Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

SAE INTERNATIONAL

ARTICLE INFO
Article ID: 10-07-04-0031
© 2023 Nanyang Technological University
doi:10.4271/10-07-04-0031

Robust Multiagent Reinforcement Learning toward Coordinated Decision-Making of Automated Vehicles

Xiangkun He,[1] Hao Chen,[1] and Chen Lv[1]

[1]Nanyang Technological University, School of Mechanical and Aerospace Engineering, Singapore

Abstract

Automated driving is essential for developing and deploying intelligent transportation systems. However, unavoidable sensor noises or perception errors may cause an automated vehicle to adopt suboptimal driving policies or even lead to catastrophic failures. Additionally, the automated driving longitudinal and lateral decision-making behaviors (e.g., driving speed and lane changing decisions) are coupled, that is, when one of them is perturbed by unknown external disturbances, it causes changes or even performance degradation in the other. The presence of both challenges significantly curtails the potential of automated driving. Here, to coordinate the longitudinal and lateral driving decisions of an automated vehicle while ensuring policy robustness against observational uncertainties, we propose a novel robust coordinated decision-making technique via robust multiagent reinforcement learning. Specifically, the automated driving longitudinal and lateral decisions under observational perturbations are modeled as a constrained robust multiagent Markov decision process. Meanwhile, a nonlinear constraint setting with Kullback–Leibler divergence is developed to keep the variation of the driving policy perturbed by stochastic perturbations within bounds. Additionally, a robust multiagent policy optimization approach is proposed to approximate the optimal robust coordinated driving policy. Finally, we evaluate the proposed robust coordinated decision-making method in three highway scenarios with different traffic densities. Quantitatively, in the absence of noises, the proposed method achieves an approximate average enhancement of 25.58% in traffic efficiency and 91.31% in safety compared to all baselines across the three scenarios. In the presence of noises, our technique improves traffic efficiency and safety by an approximate average of 30.81% and 81.02% compared to all baselines in the three scenarios, respectively. The results demonstrate that the proposed approach is capable of improving automated driving performance and ensuring policy robustness against observational uncertainties.

History

Received: 06 May 2023
Revised: 07 Jul 2023
Accepted: 14 Aug 2023
e-Available: 04 Sep 2023

Keywords

Automated vehicle, Coordinated decision-making, Reinforcement learning, Motion control, Vehicle dynamics

Citation

He, X., Chen, H., and Lv, C., "Robust Multiagent Reinforcement Learning toward Coordinated Decision-Making of Automated Vehicles," *SAE Int. J. Veh. Dyn., Stab., and NVH* 7(4):475-488, 2023, doi:10.4271/10-07-04-0031.

ISSN: 2380-2162
e-ISSN: 2380-2170

© 2023 Nanyang Technological University; Published by SAE International. This Open Access article is published under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits distribution, and reproduction in any medium, provided that the original author(s) and the source are credited.

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

Introduction

In the present-day world, automobiles have become an indispensable means of transportation. However, the increased mobility provided by motor vehicles comes at a cost, such as an increase in traffic accidents [1, 2, 3]. Fortunately, with the rise of cutting-edge technologies such as artificial intelligence (AI) [4, 5] or metaverse, automated vehicles are becoming increasingly promising as they are able to shoulder the burden and stress of human drivers, leading to enhanced traffic safety and efficiency [6, 7].

An automated vehicle involves the integration of multidisciplinary knowledge and theories, which is basically composed of sensing, decision-making, and motion control systems [8, 9, 10]. The decision-making system is the brain of automated driving, enabling an automated vehicle to reasonably determine driving behaviors according to environmental information and vehicle status [11]. In recent years, researchers have extensively reported numerous technological advances in automated driving [12, 13, 14]. Here our work centers around the decision-making system of automated vehicles.

The decision-making technologies utilized in automated vehicles can be broadly categorized into rule-based, modeling-based, and learning-based methods [15]. The finite state machine (FSM) is a widely recognized scheme for rule-based driving decision-making [16, 17]. A hierarchical FSM strategy was developed, which utilizes a meta-state machine for different situations and a substate machine for the driving state of the vehicle in [18]. An adaptive FSM scheme was presented to control automated vehicles by following a sequence of states/behaviors to reach a destination in [19]. A behavior decision-making scheme based on rules was developed by combining a fuzzy rule base for modeling basic driving elements with a probabilistic FSM in [20]. In general, such methods are simple to implement and highly interpretable. However, since these approaches lack learning skills, they rely heavily on the prior knowledge of specialists. Moreover, it is difficult to design rules in complex scenarios.

The modeling-based decision-making schemes are generally based on game theory or Markov decision process (MDP) [15]. A Stackelberg game was adopted to cope with the automated driving decision-making problem in [21]. An automated vehicle determines its actions that maximize its utility by considering the actions that the following vehicles may take. All other vehicles perform similarly in each stage of the game. The Stackelberg game theory was leveraged to model the driving decisions by taking into account various driving styles and social interaction characteristics in [22]. A method for situation-aware decision-making in urban automated driving was presented using a partially observable MDP in [23]. An intention-aware online partially observable MDP scheme was designed to handle automated driving decision-making tasks in [24]. By and large, such methods can derivate the optimal policies by mathematical models of strategic interactions among vehicles or agents. Nonetheless, they are computationally intractable for large state and action spaces, since these schemes have to solve an optimization problem for each of the different situations. Moreover, such ways lack the generalizability to unseen cases.

The learning-based decision-making approaches generally are implemented through imitation learning (IL) and reinforcement learning (RL) [25, 26]. Xu et al. advanced a framework based on combining a fully convolutional network (FCN) with long short-term memory (LSTM) to learn driving policies via a large-scale crowd-sourced video dataset. Kuefler et al. [27] developed a generative adversarial IL method to tackle automated driving decision problems. Ngai et al. [28] proposed a multiple-goal RL technique to cope with the vehicle overtaking task by double-action Q-learning. Chen et al. [29] presented an automated driving decision scheme using deep hierarchical RL. Everett et al. [30] advanced an automated driving decision method for collision avoidance by combining deep RL with LSTM. Xu et al. [31] proposed an RL technique with multi-objective approximate policy iteration to determine driving behaviors for automated vehicles on highways. You et al. [32] designed an integrated RL and deep inverse RL approach to obtain the optimal driving policy. By deep neural networks (DNNs) as powerful function approximators [33, 34], the learning-based methods can learn a single parametric representation that maps high-dimensional perceived information directly to driving strategies while generalizing unseen situations [35, 36]. However, in general, the decision-making approaches with IL have high requirements on the quantity and quality of labeled training data or demonstration data that is expensive and time-consuming to collect in practical applications. The decision-making approaches with RL differ from IL schemes in not needing labeled data or demonstration data. Instead, RL agents acquire skills to perform tasks by repeatedly interacting with the environment and learning from trial-and-error [37, 38].

Although the aforementioned decision-making methods have achieved impressive success in a series of automated driving tasks, there is still room for improvement and perfection. On the one hand, the sensing and perception information of automated vehicles may contain natural stochastic noises or perception errors that could mislead automated driving systems into taking suboptimal or even cause catastrophic failures. Nonetheless, few studies concern and cope with this challenge. On the other hand, the longitudinal and lateral decision-making behaviors (e.g., driving speed and lane changing decisions) are coupled, that is, when one of them is perturbed by unknown external disturbances, it gives rise to changes or even performance degradation in the other. In other words, the longitudinal and lateral decision-making systems require to be effectively coordinated to guarantee vehicle performance. For instance, if observational perturbations in sensor inputs cause an automated vehicle to swerve sharply, then the vehicle speed should be regulated effectively to ensure driving safety. Consequently, these two challenges greatly hinder the potential of automated vehicles. Notwithstanding, to the best of our knowledge, the robust coordinated longitudinal and lateral decision-making technique for automated driving has not yet been fully explored.

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

Based on the earlier considerations, to ensure the automated driving performance and the policy robustness against observational uncertainties, we present a novel robust coordinated decision-making (RCDM) approach for automated driving through robust multiagent reinforcement learning (RMARL). The proposed RCDM technique attempts to enable an automated vehicle to coordinate its longitudinal and lateral decision-making systems while ensuring driving policy robustness against observational uncertainties. This work's main contributions are highlighted as follows.

- A constrained robust multiagent MDP (CRMA-MDP) is presented to model the cooperative driving behaviors of the longitudinal and lateral decision-making systems of an automated vehicle under observational uncertainties. Additionally, a nonlinear constraint setting with Kullback–Leibler (KL) divergence is developed to keep the variations of the longitudinal and lateral driving policies perturbed by stochastic observational perturbations within bounds.
- A robust multiagent policy optimization (RMAPO) algorithm is advanced to enable the longitudinal and lateral decision-making systems to learn the optimal robust coordinated policies by solving the constrained optimization problem formulated via CRMA-MDP and the nonlinear constraint.

We assess the performance of the proposed method through three testing cases with varying traffic flow densities using the simulation of urban mobility (SUMO) platform [39]. Our results demonstrate that the proposed RCDM approach effectively enhances the performance of the automated vehicle while ensuring the driving policy robustness against observational uncertainties.

The remainder of this article is organized as follows. Section "Methodology" presents an illustration of the RCDM approach. Section "Technical Implementation" provides implementation details of our method. Section "Performance Evaluation" discusses the testing results and analyses. Finally, Section "Conclusion" concludes this work.

Methodology

Technical Framework

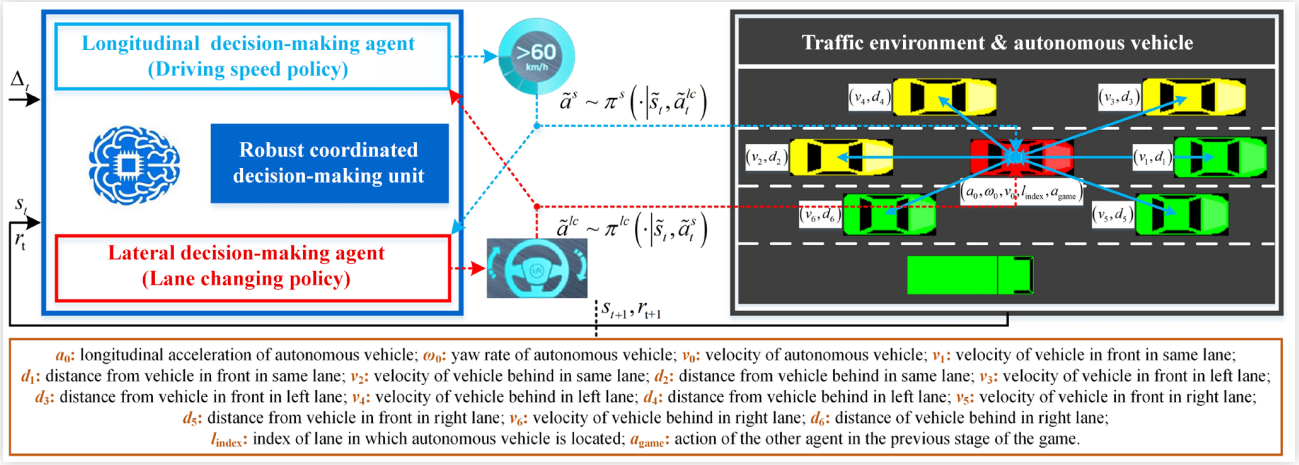
In Figure 1, we show a block diagram of the architecture for the proposed RCDM approach. The red vehicle represents the ego automated vehicle, whereas the other colored vehicles denote social vehicles. Since the longitudinal and lateral decision-making systems of an automated vehicle have a common objective (i.e., improving automated driving performance), the interaction between them can be regarded as a cooperative game (i.e., multiagent collaboration problem). The states of the longitudinal and lateral decision-making agents contain 17 dimensions. The output of the longitudinal agent is the continuous vehicle speed. Furthermore, the lateral agent executes a discrete lane changing policy, including lane keeping, left lane changing and right lane changing.

In Figure 1, the longitudinal decision-making agent (i.e., system) makes an approximately optimal response to the action from the lateral decision-making agent, while the lateral decision-making agent will also take into the response of the longitudinal decision-making agent. The action of one agent can be completely observed by the other agent at each stage of the game. Moreover, Δ_t represents observational uncertainty at the time step t, $\tilde{s}_t$ denotes the state perturbed by stochastic perturbation at the time step t, s_t and s_{t+1} represent states at the time step t and $t+1$, r_t and r_{t+1} denote the reward at the time step t and $t+1$, $\tilde{a}_t^s$ and $\tilde{a}_t^{lc}$ represent the perturbed actions of longitudinal and lateral decision-making agents at the time step t, π^s and π^{lc} denote the policies of longitudinal and lateral agents, respectively.

Problem Modeling

In this section, we model strategic interactions between the longitudinal and lateral decision-making systems of an automated vehicle as a cooperative multiagent problem. Specifically,

FIGURE 1 Illustration of our robust coordinated decision-making framework for automated driving.



© Nanyang Technological University

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

we extend existing MDP formulations for explicitly modeling the multiagent behaviors under observational uncertainties. Hence, here CRMA-MDP is presented.

Definition 1 *A CRMA-MDP is able to be characterized by a 7-tuple* $(\mathcal{S}, \mathcal{A}^n, p^n, r^n, \gamma, \Delta, C)$, where $\mathcal{S}$ *denotes the state space.* $\mathcal{A}^n : \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ *is a space of joint action, where* $\mathcal{A}_i$ *indicates the space of action that can be taken by agent* i. $p^n : \mathcal{S} \times \mathcal{A}^n \times \mathcal{S} \to \mathbb{R}$ *represents the transition probability from the current state to the next state under the joint action for all agents.* $r^n : \mathcal{S} \times \mathcal{A}^n \to \mathbb{R}$ *represents the shared reward function for each agent.* $\gamma \in (0, 1)$ *signifies the discount factor.* Δ *denotes the observational uncertainty.* C *is the constraint function.* In CRMA-MDP, all agents share the state space and the reward. Moreover, in the RCDM task, the longitudinal decision-making agent is taking speed policy π_s and the lateral decision-making agent is playing the lane changing policy π_{lc}, then both agents observe the perturbed state $\tilde{s}_t$ and take perturbed actions $\tilde{a}_t^s \sim \pi^s\left(\cdot\,|\tilde{s}_t, \tilde{a}_t^{lc}\right)$ and $\tilde{a}_t^{lc} \sim \pi^{lc}\left(\cdot\,|\tilde{s}_t, \tilde{a}_t^s\right)$ at every time step t. In addition, the transition probability $s_{t+1} = p\left(s_{t+1}|s_t, \tilde{a}_t^s, \tilde{a}_t^{lc}\right)$, the shared reward $r_t = r\left(s_t, \tilde{a}_t^s, \tilde{a}_t^{lc}\right)$ can be obtained from the environment. Therefore, CRMA-MDP for each time step can be represented by $\left(s_t, \tilde{a}_t^s, \tilde{a}_t^{lc}, r_t, s_{t+1}\right)$.

According to CRMA-MDP, our RCDM task is able to be formulated as the following constrained optimization task:

$$\begin{aligned} &\max_{\pi^s, \pi^{lc}} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r\left(s_t, a_t^s, a_t^{lc}\right)\right] \\ &\text{s.t. } \forall t, \quad \mathbb{E}\left[C^{\pi^s}\left(s_t, a_t^{lc}, \Delta_t\right)\right] \le \epsilon_s \\ &\qquad\quad \mathbb{E}\left[C^{\pi^{lc}}\left(s_t, a_t^s, \Delta_t\right)\right] \le \epsilon_{lc} \end{aligned} \qquad \text{Eq. (1)}$$

where T is the last time step, ϵ_s and ϵ_{lc} are predefined thresholds, a_t^s and a_t^{lc} denote the actions of the longitudinal and lateral decision-making agents, respectively. Additionally, the longitudinal and lateral driving policies' KL divergence-based constraints $C^{\pi^s}(\cdot)$ and $C^{\pi^{lc}}(\cdot)$ can be defined as:

$$C^{\pi^s}\left(s, a^{lc}, \Delta\right) = D_{KL}^s\left(\pi^s\left(a^s|s, a^{lc}\right)||\pi^s(\tilde{a}^s|\tilde{s}, \tilde{a}^{lc})\right) \qquad \text{Eq. (2)}$$

$$C^{\pi^{lc}}\left(s, a^s, \Delta\right) = D_{KL}^{lc}\left(\pi^{lc}\left(a^{lc}|s, a^s\right)||\pi^{lc}(\tilde{a}^{lc}|\tilde{s}, \tilde{a}^s)\right) \qquad \text{Eq. (3)}$$

$$\tilde{s} = \Delta \cdot s \qquad \text{Eq. (4)}$$

where the observational uncertainty $\Delta = 1 - \bar{\Delta}$, and $\bar{\Delta}$ is the independent Gaussian noise with variance $\bar{\sigma}^2$. Here $\bar{\sigma}$ is set to 1.00. Moreover, $C^{\pi^s}(\cdot)$ or $C^{\pi^{lc}}(\cdot)$ can measure the variations of the policies perturbed by stochastic observational perturbations. Here, to improve the driving policy robustness against perturbations, we try to maximize expected return while keeping $C^{\pi^s}(\cdot)$ and $C^{\pi^{lc}}(\cdot)$ within bounds.

For the longitudinal decision-making agent, the speed policy is represented by a Gaussian with mean and variance provided via neural networks. We denote $\pi^s\left(a^s|s, a^{lc}\right) \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and $\pi^s\left(\tilde{a}^s|\tilde{s}, \tilde{a}^{lc}\right) \sim \mathcal{N}\left(\tilde{\mu}, \tilde{\sigma}^2\right)$, where μ and $\tilde{\mu}$ are the mean of Gaussian distribution and Gaussian distribution with observational uncertainty Δ, σ and $\tilde{\sigma}$ are the variance of Gaussian distribution and Gaussian distribution with Δ respectively. Hence, the constraint $C^{\pi^s}(\cdot)$ can be written as:

$$\begin{aligned} C^{\pi^s}\left(s, a^{lc}, \Delta\right) &= D_{KL}^s\left(\pi^s\left(a^s|s, a^{lc}\right)||\pi^s(\tilde{a}^s|\tilde{s}, \tilde{a}^{lc})\right) \\ &= \log\frac{\tilde{\sigma}}{\sigma} + \frac{\left(\mu - \tilde{\mu}\right)^2 + \sigma^2 - \tilde{\sigma}^2}{2\tilde{\sigma}^2}. \end{aligned} \qquad \text{Eq. (5)}$$

For the lateral decision-making agent, the lane changing policy is discrete, and the constraint $C^{\pi^{lc}}(\cdot)$ can be expressed as:

$$\begin{aligned} C^{\pi^{lc}}\left(s, a^s, \Delta\right) &= D_{KL}^{lc}\left(\pi^{lc}\left(a^{lc}|s, a^s\right)||\pi^{lc}(\tilde{a}^{lc}|\tilde{s}, \tilde{a}^s)\right) \\ &= \sum \pi^{lc}\left(a^{lc}|s, a^s\right) \log\left(\frac{\pi^{lc}\left(a^{lc}|s, a^s\right)}{\pi^{lc}\left(\tilde{a}^{lc}|\tilde{s}, \tilde{a}^s\right)}\right). \end{aligned} \qquad \text{Eq. (6)}$$

Optimization Approach

The proposed RMAPO method is introduced in this section. RMAPO attempts to enable the longitudinal and lateral decision-making agents to learn the optimal robust coordinated driving policies.

Here the expected return of the cooperative task can be decomposed into a sum of rewards at all the time steps. Since the policy at the current time step can only affect the future objective value, the dynamic programming algorithm is able to be leveraged to solve the policy backward through time. Hence, the optimization objective of driving decision-making agents can be represented as an iterated maximization:

$$\begin{aligned} v^{\pi^s, \pi^{lc}}\left(s\right) = \max_{\pi_0^s, \pi_0^{lc}} &\left(\mathbb{E}\left[r\left(s_0, a_0^s, a_0^{lc}\right)\right]\right. \\ &\left. + \max_{\pi_1^s, \pi_1^{lc}} \left(\mathbb{E}\left[\ldots\right] + \max_{\pi_T^s, \pi_T^{lc}} \mathbb{E}[r(s_T, a_T^s, a_T^{lc})]\right)\right). \end{aligned} \qquad \text{Eq. (7)}$$

The longitudinal and lateral decision-making agents can be optimized from the last time step T. Here we require to solve the following constrained optimization task:

$$\begin{aligned} &\max_{\pi_T^s, \pi_T^{lc}} f\left(\pi_T^s, \pi_T^{lc}\right) \\ &\text{s.t. } c_T^s \ge 0, \; c_T^{lc} \ge 0 \end{aligned} \qquad \text{Eq. (8)}$$

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

with

$$c_T^s = \epsilon_s - C^{\pi^s}\left(s_T,\, a_T^{lc},\, \Delta_T\right)$$
$$c_T^{lc} = \epsilon_{lc} - C^{\pi^{lc}}\left(s_T,\, a_T^{s},\, \Delta_T\right)$$
$$f\left(\pi_T^s,\, \pi_T^{lc}\right) = \begin{cases} \mathbb{E}\left[r\left(s_T,\, \tilde{a}_T^s,\, \tilde{a}_T^{lc}\right)\right], & \text{if } c_T^s \geq 0 \text{ and } c_T^{lc} \geq 0 \\ -\infty, & \text{otherwise} \end{cases}.$$

Hence, the Lagrange function can be obtained:

$$L\left(\pi_T^s,\, \pi_T^{lc},\, \alpha_T^s,\, \alpha_T^{lc}\right) = f\left(\pi_T^s,\, \pi_T^{lc}\right) + \alpha_T^s c_T^s + \alpha_T^{lc} c_T^{lc} \quad \text{Eq. (9)}$$

where α_T^s and α_T^{lc} are Lagrange multipliers in the optimization of the speed policy and the lane changing policy at the time step T.

If we attempt to minimize $L\left(\pi_T^s,\, \pi_T^{lc}, \alpha_T^s, \alpha_T^{lc}\right)$ with respect to α_T^s and α_T^{lc} under π_T^s and π_T^{lc}, then the following equation is able to be derived:

$$f\left(\pi_T^s,\, \pi_T^{lc}\right) = \min_{\alpha_T^s \geq 0, \alpha_T^{lc} \geq 0} L\left(\pi_T^s,\, \pi_T^{lc},\, \alpha_T^s,\, \alpha_T^{lc}\right). \quad \text{Eq. (10)}$$

Here, since the constraints are convex functions and the optimization objective is linear, the strong duality holds [40]. Suppose $f\left(\pi_T^s,\, \pi_T^{lc}\right)$ is maximized, the following relation can be derived:

$$\max_{\pi_T^s,\, \pi_T^{lc}} f\left(\pi_T^s,\, \pi_T^{lc}\right) = \max_{\pi_T^s,\, \pi_T^{lc}} \min_{\alpha_T^s \geq 0, \alpha_T^{lc} \geq 0} L\left(\pi_T^s,\, \pi_T^{lc},\, \alpha_T^s,\, \alpha_T^{lc}\right) = \min_{\alpha_T^s \geq 0, \alpha_T^{lc} \geq 0} \max_{\pi_T^s,\, \pi_T^{lc}} L\left(\pi_T^s,\, \pi_T^{lc},\, \alpha_T^s,\, \alpha_T^{lc}\right). \quad \text{Eq. (11)}$$

The optimal π_T^s, π_T^{lc}, α_T^s and α_T^{lc} can be approximated iteratively. First given the current α_T^s and α_T^{lc}, compute the best policies π_T^{s*} and π_T^{lc*} which maximizes $L\left(\pi_T^s,\, \pi_T^{lc}, \alpha_T^s, \alpha_T^{lc}\right)$. Furthermore, plug in π_T^{s*} and π_T^{lc*}, and approximate α_T^{s*} and α_T^{lc*}, which minimizes $L\left(\pi_T^{s^*},\, \pi_T^{lc^*}, \alpha_T^s, \alpha_T^{lc}\right)$. Therefore, the following expression can be obtained:

$$\pi_T^{s*},\, \pi_T^{lc*} = \arg\max_{\pi_T^s,\, \pi_T^{lc}} L\left(\pi_T^s,\, \pi_T^{lc},\, \alpha_T^s,\, \alpha_T^{lc}\right) \quad \text{Eq. (12)}$$

$$\alpha_T^{s*},\, \alpha_T^{lc*} = \arg\max_{\alpha_T^s \geq 0,\, \alpha_T^{lc} \geq 0} L\left(\pi_T^{s*},\, \pi_T^{lc*},\, \alpha_T^s,\, \alpha_T^{lc}\right). \quad \text{Eq. (13)}$$

At the time step $T-1$, the constrained Q-function and the constrained value function under observational uncertainties can be defined as:

$$Q_{T-1}^{\pi^s,\, \pi^{lc}}\left(s_{T-1},\, a_{T-1}^s,\, a_{T-1}^{lc}\right) = \mathbb{E}\left[r\left(s_{T-1},\, a_{T-1}^s,\, a_{T-1}^{lc}\right)\right] + \mathbb{E}[\boldsymbol{r}\left(s_T,\, a_T^s,\, a_T^{lc}\right) - \alpha_T^s C^{\pi^s}\left(s_T,\, a_T^{lc},\, \Delta_T\right) - \alpha_T^{lc} C^{\pi^{lc}}\left(s_T,\, a_T^s,\, \Delta_T\right)]. \quad \text{Eq. (14)}$$

$$\upsilon^{\pi^s,\, \pi^{lc}}\left(s_{T-1}\right) = \mathbb{E}[Q_{T-1}^{\pi^s,\, \pi^{lc}}\left(s_{T-1},\, a_{T-1}^s,\, a_{T-1}^{lc}\right) - \alpha_{T-1}^s C^{\pi^s}\left(s_{T-1},\, a_{T-1}^{lc},\, \Delta_{T-1}\right) - \alpha_{T-1}^{lc} C^{\pi^{lc}}\left(s_{T-1},\, a_{T-1}^s,\, \Delta_{T-1}\right)]. \quad \text{Eq. (15)}$$

With Equation 14, π_T^{s*} and π_T^{lc*}, the optimal constrained Q-function under observational uncertainties can be obtained:

$$Q_{T-1}^{\pi^{s*},\, \pi^{lc*}}\left(s_{T-1},\, a_{T-1}^s,\, a_{T-1}^{lc}\right) = \mathbb{E}\left[r\left(s_{T-1},\, a_{T-1}^s,\, a_{T-1}^{lc}\right)\right] + \max_{\pi_T^s,\, \pi_T^{lc}} \mathbb{E}\left[r\left(s_T,\, a_T^s,\, a_T^{lc}\right)\right] - \alpha_T^s C^{\pi^{s*}}\left(s_T,\, a_T^{lc},\, \Delta_T\right) - \alpha_T^{lc} C^{\pi^{lc*}}\left(s_T, a_T^s, \Delta_T\right) \quad \text{Eq. (16)}$$

Hence, if we move one step back to the time step $T-1$, with Equation 16 and the duality theory, the following expression is able to be obtained:

$$\max_{\pi_{T-1}^s,\, \pi_{T-1}^{lc}} \left(\mathbb{E}\left[r(s_{T-1}, a_{T-1}^s, a_{T-1}^{lc})\right] + \max_{\pi_T^s,\, \pi_T^{lc}} \mathbb{E}\left[r(s_T, a_T^s, a_T^{lc})\right]\right) = \min_{\alpha_{T-1}^s \geq 0, \alpha_{T-1}^{lc} \geq 0} \max_{\pi_{T-1}^s,\, \pi_{T-1}^{lc}} \left(Q_{T-1}^{\pi^{s*},\, \pi^{lc*}}(s_{T-1}, a_{T-1}^s, a_{T-1}^{lc}) + \alpha_{T-1}^s c_{T-1}^s + \alpha_{T-1}^{lc} c_{T-1}^{lc}\right) + \alpha_T^s C^{\pi^{s*}}(s_T, a_T^{lc}, \Delta_T) + \alpha_T^{lc} C^{\pi^{lc*}}(s_T, a_T^s, \Delta_T). \quad \text{Eq. (17)}$$

Similarly, the optimal driving policies and the optimal Lagrange multipliers at the time step $T-1$ can be obtained by the following equations:

$$\pi_{T-1}^{s*}, \pi_{T-1}^{lc*} = \arg\max_{\pi_{T-1}^s,\, \pi_{T-1}^{lc}} \mathbb{E}[Q_{T-1}^{\pi^{s*},\, \pi^{lc*}}\left(s_{T-1},\, a_{T-1}^s,\, \boldsymbol{a}_{T-1}^{lc}\right) + \alpha_{T-1}^s c_{T-1}^s + \alpha_{T-1}^{lc} c_{T-1}^{lc}] \quad \text{Eq. (18)}$$

$$\alpha_{T-1}^{s*},\, \alpha_{T-1}^{lc*} = \arg\min_{\alpha_{T-1}^s \geq 0, \alpha_{T-1}^{lc} \geq 0} \mathbb{E}[\alpha_{T-1}^s \epsilon_s - \alpha_{T-1}^s C^{\pi^{s*}}\left(s_T,\, a_T^{lc},\, \Delta_T\right) + \alpha_{T-1}^{lc} \epsilon_{lc} - \alpha_{T-1}^{lc} C^{\pi^{lc*}}\left(s_T,\, a_T^{lc},\, \Delta_T\right)]. \quad \text{Eq. (19)}$$

In this way, the constrained optimization problem 1 can be solved recursively. However, without a limitation on the distance between the old and the new policies, to optimize agent would lead to instability with extremely large parameter updates. Here we impose the constraint via forcing the distance between the old and the new policies to stay within a small interval. Although our scheme tries to solve the cooperative task, the decentralized value function is adopted to accurately assess the decision-making behaviors of the driving

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

agents. Therefore, for the longitudinal decision-making agent, its optimization objective can be written as:

$$\max_{\theta^s} \mathbb{E}\left[J^s\left(\theta^s\right)\right] \quad \text{s.t. } \mathbb{E}\left[C^{\pi^s}\left(s, a^{lc}, \Delta\right)\right] \le \epsilon_s \qquad \text{Eq. (20)}$$

where θ^s represents the policy model parameters of the longitudinal decision-making agent, $J^s(\cdot)$ is the clipped surrogate objective of the agent, it is given via:

$$J^s\left(\theta^s\right) = \min\left(\frac{\pi^s\left(\cdot|s, a^{lc}; \theta^s\right)}{\pi^s\left(\cdot|s, a^{lc}; \theta^s_{\text{old}}\right)}\hat{A}^{\theta^s}\left(s, a^s, a^{lc}, \Delta\right), \text{clip}\left(\frac{\pi^s\left(\cdot|s, a^{lc}; \theta^s\right)}{\pi^s\left(\cdot|s, a^{lc}; \theta^s_{\text{old}}\right)}, 1-\epsilon, 1+\epsilon\right)\hat{A}^{\theta^s}(s, a^s, a^{lc}, \Delta)\right) \qquad \text{Eq. (21)}$$

where $\hat{A}^{\theta^s}(\cdot)$ is the advantage function of the longitudinal decision-making agent and ϵ is a clip ratio.

With generalized advantage estimation technique and the constraint $C^{\pi^s}(\cdot)$, the advantage function $\hat{A}^{\theta^s}(\cdot)$ in Equation 21 is able to be developed as:

$$\hat{A}^{\theta^s}\left(s_t, a^s_t, a^{lc}_t, \Delta\right) = \delta^s_t + (\gamma\lambda)\delta^s_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta^s_{T-1} \qquad \text{Eq. (22)}$$

$$\delta^s_t = \upsilon^{\text{target}}_{\phi^s} - v^{\pi^s,\pi^{lc}}\left(s_t; \phi^s\right) \qquad \text{Eq. (23)}$$

$$v^{\text{target}}_{\phi^s} = \mathbb{E}[r\left(s_t, a^s_t, a^{lc}_t\right) + \gamma v^{\pi^s,\pi^{lc}}\left(s_{t+1}; \phi^s\right) - \alpha^s C^{\pi^s}\left(s_t, a^{lc}_t, \Delta_t\right)] \qquad \text{Eq. (24)}$$

where λ is a decay factor, $v^{\text{target}}_{\phi^s}$ is a target value function of the longitudinal decision-making agent, and ϕ^s are the parameters of the evaluation model for the agent.

The lateral decision-making agent is trained based on the method similar to the longitudinal agent. Therefore, the policy model parameters of the longitudinal and lateral driving agents can be updated by maximizing the following loss functions:

$$J^s_\pi\left(\theta^s\right) = \mathbb{E}\left[J^s\left(\theta^s\right) - \alpha^s C^{\pi^s}\left(s, a^{lc}, \Delta\right)\right] \qquad \text{Eq. (25)}$$

$$J^{lc}_\pi\left(\theta^{lc}\right) = \mathbb{E}\left[J^{lc}\left(\theta^{lc}\right) - \alpha^{lc} C^{\pi^{lc}}\left(s, a^s, \Delta\right)\right] \qquad \text{Eq. (26)}$$

where θ^{lc} represents the policy model parameters of the lateral decision-making agent and $J^{lc}(\cdot)$ is the clipped surrogate objective of the agent.

The evaluation model parameters of the longitudinal and lateral agents are able to be learned by minimizing the following loss functions:

$$J^s_v\left(\phi^s\right) = \mathbb{E}\left[\left\|v^{\text{target}}_{\phi^s} - v^{\pi^s,\pi^{lc}}\left(s; \phi^s\right)\right\|^2_2\right] \qquad \text{Eq. (27)}$$

$$J^{lc}_v\left(\phi^{lc}\right) = \mathbb{E}\left[\left\|v^{\text{target}}_{\phi^{lc}} - v^{\pi^s,\pi^{lc}}\left(s; \phi^{lc}\right)\right\|^2_2\right] \qquad \text{Eq. (28)}$$

where $v^{\text{target}}_{\phi^{lc}}$ is a target value function of the lateral decision-making agent and ϕ^{lc} are the parameters of the evaluation model for the lateral agent.

The optimal dual variables of the two agents can be approximated by minimizing the following objective:

$$J^s_\alpha\left(\alpha^s\right) = \mathbb{E}\left[\alpha^s\epsilon_s - \alpha^s C^{\pi^s}\left(s, a^{lc}, \Delta\right)\right] \qquad \text{Eq. (29)}$$

$$J^{lc}_\alpha\left(\alpha^{lc}\right) = \mathbb{E}\left[\alpha^{lc}\epsilon_{lc} - \alpha^{lc} C^{\pi^{lc}}\left(s, a^s, \Delta\right)\right]. \qquad \text{Eq. (30)}$$

Technical Implementation

Algorithm

The proposed scheme optimizes both of the RL agents by the following alternating procedure. Both RL agents' initial policy model parameters are determined based on a random distribution. For each iteration, agents first require to collect the interactive data of M time steps and save them to memory $\mathcal{D}$. Furthermore, the longitudinal agent based on the policy π^s and action from the lateral agent makes a response, and then the lateral agent with the policy π^{lc} and action from the longitudinal agent takes a response. The environment involves the reward functions and the transition probability to provide the interactive information. Then the policies of the two agents are optimized alternately. Algorithm 1 outlines the proposed RCDM for automated driving in detail.

State and Action

Figure 1 shows the observations of the longitudinal and lateral driving decision-making agents. In order to guarantee the policy model's generalization, we convert the observations of the agent to the states through normalization. Aside from the states of the ego vehicle, we also leverage the relevant information of the six nearest social cars in the ego vehicle's lane and adjacent lanes on both sides. The state space of the longitudinal or lateral decision-making agent has 17 dimensions, including the relative distances between the surrounding social vehicles and the ego vehicle, the velocities of the surrounding social vehicles, the longitudinal acceleration, yaw rate, driving speed

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

ALGORITHM 1 Robust coordinated decision-making.

Initialize policy model parameters θ^s and θ^{lc}, evaluation model parameters ϕ^s and ϕ^{lc}, dual variables α^s and α^{lc}, and an empty memory $\mathcal{D}$.
for iteration step $n = 1, 2, \dots N$ **do**
  Reset state s, action $\tilde{a}^{lc}$ and memory $\mathcal{D}$.
  for environment step $t = 1, 2, \dots M$ **do**
    Determine driving speed based on π^s: $a_t^s \sim \pi^s(\cdot|s_t, a_t^{lc}; \theta^s)$.
    Select lane changing behavior based on π^{lc}: $a_t^{lc} \sim \pi^{lc}(\cdot|s_t, a_t^s; \theta^{lc})$.
    Obtain the transition from the environment: $s_{t+1} \sim p(\cdot|s_t, a_t^s, a_t^{lc})$, $r_t \leftarrow r(s_t, a_t^s, a_t^{lc})$.
    Save the transition in the memory: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t^s, a_t^{lc}, r_t, s_{t+1})\}$.
  end for
  Extract data trajectories from memory $\mathcal{D}$.
  Update policy model parameters by Eq. 25 and Eq. 26: $\theta^s \leftarrow \nabla J_\pi^s(\theta^s)$, $\theta^{lc} \leftarrow \nabla J_\pi^{lc}(\theta^{lc})$.
  Update evaluation model parameters by Eq. 27 and Eq. 28: $\phi^s \leftarrow \nabla J_v^s(\phi^s)$, $\phi^{lc} \leftarrow \nabla J_v^{lc}(\phi^{lc})$.
  Update dual variables by Eq. 29 and Eq. 30: $\alpha^s \leftarrow \nabla J_\alpha^s(\alpha^s)$, $\alpha^{lc} \leftarrow \nabla J_\alpha^{lc}(\alpha^{lc})$.
end for

© Nanyang Technological University

and lane index of the ego vehicle, and the action from the other agent in the previous stage of the game.

The action space of the longitudinal decision-making agent is the continuous driving speed. The maximum traffic speed of all vehicles is 35 m/s. Furthermore, the action space of the lateral driving agent is discrete, including lane keeping, left lane changing and right lane changing.

Reward Function

One of the challenges in this research is learning the robust coordinated longitudinal and lateral driving policies from scratch, without prior knowledge. As a result, the reward function plays a crucial role in guiding the RL agent toward learning the preferred decision-making behaviors [41].

Since our research tries to solve a cooperative task, the shared reward function for each agent is designed. Transport efficiency, comfort, and safety need to be considered in determining the driving speed and lane-change behavior. Algorithm 2 outlines the shared reward function design in detail where a_y denotes vehicle lateral acceleration, k represents dynamic factor [42], g represents gravity acceleration, and $\bar{\mu}$ denotes adhesion coefficient. The comfort in reward function is set based on the results of [43]. In terms of safety, not only collision but also vehicle dynamic stability is considered. Here, we leverage a linear two-degree-of-freedom vehicle dynamics model to describe the dynamic behaviors of the vehicle. According to a study [44], the vehicle yaw rate's upper bound is able to be written as:

$$\bar{\omega} = 0.85 \frac{\bar{\mu} g}{v_0}. \qquad \text{Eq. (31)}$$

ALGORITHM 2 Shared reward function design.

Input: State, speed, lane changing behavior.
$r(\cdot) = v_0/35$. ▷ Reward the agent for enhancing efficiency
if $d_1 < 30$ **then**
  $r(\cdot) = r(\cdot) - 0.1$. ▷ Reward the agent for overtaking
end if
if $a_0 > 1.47$ **or** $a_0 < -2$ **or** $|a_y| > 4$ **then**
  $r(\cdot) = r(\cdot) - 0.02$. ▷ Penalize uncomfortable driving
end if
if $|\omega_0| > k \cdot \bar{\mu} \cdot g/v_0$ **and** $v_0 > 30$ **then**
  $r(\cdot) = r(\cdot) - 0.02$. ▷ Penalize instable dynamics driving
end if
if the ego vehicle performs a lane change **and** $v_0 > 20$ **then**
  $r(\cdot) = r(\cdot) - a^s/350$. ▷ Penalize high-speed lane changing
end if
if the ego vehicle collides **then**
  $r(\cdot) = r(\cdot) - 0.1$. ▷ Penalize dangerous driving
end if

© Nanyang Technological University

Hence, the dynamic factor k is set as 0.85 in this work. If the measured vehicle yaw rate exceeds its upper bound and the driving speed is higher than 30 m/s, our RL agent will receive a penalty signal.

Network and Hyperparameter

The policy and evaluation neural networks are constructed with a single fully connected hidden layer of size 128, using ReLU activation functions. The key hyperparameters for the proposed algorithm are listed in Table 1.

Performance Evaluation

Environment

In this section, the experiment is implemented to assess the performance of the proposed RCDM scheme for automated vehicles. The SUMO platform has been available since 2001, and it is developed to cope with large-scale complex traffic flows and networks. SUMO is exclusively microscopic: each vehicle is explicitly modeled, has its own route, and moves through the network individually. Simulations are

TABLE 1 The main hyperparameters of our RCDM approach.

Parameters	Value	Parameters	Value
Clip ratio ϵ	0.1	Discount factor γ	0.95
Decay factor λ	0.95	Adhesion coefficient $\bar{\mu}$	0.90
Dynamic factor k	0.85	Actor learning rate l_a	0.00005
Dual learning rate l_α	0.001	Critic learning rate l_c	0.001
Constraint threshold ϵ_s	0.05	Constraint threshold ϵ_{lc}	0.001

© Nanyang Technological University

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

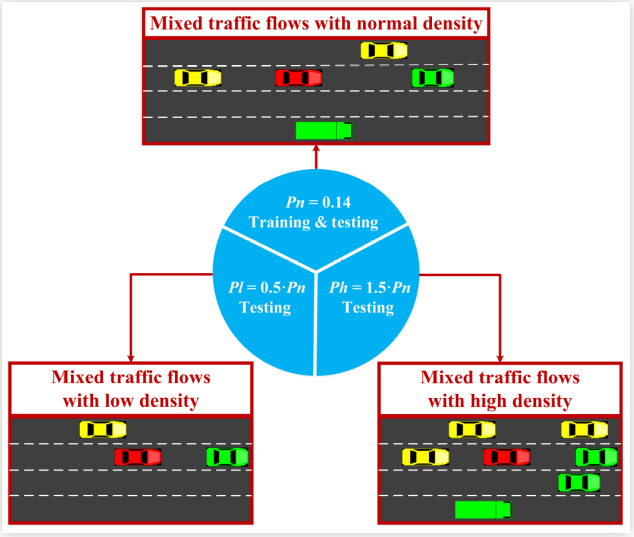
deterministic by default, but there are many ways to introduce randomness. Hence, here the SUMO platform is adopted to simulate the highway scenarios with the three stochastic mixed traffic flows based on different densities.

Figure 2 shows the proposed assessment method. *P* is utilized to represent the probability with regard to emitting a car each second. *Pn*, *Pl*, and *Ph* denote the probabilities concerning emitting a car each second in stochastic traffic flows with normal, low, and high densities, respectively. Moreover, we set *Pn*, *Pl*, and *Ph* as 0.14, 0.07, and 0.21, respectively. The proposed approach and baseline methods are evaluated both during training and testing. We train and test the policy model based on each method using the stochastic traffic flows with the normal density. Meanwhile, we leverage the stochastic traffic flows with low and high densities to test the performance of the policy models. All social vehicles are driven by the intelligent driver model (IDM) of SUMO. All evaluations are conducted on a single computer equipped with a 2.90-GHz 12-core Intel i9-8950HK CPU.

Baseline

The multiagent proximal policy optimization (MAPPO) [45] scheme is a state-of-the-art baseline in MARL. Therefore, the coordinated decision-making (CDM) method with MAPPO (CDM-MAPPO) is implemented as a state-of-the-art baseline. The CDM-MAPPO agents with observational perturbations (CDM-MAPPO-OP) during interaction with the environment are employed as the second baseline method to analyze the impact of uncertainties on the training effect. Additionally, for the same reason, our RCDM agents with observational perturbations during interaction with the environment can be denoted as RCDM-OP. Here the observational perturbation obeys Gaussian distribution.

FIGURE 2 Evaluation scheme by SUMO-based stochastic mixed traffic flows.



© Nanyang Technological University

Soft actor-critic (SAC) [46] is a state-of-the-art RL algorithm. To further benchmark the proposed approach, a longitudinal and lateral decision technique with SAC (LLDM-SAC) is implemented as a competitive baseline. In this baseline, the longitudinal and lateral decision behaviors are determined via two SAC agents that output continuous and discrete actions, respectively. Besides, to evaluate the robustness of the agents, LLDM-SAC-OP is used to represent the LLDM-SAC policy perturbed by observation noises.

Metric

We leverage the driving speed and number of collisions to measure the travel efficiency and driving safety of the automated vehicle. Furthermore, the expected return is adopted to assess the automated driving agent's comprehensive performance.

In addition, to evaluate the policy robustness against observational perturbations, with Equations 5 and 6, we develop the following metric for robustness:

$$M_r = \frac{C^{\pi^s}\left(s,a^{lc},\Delta\right)+C^{\pi^{lc}}\left(s,a^{s},\Delta\right)}{2}. \quad \text{Eq. (32)}$$

According to Equation 32, we can infer that the smaller variations of the policies perturbed by stochastic observational perturbations, then the smaller robustness metric, and stronger policy robustness.

Evaluation

Model Training We conduct five different training runs for each approach, using different random seeds and 1000 episodes, in the stochastic mixed traffic flows with the normal density. The max time step of each episode is equal to 200.

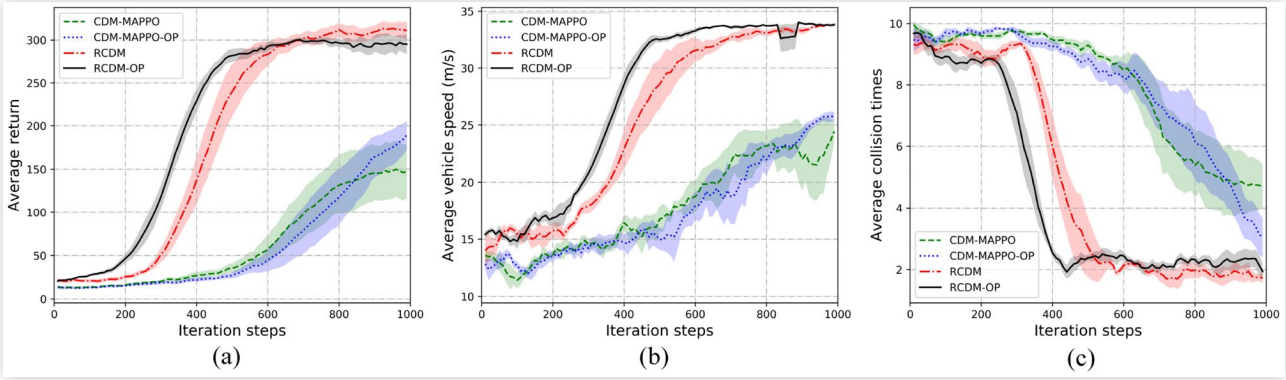
The training performance of each method in the highway scenario with the traffic flows based on the normal density is presented in Figure 3. Table 2 provides the final performance of the different methods, where the best result in each column is shown in bold. All the approaches are assessed on five trials with different random seeds. The solid curve represents the mean, and the shaded region denotes the standard deviation.

It can be seen from Figure 3 and Table 2 that our RCDM and RCDM-OP schemes outperform the baselines by a large margin, both in terms of the learning efficiency and the final performance. For instance, compared with CDM-MAPPO and CDM-MAPPO-OP, RCDM gains 125.58% and 79.23% improvements concerning the final return, respectively. Moreover, the final running speed and collision times of our RCDM and RCDM-OP schemes are superior to ones of CDM-MAPPO and CDM-MAPPO-OP approaches.

It can be found that, compared with CDM-MAPPO, CDM-MAPPO-OP can achieve better the final performance. Therefore, stochastic observational perturbations are beneficial for improving the final performance in the training phase

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

FIGURE 3 Training curves of the CDM-MAPPO, CDM-MAPPO-OP, RCDM, and RCDM-OP approaches in the normal-density traffic flows.



of the agents. One possible explanation is that incorporating observational perturbations may improve the exploration capability of the CDM-MAPPO agents during training.

However, it is obvious that stochastic observational perturbations reduce the learning efficiency of the CDM-MAPPO method. Moreover, RCDM outperforms RCDM-OP in terms of the final return and collision times. RCDM performs comparably to RCDM-OP in the final running speed.

Additionally, the average time consumption of our model for each update is about 3.00×10^{-2} s.

Model Testing We test the final models trained via each algorithm under five random seeds. The average return, vehicle speed, and collision times across every 10 episodes are leveraged to assess the automated driving agents' comprehensive performance, travel efficiency, and driving safety, respectively. Each policy model is tested for 200 episodes, and the maximum time step of each episode is 200.

The test results in Figure 4 and Table 3 demonstrate that our RCDM policy outperforms the LLDM-SAC and CDM-MAPPO policies by a large margin, both in terms of the return, speed, and safety, in the traffic flows with the low and high densities. Additionally, in the testing cases with observational perturbations, the RCDM-OP policy exceeds the performance of the LLDM-SAC-OP and CDM-MAPPO-OP policies consistently in terms of return, speed, and safety. Obviously, in contrast to the LLDM-SAC and CDM-MAPPO policies, the RCDM policy shows better robustness against observational perturbations. Additionally, our approach demonstrates consistent performance in highway scenarios with two distinct traffic densities.

More specifically, as shown in Table 3, compared with the CDM-MAPPO policy, our RCDM policy gains 59.16% and 47.08% improvements in the matter of the return and driving speed in the stochastic traffic flow with low density, respectively. Under observational perturbations, the return and travel efficiency of the RCDM-based automated vehicle are deteriorated by about 5.99% and 5.47%. The average collision times of every 10 testing episodes are increased from 0.00 to 0.11. Furthermore, compared with the CDM-MAPPO policy, the return and transport efficiency of the automated vehicle based on the CDM-MAPPO-OP policy is degenerated by about 49.43% and 14.63%. The average collision times are dramatically increased from 0.00 to 3.00. In comparison with the LLDM-SAC policy, the average return of the automated driving agent based on our RCDM policy is improved by about 34.73%. The RCDM policy performs comparably to the LLDM-SAC policy regarding transport efficiency. Unlike the LLDM-SAC-based automated vehicle, the one with the proposed RCDM policy does not cause any collisions in the stochastic traffic flows with the low density.

In the traffic flows with the high density, our method's advantages are more obvious compared to the baselines. It can be found from Figure 4 and Table 3 that the RCDM policy is superior to the LLDM-SAC and CDM-MAPPO policies in terms of all the metrics in the highway scenario with high traffic density. For example, in contrast to the CDM-MAPPO policy, the RCDM policy gains 48.78%, 36.36%, and 271.43% improvements in the matter of the return, speed, and safety, respectively. Compared with the LLDM-SAC policy, the return, driving speed, and collision times of the RCDM-based automated vehicle is improved by about 179.43%, 14.01%, and 3333.33%, respectively, in the traffic flows with the high density. Under observational perturbations, the average returns of the automated driving agents based on the LLDM-SAC, CDM-MAPPO, and RCDM policies are deteriorated by about 12.84%, 66.83%, and –2.30%, respectively. It meant our RCDM policy is least affected by observational perturbations.

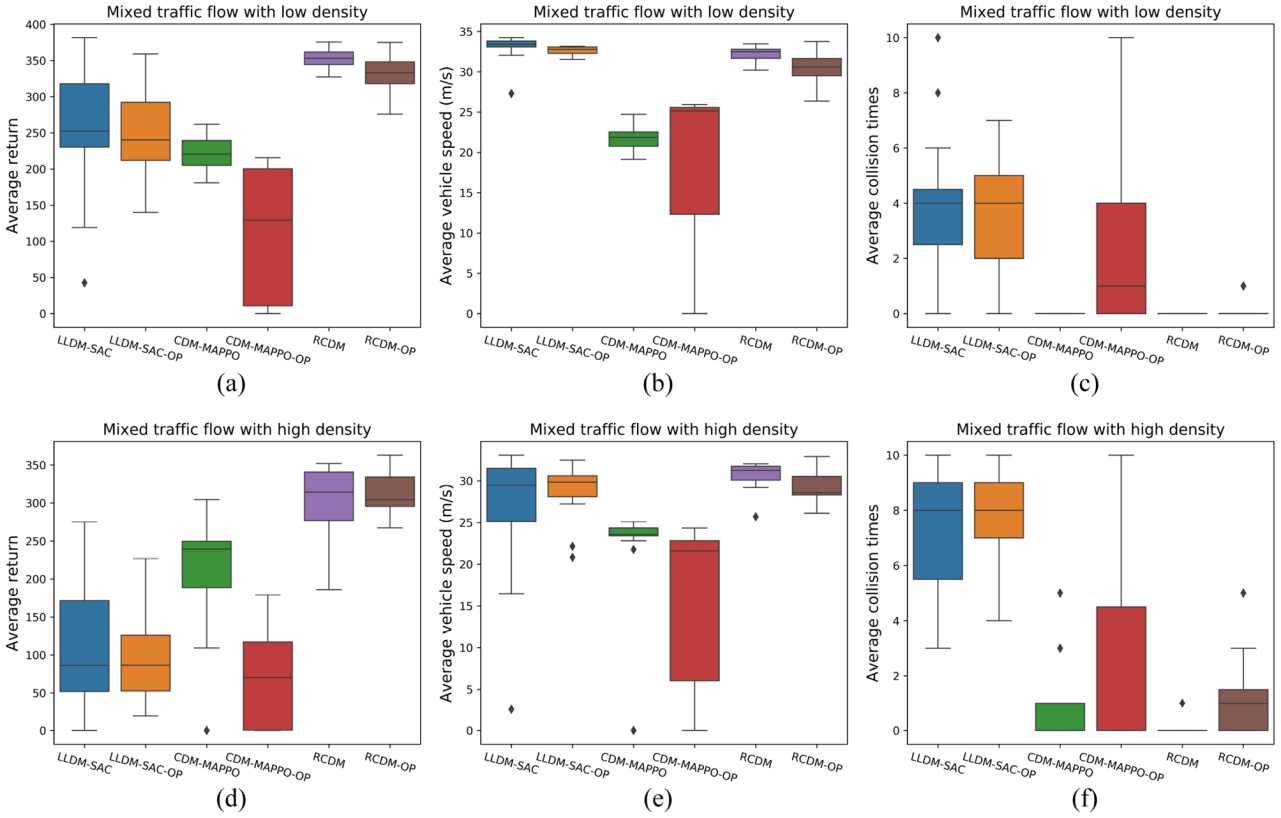
TABLE 2 Final performance for different methods in model training.

	Return	Speed	Collision times
CDM-MAPPO	148.19 ± 43.12	24.38 ± 1.51	5.00 ± 1.79
CDM-MAPPO-OP	186.52 ± 60.86	25.28 ± 1.13	3.40 ± 2.42
RCDM	**334.28 ± 52.10**	33.93 ± 0.50	**1.40 ± 1.50**
RCDM-OP	324.49 ± 28.33	**34.34 ± 0.07**	1.60 ± 0.80

© Nanyang Technological University

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

FIGURE 4 Evaluation of different policy models in the traffic flows based on the low and high densities. Here the observational perturbation obeys Gaussian distribution. (a)–(c): Average return, speed, and collision times of the different automated driving agents in the traffic flow with the low density; (d)–(f): Average return, speed, and collision times of the different automated driving agents in the traffic flow with the high density.



© Nanyang Technological University

To evaluate the impact of different types of noise on the policy model, we adopt the normal-density traffic flows to assess the performance of RL agents in the absence of noises, Gaussian noises, and Laplacian noises, respectively. We adopt the standard normal distribution to generate Gaussian noises. The Laplacian noise obeys the Laplace distribution [47]. Here the location parameter μ in Laplace(μ, b) is set to 0.00, and the scale parameter b in Laplace(μ, b) is equal to 1.00.

Figure 5 and Table 4 show the evaluation results of automated driving agents under different noise situations. In general, the results indicate that our RCDM automated driving agent surpasses the baselines via a large margin in the average return, vehicle speed, collision times, and robustness. For example, in the absence of noise, compared with the LLDM-SAC and CDM-MAPPO agents, the RCDM agent gains approximately 98.48% and 148.00% improvements concerning average return, respectively, the average vehicle speed of the RCDM agent is enhanced by about 10.01% and 62.17%, respectively, and the average collision times of the RCDM agent are decreased by about 89.79% and 43.27%, respectively.

Additionally, in general, compared with the baselines, the performance of the RCDM agent is minimally affected by the Gaussian and Laplacian noises. For instance, based on the results in Table 4, in contrast to the case without noises, the average returns of the LLDM-SAC agent disturbed by Gaussian noises and Laplacian noises are changed by about 8.40% and

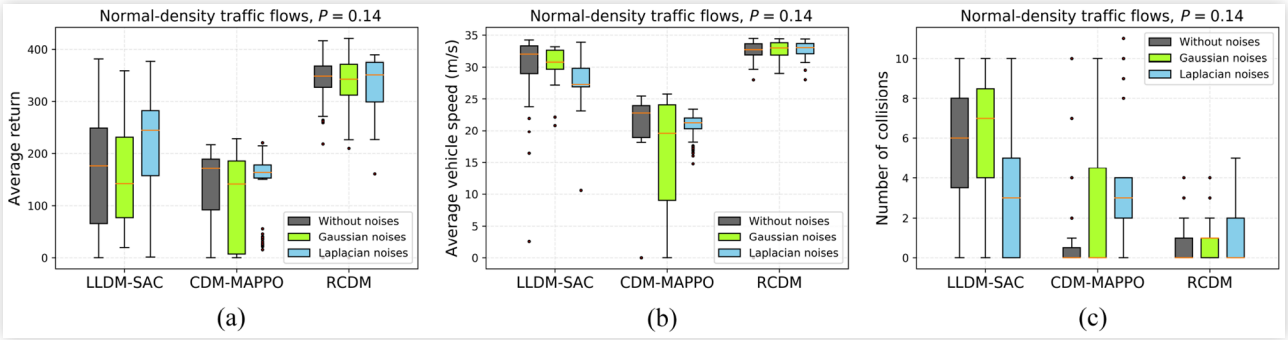
TABLE 3 Evaluation of the policy models trained via different methods in the traffic flows with low and high densities.

Environment	Metric	LLDM-SAC	LLDM-SAC-OP	CDM-MAPPO	CDM-MAPPO-OP	RCDM	RCDM-OP
Low density	Return	262.08 ± 85.08	247.45 ± 52.56	221.95 ± 22.42	112.24 ± 85.10	**353.26 ± 12.24**	332.09 ± 24.82
	Speed	**33.10 ± 1.47**	32.63 ± 0.50	21.88 ± 1.48	18.68 ± 11.17	32.18 ± 0.83	30.42 ± 1.73
	Collision times	3.63 ± 2.52	3.63 ± 1.69	**0.00 ± 0.00**	3.00 ± 3.70	**0.00 ± 0.00**	0.11 ± 0.31
High density	Return	109.49 ± 81.85	95.37 ± 52.91	205.93 ± 70.36	68.80 ± 60.84	305.72 ± 42.73	**312.16 ± 27.00**
	Speed	26.91 ± 7.29	28.84 ± 2.85	22.55 ± 5.37	15.85 ± 9.41	**30.68 ± 1.50**	29.29 ± 1.80
	Collision times	7.21 ± 2.31	7.90 ± 1.48	0.78 ± 1.36	2.21 ± 2.84	**0.21 ± 0.41**	1.05 ± 1.28

© Nanyang Technological University

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

FIGURE 5 Performance of automated driving agents in the normal-density traffic flows under different noise situations.



© Nanyang Technological University

27.16%, respectively. In comparison with the case without noises, the average returns of the CDM-MAPPO agent disturbed by Gaussian noises and Laplacian noises are changed by about 16.50% and 7.31%, respectively. In contrast to the case without noises, the average returns of the RCDM agent disturbed by Gaussian noises and Laplacian noises are changed by about 1.60% and 2.22%, respectively. Meanwhile, compared with the case without noises, the average vehicle speed of the LLDM-SAC, CDM-MAPPO, and RCDM agents disturbed by Gaussian noises and Laplacian noises are changed by about 3.30% and 5.55%, 18.87% and 3.66%, and 0.43% and 1.27%, respectively. Furthermore, in comparison with the case without noises, the collision times of the LLDM-SAC agent disturbed by Gaussian noises and Laplacian noises are changed by about 7.09% and 46.71%, respectively. Compared with the case without noises, the collision times of the CDM-MAPPO agent disturbed by Gaussian noises and Laplacian noises are changed by about 170.19% and 290.39%, respectively. In contrast to the case without noises, the collision times of the RCDM agent disturbed by Gaussian noises and Laplacian noises are changed by about 61.02% and 69.49%, respectively.

With Equation 32, we compute the policy robustness metric of each agent under Gaussian noises and Laplacian noises. Specifically, under Gaussian noises, in contrast to the LLDM-SAC and CDM-MAPPO agents, the policy robustness of the RCDM agent is enhanced by about 50.82% and 58.90%, respectively. Moreover, under Laplacian noises, compared with the LLDM-SAC and CDM-MAPPO agents, the RCDM agent gains approximately 73.29% and 77.70% improvements concerning robustness, respectively. Hence, it is clear that our technique enables the automated driving agent to learn a more robust policy compared to the baselines.

In addition, the average time consumption of our model for each inference is approximately 5.00×10^{-4} s.

Conclusion

In this work, we aim to develop the RCDM technique that enables an automated vehicle to coordinate its longitudinal and lateral decision-making systems while ensuring the driving policy robustness against observational uncertainties. Specifically, CRMA-MDP is advanced to model the longitudinal and lateral driving behaviors of the automated vehicle under observational uncertainties. Meanwhile, the nonlinear constraint setting based on KL divergence is developed to keep the variation of the driving policy perturbed by stochastic observational perturbations within bounds. In addition, the RMAPO algorithm is presented to approximate the optimal robust coordinated driving policy.

The results in three traffic flows with different densities demonstrate that our method is able to effectively coordinate the longitudinal and lateral decision behaviors

TABLE 4 Statistical results of automated driving agents in the normal-density traffic flows under different noise situations. RE: Return; SP: Speed; NC: Number of Collisions; RO: Robustness.

	LLDM-SAC			CDM-MAPPO			RCDM		
Metric	**Without noises**	**Gaussian noises**	**Laplacian noises**	**Without noises**	**Gaussian noises**	**Laplacian noises**	**Without noises**	**Gaussian noises**	**Laplacian noises**
RE	170.52 ± 111.82	156.20 ± 91.30	216.84 ± 86.87	136.47 ± 67.72	113.96 ± 81.17	146.44 ± 59.53	**338.45 ± 50.51**	333.02 ± 54.12	330.95 ± 49.94
SP	29.38 ± 6.48	30.35 ± 2.90	27.75 ± 4.44	19.93 ± 6.85	16.17 ± 9.69	20.66 ± 2.04	32.32 ± 3.55	32.46 ± 3.30	**32.73 ± 1.25**
NC	5.78 ± 2.97	6.19 ± 2.61	3.08 ± 2.79	1.04 ± 2.31	2.81 ± 4.05	4.06 ± 3.02	**0.59 ± 0.92**	0.95 ± 1.38	1.00 ± 1.38
RO ($\times 10^{-3}$)	N/A	1.83 ± 0.72	34.70 ± 9.22	N/A	2.19 ± 0.04	41.56 ± 3.53	N/A	**0.90 ± 0.13**	9.27 ± 0.61

© Nanyang Technological University

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

of an automated vehicle and shows better performance in comparison with baselines. Furthermore, the RCDM policy models have superior generalization to unseen situations and robustness against perturbations on observations.

Here, we present potential enhancements and future work to the proposed approach: (1) taking into account the influence of low-level controllers on energy efficiency, such as integrating steering and braking control systems. (2) Performing more comprehensive simulations that encompass various traffic elements. For instance, incorporating the status of traffic signals into the agent's state and reward function. (3) Exploring alternative neural network architectures, such as the Transformer, improves the generalization of models. (4) Evaluating the models trained by the proposed method using a real vehicle equipped with an edge computing system (e.g., Jetson Xavier NX 16 GB).

Acknowledgement

This work was supported by Strat-Up Grant of Nanyang Technological University.

Contact Information

Chen Lv
corresponding author
lyuchen@ntu.edu.sg

References

1. Lai, F., Huang, C., Jiang, C., and Zhang, Y., "Simulation Analysis of Automatic Emergency Braking System under Constant Steer Conditions," *SAE Int. J. Veh. Dyn., Stab., and NVH* 6(4):461-476, 2022, doi:https://doi.org/10.4271/10-06-04-0030.
2. Ren, Y., Jiang, J., Zhan, G., Li, S.E. et al., "Self-Learned Intelligence for Integrated Decision and Control of Automated Vehicles at Signalized Intersections," *IEEE Transactions on Intelligent Transportation Systems* 23(12):24145-24156, 2022.
3. Wang, Y., Wei, H., Hu, B., and Lv, C., "Robust Estimation of Vehicle Dynamic State Using a Novel Second-Order Fault-Tolerant Extended Kalman Filter," *SAE Int. J. Veh. Dyn., Stab., and NVH* 7(3), 2023, doi:https://doi.org/10.4271/10-07-03-0019.
4. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A. et al., "Human-Level Control through Deep Reinforcement Learning," *Nature* 518(7540):529-533, 2015.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J. et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30:1-11, 2017.
6. Zhao, Q., Zheng, H., Kaku, C., Cheng, F., and Zong, C., "Safety Spacing Control of Truck Platoon Based on Emergency Braking under Different Road Conditions," *SAE Int. J. Veh. Dyn., Stab., and NVH* 7(1):69-81, 2023, doi:https://doi.org/10.4271/10-07-01-0005.
7. Cao, Z., Xu, S., Peng, H., Yang, D., and Zidek, R., "Confidence-Aware Reinforcement Learning for Self-Driving Cars," *IEEE Transactions on Intelligent Transportation Systems* 23(7):7419-7430, 2022.
8. Gupta, U., Nouri, A., Subramanian, C., Taheri, S. et al., "Developing an Experimental Setup for Real-Time Road Surface Identification Using Intelligent Tires," *SAE Int. J. Veh. Dyn., Stab., and NVH* 5(3):351-367, 2021, doi:https://doi.org/10.4271/10-05-03-0024.
9. Ji, X., He, X., Lv, C., Liu, Y., and Wu, J., "Adaptive-Neural-Network-Based Robust Lateral Motion Control for Autonomous Vehicle at Driving Limits," *Control Engineering Practice* 76:41-53, 2018.
10. Peng, H. and Chen, X., "Active Safety Control of X-by-Wire Electric Vehicles: A Survey," *SAE Int. J. Veh. Dyn., Stab., and NVH* 6(2):115-133, 2022, doi:https://doi.org/10.4271/10-06-02-0008.
11. Peng, J., Zhang, S., Zhou, Y., and Li, Z., "An Integrated Model for Autonomous Speed and Lane Change Decision-Making Based on Deep Reinforcement Learning," *IEEE Transactions on Intelligent Transportation Systems* 23(11):21848-21860, 2022.
12. Wang, Y., Wei, H., Hu, B., and Lv, C., "A Review of Dynamic State Estimation of the Neighborhood System for Connected Vehicles," *SAE Int. J. Veh. Dyn., Stab., and NVH* 7(3), 2023, doi:https://doi.org/10.4271/10-07-03-0023.
13. Negash, N.M. and Yang, J., "Anticipation-Based Autonomous Platoon Control Strategy with Minimum Parameter Learning Adaptive Radial Basis Function Neural Network Sliding Mode Control," *SAE Int. J. Veh. Dyn., Stab., and NVH* 6(3):247-265, 2022, doi:https://doi.org/10.4271/10-06-03-0017.
14. Wu, J., Zhang, J., Nie, B., Liu, Y., and He, X., "Adaptive Control of PMSM Servo System for Steering-by-Wire System with Disturbances Observation," *IEEE Transactions on Transportation Electrification* 8(2):2015-2028, 2021.
15. Schwarting, W., Alonso-Mora, J., and Rus, D., "Planning and Decision-Making for Autonomous Vehicles," *Annual Review of Control, Robotics, and Autonomous Systems* 1:187-210, 2018.
16. Urmson, C., Anhalt, J., Bagnell, D., Baker, C. et al., "Autonomous Driving in Urban Environments: Boss and the Urban Challenge," *Journal of Field Robotics* 25(8):425-466, 2008.
17. Montemerlo, M., Becker, J., Bhat, S., Dahlkamp, H. et al., "Junior: The Stanford Entry in the Urban Challenge," *Journal of Field Robotics* 25(9):569-597, 2008.
18. Kurt, A. and Özgüner, Ü., "Hierarchical Finite State Machines for Autonomous Mobile Systems," *Control Engineering Practice* 21(2):184-194, 2013.

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

19. Sales, D.O., Correa, D.O., Fernandes, L.C., Wolf, D.F., and Osório, F.S., "Adaptive Finite State Machine Based Visual Autonomous Navigation System," *Engineering Applications of Artificial Intelligence* 29:152-162, 2014.
20. Hülnhagen, T., Dengler, I., Tamke, A., Dang, T. et al., "Maneuver Recognition Using Probabilistic Finite-State Machines and Fuzzy Logic," *2010 IEEE Intelligent Vehicles Symposium*, La Jolla, CA, 65-70, 2010, IEEE.
21. Li, N., Oyler, D.W., Zhang, M., Yildiz, Y. et al., "Game Theoretic Modeling of Driver and Vehicle Interactions for Verification and Validation of Autonomous Vehicle Control Systems," *IEEE Transactions on Control Systems Technology* 26(5):1782-1797, 2017.
22. Hang, P., Lv, C., Xing, Y., Huang, C., and Hu, Z., "Human-Like Decision Making for Autonomous Driving: A Noncooperative Game Theoretic Approach," *IEEE Transactions on Intelligent Transportation Systems* 22(4):2076-2087, 2020.
23. Liu, W., Kim, S.-W., Pendleton, S., and Ang, M. H., "Situation-Aware Decision Making for Autonomous Driving on Urban Road Using Online POMDP," *2015 IEEE Intelligent Vehicles Symposium (IV)*, Seoul, Korea, 1126-1133, 2015, IEEE.
24. Bai, H., Cai, S., Ye, N., Hsu, D. et al., "Intention-Aware Online POMDP Planning for Autonomous Driving in a Crowd," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, 454-460, 2015. IEEE.
25. Le Mero, L., Yi, D., Dianati, M., and Mouzakitis, A., "A Survey on Imitation Learning Techniques for End-to-End Autonomous Vehicles," *IEEE Transactions on Intelligent Transportation Systems* 23(9):14128-14147, 2022.
26. Chen, L., He, Y., Wang, Q., Pan, W., and Ming, Z., "Joint Optimization of Sensing, Decision-Making and Motion-Controlling for Autonomous Vehicles: A Deep Reinforcement Learning Approach," *IEEE Transactions on Vehicular Technology* 71(5):4642-4654, 2022.
27. Kuefler, A., Morton, J., Wheeler, T., and Kochenderfer, M., "Imitating Driver Behavior with Generative Adversarial Networks," *2017 IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, 204-211, 2017, IEEE.
28. Ngai, D.C.K. and Yung, N.H.C., "A Multiple-Goal Reinforcement Learning Method for Complex Vehicle Overtaking Maneuvers," *IEEE Transactions on Intelligent Transportation Systems* 12(2):509-522, 2011.
29. Chen, J., Wang, Z., and Tomizuka, M., "Deep Hierarchical Reinforcement Learning for Autonomous Driving with Distinct Behaviors," *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, China, 1239-1244, 2018, IEEE.
30. Everett, M., Chen, Y. F., and How, J. P., "Motion Planning among Dynamic, Decision-Making Agents with Deep Reinforcement Learning," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 3052-3059, 2018, IEEE.
31. Xu, X., Zuo, L., Li, X., Qian, L. et al., "A Reinforcement Learning Approach to Autonomous Decision Making of Intelligent Vehicles on Highways," *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 50(10):3884-3897, 2018.
32. You, C., Lu, J., Filev, D., and Tsiotras, P., "Advanced Planning for Autonomous Vehicles Using Reinforcement Learning and Deep Inverse Reinforcement Learning," *Robotics and Autonomous Systems* 114:1-18, 2019.
33. Mozaffari, S., Arnold, E., Dianati, M., and Fallah, S., "Early Lane Change Prediction for Automated Driving Systems Using Multi-Task Attention-Based Convolutional Neural Networks," *IEEE Transactions on Intelligent Vehicles* 7(3):758-770, 2022.
34. Szegedy, C., Toshev, A., and Erhan, D., "Deep Neural Networks for Object Detection," *Advances in Neural Information Processing Systems* 26:1-9, 2013.
35. Hu, Z., Xing, Y., Gu, W., Cao, D., and Lv, C., "Driver Anomaly Quantification for Intelligent Vehicles: A Contrastive Learning Approach with Representation Clustering," *IEEE Transactions on Intelligent Vehicles* 8(1):37-47, 2023.
36. Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., and López, A.M., "Multimodal End-to-End Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems* 23(1):537-547, 2022.
37. He, X., Yang, H., Hu, Z., and Lv, C., "Robust Lane Change Decision Making for Autonomous Vehicles: An Observation Adversarial Reinforcement Learning Approach," *IEEE Transactions on Intelligent Vehicles* 8(1):184-193, 2023.
38. Zhang, J., Chang, C., Zeng, X., and Li, L., "Multi-Agent DRL-Based Lane Change with Right-of-Way Collaboration Awareness," *IEEE Transactions on Intelligent Transportation Systems* 24(1):854-869, 2023.
39. Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J. et al., "Microscopic Traffic Simulation Using Sumo," *The 21st IEEE International Conference on Intelligent Transportation Systems*, Maui, HI, 2018, IEEE.
40. Boyd, S., Boyd, S.P., and Vandenberghe, L., *Convex Optimization* (Cambridge: Cambridge University Press, 2004).
41. Crosato, L., Shum, H.P.H., Ho, E.S.L., and Wei, C., "Interaction-Aware Decision-Making for Automated Vehicles Using Social Value Orientation," *IEEE Transactions on Intelligent Vehicles* 8(2):1339-1349, 2022.
42. He, X., Liu, Y., Lv, C., Ji, X., and Liu, Y., "Emergency Steering Control of Autonomous Vehicle for Collision Avoidance and Stabilisation," *Vehicle System Dynamics* 57(8):1163-1187, 2019.
43. He, X., Lou, B., Yang, H., and Lv, C., "Robust Decision Making for Autonomous Vehicles at Highway On-Ramps: A Constrained Adversarial Reinforcement Learning Approach," *IEEE Transactions on Intelligent Transportation Systems* 24(4):4103-4113, 2022.
44. Rajamani, R., *Vehicle Dynamics and Control* (New York: Springer Science & Business Media, 2011).

Downloaded from SAE International by Nanyang Tech University, Wednesday, February 07, 2024

45. Yu, C., Velu, A., Vinitsky, E., Gao, J. et al., "The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games," *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, New Orleans, Louisiana, USA, 2022.

46. Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *International Conference on Machine Learning*, Stockholm, Sweden, 1861-1870, 2018, PMLR.

47. Yuan, W., Zhuang, H., Wang, C., and Yang, M., "AGBM: An Adaptive Gradient Balanced Mechanism for the End-to-End Steering Estimation," *IEEE Transactions on Intelligent Transportation Systems* 23(9):16016-16025, 2022.

© 2023 Nanyang Technological University; Published by SAE International. This Open Access article is published under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits distribution, and reproduction in any medium, provided that the original author(s) and the source are credited.

Positions and opinions advanced in this work are those of the author(s) and not necessarily those of SAE International. Responsibility for the content of the work lies solely with the author(s).