Multi-Agent Reinforcement Learning for CAV Cooperative Merging Considering Communication Failure

Tingting Fan*[1], Edward Chung[1]

[1] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China

SHORT SUMMARY

Highway on-ramp merging sections are prone to oscillations caused by merging vehicles cutting into and disrupting the mainline flow. With the advancements in vehicle-to-everything (V2X) communications, more intelligent traffic control strategies have been studied to migrate on-ramp congestions using available information. However, most studies assume perfect V2X communications and complete information, overlooking the impact of communication failures that may degrade the effectiveness of cooperative merging strategies. In learning-based approaches, such failures can disrupt or distort the information provided to neural networks, leading to inaccurate decisions or predictions. This study introduces a Multi-Agent Reinforcement Learning (MARL) strategy with an attention mechanism, namely Attention-based Multi-Agent Proximal Policy Optimization (AMAPPO), for cooperative merging at highway on-ramp sections. Additionally, the significance of various V2X information sources and their impact on traffic control performance are evaluated. Experimental results demonstrate that our AMAPPO is more robust for impaired information compared to standard MAPPO.

Keywords: cooperative merging, connected and autonomous vehicles, multi-agent, reinforcement learning, communication failure.

1. INTRODUCTION

Highway on-ramp merging sections are prone to oscillations caused by merging vehicles cutting into the mainline flow and disrupting the traffic pattern. Drivers may face the dilemma of either rushing to pass or yielding to merging vehicles, leading to potentially competitive behaviors that can impact traffic flow negatively. Vehicle-to-Everything (V2X) communications brought new potential for migrating on-ramp congestions by considering the information sharing between merging and facilitating vehicles. Strategies such as determining the passing order based on principles like first-in-first-out (FIFO) or estimated arrival times, virtual mapping of vehicles from ramp lane or mainline to a unified lane, can reduce competition among vehicles and subsequently enhancing merging efficiency and safety. In recent times, the rise of data-driven methodologies, particularly reinforcement learning (RL) has received increasing attention in addressing merging control problems due to its scalability and adaptability in handling stochastic and uncertain environments. However, most of the research assume that the V2X communications are perfect and neglect the adverse impact of communication failure. Due to high vehicular mobility and signal fading in complex environments, the connectivity of vehicular networks is always fragile and intermittent (Ho et al., 2011).

The loss of communication can degrade the effectiveness of cooperative merging strategies. Research by Liu et al. (2023) investigated the impact of a merge vehicle losing connectivity, revealing that while preceding non-adjacent vehicles remain unaffected, adjacent vehicles may face

potential collisions, and non-adjacent rear vehicles could experience delays. Hence, the availability of timely and accurate vehicle state information is crucial for optimizing merging sequences and trajectory planning. As for the learning-based strategies, communication failure can directly disrupt or distort the information sources provided to neural networks. In most of the RL-based traffic control studies, researchers typically focus on model design and strategy setting. They often rely on complete training data from simulations or empirical datasets without considering the potential loss of information acquisition. Consequently, two critical oversights become apparent. Firstly, there is a risk of overestimating the performance of learning-based methods when operating with limited or degraded data. It remains uncertain whether control efficiency can remain comparable under such impaired information conditions. Secondly, the significant disparity between the unstable nature of real-world V2X information and the idealized training or simulation data. Transitioning from a well-trained model to real-world implementation becomes challenging due to this discrepancy.

As far as the existing literature in learning-based cooperative merging is concerned, limited attempts have been made for considering the instability of information. Kherroubi et al. (2022) proposed a blind actor-critic network that numerically approximate the immediate reward even incorporating loss and delay. In traffic signal control scenarios, Agarwal et al. (2021) update the state-action function using estimated transition probabilities and reward distributions. These works estimate the rewards or transitions in RL decisions to compensate the loss or delayed data collection. However, accurately modelling these elements in a non-stationary Partially Observable Markov Decision Process (POMDP) environment, where transitions are determined not only by environmental changes but also by other agents' decisions, remains challenging. Pang et al. (2024) used LSTM network to predict current state given stacked historical delayed states in traffic signal control. This work compensated the delayed state prior to the RL loop, which can also adaptable to POMDP environments. However, their approach assumed a fixed communication delay and accessible to the actual information, which may not always hold true in practical. Therefore, our work aims to address the aforementioned limitations, focusing on a learning-based cooperative merging control strategy in a non-stationary multi-agent environment characterized by impaired and inaccessible information. Specifically, the main contributions are listed below:

- Develop a cooperative merging strategy using Multi-Agent Reinforcement Learning (MARL) for Connected and Autonomous Vehicles (CAVs) to execute efficient and safe merging maneuvers in mixed CAVs and Connected Vehicles (CVs) scenarios.
- Evaluate the significance of various V2X information sources and their impacts on MARL-based traffic control effectiveness.
- Propose a novel MARL framework that is more robust in handling and compensating for lost or impaired observations caused by stochastic communication failures.

2. METHODOLOGY

Cooperative Merging as MARL

In this study, we depict an on-ramp merging road with mixed CAVs and CVs, as illustrated in the Fig. 1. Each CAV is controlled by a decentralized actor network, optimized through centralized critic network, establishing the framework of Multi-Agent Proximal Policy Optimization (MAPPO) (Yu et al., 2022). CVs, operated by human drivers, serve as background vehicles but are capable of transmitting real-time information to CAVs via Vehicle-to-Vehicle (V2V) communication channels. In addition, RSUs are responsible for gathering lane-specific statistical data and relaying this information to CAVs through Infrastructure-to-Vehicle (I2V) communication.

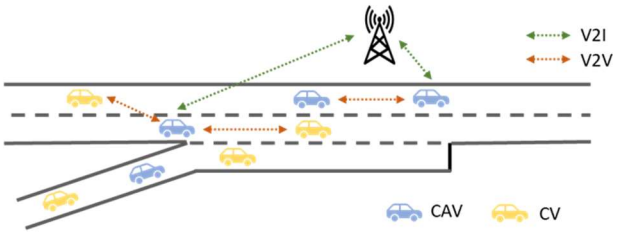

Fig. 1. On-ramp merging scenario with mixed CAVs and CVs.

AMAPPO

This work utilizes MAPPO as the primary framework, employing a centralized critic for training and decentralized actors for execution. Additionally, an attention mechanism is integrated to enhance state encoding. The framework of the proposed Attention-based MAPPO (AMAPPO) is illustrated in Fig. 2.

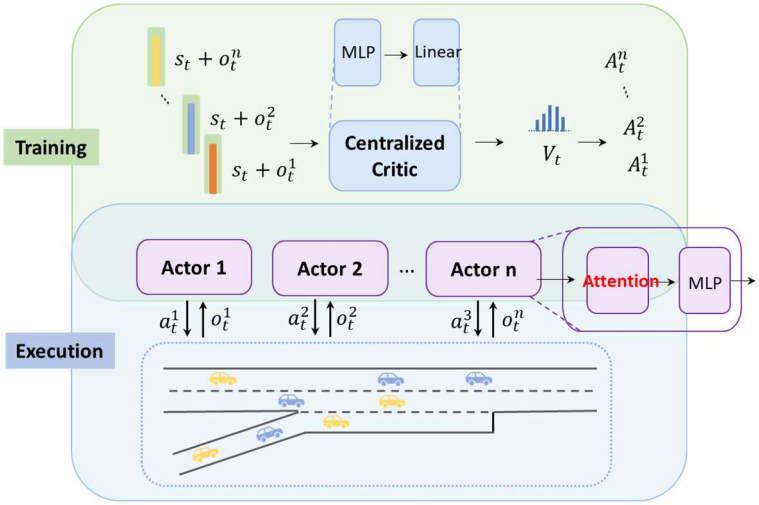

Fig. 2. The framework of AMAPPO.

During execution, each agent i control a specific CAV i, which is equipped with its individual actor network. The agent observes its local observation $\boldsymbol{o}_t^i$ and selects an action a_t^i based on the current policy. The selected actions $a_t^1, a_t^2, \ldots, a_t^n$ are then executed by the corresponding vehicles in SUMO. Subsequently, the agents receive updated observations and rewards based on SUMO's response. Throughout this process, each agent i collects a trajectory of experiences over multiple time steps T, defined as:

$$\tau_i = \{(s_t, a_t, \boldsymbol{o}_t^i, r_t, s_{t+1}, \boldsymbol{o}_{t+1}^i)\}_{t=1}^{T} \tag{1}$$

In training phase, the actor networks and the critic network are updated to optimize their parameters. The aggregated inputs $S_t^1, S_t^2, \ldots, S_t^n$ are fed into the critic network, which estimates the value function V. The objective function of the critic network is the mean squared error (MSE) loss between the predicted value and target values:

$$L(\emptyset) = \frac{1}{BN} \sum_{k=1}^{B} \sum_{i=1}^{N} \left(V_\emptyset \left(S_k^{(i)} \right) - \hat{R}_k \right)^2 \tag{2}$$

or

$$L(\emptyset) = \frac{1}{BN}\sum_{k=1}^{B}\sum_{i=1}^{N}\left(V_{\emptyset}\left(S_k^{(i)}\right) - \hat{R}_k^{(i)})^2\right) \tag{3}$$

where B is the batch size, N is the number of agents and $\widehat{\mathrm{R}}$ is reward-to-go, which is the sum of rewards after a given point in a trajectory. $\hat{R}_k$ is used when the reward is homogenous for all agents (i.e. r_t^1=$r_t^2 = \cdots = r_t^n$) and $\hat{R}_k^{(i)}$ is for the case when r_t^i is not shared between agents (i.e. $r_t^1 \neq r_t^2 \neq \cdots \neq r_t^n$). Then, the critic network parametrized by $\emptyset$ is updated by minimizing the objective function $L(\emptyset)$:

$$\emptyset \leftarrow \emptyset - \alpha\nabla_{\emptyset}L(\emptyset) \tag{4}$$

where α is the learning rate for the critic network.

Each agent's actor is updated using the PPO algorithm. The Generalized Advantage Estimation (GAE) is used to compute the advantage, reducing the variance in the policy gradient estimates. First, the Temporal Difference (TD) error is calculated using Eq. (5). The advantage is then computed by summing the discounted TD errors as Eq. (6).

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{5}$$

$$A_t^{GAE} = \delta_t + \gamma\lambda\delta_{t+1} + \gamma^2\lambda^2\delta_{t+2} + \cdots = \sum_{l=0}^{\infty}(\gamma\lambda)^l\,\delta_{t+l} \tag{6}$$

where γ is the discount factor, λ is the GAE parameter.

Then, the clipped surrogate objective $L^{CLIP}(\theta)$ for agent i is given by:

$$L^{CLIP}(\theta) = \frac{1}{B_n}\ \sum_k^B\sum_i^n min\left(\frac{\pi_\theta^i(a_t^i|o_t^i)}{\pi_{\theta_{old}}^i(a_t^i|o_t^i)}A_t^{GAE(i)}, clip(\ , 1-\epsilon, 1+\epsilon)A_t^{GAE(i)}\right) \tag{7}$$

The parameter sharing strategy is used among agents to promote learning efficiency and coordination, the shared parameter θ for agents is updated by maximizing the surrogate objectives:

$$\theta \leftarrow \theta - \alpha\nabla_{\theta}L^{CLIP}(\theta) \tag{8}$$

The attention mechanism weights the importance of different features in an input sequence, allowing the model to focus on the most relevant aspects of the state, which is especially crucial when dealing with impaired features.

Local observations

- **Ego vehicle basic info ($\boldsymbol{e_t^i}$)**:

Basic information about the ego vehicle i detected by its own sensors, such as speed, longitudinal position, lateral position, and heading.

- **Surrounding vehicles info ($\boldsymbol{e_t^{-i}}$)**:

Dynamics of surrounding vehicles obtained through Vehicle-to-Vehicle (V2V) communication. The six nearest surrounding vehicles' states relative to the ego vehicle, including their relative longitudinal position, relative lateral position, relative speed, are transmitted to the ego vehicle via V2V communications.

$$\boldsymbol{e}_t^{-i} = \left\{e_t^j \middle|\, j \in \mathcal{N}_i(t)\right\} \tag{9}$$

Where $\mathcal{N}_i(t)$ is the set of indices of the 6 nearest vehicles to vehicle i at time t. In cases where there are fewer than six vehicles in the vicinity of the ego vehicle, zero-padding is applied to ensure a consistent data structure.

- **Ego lane info ($\boldsymbol{l_t^i}$)**:

Statistical lane of ego vehicle gathered from RSUs, including road id, lane id, lane density, lane length, mean speed, current waiting times, accumulated waiting time.

Therefore, o_t^i can be represented as:

$$\boldsymbol{o_t^i} = \{\boldsymbol{e_t^i}, \boldsymbol{e_t^{-i}}, \boldsymbol{l_t^i}\} \tag{10}$$

where $i \in \{1,2,\dots,N\}, N$ is the number of CAVs.

Global state

The environment-provided global state, which consists of general global information about traffic, including all lanes statistic $\boldsymbol{l_t^k}$ and all CAVs basic information $\boldsymbol{e_t^i}$, are used for homogenous part for critic input. Apart from that, we also introduced agent-specific features for heterogenous part, i.e. local observation for each vehicle $\boldsymbol{o_t^i}$. The global state for agent i can be denoted as follows:

$$S_t^i = \{\boldsymbol{l_t^k}, \boldsymbol{e_t^i}, \boldsymbol{o_t^i}\} \tag{11}$$

Action

Each local actor network generates actions for its controlled CAV. The available actions $a \in \{0, 1, 2, 3, 4\}$ are defined as follows:

- 0: Maintain current state (remain constant)
- 1: Change lane to the left
- 2: Change lane to the right
- 3: Increase speed by 2 m/s
- 4: Decrease speed by 2 m/s

Invalid actions are masked according to the following rules to prevent the generation of infeasible maneuvers:

- If the current lane of the ego vehicle is the leftmost lane, the vehicle cannot change lane to the left.
- If the current lane of the ego vehicle is the rightmost lane, the vehicle cannot change lane to the right.
- If the ego vehicle is on a ramp, the vehicle cannot change lanes.

Reward

There are two components of reward design, instant reward incentivizes higher speeds and prompt completion, and penalizes unsafe gap distance. Final reward encourages the overall completion of all vehicles in the system, which designed to motivate cooperation and coordination among the CAVs. Conversely, the system imposes penalties for collision events.

(1) Instant reward

- Average speed for all running vehicles: $r_t^{speed} = \sum_i^n \frac{v_i}{v_{max}}$.
- Warn distance: $r_{t,i}^{warn} = \begin{cases} \frac{g-g_s}{g_s}, g < g_s \\ 0, \quad else \end{cases}$, where g is gap and g_s is safe gap.
- Ramp end penalty: $r_{t,i}^{end\ cost} = -\exp\left(-\frac{d^2}{10L}\right)$
- Time penalty: $r_{t,i}^{time} = -1$

Instant reward for each agent i at time t leads to:

$$r_t^i = r_t^{speed} + r_{t,i}^{warn} + r_{t,i}^{end\ cost} + r_{t,i}^{time} \tag{12}$$

Where r_t^{speed} is the homogenous part for all agents, and the last three terms are heterogenous between each agent i. Specifically, $r_{t,i}^{warn}$ is the penalty when the gap of vehicle between its leader or follower smaller than the predefined safety gap g_s. $r_{t,i}^{end\ cost}$ is the penalty for vehicle approach to the ramp end, the smaller the distance, denoted by d, between vehicle i and ramp end, the larger the penalty imposed on the vehicle. L is the length of accelerate lane.

(2) Final Reward

$$r_T \begin{cases} T_{max} - T, & if\ completes\ successfully \\ -200, & if\ collision\ happens \\ -100, & if\ T > T_{max} \end{cases} \tag{13}$$

Final reward is the incentive or disincentive triggered at the last time step T. In the event of successful completion, the reward is inversely proportional to the completion time T. Conversely, if a collision occurs or the time limit is exceeded, a substantial penalty is enforced,\

Loss of Observation

To account for potential communication instabilities leading to the loss of observations from surrounding vehicles via V2V and ego lane information via I2V, a probability p_{PL} is introduced, which is the likelihood that each piece of information transmitted from either surrounding vehicles or infrastructure lane data is lost during communication. The binary indicators are introduced to model the communication links considering packet loss:

$$\delta_{PL} \sim Bernoulli(1 - p_{PL}) \tag{14}$$

Therefore, the local observation with packet loss can be represented as:

$$\boldsymbol{o'}_t^i = \{ \boldsymbol{e}_t^i, \delta_{PL} \boldsymbol{e}_t^{-i}, \delta_{PL} l_t^i \} \tag{15}$$

3. RESULTS AND DISCUSSION

Simulation Settings

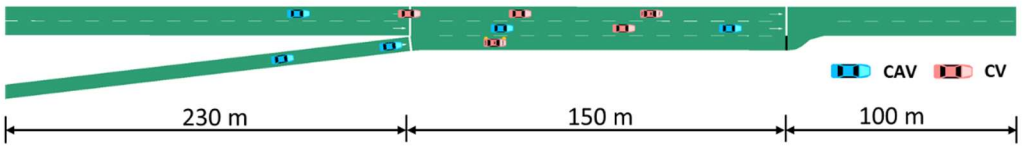

Fig. 3. The road network setting.

The case study is conducted on an on-ramp merging road network with single ramp lane connected to an acceleration lane and a two-lane mainline, as shown in Fig. 3. The traffic simulation SUMO, is used for model mixed CAV and CV environment. Five CAVs are integrated into the network with a 50% penetration rate. All vehicles, including the CAVs and CVs, commence their journeys from random starting positions and follow randomized paths within the simulated environment. CAVs are completely controlled by MARL agents and CVs are executed by SUMO default car-following and lane-changing model.

Effect of different reward settings

We evaluate the agents' performance under various reward designs to identify the optimal configuration for the MARL system. Different rewards for training are listed below:

- Global: The reward for each agent is shared among all agents. It is calculated based on the collective performance metrics, such as the average speed of all vehicles, collision penalized to all vehicles, for instant rewards and final rewards, respectively.
- Local: Rewards are determined by the agent's own incentive for efficiency and penalties for collision. There is no common final reward in this setting.
- Local and global: Combines both individual performance-based rewards and shared rewards based on overall efficiency and penalties.

As shown in Fig. 4, model with local reward achieves the highest efficiency in terms of average speed and completion duration. However, this efficiency comes at the expense of more competitive behaviour, resulting in a higher collision rate. On the other hand, global rewards prioritize overall cooperation, leading to conservative actions and a low collision rate. Thus, there exists a trade-off between traffic efficiency and safety. The combination of local and global reward can achieve a compromised performance in terms of efficiency and safety metrics.

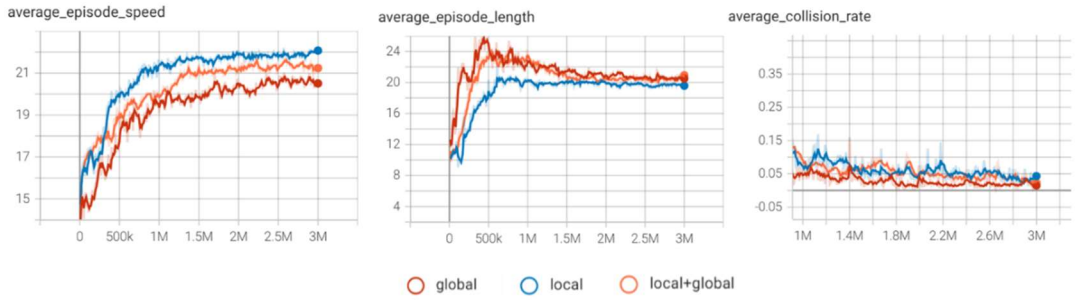

Fig. 4. Training performances of different reward setting.

Effect of different information loss

To evaluate the significant of local and global information for MAPPO cooperative merging, we modelled three scenarios outlined in Table 1. Scenario 1 operates under the perfect communication scenario, while scenario 2 and 3 are under packet loss rate $p_{PL} = 0.2$, which has probability of lost information for each communication links. For processing the lost information, we padded the lost features to zero. We neglect the communication loss between RSU and critic network, which assumed to transmitted via infrastructure-to-central, therefore all lanes states obtained from the RSU are constant accessible.

Table 1. Various scenarios of information loss

Info. Source	Local info. for decentralized actors		Global info. for centralized critic	
	Surrounding vehicles info e_t^{-i}	Ego lane info, l_t^i	All vehicles info e_t^i.	All lanes states l_t^i
Scenario 1	√	√	√	√
Scenario 2	$p_{PL} = 0.2$	$p_{PL} = 0.2$	√	√
Scenario 3	$p_{PL} = 0.2$	$p_{PL} = 0.2$	$p_{PL} = 0.2$	√

The training results under various communication scenarios are shown in Fig. 5. It is obvious that training performance is optimal when communication is perfect. The absence of local information has less impact on overall traffic performance, while the lack of global information significantly affects both traffic efficiency and safety. Thus, we can conclude that utilizing global information for the critic network is crucial for MAPPO-based cooperative merging control.

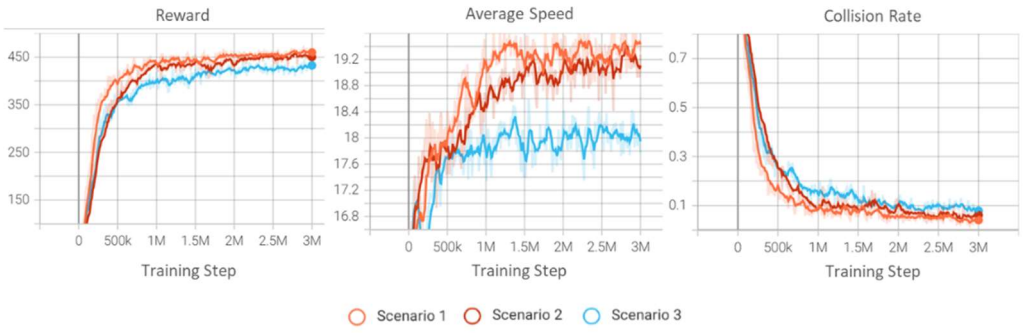

Fig. 5. Training performance in various scenarios

Comparison of AMAPPO and MAPPO

The performances of AMAPPO and MAPPO are evaluated under p_{PL} of 0.4, and a benchmark comparison was made using MAPPO with full information (p_{PL}=0). As shown in Fig. 6, while reward and collision rate are relatively less affected by severe information impairments, there was a noticeable negative impact on traffic efficiency metrics such as average speed and completion time. This outcome can be attributed to the significant weight assigned to collision penalties, prioritizing safety over traffic efficiency in the reward settings. Consequently, the trends in reward and collision rate show similar variations across different environmental conditions.

Despite these similarities, AMAPPO present improved merging efficiency in terms of average speed and completion time compared to MAPPO when information loss rate reaches 0.4, which suggests the adaptability of AMAPPO in coping with severe information impairments.

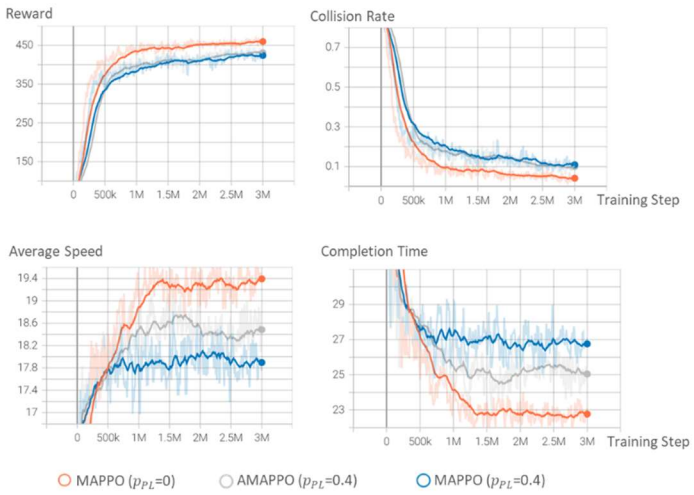

Fig. 6. Comparison of MAPPO and AMAPPO under severe information loss

4. CONCLUSIONS

This study develops a MARL-based cooperative strategy for CAVs executing efficient and safe merging maneuvers, and evaluates the impact of impaired information on the performances of MARL algorithms. Experimental results demonstrate that implementing a combination of global and local rewards can be a balance to ensure both safety and efficiency. Furthermore, our AMAPPO algorithm is more robust in scenarios with impaired observations, particularly presenting improved efficiency levels.

REFERENCES

Agarwal, M., & Aggarwal, V. 2021. Blind decision making: Reinforcement learning with delayed observations. Pattern Recognition Letters, 150, 176–182.

Ho, I. W.-H., North, R., Polak, J., & Leung, K. 2011. Effect of Transport Models on Connectivity of Interbus Communication Networks. Journal of Intelligent Transportation Systems, 15, 161–178.

Kherroubi, Z. el abidine. 2024. Novel Actor-Critic Algorithm for Robust Decision Making of CAV under Delays and Loss of V2X Data (arXiv:2405.05072).

Liu, Q., Zhang, J., Zhong, W., Li, Z., Ban, X. (Jeff), Li, S., & Li, L. 2023. Fault-Tolerant cooperative driving at highway on-ramps considering communication failure. Transportation Research Part C: Emerging Technologies, 153, 104227.

Pang, A., Wang, M., Chen, Y., Pun, M.-O., & Lepech, M. 2024. Scalable Reinforcement Learning Framework for Traffic Signal Control Under Communication Delays. IEEE Open Journal of Vehicular Technology, 5, 330–343.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), Advances in Neural Information Processing Systems 35 - 36th Conference on Neural Information Processing Systems, NeurIPS 2022 (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1787, 24611–24624.