jun. 2017, doi: 10.1016/j.trc.2017.04.001.

[11] X. Liang, X. Du, G. Wang, and Z. Han, "A deep reinforcement learning network for traffic light cycle control," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1243–1253, Feb. 2019, doi: 10.1109/TVT.2018.2890726.

[12] A. Vidali, "Simulation of a traffic light scenario controlled by a deep reinforcement learning agent," M.S. thesis, Univ. Milano-Bicocca, Milano, Italy, 2018.

[13] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," *IET Intell. Transp. Syst.*, vol. 4, no. 2, pp. 128–135, 2010, doi: 10.1049/iet-its.2009.0070.

[14] W. Genders and S. Razavi, "Using a deep reinforcement learning agent for traffic signal control," 2016, *arXiv:1611.01142*. [Online]. Available: http://arxiv.org/abs/1611.01142

[15] H. Wei, G. Zheng, H. Yao, and Z. Li, "IntelliLight: A reinforcement learning approach for intelligent traffic light control," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2496–2505, doi: 10.1145/3219819.3220096.

[16] C.-H. Wan and M.-C. Hwang, "Value-based deep reinforcement learning for adaptive isolated intersection signal control," *IET Intell. Transp. Syst.*, vol. 12, no. 9, pp. 1005–1010, 2018, doi: 10.1049/iet-its.2018.5170.

[17] W. Genders and S. Razavi, "Asynchronous n-step Q-learning adaptive traffic signal control," *J. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 319–331, Jul. 2019, doi: 10.1080/15472450.2018.1491003.

[18] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, Mar. 2020, doi: 10.1109/TITS.2019.2901791.

[19] S. Shladover, D. Su, and X.-Y. Lu, "Impacts of cooperative adaptive cruise control on freeway traffic flow," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2324, pp. 63–70, Dec. 2012, doi: 10.3141/2324-08.

[20] K. Yang, S. I. Guler, and M. Menendez, "Isolated intersection control for various levels of vehicle technology: Conventional, connected, and automated vehicles," *Transp. Res. C, Emerg. Techn.*, vol. 72, pp. 109–129, Nov. 2016, doi: 10.1016/j.trc.2016.08.009.

[21] M. Algomaiah and Z. Li, "Utilizing lane-based strategy to incorporate mixed traffic in interchange control for connected and autonomous vehicles," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 5, pp. 454–465, 2019, doi: 10.1177/0361198119839341.

[22] J. Lee, B. Park, and I. Yun, "Cumulative travel-time responsive real-time intersection control algorithm in the connected vehicle environment," *J. Transp. Eng.*, vol. 139, no. 10, pp. 1020–1029, 2013, doi: 10.1061/(ASCE)TE.1943-5436.0000587.

[23] S. Hausberger, M. Rexeis, and M. Zallinger, "Emission factors from the model PHEM for the HBEFA version 3," Graz Univ. Technol., Graz, Austria, Tech. Rep. I-20/2009, 2009.

[24] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 62, no. 2, pp. 1805–1824, 2000, doi: 10.1103/PhysRevE.62.1805.

[25] M. Treiber and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin, Germany: Springer, 2013.

[26] M. S. Adil, "Analysis and development of car-following models using xFCD," M.S. thesis, Tech. Univ. Munich, Munich, Germany, 2020.

[27] L. Xiao, M. Wang, and B. Van Arem, "Realistic car-following models for microscopic simulation of adaptive and cooperative adaptive cruise control vehicles," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2623, no. 1, pp. 1–9, Jan. 2017, doi: 10.3141/2623-01.

[28] A. Liu, H. Xiao, and L. Kan, "Using cooperative adaptive cruise control (CACC) to form high-performance vehicle streams," Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. DTFH61-13-H-00013, 2018.

[29] K. N. Porfyri, E. Mintsis, and E. Mitsakis, "Assessment of ACC and CACC systems using SUMO," *EPiC Ser. Eng.*, vol. 2, pp. 69–82, Jun. 2018, doi: 10.29007/r343.

[30] L. Song, W. Fan, and P. Liu, "Exploring the effects of connected and automated vehicles at fixed and actuated signalized intersections with different market penetration rates," *Transp. Planning Technol.*, vol. 44, no. 6, pp. 577–593, 2021, doi: 10.1080/03081060.2021.1943129.

[31] E. Mintsis, "Modelling, simulation and assessment of vehicle automations and automated vehicles' driver behaviour in mixed traffic," *TransAID Deliverable*, vol. 1, Aug. 2018, Art. no. 723390. [Online]. Available: https://www.transaid.eu/wp-content/uploads/2017/Deliverables/WP3/TransAID_D3.1_Modelling-simulation-and-assessment-of-vehicle-automations.pdf

LI SONG received the M.S. degree in transportation engineering from the Harbin Institute of Technology. He is currently a Ph.D. Research Assistant at the USDOT Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE), Department of Civil and Environmental Engineering, The University of North Carolina at Charlotte. His research interests include traffic safety, traffic control, and intelligent transportation systems.

WEI (DAVID) FAN received the Ph.D. degree in transportation engineering from The University of Texas at Austin. He is currently the Director of the USDOT Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) and a Professor with the Department of Civil and Environmental Engineering, The University of North Carolina at Charlotte. His research interests include big data analytics in transportation, connected and autonomous vehicles, shared mobility and multimodal transportation, traffic system operation and control, and transportation system analysis and network modeling. He served as the Handling Editor for *Transportation Research Record* (TRR) Inaugural Editorial Board and an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Journal of Transportation Engineering Part A: Systems* (ASCE), and the *International Journal of Transportation Science and Technology*.