Proceedings Paper:
Liu, X., Apriaskar, E. and Mihaylova, L. orcid.org/0000-0001-5856-2223 (2024) Deep reinforcement learning method for control of mixed autonomy traffic systems. In: Proceedings of the 2024 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). 2024 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 04-06 Sep 2024, Pilsen, Czech Republic. Institute of Electrical and Electronics Engineers (IEEE) ISBN 9798350368048

https://doi.org/10.1109/MFI62651.2024.10705775

Deep Reinforcement Learning Method for Control of Mixed Autonomy Traffic Systems

Xingyu Liu[1], Esa Apriaskar[2] and Lyudmila Mihaylova[3]

***Abstract*— The introduction of autonomous vehicles (AVs) presents a novel approach to regulating and optimising traffic flow through the automated control of AVs. In this context, the AV is defined as the actuator and an optimal control policy is desired to make control decisions. Deep Reinforcement Learning (DRL) is a novel method which aims to maximize the cumulative rewards given by the predefined reward function by making sequential decisions in a stochastic environment. In light of the above, we propose a DRL-based vehicular control method to train an optimal policy for the control of AV in a model-free fashion, and consequently improve the traffic efficiency with the obtained control policy. A single-lane circular road environment with both AV and human-driven vehicles is selected to serve as the mixed autonomy traffic system in the Simulation of Urban MObility (SUMO) [1] traffic simulator, and the Proximal Policy Optimization (PPO) algorithm is applied for the policy improvement. Simulation results demonstrate that our strategy is effective in mitigating the unstable stop-and-go waves, increasing 67.7% of the average driving speed and reducing 19.3% of the average energy consumption in a closed-ring road environment.**

***Index Terms*— autonomous vehicles, deep reinforcement learning, proximal policy optimization, SUMO**

I. INTRODUCTION

The rapid development of electronic and communication techniques makes it possible to induce automation in mobility systems and give the potential to achieve real-time control for autonomous mobile systems like Robots and Unmanned Aerial Vehicles. In traffic systems, the emergence of partial or full adoption of automation produces an autonomous vehicle, which is capable of being fully manipulated by the computer program without the necessity for human intervention. Because of this characteristic, it is anticipated that AVs will be employed in traffic systems such as urban road networks to reduce the impact of unsatisfied human driving behaviours and optimize global objectives like traffic efficiency as well as total energy consumption [2]. Based on this, a promising research direction today is to design the controller for each AV to attain optimal control in traffic systems [3].

[1]Xingyu Liu is with Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom xliu231@sheffield.ac.uk

[2]Esa Apriaskar is with Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom, and Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Semarang, Semarang, 50229, Indonesia eapriaskar1@sheffield.ac.uk

[3]Lyudmila Mihaylova is with Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom l.s.mihaylova@sheffield.ac.uk

While full adoption of AV is nearly impossible in the near future, a conceivably intermediate scenario could involve the integration of AVs and human-driven vehicles (HDVs), creating a mixed-autonomy traffic system. In such a system, the behaviours of AV will fully or partially influence the behaviours of other HDVs. For instance, the acceleration and deceleration of an AV will affect the driving behaviours of all HDVs behind itself on a single-lane road. It has been demonstrated that the controlled vehicles are capable of completely eliminating the stop-and-go wave, a significant source of traffic congestion, in a ring road environment [4], [5]. Furthermore, it is also proved, from the control-theoretic perspective, that a mixed traffic system consisting of HDVs and AVs in a single-lane ring road is not completely controllable but is stabilizable [6]. All these evidences provide the feasibility of designing and implementing the controller for AV to achieve optimal control in mixed-autonomy traffic systems, where the AV serves as an actuator to intervene in the state of the entire system. However, the difficulty of modelling and formulating stochastic traffic dynamics prevents researchers and practitioners from solving the underlying optimization problem using the conventional optimal control style.

Deep Reinforcement Learning (DRL) is a branch of machine learning methods [7] which addresses the problem of how an intelligent agent makes decisions of actions to maximize the cumulative reward given by the reward function in a stochastic environment. In particular, the model-free DRL method learns an optimal policy which serves as the controller via the collected dataset generated from the interaction between the agent and the environment. Thanks to its model-free property, this approach is believed to achieve optimal control of mixed-autonomy traffic systems without the necessity of modelling the implicit traffic dynamics. Based on this background, we attempt to train and evaluate such an optimal control policy for the AV with the DRL-based method to optimize global objectives in the mixed-autonomy system.

A. Main Contributions

In this article, we present a DRL-based control strategy to train and evaluate an optimal control policy for the AV in a simulated mixed autonomy ring road environment, the simplest benchmark environment as in [6] for the problem of controlling mixed-autonomy systems as shown in Figure 1. We use a similar single-ring environment setting to those in [8], [9] with the objectives of maximizing traffic efficiency, and another hybrid objective is also employed which considers both traffic efficiency and energy consumption. We

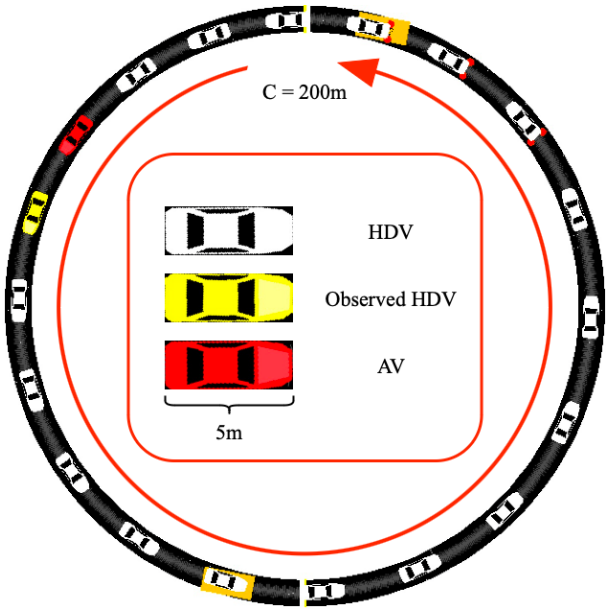

Fig. 1. **Single-lane Ring Environment.** The AV, HDV and observed HDV are displayed by the colours of red, white and yellow respectively

apply the Proximal Policy Optimization algorithm [10], an online and on-policy DRL algorithm to improve the policy, and an additional policy entropy loss is added to encourage the exploration during the training process.

In summary, our main contributions are:

1) We propose a PPO-based DRL methodology for the control of a single-lane ring road mixed-autonomy environment.
2) To encourage the exploration of PPO during training, we apply the maximum entropy strategy for policy improvement.
3) We evaluate the effectiveness of trained policy across different vehicle densities.

The rest of this article is organized as follows: Section II introduces related works, Section III presents the problem we aim to solve, Section IV presents theoretical background, Section V details our proposed approach, Section VI provides the evaluation results and Section VII gives the conclusion.

II. RELATED WORKS

A. Road-based Traffic Control

As a serious and ubiquitous problem, traffic congestion has prompted numerous researchers to design road-based traffic control strategies to improve time and energy efficiency. Traffic signal control is a major road-based approach and has been widely studied and implemented for intersections. These implementations include fixed-time control strategies which aim to reduce average delay [11], and adaptive control strategies which are presented to constantly adjust the traffic signal timing plan according to the recent traffic conditions [12]. In a specific environment, such as a freeway, ramp metering acts as the most direct control method to mitigate the onset of congestion phenomena. It facilitates the merge of the on-ramp flows with the mainstream, reducing the risk of collisions and overloaded vehicles on the road [13], [14]. The route guidance control is another strategy that also aims to enhance traffic efficiency by providing the users with information about the traffic conditions (e.g. congestion, accidents, working zones) in the alternative routes or, in some cases, indicating specific paths to follow [15]. Although some works indicate that these strategies can increase traffic efficiency, these traditional strategies rely on human drivers. With all their humanistic sense, drivers may not always comply with any suggested action resulting from these control strategies which can lead to inefficiencies and potential safety risks. The most recent strategy has been introduced, considering the emergence of AV. According to [16], AV with its communication ability to the surrounding environment offers prospective benefits that can affect driving behaviour, leading to global traffic efficiency.

B. Model-free Deep Reinforcement Learning

Model-free DRL is a branch of Reinforcement Learning algorithms that optimizes policies without needing a state transition model and reward distribution estimations. The policy gradient methods, derived from the policy gradient theorem, are a dominating portion in this category. REINFORCE is the first practical policy gradient-based algorithm that applies the Monte Carlo method to estimate the policy gradient [17]. After that, the A2C algorithm with an actor-critic framework was presented to reduce the variance of the training process introduced by the inaccurate Monte Carlo estimation [18]. To achieve a monotonic improvement in each policy update step, the Trust Region Policy Optimization (TRPO) [19] was proposed and performed excellently compared to other policy gradient-based algorithms. TRPO has the theoretical monotonic improvement property but suffers from complicated update processes and approximation errors. For this reason, the PPO algorithm [10] was invented to simplify the update process of TRPO and ease the computational load.

C. Model-free DRL for Traffic Control

As a suitable and powerful tool, the model-free DRL methods have also been studied to implement in road- and vehicle-based approaches. [20] and [21] applied DRL for optimizing traffic signal timing. [22] presented a smart re-routing technique for the AV to increase traffic efficiency at the intersection, and [23] presented a multi-agent DRL method to coordinate AVs and minimize the lost time. To facilitate the design and implementation of vehicular traffic control with DRL, a modular framework Flow [9] was published to construct a platform for the control of mixed autonomy systems. A unified DRL methodology for mixed autonomy systems [8] which works for several road scenarios, including several closed and open road networks was also presented to establish a unified DRL design scheme using a TRPO algorithm. Based on this background, our work attempts to implement a PPO algorithm in a ring road system with a single AV to maximize traffic efficiency. Some

studies, such as [24], [25] have also attempted to implement PPO extending the work of [9]. Implementation of PPO for traffic smoothing with different environments and reward functions are also presented in [26]. In this work, we propose a different reward function by adding an entropy loss in the actor's objective function to increase the exploration of the standard PPO algorithm.

III. Problem Formulation

In the single-ring road system, as shown in Figure 1, a single-lane circular road with a circumference of C meters forms the road network. There are N vehicles denoted by $\{v_1, v_2, \dots v_N\}$ driving endlessly following a counterclockwise direction, and one of these vehicles is defined as AV and all other vehicles are HDVs. To simulate the driving behaviours of the human driver, we define HDVs, which are controlled by the car-following model, and the well-studied Intelligent Driver Model (IDM) model [27] is applied to play the control role. Since the single-ring system is closed and no lane-change action is permitted, the driving behaviour of each vehicle will at least partially influence other vehicles' behaviour, therefore it is feasible to intervene in the aggregated whole traffic flow's behaviour as proved in [6].

The traffic control problem in our single-ring mixed-autonomy system is how to design a controller to manipulate the acceleration and deceleration behaviours of the AV to further improve traffic efficiency and decrease the total energy consumption. And this problem will be addressed with our DRL-based strategy.

IV. Background Methodology

A. Markov Decision Process

Markov Decision Process (MDP) serves as a mathematical representation of the decision-making process of an intelligent agent in a stochastic environment. MDP is defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, \rho_0, \gamma, H)$ which includes state space $\mathcal{S}$, action space $\mathcal{A}$, state transition function $T(s, a, s') = p(s' \mid s, a)$ for $s, s' \in S$, reward function $r(s, a, s') \in \mathbb{R}$, initial state distribution ρ_0, discounted factor $\gamma \in [0, 1]$ and horizon $H \in \mathbb{Z}_+$. Under the MDP framework, the reinforcement learning task aims to maximize the following objective

$$\max_{a_0 \dots a_{H-1} \in \mathcal{A}} \mathbb{E}_{\substack{s_0 \sim \rho_0, \\ s_{t+1} \sim T(s_t, a_t, \cdot)}} \left[\sum_{t=0}^{H-1} \gamma^t r(s_t, a_t, s_{t+1}) \right] \quad (1)$$

which represents the expected discounted cumulative reward and is going to be optimized by selecting a sequence of the optimal actions $a_0 \dots a_{H-1} \in \mathcal{A}$. The reward function $r(s, a, s') \in \mathbb{R}$ is of great importance in giving the appropriate feedback for the agent's performance in each time step.

B. Proximal Policy Optimization Approach

PPO is a policy-based model-free DRL approach which trains and updates a parametric stochastic policy π_θ with parameters θ (*e.g.* weights in a neural network). The policy parameters θ are optimized to maximize the expected cumulative discount reward

$$\max_{\theta} \mathbb{E}_{\substack{s_0 \sim \rho_0, a_t \sim \pi_\theta(\cdot \mid s_t), \\ s_{t+1} \sim T(s_t, a_t, \cdot)}} \left[\sum_{t=0}^{H-1} \gamma^t r(s_t, a_t, s_{t+1}) \right] \quad (2)$$

The policy gradient algorithms update the policy by estimating the gradient of the expected cumulative discount reward

$$\nabla_\theta J(\theta) = \mathbb{E}\left[Q_{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a \mid s) \right] \quad (3)$$

and perform gradient ascent on θ in each update step

$$\theta_{\mathrm{k}+1} \leftarrow \theta_{\mathrm{k}} + \alpha \nabla_\theta J(\theta_{\mathrm{k}}). \quad (4)$$

where α is the step size.

On this basis, the TRPO attempts to have the monotonic performance improvement of the policy by optimizing the constrained objective

$$\begin{aligned} \max_{\theta_{k+1}} \quad & \mathbb{E}_{\substack{s_0 \sim \rho_0, a_t \sim \pi_{\theta_k}(\cdot \mid s_t), \\ s_{t+1} \sim T(s_t, a_t, \cdot)}} \left[\frac{\pi_{\theta_{k+1}}(a \mid s)}{\pi_{\theta_k}(a \mid s)} A_{\pi_{\theta_k}}(s, a) \right] \\ \text{s.t.} \quad & \mathbb{E}_{\substack{s_0 \sim \rho_0, \\ s_{t+1} \sim T(s_t, a_t, \cdot)}} \left[D_{KL}\left(\pi_{\theta_k}(\cdot \mid s), \pi_{\theta_{k+1}}(\cdot \mid s) \right) \right] \leq \delta \end{aligned} \quad (5)$$

where δ is the upper bound of the mean KL divergence [19] between old policy π_{θ_k} and new policy $\pi_{\theta_{k+1}}$ to prevent θ_{k+1} from deviating too far from θ_k. And PPO performs the KL constraint by optimizing the clipped objective

$$\begin{aligned} \max_{\theta_{k+1}} \quad & \mathbb{E}_{\substack{s_0 \sim \rho_0, \\ a_t \sim \pi_{\theta_k}(\cdot \mid s_t), \\ s_{t+1} \sim T(s_t, a_t, \cdot)}} \left[\min \left(\frac{\pi_{\theta_{k+1}}(a \mid s)}{\pi_{\theta_k}(a \mid s)} A_{\pi_{\theta_k}}(s, a), \right. \right. \\ & \left. \left. \operatorname{clip}\left(\frac{\pi_{\theta_{k+1}}(a \mid s)}{\pi_{\theta_k}(a \mid s)}, 1 - \epsilon, 1 + \epsilon \right) A_{\pi_{\theta_k}}(s, a) \right) \right] \end{aligned} \quad (6)$$

where $\operatorname{clip}(x, y, z) = \max(\min(x, z), y)$, this operation is to constraint the ratio between old policy and new policy into $[1 - \epsilon, 1 + \epsilon]$ where ϵ is a hyperparameter.

V. Proposed Approach

A. MDP Definition

In our vehicle control strategy, the AV acts as the intelligent agent, and each decision-making of the AV is defined as an MDP. Due to the physical limits on the sensor or detector assembled on the agent, the state s may not be fully observable, and only a subset of state space is feasible to obtain. For example, the AV could hardly detect the behaviours of vehicles which are far away from it. Therefore, an observation function $z(s) = o \in \mathcal{O}$ is added to process the state s into a portion of the original state information o, and this also induces the partially observable MDP (POMDP) denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, \rho_0, \gamma, H, \mathcal{O}, z)$. Overall, in each training episode[1], we perform the POMDPs for the AV, the state s is composed of the speed, location and all other information of all vehicles in the ring environment, and the

[1]In Reinforcement Learning, episode means the recording of actions and states that an agent performed from a start state to an end state.

observable region is defined as the combination of the speed of AV, the speed of the vehicle in front of AV and the distance between AV and its front vehicle.

For the definition of the action space, thanks to the single-lane property of the ring environment, the lane-changing does not need to be considered for the AV's control policy, and therefore the action space of AV could be defined as a continuous acceleration space $[a_{min}, a_{max}]$. The continuous control tasks normally have the nature of the infinite horizon length, which is usually impossible to implement in practice, thereby we convert the continuous control mission to an episodic task by using a large horizon H to approximate the infinity.

B. Maximum Entropy Strategy

In our proposed PPO setting, we use an actor-critic framework [18] consisting of a policy network and a value network to reduce the variance of the training process, and the General Advantage Estimation (GAE) [10] is also applied to estimate the advantage $\hat{A}_t$. To encourage the exploration, we add an entropy loss in the actor's objective function

$$\mathrm{L}_{actor} = L(\theta) + \alpha\mathcal{H}(\cdot \mid s_t) \tag{7}$$

where $L(\theta)$ is the clipped objective as defined in equation 6, and $\mathcal{H}(\cdot \mid s_t) = \mathbb{E}_{a_t\sim\pi}\left[-\log\left(\pi\left(a_t \mid s_t\right)\right)\right]$ is the policy entropy to quantify the uncertainty of the probability distribution of policy $\pi\left(a_t \mid s_t\right)$. Maximizing the policy entropy loss augments the uncertainty of the probability distribution over action space. Consequently, sub-optimal actions are more likely to be selected, thereby further increasing the exploration of the policy.

C. Neural Network Architecture

In our strategy, we construct both the actor and critic as the Multi-Layer Perceptron (MLP) with three fully connected layers and a hidden size of 64. To handle the continuous action space, we apply the Gaussian distribution, which is fully described by its mean and standard deviation, to approximate the distribution over the action space. In this case, the policy network will receive the partially observed state information and calculate the corresponding mean and standard deviation via the forward propagation in each time step. The orthogonal initialization is applied for each network to avoid vanishing and exploding gradient problems at the beginning of the training process, and the $tanh$ activation function is also used to make the networks capture more nonlinear features.

D. Training

We acknowledge that the Sim2Real gap in DRL is almost impossible to be solved in the near future, therefore we finish the training and evaluation parts in the simulator and do not address the Sim2Real problem. Some works tried to bring the DRL methods into a real environment in another environment, such as [28] for autonomous mobile robot navigation, ensuring that the potential remains open. SUMO [1] is a microscopic traffic simulator which provides a framework for the generation and validation of any traffic scenario which simulates the real environment, including a road environment with a mixture of HDVs and AVs.

To train a policy across different vehicle densities, we build several simulation environments with varying numbers of vehicles in a range of$[N_{\min}, N_{\max}]$. In each episode, these simulations will be started individually, a warm-up horizon $H_0 = \frac{500}{\delta}$ with simulation step size δ is used to initialize each simulation, and an episode $H_1 = \frac{1000}{\delta}$ is set for the interaction and data collection. We collect N trajectories in each episode and concatenate these data before updating the policy. For the reward processing, rather than the conventional reward normalization, we apply a novel reward scaling method as presented in [29], which makes the rewards divided by the standard deviation $\hat{\sigma}_R$ of a rolling discounted sum of rewards

$$r_{\text{scaling}} = \frac{r(s_t, a_t, s_{t+1})}{\hat{\sigma}_R} \tag{8}$$

where $\hat{\sigma}_R$ is the standard deviation of the collected reward.

VI. Performance Evaluation

We train the policy with two objectives:

1) Total Travelled Distance: The reward function $r(s, a, s')$ is the average speed of all vehicles in s'.
2) Fuel Consumption: The reward function $r(s, a, s')$ is the average speed minus the average fuel consumption multiplied by a coefficient β in s', which is $r(s, a, s') = v_{average} - \beta q_{average}$, where $v_{average}$ refers to the average speed, and $q_{average}$ refers to the average fuel consumption.

We set the baseline for both cases as the situation where the IDM model controls all vehicles, and separately give the evaluation results. The hyperparameters setting of PPO and environmental parameters in both two cases are shown in Table I.

TABLE I
EXPERIMENTAL PARAMETERS SETTING FOR PPO AND RING ENVIRONMENT

PPO Hyperparameters		**Environment Parameters**	
Discount factor (γ)	0.99	Circumference C	200
GAE discount (λ)	0.95	Maximum of vehicles $N_{\max}$	19
Actor learning rate	3e-5	Minimum of vehicles $N_{\max}$	16
Critic learning rate	3e-5	Simulation step length δ	0.5
Entropy coefficient (α)	1e-4	Maximum acceleration a_{max}	1.8
PPO Clipping ϵ	0.2	Minimum acceleration a_{min}	-2.5
Epochs	10	Number of episodes	100

A. Evaluation of Total Travelled Distance

Figure 2 compares the average speeds between PPO-controlled and IDM baseline scenarios under different vehicle densities. It is evident that our PPO-trained policy improved the average speed in the environments where $N \in [17, 19]$, surpassing the IDM baseline. In the environments where N equals 16, our trained policy achieved similar performance to the baseline. In general, our trained policy

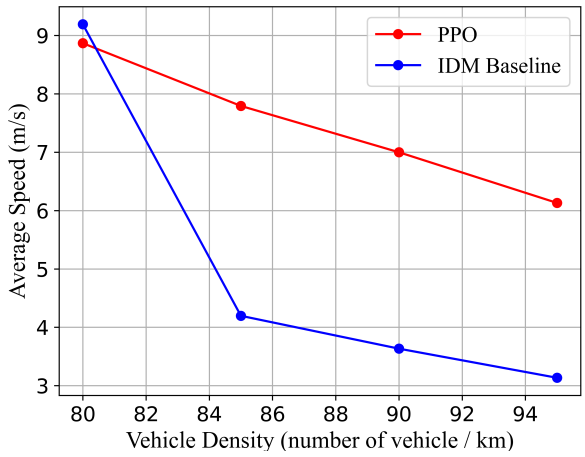

Fig. 2. **Average Speed in Different Vehicle Densities.** We compute the average speed by summing the temporary average speed of all vehicles in each time step and then divided by the cumulative time.

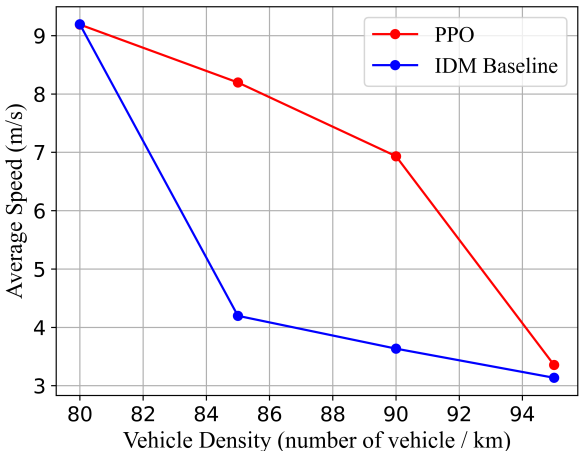

Fig. 4. **Average Speed in Different Vehicle Densities.** We compute the average speed by summing the temporary average speed of all vehicles in each time step and then divided by the cumulative time.

increased 67.7% of the average speed compared to the baseline. To investigate the effectiveness of our trained

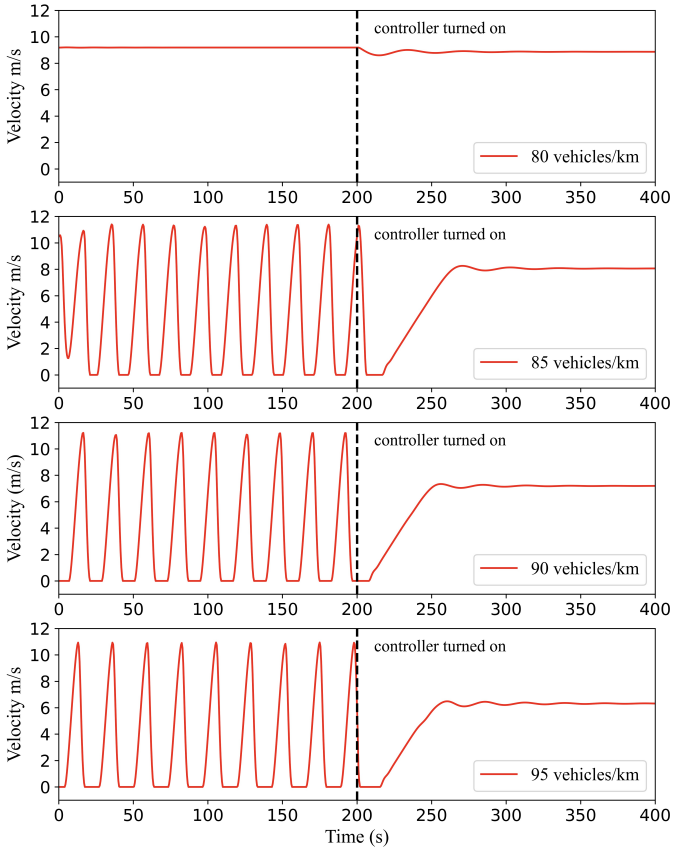

Fig. 3. **Velocity Distribution of HDV in Different Vehicle Densities** We select one of HDV to collect and display the simulated velocities throughout the simulation.

policy, we present the velocity distributions of a selected HDV as shown in Figure 4. After loading all vehicles in

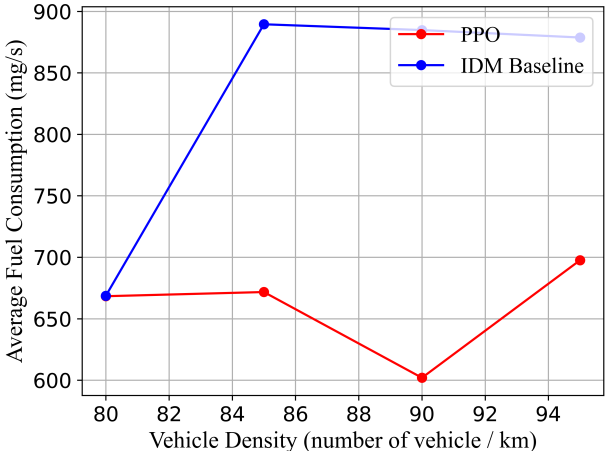

Fig. 5. **Average Fuel Consumption in Different Vehicle Densities.** We compute the average fuel consumption by summing the temporary average fuel consumption of all vehicles in each time step and then divided by the cumulative time.

each environment, we first turn off the controller to leave all vehicles uncontrolled and let them follow the IDM model, then start controlling after a certain time step. It is noticeable that before the control policy was involved, the IDM model resulted in severe stop-and-go waves, and the velocity decreased to almost zero in some time steps. After the start time, our trained control policy quickly regulated the traffic flow and suppressed the unstable stop-and-go waves. The velocities of the selected HDV gradually changed to a constant speed, which is the ideal optimal traffic flow behaviour for the single-lane ring environment.

B. Evaluation of Energy Consumption

Figure 4 shows the average speeds under the energy-saving objective in different vehicle densities. It is clear that, given a different reward function, our trained AV could still improve

the overall average speed in most vehicle densities, or at least achieve a similar performance with the baseline. And Figure 5 displays the average fuel consumption result in different vehicle densities. Our trained policy dramatically reduced the average fuel consumption in most density settings compared to the baseline. In general, our trained policy reduced 19.3% of the average fuel consumption compared to the baseline.

VII. CONCLUSIONS

This paper proposes a PPO-based DRL control strategy for controlling a single-lane mixed-autonomy road system to mitigate traffic congestion and reduce the total cost of energy. The simulation results show that our trained policy effectively improves the average speed of all vehicles. The unstable stop-and-go waves are greatly weakened and the total fuel consumption is also largely decreased via controlling of AV.

The model-free DRL has the advantages of achieving optimal control for a complex and stochastic dynamical system and automatically discovering the optimal behaviour. However, it still suffers from the difficulty of convergence and is easily stuck in a local minimum. Data efficiency is also a main drawback of on-policy DRL algorithms. In addition, the Sim2Real problem affects the real application of DRL-based methods. Future research directions include extensions to large-scale road networks and could apply off-policy or offline DRL algorithms to address data-efficiency problems.

ACKNOWLEDGEMENT

We acknowledge the Indonesian Endowment Fund for Education (LPDP) for the funding support during the PhD research of Esa Apriaskar with contract no. 20230722299644.

REFERENCES

[1] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, "Microscopic Traffic Simulation using SUMO," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2575–2582, Nov. 2018. ISSN: 2153-0017.

[2] M. Al-Turki, N. T. Ratrout, S. M. Rahman, and I. Reza, "Impacts of autonomous vehicles on traffic flow characteristics under mixed traffic environment: Future perspectives," *Sustainability*, vol. 13, no. 19, 2021.

[3] H. Yu, R. Jiang, Z. He, Z. Zheng, L. Li, R. Liu, and X. Chen, "Automated vehicle-involved traffic flow studies: A survey of assumptions, models, speculations, and perspectives," *Transportation Research Part C: Emerging Technologies*, vol. 127, p. 103101, June 2021.

[4] R. E. Stern, S. Cui, M. L. Delle Monache, R. Bhadani, M. Bunting, M. Churchill, N. Hamilton, H. Pohlmann, F. Wu, B. Piccoli, *et al.*, "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 205–221, 2018.

[5] S. Cui, B. Seibold, R. Stern, and D. B. Work, "Stabilizing traffic flow via a single autonomous vehicle: Possibilities and limitations," in *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1336–1341, IEEE, 2017.

[6] Y. Zheng, J. Wang, and K. Li, "Smoothing traffic flow via control of autonomous vehicles," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3882–3896, 2020.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[8] Z. Yan, A. R. Kreidieh, E. Vinitsky, A. M. Bayen, and C. Wu, "Unified automatic control of vehicular systems with reinforcement learning," *IEEE Transactions on Automation Science and Engineering*, 2022.

[9] C. Wu, A. R. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: A modular learning framework for mixed autonomy traffic," *IEEE Transactions on Robotics*, vol. 38, no. 2, pp. 1270–1286, 2021.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[11] S. Araghi, A. Khosravi, and D. Creighton, "Intelligent cuckoo search optimized traffic signal controllers for multi-intersection network," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4422–4431, 2015.

[12] S. Chen and D. J. Sun, "An improved adaptive signal control method for isolated signalized intersection based on dynamic programming," *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 4, pp. 4–14, 2016.

[13] I. Papamichail, M. Papageorgiou, and Y. Wang, "Motorway traffic surveillance and control," *European Journal of Control*, vol. 13, no. 2-3, pp. 297–319, 2007.

[14] K. Shaaban, M. A. Khan, and R. Hamila, "Literature review of advancements in adaptive ramp metering," *Procedia Computer Science*, vol. 83, pp. 203–211, 2016.

[15] H. Chai, H. M. Zhang, D. Ghosal, and C.-N. Chuah, "Dynamic traffic routing in a network with adaptive signal control," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 64–85, 2017.

[16] C. Diakaki, M. Papageorgiou, I. Papamichail, and I. Nikolos, "Overview and analysis of Vehicle Automation and Communication Systems from a motorway traffic management perspective," *Transportation Research Part A: Policy and Practice*, vol. 75, pp. 147–165, May 2015.

[17] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[18] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in neural information processing systems*, vol. 12, 1999.

[19] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the International conference on machine learning*, pp. 1889–1897, PMLR, 2015.

[20] L. Li, Y. Lv, and F.-Y. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 3, pp. 247–254, 2016.

[21] W. Genders and S. Razavi, "Using a deep reinforcement learning agent for traffic signal control," *arXiv preprint arXiv:1611.01142*, 2016.

[22] A. Mushtaq, I. U. Haq, M. U. Imtiaz, A. Khan, and O. Shafiq, "Traffic flow management of autonomous vehicles using deep reinforcement learning and smart rerouting," *IEEE Access*, vol. 9, pp. 51005–51019, 2021.

[23] G.-P. Antonio and C. Maria-Dolores, "Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrow's intersections," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7033–7043, 2022.

[24] E. Vinitsky, A. Kreidieh, L. L. Flem, N. Kheterpal, K. Jang, C. Wu, F. Wu, R. Liaw, E. Liang, and A. M. Bayen, "Benchmarks for reinforcement learning in mixed-autonomy traffic," in *Proceedings of The 2nd Conference on Robot Learning*, pp. 399–409, PMLR, Oct. 2018. ISSN: 2640-3498.

[25] H. Wei, X. Liu, L. Mashayekhy, and K. Decker, "Mixed-Autonomy Traffic Control with Proximal Policy Optimization," in *2019 IEEE Vehicular Networking Conference (VNC)*, pp. 1–8, Dec. 2019. ISSN: 2157-9865.

[26] N. Lichtlé, K. Jang, A. Shah, E. Vinitsky, J. W. Lee, and A. M. Bayen, "Traffic Smoothing Controllers for Autonomous Vehicles Using Deep Reinforcement Learning and Real-World Trajectory Data," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 4346–4351, Sept. 2023. ISSN: 2153-0017.

[27] R. Herman, E. W. Montroll, R. B. Potts, and R. W. Rothery, "Traffic dynamics: analysis of stability in car following," *Operations research*, vol. 7, no. 1, pp. 86–106, 1959.

[28] H. Surmann, C. Jestel, R. Marchel, F. Musberg, H. Elhadj, and M. Ardani, "Deep Reinforcement learning for real autonomous mobile robot navigation in indoor environments," May 2020. arXiv:2005.13857 [cs].

[29] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry, "Implementation matters in deep policy gradients: A case study on PPO and TRPO," *arXiv preprint arXiv:2005.12729*, 2020.