Deep Reinforcement Learning based Joint Optimization of the Traffic Signal and Autonomous Vehicles Considering Dynamic Uncertainties

Bin Zhou[1,2], Zhengyang Zhang[2], Panagiotis Angeloudis[3], and Simon Hu[1,2,*]

[1]ZJU-UIUC Institute, Zhejiang University, China
[2]College of Civil Engineering and Architecture, Zhejiang University, China
[3]Department of Civil and Environmental Engineering, Imperial College London, UK

Short summary

The increasing availability of sensor data and advancements in deep reinforcement learning (DRL) offer promising opportunities for enhancing traffic management efficiency. However, most existing DRL methods for traffic management concentrate solely on enhancing signal controls. To address this limitation, we propose a novel DRL-based joint optimization method that integrates traffic signal control with the longitudinal control of connoted autonomous vehicles (CAV). This method enhances overall traffic efficiency by synchronizing signal control with CAV behaviour. To alleviate dynamic uncertainty problems in joint optimization, this method leverages temporal prediction networks, ideal for anticipating future traffic states, and fuzzy neural networks, adept at handling information with uncertainties and errors, to extract useful and accurate traffic representations in mixed-traffic environments. A case study across various isolated intersections is conducted to validate the effectiveness of the proposed method. The results demonstrate that our method outperforms state-of-the-art methods on both synthetic and real-world datasets.
Keywords: Deep reinforcement learning, adaptive traffic signal control, intelligent traffic systems, fuzzy logic inference

1 Introduction

Traffic congestion in urban cities has been becoming a worldwide problem due to population growth and urbanization, which not only exacerbates environmental pollution but also results in significant economic losses (Q. Wu et al., 2019). One effective way to alleviate traffic congestion is by implementing more efficient traffic signal control (TSC) systems. Traditional TSC methods, such as fixed-time control and self-organizing traffic light control, have been proven inefficient due to their reliance on predefined assumptions or manually designed rules (L. Wu et al., 2021).

Subsequently, the adaptive traffic signal control (ATSC) system has proposed to alleviate traffic congestion by adjusting the signal timing dynamically based on real-time traffic data. Recently, many Deep Reinforcement Learning (DRL) methods have been utilized in the ATSC system and achieved promising results. DRL is a promising Artificial intelligence (AI), that directly interacts with and adapts to complex and dynamic traffic environments, learning optimal traffic control policies without the need for pre-defined models or assumptions (Ma et al., 2022; L. Wu et al., 2021).

With the improvement of infrastructure and the advancement of autonomous driving technologies, Connect Autonomous Vehicles (CAV) are increasingly being recognized as integral components of traffic systems. CAVs are expected to enhance traffic efficiency through Vehicle Trajectory Planning (VTP), a strategy that optimizes vehicle movements to alleviate congestion and shorten travel times by considering the timing of the intersection signal and the state of traffic flow (Li et al., 2023). Most current research in VTP is concentrated on the trajectory planning of individual CAVs or vehicle platoons. Han et al. (2020) employed gap-feedback-control to guide vehicles with varying speeds and headways into optimal trajectories, treating all CAVs in one platoon as a collective unit. Yu & Long (2022) developed a Model Predictive Control model for planning CAV trajectories based on signal timing schemes and the behavior of the leading vehicle, to reduce fuel consumption.

Notably, the penetration of autonomous driving technology into the mainstream will be gradual. We are still far from achieving a high level of CAV penetration or a fully CAV-dominant environment (Q. Guo et al., 2019). Human-Driven Vehicles (HDVs) will continue to share road resources

with CAVs for the foreseeable future, creating a mixed traffic environment of manual and autonomous driving. Consequently, more studies have focused on cooperative control for ATSC and CAV in mixed traffic scenarios, incorporating both HDVs and CAVs. remains a prominent area of focus in global traffic control theory and application (Wang et al., 2018). The mixed traffic system is a vast and intricate nonlinear, time-varying structure intertwined with various uncertainties. The challenge of dynamic uncertainty lies in the continuous variation of traffic conditions over time and response to the evolving traffic environment.
In response to the aforementioned issues, our contributions are summarized as follows:

- We propose a DRL-based method designed for cooperative control involving both traffic lights and CAVs, namely Signal-Vehicle Cooperative Control with Fuzzy Inference and Temporal Prediction (SVCCFT). The proposed method employs a parallel structure to achieve synchronized traffic management, accommodating two distinct types of agents simultaneously.
- Our method strengthens cooperative control by leveraging two primary perspectives. Firstly, it employs fuzzy inference to address uncertainties in input and output variables, accurately inferring the traffic state from imprecise collected data. Secondly, a temporal prediction module extracts valuable time information from continuous traffic flow, minimizing uncertainties related to traffic timing and optimizing the future phase strategy of the TSC.
- We validated the proposed method in two distinct scenarios. By conducting multiple experiments and comparisons with other TSC and VTP methods, we illustrated the significance of fuzzy inference and temporal prediction in the cooperative control of traffic lights and CAVs. The results indicate that the proposed SVCCFT method outperforms other TSC and VTP methods in terms of both efficiency and eco-friendly perspectives.

2 Methodology

This section primarily introduces the SVCCFT, a method designed to enhance traffic efficiency by optimizing signal timing and CAV behavior.

Signal-Vehicle Cooperative Control Problem

This paper investigates a signal-vehicle cooperative control in a single-intersection. Our focus lies on optimizing primary objects: CAV and traffic signal light. This control problem could be modeled as a standard Markov Decision Process (MDP), wherein two types of agents — representing the CAVs and the TSC — make optimal decisions. In our method, it is crucial to note that CAV agents exclusively control the frontmost CAVs in each incoming lane. To provide a clear understanding of the issue and the terminology involved, we've illustrated four typical intersection scenarios in Fig. 1. If an incoming lane does not contain CAVs, the CAV agents refrain from controlling any vehicles in that lane, thereby maintaining the natural flow of traffic, as demonstrated in Fig. 1 (d).
Typically, the MDP problem is formulated by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. At time t, the agent observes the environment state $s_t \in \mathcal{S}$ and then chooses an action $a_t \in \mathcal{A}$ based on s_t. After taking action, the agent receives feedback in the form of numerical reward r_t, reflecting the effectiveness of the action. The agent's objective is to develop a policy π mapping from states to actions $a = \pi(s)$ with the ultimate goal of maximizing the cumulative discounted future rewards. The calculation for the return G_t of the cumulative discounted future rewards can be expressed as follows:

$$G_t = \sum_{k=0}^{T} \gamma^k r_{t+k} \tag{1}$$

where T represents the horizon and γ is a discount factor for future rewards between 0 and 1.
In our study, we implement the Proximal Policy Optimization (PPO) algorithm, as proposed by Schulman et al. (2017), for both types of agents we are examining. A key advantage of PPO is its versatility, it performs effectively in scenarios that require either discrete or continuous actions. This versatility is especially beneficial in our application, where CAV agents use continuous actions for speed control and traffic signal agents employ discrete actions for signal timing. Our study focuses exclusively on the longitudinal acceleration and deceleration behavior of CAVs without lane-changing operations, a scope that aligns with the research conducted by J. Guo et al. (2023).

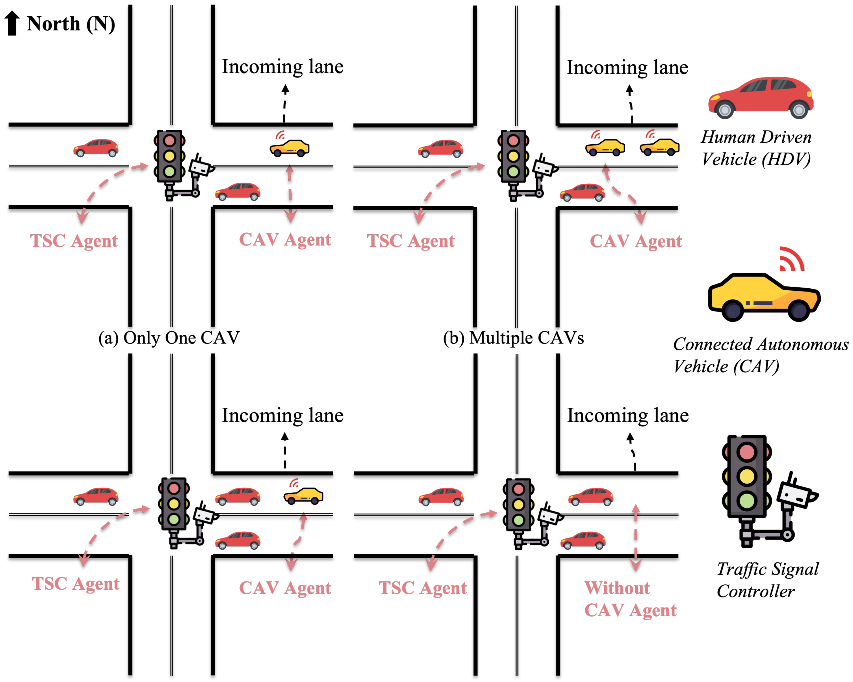


Figure 1: Four typical intersection scenarios of Cooperative Control.

PPO is designed to optimize policy decision-making while balancing exploration and exploitation. Typically, the objective function in PPO is expressed as follows:

$$L(\theta) = E_t \left[\min \left(z_t(\theta)\hat{A}_t, clip\left(z_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t \right) \right] \tag{2}$$

where $z_t(\theta)$ is the probability ratio of the new policy to the old, θ denotes the parameters of our neural network that approximates the policy. In the context of our DRL design, actions like phase switching for traffic signals or acceleration adjustments for CAVs are determined by this policy. The advantage estimation $\hat{A}_t$ is pivotal for updating the policy. It is calculated based on rewards, such as reduced pressure at intersections or improved efficiency of CAV platoons, and value functions, which reflect the long-term benefits of actions under the current policy. The hyperparameter ϵ specifies the clipping range, ensuring that the updates to the policy are not too drastic. In the context of our DRL design for traffic lights and CAVs, θ represents the parameters of our neural network that approximates the policy. Actions such as phase switching (for TSC) or acceleration adjustments (for CAVs) are guided by this policy. The advantage estimates $\hat{A}_t$, crucial for policy update, is computed based on the rewards (like reduced intersection queue length or improved CAV platoon efficiency) and value functions, reflecting the long-term benefit of actions under the current policy.

Overall Architecture

The overall architecture of the SVCCFT model is illustrated in Fig. 2. It consists primarily of a Fuzzy Inference module and a Temporal Prediction module, indicated by the grey dashed boxes. The Fuzzy Inference module relies on the Fuzzy Neural Network (FNN), established as an efficient approach for addressing the uncertainty inherent in traffic information (Bi et al., 2017). Furthermore, the temporal prediction module is grounded in the Recurrent Neural Network (RNN), distinguished as a powerful tool for temporal prediction, adept at recognizing and forecasting time patterns—a critical capability for decoding temporal states (L. Wu et al., 2021). It is crucial to note that multiple CAV agents, each assigned to an individual CAV, collaborate to manage the CAVs. Furthermore, our method exclusively addresses the longitudinal behavior of the CAVs, omitting lane-changing operations, in line with the method proposed by J. Guo et al. (2023). All operations of these agents, encompassing the TSC agent and CAVs agents, are performed concurrently, ensuring a cohesive and efficient traffic system. To provide further clarity on our method, we outline the design of actions, states, and rewards for two distinct types of DRL agents as follows:

A. State: *CAV state* includes original information (speed, acceleration, distance to the intersection, signal phase) and inference of accurate traffic information from the Fuzzy Inference Module. *TSC state* encompasses real-time information and prediction information. Real-time traffic information comprises both signal-related data (current phase) and vehicle-related data, including the number of vehicles per lane and detailed information about the closest CAV approaching the intersection. This detailed information contains speed, acceleration, distance to the intersection, and lane ID. Conversely, prediction information refers to the forecasted number of vehicles entering and leaving the intersection within the next second. ***B. Action:*** *CAV action* focuses on the lead CAV in each lane, using continuous acceleration within $[-3m/s^2, 3m/s^2]$ for effective platoon formation. Maximum speed is capped at $15m/s$. *TSC action* adopts phase switching, restricting actions to a binary set (1 for changing the current phase, 0 for maintaining). Time constraints with a 3-second yellow phase enhance road safety and traffic flow. ***C. Reward:*** *CAV reward* considers information from all CAVs in the same lane as the given CAV agent. It encourages CAVs to operate within a predefined safe speed range and achieve smooth acceleration and deceleration. For the given CAV agent C, the reward of the CAV agent, denoted as R_C, is defined as follows:

$$R_C = -\frac{\sum_{n \in N} (v^* - v_n)}{v^* \times N} - \sqrt{\frac{\sum_{n \in N} a_n^2}{a^{*2} \times N^2}} \tag{3}$$

where v^* denotes the predefined safe speed limitation of the CAV, a^* represents the CAV's predefined maximum acceleration, and N denotes the number of CAVs in the given lane.

TSC reward is the penalty of intersection pressure, normalized by average lane capacity c. The intersection pressure is defined as the difference between the sum of the number of vehicles on the incoming lanes N_{in} and the sum of the number of vehicles on the outgoing lanes N_{out}. The final reward is computed as follows:

$$R_L = -\frac{N_{in} - N_{out}}{c} \tag{4}$$

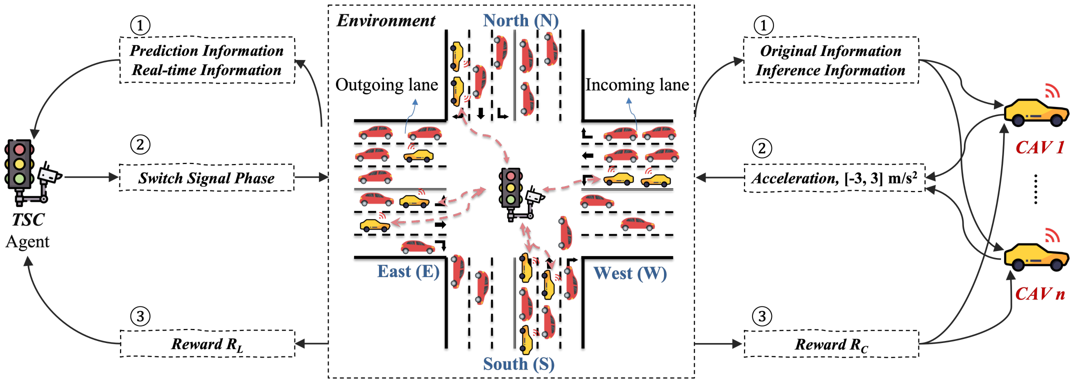


Figure 2: Framework of our proposed method. All processes with the same number occur in parallel.

3 Results and discussion

Simulation scenarios

Our study employs the Flow simulation platform, as proposed by C. Wu et al. (2021), which leverages the capabilities of Simulation of Urban MObility (SUMO) (Lopez et al., 2018) for detailed and realistic traffic microsimulation, and the Reinforcement Learning Library (RLlib) for implementing sophisticated DRL algorithms. For the experiments, we used both synthetic and real-world traffic datasets with a 10% CAV penetration rate. In the case of DRL-based methods, we employed 18 episodes for the training of each method.

- **Synthetic dataset:** Following the dataset outlined by J. Guo et al. (2023), our road network is constructed with each edge comprising two roads running in opposite directions. The length of each road is set at 300 meters, and the maximum speed limit is established at 54 km/h. Our synthetic dataset only contains four directions of go-straight traffic flows. Specifically, the traffic flow rates, measured in the number of vehicles per hour for each road, are as follows: 288 from North to South, 240 from West to East, 192 from East to West, and 120 from South to North.

- **Real-world datasets:** The dataset employed was derived from a real-world scenario at the intersection of Longwang Sha Road and Shanhu Sha Road in Hangzhou, China. This intersection comprises four directions, each spanning 300 meters long and featuring two lanes. Traffic movements on the incoming lanes include left-turn and shared straight/right-turn. The traffic scenario, as depicted in Table 1, reflects the actual traffic conditions.

Table 1: Configurations for Real-world Dataset.

Directions	Flow (Veh/h/lane)	Left Turn Rate
SN	267	0.11
NS	240	0.16
WE	180	0.20
EW	275	0.14

Evaluation metrics and compared algorithms

The performance of methods is assessed using four metrics, divided into two categories: *Traffic efficiency* and *Environmental friendliness. Traffic efficiency* includes 'Travel Time', measuring vehicles' travel duration, and 'Delay', indicating waiting time at intersections. *Environmental friendliness* encompasses 'Fuel Consumption', the average fuel used per 100 kilometres, and 'CO_2 Emissions', the carbon dioxide emitted per kilometre. These metrics are derived using FLOW's integrated vehicle information and emission models.
In our study, we compare SVCCFT with several baseline methods, including several state-of-the-art methods, categorized into two groups: Non-DRL Based Method, and DRL Based Method.
Non-DRL Based Method

- FixedTime: A conventional method with pre-set signal phases. Green phases last 40 seconds and yellow phases 3 seconds.
- GLOSA: A non-DRL method adjusting CAV speed according to traffic light phases and CAV status (Suzuki & Marumo, 2018).

DRL Based Method

- PressLight: A state-of-the-art DRL method minimizing intersection pressure, based on Max-Pressure theory, utilizing a pressure-centric reward function (Wei et al., 2019).
- FlowCAV: A DRL-based model controlling lead CAV speed per lane to form platoons and enhance traffic efficiency (C. Wu et al., 2021).
- CoTV: A DRL method for cooperative control, focusing on harmonizing traffic light controllers with lead CAVs in each lane to balance traffic efficiency and environmental considerations (J. Guo et al., 2023).

Experiment results

Table 2 presents the comparative results in the synthetic dataset. The proposed SVCCFT outperforms all the other methods in four different evaluation metrics. In contrast to the conventional fixed time method, our method has achieved an improvement of over 15% in reducing fuel consumption and carbon dioxide emissions, and more than a 17% enhancement in average travel time and average delay. GLOSA, which is a non-DRL based method, also needs to rely on deterministic algorithms and lacks the necessary flexibility to adapt to the dynamic changes of real-world traffic conditions. Therefore, although GLOSA tries to optimize the timing of traffic lights and vehicle speed at the same time, it has failed to effectively solve the changing traffic scenarios. DRL-based methods show their advantages because they do not rely on predetermined assumptions about the environment, are not limited by deterministic formulas, and show commendable adaptability. PressLight and CoTV have made significant improvements in all benchmarks, reducing fuel consumption and CO_2 emissions by more than 12% and 13% respectively. However, the performance of FlowCAV is slightly lower than that of the fixed-time strategy, mainly due to its static traffic light strategy and the CAV agent's neglect of the current traffic signal.

Table 3 displays the results of our analysis of the real-world dataset. Our method has yielded impressive results, surpassing the FixedTime method by improving over 30% in all four metrics. Particularly noteworthy is the remarkable 50% enhancement achieved in the delay metric. Consistent with the findings in Table 2, DRL-based methods generally outperform non-DRL based methods. Among them, methods with signal-vehicle cooperative control, such as COTV and our method, are leading the way. Furthermore, our methods outperform all other methods across all metrics. Overall, our method has accomplished its primary system objectives, leading to a reduction in travel time and delay, lowered fuel consumption, and minimized CO_2 emissions.

To evaluate the impact of various penetration rates, we conducted six experiments on CAV penetration rates using real-world datasets: 10%, 20%, 40%, 60%, 80%, and 100%. The results, as illustrated in Fig. 3, demonstrate that our method surpasses a competitive method across all penetration rates and maintains consistent performance throughout the various CAV penetration rates. Our method exhibits relative stability in terms of fuel consumption, CO_2 emission, travel time, and delay. This stability suggests that our method can adapt to variations in traffic conditions and the proportion of CAV, ensuring reliable and efficient traffic flow management and environmental impact. Such stability is especially critical for overall traffic management and environmental impact.

Table 2: Results on the synthetic dataset. The displayed percentages represent improvements compared to the fixed time method. The best results are in bold.

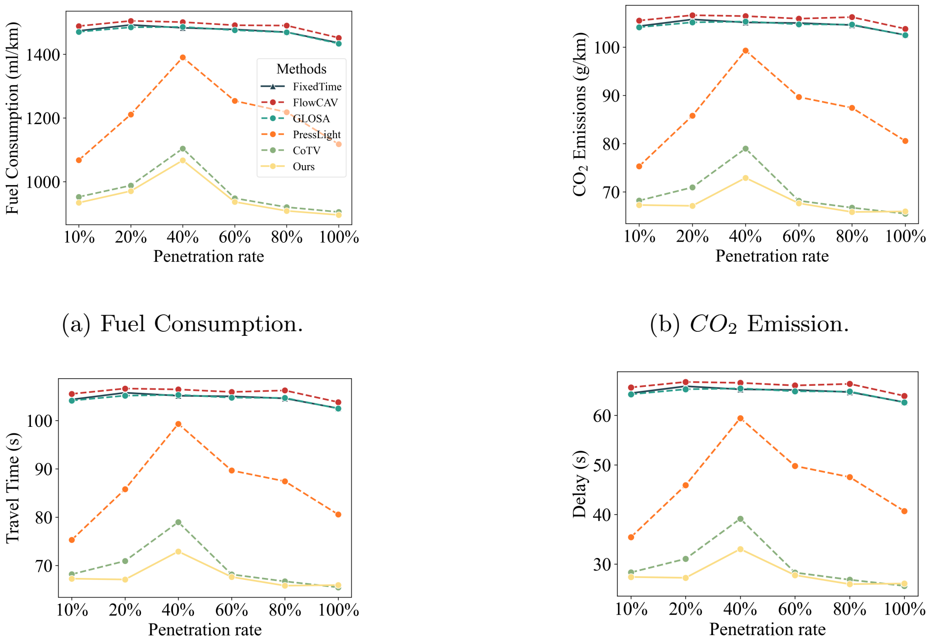
Methods	**Fuel consumption(ml/km)**	CO_2 **emission (g/km)**	**Travel time (s)**	**Delay (s)**
FixedTime	838.95	263.03	58.99	17.21
GLOSA	728.28 -13.19%	228.33 -13.19%	50.27 -14.79%	8.49 -50.69%
PressLight	732.52 -12.69%	229.66 -12.69%	49.90 -15.41%	8.12 -52.83%
FlowCAV	841.07 +0.25%	263.69 +0.25%	59.14 +0.26%	17.36 +0.90%
CoTV	727.58 -13.28%	228.11 -13.27%	49.27 -16.47%	7.49 -56.46%
Ours	**771.15** **-15.19%**	**223.08** **-15.19%**	**48.94** **-17.04%**	**7.16** **-58.41%**

Table 3: Results on the real-world dataset. The displayed percentages represent improvements compared to the fixed time method. The best results are in bold.

Methods	**Fuel consumption(ml/km)**	CO_2 **emission (g/km)**	**Travel time (s)**	**Delay (s)**
FixedTime	1472.85	461.76	104.32	64.44
GLOSA	1469.31 -0.24%	460.65 -0.24%	104.08 -0.23%	64.20 -0.37%
PressLight	1354.33 -8.05%	424.60 -8.05%	96.57 -7.42%	56.70 -12.02%
FlowCAV	1487.33 +0.98%	466.30 +0.98%	105.43 +1.06%	65.55 +1.72%
CoTV	1011.09 -31.35%	316.99 -31.35%	72.30 -30.69%	32.43 -49.67%
Ours	**970.67** **-34.10%**	**291.66** **-36.84%**	**67.09** **-35.69%**	**27.21** **-57.77%**

4 Conclusions

This paper presents an innovative method for vehicle-infrastructure collaborative control that incorporates fuzzy reasoning and temporal forecasting into a DRL model, specifically designed to



(a) Fuel Consumption.

(b) CO_2 Emission.

(c) Travel Time.

(d) Delay.

Figure 3: Performance on CAV Penetration Rate Variation.

address dynamic uncertainty issues in traffic management. Our method employs a parallel framework to coordinate the interaction between the CAV and the traffic light controller, ensuring smooth traffic flow while accommodating two distinct agent types. Our experiments have demonstrated that this method effectively reduces fuel consumption and CO_2 emissions while improving traffic efficiency at a single intersection. Additionally, our method has been rigorously tested under varying CAV penetration rates, indicating its potential to establish a sustainable intelligent transportation system.

Acknowledgements

This study was supported in part by the National Key R&D Program of China (2023YFB4302600), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2023C03155), the National Natural Science Foundation of China key program project (52131202), the Smart Urban Future (SURF) Laboratory, Zhejiang Province, the Sustainable Smart Liveable Cities Research Centre, Zhejiang Province, Zhejiang University Global Partnership Fund, the ZJU-UIUC Joint Research Centre Project of Zhejiang University (DREMES202001), the Zhejiang University Sustainable Smart Livable Cities Alliance (SSLCA). It was led by Principal Investigator Simon Hu.

References

Bi, Y., Lu, X., Sun, Z., Srinivasan, D., & Sun, Z. (2017). Optimal type-2 fuzzy system for arterial traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, *19*(9), 3009–3027.

Guo, J., Cheng, L., & Wang, S. (2023). Cotv: Cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*.

Guo, Q., Li, L., & Ban, X. J. (2019). Urban traffic signal control with connected and automated vehicles: A survey. *Transportation research part C: emerging technologies*, *101*, 313–334.

Han, X., Ma, R., & Zhang, H. M. (2020). Energy-aware trajectory optimization of cav platoons through a signalized intersection. *Transportation Research Part C: Emerging Technologies*, *118*, 102652.

Li, J., Yu, C., Shen, Z., Su, Z., & Ma, W. (2023). A survey on urban traffic control under mixed traffic environment with connected automated vehicles. *Transportation research part C: emerging technologies*, *154*, 104258.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., ... Wießner, E. (2018). Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (itsc)* (pp. 2575–2582).

Ma, D., Zhou, B., Song, X., & Dai, H. (2022). A deep reinforcement learning approach to traffic signal control with temporal traffic pattern mining. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Suzuki, H., & Marumo, Y. (2018). A new approach to green light optimal speed advisory (glosa) systems for high-density traffic flowe. In *2018 21st international conference on intelligent transportation systems (itsc)* (pp. 362–367).

Wang, Y., Yang, X., Liang, H., Liu, Y., et al. (2018). A review of the self-adaptive traffic signal control system based on future traffic environment. *Journal of Advanced Transportation*, *2018*.

Wei, H., Chen, C., Zheng, G., Wu, K., Gayah, V., Xu, K., & Li, Z. (2019). Presslight: Learning max pressure control to coordinate traffic signals in arterial network. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1290–1298).

Wu, C., Kreidieh, A. R., Parvate, K., Vinitsky, E., & Bayen, A. M. (2021). Flow: A modular learning framework for mixed autonomy traffic. *IEEE Transactions on Robotics*, *38*(2), 1270–1286.

Wu, L., Wang, M., Wu, D., & Wu, J. (2021). Dynstgat: Dynamic spatial-temporal graph attention network for traffic signal control. In *Proceedings of the 30th acm international conference on information & knowledge management* (pp. 2150–2159).

Wu, Q., Shen, J., Yong, B., Wu, J., Li, F., Wang, J., & Zhou, Q. (2019). Smart fog based workflow for traffic control networks. *Future Generation Computer Systems*, *97*, 825–835.

Yu, M., & Long, J. (2022). An eco-driving strategy for partially connected automated vehicles at a signalized intersection. *IEEE transactions on intelligent transportation systems*, *23*(9), 15780–15793.