Cole-Puli International Conference Publishing Group

Science, Engineering and Technology Proceedings

Article

2025 International Conference on Natural Sciences, Agricultural Economics, Biomedicine and Sustainable Development (AEBSD 2025)

Multi-Agent Reinforcement Learning for Autonomous Driving in Signal-Free Intersections and Roundabouts

Haoran Tang [1,*]

[1] University of Birmingham, Birmingham, UK
* Correspondence: Haoran Tang, University of Birmingham, Birmingham, UK

Abstract: Signal-Free Intersections & Roundabouts: One of the most difficult problems that a self-driving car faces is the proper functioning at intersections with no traffic lights. It is because there are no traffic lights, which assist us in making quick decisions, unlike just using lidar sensors and cameras for fixed-lane driving. In these situations the current research on multi-agent reinforcement learning (MARL) does not extent to such domains because of issues like non-stationarity and partial observability. The novel hierarchical learning framework for related communications integrates hierarchical learning structures and complex cost functions, which can better adjust a more adaptive response to safety and traffic flow efficiency.

Keywords: signal-free intersections; MARL; LCF

1. Introduction

The promising advantages of urban transportation improvement sparked the strong interest in the possibility that autonomous driving would revolutionize the sector by making traffic more efficient, car crashes less, and that travel times would be shortened. One of the more difficult tests there is for this area are signal-free intersections and roundabouts. Consequently, such environments necessitate coordinated behavior among many agents for navigation and tend to be dynamic with no clear points of reference. Intersections in particular are known to be the main causes of around 40% of all motor vehicle accidents with the number of serious crashes from which more than half of the victims become over half of such incidents and fatalities are the largest category of the death cause in traffic making as much as one-fifth of all the fatal ones [1]. Very few studies focus on the interactions at unsignalized intersections or roundabouts, dealing with things like negotiating the right of way, and the speeds being variable, required to yield or maneuver the traffic circles. The study suggests a different method for MARL which is capable of real-time adapting the Local Coordination Factor (LCF) to local environmental signals such as neighbors' proximity and behavior. Thus, the agents get to feel the impact of their actions if they were to adjust their strategies based on live traffic conditions, which would actually let them be safer and by doing so a better flow of the traffic will be possible.

The primary difficulty in managing these environments lies in coordinating the actions of a large number of autonomous vehicles (AVs) in the face of environmental uncertainty, non-connected agents (e.g., human drivers and pedestrians), and the need for real-time decision-making [2]. Assessments of traditional MARL methods, examples of which include Markov games and independent Q-learning, have resulted in the

Received: 11 November 2025
Revised: 26 November 2025
Accepted: 13 December 2025
Published: 20 December 2025

realization that these techniques cannot maintain similar scalability when dealing with numerous agents and fast-changing interactions in such settings [3].

2. Related Work

2.1. Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning outlines the conceptual layout of the development of autonomous systems operating in shared environments. Alongside Markov Games and Independent Q-learning, as the main techniques, Joint Action Learning and Cooperative Learning contributed to the overall progress of multi-agent systems [4]. However, the methods still have difficulties in successfully adapting to real-world scenarios within which a large number of agents interact concurrently [5]. The learning process in these types of environments is adversely affected by the presence of non-stationarity and partial observability, two factors that severely constrain algorithmic performance [6].

Multitude of efforts has been dedicated to enhancing the reliability of Multi-Agent Reinforcement Learning (MARL) algorithms such that they could overcome challenges posed by mixed cooperative-competitive environments. One such example is the Multi-Agent Actor-Critic (MAAC) family of methods which in general, manage to solve the issues mentioned in both [6]. These methods have been upgraded by including an agent-cooperation module to lighten the system load and enhance their efficiency, but they are still unable to function properly in big and complicated traffic scenarios such as an intersection without signals or a roundabout [7].

2.2. Traffic Simulation and Autonomous Driving

Simulation environments play a major role in the creation and the review of autonomous vehicle algorithms. One such example is virtual reality testbeds (like CARLA), which practically replicate daily urban driving and allow standard experiments to be conducted for different types of maneuvers [8]. However, the accurate simulation of these complex situations, especially signal-free intersections and roundabouts, is still a challenge due to their high variability and the intricate nature of interactions.

2.3. Dynamic Local Coordination in MARL

Dynamic local coordination is an innovative concept that could solve the problems of scalability and stability in MARL, which is a dense traffic situation without signals. Using Local Coordination Factor (LCF) to adjust the agents' behavior to the changing conditions and density of the local population, they synchronize their actions with the nearest surroundings. Interacting with the presence and actions of neighboring units that are close enough to reach you both a suitable blend of cooperative and competitive strategies that are likely to lead to an increase in your personal safety as well as system-level throughput [9].

3. Methodology

3.1. Problem Formulation

Due to the characteristics of these surroundings, the agents are required to be flexible and engage in reactive coordination. Therefore, we present a combination means-the Local Coordination Factor (LCF)-to our MARL system architecture for accomplishing local coordination. Being updated on the ground with the latest situational awareness, LCF changes depending on the distance and the behavior of the agents, thus enabling individual adaptation that is in line with the overall goals of the system (for instance, decreasing the aggregate time of travel and alleviating the flow of interruptions).

As the changes in such environments are the main reasons that flexible and responsive agent coordination is needed, we have come up with a new way of our MARL

framework which is represented by an added mechanism named Local Coordination Factor (LCF). The LCF, in return, is very much connected with the on-the-spot situational awareness and it changes according to the closeness of agent density and as well to the behavior changes. This is the way each agent is not just able to change their tactics given the local environment but also in a manner that is compatible with the overall system goals (e.g. lower total travel time and traffic flow interruptions).

3.2. Local Coordination Factor (LCF)

We, therefore, come up with a new solution to properly coordinate agent behavior, that we named the Local Coordination Factor (LCF). The LCF depends on the state of neighboring agents that are within a given distance $\mathrm{d_n}$ and allows each agent to dynamically modify its behavior depending on local environment conditions. The mathematical expression for such definition is:

$$\mathrm{LCF_{\varphi i}} = \mathrm{f}\left(\mathrm{S_{\mathcal{N}_{\{d_n\}(i,t)}}}\right)$$

Where $\mathrm{S_{\mathcal{N}_{\{d_n\}(i,t)}}}$ represents the state of neighboring agents in a $\mathrm{d_n}$ radius around agent i at time t. The function f updates the LCF in a dynamic way based on both population and activity of neighboring agents, allowing an agent to be balanced among selfishness and cooperative behaviors.

The local reward in a neighborhood of agent $\boldsymbol{t}$ at time step $\boldsymbol{t}$ is denoted standardly as:

$$r_{i,t}^{N} = \frac{\sum_{j \in \mathcal{N}_{d_n}(i,t)} r_{j,t}}{|\mathcal{N}_{d_n}(i,t)|}$$

$\mathcal{N}_{d_n}(i,t) = \{j : ||Pos(i) - Pos(j)|| \leq d_n\}$ is the set of neighbor agents, this reward provides a mean of the pay-off performance within neighbor agents to help learn them cooperate.

The coordinated reward for the agent is then defined as:

$$\mathrm{r_{i,t}^{C}} = \alpha \cos(\phi)\, \mathrm{r_{i,t}} + \beta \sin(\phi)\, \mathrm{r_{i,t}^{N}}$$

Where $\mathrm{r_{i,t}}$: The reward of an individual and $\mathrm{r_{i,t}^{N}}$: the neighborhood rewards the constants $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ control the influence of single rewards vs. neighborhoods' rewards.

The agent, in turn updates its strategy via the LCF that is adapting to local condition and maximizes the coordinated reward $r_{\{i,t\}}^{C}$. While in a group of many neighboring agents, the LCF will cause it to become more cooperative and try to achieve global optimal solution for traffic flow. In contrast, the LCF allows for more competitive behavior in lower density scenarios concentrating on individual rewards.

This adaptive modification of the LCF enables each agent's behavior to be adjusted according to environmental conditions on-site, thereby promoting safety and traffic flow in more complicated signal-free scenarios.

3.3. Reinforcement Learning Environment

In this study, traffic signal-free intersections and roundabouts are simulated through reinforcement learning environment for self-driving vehicles. Environment setup is as follows:

Agents: The agents in this environment are autonomous cars, referred to as target vehicles. Each agent operates independently but must coordinate its actions with others to navigate safely and efficiently.

Actions: Each agent has two action controls:

1) Steering: A continuous action in the range [-1, 1], where -1 corresponds to a full left turn and 1 to a full right turn.
2) Acceleration/Deceleration: A continuous action in the range [-1, 1], where -1 represents full braking (backwards) and 1 represents full throttle (forwards).

State: The state observed by each agent includes:

1) Steering, Heading, and Velocity: These are the basic control states of the vehicle.

2) Relative Distance to Boundaries: The distance to the left and right road boundaries and the proximity to other vehicles.
3) Lidar-like Point Clouds: A 240-dimensional vector (72-dimensional in MARL) representing 2D Lidar-like point clouds. Point clouds of environment around the agent, each item in vector is how far away closest obstacle given 50 meter radius.
4) Target Vehicle State Summary: A vector summarizing the state of the target vehicle, including its steering, heading, velocity, and distances to boundaries.
5) Navigation Information: Checkpoint information is provided on the path to get the vehicle to its destination. The checkpoints on the route are about 50 meters apart. The observations of the vehicle are then updated in relation to where other checkpoints will be.

3.4. Reward Function

Our MARL framework rewards to consider three objectives: efficiency of driving, safety and compliance with traffic rules. The reward is typically a dense driving reward and then sparse terminal rewards.

3.4.1. Reward I: Driving Efficiency and Safety

Reward I is composed of three parts:

$$R = c_1 R_{disp} + c_2 R_{speed} + R_{termination}$$

1) Displacement Reward (R_{disp}): Encourages forward movement by rewarding the longitudinal movement of the agent between consecutive time steps. It is calculated as $R_{disp} = d_t - d_{t-1}$, where d_t and d_{t-1} represent the agent's positions at consecutive time steps.
2) Speed Reward (R_{speed}): Incentivizes the agent to maintain high speed, calculated as $R_{speed} = \frac{V_t}{V}$, where vt is the current velocity and vmax is the maximum allowable speed.
3) Terminal Reward ($R_{termination}$): A sparse reward applied at the last time step. If the agent successfully reaches its destination, $R_{termination} = +10$; if it crashes or violates traffic rules, $R_{termination} = -5$.

3.4.2. Reward II: Advanced Control and Safety

Reward II introduces additional control-focused rewards:

$$R = c_1 R_{disp} + c_2 R_{lateral} + c_3 R_{heading} + c_4 R_{steering} - c_5 P_{collision} + R_{termination}$$

1) Displacement Reward (R_{disp}): Similar to Reward I, this rewards forward movement.
2) Lateral Reward ($R_{lateral}$): Encourages the agent to stay close to the reference trajectory, rewarding lane-keeping and penalizing deviations.
3) Heading Reward ($R_{heading}$): Encourages alignment with the road's heading direction by rewarding the agent for maintaining the correct heading.
4) Steering Reward ($R_{steering}$): Penalizes large steering adjustments at high speeds to promote smoother control.
5) Collision Penalty ($P_{collision}$): Penalizes the agent for any collisions with other vehicles or objects.
6) Terminal Reward ($R_{termination}$): Similar to Reward I, this reward is applied at the end of the episode based on the agent's final state.

3.5. Cost Function

Alongside the reward functions, a cost function is defined to punish unsafe actions. The cost function is used to make sure that all of these agents prioritize safety.

1) Collision Cost (C_{coll}): A penalty is applied when an agent collides with another vehicle, human, or object. This is expressed as $\mathrm{C}coll = -\beta 1 \times collision\ event$, where $\beta 1$ is a large negative constant representing the severity of collisions.
2) Energy Consumption Cost (C_{energy}): To promote energy-efficient driving, a cost is applied based on energy usage. This is calculated as $C_{energy} = -\beta 2 \times energy\ used$, where $\beta 2$ represents the cost associated with energy consumption.

3.6. Transition and Sequence of Events

State transitions are done through DQN (Deep Q-Network) where the agents make decisions based on Q-values. Once an agent leaves the environment either by reaching its destination or via a collision, a new car will spawn to replace it keeping and endless flow of agents moving through the environment.

4. Experiments

4.1. Experiment Setup

Simulated Environment: We used Meta Drive to create realistic urban driving situations with different road layouts, traffic densities and environmental conditions. Meta Drive allowed us to simulate difficult traffic primitives, such as a signal-free intersection and roundabout, which are the leading factors for the stability test of autonomous driving systems.

Scenario Configuration: We have created scenarios of controlled behavior with three different traffic situations: light, medium, and heavy. We also looked at several roundabout designs after going beyond the fundamental intersection. In each case, traffic lights were not used, and agents were presumed to continue a safe, efficient flow.

Hardware configuration. The machine with specifications such as [CPU], [RAM], and [GPU] under [OS] was used to run all the experiments and simulations were conducted on [platform/version]:

1) GPU: RTX 3090 / 24 GB
2) Memory: 80 GB
3) CPU: AMD EPYC 7642

4.2. Evaluation Metrics

We conducted tests with various combinations of Meta Drive. To evaluate our MARL framework in practice, we report two aggregate indicators.

1) **Collision Rate:** The number of vehicle-to-vehicle collisions in the simulation; the less the number of collisions, the safer the system. We have decided to present the average collision rate, i.e., the number of collisions per second, resulting from the counting process of all contacts and the elapsed simulation time being used for normalization.
2) **Success Rate:** The fraction of the vehicles that successfully reach their targets without crashes or violations of the traffic rules, which, in turn, indicates the quality of overall task completion.
3) **Mean Acceleration (Maximum Value):** We present the average longitudinal acceleration per vehicle along with the highest instantaneous acceleration that each vehicle experienced. The lower the values employed, the more the control will be smooth, the stability will be better, and the comfort of the ride will be greater.
4) **Environment Reward**: One of the core features of reinforcement learning is the reward function that essentially depicts how agents have performed in the environment. Hence, the maximum and average reward values were obtained

during agents' execution of various tasks, which allowed us to have an overall performance comparison of their efficiency/effectiveness across different scenarios.

5) **Episode Length (Maximum Value):** This measurement is introduced to represent the duration of a single worst-case performance by an agent. Limiting episode lengths is important because the longer the line, the worse a virtual I-paddler task execution or poor decision making become more visible.

4.3. Experimental Results

1) Based on our experimental results using the Meta Drive simulation, the MARL framework has been demonstrated to outperform the baselines under various parameter settings.
2) **Safety:** From collision frequency, it can be recognized that the number of collisions was DATA% less, showing that the method is efficient for lowering this number.
3) **Success rate:** The success rate of the proposed MARL framework was 66.47% which was higher than that of the traditional approach.
4) **Acceleration control:** The success rate of the proposed MARL framework was 66.47%, which was higher than that of the traditional approach.
5) **Environmental reward:** The largest environmental reward visible was 1.37×104 (≈13,700). This shows that our method was able to achieve a high performance of the task. The higher the rewards, given our shaping, are a reflection of the more efficient progress and the safer interaction patterns during the execution of the task.
6) **Episode length:** The maximum duration of the longest episode was 319.9 steps. Such long horizons denote scenario intricacies and, in some instances, decision-making inefficiencies (e.g., congestion or overly conservative policies). We consider the length of the episode as an indicator of the difficulty of the task and the effectiveness of the policy.

We present the experimental results in tables and figures to enable side-by-side comparison across algorithms. Visual summaries are reported over standard metrics-e.g., environmental reward, success rate, collision rate, travel time/throughput, and comfort measures such as mean and peak acceleration-to provide a clear view of relative performance (As shown in Table 1).

Table 1. Comparative performance of different reinforcement learning algorithms across safety, efficiency, and comfort metrics.

Metric	Curriculum Learning (CL)	Independent Proximal Policy Optimization (IPPO)	CCPPO Mean Field (MF)	CCPPO Concat (CONCAT)	Proposed MARL
Collision Rate	0.1687	0.2023	0.2543	0.2883	0.2692
Success Rate	0.7847	0.7493	0.6749	0.636	0.6647
Mean Acc (Max)	0.7681	0.675	0.6282	0.7332	0.5091
Eny Reward (MAX)	2.7166e+4	3.1457e+4	3.0647e+4	2.9418e+4	1.37e+4
Epi Length (MAX)	623.1	623.1	488	552.5	319.9

4.4. Collision Rate Graph

The picture presents how the collision rate reduces over time for both the intersection (orange) and the roundabout (blue). Both lines trend downwards, but the intersection is characterized by larger, higher-frequency oscillations, which are in agreement with the greater environmental complexity and the higher number of conflict points. On the other hand, the roundabout is showing a less bumpy, more steady decline of the collision rate, thus suggesting more space for the vehicles and interactions that are better balanced.

4.5. Success Rate Graph

On this graph, the success rate, i.e., the percentage of episodes completed without collisions or violations, in two different environments is represented. As can be seen from the picture, the roundabout (blue) keeps a high level of success rate with a small variance throughout which means that performance is stable and reliable. On the contrary, the intersection (orange) has a very noisy success rate that varies significantly with time, thus indicating that in such a setting, maintaining steady performance is challenging.

4.6. Acceleration Mean (Max) Graph

As shown in Figure 1. The diagrams refer to the change of the average acceleration over time with the maxima marked (car pictograms represent the locations) for the intersection (orange) and the roundabout (blue). The roundabout pattern is going to be calm and rather consistent, indicating the smooth driving of the vehicle. Conversely, the intersection is all over the place with very big ups and downs-quite frequent stop-and-go manner as well as properly close gaps of the vehicles might be the reason here-thus, the local traffic control could be more effective with regards both the safety and the comfort aspects if the security measure is scaled down here.

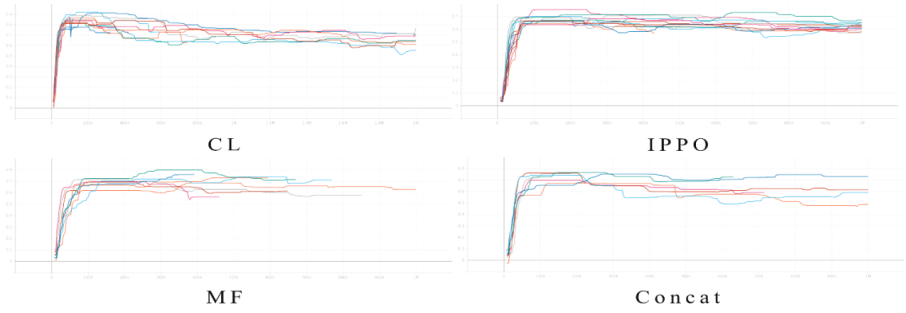

Figure 1. Acceleration Mean (Max).

4.7. Environmental Reward (Max) Graph

The environmental reward chart shows that the roundabout environment (blue line) is always giving a higher and more stable reward, which means better agent performance. The intersection environment (orange line) is characterized by higher reward variances and might benefit from further fine tuning (As shown in Figure 2).

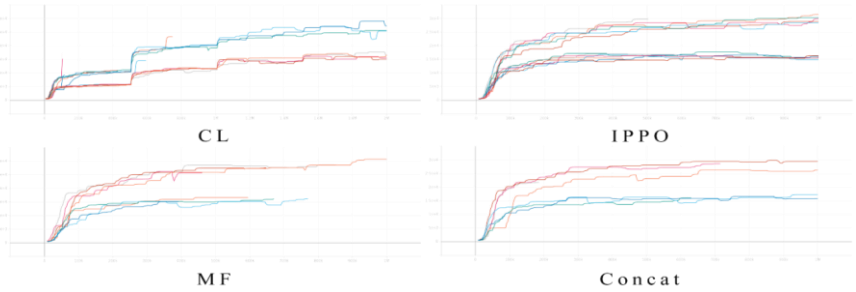

Figure 2. Environmental Reward (Max).

4.8. Episode Length (Max) Graph

As shown in Figure 3. The intersection environment depicted by the orange line can be better understood as being more able to explore the state space and the episodes being jittery for longer, thus pointing to a greater complexity. On the other hand, the roundabout environment shown by the blue line reveals shorter episode durations with smaller variances suggesting that the agents have an easier path.

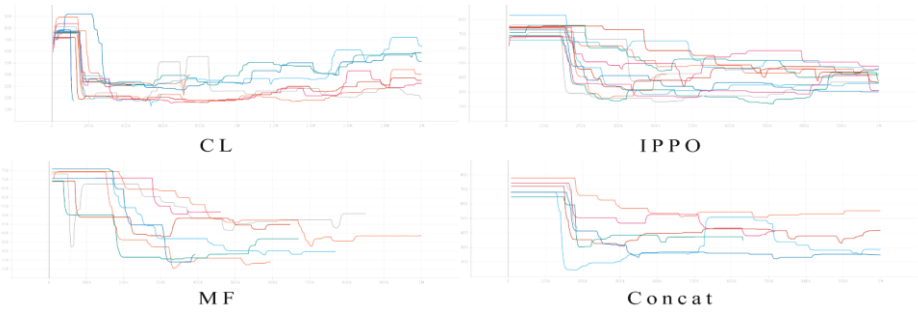

Figure 3. Episode Length (Max).

5. Conclusions

5.1. Result Analysis

For testing purposes, we conducted our experiments in the MetaDrive testbed, and the experimental results showed that the MARL framework dominates in general over the baseline models in various scenarios, all metrics measured. Basically, these upgrades may be visible through the changes in the collision rate, success rate, mean acceleration, environmental reward, episode length besides. Moreover, this is the evidence that our framework is capable of optimizing the traffic flow even in such places like the signal-free areas with complicated topology from the navigation perspective.

1) **Collision Rate:** The framework was the main reason for the system to have low collision rates, i.e., the indicators that the system was safer. We could have restricted the number of collisions by using more advanced learning methods that allowed not only efficient traffic flow but also the production of a safer driving environment.
2) **Rate of Success:** As a result of our approach, the corresponding success rate was increased in the majority of cases, thus, making sure that other road users reached the vehicles' destination in a collision- and violation-free manner while being logged under autonomous driving systems.
3) **Mean Acceleration:** The lack of abrupt acceleration changes reveals that our model promotes the issuance of driver control actions, which are very important in situations of heavy traffic, where rapid changes of speed may lead to the occurrence of inefficiencies or even accidents.
4) **Environmental Reward:** Our design enjoys greater environmental rewards, which implies that safety and speed/efficiency were better balanced in our framework, thus the agents became more efficient in different traffic environments.
5) **Reduced length of episodes:** The reduction in episodes length as a result of improved performance in task solving, was very significant for the cases with high volume traffic. The faster episodes represented by blue bars indicate a better performance, as the vehicles were able to move through complex scenarios with much less difficulty and thus got them closer (in less time) to their targets.

5.2. Limitations and Future Work

However, except the powerful results of our framework on Meta Drive, we also detected its limitationsEven though our framework yielded excellent results on Meta Drive, we also observed its restrictions:

1) **Scalability:** The framework was quite effective in scenarios of medium scale, however, it can not be extended to large networks having hundreds of agents. The development of more efficient learning algorithms to tackle such issues can be a direction for future work.
2) **Real-world Applicability:** Simulation-based methods may not be able to fully represent the features of actual driving. For example, other vehicles are unpredictably reacting to the mistakes of drivers, the different conditions of the road, and the emergent properties. Our following research steps should be to verify the extent to which this model can be dependable in the actual world or in more complex simulations.
3) **Computational Resources:** The processing power consumed by our experiments was substantial which also demonstrates that the framework needs to be further optimized before it can be deployed in resource-constrained environments. The issue of how future studies will find the best way of computing efficiency and the implementation of the framework will be resolved.

References

1. R. Chandra, and D. Manocha, "Gameplan: Game-theoretic multi-agent planning with human drivers at intersections, roundabouts, and merging," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2676-2683, 2022. doi: 10.1109/lra.2022.3144516
2. M. Mamlouk, and B. Souliman, "Effect of traffic roundabouts on accident rate and severity in Arizona," *Journal of Transportation Safety & Security*, vol. 11, no. 4, pp. 430-442, 2019. doi: 10.1080/19439962.2018.1452812
3. M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," In *Machine learning proceedings 1994*, 1994, pp. 157-163. doi: 10.1016/b978-1-55860-335-6.50027-1
4. L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156-172, 2008. doi: 10.1109/tsmcc.2007.913919
5. A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," In *Conference on robot learning*, October, 2017, pp. 1-16.
6. R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.
7. J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5068-5078, 2021. doi: 10.1109/tits.2020.3046646
8. X. Ma, J. Li, M. J. Kochenderfer, D. Isele, and K. Fujimura, "Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships," In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May, 2021, pp. 6064-6071. doi: 10.1109/icra48506.2021.9562006
9. 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, 2023. doi: 10.1109/tte.2023.3290069