IEEE Open Journal of
Intelligent Transportation
Systems

Received 21 February 2023; revised 25 April 2023, 17 June 2023, and 25 July 2023; accepted 29 July 2023. Date of publication 1 August 2023;
date of current version 22 August 2023.

Digital Object Identifier 10.1109/OJITS.2023.3300748

A Hierarchical Framework for Multi-Lane Autonomous Driving Based on Reinforcement Learning

XIAOHUI ZHANG[1], JIE SUN[2], YUNPENG WANG[3,4], AND JIAN SUN[1]

[1]Department of Traffic Engineering and Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 200092, China

[2]School of Civil Engineering, The University of Queensland, Brisbane, QLD 4072, Australia

[3]Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Beihang University, Beijing 100191, China

[4]School of Transportation Science and Engineering, Beihang University, Beijing 100191, China

CORRESPONDING AUTHOR: J. SUN (e-mail: sunjian@tongji.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 52125208 and Grant 52232015,
and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100.

ABSTRACT This paper proposes a hierarchical framework integrating deep reinforcement learning (DRL) and rule-based methods for multi-lane autonomous driving. We define an instantaneous desired speed (IDS) to mimic the common motivation for higher speed in different traffic situations as an intermediate action. High-level DRL is utilized to generate IDS directly, while the low-level rule-based policies including car following (CF) models and three-stage lane changing (LC) models are governed by the common goal of IDS. Not only the coupling between CF and LC behaviors is captured by the hierarchy, but also utilizing the benefits from both DRL and rule-based methods like more interpretability and learning ability. Owing to the decomposition and combination with rule-based models, traffic flow operations can be taken into account for individually controlled automated vehicles (AVs) with an extension of traffic flow adaptive (TFA) strategy through exposed critical parameters. A comprehensive evaluation for the proposed framework is conducted from both the individual and system perspective, comparing with a pure DRL model and widely used rule-based model IDM with MOBIL. The simulation results prove the effectiveness of the proposed framework.

INDEX TERMS Autonomous driving, decision making, deep reinforcement learning, car following, lane changing, rule-based policy, traffic adaptive.

I. INTRODUCTION

AUTONOMOUS vehicles (AVs) that drive themselves without the need for human intervention hold enormous potentials to increase the safety and efficiency of the transportation system and provide better mobility services for people and goods. As one of the most challenging tasks for autonomous driving is the decision making for safe, comfortable, and efficient vehicle maneuvers on a multi-lane highway [1]. What makes multi-lane driving significantly more challenging is the fact that multi-vehicle interaction happens both laterally and longitudinally and requires coordination between lateral and speed control [2], [3], [4]. The coupling control in a multi-lane environment needs to take into account more potential vehicles in multiple lanes, which enlarges the complexity.

In traffic flow research community, longitudinal (car following) and lateral (lane changing) behavior are usually modeled with rule-based methods separately. Numerous car-following (CF) models have been developed as an ordinary differential equation, such like Newell [5], Gipps' model [6], Intelligent Driver Model (IDM) [7] and Optimal Velocity Model (OVM) [8]. While relatively fewer studies focus on modeling lane-changing (LC) behavior due to its complexity [2], much less efforts have been made in integrating

The review of this article was arranged by Associate Editor Jiwon Kim.

CF and LC models for multi-lane driving scenarios [9]. Although these rule-based models are easy to interpret and calibrate with terms in physical meaning and mathematically tractable for safety guarantee, they are generally not compatible with complex traffic scenarios involving a variety of interactive agents which affect the subject vehicle's driving behavior.

With the recent advances of machine learning methods, more studies opt for adopting learning-based approaches for multi-lane driving decision making [10], [11], [12], [13], [14], [15] rather than conventional rule-based approaches, as learning-based approaches can achieve better fitting or optimizing performance. Among those, deep reinforcement learning (DRL) is a recently emerging approach that aims to learn an optimal driving policy by gaining more rewards through interacting with the traffic environment [16]. Many task-specific DRL applications emerge for CF or LC behavior modelling (please see a review in [17]), and DRL based multi-lane driving models which can output the CF and LC decisions directly have also been proposed (please see a review in [18]). However, these models are found to lack robustness due to multiple interactive vehicles and the lack of a causal model [19]. Additionally, another critical deficiency of learning-based approaches is the lack of interpretability [20], which is essential for studying individual behavior, in particular, AV driving, for reasons like trustability and safety of AV and transparency in governance. Moreover, current AV models based on DRL for multi-lane driving usually consider the optimal performance of the subject vehicle [18] and the performance of traffic system separately, where the optimization of traffic system adopts centralized control (i.e., all AVs are controlled to the optimize the traffic flow regardless their individual performance) rather than decentralized control (i.e., each AV achieves optimal performance and jointly contributes to the traffic flow optimization). Given that negative impacts on traffic system that have been discussed in previous AV studies [21], [22], [23], an efficient model capable of considering the traffic flow operation with individually optimized vehicles is thus of significance.

Considering the challenges of rule-based and DRL-based models for multi-lane AV driving modeling in terms of generalization, interpretation, and robustness, we aim to explore how to leverage the strengths of both DRL and rule-based methods to address the aforementioned challenges. To this end, we develop a hierarchical decision-making framework for multi-lane driving considering both individual and system performance optimization. Specifically, there are mainly four contributions of this study.

1) We combine the DRL model and rule-based policy for multi-lane driving framework in a hierarchy to inherit the advantages of both methods, where DRL is used in the high level, and the rule-based CF and LC models are developed in the low level.

2) We define the instantaneous desired speed (IDS) as the intermediate action to synergize DRL and rule-based methods. It is the direct output of DRL, which depicts inherent pursuit of speed and motivates both longitudinal and lateral movements.

3) We extend the hierarchical framework with traffic flow adaptive (TFA) strategy based on the exposed parameters in the framework, for optimizing the mixed traffic flow.

4) The performance is demonstrated in several typical traffic scenarios, including ring roads and an on-ramp bottleneck, and compared with the classical DRL model and widely used rule-based model IDM with MOBIL.

The rest of the paper is structured as follows: Section II reviews the related literature; Section III introduces the hierarchical framework; Section IV presents the training results of the framework; Section V applies the proposed framework for traffic flow optimization and discusses the evaluation results; Section VI presents the conclusions of this work.

II. LITERATURE REVIEW

With the recent massive progress of DRL, more DRL-based studies emerge in the domain of autonomous driving. This section first introduces DRL-based AV studies, followed by system-optimal DRL modeling. Lastly, we review the studies on hybrid models combining learning-based and rule-based methods.

A. DRL MODELS FOR AV DRIVING

Most DRL models for AV in the literature only concentrate on single-task driving, e.g., CF [24], LC [25], overtaking [26], ramp merging [10], intersection crossing [27] and single-lane trajectory planning [28]. For multi-lane driving, although the driving decision of CF and LC are usually generated simultaneously, the coupling between LC and CF cannot be guaranteed, while CF and LC are strongly related according to previous studies [29], [30].

By creating a policy hierarchy, Hierarchical RL (HRL) boosts both learning efficiency and solution quality [31]. Generally, HRL can be categorized into option-based [32] and goal-based framework [33]. For multi-lane driving, option-based HRL is applied in [34], where two DRL algorithms are adopt to select lane-changing actions in the upper layer and process car-following decisions in the lower layer respectively. Nonetheless, only one option (e.g., left LC, right LC and lane keeping) can be trained at a time, which makes it inefficient and the mutual interdependence of CF and LC is not fully utilized. The goal-based HRL uses a common goal to communicate subtasks, activated by a higher-level manager and implemented by a lower-level worker. It is suitable for multi-lane driving as CF and LC are motivated by inherent pursuit of optimal speed. Thus, we adopt the basic ideas of goal-based HRL in this work and combine it with rule-based models. Regarding more detailed DRL models for multi-lane driving, recent surveys are presented in [17], [18].

B. DRL MODELS FOR TRAFFIC FLOW OPTIMIZATION

Many studies have been completed for traffic flow optimization through cooperative AVs based on multi-agent

reinforcement learning (MARL), where the movement of AVs is managed by a central controller to collectively regulate traffic flow and smooth out traffic jams. The series of studies by the FLOW project team are representative works of system-optimal MARL models [35], [36], [37], [38], [39], [40]. However, centralized control overlooks the performance of each vehicle, in addition to that it can only be realized in far future with high penetration rate of connected automated vehicles. With elaborated design for reward function consisting with multiple objectives like traffic flow stability, the distributed control based on DRL model is applied to take into account both traffic flow operation and driving performance of single AV [41], [42], [43], [44], [45]. However, these studies generally neglected LC behavior for the sake of simplicity, which is one of the most basic driving behaviors and is more likely to disturb traffic flow [46].

C. HYBRID MODELS COMBINING LEARNING AND RULE-BASED METHODS

This paper aims to combine DRL with rule-based methods for improved multi-lane driving modeling. This type of hybrid models can utilize prior knowledge stemming from our observational, empirical, physical or mathematical understanding of the world as rules to enhance the performance of a learning algorithm.

Despite the advantages of the hybrid models, only little effort is made in driving behavior modeling. For example, to connect with classical rule-based CF models, neural networks of diverse architectures are employed [47]. To achieve a higher prediction accuracy, a rule-based CF model can be encoded into the neural network [48]. For DRL, common ways to embed domain knowledge include designing new reward function [43] in loss function, injecting rule-based constraints as the safeguard [49] and learning the parameters of the rules with DRL [50]. However, the coupling relationship between CF and LC is hard to express as reward function.

In summary, this paper aims to explore a novel hierarchical multi-lane driving framework based on DRL, which utilizes goal-based HRL and rule-based methods to capture the coupling CF and LC behaviors and realize both individual-level and system-level improvement with adaptation to different traffic scenarios and conditions through more exposed parameters in the framework.

III. METHODOLOGY

A. PRELIMINARIES OF DRL

DRL is the integration of reinforcement learning (RL) and deep neural networks, in which the core idea is to find the optimal policy for an agent by interacting with the environment [16]. Markov Decision Process (MDP) is the theoretical basis of the RL, which can be expressed as a tuple of $\{S, A, R, P, \rho_0\}$, where: S is a set of states; A is a set of actions; R is the reward function, with reward $r_t = R(s_t, a_t, s_{t+1})$; P is the transition probability function, with $P(s'|s, a)$ being the probability of transition into the state s' starting from s and taking action a; and ρ_0 is the starting state distribution.

A policy is a rule set/network used by an agent to decide which action to take, as a stochastic policy can be denoted as $a_t \sim \pi_\theta(\cdot|s_t)$, where θ is the parameter set of the policy. We can define the expected return of stochastic policy as (1), where ε is the discounted factor and τ is a sequence of states and actions. The goal of DRL is to find a policy that maximizes the expected return, which can be expressed by (2), where π^* is the optimal policy. Moreover, value functions are always estimated to assess the quality of current state-action pair or state rather than waiting for the long-term result, as in (3) and (4) respectively.

$$J(\pi) = E_{\tau\sim\pi}\left(\sum_t \varepsilon^t r_t\right) \tag{1}$$

$$\pi^* = \arg\max_\pi J(\pi) \tag{2}$$

$$Q^\pi(s, a) = E_{\tau\sim\pi}\left(\sum_{t=1} \varepsilon^t r_t | s_0 = s, a_0 = a\right) \tag{3}$$

$$V^\pi(s) = E_{a\sim\pi}\left(Q^\pi(s, a)\right) \tag{4}$$

To get the optimal policy, policy optimization is one of the commonly used and powerful methods in modern DRL, which optimizes the parameter set θ directly by gradient ascent on the performance objective $J(\pi)$, as in (5).

$$\theta_{k+1} = \theta_k + \nabla_\theta J(\pi_\theta)|_{\theta_k} \tag{5}$$

B. HIERARCHICAL FRAMEWORK WITH TRAFFIC-ADAPTIVE EXTENSION

Although we seek to combine DRL with rule-based models to introduce inductive bias in the paradigm of PIML, the way to design the new architecture remains an open question. Based on the basic idea of HRL, a hierarchical framework composed of DRL and rule-based models is desired. Considering that learning-based techniques specialize in automatically extracting features from the large volumes of multi-dimensional data, DRL should be placed on the higher level to recognize and react to different traffic situations through processing the observed states of environment. Meanwhile, rule-based models have explicit rules, enabling further embedding hard constraints to ensure safety (e.g., responsibility-sensitive safety), and some rule-based models even have collision-free nature (e.g., the well-known safety distance model Gipps' model), while serious variance problem hinders imposing the hard constraints into the DRL algorithm [51] that may lead to collisions. Thus, it is also appropriate to plant rule-based models in the lower level to generate realistic vehicle movements.

Specifically, a hierarchical policy for multi-lane driving is proposed as shown in Fig. 1. It consists of a collection of low-level rule-based decisions including the CF policy and the three-stage LC policy, and a high-level DRL strategy to output the intermediate goal to govern them. Instantaneous desired speed (IDS) is proposed as the intermediate goal to

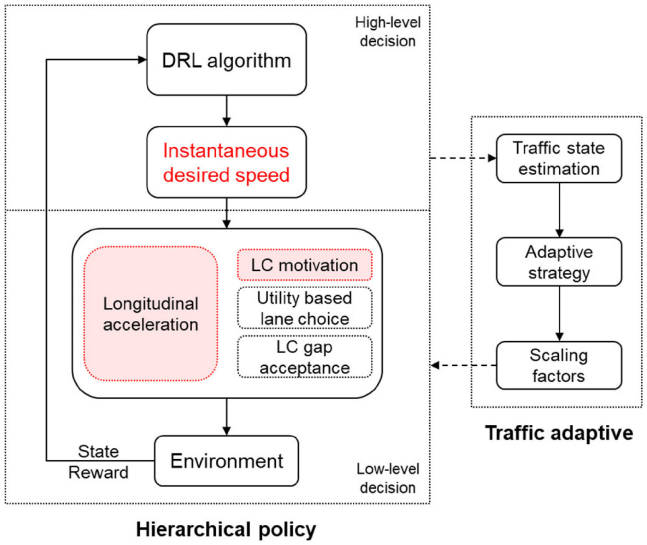

FIGURE 1. The framework of proposed HIDS-DRL with traffic-adaptive extension.

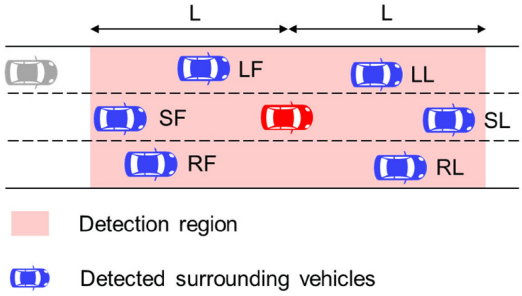

FIGURE 2. Input states with surrounding vehicles.

depict the common motivation of CF and LC. In summary, it is a *Hierarchical framework centered in Instantaneous Desired Speed for multi-lane driving based on DRL* (referred to as HIDS-DRL hereafter).

With the decomposition in a hierarchical structure and the combination of rule-based methods, critical parameters with physical meaning (e.g., IDS) are exposed. A traffic flow adaptive strategy is further designed to consider the traffic flow operations by adjusting the exposed critical parameters for different traffic states. Note that although the TFA strategy is not included in the learning framework, it is an imperative extension of the proposed HIDS-DRL model which is also strongly connected with the model. As mentioned in the Introduction, given the limitation of the existing DRL-based AV studies that the optimization of individual AV performance and traffic flow are separately studied and sometimes contradicted with each other, one main purpose of this study is to develop an efficient model capable of jointly considering the individually optimal driving performance and traffic flow optimization. This is feasible with the exposed critical parameters in the rule-based CF and LC models due to the well-designed hierarchical structure of the HIDS-DRL while other DRL models are hard to be adaptive.

Overall, by combining DRL and domain knowledge of traffic flow theory, i.e., a simple DRL agent with intuitive CF and LC models and a TFA extension, we demonstrate that the HIDS-DRL model can achieve competitive individual performance and optimized system performance simultaneously (see details in Sections IV and V).

C. HIERARCHICAL MODEL FOR MULTI-LANE DRIVING

1) HIGH-LEVEL MOTIVATION WITH IDS

Desired speed assumes that each driver has a desired driving speed, and the driver seeks to minimize the speed difference between the actual speed and the desired speed. It has been widely adopted in many well-known CF models and their variants, such as the IDM and the OVM. However, conventional desired speed in IDM or the parameters for calculating desired speed in other CF models are generally fixed, e.g., desired time headway. Hence it is not capable of adapting to different traffic situations. In terms of the optimal velocity in the class of OVM models, it is usually a function of spacing headway and changing constantly. Yet the optimal velocity is only valid for CF behavior and not comparable with other speeds, since it only takes the spacing headway relative to the leading vehicle in the subject lane into consideration.

Generally, pursuing higher speed is the motivation of both CF and LC behavior. We thus propose the IDS, which stands for the varying desired speed at each time step for acceleration and lane change motivations considering the joint impacts of surrounding vehicles. Longitudinal acceleration is generated to diminish the deviation from the IDS. Comparing to the speeds in adjacent lanes, LC is triggered when exceeding IDS. Despite the assumption of IDS to activate both CF and LC behaviors, the ideal IDS can be depicted by DRL with such CF and LC models combined in the training loop, as in the hierarchical policy of Fig. 1.

The intermediate goal IDS v_t^* at the time t is generated directly by DRL algorithm, i.e., $a_t = v_t^*$. The state of the DRL agent is defined as the gap and relative speed of surrounding vehicles and its own speed in a typical multi-lane expressway. Specifically, we identify the closest preceding and following vehicles in the current lane and adjacent lanes within a certain range L as the surrounding vehicles, as shown in Fig. 2. The state s_t of AV at time t is formulated as: $s_t = (g_t^{\mathrm{LL}}, \Delta v_t^{\mathrm{LL}}, g_t^{\mathrm{LF}}, \Delta v_t^{\mathrm{LF}}, g_t^{\mathrm{RL}}, \Delta v_t^{\mathrm{RL}}, g_t^{\mathrm{RF}}, \Delta v_t^{\mathrm{RF}}, g_t^{\mathrm{SL}}, \Delta v_t^{\mathrm{SL}}, g_t^{\mathrm{SF}}, \Delta v_t^{\mathrm{SF}}, v_t)$, where g_t^i and Δv_t^i denote the bumper-to-bumper gap and the speed difference between the surrounding vehicle i and the subject AV, respectively, as $i \in \{\mathrm{LL}, \mathrm{LF}, \mathrm{RL}, \mathrm{RF}, \mathrm{SL}, \mathrm{SF}\}$ represents the leading vehicle and following vehicle in left adjacent, right adjacent, and subject lane respectively (for example, *LL* denotes the leading vehicle in the left adjacent lane); v_t is the speed of the subject AV. Note that the state which feeds into the DRL agent is produced by the interaction between the final action of low-level policies and the environment, where the final actions encompass continuous longitudinal acceleration and discrete lane-changing decision. Overall, the high-level DRL takes the attributes of surrounding vehicles as input

states and outputs the IDS, which is passed to the low-level controller.

2) LOW-LEVEL DECISION WITH RULE-BASED POLICIES

With IDS, the final actions can be obtained from low-level policies, which are composed of a rule-based CF model and a three-stage LC model tailored for IDS. Specifically, the final actions of the AV agent are ($\dot{v}$, lcd), where $\dot{v}$ denotes AV's longitudinal acceleration and lcd denotes AV's lane-changing decision from the lateral decision set {left, keep, right}.

CF policy: Inspired by IDM and OVM, an intuitive asymmetric CF policy based on IDS is developed, which responds to IDS actively and obtains corresponding acceleration with the consideration of dynamic constraint and comfort, as its mathematical form is:

$$\dot{v}_t^i = k_i\left(1 - \left(\frac{v_t}{v_t^*}\right)^{\delta_i}\right) \tag{6}$$

where:

$\dot{v}_t$, v_t are acceleration and current speed of the subject vehicle at the time t, respectively;

v_t^* is the IDS at the time t;

δ is the exponent that reflects the degree of response to the IDS;

i indicates acceleration ($i = a$) or deceleration ($i = b$);

k is the acceleration multiplier, as k_a refers to the maximum acceleration and k_b refers to the comfortable braking deceleration.

Intuitively, the potential increment/decrease of speed is $v^* - v$, which could be normalized by v^* as $(v^* - v)/v^* = 1 - v/v^*$. The varying reaction exponent δ is considered for acceleration and deceleration, respectively. While the feasible range of acceleration and deceleration are different, k_i is added as the acceleration multiplier.

LC policy: The whole LC decision process can be divided into three steps, i.e., motivation generation, selection of target lane and gap acceptance [52], [53]. As mentioned before, the motivation of LC is to pursue higher speed as well. Since IDS can be regarded as a potential target speed that the subject vehicle will get soon in the subject lane, while the average speed of neighboring vehicles is stable in a short term. Therefore, LC is triggered when the average speed of vehicles on the adjacent lanes exceeds the IDS as in (7).

$$\bar{v}_{adj} > \gamma v_t^* \tag{7}$$

where:

$\bar{v}_{adj}$ is the average speed of vehicles in the adjacent lane, e.g., $\bar{v}_{adj} = \frac{v^{\mathrm{LL}} + v^{\mathrm{LF}}}{2}$;

γ is the threshold factor for LC motivation, $\gamma = 1$ by default;

v_t^* is the IDS at the time t.

Additionally, lane choice will be made if the average speeds of both adjacent lanes are larger than the subject vehicle's IDS. Based on the lane choice model with utility theory proposed in [54], we develop a simple lane choice policy considering the gap and the speed:

$$U = c_v \times \frac{\bar{v}_{adj}}{v_{\mathrm{coef}}} + c_g \times \frac{g^L}{g_{\mathrm{coef}}} \tag{8}$$

where:

U is the total utility of the adjacent lane;

$\bar{v}_{adj}$ is the average speed of vehicles in the adjacent lane, e.g., $\bar{v}_{adj} = \frac{v^{\mathrm{LL}} + v^{\mathrm{LF}}}{2}$;

g^L is the gap with the preceding vehicle in the adjacent lane;

c_v, c_g are the weights of speed and gap;

v_{coef}, g_{coef} are the coefficients to normalize speed and gap.

After selecting the lane with a larger utility, the gap on that lane should be verified to make a safe LC. We consider 1s as the safety time headway threshold, which means LC will be executed only if the time gap from the subject vehicle to the vehicle ahead and behind exceed 1s [55]. As we mainly focus on decision making in this paper, LC execution is simplified to be a lateral movement with a fixed duration [56].

D. TRAFFIC FLOW ADAPTIVE STRATEGY

To take traffic system operation into consideration, we incorporate a heuristic method of TFA strategy in [57] to extend the capability of the HIDS-DRL model for traffic system optimization.

Specifically, the principles of TFA strategy based on the individual vehicle are as follows: 1) Decelerate slowly and depress lane changing when approaching jam, to avoid forming backward shockwave and propagating it laterally; 2) Keep an agile driving style when leaving congested traffic or entering a bottleneck, to reduce the possibility of capacity drop and thereby increase the throughput. Therefore, harnessing the exposed parameters in the HIDS-DRL model, the TFA strategies can be applied by scaling those parameters to adjust the driving behavior.

To implement the TFA strategy, the real-time local traffic state needs to be estimated firstly and various driving styles based on traffic states can then be applied. As per [58], the exponential moving average (EMA) of speed $v_{\mathrm{EMA}}(t)$ is adopted to smooth the short-term fluctuations of the subject vehicle's speed $v(t)$, so that different local traffic states can be identified with the average velocity, with the formulation of $v_{\mathrm{EMA}}(t)$ and its update with relaxation time ξ shown in (9) and (10), respectively. It is approaching a jam when $v(t) - v_{\mathrm{EMA}}(t) < -v_{\mathrm{th}}$, and leaving a jam when $v(t) - v_{\mathrm{EMA}}(t) > v_{\mathrm{th}}$, where the threshold $v_{\mathrm{th}} = 10\mathrm{km/h}$ and $\xi = 5\mathrm{s}$ [58].

$$v_{\mathrm{EMA}}(t) = \frac{1}{\tau}\int_{-\infty}^{t} e^{-(t-t')/\xi} v(t')dt' \tag{9}$$

$$\frac{d}{\mathrm{d}t} v_{\mathrm{EMA}}(t) = \frac{v(t) - v_{\mathrm{EMA}}(t)}{\xi} \tag{10}$$

For the identification of entering a bottleneck, position of the bottleneck (x_s, x_e) can be obtained from digital map. Thus, it is in the bottleneck when the position of the subject vehicle $x(t)$ fulfills the spatial criteria that $x_s < x(t) < x_e$. The driving strategies are encoded by critical parameters k_a

TABLE 1. Traffic flow adaptive driving strategies.

Situation	λ_{k_a}	λ_{k_b}	λ_{v^*}	λ_γ	Driving policy
Free flow	1	1	1	1	Default
Approaching a jam	1	0.7	1	1.5	Decelerate slowly and reduce LC
Entering the bottleneck	1.5	1	1.2	1	Breakdown prevention
Leaving a jam	2	1	1	1	Increase speed rapidly
Congested traffic	1	1	1	1	Default

(increasing value denotes increasing agility/responsiveness), k_b (increasing value denotes increasing aggressiveness in brake), v^* (increasing value denotes increasing aggressiveness in CF and a more conservative style in LC), and γ (increasing value denotes a conservative LC style). For each state, the driving policies are parameterized complying with the TFA driving strategies, as presented in Table 1. The multipliers λ_a, λ_b, λ_{v^*} and λ_γ of the parameters k_a, k_b, v^* and γ, respectively, apply to each situation (e.g., new deceleration multiplier $k_b{}'$ equals to k_b multiplied by λ_{k_b}, namely $k_b{}' = \lambda_{k_b} \times k_b$). For example, when approaching a jam, the strategy reduces the comfortable deceleration to 70% and makes the threshold factor for LC 1.5 times to conduct gently decelerating to stop and reduce LC, aiming to dampen the backward forming shockwave in multiple lanes. In particular, since the focus of this study lies in the feasibility of the HIDS-DRL model to improve traffic flow with traffic-adaptive strategies, the specific values of parameters are determined empirically as an example and not necessary to be optimal.

IV. TRAINING AND EVALUATION OF HIDS-DRL MODEL

A. MODEL TRAINING SETUP

Training environment: Following the setup in previous studies [59], we build a three-lane expressway with a length of 1km in a commonly-used microscopic traffic simulator VISSIM. Referring to the code for design of urban road engineering, for each episode the traffic flow is randomly set within the range of low and middle level of service (LOS), to produce more realistic CF and LC behaviors, since lower flow may lead to more free-flow driving and heavier flow may result in a traffic jam. The simulation resolution is 0.1s and the specific AV enters the road after 100s of warm up time. A new episode will begin if a collision occurs or the AV leaves the road.

DRL algorithm: In general, any model-free DRL approach can be used to train the high-level DRL. In this paper, we adopt the policy optimization method based on trust region (TRPO), which has been proved reliable for both discrete and continuous missions [60], [61], by bounding the size of the policy update and the changes in state distributions that guarantees improvements in policy.

Reward function: In this study, various rewards are determined for different conditions. Specifically, we penalize risky time headway (less than safe time headway T_{safe}) with a value of -1 and further penalize dangerous spacing (less than minimum safety distance) with a value of -10 for safety. Regarding driving efficiency, we reward AV's high speed with a value of normalized speed value and penalize its deceleration out of CF range with a negative value of normalized gap. As for comfort, negative normalized reward is given when AV exceeds perceptible jerk J_p(1.5 m/s^3) and a reward of -1 is given when AV exceeds acceptable comfort jerk J_a(5 m/s^3). The reward for each condition is generally bounded between -1 and 1 except in the worst case that spacing is less than minimum safety distance, as the complete reward function is listed in Table 2.

TABLE 2. Rewards for different conditions.

Definition	Condition	Reward
$\frac{g^{SL}}{v} < T_{safe}$	Time headway is less than safe time headway	-1
$g^{SL} < \frac{v^2}{2b_{max}} - \frac{(v^{SL})^2}{2b_{max}}$	Spacing is less than minimum safety distance	-10
v	Indicating efficiency	$\frac{v}{v_{coef}}$
$\dot{v} < 0$ and $\left(\frac{g^{SL}}{v} > T_{CF}\ or\ g^{SL} > s_{CF}\right)$	Out of CF range	$-\frac{g^{SL}}{g_{coef}}$
$\lvert j\rvert > J_p$	Exceeding perceptible jerk	$max\left(-\frac{\lvert j\rvert}{20}, -1\right)$
$\lvert j\rvert > J_a$	Exceeding acceptable comfort jerk	-1

TABLE 3. Model parameters setting.

High-level DRL with IDS	**Hidden layers**	**Hidden neuron**	L (m)	γ
	2	400	100	1
Low-level policies	k_a (m/s^2)	k_b (m/s^2)	δ_a	δ_b
	1.4	2	4	0.5
	c_v	c_g	v_{coef} (m/s)	g_{coef} (m)
	1	1	15	150
TRPO and Reward	**Total time steps**	σ	b_{max} (m/s^2)	T_{safe} (s)
	2000000	0.002	8	1
	T_{CF} (s)	s_{CF} (m)	J_p (m/s^3)	J_a (m/s^3)
	5	60	1.5	5

Parameters: All parameters used in the model are summarized in Table 3.

B. MODEL TRAINING RESULTS

The total reward per episode [62] and the entropy of policy [63] that were commonly used in previous studies are adopted to evaluate the convergence and stability of model training. As shown in Fig. 3, the average total reward for each episode first increases and tends to converge after about 1,000 episodes. The policy entropy, which is a measure of uncertainty in information theory and can be used as an assessment for the chaos of probability distribution, gradually

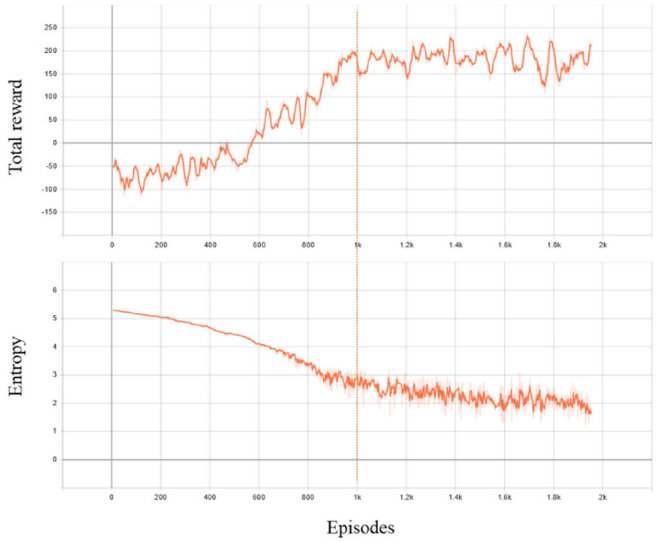

FIGURE 3. Changing of total reward and entropy during training.

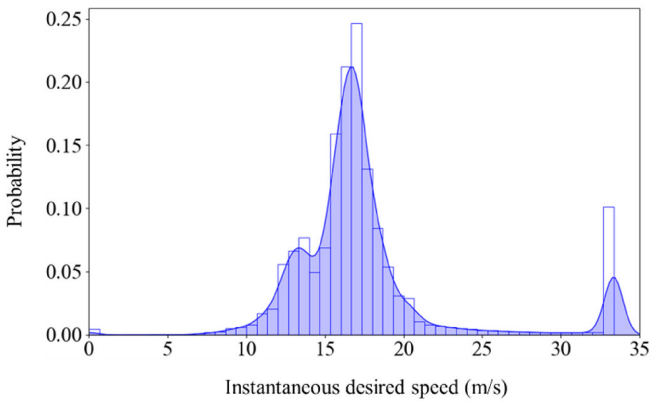

FIGURE 4. The distribution of instantaneous desired speed.

decreases and gets stabilized after about 1,000 episodes. The training results suggest that a strategy with regularity and convergent total reward is learned.

C. RESULTS OF IDS

With a stably convergent DRL model, we investigate the characteristics of the core element IDS. The distribution of IDS is shown in Fig. 4, which indicates a mean value of about 17m/s while the maximum value is 33.3m/s (120km/h) as the predetermined free-flow speed. The distribution of IDS demonstrates the diversity and feasibility of the proposed hierarchical policy.

To further analyze the relationship between the IDS and the influencing factors, the Pearson correlation coefficients between output IDS and all the input states are demonstrated in Fig. 5. The speed difference Δv^{SL}, speed of the subject vehicle v and gap g^{SL} are the most related factors as expected, and following are the attributes of the following vehicle in the subject lane. While the attributes of vehicles in adjacent lanes have smaller correlation coefficients with almost symmetrical impacts.

FIGURE 5. Correlation coefficients between IDS and input states.

D. BASELINE MODELS

To further quantify the model performance for a single AV, we first adopt a pure rule-based model including the widely used CF model IDM and the LC model MOBIL (Minimizing Overall Braking Induced by Lane Changes) [64], referred to as IDM-MOBIL below. The mathematical form of IDM is shown in (11) and (12).

$$\dot{v}_t = \alpha\left(1 - \left(\frac{v_t}{v_0}\right)^4 - \left(\frac{g_t^*}{g_t^{\mathrm{SL}}}\right)^2\right) \tag{11}$$

$$g_t^* = g_0 + v_t T + \frac{v_t \Delta v_t^{\mathrm{SL}}}{2\sqrt{\alpha\beta}} \tag{12}$$

where v_0 is the desired speed, g_t^* is the desired spacing gap, g_0 is the minimum gap at the standstill situation, T is the desired time gap, α is the maximum acceleration and β is the comfortable deceleration. Other variables have the same meanings as those introduced before. The parameters of IDM are calibrated using the reconstructed NGSIM dataset with the recommended Normalized Root Mean Squared Error (NRMSE) adopted as the objective function [65]. The calibrated parameter values are: $v_0 = 20.9m/s$, $T = 1.37s$, $\alpha = 0.97m/s^2$, $\beta = 1.85m/s^2$, and $g_0 = 2.14m$.

For MOBIL model, a LC is motivated when (13) and (14) are satisfied. Remaining LC steps including lane choice, gap acceptance and LC execution are the same with the proposed model in this paper.

$$\widetilde{a}_{\mathrm{new}} \geq -b_{\max} \tag{13}$$

$$\widetilde{a} - a + p(\widetilde{a}_{\mathrm{new}} - a_{\mathrm{new}} + \widetilde{a}_{\mathrm{old}} - a_{\mathrm{old}}) > \Delta a_{\mathrm{th}} \tag{14}$$

where $\widetilde{a}_j$, a_j are the acceleration of the vehicle j after LC and before LC, and $j = new$ means the new follower in the target lane while $j = old$ means the old follower in the subject lane; $\widetilde{a}$, a are the acceleration of the subject vehicle after LC and before LC. All the accelerations after LC are predicted with the above calibrated IDM. We use typical values for other parameters for realistic LC behavior as per [66], i.e., maximum deceleration $b_{\max} = 8m/s^2$, politeness factor $p = 0.5$ and threshold $\Delta a_{\mathrm{th}} = 0.2m/s^2$. Thus, the calibrated IDM-MOBIL represents human drivers' behavior.

Additionally, a pure DRL model for multi-lane driving [59] is also chosen as the baseline model for comparison (referred to as Pure-DRL below). Using the identical input states of surrounding vehicles as in this paper, it is also trained with the same environment and similar reward setting to achieve safety, high efficiency and driving comfort, whereas the DRL algorithm outputs the final decision of longitudinal acceleration and lateral LC decision directly.

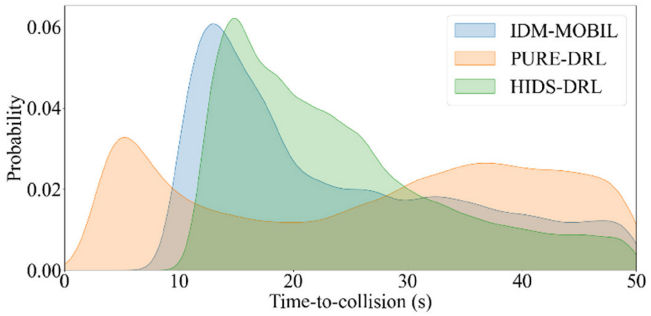

FIGURE 6. TTC distribution for three models.

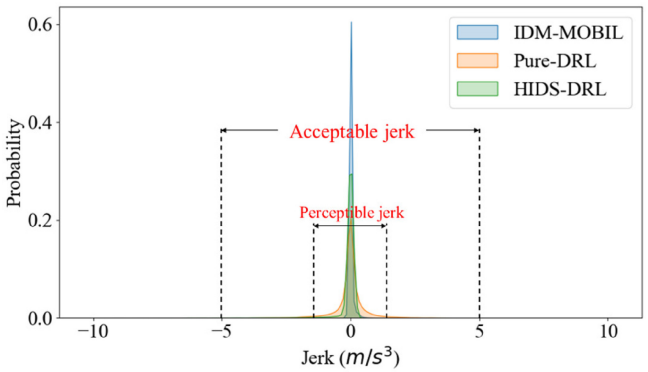

FIGURE 7. Jerk distribution for three models.

It is also noted that other parameters such as maximum acceleration of Pure-DRL are consistent with HIDS-DRL to ensure a fair and reliable comparison, while the parameters of the IDM-MOBIL are calibrated with human-driving data and as a whole works as the baseline model of human-like AV.

E. PERFORMANCE OF HIDS-DRL MODEL

All the models are evaluated with the simulation results in the training environment for 500 episodes for three aspects, i.e., safety, comfort, and efficiency. We first assess the safety risk with time-to-collision (TTC). For the sake of clarity, only TTC values between 0 and 50 s are included as shown in Fig. 6. The proposed HIDS-DRL model and the IDM-MOBIL model have satisfied safety performance while the HIDS-DRL model has even larger TTCs, most of which are larger than 10 s. Although no collision occurs for all three models, the Pure-DRL model shows worst safety performance as many small TTCs (0∼10s) arise. Given that higher TTC values correspond to lower crash risks, the results of TTC show that our proposed HIDS-DRL model achieves more safety than Pure-DRL and is comparable to IDM-MOBIL.

Comfort is evaluated by jerk in this study as shown in Fig. 7. The jerks of all three models achieve the acceptable ride comfort. Only 0.2% of the jerks in IDM-MOBIL exceed the perceptible value, with 5.7% for Pure-DRL and 2.8% for HIDS-DRL. The results of jerk indicate that the HIDS-DRL model can maintain more comfortable than the Pure-DRL model with the combination of rule-based methods

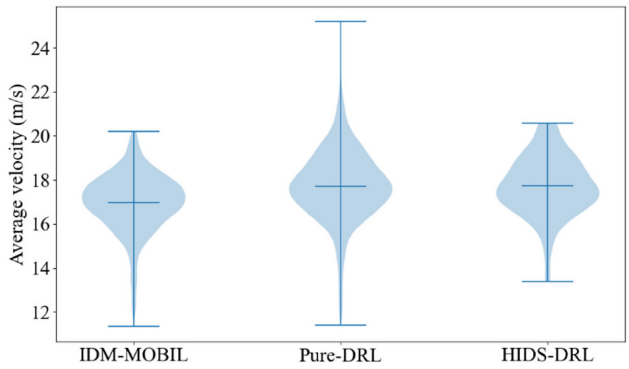

FIGURE 8. Average velocity distribution for three models.

but just below the pure rule-based IDM-MOBIL model. The higher jerk of DRL-based models compared to the IDM-MOBIL model may be due to the fact that the policy of these models is neural network, which is highly nonlinear, non-differentiable and complicated. Consequently, the output of IDS and the acceleration calculated based on the IDS in HIDS-DRL or the acceleration directly output from the Pure-DRL is less smooth than analytic models, which leads to slightly higher jerk.

For efficiency, the distributions of average velocity for simulation results in each episode are presented in Fig. 8 in terms different models. Fig. 8 shows that the behavior with the Pure-DRL model is more aggressive with longer range of velocity, while the IDM-MOBIL model obtains more low velocities. The HIDS-DRL model has the largest overall average velocity of 17.73 m/s, which is 4.5% higher than the IDM-MOBIL model (16.87m/s) and also slightly higher than the Pure-DRL (17.71 m/s). ANOVA analysis also reveals that the differences in average velocity between the HIDS-DRL and IDM-MOBIL models is statistically significant ($P = 0.000$) while that between the HIDS-DRL and Pure-DRL models is statistically equal ($P = 0.895$). Therefore, HIDS-DRL is considered as efficient as Pure-DRL and better than IDM-MOBIL. In addition, we calculate the average LC frequency for the three models, where the result of HIDS-DRL is 0.79 veh/km/ln, which is higher than 0.20 veh/km/ln for IDM-MOBIL but lower than 3.62 veh/km/ln for Pure-DRL. It implies that LC decisions generated by HIDS-DRL are more effective than Pure-DRL, owing to the clear LC motivation activated by IDS.

Moreover, with the distributions of time headway shown in Fig. 9, the average time headway of HIDS-DRL is 1.5 s compared with 2 s for IDM-MOBIL, which is expected to be more conservative as the baseline human-like model, and 1.2 s for Pure-DRL. This further validates the performance of three models in terms of efficiency.

In a nutshell, the HIDS-DRL learns satisfying driving strategies with the hierarchical framework integrating rule-based model with DRL model. It is safer and more comfortable than the Pure-DRL model while maintaining efficiency as the Pure-DRL model other than conservative IDM-MOBIL.

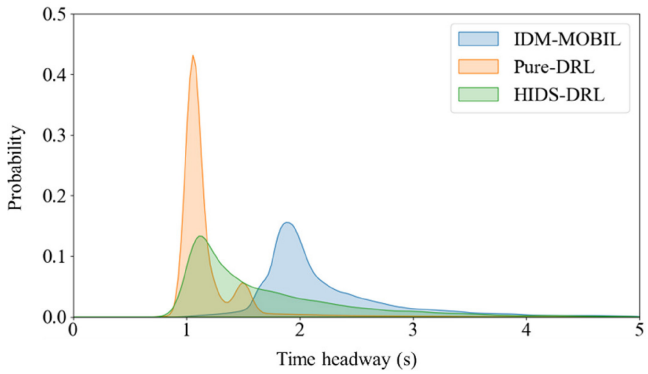

FIGURE 9. Time headway distribution for three models.

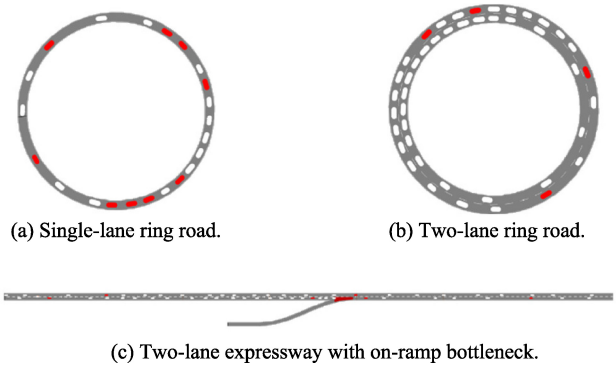

FIGURE 10. Test scenarios for system performance evaluation.

TABLE 4. Baseline methods for traffic system evaluation.

No	Model	Traffic adaptive strategy
1	IDM-MOBIL	None
2	IDM-MOBIL	Adjusting parameters k_a, k_b, T as in [58]
3	Pure-DRL	None
4	Pure-DRL	Adjusting multiplier for acceleration
5	HIDS-DRL	None
6	HIDS-DRL	Adjusting parameters k_a, k_b, v^*, γ

V. TRAFFIC FLOW OPTIMIZATION USING HIDS-DRL

As some studies have shown AVs impose possible negative impacts on the traffic system [21], [22], [23], how to alleviate the potential negative system impacts through the user-oriented driving model remains unresolved, since centralized control is not applicable at least in a near future while AVs are equipped with varying algorithms and belong to different companies. We thus extend the HIDS-DRL model with a traffic flow adaptive strategy to further accommodate and optimize the traffic flow. Three typical test scenarios are built within VISSIM, and the traffic operation is evaluated with various strategies as shown in TABLE 4. It is noted that the traffic adaptive strategy for the proposed DRL model is the same as in the TABLE 1, while it is reduced to adjust only the outputted acceleration with multiplier factor for Pure-DRL as there is no more exposed parameters, which is also one motivation of this work to accommodate adaptability and flexibility in DRL. For the IDM-MOBIL model, we implement the adaptive strategy as in [58] because of the validated performance and for consistency. It is remarked that the common factors $(\lambda_{k_a}, \lambda_{k_b})$ are identical for all the models.

A. TYPICAL TRAFFIC SCENARIOS

We develop three typical traffic scenarios, i.e., single-lane ring road, two-lane ring road and on-ramp bottleneck, to evaluate the performance of the models in optimizing traffic flow. The single-lane ring road is usually used to explore the stability of CF models. As early as 2008, Japanese scholars have conducted on-site human driving experiments, which verify that a single-lane road with a certain density will produce phantom traffic jams [67]. A ring with a circumference of 230m was built, with 22 vehicles placed on the ring road and drivers instructed to cruise at 30km/h. The results suggest that the phenomenon of stop-and-go emerges due to the reaction time of human drivers and vehicles' limited accelerating characteristics. Thus, we choose the single-lane ring road to assess the CF behavior. Furthermore, the single-lane ring road is extended to double lanes to testify the CF and LC decisions with doubled number of vehicles.

Although ring road is a simple closed system for evaluation, it is not realistic overall. To attain a more accurate assessment of the real-world traffic flow, an open road network with a bottleneck is built. Particularly, this paper utilizes a two-lane expressway of 1 km with an on-ramp bottleneck as the open testing system. AVs are placed on the main road to test whether the CF and LC behavior of AVs can eliminate bottleneck congestion. As per [68] and the definition of LOS, the input flow of the two-lanes expressway is set to 3500 pcu/h, and the single-lane on-ramp is 200 pcu/h. The first-in-first-out (FIFO) principle is applied in the conflict area, which means the priority of leaving is determined by the order of arriving, to accelerate the occurrence of congestion.

All the three scenarios are implemented in VISSIM, as shown in Fig. 10. By randomly selecting a certain number of AVs in each episode (the proportion of AV in the mixed traffic is fixed in a relative low penetration rate 5%), the average system performance is assessed through repeated experiments where for each experiment the data is collected from stable periods (100∼600s for the ring road as closed systems can achieve quick convergence, and 100∼3700s for the on-ramp bottleneck).

With respect to the operation efficiency of the whole traffic system, we use the average cumulative delay per vehicle as evaluation indicator. The cumulative delay is defined as the difference between the actual travel time and the ideal travel time [69], where the ideal travel time is the potential minimum travel time given the speed limits of road links (30 km/h for the ring road and double-lane ring road, and 50 km/h for the on-ramp bottleneck in the simulations) regardless of vehicles' desired speeds.

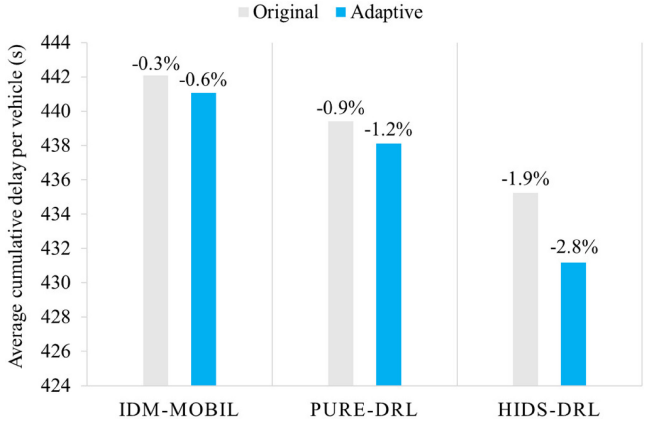

FIGURE 11. Comparison for system performance at single-lane ring road.

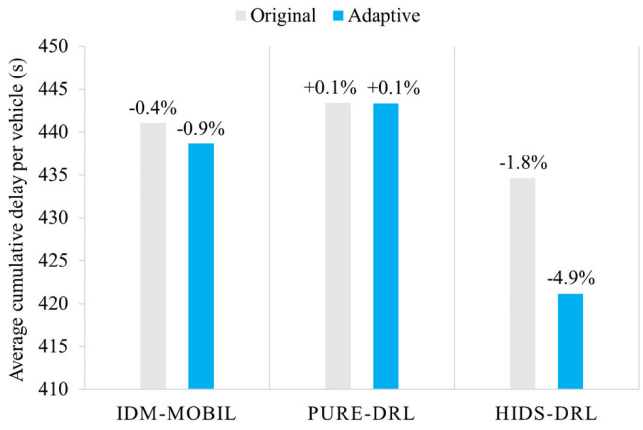

FIGURE 12. Comparison for system performance at double-lane ring road.

B. EVALUATION RESULTS

The model performance of traffic system optimization is evaluated for all the baseline methods in Table 4, and the percentage change relative to the pure human-driving vehicle (HDV) system is calculated for better comparison. Figs. 11-13 are average cumulative delays and corresponding percentage changes at single-lane ring road, double-lane ring road and on-ramp bottleneck, respectively.

For single-lane ring road in Fig. 11, the proposed HIDS-DRL has the best performance (−1.9%), following by Pure-DRL (−0.9%) and IDM-MOBIL (−0.3%). With traffic adaptive strategy, the cumulative delay further drops for all the models but achieves the largest fall (−0.9%) for HIDS-DRL.

As for the double-lane ring road scenario, as shown in Fig. 12, what stands out is that the Pure-DRL slightly deteriorates the operation efficiency of traffic system even with traffic adaptive strategy. The HIDS-DRL still performs best (−1.8%) while IDM-MOBIL almost remains the same (−0.4%) with the HDV system. Traffic adaptive strategy significantly improves the HIDS-DRL (from −1.8% to −4.9%), and little enhancement is obtained for IDM-MOBIL with traffic adaptive strategy.

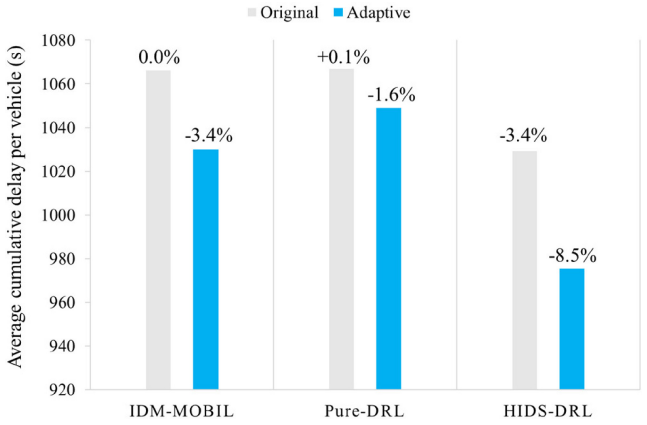

FIGURE 13. Comparison for system performance at on-ramp bottleneck.

At on-ramp bottleneck, as shown in Fig. 13, the HIDS-DRL shows the largest fall in cumulative delay, while there is nearly no difference with IDM-MOBIL and even worse with Pure-DRL. Extended by TFA strategy, the HIDS-DRL obtains 8.5% decrease in delay with only 5% of AVs, following by IDM-MOBIL with 3.4% decrease (noted that the improvement performance of IDM-MOBIL in our study and the work in [58] is not comparable mainly because the simulation settings of on-ramp scenarios in two studies are considerably different in terms of road layout, traffic demand, traffic composition (trucks were included in [58]).) and Pure-DRL with 1.6% decrease. It is also worth noting that without the TFA strategy, the HIDS-DRL model can still achieve fairly good performance in optimizing the whole traffic system (equivalent to the IDM-MOBIL with TFA strategy), which demonstrate its inherent capability of both individual behavior and traffic flow optimization.

We also would like to point out that the experiments are based on a typical set of TFA strategies, and the optimization performance may vary for different parameters. Without loss of generality, we have tested another set of parameters in TFA strategy at on-ramp bottleneck, where the HIDS-DRL model still achieves the largest fall of 7.5% in average cumulative delay relative to HDV, compared to 3% and 2.6% decrease for the Pure-DRL and IDM-MOBIL model respectively. It further demonstrates the superiority of our proposes model. However, it is worth noting that the Pure-DRL model outperforms the IDM-MOBIL model with the new parameter set, which partially indicates the significance of finding optimal parameters when comparing TFA strategies with different driving models (this is beyond the scope of this work and can be explored in future research).

C. DISCUSSION

In summary, no significant reduction in cumulative delay is found with the IDM-MOBIL model comparing to the HDV system. This is expected because the IDM model is calibrated with human-driving data and the parameters of MOBIL are chosen for a realistic behavior. Thus, the

IDM-MOBIL model behaves like a human. Regarding the Pure-DRL model, it improves the single-lane ring road traffic but negatively impacts the double-lane ring road and on-ramp bottleneck. The more congested traffic in double-lane ring road than on-ramp bottleneck leads to the deterioration results in double-lane ring road even with traffic adaptive strategy. The performance of multi-lane driving underlines the generalization problem of the Pure-DRL model in the LC decision. It validates the aggressive CF behavior and some improper LC of the Pure-DRL again. With the HIDS-DRL model, the decision making of LC is built upon IDS to pursue higher speed, as such transparent and clear logic of LC results in a more robust DRL model. Meanwhile, the IDS generated from the high-level DRL algorithm exploits the learning capability of DRL to understand the observations better, rather than manually construct rules as the pure rule-based model IDM-MOBIL. All the results justify that applying the HIDS-DRL model leads to more efficient traffic flow than other models.

Furthermore, by incorporating the TFA strategy, the traffic flow efficiency is further enhanced for all the models, which proves that the traffic adaptive strategy is applicable for DRL-based models through tunning critical parameters. Additionally, more progress is achieved in HIDS-DRL with the TFA strategy. The reason for effective extension in HIDS-DRL is twofold. Firstly, the hierarchical framework combined rule-based policies enable the exposure of more critical parameters for further adjustments. In contrast, only multipliers of acceleration can be adjusted for the Pure-DRL model which is a common condition for other DRL models. Secondly, the LC behavior is integrated in the framework aiming to achieve higher speed given the surrounding traffic environment. The explicit LC motivation and the critical parameters among it with physical meanings such as the threshold factor for LC motivation γ, make it easy to adapt the LC behavior for traffic flow optimization. Instead, the Pure-DRL model only obtains the final LC decision and only fully compliance or rejection is allowed for LC behavior, which is not flexible and effective. Given that the original traffic adaptive strategy for IDM-MOBIL model only works for the CF module, it is expected that the TFA strategy provides more room for improvement with HIDS-DRL than IDM-MOBIL. Overall, the proposed novel hierarchical framework is credited for the success application of traffic adaptive strategy.

VI. CONCLUSION

In this study, we developed a novel multi-lane autonomous driving decision-making framework combining the DRL with rule-based policies in a hierarchy, where the IDS is proposed to activate CF and LC behaviors integrally. Moreover, the TFA driving strategy is incorporated to take into consideration the traffic flow operation. We then evaluated the proposed models in three typical scenarios. The main conclusions of this paper are as follows.

1) With the decomposition and the IDS defined to integrate CF and LC for the pursuit of higher speed, the HIDS-DRL model can be learned efficiently and interpretably for multi-lane driving, to benefit from both DRL and rule-based methods.

2) Compared to Pure-DRL model and rule-based model IDM-MOBIL, the proposed HIDS-DRL model achieves higher speed with the premise of safety and comfort, verifying its ability to handle multi-lane driving with satisfying efficiency.

3) Applying the traffic adaptive strategy via exposed critical parameters, the HIDS-DRL model further improves the traffic efficiency, with a maximum of 8.5% decrease in cumulative delay with only 5% AV penetration compared to other baseline models. The results validate that the proposed framework is effective in both individual behavior and traffic flow optimization.

REFERENCES

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020, doi: 10.1109/ACCESS.2020.2983149.

[2] Z. Zheng, "Recent developments and research needs in modeling lane changing," *Transp. Res. Part B Methodol.*, vol. 60, pp. 16–32, Feb. 2014, doi: 10.1016/j.trb.2013.11.009.

[3] C. Ziegler and J. Adamy, "Anytime tree-based trajectory planning for urban driving," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 48–57, 2023, doi: 10.1109/OJITS.2023.3235986.

[4] Z. Wang, J. Guo, Z. Hu, H. Zhang, J. Zhang, and J. Pu, "Lane transformer: A high-efficiency trajectory prediction model," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 2–13, 2023, doi: 10.1109/OJITS.2023.3233952.

[5] G. F. Newell, "Nonlinear effects in the dynamics of car following," *Oper. Res.*, vol. 9, no. 2, pp. 209–229, Apr. 1961, doi: 10.1287/opre.9.2.209.

[6] P. G. Gipps, "A behavioural car-following model for computer simulation," *Transp. Res. Part B Methodol.*, vol. 15, no. 2, pp. 105–111, Apr. 1981, doi: 10.1016/0191-2615(81)90037-0.

[7] 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, Aug. 2000, doi: 10.1103/PhysRevE.62.1805.

[8] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama, "Dynamical model of traffic congestion and numerical simulation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 2, pp. 1035–1042, Feb. 1995, doi: 10.1103/PhysRevE.51.1035.

[9] T. Toledo, H. N. Koutsopoulos, and M. Ben-Akiva, "Integrated driving behavior modeling," *Transp. Res. Part C Emerg. Technol.*, vol. 15, no. 2, pp. 96–112, Apr. 2007, doi: 10.1016/j.trc.2007.02.002.

[10] T. Nishi, P. Doshi, and D. Prokhorov, "Merging in congested freeway traffic using multipolicy decision making and passive actor-critic learning," *IEEE Trans. Intell. Veh.*, vol. 4, no. 2, pp. 287–297, Jun. 2019, doi: 10.1109/TIV.2019.2904417.

[11] P. Wang, C.-Y. Chan, and A. de La Fortelle, "A reinforcement learning based approach for automated lane change maneuvers," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Jun. 2018, pp. 1379–1384, doi: 10.1109/IVS.2018.8500556.

[12] D. F. Xie, Z. Z. Fang, B. Jia, and Z. He, "A data-driven lane-changing model based on deep learning," *Transp. Res. Part C Emerg. Technol.*, vol. 106, pp. 41–60, Sep. 2019, doi: 10.1016/J.TRC.2019.07.002.

[13] X. Zhang, J. Sun, X. Qi, and J. Sun, "Simultaneous modeling of car-following and lane-changing behaviors using deep learning," *Transp. Res. Part C Emerg. Technol.*, vol. 104, pp. 287–304, Jul. 2019, doi: 10.1016/j.trc.2019.05.021.

[14] H. Zheng et al., "Learning-based safe control for robot and autonomous vehicle using efficient safety certificate," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 419–430, 2023, doi: 10.1109/OJITS.2023.3280573.

[15] N. A. Spielberg, M. Templer, J. Subosits, and J. C. Gerdes, "Learning policies for automated racing using vehicle model gradients," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 130–142, 2023, doi: 10.1109/OJITS.2023.3237977.

[16] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998, doi: 10.1109/TNN.1998.712192.

[17] N. Parvez Farazi, B. Zou, T. Ahamed, and L. Barua, "Deep reinforcement learning in transportation research: A review," *Transp. Res. Interdiscip. Perspect.*, vol. 11, Sep. 2021, Art. no. 100425, doi: 10.1016/j.trip.2021.100425.

[18] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022, doi: 10.1109/TITS.2021.3054625.

[19] P. de Haan, D. Jayaraman, and S. Levine. "Causal confusion in imitation learning." May 2019. [Online]. Available: http://arxiv.org/abs/1905.11979.

[20] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[21] D. Chen, A. Srivastava, S. Ahn, and T. Li, "Traffic dynamics under speed disturbance in mixed traffic with automated and non-automated vehicles," *Transp. Res. Part C Emerg. Technol.*, vol. 113, pp. 293–313, Apr. 2020, doi: 10.1016/j.trc.2019.03.017.

[22] J. Sun, Z. Zheng, and J. Sun, "Stability analysis methods and their applicability to car-following models in conventional and connected environments," *Transp. Res. Part B Methodol.*, vol. 109, pp. 212–237, Mar. 2018, doi: 10.1016/J.TRB.2018.01.013.

[23] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," *Transp. Res. Part C Emerg. Technol.*, vol. 71, pp. 143–163, Oct. 2016, doi: 10.1016/j.trc.2016.07.007.

[24] M. Zhu, X. Wang, and Y. Wang, "Human-like autonomous car-following model with deep reinforcement learning," *Transp. Res. Part C Emerg. Technol.*, vol. 97, pp. 348–368, Dec. 2018, doi: 10.1016/j.trc.2018.10.024.

[25] A. Alizadeh, M. Moghadam, Y. Bicer, N. K. Ure, U. Yavas, and C. Kurtulus, "Automated lane change decision making using deep reinforcement learning in dynamic and uncertain highway environment," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 1399–1404, doi: 10.1109/ITSC.2019.8917192.

[26] M. Kaushik, V. Prasad, K. M. Krishna, and B. Ravindran, "Overtaking maneuvers in simulated highway driving using deep reinforcement learning," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Jun. 2018, pp. 1885–1890, doi: 10.1109/IVS.2018.8500718.

[27] D. Isele, R. Rahimi, A. Cosgun, K. Subramanian, and K. Fujimura, "Navigating occluded intersections with autonomous vehicles using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2034–2039, doi: 10.1109/ICRA.2018.8461233.

[28] N. Rajesh, Y. Zheng, and B. Shyrokau, "Comfort-oriented motion planning for automated vehicles using deep reinforcement learning," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 348–359, 2023, doi: 10.1109/OJITS.2023.3275275.

[29] C. Wang and B. Coifman, "The effect of lane-change maneuvers on a simplified car-following theory," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 3, pp. 523–535, Sep. 2008, doi: 10.1109/TITS.2008.928265.

[30] Y. Zhang, Q. Lin, J. Wang, S. Verwer, and J. M. Dolan, "Lane-change intention estimation for car-following control in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 276–286, Sep. 2018, doi: 10.1109/TIV.2018.2843178.

[31] A. van den Bosch, B. Hengst, J. Lloyd, R. Miikkulainen, H. Blockeel, and H. Blockeel, "Hierarchical reinforcement learning," in *Encyclopedia of Machine Learning*, Boston, MA, USA: Springer, 2011, pp. 495–502, doi: 10.1007/978-0-387-30164-8_363.

[32] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artif. Intell.*, vol. 112, nos. 1–2, pp. 181–211, Aug. 1999, doi: 10.1016/S0004-3702(99)00052-1.

[33] S. Nasiriany, V. H. Pong, S. Lin, and S. Levine, "Planning with goal-conditioned policies," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Nov. 2019, pp. 14814–14825. Accessed: Oct. 16, 2022. [Online]. Available: http://arxiv.org/abs/1911.08453.

[34] J. Peng, S. Zhang, Y. Zhou, and Z. Li, "An integrated model for autonomous speed and lane change decision-making based on deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21848–21860, Nov. 2022, doi: 10.1109/TITS.2022.3185255.

[35] C. Wu, A. M. Bayen, and A. Mehta, "Stabilizing Traffic with Autonomous Vehicles," in *Proc. IEEE Int. Conf. Rob. Autom. (ICRA)*, May 2018, pp. 1–7, doi: 10.1109/ICRA.2018.8460567.

[36] A. R. Kreidieh, C. Wu, and A. M. Bayen, "Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning," in *Proc. 21st IEEE Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 1475–1480, doi: 10.1109/ITSC.2018.8569485.

[37] C. Wu et al., "Framework for control and deep reinforcement learning in traffic," in *Proc. 20th IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2017, pp. 1–8, doi: 10.1109/ITSC.2017.8317694.

[38] C. Wu, A. Kreidieh, E. Vinitsky, and A. M. Bayen, "Emergent behaviors in mixed-autonomy traffic," in *Proc. Annu. Conf. Robot Learn.*, 2017, pp. 398–407. Accessed: Jul. 23, 2023. [Online]. Available: https://proceedings.mlr.press/v78/wu17a.html

[39] E. Vinitsky, K. Parvate, A. Kreidieh, C. Wu, and A. Bayen, "Lagrangian control through deep-RL: Applications to bottleneck decongestion," in *Proc. 21st IEEE Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 759–765, doi: 10.1109/ITSC.2018.8569615.

[40] C. Wu, A. R. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: A modular learning framework for mixed autonomy traffic," *IEEE Trans. Robot.*, vol. 38, no. 2, pp. 1270–1286, Apr. 2022, doi: 10.1109/TRO.2021.3087314.

[41] G. Wang, J. Hu, Z. Li, and L. Li, "Harmonious lane changing via deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4642–4650, May 2022, doi: 10.1109/TITS.2020.3047129.

[42] X. Qu, Y. Yu, M. Zhou, C.-T. Lin, and X. Wang, "Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: A reinforcement learning based approach," *Appl. Energy*, vol. 257, Jan. 2020, Art. no. 114030, doi: 10.1016/j.apenergy.2019.114030.

[43] H. Shi, Y. Zhou, K. Wu, X. Wang, Y. Lin, and B. Ran, "Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment," *Transp. Res. Part C Emerg. Technol.*, vol. 133, Dec. 2021, Art. no. 103421, doi: 10.1016/j.trc.2021.103421.

[44] H. Shi, D. Chen, N. Zheng, X. Wang, Y. Zhou, and B. Ran, "A deep reinforcement learning based distributed control strategy for connected automated vehicles in mixed traffic platoon," *Transp. Res. Part C Emerg. Technol.*, vol. 148, Mar. 2023, Art. no. 104019, doi: 10.1016/j.trc.2023.104019.

[45] L. Jiang, Y. Xie, N. G. Evans, X. Wen, T. Li, and D. Chen, "Reinforcement Learning based cooperative longitudinal control for reducing traffic oscillations and improving platoon stability," *Transp. Res. Part C Emerg. Technol.*, vol. 141, Art. no. 103744, Aug. 2022, doi: 10.1016/j.trc.2022.103744.

[46] Z. Zheng, S. Ahn, D. Chen, and J. Laval, "Freeway traffic oscillations: Microscopic analysis of formations and propagations using Wavelet Transform," *Transp. Res. Part B Methodol.*, vol. 45, no. 9, pp. 1378–1388, Nov. 2011, doi: 10.1016/J.TRB.2011.05.012.

[47] F. Wu and D. B. Work, "Connections between classical car following models and artificial neural networks," in *Proc. 21st IEEE Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3191–3198, doi: 10.1109/ITSC.2018.8569333.

[48] Z. Mo, R. Shi, and X. Di, "A physics-informed deep learning paradigm for car-following models," *Transp. Res. Part C Emerg. Technol.*, vol. 130, Sep. 2021, Art. no. 103240, doi: 10.1016/j.trc.2021.103240.

[49] Z. Cao, S. Xu, X. Jiao, H. Peng, and D. Yang, "Trustworthy safety improvement for autonomous driving using reinforcement learning," *Transp. Res. Part C Emerg. Technol.*, vol. 138, May 2022, Art. no. 103656, doi: 10.1016/j.trc.2022.103656.

[50] A. Likmeta, A. M. Metelli, A. Tirinzoni, R. Giol, M. Restelli, and D. Romano, "Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving," *Rob. Auton. Syst.*, vol. 131, Sep. 2020, Art. no. 103568, doi: 10.1016/j.robot.2020.103568.

[51] S. Shalev-Shwartz, S. Shammah, and A. Shashua. "Safe, multi-agent, reinforcement learning for autonomous driving." Oct. 2016. Accessed: Jul. 23, 2023. [Online]. Available: http://arxiv.org/abs/1610.03295.
[52] P. Hidas, "Modelling lane changing and merging in microscopic traffic simulation," *Transp. Res. Part C Emerg. Technol.*, vol. 10, nos. 5–6, pp. 351–371, Oct. 2002, doi: 10.1016/S0968-090X(02)00026-8.
[53] J. Sun, J. Ouyang, and J. Yang, "Modeling and analysis of merging behavior at expressway on-ramp bottlenecks," *Transp. Res. Rec.*, vol. 2421, no. 1, pp. 74–81, 2014, doi: 10.3141/2421-09.
[54] T. Toledo, C. F. Choudhury, and M. E. Ben-Akiva, "Lane-changing model with explicit target lane choice," *Transp. Res. Rec.*, vol. 1934, no. 1, pp. 157–165, Jan. 2005, doi: 10.1177/0361198105193400117.
[55] J. Sun, J. Sun, and Z. Li, "Study on traffic characteristics for a typical expressway on-ramp bottleneck considering various merging behaviors," *Physica A Stat. Mech. Appl.*, vol. 440, pp. 57–67, Dec. 2015, doi: 10.1016/j.physa.2015.08.007.
[56] T. Toledo and D. Zohar, "Modeling duration of lane changes," *Transp. Res. Rec.*, vol. 1999, pp. 71–78, Jan. 2007, doi: 10.3141/1999-08.
[57] M. Treiber and A. Kesting, *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin, Germany: Springer, 2013. doi: 10.1007/978-3-642-32460-4.
[58] A. Kesting, M. Treiber, M. Schönhof, and D. Helbing, "Adaptive cruise control design for active congestion avoidance," *Transp. Res. Part C Emerg. Technol.*, vol. 16, no. 6, pp. 668–683, 2008, doi: 10.1016/j.trc.2007.12.004.
[59] Y. Ye, X. Zhang, and J. Sun, "Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment," *Transp. Res. Part C Emerg. Technol.*, vol. 107, pp. 155–170, Oct. 2019, doi: 10.1016/j.trc.2019.08.011.
[60] A. Rajeswaran et al., "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Proc. 14th Conf. Rob. Sci. Syst. XIV*, Jun. 2018. doi: 10.15607/RSS.2018.XIV.049.
[61] A. A. Rusu, M. Vecerik, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell, "Sim-to-real robot learning from pixels with progressive nets." Oct. 2016. Accessed: Jul. 23, 2023. [Online]. Available: http://arxiv.org/abs/1610.04286.
[62] V. Mnih et al. "Playing atari with deep reinforcement learning." Dec. 2013. Accessed: Jul. 23, 2023. [Online]. Available: http://arxiv.org/abs/1312.5602
[63] B. Xin, H. Yu, Y. Qin, Q. Tang, and Z. Zhu, "Exploration entropy for reinforcement learning," *Math. Problem Eng.*, vol. 2020, Jan. 2020, Art. no. 2672537, doi: 10.1155/2020/2672537.
[64] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model MOBIL for car-following models," *Transp. Res. Rec.*, vol. 1999, no. 1, pp. 86–94, Jan. 2007, doi: 10.3141/1999-10.
[65] V. Punzo, Z. Zheng, and M. Montanino, "About calibration of car-following dynamics of automated and human-driven vehicles: Methodology, guidelines and codes," *Transp. Res. Part C Emerg. Technol.*, vol. 128, Jul. 2021, Art. no. 103165, doi: 10.1016/J.TRC.2021.103165.
[66] M. Treiber and A. Kesting, "Modeling lane-changing decisions with MOBIL," in *Traffic and Granular Flow '07*. Berlin, Germany: Springer, 2009, pp. 211–221, doi: 10.1007/978-3-540-77074-9_19.
[67] Y. Sugiyamal et al., "Traffic jams without bottlenecks–experimental evidence for the physical mechanism of the formation of a jam," *New J. Phys.*, vol. 10, no. 3, Mar. 2008, Art. no. 33001, doi: 10.1088/1367-2630/10/3/033001.
[68] M. Treiber and A. Kesting, "Evidence of convective instability in congested traffic flow: A systematic empirical and theoretical investigation," *Transp. Res. Part B Methodol.*, vol. 45, no. 9, pp. 1362–1377, Nov. 2011, doi: 10.1016/j.trb.2011.05.011.
[69] *Highway Capacity Manual 7th Edition*. Washington, DC, USA: Nat. Acad. Press, 2022, doi: 10.17226/26432.

XIAOHUI ZHANG received the B.S. and M.S. degrees in transportation engineering from Tongji University, China, in 2020, where she is currently pursuing the Ph.D. degree with the Department of Traffic Engineering. Her main research interests include connected and automated vehicles, traffic flow theory and modeling, and emerging technology for intelligent traffic system integrated machine learning.

JIE SUN received the Ph.D. degree in transportation engineering from Tongji University in 2019. He is currently working as a Postdoctoral Research Fellow with the School of Civil Engineering, The University of Queensland, Brisbane, Australia. His main research interests include traffic flow theory and modeling, traffic simulation, and connected and automated vehicles.

YUNPENG WANG received the B.Sc., M.Sc., and Ph.D. degrees from Jilin University, Changchun, China, in 1988, 1994, and 1997, respectively. From 1988 to 2008, he served as the Dean of the School of Transportation and the Director of the Science and Technology Department, Jilin University and the Vice President of the Changchun University of Technology. Since 2009, he has been a Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China, where he is currently the President. He has published over 100 research articles. His research interests include intelligent transportation control, cooperative vehicle infrastructure systems, and traffic emergency management systems. He is the Academician of the Chinese Academy of Engineering also a Cheung Kong Scholar Professor, and the Subject Expert of the National High Technology Research and Development Program ("863" Program) of China.

JIAN SUN received the Ph.D. degree from Tongji University in 2006. Subsequently, he was with Tongji University as a Lecturer and then promoted to the position as a Professor in 2011, where he is currently a Professor with the College of Transportation Engineering and the Dean of the Department of Traffic Engineering. His main research interests include traffic flow theory, traffic simulation, connected and automated vehicles, and intelligent transportation systems.