applied sciences

MDPI

Article

Enhancing Traffic Efficiency Through Deep Reinforcement Learning-Based Traffic Signal Control with Cooperative Connected and Autonomous Vehicles

Le Dinh Nghiem [1], Sang Hoon Bae [1], Pham Minh Thao [2] and Kyoung Kuk Yoon [2,*]

[1] Spatial Information Systems Engineering, Department of Earth and Environmental System Sciences, Pukyong National University, Busan 48513, Republic of Korea; nghiemld@pukyong.ac.kr (L.D.N.); sbae@pknu.ac.kr (S.H.B.)

[2] Division of Maritime AI & Cyber Security, National Korea Maritime & Ocean University, Busan 49112, Republic of Korea; sqrtthao96@g.kmou.ac.kr

* Correspondence: kkyoon@kmou.ac.kr

Abstract

Optimizing traffic performance using artificial intelligence (AI) has consistently been a prominent direction in the development of intelligent transportation systems. While numerous studies have proposed methodologies for integrating cooperative connected and autonomous vehicles (CCAVs) with traffic signal systems via V2X communication, they often rely on simplified control strategies or lack effective coordination between signal timing and vehicle behavior. In this study, we propose a novel, integrated traffic signal control strategy combined with CAVs using deep reinforcement learning. Our key differentiation lies in the simultaneous optimization of signal phases using the Soft Actor–Critic (SAC) algorithm and the regulation of CCAVs via cooperative adaptive cruise control and Green Light Optimal Speed Advisory. This dual approach allows the signal controller to leverage rich state information from CAVs and the road infrastructure, enabling more anticipatory and cooperative decisions. The proposed approach is implemented and evaluated through various scenarios using the Simulation of Urban MObility (SUMO) platform. The results demonstrate the superior learning performance and robustness of the proposed model. Specifically, our proposed model achieves a significant reduction in average vehicle waiting time by up to over 80% compared to baseline models under high-demand scenarios (4800–6000 veh/h). These findings underscore the critical importance of joint optimization in future intelligent transportation systems, paving the way for more resilient urban traffic management.

Keywords: deep reinforcement learning; adaptive traffic signal control; cooperative connected and autonomous vehicles; green light optimal speed advisory; cooperative adaptive cruise control; soft actor–critic

Academic Editors: Artur Kierzkowski, Tomasz Nowakowski, Agnieszka A. Tubis, Franciszek Restel, Tomasz Kisiel, Anna Jodejko-Pietruczuk, Mateusz Zając, Viktoria Ivannikova, Michał Stosiak and Andrija Vidović

Received: 2 February 2026
Revised: 4 March 2026
Accepted: 5 March 2026
Published: 7 March 2026

1. Introduction

The rapid growth of urban populations, especially in metropolitan areas, has placed heavy demands on transportation infrastructures. Along with economic development and technological progress, the surge in private vehicle ownership and use has outstripped the expansion of transport facilities, creating a mismatch between supply and mobility needs. As a result, traffic congestion has emerged as an unavoidable issue, becoming increasingly intricate and difficult to resolve. Its negative impacts span across the economy, public health, and the environment [1,2]. Thus, alleviating congestion continues to be a central research priority in global traffic engineering.

To address this challenge, many large cities have adopted adaptive traffic signal control (ATSC) systems such as SCOOT [3] and SCAT [4] to regulate intersection flows. These rule-based systems rely on data collected from conventional sensing devices. Alternative approaches, including fuzzy logic [5–7] and genetic algorithms [8,9], have also been investigated. Nevertheless, during this stage, technological constraints limited the quality of data acquisition, hindering the design of truly optimal traffic management strategies.

Recent advances in both wired and wireless networking have made it possible to efficiently collect and transmit massive amounts of traffic data. Within this context, the Internet of Things (IoT) has been recognized as a key enabler for smarter congestion mitigation [10]. Vehicular Ad Hoc Networks (VANETs) [11], as a specific type of wireless ad hoc system, provide protocols for reliable vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications, strengthening both the adaptability and robustness of traffic control systems. Traditional control approaches often depend on detailed dynamical models to generate control signals. However, the scale of available datasets and the inherent uncertainty in traffic systems reduce the practicality of such model-based strategies. Consequently, attention has shifted toward intelligent and adaptive control methods capable of making effective decisions without explicit reliance on exact system models.

Reinforcement learning (RL), grounded in Markov decision processes (MDPs), offers a strong foundation for tackling such uncertain control problems [12]. RL has proven essential in advancing intelligent traffic management, but classical RL techniques store state–action values in tabular form, which becomes computationally infeasible for large-scale applications such as traffic signal optimization. A major breakthrough in this regard has been the incorporation of deep neural networks (DNNs). This integration, referred to as deep reinforcement learning (DRL) [13], combines RL principles with the representational power of DNNs, enabling the effective handling of large and complex state–action spaces. Accordingly, DRL [14,15] has been increasingly applied in traffic control research. Seminal contributions to this field include the works of Genders and Razavi [16], Gao et al. [17], and Liang et al. [18], which focus on training autonomous agents to manage isolated, simulated traffic intersections. These foundational approaches share a common methodological framework, employing Convolutional Neural Networks (CNNs) to represent the agent's policy and utilizing Q-learning as the core training algorithm. The primary distinctions among them emerge in the design of the reward functions, the architectural complexity of the simulated intersection. The same work is extended to a more general concept discussing several on-policy, off-policy RL algorithms on various state, action, and reward definitions in an experimental view [19]. Along with the three state representations and the variable-phase action model in [20], the authors also tested their models with the binary action model in a fixed green-phase cycle and four reward functions, including waiting time and queue length. However, DRL models may perform poorly during early training, potentially causing suboptimal or unsafe signal timings. Effective DRL training requires extensive simulation data and computational resources, which can be a barrier for real-time applications.

In parallel, the rise in connected and autonomous vehicles (CAVs) and vehicle-to-everything (V2X) technologies [21,22] introduces new opportunities for congestion reduction. Traffic performance is strongly influenced by driver behavior, and human-driven vehicles (HDVs) are a major source of unpredictability and inefficiency. Cooperative connected and autonomous vehicles (CCAVs) [22,23] mitigate these shortcomings by operating autonomously and coordinating with other vehicles and infrastructure. This not only improves the mobility of individual vehicles but also enhances collective traffic flow, reducing

https://doi.org/10.3390/app16052576

disorder and amplifying the benefits of adaptive signal control. By utilizing V2X communication, CACC enables continuous data exchange (e.g., position, velocity, acceleration, lane change intentions), allowing for synchronized platoon behavior with shorter and more stable headways. This enhances safety, increases road capacity, and reduces fuel consumption. Many studies have explored CACC through simulation and performance analysis. Wang et al. applied first-order consensus algorithms for weighted and constrained distance control [24], later extending this to second-order consensus with predecessor-following and predecessor–leader-following topologies. Third-order consensus approaches further incorporated longitudinal acceleration differences [25–27]. An MPC-based scheme was also proposed to optimize cooperative driving at signalized intersections [28]. PLEXE supports flexible platoon formation using models like Krauss [29], IDM [30], and ACC/CACC [31], with its mechanisms and applicability analyzed in existing traffic flow and vehicle control studies.

Despite significant progress in both deep reinforcement learning for traffic control and the development of connected and autonomous vehicles, the synergy between these two domains has yet to be fully realized. A careful examination of the current literature exposes three fundamental limitations that our work aims to overcome. First, existing DRL-based traffic signal control studies predominantly treat vehicles as passive elements whose behavior is exogenously determined rather than actively coordinated. While researchers have made significant progress in optimizing signal timing using DRL algorithms [32–36], the underlying traffic dynamics remain governed by human driving models (e.g., Krauss) that introduce stochasticity and instability. This creates a fundamental asymmetry: the signal controller adapts to traffic conditions, but vehicles do not adapt to the signal controller beyond basic stop–go responses. The result is a one-way adaptation rather than genuine bidirectional coordination. Second, current approaches to integrating CAVs with traffic signals exhibit a decoupled control architecture. Although several studies have proposed methodologies for connecting CAVs with traffic signal systems via V2X communication [37–40], they typically implement hierarchical or sequential control schemes where signal timing and vehicle behavior are optimized independently. For instance, GLOSA algorithms compute advisory speeds based on fixed signal schedules, while CACC controllers maintain platoon stability without explicit awareness of downstream signal timing. This separation ignores the synergistic potential of simultaneous optimization, where signal decisions could anticipate vehicle responses and vehicle trajectories could actively shape signal timing decisions. Third, the exploration–exploitation trade-off in DRL-based signal control remains inadequately addressed for highly dynamic traffic environments. Value-based methods (e.g., DQN) and deterministic policy gradient methods (e.g., DDPG) often converge to suboptimal policies or exhibit instability under varying demand patterns [41]. While these algorithms have demonstrated promise in controlled settings, their performance degrades when traffic conditions become highly stochastic, precisely the conditions where adaptive control is the most needed. The lack of explicit exploration mechanisms and susceptibility to overestimation bias limit their practical applicability.

Motivated by these limitations, this paper proposes a tightly coupled cyber–physical control framework that jointly optimizes traffic signal timing and cooperative vehicle longitudinal behavior at signalized intersections. The main contributions of this work are summarized as follows:

- We formulate the ATSC problem as a Markov decision process and employ the SAC algorithm to learn robust signal control policies under highly dynamic traffic conditions. By incorporating a maximum entropy objective, the proposed signal controller balances reward maximization with policy stochasticity, leading to improved exploration and training stability compared to value-based and deterministic policy gradient methods. More importantly, the entropy-regularized formulation allows the signal controller to exhibit anticipatory behavior, adapting phase selections proactively in response to evolving queue dynamics shaped by cooperative vehicle motion.

- This study further develops a cooperative connected and autonomous vehicle (CCAV) control scheme that combines V2V-based cooperative adaptive cruise control with a Green Light Optimal Speed Advisory algorithm designed in this work. While CACC is adopted to maintain stable platoon coordination, the proposed GLOSA strategy utilizes real-time signal information via V2I communication to enable anticipatory speed adjustment when approaching intersections. The integration of CACC and the developed GLOSA forms a unified CCAV framework that improves trajectory smoothness and reduces unnecessary stopping.
- We propose a unified cyber–physical control framework that jointly optimizes traffic signal timing and connected autonomous vehicle (CAV) behavior at a signalized intersection. Unlike existing studies that treat adaptive traffic signal control (ATSC) and vehicle control as two independent layers, the proposed framework establishes a closed-loop interaction between infrastructure-level decisions and vehicle-level longitudinal control. Specifically, the traffic signal controller, trained using Soft Actor–Critic (SAC), explicitly exploits real-time traffic states influenced by cooperative adaptive cruise control (CACC), while CAVs dynamically adjust their motion in response to signal decisions via V2I communication. This bidirectional coupling enables genuine cooperative control between signals and vehicles rather than simple coexistence.
- Extensive simulations are conducted in SUMO across a wide range of traffic demand levels and CAV penetration rates. The proposed SAC_CCAV framework is systematically compared with representative baseline methods, including DQN, DDPG, and A2C. The experimental results demonstrate that SAC_CCAV consistently outperforms all baselines in terms of cumulative reward, average waiting time, throughput, and average speed. Notably, the proposed method achieves an over 80% reduction in average waiting time under high-demand conditions and exhibits strong robustness as traffic demand and system complexity increase.

The remainder of this paper is organized as follows. Section 2 formulates the traffic signal control problem, describing the intersection geometry, signal phase configuration, and key assumptions underlying our approach. Section 3 presents the proposed methodology in two parts: first, the deep reinforcement learning framework for adaptive signal control using the Soft Actor–Critic algorithm, including the definition of the state space, action space, and reward function, and second, the cooperative connected and autonomous vehicle control layer integrating cooperative adaptive cruise control and Green Light Optimal Speed Advisory. Section 4 describes the experimental setup in SUMO and provides comprehensive results comparing the proposed SAC_CCAV framework against baseline methods across various traffic demand scenarios and CAV penetration rates. Finally, Section 5 concludes this paper and outlines directions for future research.

2. Statement of Problems

We study a four-way signalized intersection with three lanes on each incoming approach. The intersection consists of four entry edges (E1~E4), representing the east, west, north, and south directions, and four corresponding exit edges (−E1~−E4). Each entry edge contains three lanes: lane 0 is dedicated to left-turn movements, lane 1 is reserved for through movements, and lane 2 supports both through and right-turn movements. This configuration results in twelve distinct traffic movements within the intersection (see Figure 1). All lanes are assigned the same maximum speed limit.

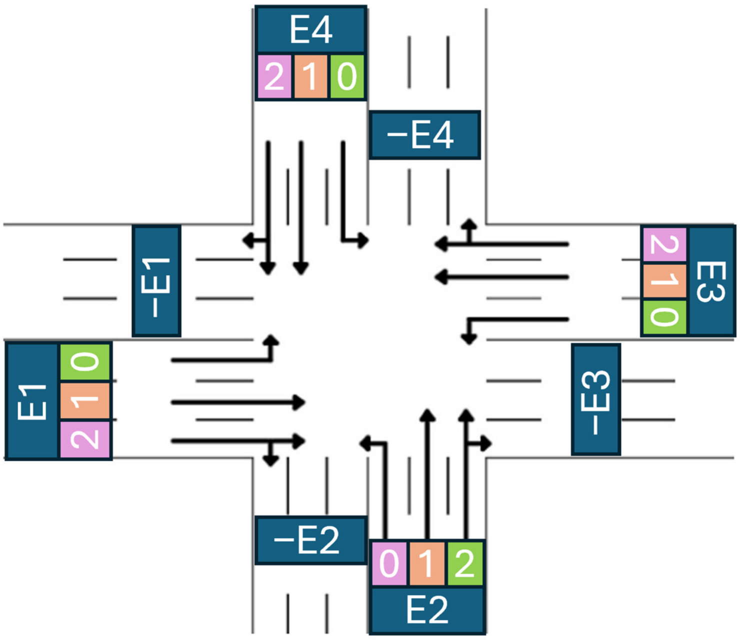

Figure 1. Studied intersection.

To prevent conflicts among traffic flows, the intersection is governed by four traffic signal phases (illustrated in Figure 2). At any given time, only one phase is active (green light), while the others remain inactive (red lights). Signal transitions follow a binary control mechanism. Before switching phases, the currently active phase must pass through a 3 s yellow interval to ensure safety. During this period, vehicles already inside the intersection may continue, while those waiting behind the stop line must remain stationary. In this study, to ensure pedestrian safety and optimize vehicular traffic efficiency, the intersections are assumed to be equipped with pedestrian overpasses or underpasses. This configuration eliminates conflicts between pedestrians and vehicles, allowing the proposed DRL-based control model to focus solely on vehicular signal optimization. Consequently, the implementation and evaluation of the control algorithms become more straightforward and effective.

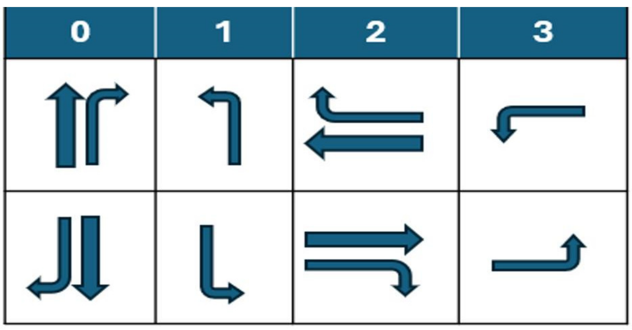

Figure 2. Phases of traffic lights.

We assume that each lane is equipped with reliable traffic detection devices, providing stable and noise-free measurements for signal control. Additionally, vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication technologies are fully available to all connected and autonomous vehicles (CAVs). This ensures the seamless exchange of information without interference or delay, enabling cooperative control for improved traffic safety. For human-driven vehicles (HDVs), strict adherence to traffic regulations is assumed at all times. Furthermore, abnormal situations such as vehicle breakdowns in the middle of the road or other disruptions to traffic flow are excluded from consideration.

3. Methodology

In this section, we present the methodological framework of the proposed cooperative traffic control system. The overall architecture consists of two tightly interacting layers: an adaptive traffic signal controller based on deep reinforcement learning and a cooperative connected and autonomous vehicle (CCAV) control layer. The signal control problem is formulated as a Markov decision process and optimized using the Soft Actor–Critic algorithm, while the vehicle layer integrates cooperative adaptive cruise control (CACC) with a signal-aware Green Light Optimal Speed Advisory (GLOSA) strategy. Through V2V and V2I communication, the two layers influence each other, enabling coordinated infrastructure–vehicle interaction at the intersection. The following subsections describe the signal control formulation and the CCAV framework in detail.

3.1. Adaptive Traffic Signal Control Using Deep Reinforcement Learning

Markov decision processes are optimization models for decision-making under uncertainty, where actions influence both immediate and future rewards by affecting subsequent states. MDPs provide a formal framework for modeling sequential decision-making problems in stochastic environments, allowing agents to learn from interactions to achieve desired goals. We formulate the ATSC problem as a Markov decision process (MDP):

$$M = (S,\ A,\ P,\ R,\ \gamma) \tag{1}$$

where S represents traffic states (queue lengths, waiting times, signal phase), A is the set of available signal phases, P denotes transition probabilities from state s to s′, R is the reward function penalizing congestion and γ is the discount factor.

In this study, we employ the Soft Actor–Critic (SAC) algorithm to learn an optimal strategy for traffic signal control. SAC is an off-policy Actor–Critic algorithm within the framework of Maximum Entropy Reinforcement Learning.

In essence, the objective of SAC is not only to maximize the cumulative return but also to maximize the entropy of the policy. This means that the agent is encouraged to achieve high rewards while maintaining as much randomness as possible during the learning process. SAC adopts an Actor–Critic architecture consisting of two critic networks (Q-functions) and one actor network (policy function). To stabilize training, it also leverages target networks similar to DDPG and TD3.

The objective function of SAC is extended by adding an entropy term to the standard objective, expressed as follows:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[\gamma^t\left(r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot \mid s_t))\right)\right] \tag{2}$$

where $\mathcal{H}(\pi(\cdot \mid s_t))$ denotes the entropy of the policy at state s_t, and $\alpha \in (0,\ 1)$ is the temperature parameter controlling the trade-off between reward maximization and entropy maximization.

A notable feature of SAC is the use of two independent Q-functions, with the minimum of their predicted values taken for learning. This design reduces the overestimation bias commonly observed in Q-function approximations.

The critic (value network) is trained to estimate both the Q-function Q(s,a) and the value function V(s) as Equations (3) and (4):

$$Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p}[V(s_{t+1})] \tag{3}$$

$$V(s_t) = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)] \tag{4}$$

The loss function for updating the Q-networks is given by the following:

$$L_Q(\phi_i) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}}\left[\left(Q_{\phi_i}(s,a) - y(r,s')\right)^2\right] \tag{5}$$

$$y(r,s') = r + \gamma(1-d)\left(\min_{j=1,2} Q_{\phi_{\text{targ},j}}(s', \tilde{a}') - \alpha \log \pi_\theta(\tilde{a}' \mid s')\right), \ \ \tilde{a}' \sim \pi_\theta(\cdot \mid s') \tag{6}$$

where d denotes the terminal indicator of an episode (d = 0 if not terminated; d = 1 otherwise).

The target value network is updated slowly from the main value network to ensure training stability. The parameters $\overline{\phi}$ in the target value network are updated according to Equation (7) using the soft update coefficient τ.

$$\overline{\phi} \leftarrow \tau\phi + (1-\tau)\overline{\phi} \tag{7}$$

The actor (policy network) learns a stochastic policy π_θ (a|s). Unlike DDPG, which outputs a deterministic action, the SAC policy outputs a probability distribution (typically Gaussian) over actions. Actions are then sampled from this distribution during execution. The policy loss for updating the actor is defined as Equation (8):

$$L_\pi(\theta) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\mathbb{E}_{a_t \sim \pi_\theta}\left[\alpha \log \pi_\theta(a_t \mid s_t) - \min_{j=1,2} Q_{\phi_j}(s_t, a_t)\right]\right] \tag{8}$$

To implement SAC for traffic signal control, we define the state space, action space, and reward function as follows.

The state space represents input information collected from the environment that influences the controller's decision-making. Various factors determine which traffic movements are allowed or restricted. We define the state space as Equation (9):

$$s_t = \{f_i^t, h_i^t, w_i^t, p_i^t\}_{i=1}^{N_l}, \tag{9}$$

where N_l is the total number of incoming lanes. Specifically, f_i^t denotes the traffic flow of lane i, h_i^t represents the number of queued vehicles, w_i^t is the cumulative waiting time of vehicles in the lane, and p_i^t indicates the current signal phase status.

Each component of the state vector serves a distinct control purpose. Traffic flow f_i^t reflects incoming demand pressure, queue length h_i^t captures instantaneous congestion, and cumulative waiting time w_i^t represents long-term inefficiency and delay accumulation. The inclusion of signal phase information p_i^t provides the agent with phase memory, which helps prevent oscillatory switching behavior and promotes smoother phase transitions. Collectively, this state representation enables the agent to perceive both short-term dynamics and longer-term congestion trends, which is essential for anticipatory signal control.

The action space Equation (10) is defined as a discrete set corresponding to the predefined signal phases illustrated in Figure 2. At each control interval Δt, the agent selects one phase to activate. Phase durations are not explicitly optimized; instead, the controller determines phase selection at fixed time steps, while safety constraints are enforced through a mandatory yellow interval before any phase change.

$$a = [0, 1, 2, 3] \tag{10}$$

The reward is designed to achieve two primary objectives: minimizing the number of vehicles waiting at intersections and reducing their total waiting times. The reward function is defined by Equation (11):

$$r = \mu w_{avg} + \nu h_{avg} \tag{11}$$

where w_{avg} and h_{avg} denote the average waiting time and the average queue length across lanes after each control interval Δt, respectively. This reward formulation encourages the controller to balance immediate congestion relief with long-term delay reduction. Unlike pressure-based rewards that may favor aggressive phase switching, the use of averaged metrics promotes scale invariance across different traffic demand levels and yields smoother control behavior. Moreover, by incorporating both queue length and waiting time, the reward captures complementary aspects of traffic inefficiency, enabling the agent to learn signal policies that reduce stop-and-go behavior and improve overall traffic flow stability.

3.2. Cooperative Connected and Autonomous Vehicles (CCAVs)

Connected and autonomous vehicles (CAVs) provide new opportunities to improve traffic efficiency and stability by enabling cooperative vehicle control through communication and automation. In this study, we adopt a cooperative connected and autonomous vehicle (CCAV) framework that integrates both vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication to coordinate vehicle motion at signalized intersections, as shown in Figure 3. The proposed CCAV control layer consists of two complementary components: (i) cooperative longitudinal control among vehicles based on cooperative adaptive cruise control (CACC) and (ii) signal-aware speed planning based on Green Light Optimal Speed Advisory (GLOSA). Together, these components enable coordinated vehicle behavior that is both string-stable and traffic-signal-aware.

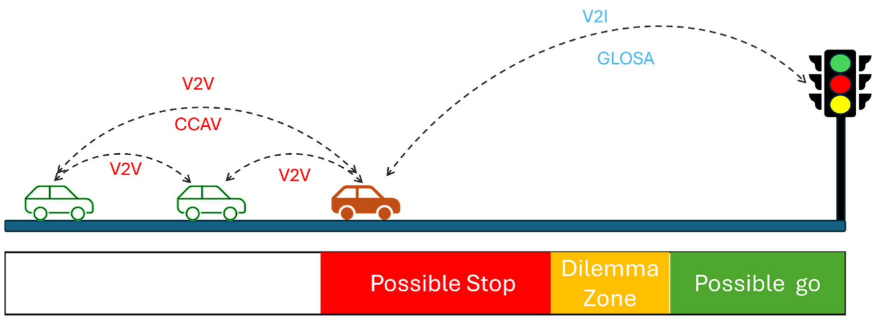

Figure 3. Proposed CCAV structure.

3.2.1. CACC-Based Cooperative Vehicle Control

Within the CCAV framework, CAVs are equipped with cooperative adaptive cruise control (CACC) to regulate their longitudinal dynamics through V2V communication. By sharing real-time information such as position, velocity, and acceleration, CACC enables vehicles to operate in a coordinated platoon-based manner, reducing the randomness and instability typically observed in human-driven traffic.

The CACC controller adopts a hybrid control structure composed of speed control and gap control modes. When the inter-vehicle spacing exceeds a predefined threshold, vehicles operate in speed control mode, where the primary objective is to minimize relative velocity differences with respect to the preceding vehicle. The acceleration command for vehicle i at time step k is given by

$$a_{i,k+1} = k_1(v_d - v_{i,k}), \quad k_1 > 0 \tag{12}$$

When the speed deviation and the distance between two vehicles are smaller than 0.1 m/s and 0.2 m, respectively, the gap control mode is activated. The velocity at the next time step is calculated as Equation (13):

$$v_{i,k+1} = v_{i,k} + k_2 e_{i,k} + k_3 \dot{e}_{i,k}, \quad k_2, k_3 > 0 \tag{13}$$

$$\dot{e}_{i,k} = v_{i-1,k} - v_{i,k} - t_d a_{i,k} \tag{14}$$

where $e_{i,k}$ represents the distance deviation, and t_d denotes the desired time gap of the CACC controller.

Through this cooperative control mechanism, CACC suppresses stop-and-go oscillations and ensures stability within vehicle platoons. However, while CACC effectively stabilizes vehicle interactions, it does not explicitly account for downstream traffic signal timing. To address this limitation, a signal-aware speed planning module is incorporated, as described in the following subsection.

3.2.2. Green Light Optimal Speed Advisory (GLOSA)

While CACC ensures stable and cooperative longitudinal control among vehicles, it does not explicitly exploit traffic signal timing information. To enable proactive vehicle–signal coordination, we incorporate a Green Light Optimal Speed Advisory (GLOSA) strategy within the CCAV framework. The objective of GLOSA is to compute an advisory speed that allows a vehicle to traverse a signalized intersection during the green phase while avoiding unnecessary stopping, excessive acceleration, and abrupt deceleration, particularly under actuated signal control where future phase durations are uncertain.

Consider an ego vehicle approaching a single actuated signalized intersection. At a given decision time t_0, the vehicle state is characterized by its distance to the stop line d and its current speed v_0. The decision variable of the GLOSA problem is chosen as the arrival time at the stop line t_a.

Assuming quasi-constant speed over the short approach distance, the advisory speed corresponding to a candidate arrival time is given by

$$v(t_a) = \frac{d}{t_a} \tag{15}$$

The average acceleration required to transition from the current speed v_0 to the advisory speed is approximated as

$$a(t_a) = \frac{v(t_a) - v_0}{t_a} \tag{16}$$

To ensure passenger comfort and vehicle safety, the advisory speed must satisfy kinematic constraints:

$$v_{\min} \le v(t_a) \le v_{\max}, a_{\min} a(t_a) \le a_{\max}$$

where v_{min} and v_{max} denote speed limits, and a_{min} and a_{max} represent comfortable deceleration and acceleration bounds, respectively.

Unlike fixed-time signals, actuated traffic signals do not operate with a predetermined cycle length. Instead, the green phase may be extended in discrete steps based on detector actuation. Let τ_{sw} denote the remaining time until the earliest possible phase change from green to yellow, and let g represent the green extension step.

Since the total green duration is uncertain, the end of the green phase is bounded as

$$t_G \in [t_G^{\min}, t_G^{\max}]$$

where

$$t_G^{\min} = \tau_{sw}, t_G^{\max} = \tau_{sw} + \Delta_{\max}(g)$$

To account for uncertainty in green extensions, a risk-aware green horizon is defined as

$$t_G^{adv} = t_G^{\min} + \eta(t_G^{\max} - t_G^{\min}) \quad (17)$$

where $\eta \in [0,1]$ is a conservativeness parameter. Setting $\eta = 0$ yields a fully robust but conservative strategy, while larger values allow for a more aggressive exploitation of potential green extensions.

When approaching the end of the green phase, vehicles may encounter a dilemma zone, where neither stopping comfortably nor passing the intersection safely is feasible. Let t_Y and t_R denote the start times of the yellow and red phases, respectively, with a fixed yellow duration T_Y.

The feasibility of stopping is determined by the minimum deceleration required to stop within distance d:

$$a_{stop} = -\frac{v_0^2}{2d} \quad (18)$$

Stopping is feasible if $a_{stop} \geq a_{min}$.

Similarly, the feasibility of passing the intersection before the onset of red is assessed by the required average acceleration:

$$a_{go} = \frac{2(d - v_0 t_R)}{t_R^2} \quad (19)$$

Passing is feasible if $a_{go} \leq a_{max}$.

The vehicle is considered to be in a dilemma zone when stopping is infeasible and passing is infeasible simultaneously. To prevent unsafe or uncomfortable behavior, arrival times within the interval $[t_R]$ are explicitly excluded from the feasible solution set.

Based on the above considerations, the GLOSA problem is formulated as a one-dimensional constrained optimization over the arrival time t_a:

$$\min_{t_a \in (0, t_G^{adv})} J(t_a) = w_t t_a + w_a a(t_a)^2 + w_v(v(t_a) - v_0)^2 \quad (20)$$

subject to vehicle kinematic constraints and dilemma zone exclusion.

The cost function balances travel efficiency, acceleration smoothness, and speed deviation from the current velocity. The optimal arrival time t_a^* is obtained by evaluating feasible candidates, and the optimal advisory speed is then computed as

$$v^* = \frac{d}{t_a^*} \quad (21)$$

Within the proposed CCAV framework, the advisory speed generated by GLOSA serves as a high-level reference for the vehicle control layer. For platoon leaders or vehicles operating in free-flow conditions, the GLOSA speed directly guides longitudinal motion. For follower vehicles, the CACC controller ensures that inter-vehicle spacing and safety constraints are preserved while tracking the GLOSA-adjusted speed profile.

Moreover, the GLOSA strategy is tightly coupled with the SAC-based traffic signal controller through V2I communication. Signal phase decisions determine the green horizon used by GLOSA, while smoother vehicle arrivals and reduced stopping behavior reshape queue dynamics and waiting time distributions observed by the signal controller. This bidirectional interaction enables anticipatory vehicle–signal coordination and contributes significantly to reducing delay and improving traffic stability at signalized intersections.

4. Experiment and Results

This section presents the experimental setup and performance evaluation of the proposed SAC_CCAV framework. First, the simulation environment, implementation details, and training configurations are described. Next, comparative experiments are conducted against representative DRL-based methods (DQN, DDPG, A2C) and the model-based Max-Pressure (MP) controller under varying traffic demand levels. Finally, the impact of CCAV penetration rates is analyzed to assess the robustness and scalability of the proposed approach.

4.1. Setup Environment

Selecting an appropriate simulation tool is critical to implementing the model and conducting reliable evaluations. The research community has access to a variety of traffic simulation tools, such as GLD [42], VISSIM [43], etc. These tools are widely utilized to design and evaluate traffic management strategies within simulation environments.

In this study, we selected SUMO [44] to perform simulations and evaluate the proposed model across a range of scenarios. SUMO, an open-source program, is a valuable tool for researching, implementing, and assessing traffic management strategies, such as traffic light control systems or direct vehicle operation on road networks. SUMO supports simulation at various detail levels, including macroscopic and microscopic scales. It can interact dynamically with other software tools and supports the use of different programming languages for algorithm implementation via an efficient API known as Traci. Furthermore, SUMO allows for the simulation of scenarios ranging from single roads and simple intersections to large-scale and complex networks.

To verify the functionality and performance of the proposed model, we conducted simulations and evaluations across a wide array of scenarios.

To simplify and focus on evaluating the efficiency of the proposed model and the impact of integrating various vehicle control methods, we employ our proposed algorithm to manage traffic lights in conjunction with two specific vehicle types: human-driven vehicles (HDVs) and cooperative connected and autonomous vehicles (CCAVs). For HDVs, we utilize the Krauss model to objectively describe the randomness and instability of traffic flow, aligning with real-world traffic conditions and reflecting the uncertainty in human-controlled vehicle behavior. Table 1 summarizes the fundamental vehicle parameters adopted in the simulation environment, including vehicle length, maximum acceleration and deceleration, desired speed, and reaction-related parameters. These settings ensure consistency across vehicle types while preserving the distinct behavioral characteristics of HDVs and CCAVs.

Table 1. Parameters of vehicle.

Parameter	Unit	HDVs	CCAVs
Maximum velocity (v_max)	m/s	20	20
Max acceleration (a_max)	m/s^2	3	3
Min acceleration (a_min)	m/s^2	−6	−6
Length of vehicle	m	5	5
Car-following model		Krauss	CACC

To implement the SAC algorithm for the ATSC system, we set up the following parameters. We simulate the traffic network for one hour, which corresponds to one episode, with a time step of Δt = 10 s. Thus, each episode consists of 360 time steps. We train the model for 200 episodes.

According to Equation (9), the state vector consists of lane-level traffic flow, queue length, and cumulative waiting time for each incoming lane, together with the current signal phase. For the considered four-approach intersection with three lanes per approach, this results in a 40-dimensional state representation. This design allows the agent to capture both instantaneous traffic conditions and accumulated congestion information while maintaining a compact input structure for efficient learning.

We use a lane area detector (LAD) to measure the number of vehicles waiting in each lane. The waiting time for vehicles at each intersection is recorded per second, provided the vehicle speed within the detector's measurement area registers as 0.1 m/s. A inductive loop detector (ILD) is placed at the entry points of each road segment to collect traffic flow data at both ends. The LAD and ILD are two measurement tools available in SUMO. Within the 50 m detection range, the maximum observable queue is approximately 10 vehicles, whereas individual waiting time may exceed 100 s. Given the action interval of Δt = 10 s, these two metrics differ significantly in magnitude. Following established practices in RL-based traffic signal control [45–47], the weighting coefficients in Equation (11) are set to $\mu = 1/10$ and $\nu = 1$ to balance their influence and prevent numerical bias during training.

$$r = \frac{w_{avg}}{10} + h_{avg} \tag{22}$$

In the context of traffic signal control, the temperature parameter α influences how conservatively or adaptively the controller selects signal phases under uncertain traffic conditions. A larger α encourages more exploration phase selection, which can help the controller adapt to fluctuating arrival patterns, especially under varying CCAV penetration rates. Conversely, a smaller α leads to more deterministic behavior, prioritizing immediate congestion reduction based on current traffic states. In this study, the temperature parameter α is treated as a fixed hyperparameter (α = 0.2). Since the action space consists of only four signal phases and the reward scale remains bounded across demand scenarios, a fixed α provides stable exploration during training. Table 2 lists the architectural settings and training parameters adopted for the SAC agent in the traffic signal control framework.

The proposed Soft Actor–Critic (SAC) agent adopts a fully connected multilayer perceptron (MLP) architecture for both the actor and critic networks, which is suitable for the structured vector-based traffic state representation used in this study, as shown in Figure 4. The input state consists of lane-level traffic flow, queue length, cumulative waiting time, and the current signal phase. All state variables are normalized prior to being fed into the networks to improve training stability. The current signal phase is encoded using a one-hot representation to avoid introducing artificial ordinal relationships between phases.

The actor network parameterizes a categorical policy over the discrete action space, where each action corresponds to one of the four signal phases. The network contains two hidden layers with 256 neurons each, followed by Rectified Linear Unit (ReLU) activation functions. The output layer produces four logits, which are transformed into a probability distribution using the softmax function.

To mitigate overestimation bias, two independent critic networks are employed following the double Q-learning principle. Each critic shares the same MLP structure as the actor (two hidden layers with 256 neurons and ReLU activation) but outputs four Q-values corresponding to all possible actions. The Q-value associated with the selected action is used for loss computation. Target networks are maintained for both critics and updated via soft updates with a smoothing coefficient τ. The minimum of the two target Q-values is used in the Bellman backup to enhance learning stability.

Table 2. The hyperparameters of the proposed SAC-based signal controller.

Hyperparameter			Value
Layer/activation function	Actor network	Input (state)	40
		Hidden 1	256, ReLu
		Hidden 2	256, ReLu
		Output	4 logits, followed by SoftMax activation
	Critic network	Input (state)	40
		Hidden 1	256, ReLu
		Hidden 2	256, ReLu
		Output	4 Q-values corresponding to all discrete actions
Batch size			128
Replay buffer size			1,000,000
Discount factor γ			0.99
Soft update τ			0.005
Temperature α			0.2
Optimizer			Adam
Learning rate (actor/critic) η			3×10^{-4}
Initial random steps			5000
Steps per learning update			4
Fixed network update frequency			8000
Target entropy			$0.98 \times \log(\mid A \mid)$

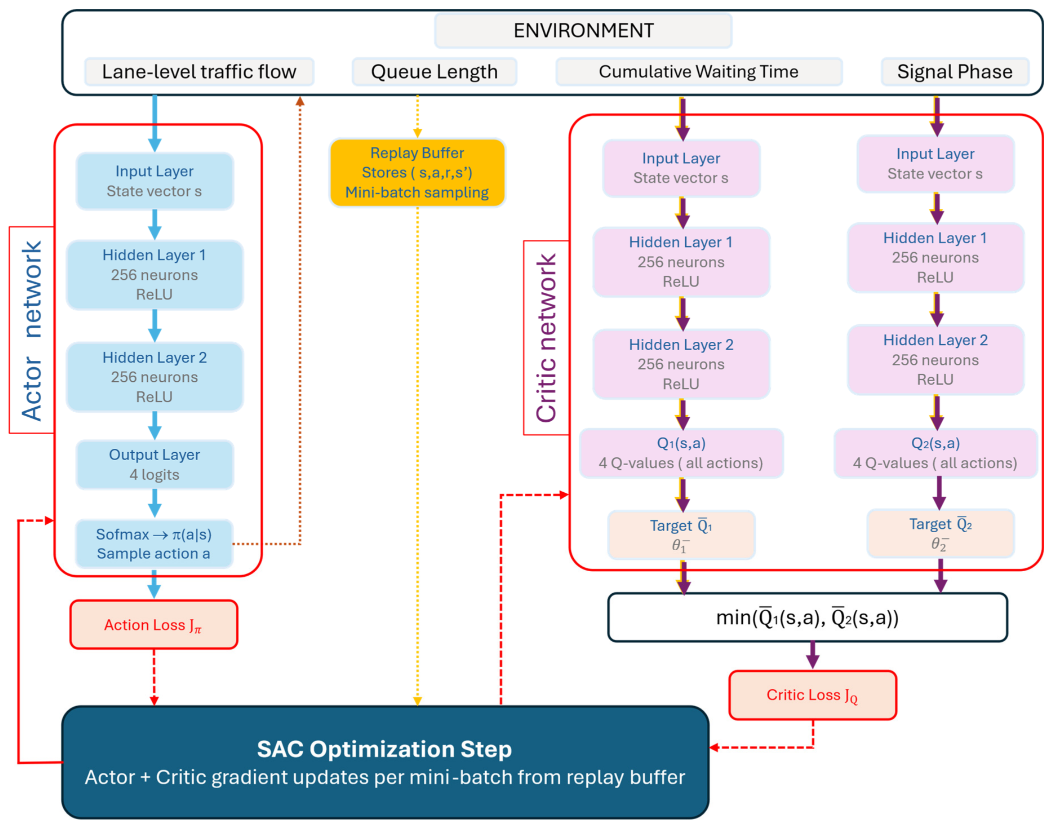

Figure 4. Proposed SAC architecture.

4.2. Results

All learning-based controllers evaluated in this study were implemented using deep neural networks as function approximators to ensure a fair and consistent comparison framework. To benchmark the proposed SAC_CCAV method, we selected three widely adopted deep reinforcement learning algorithms, Deep Q-Network (DQN), Deep Deterministic Policy Gradient (DDPG), and Advantage Actor–Critic (A2C), as well as the classical Max-Pressure (MP) control strategy. These methods represent distinct categories of traffic signal control approaches, including value-based learning, deterministic policy gradient methods, Actor–Critic architectures, and model-based analytical control, thereby providing comprehensive performance references under varying traffic conditions.

DQN is a value-based reinforcement learning algorithm that approximates the state–action value function using a deep neural network. It selects actions by estimating the expected cumulative reward associated with each discrete action and choosing the one with the highest Q-value. While DQN has been extensively applied to traffic signal control problems with discrete phase selection, its performance may deteriorate in highly dynamic environments due to instability issues and limited sample efficiency [48].

DDPG extends deterministic policy gradient methods to continuous control problems by combining an actor network that outputs deterministic actions and a critic network that evaluates them. Although DDPG is primarily designed for continuous action spaces, it has been applied to traffic signal optimization scenarios where signal timing parameters are treated as continuous variables [49].

A2C belongs to the Actor–Critic family and jointly optimizes a policy network (actor) and a value network (critic). By integrating policy gradient updates with value function estimation, A2C improves training stability compared to purely value-based approaches. It has been widely used in traffic signal control research due to its balance between computational efficiency and learning performance, particularly in environments with moderate state–action complexity [50].

In addition to DRL-based baselines, the Max-Pressure (MP) controller is included as a representative model-based method grounded in traffic flow theory. MP determines signal phases by maximizing the pressure difference between incoming and outgoing queues, thereby aiming to stabilize network throughput. It has strong theoretical guarantees under certain conditions and is recognized for its robustness in congested networks [51].

To ensure a fair comparison, all learning-based controllers (SAC, DQN, DDPG, and A2C) operate under identical sensing and state information. Specifically, each algorithm receives the same 40-dimensional state vector consisting of lane-level traffic flow, queue length, cumulative waiting time, and current signal phase, collected from the same SUMO detectors. No controller is provided with privileged access to individual CAV trajectories, future signal timing, or additional vehicle-level information.

4.2.1. Training Reward Comparison

Figure 5 illustrates the cumulative training reward per episode for all compared algorithms. The proposed SAC_CCAV model demonstrates rapid convergence and stable learning behavior, reaching a high reward level within approximately the first 50 training episodes. After convergence, the cumulative reward stabilizes around approximately −100, with only minor fluctuations, indicating consistent policy performance throughout the remainder of the training process.

In comparison, the baseline algorithms exhibit significantly inferior convergence characteristics. DDPG initially achieves moderate rewards during early training but quickly reaches a plateau, stabilizing at a suboptimal level of approximately −900, with persistent oscillations and no further improvement. DQN performs considerably worse, with cumulative rewards remaining consistently low throughout training, fluctuating around -4000, indicating difficulty in learning effective signal control policies under stochastic traffic dynamics. A2C shows the most unstable behavior: although it initially attains reward values around -1000, the training process collapses after approximately 75 episodes, with rewards dropping sharply below -5000, reflecting divergence and failure to converge to a viable control strategy.

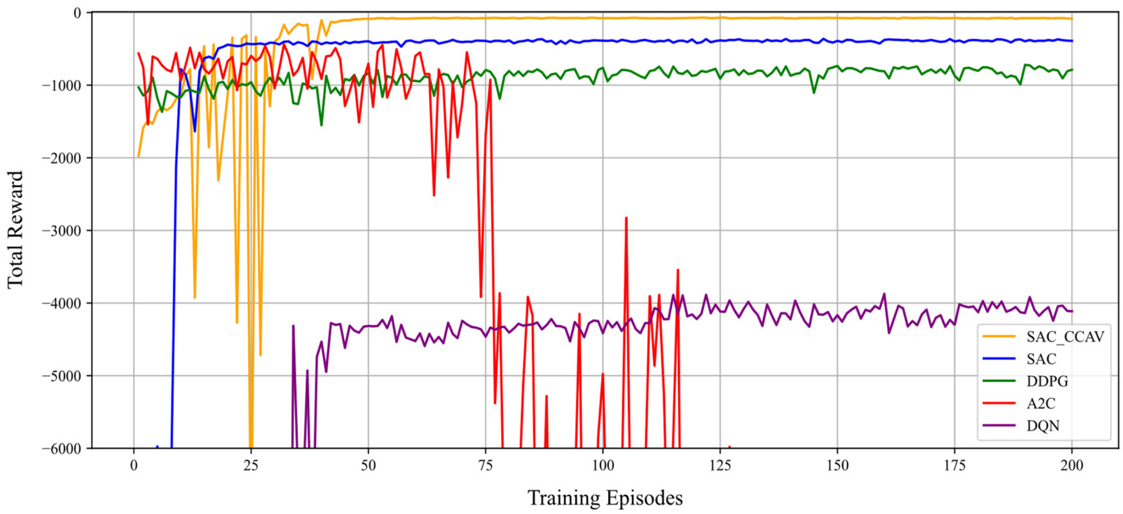

Figure 5. Training reward comparison.

The pronounced performance gap among the algorithms highlights both algorithmic and system-level effects. From an algorithmic perspective, the entropy-regularized SAC framework provides improved exploration and training stability compared to value-based (DQN) and deterministic policy gradient (DDPG) methods. From a system perspective, the integration of the CCAV layer plays a critical role. Through CACC, speed fluctuations within vehicle platoons are reduced, and through GLOSA, signal-induced stopping events are mitigated. These cooperative mechanisms substantially decrease traffic state variability, leading to smoother reward landscapes and more predictable state transitions. As a result, SAC_CCAV converges to a substantially higher-performing and more stable policy than the baseline methods.

4.2.2. Evaluation

After obtaining the trained models, we evaluated their performance under different traffic demand scenarios using key indicators such as vehicle waiting time, the number of vehicles passing through the intersection, and the average vehicle speed. To ensure statistical reliability and mitigate the influence of randomness in network initialization, traffic generation, and mini-batch sampling, each algorithm was executed over 10 independent runs with different random seeds. For each performance metric, the mean value and standard deviation were computed across the 10 runs.

Table 3 presents the cumulative rewards obtained under different traffic demand scenarios. It can be observed that our proposed model consistently outperforms all baseline methods across all traffic demand levels. At low demand (1200 veh/h), SAC_CCAV achieves a cumulative reward of –96.82, whereas at high demand (6000 veh/h), the model still maintains a reward of –8122.72. In contrast, both DDPG and DQN demonstrate poor traffic performance once demand exceeds 2400 veh/h. This indicates that SAC_CCAV exhibits strong robustness, achieving effective control across a wide range of traffic conditions, from low to high demand.

Table 3. Reward comparison in terms of different traffic demands.

Traffic Volume (Veh/h)	SAC_CCAV	DDPG	DQN
1200	−96.82	−893.73	−156.83
2400	−201.5	−228.35	−378
3600	−3857.17	−75,098.08	−35,018.5
4800	−7956.42	−84,496.06	−55,560.63
6000	−8122.72	−71,702.32	−58,859.68

As expected, the number of vehicles able to complete their trips through the network is substantially higher when applying SAC_CCAV. For traffic demand levels below 2400 veh/h, nearly all vehicles successfully finish their journeys. When demand increases and traffic density becomes more congested, regulation becomes significantly more challenging. Nevertheless, in such high-demand scenarios, SAC_CCAV clearly demonstrates its advantages, enabling a considerably higher number of vehicles to complete their trips compared to DDPG and DQN, as reported in Table 4.

Table 4. Number of vehicles passing.

Traffic Volume (Veh/h)	SAC_CCAV	DDPG	DQN	MP
1200	1200± 0	1190 ± 8	1127 ± 25	1185 ± 12
2400	2332 ± 15	2100 ± 45	2216 ± 52	2250 ± 30
3600	2754 ± 28	1968 ± 110	1942 ± 135	2350 ± 85
4800	3936 ± 45	1827 ± 156	1463 ± 207	2400 ± 154
6000	4165 ± 63	1837 ± 213	1087 ± 269	2936 ± 180

Furthermore, Table 5 illustrates the total waiting time of vehicles. The application of SAC_CCAV significantly reduces cumulative waiting times compared to other control models, highlighting its effectiveness in improving overall traffic flow efficiency.

Table 5. Waiting time comparison.

Traffic Volume (Veh/h)	SAC_CCAV	DDPG	DQN	MP
1200	770.033 ± 45.2	1073.39 ± 180.5	1477.112 ± 95.1	1120.3 ± 80.6
2400	874.308 ± 62.8	5972.27 ± 520.3	1586.545 ± 140.3	1450.2 ± 102.4
3600	4038.87 ± 210.5	55,045.50 ± 3100	24,683.89 ± 1800	12,500.36 ± 980
4800	4898.13 ± 305.4	60,922.93 ± 4200	39,082.4 ± 2500	19,800.6 ± 1300.5
6000	4999.72 ± 480.6	50,601.9 ± 3800	40,604.54 ± 2900	22,100.4 ± 1800

Figure 6 illustrates the average vehicle speed within the intersection under a traffic demand of 5000 veh/h. At the beginning of the simulation, vehicles maintain relatively high speeds because traffic density is low, allowing the signal controller to regulate flows efficiently without causing congestion. As traffic demand increases and queues gradually accumulate, the average vehicle speed decreases for all controllers due to growing interaction among vehicles and limited discharge capacity at the intersection.

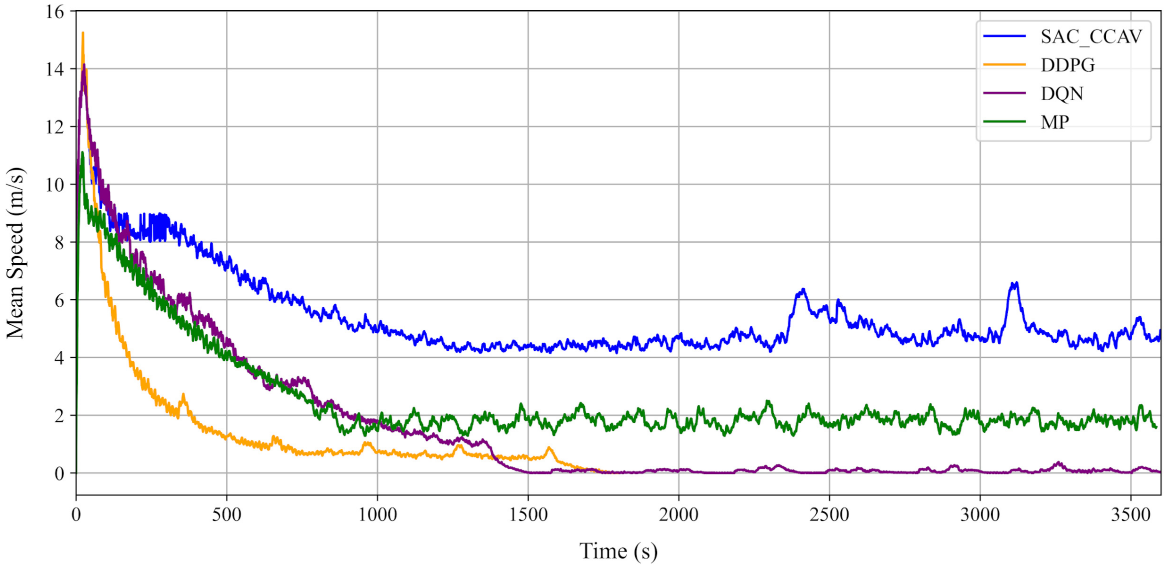

Figure 6. Average speed comparison.

Among the compared methods, SAC_CCAV maintains the highest and most stable average speed after approximately 500 s, fluctuating around 4–6 m/s even under sustained high demand. This indicates that the cooperative interaction between adaptive signal control and CCAV trajectory smoothing effectively mitigates severe congestion.

The Max-Pressure (MP) controller shows moderate performance. Although the average speed under MP declines as congestion builds up, it stabilizes around 1.5–2 m/s rather than collapsing to zero. This behavior suggests that MP is capable of preventing complete gridlock by balancing queue pressures across approaches; however, it lacks the adaptability to dynamically respond to rapidly changing traffic states, leading to persistent low-speed operation under heavy demand.

In contrast, both DDPG and DQN exhibit significant performance degradation. Their average speeds decrease sharply and approach nearly 0 m/s after prolonged operation, indicating that these learning-based controllers fail to maintain stable signal policies under high-demand conditions. This severe reduction in speed corresponds to longer queues, increased waiting times, and a lower number of completed trips observed in the previous performance metrics.

The results demonstrate that while MP provides a certain level of robustness compared to DDPG and DQN, it remains limited under extreme traffic demand. The proposed SAC_CCAV framework achieves superior performance by combining entropy-regularized signal learning with cooperative vehicle control, thereby sustaining higher traffic mobility and preventing speed collapse.

The string stability of the CCAVs was evaluated by quantifying the amplification of speed and acceleration disturbances between consecutive vehicles (see Appendix A). The maximum speed disturbance amplification gain across adjacent vehicle pairs was $G_v^{max} = 0.587 \pm 0.093$, and the corresponding acceleration gain was $G_a^{max} = 0.724 \pm 0.187$. As both values remain well below the theoretical stability threshold of unity, traffic disturbances introduced by the leading vehicle are attenuated rather than amplified as they propagate downstream. This confirms that the implemented CCAV strategy satisfies the fundamental condition of string stability.

From a dynamic perspective, the attenuation of speed fluctuations indicates that follower vehicles respond smoothly to upstream deceleration and acceleration events without generating oscillatory amplification. The reduced acceleration gain further implies moderate control effort and the suppression of stop-and-go wave formation. Consequently, inter-vehicle spacing remains bound and well-regulated, enabling coherent and coordinated platoon motion. Compared to typical human-driven traffic, where disturbances often amplify along the vehicle string, the automated coordination effectively filters high-frequency fluctuations and dampens traffic oscillations.

Importantly, this stabilized longitudinal behavior significantly enhances the effectiveness of the upper-layer signal control. By smoothing vehicle arrivals and reducing shockwave propagation toward the intersection, the CCAV framework decreases the traffic state variability perceived by the SAC-based signal controller. The resulting improvement in flow predictability enables more consistent phase selection, reduces unnecessary stopping events, and enhances discharge efficiency during green phases. Therefore, the demonstrated string-stable vehicle behavior not only improves platoon-level smoothness and energy efficiency but also plays a critical enabling role in the overall performance gains achieved by the integrated SAC_CCAV framework.

4.2.3. Impact of CCAV Penetration Rate

Table 6 summarizes the impact of different CCAV penetration rates on traffic performance at the signalized intersection. As the proportion of CCAVs increases from 25% to 100%, consistent improvements are observed across all evaluated performance metrics. Specifically, the average vehicle speed increases from 3.76 m/s to 5.43 m/s, while the number of vehicles successfully passing through the intersection rises from 1223 to 4310. At the same time, the average travel time decreases markedly from 263.9 s to 182.1 s. These results indicate that higher levels of vehicle connectivity and automation are associated with substantial gains in intersection efficiency and throughput under the considered traffic conditions.

Table 6. The simulation results with different CCAV rates.

CCAV Rate (%)	Average Speed (m/s)	Number of Vehicles Passed (vehs)	Average Travel Time (s)
25	3.76	1223	263.922
50	4.20	2339	235.378
75	4.30	2942	229.904
100	5.43	4310	182.067

The observed trends can be explained by the increasing effectiveness of cooperative vehicle behavior as the CAV penetration rate grows. With a larger proportion of CAVs, V2V-based (CACC) communication more effectively stabilizes platoon dynamics, mitigates stop-and-go oscillations, and maintains smoother inter-vehicle spacing. In parallel, V2I-based Green Light Optimal Speed Advisory (GLOSA) enables a greater share of vehicles to adjust their approach speeds according to real-time signal phase and timing information, thereby reducing unnecessary stopping and idle time at red signals. As a result, vehicle trajectories become smoother and queue evolution more predictable, which allows the SAC_CCAV signal controller to operate under less volatile traffic states and exploit its learned policy more effectively.

4.3. Discussion

The experimental results demonstrate that the performance gains of the proposed SAC_CCAV framework are not solely due to the use of a more advanced reinforcement learning algorithm but rather stem from the coordinated interaction between infrastructure-level signal control and vehicle-level cooperative behavior. Compared with conventional DRL-based traffic signal control approaches that optimize signal timing independently [32–34], the proposed method actively reshapes traffic dynamics through CCAV coordination before congestion propagates to the intersection level. This bidirectional interaction reduces traffic state variability, which is a known challenge for value-based and deterministic policy gradient methods under highly stochastic demand conditions [41].

Furthermore, while model-based approaches such as Max-Pressure (MP) provide theoretical guarantees in stabilizing queue pressure [51], they lack adaptability to evolving traffic patterns and vehicle behavior heterogeneity. The superior performance of SAC_CCAV under high-demand scenarios indicates that entropy-regularized policy learning, combined with cooperative vehicle smoothing via CACC and GLOSA, enables more anticipatory and stable control decisions. These findings highlight the importance of the joint optimization of signals and vehicles, suggesting that future intelligent transportation systems should move beyond isolated signal adaptation toward integrated cyber–physical coordination.

5. Conclusions

This paper investigated a cooperative traffic control paradigm that jointly optimizes adaptive traffic signal control and connected autonomous vehicle behavior at a signalized intersection. By integrating a Soft Actor–Critic (SAC)-based signal controller with a cooperative connected and autonomous vehicle (CCAV) framework, the proposed SAC_CCAV approach establishes a tightly coupled vehicle–infrastructure control loop that leverages both vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications.

At the infrastructure level, traffic signal control is formulated as a Markov decision process and optimized using an entropy-regularized reinforcement learning framework. The use of SAC enables stable and anticipatory signal phase selection under stochastic and time-varying traffic demand. At the vehicle level, the CCAV framework combines cooperative adaptive cruise control (CACC) and Green Light Optimal Speed Advisory (GLOSA) to achieve cooperative and signal-aware longitudinal motion. While CACC enhances string stability and suppresses stop-and-go oscillations through V2V coordination, GLOSA exploits real-time signal phase and timing information to reduce unnecessary stopping and idle time at red signals.

Extensive microscopic simulations conducted in SUMO demonstrate that the proposed SAC_CCAV framework consistently outperforms representative deep reinforcement learning baselines, including DQN, DDPG, and A2C, as well as the classical Max-Pressure (MP) controller, across a wide range of traffic demand levels. The results show substantial improvements in cumulative reward, intersection throughput, total waiting time, and average vehicle speed. In particular, under high-demand conditions, the proposed method achieves more than an 80% reduction in vehicle waiting time while maintaining stable traffic flow, whereas the learning-based baselines suffer from severe congestion and MP stabilizes at a relatively low operational speed without fully mitigating queue accumulation.

Beyond quantitative performance improvements, the results provide insight into the fundamental mechanisms underlying the observed gains. The performance advantage of SAC_CCAV arises from the synergistic interaction between signal-level decision-making and vehicle-level cooperative control. By smoothing vehicle arrivals at signalized intersections and stabilizing platoon discharge during green phases, the CCAV layer reduces traffic state variability and improves the predictability of system dynamics. This, in turn, enhances the learning efficiency and robustness of the SAC-based signal controller, highlighting the importance of joint optimization rather than the isolated control of infrastructure or vehicles.

Despite its promising performance, this study focuses on a single isolated intersection under ideal communication conditions. Several limitations should be acknowledged. First, the experimental validation is conducted in a microscopic simulation environment, and real-world deployment may introduce additional uncertainties not captured in SUMO. Second, the communication process between vehicles and infrastructure is assumed to be reliable, without explicitly modeling packet loss, transmission delay, or bandwidth constraints, which may influence cooperative performance in practice. Third, the entropy coefficient in the SAC framework is treated as a fixed parameter, and although it provides stable training behavior in the considered scenarios, adaptive tuning strategies may further enhance robustness under highly non-stationary traffic patterns. In addition, the current study does not explicitly consider computational constraints associated with real-time deployment.

Several directions for future research are therefore identified. Extending the proposed framework to multi-intersection and corridor-level networks will be essential to assess scalability and network-wide coordination effects. Moreover, more realistic conditions, such as partial observability, sensor noise, communication delays, and mixed traffic with heterogeneous automation levels, should be incorporated to evaluate robustness in real-world deployments. Finally, integrating safety-aware constraints and hybrid control architectures, such as combining deep reinforcement learning with model predictive control or rule-based safety supervisors, represents a promising avenue for practical implementation in intelligent transportation systems.

Author Contributions: L.D.N. developed the methodology and prepared the original draft; P.M.T. validated the obtained results; S.H.B. and K.K.Y. reviewed and edited the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by HLB OceanTech Co., Ltd. under a commissioned research project entitled "Development of Motion Stabilization Technology" conducted through the Industry–Academic Cooperation Foundation of Korea Maritime and Ocean University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

ATSC	Adaptive Traffic Signal Control
CCAV	Cooperative Connected and Autonomous Vehicle
CACC	Cooperative Adaptive Cruise Control
GLOSA	Green Light Optimal Speed Advisory
V2V	Vehicle-to-Vehicle Communication
V2I	Vehicle-to-Infrastructure Communication
V2X	Vehicle-to-Everything Communication
DRL	Deep Reinforcement Learning
RL	Reinforcement Learning
SAC	Soft Actor–Critic
DQN	Deep Q-Network
DDPG	Deep Deterministic Policy Gradient
A2C	Advantage Actor–Critic
MP	Max-Pressure
MDP	Markov Decision Process
HDV	Human-Driven Vehicle

Appendix A. String Stability Metrics for CCAVs

In longitudinal vehicle control, string stability refers to the property that disturbances (e.g., speed or acceleration fluctuations) introduced by a leading vehicle do not amplify as they propagate downstream along a platoon. Formally, a platoon is considered string-stable if the disturbance magnitude of each following vehicle does not exceed that of its preceding vehicle.

- *Speed Disturbance Amplification Gain* G_v

Let $v_i(t)$ denote the speed of vehicle i, and let $v^{-}{}_i$ represent its time-averaged speed over the evaluation window T. The speed fluctuation signal is defined as follows:

$$\widetilde{v}_i(t) = v_i(t) - v^{-}{}_i$$

The root mean square (RMS) of the fluctuation is computed as follows:

$$RMS(\widetilde{v}_i) = \sqrt{\frac{1}{T}\int_0^T \widetilde{v}_i^2(t)\,dt}$$

For each adjacent vehicle pair $(i-1, i)$, the speed disturbance amplification gain is defined as the following:

$$G_v(i) = \frac{RMS(\widetilde{v}_i)}{RMS(\widetilde{v}_{i-1})}$$

The maximum gain across the platoon is then as follows:

$$G_v^{max} = max_{i\geq 2} G_v(i)$$

A necessary condition for string stability is $G_v^{max} < 1$ which indicates the attenuation of speed disturbances downstream.

- *Acceleration Disturbance Amplification Gain* G_a

Similarly, let $a_i(t)$ denote the acceleration of vehicle i. The RMS acceleration magnitude is computed as follows:

$$RMS(a_i) = \sqrt{\frac{1}{T}\int_0^T a_i^2(t)\,dt}$$

The acceleration disturbance amplification gain is defined as the following:

$$G_a(i) = \frac{RMS(a_i)}{RMS(a_{i-1})}$$

and

$$G_a^{max} = max_{i\geq 2} G_a(i)$$

A value $G_a^{max} < 1$ indicates that acceleration fluctuations are not amplified along the platoon, implying moderated control effort and the suppression of oscillatory behavior.

References

1. Fattah, M.A.; Morshed, S.R.; Kafy, A.A. Insights into the socio-economic impacts of traffic congestion in the port and industrial areas of Chittagong city, Bangladesh. *Transp. Eng.* **2022**, *9*, 100122. [CrossRef]
2. Xu, S.; Sun, C.; Liu, N. Road congestion and air pollution-Analysis of spatial and temporal congestion effects. *Sci. Total Environ.* **2024**, *945*, 173896. [CrossRef]
3. Hunt, P.B.; Robertson, D.I.; Bretherton, R.D.; Royle, M.C. The SCOOT on-line traffic signal optimisation technique. *Traffic Eng. Control.* **1982**, *23*, 190–192.

4. Sims, A.G.; Dobinson, K.W. The Sydney Coordinated Adaptive Traffic (SCAT) System Philosophy and Benefits. *IEEE Trans. Veh. Technol.* **1980**, *29*, 130–137. [CrossRef]
5. Koukol, M.; Zajíčková, L.; Marek, L.; Tuček, P. *Fuzzy Logic in Traffic Engineering: A Review on Signal Control*; Koukol, M., Zajíčková, L., Marek, L., Tuček, P., Eds.; Mathematical Problems in Engineering; Wiley Online Library: Hoboken, NJ, USA, 2015. [CrossRef]
6. Esmaeili, M.; Anjomshoae, A.; Shahsavari-Pour, N.; Srisurin, P.; Banomyong, R. An optimization similarity fuzzy inference method for traffic signal control at an isolated intersection. *Multimodal Transp.* **2025**, *4*, 100234. [CrossRef]
7. Jafari, S.; Shahbazi, Z.; Byun, Y.C. Improving the Performance of Single-Intersection Urban Traffic Networks Based on a Model Predictive Controller. *Sustainability* **2021**, *13*, 5630. [CrossRef]
8. Yektamoghadam, H.; Nikoofard, A.; Behzadi, M.; Khosravy, M.; Dey, N.; Witkowski, O. Multi-criteria evolutionary optimization of a traffic light using genetics algorithm and teaching-learning based optimization. *Expert. Syst.* **2024**, *41*, e13487. [CrossRef]
9. Jamal, A.; Rahman, M.T.; Al-Ahmadi, H.M.; Ullah, I.; Zahid, M. Intelligent Intersection Control for Delay Optimization: Using Meta-Heuristic Search Algorithms. *Sustainability* **2020**, *12*, 1896. [CrossRef]
10. Mutambik, I. IoT-Enabled Adaptive Traffic Management: A Multiagent Framework for Urban Mobility Optimisation. *Sensors* **2025**, *25*, 4126. [CrossRef]
11. Rizwan, A.; Karras, D.A.; Dighriri, M.; Kumar, J.; Dixit, E.; Jalali, A.; Mahmoud, A. Simulation of IoT-based Vehicular Ad Hoc Networks (VANETs) for Smart Traffic Management Systems. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 3378558. [CrossRef]
12. Sutton, R.S.; Barto, A.G. *Reinforcement Learning Second Edition: An Introduction*; Adaptive Computation and Machine Learning Series; The MIT Press: Cambridge, MA, USA; London, UK, 2014.
13. Ladosz, P.; Weng, L.; Kim, M.; Oh, H. Exploration in deep reinforcement learning: A survey. *Inf. Fusion* **2022**, *85*, 1–22. [CrossRef]
14. Haydari, A.; Yilmaz, Y. Deep Reinforcement Learning for Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11–32. [CrossRef]
15. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A.A.A.; Yogamani, S.; Pérez, P. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 4909–4926. [CrossRef]
16. Genders, W.; Razavi, S. Using a deep reinforcement learning agent for traffic signal control. *arXiv* **2016**, arXiv:1611.01142. [CrossRef]
17. Zhao, R.; Wang, K.; Li, Y.; Fan, Y.; Gao, F.; Gao, Z. Safe Multi-Agent Deep Reinforcement Learning for the Management of Autonomous Connected Vehicles at Future Intersections. *IEEE Trans. Parallel Distrib. Syst.* **2025**, *36*, 1744–1761. [CrossRef]
18. Kuang, L.; Zheng, J.; Li, K.; Gao, H. Intelligent traffic signal control based on reinforcement learning with state reduction for smart cities. *ACM Trans. Internet Technol.* **2021**, *21*, 1–24. [CrossRef]
19. Zhang, X.; Riedel, T. Urban traffic control: Present and the future. *Int. J. Urban Sci.* **2017**, *21*, 87–100. [CrossRef]
20. Ibrokhimov, B.; Kim, Y.J.; Kang, S. Biased Pressure: Cyclic Reinforcement Learning Model for Intelligent Traffic Signal Control. *Sensors* **2022**, *22*, 2818. [CrossRef]
21. Khezri, R.; Steen, D.; Tuan, L.A. A Review on Implementation of Vehicle to Everything (V2X): Benefits, Barriers and Measures. In Proceedings of the 2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Novi Sad, Serbia, 10–12 October 2022. [CrossRef]
22. Shiwakoti, N.; Stasinopoulos, P.; Fedele, F. Investigating the state of connected and autonomous vehicles: A literature Review. *Transp. Res. Procedia* **2020**, *48*, 870–882. [CrossRef]
23. Kopelias, P.; Demiridi, E.; Vogiatzis, K.; Skabardonis, A.; Zafiropoulou, V. Connected & autonomous vehicles—Environmental impacts—A review. *Sci. Total Environ.* **2020**, *712*, 135237. [CrossRef]
24. Xi, J.; Wang, C.; Liu, H.; Wang, L. Completely Distributed Guaranteed-Performance Consensualization for High-Order Multiagent Systems with Switching Topologies. *IEEE Trans. Syst. Man. Cybern. Syst.* **2019**, *49*, 1338–1348. [CrossRef]
25. Hanif, O.; Gruber, P.; Sorniotti, A.; Montanaro, U. Network-Based Fractional-Order Control Algorithms for Vehicle Platooning. *IEEE Access* **2025**, *13*, 139992–140011. [CrossRef]
26. Elahi, A.; Alfi, A.; Modares, H. H_∞ Consensus of Homogeneous Vehicular Platooning Systems With Packet Dropout and Communication Delay. *IEEE Trans. Syst. Man. Cybern. Syst.* **2022**, *52*, 3680–3691. [CrossRef]
27. He, C.; Li, Y.; Xu, J.; Hao, W. Third-order model based nonlinear longitudinal control for heterogeneous connected vehicle platoon. In Proceedings of the 2019 IEEE International Conference on Real-Time Computing and Robotics, Irkutsk, Russia, 4–9 August 2019; pp. 106–111. [CrossRef]
28. Huang, S.; Sun, D.; Zhao, M. Distributed MPC-Based Hierarchical Cooperative Control for Mixed Vehicle Groups with T-CPS in the Vicinity of Traffic Signal Light. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 8003–8016. [CrossRef]
29. Krauss, S. Microscopic Modeling of Traffic Flow: Investigation of Collision Free Vehicle Dynamics. Ph.D. Thesis, University of Cologne, Cologne, Germany, 1997.
30. Treiber, M.; Hennecke, A.; Helbing, D. Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E* **2000**, *62*, 1805. [CrossRef]

31. Segata, M.; Lo Cigno, R.; Hardes, T.; Heinovski, J.; Schettler, M.; Bloessl, B.; Sommer, C.; Dressler, F. Multi-Technology Cooperative Driving: An Analysis Based on PLEXE. *IEEE Trans. Mob. Comput.* **2023**, *22*, 4792–4806. [CrossRef]
32. Ma, D.; Zhou, B.; Song, X.; Dai, H. A Deep Reinforcement Learning Approach to Traffic Signal Control with Temporal Traffic Pattern Mining. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11789–11800. [CrossRef]
33. Li, M.; Pan, X.; Liu, C.; Li, Z. Federated deep reinforcement learning-based urban traffic signal optimal control. *Sci. Rep.* **2025**, *15*, 11724. [CrossRef]
34. Fan, C.; Wang, F.; Zhou, B.; Ma, D. A Centralized Reinforcement Learning-Based Method for Traffic Signal Optimization Using an Adaptive Sequential Decision. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 13201–13216. [CrossRef]
35. Ma, Z.; Cui, T.; Deng, W.; Jiang, F.; Zhang, L. Adaptive Optimization of Traffic Signal Timing via Deep Reinforcement Learning. *J. Adv. Transp.* **2021**, *2021*, 6616702. [CrossRef]
36. Li, L.; Lv, Y.; Wang, F.Y. Traffic signal timing via deep reinforcement learning. *IEEE/CAA J. Autom. Sin.* **2016**, *3*, 247–254. [CrossRef]
37. Shafik, A.K.; Eteifa, S.; Rakha, H.A. Optimization of Vehicle Trajectories Considering Uncertainty in Actuated Traffic Signal Timings. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 7259–7269. [CrossRef]
38. Typaldos, P.; Kalogianni, I.; Mountakis, K.S.; Papamichail, I.; Papageorgiou, M. Vehicle Trajectory Specification in Presence of Traffic Lights with Known or Uncertain Switching Times. *Transp. Res. Rec.* **2020**, *2674*, 53–66. [CrossRef]
39. Khanmohamadi, M.; Guerrieri, M. Smart Intersections and Connected Autonomous Vehicles for Sustainable Smart Cities: A Brief Review. *Sustainability* **2025**, *17*, 3254. [CrossRef]
40. Zhang, X.; Fang, S.; Shen, Y.; Yuan, X.; Lu, Z. Hierarchical Velocity Optimization for Connected Automated Vehicles with Cellular Vehicle-to-Everything Communication at Continuous Signalized Intersections. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 2944–2955. [CrossRef]
41. Mao, F.; Li, Z.; Li, L. A Comparison of Deep Reinforcement Learning Models for Isolated Traffic Signal Control. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 160–180. [CrossRef]
42. Prashanth, L.A.; Bhatnagar, S. Reinforcement learning with function approximation for traffic signal control. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 412–421. [CrossRef]
43. Fellendorf, M.; Vortisch, P. Microscopic Traffic Flow Simulator VISSIM. *Int. Ser. Oper. Res. Manag. Sci.* **2010**, *145*, 63–93. [CrossRef]
44. Lopez, P.A.; Behrisch, M.; Bieker-Walz, L.; Erdmann, J.; Flötteröd, Y.-P.; Hilbrich, R.; Lücken, L.; Rummel, J.; Wagner, P.; Wiessner, E. Microscopic Traffic Simulation using SUMO. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018.
45. Chu, T.; Wang, J.; Codeca, L.; Li, Z. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1086–1095. [CrossRef]
46. Wang, T.; Zhu, Z.; Zhang, J.; Tian, J.; Zhang, W. A large-scale traffic signal control algorithm based on multi-layer graph deep reinforcement learning. *Transp. Res. Part C Emerg. Technol.* **2024**, *162*, 104582. [CrossRef]
47. Tan, T.; Bao, F.; Deng, Y.; Jin, A.; Dai, Q.; Wang, J. Cooperative deep reinforcement learning for large-scale traffic grid signal control. *IEEE Trans. Cybern.* **2020**, *50*, 2687–2700. [CrossRef]
48. Bohra, A.R.; Selvi, T.; Vishal, V.; Sunkara, P.S.; Sunku, B.; Jignesh, B.N. Reinforcement Learning for Adaptive Traffic Signal Control Using Deep Q-Networks. In Proceedings of the 1st International Conference on Sustainable Energy Technologies and Computational Intelligence: Towards Sustainable Energy Transition, SETCOM 2025, Gandhinagar, India, 21–23 February 2025. [CrossRef]
49. Casas, N. Deep Deterministic Policy Gradient for Urban Traffic Light Control. *arXiv* **2017**, arXiv:1703.09035. [CrossRef]
50. Xiong, Y.; Xu, K.; Zheng, G.; Li, Z. Learning traffic signal control from demonstrations. *Int. Conf. Inf. Knowl. Manag. Proc.* **2019**, *4*, 2289–2292. [CrossRef]
51. Ma, D.; Xiao, J.; Song, X.; Ma, X. A back-pressure-based model with fixed phase sequences for traffic signal optimization under oversaturated networks. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 5577–5588. [CrossRef]