Synergistic Reinforcement Learning Models for Pedestrian-Friendly Traffic Signal Control

Desong Chen, Junyan Hu, *Senior Member, IEEE*, Hao Zhang, and Boli Chen, *Senior Member, IEEE*

***Abstract*— Traffic signal control is essential for managing urban traffic, reducing congestion, and minimizing environmental impact by optimizing both vehicular and pedestrian flow. This paper investigates the application of Reinforcement Learning (RL) in traffic signal control within mixed traffic environments, emphasizing the development of a synergistic RL approach, named Advantage Actor-Critic with Maximum Pressure (A2CMP). A2CMP leverages actor-critic techniques in combination with real-time pressure metrics to dynamically adjust traffic signals based on prevailing traffic conditions. Additionally, the paper introduces a pedestrian-friendly phase-skipping mechanism for further enhancing the efficiency of the proposed algorithm in real-world traffic management. Simulation results across diverse traffic scenarios show significant reductions in CO_2 emissions and waiting time. Particularly, A2CMP can reduce waiting time by 12% compared to other RL-based algorithms.**

I. Introduction

Managing vehicle throughput amid growing urban traffic congestion has become a major challenge for cities around the world. In the European Union, traffic congestion has significant economic consequences, costing billions of euros annually and affecting both environmental and financial sustainability. Key contributors to congestion include the volume of traffic entering intersections exceeding that of traffic exiting, cross-blocking caused by downstream lane obstructions, and green idling, where green lights occur without vehicle movement [1]. Traditional traffic signal control (TSC) systems, which were based on fixed-time intervals, are unable to adapt to fluctuating traffic conditions. To address this, adaptive TSC systems such as the Split Cycle Offset Optimization Technique (SCOOT) and the Sydney Coordinated Adaptive Traffic System were developed [2]. These systems, which rely on sensor data or manual adjustments, have been successfully implemented in many cities but come with higher costs and technical complexities. With advancements in communication and control strategies, ongoing research in TSC remains essential [3].

Max-Pressure (MP) control, initially developed for packet scheduling in wireless networks [4], was first adapted for urban traffic by [5], significantly outperforming fixed-time controls [6]. While the MP controller offers several advantages, including ease of implementation and certain stability properties, it lacks optimality guarantees, which motivates the exploration of optimization-based approaches. Typically, optimal control-based TSC frames the problem as a model-based optimization challenge. However, these model-based methods are often limited by assumptions that do not adequately reflect real-world complexities [7]. In contrast, Reinforcement Learning (RL) learns optimal strategies from data and has the potential to overcome these limitations by directly engaging with the dynamic nature of complex traffic systems, making it particularly well-suited for TSC [8].

Pioneering work in single intersection RL began with the advancement of adaptive methods [9], which was later extended to multi-intersection frameworks [8], [10], [11]. These advancements build upon the foundation of single-agent algorithms to explore coordinated interactions between multiple agents [12], and existing RL-based TSC commonly includes value-based methods such as Q-learning, policy-based approaches like policy gradients, and Actor-Critic methods, as reviewed in [13]. Regardless of single or multi-agents, most RL-based TSC research has largely overlooked the integration of pedestrian considerations, with limited focus on active management of pedestrian signals. Given the essential role of pedestrian signals in urban traffic environments, neglecting their impact could result in suboptimal signal control strategies, reduced intersection efficiency, and even unsafe behavior, such as jaywalking due to prolonged pedestrian wait times [14], [15], [16]. Current RL-based TSC simulations typically synchronize green pedestrian signals with vehicle phases [17]. However, this practice may unintentionally delay right-turning vehicles, exacerbating congestion [18]. Some studies suggest dedicated pedestrian-only phases to improve overall throughput, but these often increase vehicle wait times [19], [20].

Motivated by the limitations of recent advancements in TSC, in this paper, we propose a RL-based traffic signal control model considering both pedestrians and vehicles. Our key contributions are: 1) A hybrid traffic signal control algorithm is proposed for mixed traffic scenarios, dynamically adjusting phase duration by integrating the MP algorithm with RL techniques. Unlike existing RL-based TSC approaches, which enforce a minimum green period for pedestrian crossings regardless of pedestrian volume—potentially reducing overall traffic efficiency—the proposed algorithm incorporates a phase-skipping technique. Pedestrian phases are skipped until both the aggregated waiting time and pedestrian count reach a specified threshold, optimizing traffic flow without compromising pedestrian needs; and 2) a comprehensive assessment of the algorithm's performance was conducted in SUMO using metrics for both traffic efficiency and environmental sustainability. Our results show that the proposed method outperforms existing RL-based TSC algorithms, delivering improvements for all road users.

The paper is organized as follows: Section II outlines the

This work has been supported by the State Key Laboratory of Intelligent Green Vehicle and Mobility under Project No. KFZ2405.

D. Chen, and B. Chen are with the Department of Electronic and Electrical Engineering, University College London, London, UK. (desong.chen.23@ucl.ac.uk; boli.chen@ucl.ac.uk).

J. Hu is with the Department of Computer Science, Durham University, Durham, UK. (junyan.hu@durham.ac.uk)

H. Zhang is with the State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China. (hao_thu@foxmail.com)

problem statement and presents the necessary preliminaries. Section III, elaborates on the proposed methodology. Section IV presents the experimental setup and analyzes experimental results. The paper is finally concluded in Section V.

II. Problem Statement and Preliminaries

As illustrated in Fig. 1, we consider a single-agent TSC problem for a four-way road intersection. Each road consists of two lanes: the inner lanes are designated for through and left-turning traffic, while the outer lanes accommodate through and right-turning vehicles. Sidewalks are positioned

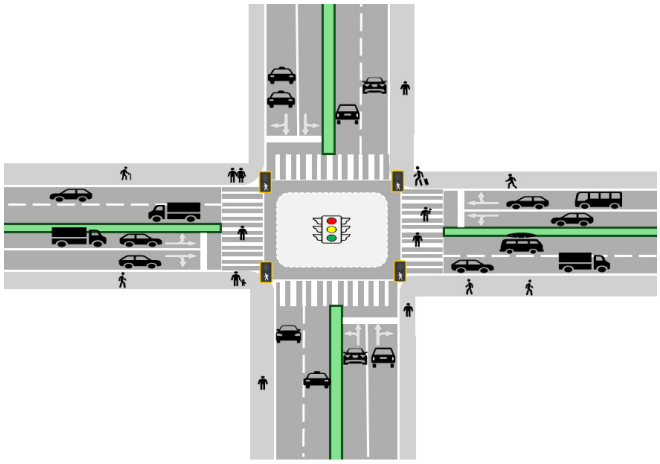

Fig. 1. Four-way intersection with pedestrian crossing.

along the outer edges. The traffic signals at the four-way intersection manage a total of eight vehicle lanes and four pedestrian crossings, using six phases (as illustrated in Fig (3)) to approximate the overall phase space. Each vehicle and pedestrian passing through the intersection is regulated by the respective traffic signals. The minimum green light duration is set based on the time needed for pedestrians to cross safely, with fixed yellow and red clearance intervals between green and red lights to ensure traffic safety. By optimally assigning the duration of each traffic signal phase, the goal is to optimize traffic flow, maximize throughput, and minimize congestion. The intersection environment encompasses current phase P, phase duration D, waiting time κ, and queue length q. Real-time state information s_t from the environment is assumed to be available to the central coordinator, who is responsible for traffic light control.

Next, the MP algorithm is briefly recalled as it will be utilized in the proposed algorithm. In principle, the MP algorithm generates a vector named pressurelist, PL, which records the phase priorities as follows: $PL = [\text{link}_i, \text{link}_j, \text{link}_p, \text{link}_q]$. The links are prioritized from highest to lowest and stored in the PL. As shown in Fig. 2, the link pressure $x_{ij}(t)$ counts for waiting vehicle and pedestrian numbers from link i to j, and it is calculated by

$$x_{ij}(t+1) = x_{ij}(t) + a_{ij}(t) - \min\{x_{ij}(t), c_{ij}s_{ij}(t)\}, \quad \forall (i,j) \in \mathcal{M} \tag{1}$$

where $a_{ij}(t)$ represents the exogenous demand from link i to link j. The parameter c_{ij} stands for the saturation flow rate, which represents the maximum possible flow through the link. The signal state $s_{ij}(t)$ takes a value of 0 or 1, corresponding to red or green phases. Additionally,

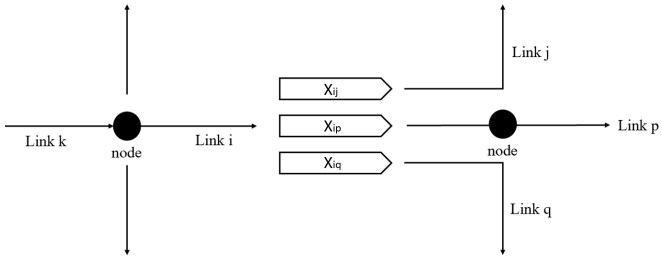

Fig. 2. MP store-and-forward model [6].

$\mathcal{M}$ represents the set of all movements. Based on the above discussion, TSC should actively consider pedestrian lights in a simulation environment closely resembling real-world conditions, rather than focusing solely on vehicle light optimization. While hybrid action spaces can significantly enhance TSC performance, actor-critic hybrid algorithms have not yet been fully explored. In this paper, we aim to bridge these gaps.

III. Main Methodology

A. Pedestrian-friendly signaling system

The primary goal of the pedestrian-friendly signaling system is to dynamically adjust signal timing based on real-time pedestrian demand, making the system responsive to varying traffic and pedestrian conditions. Traditional pedestrian signals are often based on fixed timing, which can lead to inefficiencies, particularly when pedestrian volumes fluctuate throughout the day. In contrast, a dynamic system can enhance traffic flow by optimizing signal timing to accommodate both vehicles and pedestrians effectively. When a pedestrian presses the button and remains within a designated waiting area, the system can classify them as "waiting" and adjust the signal timing accordingly. The pedestrian-friendly mechanism considers the pedestrian's position, intended crossing route, and waiting time, which are then used to calculate and optimize signal timing. As shown in Fig. 3, the 6 phases include straight-through and left-turn movements, followed by an all-direction phase. Phase 2 is subsequently succeeded by clearance intervals before transitioning to equivalent movements in the opposite direction. If no pedestrians meet the waiting criteria, Phases 1 and 4 are skipped, and the sequence follows the alternative path indicated by the purple loop in Fig.3. Otherwise, the sequence will continue to loop according to the phase number shown in Fig. 3.

This pedestrian-friendly mechanism can be implemented using a combination of sensors, such as LIDAR, infrared cameras, or even smart city technologies that connect directly with traffic management systems. By implementing centralized training on a single agent, the system can effectively manage complex urban traffic scenarios.

B. Synergistic RL-based signal control

To leverage the flexibility and optimality of the RL algorithm while maintaining the baseline performance of MP, a hybrid algorithm called A2CMP is proposed, as illustrated in Fig 4. It can be seen that both the RL and MP controllers operate concurrently to facilitate action selection. The design of the algorithm is detailed below.

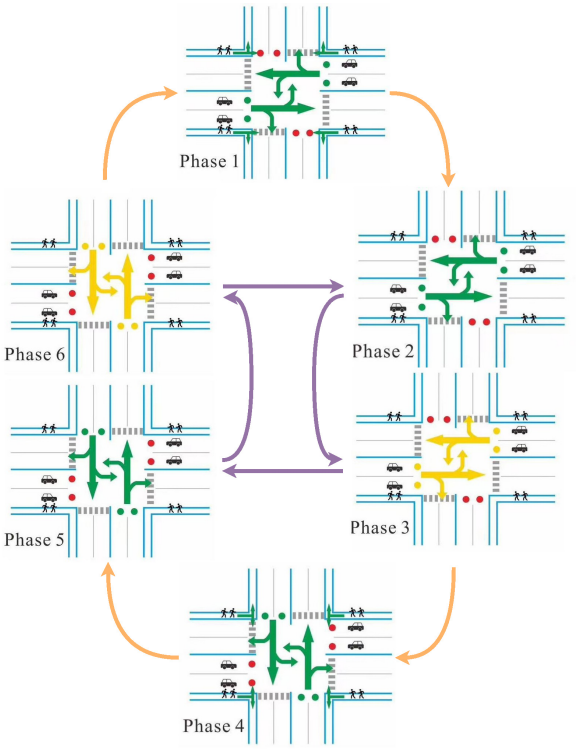

Fig. 3. Pedestrian-friendly signaling system for a standard four-way intersection with eight incoming lanes.

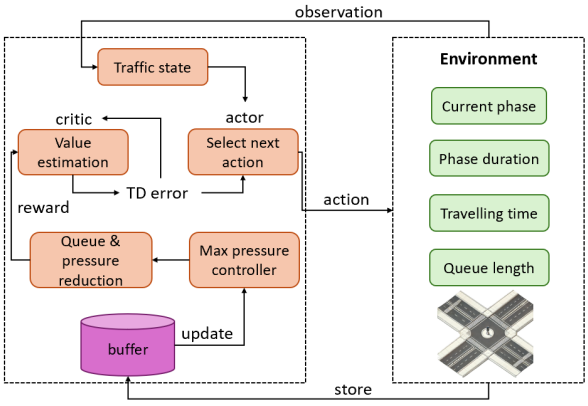

Fig. 4. Schematic framework of the proposed A2CMP.

1) State (Observation): The state s_t consists of a vehicle matrix $V_m(t) \in \mathbb{R}^{|L|\times|x_v|}$, which represents the number of waiting vehicles x_v (with velocity under threshold) for each lane $l \in L$. Meanwhile, a pedestrian matrix $V_p(t) \in \mathbb{R}^{f|L|\times|x_p|}$ captures the number of waiting pedestrians x_p with $f|L|$, reflecting the typically smaller number of pedestrian pathways relative to vehicle lanes. The phase is represented by a 3D matrix $N \in \mathbb{R}^{|P|\times|D|\times|L|}$, which encodes the current phase configuration, including the current phase P, phase durations D, and L. Transition dynamics, denoted by probability $Pr(s_{t+1}|s_t, u_t)$, are learned implicitly through interactions with the environment rather than through explicit modeling. Finally, the intersection state is defined as $s_t = (V_m(t), V_p(t), N)$.

2) Action: The agent selects a parameterized action u_t from the hybrid action space $U \in \mathbb{R}^{|P|\times|D|}$ based on the observation o_t, drawing from a candidate action space according to the policy $\pi(u|o)$. u_t is then remapped to a real-valued range, ensuring a constrained phase sequence and duration.

3) Reward: When the environment executes the selected action, it transitions to a new state s_{t+1} and provides a reward r_{t+1} to the agent. The reward function $r(s_t, u_t)$ is designed to reduce congestion for both pedestrians and vehicles, with values normalized to be non-positive. Traffic flow is optimized effectively by balancing immediate and future rewards through a discount factor $\gamma \in [0, 1]$. Specifically, the reward function is designed as follows

$$r_t = -W_q^\top q - W_\kappa^\top \kappa + r_{mp,t+1} \tag{2}$$

where $q = [q_p, q_v]^\top$, $\kappa = [\kappa_p, \kappa_v]^\top$ with q_p, q_v denoting the total queue lengths of pedestrians and vehicles, respectively, and κ_p, κ_v represent their corresponding waiting times. The weight vectors $W_q = [w_{q,p}, w_{q,v}]^\top$ and $W_\kappa = [w_{\kappa,p}, w_{\kappa,v}]^\top$ determine the relative importance of each component in the reward function.Moreover, r_{mp} is a component proportional to the rank P within the prioritization list PL which accounts for both vehicle- and pedestrian-based phase prioritization.

This reward incentivizes alignment between the actions taken by the MP and A2C controllers. Consequently, the proposed TSC algorithm will update by value gradient $L(w)$ and policy gradient $L(\delta)$ as follows:

$$\begin{aligned} L(w) &= \frac{1}{2|\mathcal{B}|}\sum_{t\in\mathcal{B}}(\tilde{R}_t - V_w(\tilde{s}_t))^2, \\ L(\delta) &= -\frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}}\log\pi_\delta(u_t|\tilde{s}_t)\tilde{A}_t - \beta\sum_{u\in U}\pi_\delta\log\pi_\delta(u|\tilde{s}_t), \\ \tilde{A}_t &= \tilde{R}_t - V_w(\tilde{s}_t), \end{aligned} \tag{3}$$

where $V_w(\tilde{s}_t)$ represents the value function and the current total reward is $\tilde{R}_t$. The policy parameters δ define the behavior policy π_δ, which informs action selection. The value loss function measures the difference between the total reward and the estimated value, while the policy loss incorporates the advantage function $\tilde{A}_t$ and a regularization parameter β to encourage exploration.

The pseudo-code of the proposed algorithm is given in Algorithm 1. As it can be noticed, α is a weighting factor that influences the contribution of the critic, while η_w and η_δ are learning rates for the critic and actor networks, respectively. $\mathcal{B}$ represents the minibatch buffer that stores collected experiences for training. The parameter T defines the number of time steps per training episode k. $\hat{R}$ represents the estimated return at the update time τ.

IV. Simulation Validation

A. Environment setup and benchmark algorithms

This paper utilizes a simulation model based on SUMO to compare the proposed A2CMP method[1]. Fig. 5 illustrates the networks used in the simulation. The single intersection is derived from a four-way intersection in downtown Washington, D.C. The network topology is directly extracted from the OpenStreetMap dataset.

[1]The code for implementing the A2CMP can be accessed from GitHub repository https://github.com/chendesong/A2CMP.git.

Algorithm 1 A2CMP

Parameters: α, β, γ, T, $|\mathcal{B}|$, η_w, η_δ, PL;
Output: w, δ
Set up: s_0, π_{-1}, $t = 0$, $k = 0$, $\mathcal{B} = \emptyset$, $PL = \emptyset$;
repeat
  Calculate $x_{i,j}(t)$ using (1)
  Append $(u_t(mp), x_{i,j}(t))$ to PL;
  Exploration:
  for single intersection **do**
    Sample u_t from π_t;
    Receive r_t and s_{t+1};
  end for
  Update $\mathcal{B} \leftarrow \mathcal{B} \cup \{(t, s_t, \pi_t, u_t, r_t, s_{t+1})\}$;
  Increment $t \leftarrow t + 1$, $k \leftarrow k + 1$;
  if $t = T$ **then**
    Set up s_t, π_{-1}, $t \leftarrow 0$;
  end if
  if $k = |\mathcal{B}|$ **then**
    Evaluate $\hat{R}_\tau$, $\forall \tau \in \mathcal{B}$;
    Evaluate $\tilde{R}_\tau$, Bonus r_{mp};
    Update $\eta_w \nabla L(w)$ using (3);
    Update $\eta_\delta \nabla L(\delta)$ using (3);
    Initialize $\mathcal{B} \leftarrow \emptyset$, $k \leftarrow 0$;
  end if
until Total timesteps reached

Fig. 5. Network structure of a four-way single intersection in Washington, D.C., used for numerical experiments.

In our SUMO-based TSC simulation [21], the evaluation of CO_2 emissions is derived using models such as the Handbook Emission Factors for Road Transport and Passenger car and Heavy-duty Emission Model [22]. These models accurately estimate emissions by considering vehicle types, driving patterns, and fuel consumption. Additional performance parameters, such as queue length and waiting time, are obtained through the simulation's output. By analyzing the delay and congestion metrics for both vehicles and pedestrians, a comprehensive understanding of traffic efficiency and environmental impact is achieved.

The traffic demand was calibrated using historical video data from the evening peak hours, with one-hour traffic and pedestrian flow as representative samples [23]. Four datasets of synthetic vehicle and pedestrian flows shown in Fig 6 were utilized to compare the differences between peak and off-peak periods. The simulation model was based on vehicle and pedestrian parameters provided in Table I.

We compared our method with five different approaches,

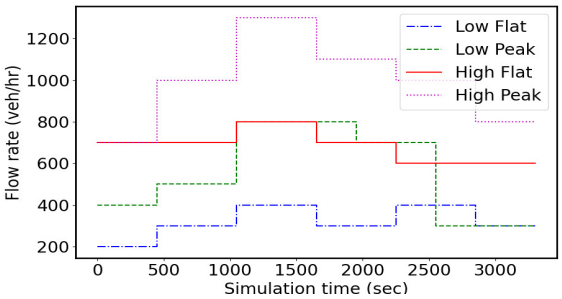

Fig. 6. Configurations of four different flow demands for synthetic traffic data in TSC experiments.

TABLE I
DEFAULT SETTINGS FOR VEHICLES, PEDESTRIANS, AND TRAFFIC LIGHTS

Parameter	Value
Vehicle Settings	
Acceleration (max *accel*)	2.6 m/s^2
Deceleration (max *decel*)	4.5 m/s^2
Sigma (*sigma*)	0.5
Length (*length*)	5 m
Min Gap (*minGap*)	2.5 m
Max Speed (*maxSpeed*)	13.89 m/s
Impatience Level	0.5
Pedestrian Settings	
Acceleration (*accel*)	1.3 m/s^2
Deceleration (*decel*)	1.5 m/s^2
Sigma (*sigma*)	0.5
Length (*length*)	0.25 m
Min Gap (*minGap*)	0.5 m
Max Speed (*maxSpeed*)	1.5 m/s
Impatience Level	0.5
Traffic Light Settings	
Minimum Green Time (*min_green*)	5 s
Maximum Green Time (*max_green*)	50 s
Yellow Light Time (*yellow_time*)	2 s
Red Clearance Interval (*clear_time*)	1 s
Delta Time (*delta_time*)	5 s
Max Depart Delay (*max_depart_delay*)	3000 s

including two RL methods, DQN and SAC, and two baseline TSC approaches, fixed-time and SCOOT. To ensure a fair comparison, all RL methods are trained without a pre-training process. DQN addresses the issues faced by traditional Q-learning and Sarsa and utilizes deep neural networks to represent the Q function, allowing it to handle high-dimensional state spaces and generalize effectively. In addition, the use of replay experience and a target network enhances learning efficiency and stability. The powerful feature extraction capabilities of deep learning make DQN more capable of avoiding local optima and finding globally optimal strategies. SAC is an off-policy algorithm that achieves stable policy optimization through dual Q-networks and policy entropy regularization. Compared to an on-policy algorithm, SAC is better suited for complex continuous control scenarios. By benchmarking against SAC, we rigorously evaluate the performance of A2CMP in challenging environments.

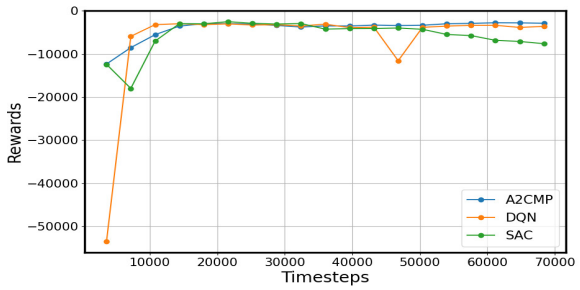

Fig. 7. Training performance and reward convergence of A2CMP and other RL baselines across the learning episodes.

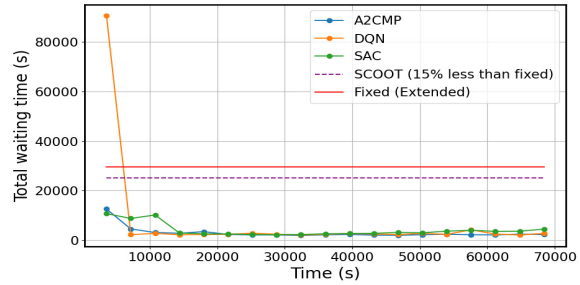

Fig. 8. Comparison of total waiting time between the proposed A2CMP method and established benchmarks.

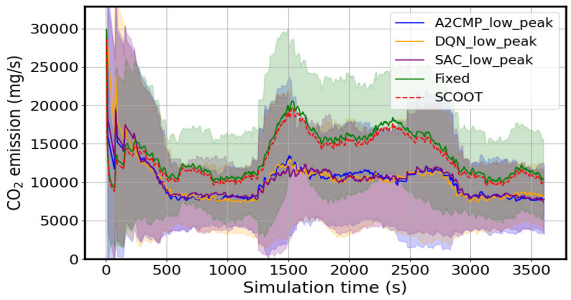

Fig. 9. Comparison of CO_2 emissions between the proposed A2CMP method and established benchmarks.

For the five controllers mentioned, we assess traffic signal performance using queue length and CO_2 emissions as metrics. The action interval is set to 5 seconds, with each run consisting of a one-hour simulation. Results are reported as the average of the last 5 runs during testing.

B. Simulation Results

For illustrative purposes, we begin by presenting results using data from the low-peak period, as shown in Fig. 6. To reveal the benefit of the pedestrian-friendly signaling system, the proposed A2CMP is compared with the same TSC but without such a signaling mechanism (as such, only the signals for the vehicles are controlled while the signals for the pedestrians simply follow). The results show that the proposed solution can save 4.65% waiting time for both vehicles and pedestrians. In the following, the pedestrian-friendly mechanism is integrated in all benchmark methods for a fair comparison.

The convergence comparison is performed by calculating the average reward of the agent,$r = \frac{1}{T}\sum_{t=0}^{T-1} r_t$, in each training episode, which quantifies the task completion level. A2CMP begins to gradually converge during episode 4, with a convergence speed slightly slower than the DQN algorithm shown in Fig. 7. However, the results of A2CMP after convergence are more stable compared to SAC and DQN. This increased stability can be attributed to the inherent advantage of the A2CMP algorithm, which combines the benefits of the actor-critic architecture with a more refined pressure control mechanism. This combination helps in reducing the variance of policy updates, leading to smoother and more consistent performance after convergence.

In terms of actual CPU computation time, A2CMP requires approximately 33 seconds per run, while DQN takes around 20 seconds. SAC, due to its complexity, needs about 5 minutes to complete a single run. In Fig. 8, A2CMP outperforms the existing fixed-time traffic signal control by reducing the total waiting time by 91%. These results are averaged over multiple evaluation episodes to ensure statistical reliability. It also shows significant improvement over SCOOT. When comparing with other RL-based algorithms, the waiting time of the A2CMP is 15% less than DQN and 40% less than SAC. From an environmental perspective, the reduction in CO_2 emissions is not as obvious as the reduction in waiting time. Specifically, although the RL algorithm can save about 25% of CO_2 emissions compared to fixed-time control, the CO_2 emissions produced by the three RL algorithms are largely similar. This can be attributed to two main factors: marginal benefit and vehicle technology. First, once traffic flow optimization reaches a certain threshold, the marginal gains in reducing CO_2 emissions diminish. Second, simulation does not account for renewable vehicle models, which may limit the observed emission reduction, even with optimized traffic flow.

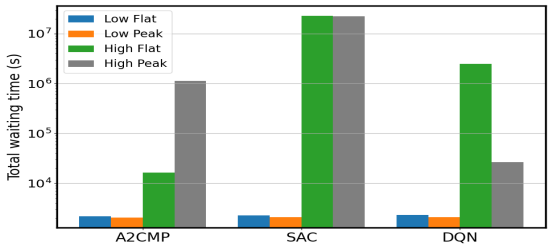

Fig. 10. Validation results comparing the proposed A2CMP method with SAC and DQN under different flow demand scenarios.

Fig. 10 illustrates the performance of three RL TSCs across four different traffic flow scenarios given in Fig. 6. The primary metrics evaluated include the total wait time of road users at intersections. The data on total waiting time and total queued number align well. In low-density scenarios, A2CMP and DQN perform similarly, both outperforming the SAC algorithm. In high-density scenarios, A2CMP generally surpasses DQN. In conditions of excessive traffic saturation, DQN slightly outpaces A2CMP in terms of traffic throughput. The lag in response exhibited by A2C is a key factor contributing to its suboptimal performance in highly discrete environments. From a sustainability perspective, A2CMP marginally surpasses other algorithms. Specifically, A2CMP can cut CO_2 emissions by 1.84% and 4.70% compared to DQN and SAC, respectively. Overall, A2CMP proves to be a reliable algorithm, delivering the best overall performance across the four metrics in single intersection scenarios.

To further evaluate the performance of the proposed

TABLE II
COMPARISON OF ALGORITHMS ACROSS TWO CITIES

Controller	City	Total waiting time (min)	CO_2 emissions (kg/s)
DQN	**Monaco**	313.7	6.5
	London	176.8	4.4
SAC	**Monaco**	379.6	6.5
	London	210.5	4.1
A2CMP	**Monaco**	273.9	6.4
	London	157	4.3
SCOOT	**Monaco**	1032.4	8.7
	London	711.5	5.9
Fixed Time	**Monaco**	1215.9	10.2
	London	839	6.9

A2CMP across different road intersection configurations, we conducted tests at a three-way intersection in London and a five-way intersection in Monaco. The comparative results are shown in Table II. The vehicle dataset for Monaco is obtained from SUMO scenarios [24], while the London dataset is derived from traffic signal camera data [25]. The results obtained align closely with the observations from previous results. Moreover, A2CMP and DQN are computationally efficient in all case studies, enabling faster decision-making than other RL-based approaches. These findings validate the superior performance and robustness of the A2CMP algorithm compared to conventional RL-based traffic management approaches, highlighting its potential for effective real-world implementation in mixed-traffic environments.

V. CONCLUSION AND FUTURE WORK

The key contributions of this paper include the development of a pedestrian-friendly signaling system and the A2CMP algorithm, which combines DRL and MP, to optimize traffic signal control in complex urban settings. A2CMP effectively reduces queue lengths, CO_2 emissions, and overall waiting time, all while maintaining acceptable computational costs, proving to be a robust and adaptable solution for urban traffic management. The results demonstrate A2CMP's superiority over traditional and other RL-based methods, though there remains room for improvement under conditions of excessive congestion. Future research will explore the real-world applicability of A2CMP, particularly its performance in handling complex traffic scenarios such as accidents and signal noise, which have not been addressed in this paper. These advancements will be crucial for developing efficient and safe intelligent transportation systems in real-world environments.

REFERENCES

[1] H. Wei, G. Zheng, V. Gayah, and Z. Li, "A survey on traffic signal control methods," *arXiv preprint arXiv:1904.08117*, 2019.

[2] R. Zhu, S. Wu, L. Li, P. Lv, and M. Xu, "Context-aware multiagent broad reinforcement learning for mixed pedestrian-vehicle adaptive traffic light control," *IEEE Transactions on Intelligent Transportation Systems*, no. 1, pp. 1–10, 2022.

[3] 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.

[4] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[5] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 177–195, 2013.

[6] X. Wang, Y. Yin, Y. Feng, and H. X. Liu, "Learning the max pressure control for urban traffic networks considering the phase switching loss," *Transportation Research Part C: Emerging Technologies*, vol. 140, p. 103697, 2022.

[7] H. Zhao, C. Dong, J. Cao, and Q. Chen, "A survey on deep reinforcement learning approaches for traffic signal control," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108100, 2024.

[8] M. Yazdani, H. Parineh, M. Sarvi, S. Asadi Bagloee, N. Nassir, and J. Price, "Intelligent vehicle pedestrian light (IVPL): A deep reinforcement learning approach for traffic signal control," *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 102942, 2023.

[9] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. Cambridge: MIT Press, 1998, vol. 135.

[10] Y. Chen and C. G. Cassandras, "Adaptive traffic light control for competing vehicle and pedestrian flows," in *2024 European Control Conference (ECC)*, 2024, pp. 1875–1880.

[11] P. Agand, A. Iskrov, and M. Chen, "Deep reinforcement learning-based intelligent traffic signal controls with optimized co2 emissions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[12] W. Genders and S. Razavi, "An open-source framework for adaptive traffic signal control," *Journal of Transactions on Intelligent Transportation Systems*, vol. X, no. X, August 2019, arXiv:1909.00395.

[13] V. Gayah, H. Wei, and G. Zheng, "Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 2, pp. 12–23, 2020.

[14] Y. Chunhui, M. Wanjing, H. Ke, and Y. Xiaoguang, "Optimization of vehicle and pedestrian signals at isolated intersections," *Transportation Research Part B: Methodological*, vol. 98, pp. 135–153, 2017.

[15] M. M. Ishaque and R. B. Noland, "Trade-offs between vehicular and pedestrian traffic using micro-simulation methods," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 9, pp. 857–873, 2007.

[16] Q. Yang and R. F. Benekohal, "Multi-objective traffic signal optimization for emissions reduction and delay minimization using evolutionary algorithms," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 1, pp. 82–98, 2011.

[17] R. Zhu, S. Wu, L. Li, P. Lv, and M. Xu, "Context-aware multiagent broad reinforcement learning for mixed pedestrian-vehicle adaptive traffic light control," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 19 694–19 705, 2022.

[18] T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang, and D. O. Wu, "Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8243–8256, 2020.

[19] W. Ma, Y. Liu, and K. L. Head, "Optimization of pedestrian phase patterns at signalized intersections: A multi-objective approach," *Journal of Advanced Transportation*, vol. 48, no. 8, pp. 1138–1152, 2014.

[20] K. Xu, J. Huang, L. Kong, J. Yu, and G. Chen, "Pv-tsc: Learning to control traffic signals for pedestrian and vehicle traffic in 6g era," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7552–7563, 2023.

[21] M. Noaeen, A. Naik, L. Goodman, J. Crebo, T. Abrar, Z. S. H. Abad, A. L. C. Bazzan, and B. H. Far, "Reinforcement learning in urban network traffic signal control: A systematic literature review," *Expert Systems with Applications*, vol. 199, p. 116830, 2022.

[22] O. Pribyl, R. Blokpoel, and M. Matowicki, "Addressing eu climate targets: Reducing co2 emissions using cooperative and automated vehicles," *Transportation Research Part D: Transport and Environment*, vol. 86, p. 102437, 2020.

[23] D. of Columbia Government, "2020 traffic volume dataset," https://opendata.dc.gov/datasets/DCGIS::2020-traffic-volume/about, 2020.

[24] C. Sommer, D. Eckhoff, R. German, and F. Dressler, "A survey of networking solutions in commercial road traffic telematics," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 786–797, 2013, accessed: Aug. 29, 2024. [Online]. Available: https://sumo.dlr.de/docs/Data/Scenarios.html

[25] Department for Transport, "Road traffic statistics," https://www.gov.uk/government/collections/road-traffic-statistics, accessed: Aug. 28, 2024.