Jerk-minimized Autonomous Driving Strategy with Deep Reinforcement Learning

Jaehwi Lee, Chanin Eom, Dongsu Lee, Minhae Kwon

***Abstract*—With the increasing focus on autonomous driving research, road environments have evolved into new types of multi-agent systems. Deep Reinforcement Learning (RL) has been regarded as a potential solution for enabling successful decision-making strategies in autonomous vehicles. However, conventional RL-based approaches struggle to provide ride comfort, which is important for practicality. This challenge arises from the conventional reward design, which incurs a high reward value at the fixed target (e.g., achieving target speed). The agent with such a reward design is trained to quickly achieve this target because the objective of RL is maximizing reward. It leads to the driving strategy with rapid acceleration or deceleration, which can reduce ride comfort. In this study, we aim to develop an autonomous driving strategy related to ride comfort. To achieve this, we employ a reward component based on jerk, which is proportional to the differences in acceleration. This reward component can encourage the agent to adopt a smooth driving strategy by penalizing rapid acceleration or deceleration. The simulation results demonstrate the proposed reward component can effectively reduce the jerk regardless of the driving scenarios.**

I. INTRODUCTION

Autonomous driving has been a remarkable improvement as the emerging transportation system. This achievement has evolved into a new type of multi-agent system, in which the interaction between human vehicles and autonomous vehicles is frequent. In this trend, autonomous vehicles are required to provide a successful decision-making strategy while ensuring high ride comfort.

Enhancing the decision-making performance of autonomous vehicles has been widely studied [1], [2]. One potential approach is deep RL because it can find optimal behavior through interactions with the environment. Additionally, deep RL effectively enhances the scalability of the autonomous driving strategy by leveraging the generalization ability of deep neural network [3]. From these perspectives, numerous studies have attempted to make an RL-based autonomous driving strategy in various traffic scenarios [4], [5].

Despite the achievement of the deep RL-based approach, most conventional autonomous driving strategies face challenges in providing ride comfort. This hurdle stems from the reward design of the conventional approach. Specifically, many studies utilize high rewards when autonomous vehicles achieve certain target speeds [6]–[8] or impose high penalties when encountering unsafe environments [9]–[11]. Since the primary objective of RL is to maximize reward, autonomous vehicles designed with such reward systems strive to reach their targets as rapidly as possible. Conversely, these vehicles are trained to frequently perform high acceleration or deceleration, which can disrupt ride comfort.

Employing the reward term designed to minimize jerk value can offer a simple yet effective solution for improving ride comfort. This is because the jerk value is proportional to the magnitude of acceleration differences, which is related to ride comfort [12]. From this perspective, some studies incorporate the jerk-related reward component during RL-based training [13], [14]. However, such studies only take into account the specific driving scenario rather than providing a comprehensive analysis. The effectiveness of employing jerk penalties should be evaluated across various driving scenarios because ride comfort is always important regardless of the road environment.

In this study, we aim to develop an RL-based autonomous driving strategy to minimize jerk. To achieve this, we incorporate the reward component for jerk minimization. In addition, we employ the considerably designed POMDP setting, which can be utilized across various road scenarios. By leveraging this POMDP setting, we provide a comprehensive analysis of the proposed reward design in various autonomous driving scenarios.

II. PROPOSED SOLUTION

The main objective of this work is to develop a jerk-minimized autonomous driving strategy. In this section, we provide a carefully designed POMDP model to achieve it. A POMDP can be defined as a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \Omega, \gamma >$, which includes states $s_t \in \mathcal{S}$, actions $a_t \in \mathcal{A}$, state transition probabilities $\mathcal{T}(s_{t+1}|s_t, a_t)$, rewards $\mathcal{R}(s_t, a_t, s_{t+1})$, observations $o_t \in \mathcal{O}$, observation probabilities $\Omega(o_t|s_{t+1})$, and a discount factor $\gamma \in [0, 1)$.

A. Driving Scenarios

In this subsection, we provide details related to the road scenario. We take into account the following three driving scenarios, in which unsuccessful decision-making of the agent can lead to high-jerk value. The illustration of each driving scenario is provided in Figure 1.

Highway: This scenario comprises numerous vehicles with varying target speeds. This diversity necessitates the agent to frequently accelerate or decelerate to adaptively maneuver its driving route, potentially resulting in high jerks.

Cut-in: In this scenario, the agent has a higher target speed than other vehicles on the road. Therefore, the agent should

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00278812).

All authors are with the Department of Intelligent Semiconductors and Minhae Kwon is with School of Electronic Engineering, Soongsil University, Seoul 06978, South Korea (e-mail: dlwognl99@soongsil.ac.kr, eci0623@soongsil.ac.kr, movementwater@soongsil.ac.kr, minhae@ssu.ac.kr) (*Corresponding author: Minhae Kwon*).

IEEE ICRA (International Conference on Robotics and Automation) Workshop on MAD-GAMES: Multi-agent Dynamic Games, Yokohama, Kanagawa, Japan. 2024.

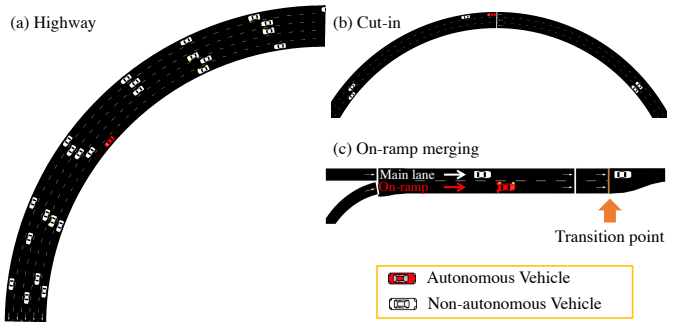

Figure 1. Illustrative examples of road driving scenario (a) Highway, (b) Cut-in, (c) On-ramp merging

overtake to achieve its target speed. The overtaking action can cause high jerks.

On-ramp merging: In this scenario, the agent passes through the on-ramp lane, which merges with the main lanes after a certain distance. The agent should adapt its speed to determine the merging timing without experiencing high jerks and avoiding collisions.

B. Road State Model

The road has M transition points and N vehicles. Let $\mathcal{M} = \{1, 2, \ldots, M\}$ represent the set of transition points, and $C = C_{NAV} \cup C_{AV}$ represent the set of vehicles driving on the road. There are $N-1$ non-autonomous vehicles $C_{NAV} = \{c_i \mid i \neq N\}$ and an autonomous vehicle $C_{AV} = \{c_i \mid i = N\}$.

In this situation, the state $s_t \in \mathcal{S}$ can be defined as follows.

$$s_t = [\mathbf{v}_t{}^T, \mathbf{p}_t^T, \mathbf{k}_t^T, \mathbf{d}_t^T]^T \tag{1}$$

In (1), $\mathbf{v}_t = [v_{t,1}, \cdots, v_{t,i}, \cdots, v_{t,N}]^T$ and $\mathbf{p}_t = [p_{t,1}, \cdots, p_{t,i}, \cdots, p_{t,N}]^T$ represent the absolute speed and absolute position vectors, respectively. A lane number vector of vehicles is represented by $\mathbf{k}_t = [k_{t,1}, \cdots, k_{t,i}, \cdots, k_{t,N}]^T$. Finally, $\mathbf{d}_t = [d_{t,1}, \cdots, d_{t,i}, \cdots, d_{t,N}]^T$ denotes the distance to the nearest front transition point of each vehicle.

C. Partially Observable Markov Decision Process

In this study, the agent can observe across the H lane within a front/rear distance V and make decisions based on these partial observations. Specifically, the agent can only access the driving information within its observable area, which is represented as the green shaded region in Fig. 2 (a). Based on this observable area, the observable vehicle set $C_{t,obs}$ always satisfies the following conditions.

$$C_{t,obs} = \begin{cases} |p_{t,i} - p_{t,N}| \leq V \\ |k_{t,i} - k_{t,N}| \leq \frac{H-1}{2} \end{cases}$$

The observable vehicle set $C_{t,obs}$ can be defined as the set of front vehicles L_t and the set of rear vehicles F_t as follows.

$$C_{t,obs} = L_t \cup F_t,$$
$$where\ L_t = \bigcup_{h=1}^{H} L_{t,h}, \quad F_t = \bigcup_{h=1}^{H} F_{t,h} \tag{2}$$

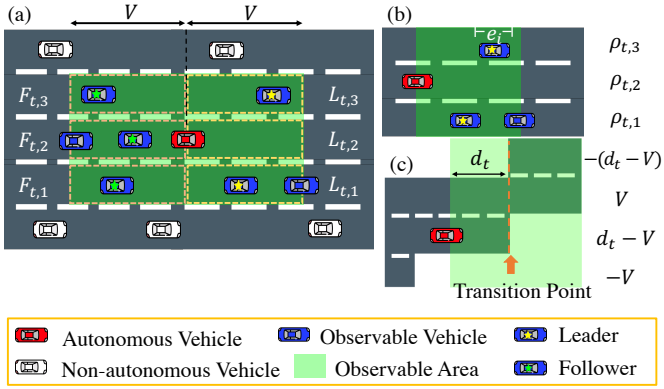

Figure 2. Illustrative examples of the observation space (a) observable area of an autonomous vehicle, (b) traffic density per lane and (c) lane existence

In (2), $L_{t,h} \subset L_t$ and $F_{t,h} \subset F_t$ denote the set of front and rear vehicles within the observable area for each lane h. The nearest vehicle to the autonomous vehicle per front and rear lane at time t is defined as leader $l_{t,h} \in L_{t,h}$, and follower $f_{t,h} \in F_{t,h}$.

Observation: At time t, the observation information of the agent, denoted by $o_t \in \mathcal{O}$, is defined as follows.

$$o_t = [v_{t,N}, \Delta\mathbf{v}_t^T, \Delta\mathbf{p}_t^T, \boldsymbol{\rho}_t^T, \boldsymbol{\zeta}_t^T]^T$$

Herein, $v_{t,N}$ denotes the absolute speed of the agent. $\Delta\mathbf{v}_t = [\Delta v_{t,l_1}, \cdots, \Delta v_{t,l_h}, \cdots, \Delta v_{t,l_H}, \Delta v_{t,f_1}, \cdots, \Delta v_{t,f_h}, \cdots, \Delta v_{t,f_H}]^T$ and $\Delta\mathbf{p}_t = [\Delta p_{t,l_1}, \cdots, \Delta p_{t,l_h}, \cdots, \Delta p_{t,l_H}, \Delta p_{t,f_1}, \cdots, \Delta p_{t,f_h}, \cdots, \Delta p_{t,f_H}]^T$ denote the relative speed and relative distance of the agent to per lane leader and follower. $\boldsymbol{\rho}_t = [\rho_{t,1}, \cdots, \rho_{t,h}, \cdots, \rho_{t,H}]^T$ denotes the density of vehicles per lane in the agent's front observable area, defined as the percentage of vehicles occupying a lane h as follows.

$$\rho_{t,h} = \frac{\Sigma_{i=1}^{|L_{t,h}|}(e_i + \delta_0)}{V} \tag{3}$$

In (3), $|L_{t,h}|$ denotes the number of vehicles observed in the front h lane, e_i denotes the length of the i-th vehicle in lane h, and δ_0 denotes the minimum safe distance between vehicles. An illustration of lane density is shown in Fig. 2 (b). Lastly, $\boldsymbol{\zeta}_t = [\zeta_{t,1}, \cdots, \zeta_{t,h}, \cdots, \zeta_{t,H}]^T$ denotes the lane existence in the observable area in front of the agent. It is defined by the front observable distance V and the remaining distance d_t to the transition point at time t. In particular, the existence of lane $\zeta_{t,h}$ is determined by $d_t - V$ if the current lane merges after d_t, whereas if the lane splits after d_t, $\zeta_{t,h}$ becomes $-(d_t - V)$. If there is no transition point within the observable distance V, $\zeta_{t,h}$ becomes $-V$ or V: V if the lane is connected, and $-V$ if the lane is not connected. The detailed illustration of the lane existence observation is presented in Fig. 2 (c).

Action: At time t, the action of agent $a_t \in \mathcal{A}$ is defined as follows.

$$a_t = \{a_{t,acc}, a_{t,lc}\},$$

where $a_{t,acc}$ denotes acceleration control action, and $a_{t,lc}$ denotes lane change action. The acceleration control action $a_{t,acc} \in [a_{\min}, a_{\max}]$ is defined within the continuous range of minimum acceleration $a_{\min}$ and maximum acceleration $a_{\max}$. The lane change action $a_{t,lc} \in \mathcal{A}_{lc} = \{-1, 0, 1\}$ is defined as a discrete value, where each value represents the direction of lane change for the agent. Specifically, $a_{t,lc} = -1$ means a lane change to the right, $a_{t,lc} = 1$ is a lane change to the left, and $a_{t,lc} = 0$ is a maintaining the current lane.

Reward: The agent performs an action a_t in the current state s_t, and receives a reward $r_t = \mathcal{R}(s_t, a_t, s_{t+1})$. The reward function $\mathcal{R}(s_t, a_t, s_{t+1})$ is defined as follows.

$$\mathcal{R}(s_t, a_t, s_{t+1}) = \mathcal{R}_{t,jerk} + \mathcal{R}_{t,driving} + \mathcal{R}_{t,collision} \quad (4)$$

In (4), $\mathcal{R}_{jerk}$ represents a reward component associated with a jerk, $\mathcal{R}_{driving}$ evaluates how well the agent performs under the general driving situation, and $\mathcal{R}_{collision}$ is a component for penalizing an accident. In the remainder of this subsection, we provide details of each reward component.

1) Jerk Minimization: The jerk reward component $\mathcal{R}_{t,jerk}$ is defined as follows.

$$\mathcal{R}_{t,jerk} = -\eta_{jerk} \left| \frac{a_{t,acc} - a_{t',acc}}{\Delta t} \right|, \quad (5)$$

where $-\eta_{jerk}$ represents the coefficient of this component, and $\Delta t = t' - t$ is the timestep interval. Through this component, the agent incurs a higher penalty as the difference in acceleration (i.e., $|a_{t,acc} - a_{t',acc}|$) increases. This component can mitigate high jerks during driving because the difference in acceleration directly corresponds to the jerk.

2) General Driving Ability: $\mathcal{R}_{t,driving}$ is defined as a linear combination of reward terms, where each reward term evaluates the overall driving circumstance.

$$\mathcal{R}_{t,driving} = \Sigma_{i=1}^{5} \eta_i \mathcal{R}_{t,i} \quad (6)$$

In (6), η_i is coefficients determining the importance of each term. The first term, denoted as $\mathcal{R}_{t,1}$, the agent learns to approach the target speed v^* while ensuring it does not exceed the speed limit $\bar{v}$.

$$\mathcal{R}_{t,1} = \begin{cases} \frac{v_{t+1,N}}{v^*}, & v_{t+1,N} \leq v^* \\ \frac{\bar{v} - v_{t+1,N}}{\bar{v} - v^*}, & v_{t+1,N} > v^* \end{cases}$$

The agent receives the maximum reward when driving close to the target speed v^*. Conversely, the agent incurs a penalty if it exceeds the speed limit $\bar{v}$.

The second term $\mathcal{R}_{t,2}$ denotes the lane change penalty which is defined as follows.

$$\mathcal{R}_{t,2} = \begin{cases} -1, & |a_{t,lc}| = 1 \\ 0, & |a_{t,lc}| = 0 \end{cases} \quad (7)$$

This term gives a constant penalty for every lane change action of the agent, thereby discouraging the meaninglessly frequent lane changes.

Both $\mathcal{R}_{t,3}$ and $\mathcal{R}_{t,4}$ pertain to safe driving. $\mathcal{R}_{t,3}$ encourage the agent not to violate the safety distance to the same lane leader $\delta^*_{t+1,\hat{l}}$.

$$\mathcal{R}_{t,3} = \min \left[0, 1 - \left(\frac{\delta^*_{t+1,\hat{l}}}{\Delta p_{t+1,\hat{l}}} \right)^2 \right] \quad (8)$$

Similarly, $\mathcal{R}_{t,4}$ encourages the agent to avoid violating safety distances to the same lane follower $\delta^*_{t+1,\hat{f}}$ when changing lanes (i.e., $|a_{t,lc}| = 1$).

$$\mathcal{R}_{t,4} = |a_{t,lc}| \min \left[0, 1 - \left(\frac{\delta^*_{t+1,\hat{f}}}{\Delta p_{t+1,\hat{f}}} \right)^2 \right], \quad (9)$$

In (8), (9), each safety distance $\delta^*_{t+1,\hat{l}}$, $\delta^*_{t+1,\hat{f}}$ is calculated by the intelligent driver model [15].

The last term $\mathcal{R}_{t,5}$, is intended to mitigate delayed merges from the ramp lane to the main lane during the on-ramp merging scenario. This term is defined based on the observation information, ζ_t, and is as follows.

$$\mathcal{R}_{t,5} = \begin{cases} \zeta_{t+1,\hat{h}}, & \zeta_{t+1,\hat{h}} < 0 \\ 0, & \zeta_{t+1,\hat{h}} \geq 0 \end{cases} \quad (10)$$

In (10), $\zeta_{t+1,\hat{h}}$ denotes the existence of a lane within the observable area in front of the agent, corresponding to the lane it is driving in. When the road in front of the observable area does not exist or disconnects (i.e., $\zeta_{t+1,\hat{h}} < 0$), the agent receives a higher penalty for driving closer to the transition point of the ramp lane.

3) Collision Avoidance: Lastly, the agent is penalized, denoted $\mathcal{R}_{t,collision}$, for driving action that results in a collision which is defined as follows.

$$\mathcal{R}_{t,collision} = \begin{cases} -\eta_{collision}, & \text{Collision} \\ 0, & \text{Otherwise} \end{cases} \quad (11)$$

It is noteworthy that $\eta_{collision}$ is the maximum value among the other coefficients, indicating that the agent incurs the highest penalty for a collision.

III. SIMULATION RESULTS

In this section, we analyze the driving strategies of autonomous vehicles in various environments based on jerk. First, we outline the simulation setup and then provide details regarding the baselines and a performance metric. Finally, we present a comparison of driving performance across different scenarios.

A. Simulation Setup

In this study, the simulation setting is performed with the FLOW framework, which is built upon the SUMO traffic control simulator [16]. The observable distance is set to $V = 30m$ units, with $H = 3$ observable lanes. The target speed of the agent is $v^* = 49.27km/h$, while the road speed limit is $\bar{v} = 116km/h$. Maximum and minimum accelerations are defined as $a_{\max} = 5.4m/s^2$ and $a_{\min} = -5.4m/s^2$, respectively. Regarding training settings, the agent undergoes

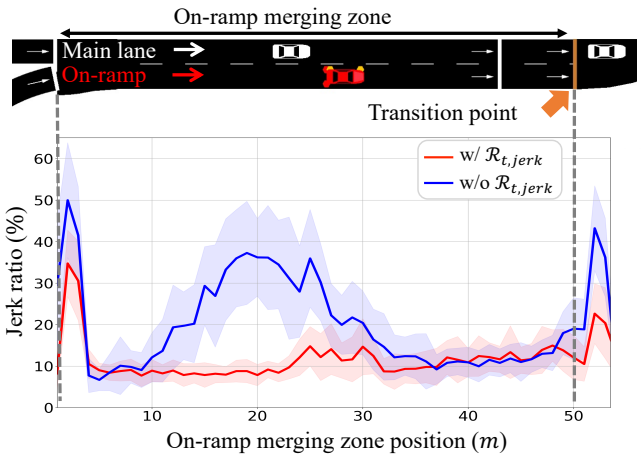

Figure 3. Jerk ratio by on-ramp merging zone position

Table I: Average jerk ratio across the driving scenarios

Scenario	w/ $\mathcal{R}_{t,jerk}$	w/o $\mathcal{R}_{t,jerk}$
Highway	8.816 ± 0.178	14.175 ± 1.624
Cut-in	9.014 ± 0.224	10.973 ± 0.255
On-ramp merging	10.354 ± 0.413	11.567 ± 0.493

700 episodes, each comprising $3000ts$, where $1ts$ corresponds to $0.1second$. To train the agent, we employ the Deep Deterministic Policy Gradient (DDPG) algorithm [17].

B. Baselines and Comparison Metric

In this subsection, we provide the baseline and comparison metric for experiments.

1) Baselines: To illustrate the impact of the proposed jerk penalty, we specifically examine the difference when employing the jerk component $\mathcal{R}_{t,jerk}$, as outlined below.

- w/ $\mathcal{R}_{t,jerk}$: It refers to the proposed autonomous driving strategy, which is trained through all the reward components in (4).
- w/o $\mathcal{R}_{t,jerk}$: This strategy is trained using the reward components in (6) and (11). It provides the baseline performance when the jerk penalty in (5) is not considered during training.

2) Metric: To evaluate the ride comfort of the autonomous driving strategies, we measure the jerk ratio, which is defined as follows.

$$\text{Jerk ratio } (\%) = \frac{\text{Measured value of jerk}}{\text{Maximum value of jerk}} \times 100 \qquad (12)$$

This metric measures the percentage of measured jerk value compared to the maximum jerk value possible in the scenario settings.

C. Driving Performance Comparison

In this subsection, we provide a driving performance comparison between w/ $\mathcal{R}_{t,jerk}$ and w/o $\mathcal{R}_{t,jerk}$. First, we present the jerk ratio defined in (12) across driving scenarios. Next, we take a close look at driving characteristics during the episode.

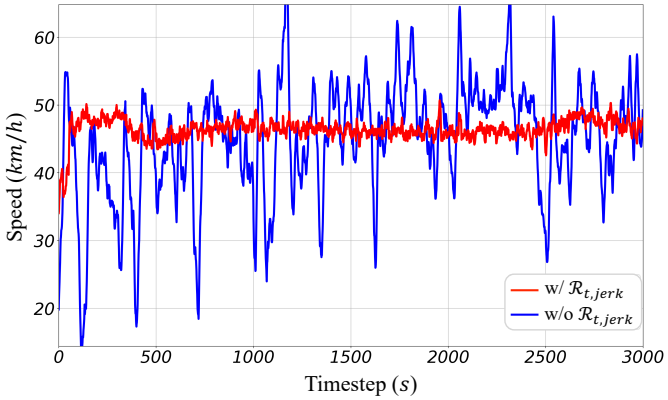

Figure 4. Speed by timestep during an episode in the highway scenario

1) Jerk Ratio Comparison: Fig. 3 illustrates the jerk ratio by road position in the on-ramp merging scenario. The solid line and shaded area represent the average jerk ratio and one standard deviation across the five random seeds. In this scenario, the agent should determine its merging time in the range between $10m$ and $30m$. It is noteworthy that the agent determines the merging time in this range. During this time, the agent accelerates, decelerates, and changes a lane, causing the jerk. This is why the jerk ratio of w/o $\mathcal{R}_{t,jerk}$ exhibits the greatly increasing jerk ratio in the range between $10m$ and $30m$. Additionally, upon reaching the transition point after failing to merge, a high jerk is caused by a sharp deceleration to merge safely. Interestingly, w/ $\mathcal{R}_{t,jerk}$ exhibits a significantly lower jerk value compared to w/o $\mathcal{R}_{t,jerk}$ on almost all positions. Consequently, we confirmed that incorporating the jerk reward component can reduce jerk during merge driving.

A numerical result of the jerk ratio across the entire road scenarios is provided by Table I. From the table, the proposed solution (w/ $\mathcal{R}_{t,jerk}$) achieves the lowest jerk ratio regardless of the driving scenarios. Specifically, the proposed solution reduces the jerk ratio by an average of 37.81% in the highway scenario, 17.85% in the cut-in scenario, and 10.49% in the on-ramp merging scenario. This means that our proposed POMDP model is effective in learning autonomous driving strategies that minimize jerk across various dynamic driving scenarios.

2) Driving Characteristic Analysis: Figure 4 illustrates the speed change according to the timestep in the highway scenario. This result effectively exhibits distinct driving characteristic differences between w/ $\mathcal{R}_{t,jerk}$ and w/o $\mathcal{R}_{t,jerk}$. Specifically, the proposed w/o $\mathcal{R}_{t,jerk}$ keeps stable driving without a drastic change in velocity. However, w/o $\mathcal{R}_{t,jerk}$ frequently displays rapid acceleration and deceleration. This result confirms that the proposed solution can provide more stable driving for autonomous vehicles. We conjecture that this originates from employing the jerk penalty. This is because the agent with a jerk penalty mitigates drastic velocity change by incurring a high penalty.

IV. CONCLUSIONS

In this study, we proposed an autonomous driving strategy to minimize jerk. For this purpose, we employed a jerk reward

component and a POMDP model that can be applied to various road driving scenarios. Simulation results confirmed that our proposed model achieves a lower jerk ratio in all scenarios. The proposed model also exhibited stable driving characteristics without rapid speed changes.

REFERENCES

[1] K. Yang, B. Li, W. Shao, X. Tang, X. Liu, and H. Wang, "Prediction failure risk-aware decision-making for autonomous vehicles on signalized intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12 806–12 820, 2023.

[2] Z. Zhu and H. Zhao, "A survey of deep RL and IL for autonomous driving policy learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 043–14 065, 2021.

[3] D. Lee and M. Kwon, "Stability analysis in mixed-autonomous traffic with deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 3, pp. 2848–2862, 2022.

[4] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 740–759, 2022.

[5] B. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2022.

[6] J. Wu, Z. Song, and C. Lv, "Deep reinforcement learning-based energy-efficient decision-making for autonomous electric vehicle in dynamic traffic environments," *IEEE Transactions on Transportation Electrification*, vol. 10, no. 1, pp. 875–887, 2024.

[7] J. Wu, Z. Huang, W. Huang, and C. Lv, "Prioritized experience-based reinforcement learning with human guidance for autonomous driving," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 855–869, 2024.

[8] P. Maramotti, A. Capasso, G. Bacchiani, and A. Broggi, "Tackling real-world autonomous driving using deep reinforcement learning," in *IEEE Intelligent Vehicles Symposium*, 2022, pp. 1274–1281.

[9] Y. Jin, Z. Ji, D. Zeng, and X. Zhang, "Vwp:an efficient drl-based autonomous driving model," *IEEE Transactions on Multimedia*, vol. 26, pp. 2096–2108, 2024.

[10] S. Chen, Y. Sun, D. Li, Q. Wang, Q. Hao, and J. Sifakis, "Runtime safety assurance for learning-enabled control of autonomous driving vehicles," in *International Conference on Robotics and Automation*, 2022, pp. 8978–8984.

[11] H. Liu, Z. Huang, J. Wu, and C. Lv, "Improved deep reinforcement learning with expert demonstrations for urban autonomous driving," in *IEEE Intelligent Vehicles Symposium*, 2022, pp. 921–928.

[12] I. Jacobson, L. Richards, and A. Kuhlthau, "Models of human comfort in vehicle environments," *Human Factors in Transport Research Edited by DJ Oborne, JA Levis*, vol. 2, 1980.

[13] O. ElSamadisy, T. Shi, I. Smirnov, and B. Abdulhai, "Safe, efficient, and comfortable reinforcement-learning-based car-following for avs with an analytic safety guarantee and dynamic target speed," *Transportation research record*, vol. 2678, no. 1, pp. 643–661, 2024.

[14] Y. Lin, J. McPhee, and N. Azad, "Anti-jerk on-ramp merging using deep reinforcement learning," in *IEEE Intelligent Vehicles Symposium*, 2020, pp. 7–14.

[15] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E*, vol. 62, no. 2, p. 1805, 2000.

[16] J. Erdmann, "SUMO's lane-changing model," *Modeling Mobility with Open Data*, pp. 105–123, 2015.

[17] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning." in *International Conference on Learning Representations*, 2016.