

Received 25 May 2025, accepted 2 June 2025, date of publication 6 June 2025, date of current version 13 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3577165

RESEARCH ARTICLE

Simultaneous Optimization of Traffic Signal Control and Vehicle Platooning Scheme Based on Connected and Automated Vehicle (CAV) Technology

JIWON KANG¹, YONGJUN CHOI¹, AND KEEMIN SOHN^{1,2}

¹Department of Urban Engineering, Chung-Ang University, Seoul 06974, South Korea

²Department of Smart City, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Keemin Sohn (kmsohn@cau.ac.kr)

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2023, and in part by Korea Agency for Infrastructure Technology Advancement (KAIA) Grant funded by the Ministry of Land, Infrastructure and Transport, under Grant RS-2022-00141102.

ABSTRACT Connected and automated vehicle (CAV) technologies can improve the traffic operation and management in urban transportation systems. In particular, a great advantage is expected when CAV technologies are applied to the traffic signal control for intersections. So far in the academy, researchers have dealt with platooning vehicles moving in an approach to an intersection to reduce stops and delays without traffic signals. The present study focuses on synchronizing stopped vehicles at a stop line in severely congested traffic condition. If vehicles in a queue move synchronously when green light initiates, the traffic throughput can be maximized without start-up delays. A reinforcement learning (RL) algorithm is devised to simultaneously optimize both the traffic signal control and the platooning scheme in a unified modelling framework. The proposed approach outperforms both an adaptive traffic signal control without platooning and a fully automated intersection management (AIM) without traffic lights, as traffic flows are oversaturated.

INDEX TERMS Connected and autonomous vehicle (CAV), traffic signal control, platoon-based traffic control, reinforcement learning, throughput maximization.

I. INTRODUCTION

The landscape of traffic operation is undergoing a significant transition due to the advent of connected and automated vehicles (CAVs). Leveraging these technological advancements, traffic flow can be controlled, managed, and optimized in real time to an unprecedented degree.

A key area that stands to benefit from this transformation is the management of intersection controls. Intersection controls, pivotal in ensuring smooth urban transportation, constitute bottlenecks due to inefficient capacity and lack of real-time adaptability. The potential of CAVs and their

integrated connectivity offers a promising solution to this issue by implementing vehicle platooning.

Vehicle platooning is a driving strategy that synchronizes CAVs in a simultaneous manner, enabling the efficient use of road capacity and reducing congestion. These platoons, connected via vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication, can interact with each other and the traffic infrastructure, sharing crucial data that allows them to adapt to traffic dynamics swiftly.

The present study starts with the doubt that the performance of a fully automated intersection management (AIM) without traffic signals could deteriorate, as traffic flows are saturated. Instead, optimizing traffic signals with the platooning CAVs could improve the intersection capacity and reduce traffic delays. Many previous studies have already

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Quan.

automated the platooning scheme in given pretimed traffic signals [1], [2]. The trajectory of CAVs is planned, so that the moving platoon can pass through intersection without stops [3]. Other researchers integrated the optimization of traffic signal control with the control of vehicle trajectories under the condition where all vehicles are equipped with communication tool or human-driven vehicles (HDV) are mixed [4], [5], [6], [7]. Most previous methods were tested in unsaturated or near-saturated traffic conditions, except for a few studies that dealt with oversaturated traffic conditions [8].

The present study explores a practical concept of traffic signal to maximize the number of vehicles that passes through an isolated intersection in severely congested traffic conditions, by platooning stopped vehicles in queues at a stop line. This scheme differs from the conventional platooning strategy that allows to group running vehicles into a platoon to avoid stops and delays. We assumed that vehicles stopped at stop lines of an intersection can be synchronized as green light initiates, while simultaneously optimizing the platooning scheme and traffic signal control. Although previous studies separately considered moving and stationary platoons [7], [9], their focus was on controlling the moving platoon to pass through the intersection during the remaining green signal time, after the stationary queue has been cleared. The primal interest of the present study is in synchronously moving the stopped platoon on the initiation of green time without start-up delays when the traffic is severely congested. The proposed method was not developed to replace the existing method in all cases. Unless long queues are observed in all approaches, the existing optimization method focusing on moving vehicles can be a robust option for optimization [7], [8], [9].

The present study adopts a deep reinforcement learning (RL) model to simultaneously optimize both the vehicle platooning scheme and the traffic signal control. The distinction from other platooning methods is that only vehicles stopped in a queue are controlled in a synchronous manner for congested traffic conditions. The action of RL thus adds only two control variables for platooning scheme: the platoon length that determines how many stopped vehicles to move together and the platoon moving speed after the ignition of green light. RL algorithm is much simpler than mixed integer linear program (MILP) or dynamic program in that no platoon trajectory plan is necessary at a separate level from the optimization of traffic signal timing. Such simplicity is the greatest merit of the proposed method.

The utility of adopting RL algorithm in the present study is three-fold. First, a single RL model jointly optimizes the traffic signal and platooning scheme without hierarchy. Existing studies depended on multi-step or multi-level optimization [8], [9]. Second, the proposed RL model learns an optimal policy directly from state transitions in experiments. This model-free learning removes the need of vehicle trajectory planning for CAVs and HDVs. The learned policy returns the optimal signal timing and platooning scheme for a given traffic state every cycle of traffic signal. Last,

even though the RL model takes a long time to be learned, the inference of the learned model is not burdensome for online implementation. Compared to traditional optimization approaches such as MILP, which require repeated reformulation and solving during each planning cycle, reinforcement learning RL model enables real-time inference through learned policies. Moreover, the computational complexity of MILP increases exponentially as the number of vehicles grows, making it impractical under severe congestion. This model-free approach reduces computational overhead and enhances adaptability to dynamic traffic conditions. These characteristics make RL particularly suitable for managing complex, uncertain environments such as severely congested intersections.

The environment where every vehicle is connected under control cannot be realized in the near future. We thus assume that vehicles are connected only within a proximate area of an intersection. Each vehicle in a platoon reverts to its own independent operation after the platoon passes through the intersection area. To ensure safety post-deployment, the gap between successive vehicles in a queue is set when they join the queue considering the next platoon moving speed. Through simulation experiments, we delve into the practical performance of the proposed optimization scheme that integrates the traffic signal control with the platooning scheme for vehicles.

The next section briefly introduces the previous studies that have dealt with the optimization of traffic control assuming a connected environment. The third section describes the modeling framework to optimize the proposed traffic control scheme based on RL algorithm. The fourth section set up the experimental environment for the proposed RL model. The results are discussed and compared with those from reference models in the fifth section. In the last section, conclusions are drawn, and the further extensions are suggested.

II. LITERATURE REVIEW

CAVs and their operation at intersections have been the subjects of various scholarly research. Shladover et al. [10] conducted pioneering work on automated highway systems, exploring the idea of CAVs and their implications for traffic efficiency. Van Arem et al. [11] extended this work by exploring the potential benefits of vehicle platooning, primarily in terms of increased road capacity and reduced traffic congestion.

A comprehensive survey by Li et al. [12] classifies recent developments in urban traffic control under mixed traffic environments with CAVs. The survey categorizes prior work into several key areas, including trajectory-aware signal optimization, route and speed planning under mixed traffic, cooperative control at both intersection and corridor levels, and control under perception and communication uncertainty. Their framework also identifies core challenges such as scalability, partial connectivity, and real-time adaptability. Building upon this classification, the present study targets an overlooked operational phase—synchronizing departures

of queued CAVs under oversaturated conditions—using reinforcement learning.

Within this broader context, one of the earliest fully connected control concepts was the AIM framework. Assuming the fully connected environment, Dresner and Stone [13] proposed the idea of AIM and introduced a reservation-based system that would enable vehicles to cross intersections without stopping, reducing the delays associated with traditional traffic signals. Following these, recent years have seen a shift toward cooperative strategies that combine V2V and V2I communications. Lee and Park [14] built upon this concept by incorporating V2I communications for improved intersection efficiency. Their work demonstrated how utilizing V2I communication could significantly improve traffic efficiency and reduce waiting times at intersections without traffic signals. While significant research efforts have been devoted to intersection management in CAV environments, the specific problem of passing vehicles through intersections without delay requires the fully connected and automated environment that cannot be realized in the near future.

Moreover, even if every vehicle is equipped with a communication unit in the future, it should be noted that the capacity of AIM will never be infinite. In severely congested urban areas, even an advanced AIM method cannot help reaching the capacity limitation. The capacity of AIM system can handle at most double the volume of a fixed traffic signal under ideal conditions [15].

Several researchers have examined platoon-based traffic signal control in constrained environments [16], [17]. They postulated that traffic signal timings can be optimized by treating incoming traffic at intersections as platoons, leveraging CAV technology to detect vehicle platoons approaching from upstream. Conversely, some studies have proposed a planning-based approach to traffic signal control, adjusting the arrival of individual vehicles to the current signal timing [1], [2].

Whereas both platoon- and planning-based traffic signal controls view vehicle operations as the state variable to be estimated, the signal vehicle coupled control (SVCC) considers both signal phases and vehicle operations as control variables for optimization [4], [5], [6], [7], [8], [9], [18], [19]. More concretely, the signal timing affects the movement of individual vehicles in SVCC, and at the same time, vehicle trajectories are the main input for setting the optimal signal. Such integrated optimization has depended on a bi-level programming scheme [19], [20], [21].

Feng et al. [19] used a bi-level program wherein vehicle delays are minimized at the upper level, and vehicle trajectories are optimized at the lower level to minimize fuel consumption and emissions. Tajalli and Hajbabaie [7] set up a weighted objective function wherein the first term prevents the queue buildup at the minor approaches, and the second term minimizes the number of stops at intersections. This unified optimization framework, however, demands to find proper weights for the two terms in advance. These two

mentioned studies allow for the mixed operation of CAVs and HDVs.

Ma et al. [8] introduced the shared-phase-dedicated-lane based intersection control in the mixed traffic condition. They extended the original traffic signal control scheme [22] that devised the concept of blue phase. Phases and lanes were dedicated for CAVs by sharing the dedicated lanes with HDVs during shared phases. The extended modelling framework included the three-level hierarchy. The upper-level model found an optimal barrier duration of the standard NEMA signal specification, and the middle level model optimized phase sequence and duration. At the lower level, the platooning and trajectory plan were made based on the outcomes from the middle level, and the travel time of CAVs derived from the lower-level model fed the middle level model.

Traditional optimization methods such as MILP and trajectory-based planning have proven effective under moderate or near-saturated conditions, where system dynamics are deterministic and full trajectory information is available. Such methods, including the single-layer MILP approach by Liu et al. [23], are well-suited for structured settings but less adaptable to oversaturation. However, these approaches often rely on hierarchical formulations, pre-defined inputs, and offline computation, making them difficult to apply in real-time or under severe congestion. Furthermore, as traffic becomes denser, the computational complexity of MILP grows exponentially, limiting its feasibility in highly saturated environments.

In contrast, the present study targets a distinct problem: synchronizing departures of queued CAVs at the stop line during green initiation—an operational phase that has received little attention in prior research. Unlike planning trajectories for approaching vehicles, this problem requires reactive, state-dependent control under high-density and uncertain conditions.

To address this challenge, we adopt a RL framework capable of learning effective control policies through direct interaction with the environment. RL enables real-time policy inference without explicit trajectory prediction, and has been shown to outperform conventional optimization in dynamic traffic scenarios. Chu et al. [24] demonstrated the superiority of multi-agent RL in large-scale networks, and Milić et al. [25] provided a comprehensive review confirming RL's adaptability under uncertainty and variability.

Recent studies have further emphasized the importance of robust RL strategies under practical deployment constraints. Li et al. [26] proposed a cooperative perception-based signal control method effective under low CAV penetration, while Shi et al. [27] developed a distributed RL framework resilient to communication failures.

Building on these advantages, our model uses RL to jointly manage signal timing and platoon synchronization, offering a scalable, data-driven alternative to traditional methods in oversaturated traffic environments.

Existing methods to jointly control platooning and traffic signal timing were tested against unsaturated and near saturated traffic conditions. The main target of the proposed method is a severely congested intersection. Ma et al. [8] tested their model against over-saturated traffic conditions where the demand inflow reached up to 3,600 vph for each 4-lane approach, and Tajalli and Hajbabaie [7] tested their model against traffic demand of 1,080 vph for each 2-lane approach. The present study tested the demand inflow up to 8,800 vph for each 4-lane approach. We empirically confirmed that our method yields the maximum throughput larger than that of Ma et al. [8] under the same condition.

It is obvious that no matter how efficient an intersection management scheme is under a fully connected environment, the scheme may fail as the traffic demand flow exceeds its capacity limit. The present study aims to address this gap and provides a practical methodology that can be applied to the congested traffic condition. This study is anchored in the SVCC concept and devised an integrated optimization method of traffic signal timing alongside vehicle platooning management within a unified RL modelling framework.

III. MODELING FRAMEWORK

A. BASIC CONCEPT

For individual intersections, the present study dynamically optimizes both the traffic signal control and the vehicle platooning scheme within a single unified modelling framework. The platooning strategy envisions that only vehicles halted at a stop line will be the target to simultaneously move as the green light activates. Stopped vehicles within each lane group that is defined by shared signaling and capacity, are consolidated into a platoon. For a comprehensive definition of a lane group, refer to the Highway Capacity Manual (HCM) [28].

Simultaneously, decision variables for signal control, such as cycle and phase ratios, are optimized in conjunction with platooning variables. These platooning variables encompass the platoon's length and movement speed. The platoon length defines the maximum distance within which stopped vehicles can move together at the initiation of the green light. When an approach of an intersection has a long queue, the decision variable plays a role of cutting the stopped queue to determine how many vehicles need to be discharged. If the intersection of interest is not saturated, the RL algorithm is likely to select the entire queue as a platoon.

While the speed of the moving platoon is determined, the spacing between vehicles is ascertained based on the speed to guarantee safety upon platoon deployment. In other words, to secure the safety to avoid collision while a platoon moves, a vehicle joining an existing queue stops a distance from the preceding vehicle. The spacing is determined linked with the platoon moving speed and is kept while passing through the intersection.

Eq. (1) illustrates the relationship between spacing and platoon movement speed. Here, the safety spacing, $s(v)$, correlates to the platoon speed v . Both δ_0 and δ_1 are constants

set to 0.1 s and 0.4 m, respectively [29].

$$s(v) = \delta_0 + \delta_1 v \quad (1)$$

An integrated adaptive traffic signal control is merged with a vehicle platooning strategy for managing vehicles stopped at stop lines. Fig. 1 delineates the basic concept of the proposed method, combining the optimal traffic signal control with a strategic platoon formation and movement scheme. The uppermost row illustrates the vehicle queue at the end of a red signal. During this red phase, vehicles align considering the spacing associated with the platoon-moving speed for the upcoming green phase. The subsequent two rows depict the queue transitioning as a platoon at the control speed upon green signal initiation. The bottom row captures the moment the green light terminates, with the control system dictating its duration. After the platoon passes through the communication boundary, the control is returned to individual vehicles and the platoon releases.

The decision variables of platooning scheme are fully correlated with the traffic signal timing. Intuitively, the platoon length should be proportional to the duration of the signal phase assigned to it. Moreover, spacing between vehicles in a platoon, which is determined by the platoon moving speed, also affects the duration of the corresponding signal phase. Jointly optimizing two types of decision variables is thus inevitable. It is, however, difficult to do so using a formal mathematical programming. This is a rationale for mobilizing a RL-based algorithm in this study.

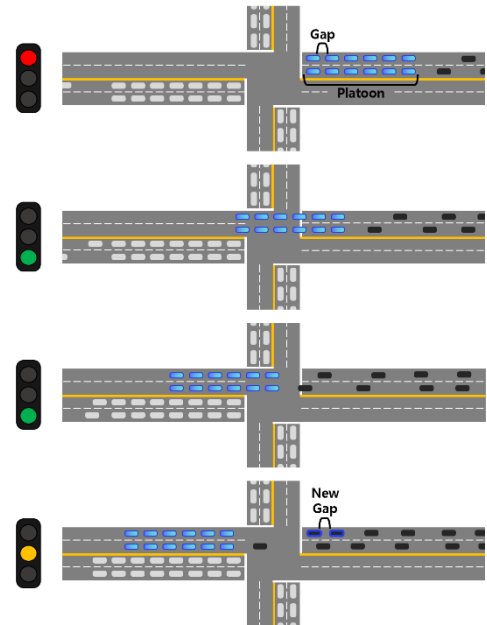


FIGURE 1. The concept of optimizing the traffic signal control with the platooning strategy.

B. DEFINITION OF STATE, ACTION, AND REWARD

RL algorithm solves a Markov decision problem (MDP) that is defined as a state set $\{s_t\}$, an action set $\{a_t\}$, a set

of transition probabilities $\{P(s_{t+1}, s_t, a_t)\}$, and a discount rate γ . s_t denotes a state of environment at time t , and a_t is an action taken at time t . $P(s_{t+1}, s_t, a_t)$ represents the transition probability that the environment's state would change to s_{t+1} in the next time interval after taking a_t in the current state s_t . However, in most cases, the transition probability cannot be modeled explicitly. The present study adopts a model-free RL algorithm, so the model is implicitly learned from experimented transitions. While implementing the algorithm, a transition tuple (s_t, a_t, r_t, s_{t+1}) observed at each time interval is continuously stored in the replay set (D) and resampled for learning in the forthcoming time intervals. The strongest merit of RL resides in this self-generation of training data. Basically, the model learns by remembering good and bad experiences through immediate rewards (r_t).

In the present study, the state is comprised of queue length and delay for each lane group. An action determines a cycle length and signal phase ratios for the traffic signal control and chooses platoon length and speed for platoon optimizing scheme. A discount rate (γ) should be chosen to be less than 1, which adjusts the future reward to be equivalent to the present value. The proposed RL algorithm takes an action every cycle of traffic signal, so the time interval varies while implementing RL algorithm.

As a summary, an action includes four control variables such as the traffic signal cycle length, phase ratio, platoon length, and platoon speed. The phase ratio is assigned to each signal phase, and the platoon length and speed are specific to each lane group that moves together while sharing the same signal phase. The proposed RL algorithm returns an optimal action every cycle of traffic signals. An optimal action for the next cycle can be obtained at the beginning of the current cycle because the inference time is much shorter than its counterparts such as MILP or dynamic program. This scheme ensures that the spacing between consecutive vehicles, when forming a queue in the current cycle, can be adjusted to match the platoon moving speed for the next cycle [see Eq. (1)].

The most important determination for a RL algorithm is how to define the immediate reward. The throughput maximization is the goal of traffic signal control when traffic flows are oversaturated [30], [31], [32], [33]. We confined the immediate reward to +1 and -1 following the original experiments of the deep Q-network [34]. The reward is set at +1 when the average throughput in the current cycle exceeds a threshold value; otherwise, it is set at -1. This binary reward structure simplifies policy learning but places critical importance on how the threshold is defined.

In early experiments, we used the throughput from the immediately preceding episode as the baseline threshold. However, this approach resulted in unstable learning, as short-term traffic fluctuations frequently triggered reward reversals. To mitigate this, we explored a moving average as a more stable reference. While effective at smoothing noise, the choice of window size introduced trade-offs: shorter windows remained volatile, while longer windows reduced responsiveness to recent performance trends.

Our final design adopts a hybrid strategy. During the initial stage of learning (episodes 1–100), we use a cumulative average over the previous episode as threshold to determine the reward, which provides smooth early feedback. After episode 100, we fix the window size to $N = 100$ for determining the threshold, allowing the reward signal to reflect recent changes while maintaining stability. This approach improved convergence and helped the agent learn more reliable policies. At the final stage, we observed that an almost constant threshold was applied for a stable convergence. As a hyper-parameter, the schedule for time window was empirically selected after testing many plausible alternatives. The next section will showcase the altering schedule for the window size.

C. SOFT ACTOR-CRITIC RL ALGORITHM

A soft actor-critic (SAC) RL algorithm was adopted to dynamically optimize both the traffic signal control and platooning scheme [35]. SAC is an advanced off-policy RL algorithm that is efficient for continuous state and action spaces. The state space is continuous since the queue length and traffic delay are chosen to be the state for each lane group. The proposed RL model also has a continuous action space consisting of cycle length, phase ratios, platoon length, and platoon moving speed.

SAC integrates the actor-critic framework with an entropy regularization and maximizes the combination of expected return and entropy. This dual optimization offers a robust balance between exploration and exploitation and enhances sample efficiency and stability. SAC uses neural approximators for action value function and policy function, and, for training, alternately updates these neural networks using the stochastic gradient descent algorithm. In addition, the temperature parameter, which plays a role of balancing entropy and has a great influence on the convergence, is optimized together with the neural approximators. In particular, instead of directly approximating a policy function with a neural network, an action is assumed to follow the Gaussian distribution, and its mean is reparametrized by using a neural network [35].

Eq. (2) denotes the return (R) be maximized in SAC to obtain an optimal policy. The objective of the standard RL is to maximize the future cumulative rewards. On the other hand, SAC adds an entropy to this standard return. Including an entropy widely explores policy domain, captures multiple modes of an optimal policy, and improves learning speed. The first term in Eq. (2) denotes the standard return, and the second term indicates the entropy. $r(s_t, a_t)$ denotes the reward function that returns an immediate reward after taking an action (a_t) in the current state (s_t). ρ_π denotes the state-action marginal of the trajectory distribution induced by a policy $\pi(a_t|s_t)$. α denotes the temperature parameter to be optimized during training and inference times. This hyper-parameter balances the standard return and the entropy of policy.

$$R = \sum_t E_{(s_t, a_t) \sim \rho_\pi} [r(s_t, a_t) + \alpha \log \pi(s_t, a_t)] \quad (2)$$

The policy-based RL algorithm alternates between the policy evaluation and the policy improvement. The policy evaluation step of SAC algorithm is to determine parameters (θ) of the soft action value function (Q_θ). Eq. (3) denotes the loss function representing the soft Bellman residual for action value function. D is a replay set that stores previous transition experiences, which allows for off-policy learning.

$$J_Q(\theta) = E_{(s_t, a_t) \sim D} \left[\frac{1}{2} (Q_\theta(s_t, a_t) - r(s_t, a_t) - \gamma E_{s_{t+1} \sim p} [V_{\bar{\theta}}(s_{t+1})])^2 \right] \quad (3)$$

$V_{\bar{\theta}}(s_{t+1})$ is the target soft state value function, and Eq. (4) represents the sum of the soft state-action value function and the entropy of policy.

$$E_{s_t \sim p} [V_{\bar{\theta}}(s_t)] = Q_{\bar{\theta}}(s_t, a_t) + \alpha \log \pi_\phi(a_t | s_t) \quad (4)$$

For each policy evaluation step, both parameters $\bar{\theta}$ and ϕ should be fixed, so that minimizing the loss function [Eq. (3)] can be only with respect to θ . The algorithm uses the incumbent ϕ determined in the previous policy improvement step. $\bar{\theta}$ is updated for every policy evaluation step using a simple smoothing rule denoted by Eq. (5). After initializing $\bar{\theta}$ with the starting value of θ , it is updated by an exponentially moving average of the original θ . τ should range from 0 to 1 acting as the hyper-parameter to balance between behavioral and target parameters.

$$\bar{\theta} = \tau \theta + (1 - \tau) \bar{\theta} \quad (5)$$

The other loss function is for the policy improvement to determine ϕ . In the policy improvement step, SAC updates the policy function towards the exponent of the new action value function. Practically, the policy function is assumed to follow the Gaussian distribution, and the mean of the distribution is reparametrized by a neural network. The discrepancy between the two distribution functions is measured by the Kullback-Leibler (K-L) divergence. The divergence can be reduced to Eq. (6). In this loss function, ε_t is a random deviate from the unit Gaussian distribution. $f_\phi(\varepsilon_t; s_t)$ is the reparametrized function for the mean of policy distribution. The policy function (π_ϕ) is parameterized by ϕ and is thus implicitly defined in terms of $f_\phi(\varepsilon_t; s_t)$. The loss function is minimized to update ϕ with the other parameter (θ) fixed at the incumbent values derived in the previous policy evaluation step.

$$J_\pi(\phi) = E_{s_t \sim D, \varepsilon_t \sim \mathcal{N}} [\alpha \log \pi_\phi(f_\phi(\varepsilon_t; s_t) | s_t) - Q_\theta(s_t, f_\phi(\varepsilon_t; s_t))] \quad (6)$$

The original SAC algorithm assumes the temperature parameter is fixed during training [36]. A more recent amelioration encompasses choosing an optimal temperature parameter for each policy improvement step while training [37]. In principle, constraining the entropy with a fixed lower bound ($\bar{\mathcal{H}}$), the temperature parameter (α) is regarded as a Lagrangian dual variable that can be optimized

along with primal variables (θ and ϕ). Eq. (7) denotes the additional loss function to update α for each policy iteration.

$$J(\alpha) = E_{a_t \sim \pi_\phi} [-\alpha \log \pi_\phi(a_t | s_t) - \alpha \bar{\mathcal{H}}] \quad (7)$$

The lower bound ($\bar{\mathcal{H}}$) is another hyper-parameter that can be set as the negative value of action dimension. The present RL environment has 21 different actions (a cycle length, 4 signal phase ratios, 8 platoon lengths, and 8 platoon moving speeds), so the value was set at -21 . On the other hand, other hyper-parameters are empirically determined after testing as many candidate values as possible. The next section will introduce hyper-parameters applied to the present experiments.

The overall algorithm starts with randomly initializing parameters (θ , ϕ , and $\bar{\theta}$). For each iteration, an action (a_t) is sampled from the incumbent policy (π_ϕ), traffic simulation steps proceed during the current cycle length to sample the next state (s_{t+1}), the replay set (D) is updated with the new transition [$s_t, a_t, r(s_t, a_t), s_{t+1}$]. If the replay set reaches the predefined size, the oldest transition is discarded. After completing this experience collection from the environment with the current policy, stochastic gradient steps repeat alternately for each loss function and update both neural networks and the temperature parameter from batches sampled from the replay set (D). For each gradient step, once two different sets of model parameters (θ and ϕ) and the temperature parameter (α) are updated in turn, and $\bar{\theta}$ is updated using Eq. (5). For more detailed explanation of the algorithm, readers can refer to [36].

The neural architecture of the soft action value function (Q_θ) and the reparametrized function for the mean of policy [$f_\phi(\varepsilon_t; s_t)$] are empirically chosen after testing many plausible specifications, also considering the parsimony of model architecture. Table 1 shows the specification of the approximation functions that were applied to present experiments. The model training was performed on a single GPU, which was a NVIDIA GeForce GTX 1080Ti with 11 GB of GDDR5 memory.

In the proposed model, the RL agent determines four interrelated control variables: platoon speed, platoon length, signal cycle length, and green-time allocation. These variables interact nonlinearly under oversaturated traffic conditions. For example, selecting a higher platoon speed requires greater spacing between CAVs to maintain safety, which reduces the number of vehicles in a platoon. Conversely, lower speeds enable denser packing and longer platoons, which may require longer green times. Rather than encoding these trade-offs explicitly, the SAC agent learns them implicitly through repeated interaction with the environment. By observing how different combinations of control variables impact throughput, the agent gradually refines a policy that balances conflicting objectives without manual parameter tuning or optimization constraints.

TABLE 1. The architecture of neural networks to approximate value and policy functions.

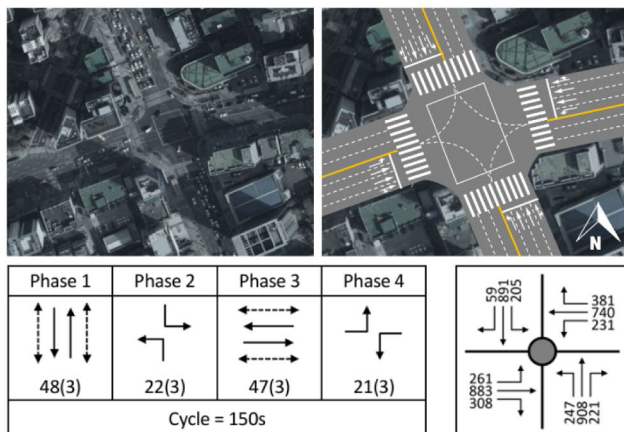
Function	Model specification
Soft action value function (Q_θ)	Input : $\langle s, a \rangle (2+4)$
	[Fully connected(32), relu]
	[Fully connected(32), relu]
	[Fully connected(32), relu]
	[Fully connected(16), relu]
	[Fully connected(1), linear]
	Output : $Q(s, a)$
Reparametrized function for the mean of policy [$f_\phi(\epsilon_t; s_t)$]	Input : $\langle s \rangle (2)$
	[Fully connected(64), relu]
	[Fully connected(32), relu]
	[Fully connected(16), relu]
	[Fully connected(21), tanh]
	[Fully connected(21), softplus]
	Output : $[\mu_1, \dots, \mu_{21}, \sigma_1^2, \dots, \sigma_{21}^2]$

IV. SIMULATION EXPERIMENT SET-UPS

A. TESTBED AND TRAFFIC CONDITIONS

A real intersection in the Seoul metropolitan area serves as the testbed for this study. Fig. 2 illustrates the layout of the four-leg intersection. The figure also displays the traffic volumes for each direction, determined based on the turning rates observed in actual traffic flows. The SUMO, an open-source traffic simulator, was employed to train the RL algorithm.

While detailed two-dimensional movement models are commonly applied to simulate vehicle behavior near intersections, this study focuses solely on the departure behavior of queued vehicles at the stop line. In this context, lateral movements are negligible, and vehicle motion is effectively constrained within fixed lane groups determined by signal phases. Based on the operational definition in the Highway Capacity Manual, we model departures as one-dimensional forward movement within each lane group. This assumption simplifies the control structure and remains consistent with the physical behavior of signalized intersections.

**FIGURE 2.** The layout of the testbed intersection.

The training consisted of 250 episodes, with each episode representing an hour of simulation time. The total simulation time for each scenario amounted to approximately 250 hours.

Even though the RL model takes a long time (≈ 12.5 hours) to be trained, the inference of the model is no burden to online implementation since no simulation is engaged. Notably, existing methods should repeatedly solve a time-consuming optimization problem every planning period.

Three traffic conditions were considered in terms of the level of saturation: undersaturated (mirroring actual traffic conditions), near saturated, and oversaturated conditions. Table 2 lists these scenarios along with their associated demand factors and LOS classifications. A single scenario was set up for each of the former two conditions. The traffic volumes observed during the morning peak hour were deemed representative of an undersaturated condition, given that the current service level is classified as “C” according to the Korea Highway Capacity Manual (KHCM) [38]. The near-saturated condition represents a slightly congested traffic scenario. The traffic volume of the scenario was generated by amplifying the observed traffic volumes using demand factors while retaining existing turning patterns. The modified traffic scenario yields a service level of “E” under the current pretimed signal operation. This near-saturated traffic volume aligned with that applied in the previous studies [7], [8], [9].

The oversaturated traffic condition signifies three intensely congested scenarios, where the service level becomes immeasurable under the pretimed traffic signal control. The most seriously congested scenario was set by reflecting the capacity of uninterrupted traffic flows. That is, the upstream demand flow was set at 2,200 vph for each lane, which means that forming a queue at stop line is inevitable for each intersection approach. The remaining two oversaturated scenarios were established by interpolating traffic inflow volumes between the worst scenario and the near-saturated scenario. As far as we ascertain, such severely congested conditions have never been tested with a CAV technology in previous studies.

TABLE 2. Traffic demand scenarios with applied demand factors and resulting LOS.

Scenario	Demand Factor	LOS**
Under saturated*	1	C
Near-saturated	2	E
Over saturated 1	3	FF
Over saturated 2	4	FFF
Over saturated 3	Max	FFF

*: observed traffic volumes of the real intersection

**: calculated based on actual pretimed(fixed) signal operation, following KHCM guidelines

B. REFERENCE METHODS

To validate the utility of the proposed RL model, three reference methods were employed. Firstly, as a baseline, the prevailing pretimed traffic signal control scheme was examined, in which the duration of traffic signal phases is assigned

proportional to the ratio of traffic volume to the saturation level of each lane group. This conventional control method struggles to adapt to traffic flows that vary in an irregular pattern.

The second method employs a SAC model but omits the platooning scheme [36]. This adaptive traffic signal control is anticipated to markedly improve traffic signal control performance compared to the baseline, as it can adjust to the stochasticity and variability in traffic flows. The performance of the method, however, deteriorates when the traffic demand surges. By employing this approach, the distinct impact of the platooning scheme can be assessed.

The third method involves an AIM method devoid of traffic lights. This approach reserves vacant slots in the time-space domain to ensure vehicles pass without stops or conflicts. Dresner and Stone [13] formulated a simple AIM method based on this concept, adopting the straightforward first-in-first-served (FIFS) rule. Upon approaching the intersection, a vehicle reserves a future trajectory that doesn't intersect with those of preceding vehicles. Should an incoming vehicle's path conflict with already reserved slots in the time-space domain, it must decelerate or stop to guarantee a safe crossing through the contentious zone.

While these rudimentary rules facilitate increased traffic throughput without traffic signals, its benefits persist only until traffic flow reaches saturation. Once traffic volumes near the capacity of this automated control, forming traffic queues becomes unavoidable. As mentioned earlier in the literature review, the capacity of AIM is at most double the capacity of pretimed traffic signal control scheme even under an ideal condition [15]. In a severely congested condition, the AIM cannot be a solution to resolve it. To address such oversaturated conditions, the proposed approach that combines both traffic signal control and a platooning scheme is more effective. The three aforementioned reference models were employed on the testbed, maintaining consistent traffic conditions with the proposed method.

C. HYPER-PARAMETERS APPLIED TO THE PRESENT EXPERIMENTS

Table 3 lists the hyper-parameters applied to the present experiments. The hyper-parameters to implement SAC algorithm are similar to those chosen by the original paper [36]. The reward threshold was, however, determined based on moving averages uniquely in the present study. The schedule for altering moving window size is presented in the table. Hyper-parameters for traffic simulations were also selected after testing as many plausible set-ups as possible.

V. RESULTS AND COMPARISONS

A. CONVERGENCE OF SAC

The proposed RL algorithm demonstrates convergence. The average reward for each episode is depicted in Fig. 3. The convergence of RL algorithms can be evaluated by

TABLE 3. Hyper-parameters applied to the present experiments.

Hyper-parameter		Applied value
SAC algorithm	Optimizer	Adam
	Actor Learning rate	0.0001
	Critic Learning rate	0.001
	Discount rate (γ)	0.95
	Replay set size	20,000
	Batch size	32
	Entropy bound ($\bar{H}$)	-21
	Target smoothing (τ)	0.001
	Target update interval	1
	Gradient step	1
Traffic simulation	Window size to determine reward threshold	Gradually increasing from 1 to 100 (Fixed at 100 after 100th episode)
	Simulation step	1 sec.
	Episode length	3,600 sec.
	Cycle length	100-200 sec.
	Amber time	3 sec.
	Platoon length	100-500 m
	Platoon moving speed	30-50 kph

using the average reward. The average reward was computed for each episode and plotted with orange line. The length of an episode is 3,600 seconds in the present experiment. Every signal cycle returns an immediate reward. Rewards were summed across all cycles in the episode, and the average reward was computed by dividing the sum by the number of cycles during the episode.

Fig. 3 shows that the average reward considerably fluctuates, so the moving average with the window size of 50 episodes was added in brown. The moving average line clearly demonstrates the convergence trend. The fluctuation in reward across episodes is general in RL, because the model is learned through exploration. The RL algorithm needs to test a random action at a certain proportion, while decreasing the proportion as the number of episodes increases.

The average reward went through an initial decay, which might be due to applying a smaller window size to choosing a reward threshold at the onset of the training process. The final average reward converged to a positive value close to +1, which indicates the algorithm's tendency to select actions that enhance throughput on average. Notably, an optimal policy does not always return the action that results in the positive reward, since the RL algorithm is stochastic.

While the average reward doesn't provide an immediate understanding of traffic control performance, we assessed the average throughput in vehicles per second as a function of episodes. Fig. 4 demonstrates that the average throughput stabilizes at a certain level, indicative of the maximum throughput achievable by the proposed control scheme. Notably, a pronounced jump in throughput is observed during the model's training phase. Less fluctuations are observed in the throughput than in the reward variation, which implies the stochasticity in reward does no harm to the model performance.

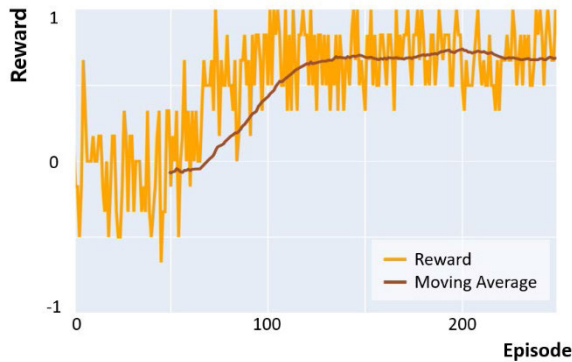


FIGURE 3. The convergence of the average reward of the proposed RL algorithm.

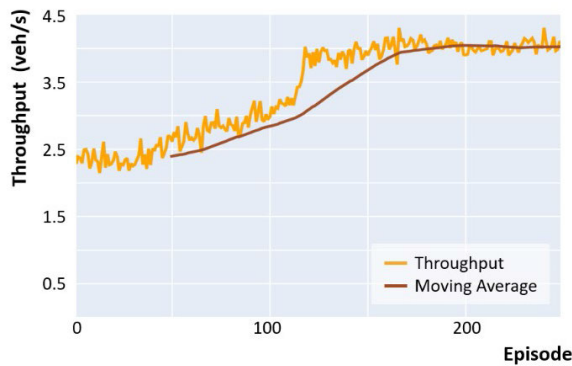


FIGURE 4. The convergence of the average throughput for the proposed RL algorithm.

B. COMPARISON WITH REFERENCE METHODS

The performance of models was measured by the throughput, which is formally defined as the number of vehicles that can pass through the intersection area during a unit time. This performance measure adequately evaluates oversaturated traffic conditions in an urban area. According to previous studies [31], [32], [33], the goal of traffic signal control differs from the level of congestion. Minimizing delays is paramount in managing uncongested traffic situations. Conversely, as traffic flows get saturated, maximizing throughput becomes the primary goal. The proposed traffic controller specifically targets oversaturated conditions, where fully automated control without signals might prove ineffective.

Table 4 presents the maximum throughput for each traffic control method post-convergence. Irrespective of the congestion level, both adaptive traffic control methods and the AIM approach surpassed the existing pretimed operation. However, comparing these three advanced methods necessitates caution, as their performance rankings shift based on varying congestion levels.

Fig. 5 illustrates the maximum throughput of each method in relation to the traffic congestion level. The throughput of each method approaches its own capacity as traffic condition worsens. No matter how much the demand inflow increases, the throughput cannot exceed its own capacity. The convergence pattern to the capacity differs by applied

TABLE 4. The performance of models measured by the maximum throughputs.

	(vehicles/second)			
	AIM operation	RL with platooning	RL w/o platooning	Fixed operation
Under-saturation	1.46	1.43	1.41	1.41
Near-saturation	2.86	2.72	2.45	2.20
Over-saturation 1	3.61	3.67	2.78	2.22
Over-saturation 2	3.63	4.05	2.79	2.24
Over-saturation 3	3.64	4.07	2.79	2.24

methods. The pretimed operation most quickly reaches its throughput ceiling at the level of near saturation. The AIM and RL without platoon approach their capacities at the first oversaturation level. The integrated RL model, on the other hand, is the last to approach maximum throughput.

Fig. 6 illustrates the convergence trend of each congestion level in relation to the number of episodes. The throughput of the proposed method converges to its own capacity as RL algorithm proceeds through episodes. The convergence pattern of the proposed method is stable before the traffic flow is saturated. After saturation, the throughput of the proposed method more fluctuates as the traffic congestion worsens. However, the fluctuation is less than that of AIM when traffic is oversaturated.

Comparing the two RL-based traffic signal control models, the integrated version incorporating the platooning scheme outperforms its counterpart without the scheme. The inclusion of the platooning strategy resulted in a 45.88% increase in throughput after 250 training episodes. This demonstrates the pure effect of platooning scheme, which supports the utility of the present study.

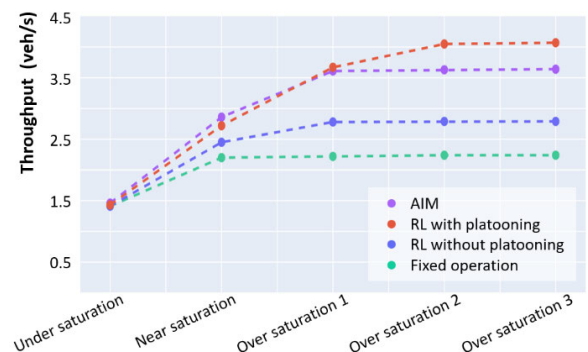


FIGURE 5. The convergence of the average throughput to the capacity of each method.

As anticipated, the AIM traffic control surpassed the two RL models in both under- and near-saturated traffic scenarios. However, since the performance was measured by

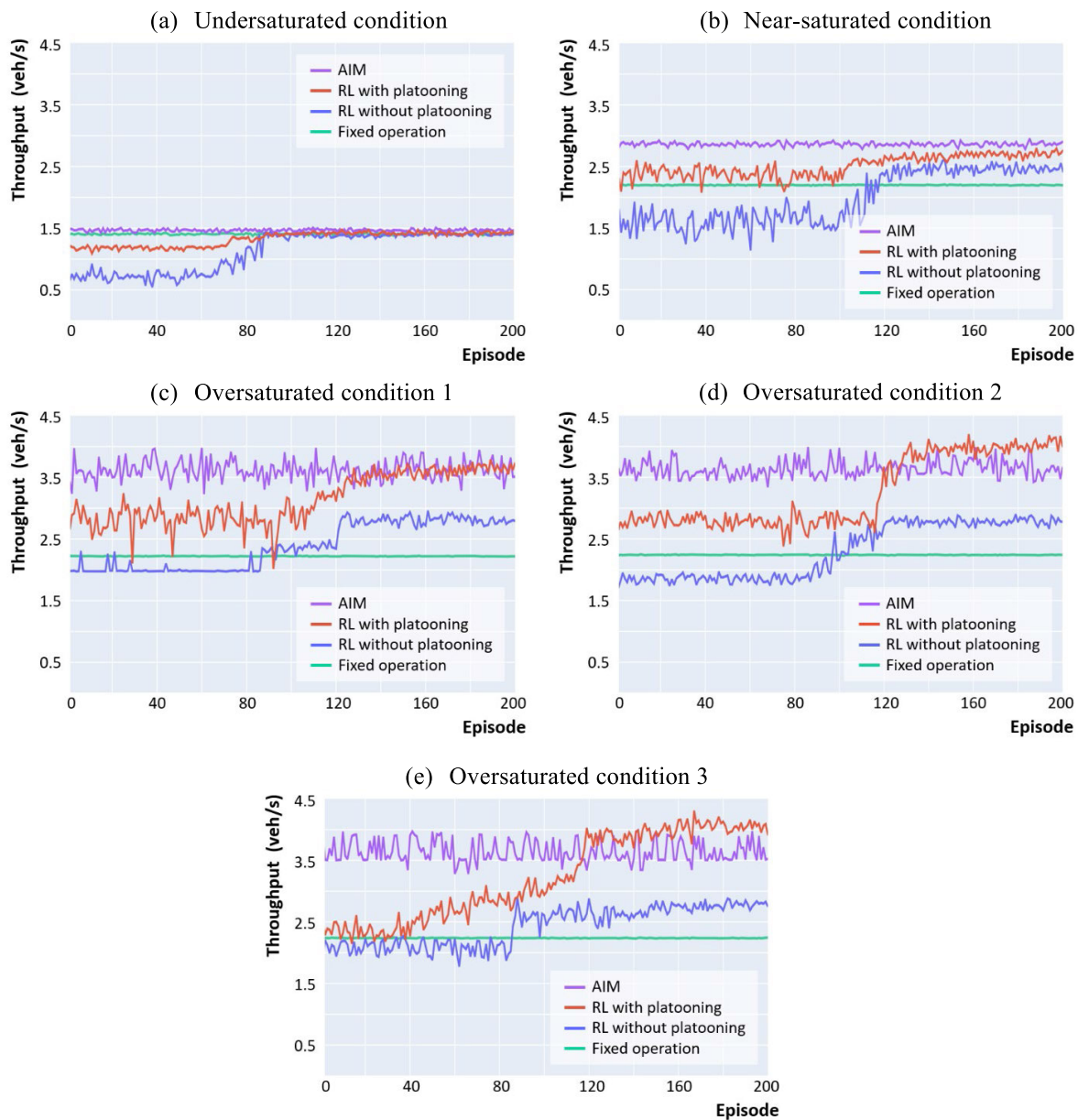


FIGURE 6. The comparison of maximum throughput for different traffic control methods.

throughput, the difference between methods is minimal in the undersaturated condition. If the performance is measured by delay, clearly AIM will outperform other methods. This automated control adeptly manages conflicting traffic flows, minimizing stops and delays, as long as the travel demand remains within the time-space control capacity.

However, with this automated control, as traffic volumes exceed a threshold ensuring uninterrupted operation, flows can intersect, necessitating stops to resolve conflicts. The capacity of AIM is at most double that of pretimed control even for the ideal condition [15]. Under intensely congested conditions typical of major cities, this control method might struggle with substantial traffic volumes converging at an intersection. Upon reaching a decision line to reserve a passing trajectory, vehicles may find the time-space domain fully

occupied. In such situations, they're compelled to wait for available slots, resulting in traffic delays. These shortcomings of AIM highlight the merits of our proposed model, which concurrently optimizes traffic lights and integrates the platooning scheme to maximize the throughput.

The AIM presents another limitation: it mandates a completely connected environment where every vehicle within the intersection's proximity must be controlled. The introduction of even a single HDV could disrupt and potentially derail system automation. Given this constraint, the immediate realization of such a fully automated control seems improbable.

As expected, the proposed RL model excels over the AIM system under extreme congestion. However, it lagged behind the AIM approach during undersaturated or near-saturated traffic scenarios. Consequently, the primary advantage of our

RL model is its potential to act as an alternative to AIM, particularly during peak traffic periods marked by heightened demand.

C. PERFORMANCE VARIATION WITH MIXED TRAFFIC

The proposed RL model presents significant practical benefits. Contrary to the AIM which demands a comprehensive connected environment, our RL model accommodates HDVs within a platoon. In scenarios where HDVs are interspersed among connected vehicles, they naturally adopt typical human car-following behaviors. We postulate that a CAV, when positioned immediately behind a HDV, will modulate its speed and acceleration to prevent conflicts with the leading HDV, subsequently assuming the role of leader for the trailing connected vehicles. These succeeding connected vehicles can then seamlessly align their movements with this leader. Details of how to move a platoon that includes HDVs through an intersection are as follows.

If a HDV leads a platoon, it runs as a normal vehicle in SUMO. When a HDV follows a preceding CAV or HDV, SUMO controls it according to the car-following model (IDM). We set up a strong assumption for a CAV behind a HDV in a platoon that the CAV can sense the distance from the leading HDV. The CAV was controlled so that the distance should be maintained longer than the given control spacing, and its speed cannot exceed the current synchronized platoon speed. That is,

$$v_{i+1}(t) = \min(v_{f\ i+1}(t), v_s(t)) \quad (8)$$

where $v_{i+1}(t)$ is the speed of a CAV behind a HDV at time t , $v_{f\ i+1}(t)$ is the CAV's speed that is needed to maintain the given control spacing, and $v_s(t)$ is the current synchronized platoon speed. Last, when a CAV follows another CAV, the follower's speed is synchronized with the leader's speed with the control spacing maintained. We also assumed throughout the entire experiments that the communication latency is negligible.

In this study, we assume that the RL agent can observe full traffic state information through a surveillance system, consistent with recent advances in "state-of-the-art" vision technology. At the simulation level, CAVs are assumed to detect surrounding vehicles (including HDVs) using onboard sensors such as radar or lidar. This allows each CAV to estimate the position and speed of nearby vehicles regardless of their connectivity.

HDVs were modeled using SUMO's default car-following settings, which include randomized desired speeds and acceleration from uniform distributions with added noise. This introduces basic variability in longitudinal movement among drivers. However, SUMO does not natively support individual differences in driver reaction time τ , which is another critical aspect of driver heterogeneity. To address this, we manually assigned each HDV a reaction delay τ sampled from a truncated lognormal distribution with mean 1.3 seconds, standard deviation 0.6 seconds, and an upper limit of 4.0 seconds. This follows empirical findings on reaction time variability [39],

TABLE 5. The performance decay according to the penetration of human-driven vehicles.

Models	Traffic condition	Maximum throughput (veh/s) for each penetration rate of human-driven vehicles			
		10%	20%	30%	50%
RL with platooning	Under-saturation	1.39	1.37	1.25	1.24
	Near-saturation	2.62	2.59	2.52	2.32
	Over-saturation 1	3.51	3.38	2.96	2.77
	Over-saturation 2	3.78	3.69	3.32	2.83
	Over-saturation 3	3.79	3.71	3.32	2.85

[40], [41] and provides a more realistic representation of mixed traffic behavior.

Allowing HDVs within a platoon could potentially impair the efficacy of the proposed RL control. We delved into understanding this potential performance degradation as the penetration rate of human-driven vehicles increases. Table 5 provides insights into the performance degradation stemming from the incorporation of human-driven vehicles. Simulation results show that when HDV penetration exceeds 50%, the performance of the RL-with-platooning model becomes nearly indistinguishable from that of the RL-without-platooning baseline. This result indicates that frequent interruptions from non-cooperative vehicles severely limit the formation of effective platoons. Consequently, we assessed four distinct penetration rates: 10%, 20%, 30%, and 50%. With a 10% integration of human-driven vehicles into a platoon, there's a reduction in maximum throughput by 0.28 vehicles per second. This decrement grows to 0.36 vehicles per second for a 20% penetration rate and further extends to 0.85 vehicles per second for the 30% rate. Notably, even with the inclusion rate of human-driven vehicles reaching up to 20%, our proposed methodology is still competitive with the AIM model in scenarios of extreme congestion. When the penetration rate is 50% for the oversaturated scenario, the maximum throughput (=2.85) is almost identical to that (=2.79) of the RL algorithm without platooning scheme.

Ma et al. [8] tested several oversaturated traffic conditions for an intersection that has almost the same configuration as ours. They assumed the half penetration rate of CAVs and resulted in the maximum throughput of 8,000 vph for the intersection. The present method passed 2.85 vehicles per second for the almost same condition, which is tantamount to 10,260 vph. Consequently, it turned out that our proposed approach can pass 28.25% more vehicles at the same condition.

D. PERFORMANCE UNDER IMBALANCED TRAFFIC DEMAND

In real-world urban intersections, traffic demand is often asymmetrically distributed across approaches. However, the

TABLE 6. Performance comparison of 4-phase signal control and 8-phase signal control under imbalanced traffic demand conditions.

	(vehicles/second)	
	4-phase signal control	8-phase signal control
Under-saturation	1.42	1.44
Near-saturation	2.61	2.76
Over-saturation 1	3.59	3.68
Over-saturation 2	4.04	4.06
Over-saturation 3	4.06	4.10

experiments presented in previous sections were based on the average daily traffic volumes observed at the real-world testbed intersection. Even in the severely congested scenario, congestion was simulated by applying a uniform demand factor across all approaches, which led to relatively balanced traffic demand among the four legs of the intersection.

To ensure the practicality and applicability of the proposed model in more realistic and diverse traffic conditions, we conducted additional experiments under intentionally imbalanced demand scenarios. In these scenarios, while maintaining the same total traffic volume as in the original experiments for fair comparison, we introduced directional imbalance by applying a directional demand factor of 0.65. This adjustment resulted in uneven traffic volumes across lane group movements.

To handle such imbalanced traffic demand, we employed an 8-phase signal control scheme based on the NEMA-compliant dual-ring structure, which allows independent control of each lane group's movement. The core SAC agent architecture, including all hyperparameters, was kept identical to previous experiments. The only modification was in the agent's action space: it was extended from selecting among 4-phase signals to selecting among 8-phase signals.

The 8-phase structure enables the model to more effectively allocate green time according to actual lane group demands. As a result, it can process traffic more efficiently under imbalanced demand conditions compared to the conventional 4-phase system.

Table 6 presents a comparison between the 4-phase and 8-phase signal structures under the imbalanced demand scenario. When the traffic demand is not saturated, the 8-phase signal structure outperformed the 4-phase system, improving intersection throughput by up to 0.15veh/s (540vph). However, as demand approaches saturation, the performance difference between the two signal schemes becomes negligible. As congestion becomes more severe and all approaches operate at or near capacity, even the approaches with lower demand begin to suffer from considerable delays, diminishing

the operational advantage of more flexible signal control. Consequently, under such over-saturated conditions, both the 4-phase and 8-phase signal structures operate near their capacity limits, and their relative effectiveness converges.

In summary, the 8-phase signal structure provides clear operational benefits under unsaturated and asymmetrically distributed traffic demand. However, its advantage diminishes under gridlocked conditions. Therefore, for urban intersections that exhibit directional traffic imbalance during off-peak periods, the adoption of an 8-phase control scheme can enhance traffic efficiency. On the other hand, during peak hours with over-saturated demand, a well-optimized 4-phase signal plan can still perform comparably.

VI. CONCLUSION

A. SUMMARY OF CONTRIBUTIONS

A novel integrated traffic signal and platooning control scheme has been formulated utilizing a deep RL algorithm. This platooning strategy specifically targets vehicles halted at the intersection's stop lines, optimizing traffic throughput during times of heavy congestion. Our proposed model surpasses both the conventional fixed operation and an RL approach devoid of platooning. Given the constraints of AIM control under extreme congestion, our scheme presents a viable alternative. Crucially, while automated controls necessitate a completely connected vehicle environment, our model accommodates platoons that comprise a mix of human-driven and connected vehicles. Considering the challenges in achieving a fully connected environment anywhere, our approach emerges as a more pragmatic solution for an intersection area.

Furthermore, the proposed framework demonstrated consistent performance advantages even under imbalanced traffic demand scenarios, where traffic flows are unevenly distributed across movement directions. In particular, under congested conditions with heavy traffic volumes, it consistently maintained superior throughput and operational efficiency across different signal control systems. These results demonstrate that our proposed framework is a practical method in saturated and dynamically changing urban traffic environments.

B. LIMITATIONS AND FUTURE WORK

This study offers a foundational step toward intelligent intersection control in congested environments, but meaningful real-world impact will require broader algorithmic generalization and deeper societal integration.

1) OPERATIONAL LIMITATIONS AND ALGORITHMIC EXTENSIONS

The proposed model assumes full traffic state observability, enabled through infrastructure-based surveillance and onboard CAV sensors. This setting simplifies the problem and allows the RL agent to operate under ideal conditions. However, such assumptions limit real-world applicability. Future work should explicitly address partial observability

by incorporating frameworks such as POMDPs or belief-state tracking.

Additionally, the reward function used in this study is tailored to oversaturated traffic scenarios, focusing exclusively on throughput maximization. While appropriate under severe congestion, this design may not reflect broader operational efficiency under moderate or fluctuating conditions. To improve adaptability, composite or regime-aware reward functions that dynamically balance throughput, delay, and queue length—as proposed by Liang et al. [42] and Kim et al. [31]—could be incorporated.

Moreover, the current framework is tested on a single intersection under uniform geometry and flow. Although varying traffic demand levels and signal configurations were explored, broader validation is required across diverse intersection layouts, lane structures, and traffic patterns. Integrating multiple intersections and scaling the model to a network level would provide a more comprehensive evaluation of system performance despite the increased computational complexity [43], [44]. Both single- and multi-agent approaches are used to manage multiple intersections. Recent work also applies meta-RL to adapt to traffic fluctuations, pointing to the need for more practical and scalable traffic signal control models. Robustness to environmental variability such as incidents or adverse weather should also be considered, as highlighted by Sun et al. [45], Zeinaly et al. [46], and Alruwaili and Xie [47].

2) PRACTICAL DEPLOYMENT AND SOCIETAL CONSIDERATIONS

While the proposed framework demonstrates strong performance under idealized conditions, practical deployment in urban settings introduces substantial challenges. Reliable V2X communication and sensor infrastructure entail high installation and maintenance costs, which may not be feasible in many cities [48], [49]. Moreover, full CAV penetration is unlikely in the near future. To ensure robustness in mixed traffic environments, future systems must tolerate partial connectivity and communication delays. Recent studies offer promising directions using distributed, decentralized, or delay-resilient RL control [26], [27], [50], [51].

In parallel, societal and ethical aspects must also be addressed. Automated traffic systems risk exacerbating inequities if their benefits disproportionately favor certain groups. Recent work has highlighted the need for fairness-aware control algorithms [52], distributional equity analysis [53], and user trust through transparent and human-centered interaction [54]. To support responsible and inclusive deployment, future extensions of this work should incorporate multi-objective reward formulations and user-centered evaluation frameworks.

REFERENCES

- [1] B. Beak, K. L. Head, and Y. Feng, "Adaptive coordination based on connected vehicle technology," *Transp. Res. Rec.*, vol. 2619, no. 1, pp. 1–12, Jan. 2017, doi: [10.3141/2619-01](#).

- [2] W. Li and X. J. Ban, "Traffic signal timing optimization in connected vehicles environment," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Los Angeles, CA, USA, Jun. 2017, pp. 1330–1335, doi: [10.1109/IVS.2017.7995896](#).
- [3] X. Liu, G. Zhao, N. Masoud, and Q. Zhu, "Trajectory planning for connected and automated vehicles: Cruising, lane changing, and platooning," 2020, *arXiv:2001.08620*.
- [4] K. Yang, S. I. Guler, and M. Menendez, "Isolated intersection control for various levels of vehicle technology: Conventional, connected, and automated vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 72, pp. 109–129, Nov. 2016, doi: [10.1016/j.trc.2016.08.009](#).
- [5] Y. Guo, J. Ma, C. Xiong, X. Li, F. Zhou, and W. Hao, "Joint optimization of vehicle trajectories and intersection controllers with connected automated vehicles: Combined dynamic programming and shooting heuristic approach," *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 54–72, Jan. 2019, doi: [10.1016/j.trc.2018.11.010](#).
- [6] P. Li and X. Zhou, "Recasting and optimizing intersection automation as a connected-and-automated-vehicle (CAV) scheduling problem: A sequential branch-and-bound search approach in phase-time-traffic hyper-network," *Transp. Res. B, Methodol.*, vol. 105, pp. 479–506, Nov. 2017, doi: [10.1016/j.trb.2017.09.020](#).
- [7] M. Tajalli and A. Hajbabaie, "Traffic signal timing and trajectory optimization in a mixed autonomy traffic stream," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6525–6538, Jul. 2022, doi: [10.1109/TITS.2021.3058193](#).
- [8] W. Ma, J. Li, and C. Yu, "Shared-phase-dedicated-lane based intersection control with mixed traffic of human-driven vehicles and connected and automated vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 135, Feb. 2022, Art. no. 103509, doi: [10.1016/j.trc.2021.103509](#).
- [9] Y. Zou, F. Zheng, Z. Fan, and Y. Tang, "Integrated control of traffic signal and automated vehicles for mixed traffic: Platoon-based bi-level optimization approach," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2107–2113, doi: [10.1109/ITSC55140.2022.9922535](#).
- [10] S. E. Shladover, C. A. Desoer, J. K. Hedrick, M. Tomizuka, J. Walrand, W.-B. Zhang, D. H. McMahon, H. Peng, S. Sheikholeslam, and N. McKeown, "Automated vehicle control developments in the PATH program," *IEEE Trans. Veh. Technol.*, vol. 40, no. 1, pp. 114–130, Feb. 1991, doi: [10.1109/25.69979](#).
- [11] B. van Arem, C. J. G. van Driel, and R. Visser, "The impact of cooperative adaptive cruise control on traffic-flow characteristics," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 4, pp. 429–436, Dec. 2006, doi: [10.1109/TITS.2006.884615](#).
- [12] J. Li, C. Yu, Z. Shen, Z. Su, and W. Ma, "A survey on urban traffic control under mixed traffic environment with connected automated vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 154, Sep. 2023, Art. no. 104258.
- [13] K. Dresner and P. Stone, "A multiagent approach to autonomous intersection management," *J. Artif. Intell. Res.*, vol. 31, pp. 591–656, Mar. 2008.
- [14] J. Lee and B. Park, "Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 81–90, Mar. 2012, doi: [10.1109/TITS.2011.2178836](#).
- [15] R. Tachet, P. Santi, S. Sobolevsky, L. I. Reyes-Castro, E. Frazzoli, D. Helbing, and C. Ratti, "Revisiting street intersections using slot-based systems," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, Art. no. e0149607, doi: [10.1371/journal.pone.0149607](#).
- [16] K. Pandit, D. Ghosal, H. M. Zhang, and C.-N. Chuah, "Adaptive traffic signal control with vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1459–1471, May 2013, doi: [10.1109/TVT.2013.2241460](#).
- [17] Q. He, K. L. Head, and J. Ding, "PAMSCOD: Platoon-based arterial multi-modal signal control with online data," *Transp. Res. C, Emerg. Technol.*, vol. 20, no. 1, pp. 164–184, Feb. 2012, doi: [10.1016/j.trc.2011.05.007](#).
- [18] Z. Li, L. Eleftheriadou, and S. Ranka, "Signal control optimization for automated vehicles at isolated signalized intersections," *Transp. Res. C, Emerg. Technol.*, vol. 49, pp. 1–18, Dec. 2014, doi: [10.1016/j.trc.2014.10.001](#).
- [19] Y. Feng, C. Yu, and H. X. Liu, "Spatiotemporal intersection control in a connected and automated vehicle environment," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 364–383, Apr. 2018, doi: [10.1016/j.trc.2018.02.001](#).
- [20] B. Xu, X. J. Ban, Y. Bian, W. Li, J. Wang, S. E. Li, and K. Li, "Cooperative method of traffic signal optimization and speed control of connected vehicles at isolated intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1390–1403, Apr. 2019, doi: [10.1109/TITS.2018.2849029](#).

- [21] M. Liu, J. Zhao, S. P. Hoogendoorn, and M. Wang, "An optimal control approach of integrating traffic signals and cooperative vehicle trajectories at intersections," *Transportmetrica B, Transp. Dyn.*, vol. 10, no. 1, pp. 971–987, Dec. 2022, doi: [10.1080/21680566.2021.1991505](https://doi.org/10.1080/21680566.2021.1991505).
- [22] D. Rey and M. W. Levin, "Blue phase: Optimal network traffic control for legacy and autonomous vehicles," *Transp. Res. B, Methodol.*, vol. 130, pp. 105–129, Dec. 2019, doi: [10.1016/j.trb.2019.11.001](https://doi.org/10.1016/j.trb.2019.11.001).
- [23] M. Liu, J. Zhao, S. Hoogendoorn, and M. Wang, "A single-layer approach for joint optimization of traffic signals and cooperative vehicle trajectories at isolated intersections," *Transp. Res. C, Emerg. Technol.*, vol. 134, Jan. 2022, Art. no. 103459.
- [24] T. Chu, J. Wang, L. Codecá, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, Mar. 2020.
- [25] M. Miletić, E. Ivanjko, M. Gregurić, and K. Kušić, "A review of reinforcement learning applications in adaptive traffic signal control," *IET Intell. Transp. Syst.*, vol. 16, no. 10, pp. 1269–1285, Oct. 2022.
- [26] W. Li, T. Zhu, and Y. Feng, "A cooperative perception based adaptive signal control under early deployment of connected and automated vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 169, Dec. 2024, Art. no. 104860.
- [27] H. Shi, Y. Zhou, X. Wang, S. Fu, S. Gong, and B. Ran, "A deep reinforcement learning-based distributed connected automated vehicle control under communication failure," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 37, no. 15, pp. 2033–2051, Dec. 2022.
- [28] National Academies of Sciences, Engineering, and Medicine, *Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis*. Washington, DC, USA: The National Academies Press, 2022, doi: [10.17226/26432](https://doi.org/10.17226/26432).
- [29] M. Amirholly, M. Nourinejad, and H. O. Gao, "Optimal traffic control at smart intersections: Automated network fundamental diagram," *Transp. Res. B, Methodol.*, vol. 137, pp. 2–18, Jul. 2020, doi: [10.1016/j.trb.2019.10.001](https://doi.org/10.1016/j.trb.2019.10.001).
- [30] B.-K. Choi, "Adaptive signal control for oversaturated arterials," Ph.D. dissertation, Dept. Transp. Planning Eng., Polytech. Univ., Brooklyn, NY, USA, 1997.
- [31] G. Kim, J. Kang, and K. Sohn, "A meta-reinforcement learning algorithm for traffic signal control to automatically switch different reward functions according to the saturation level of traffic flows," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 38, no. 6, pp. 779–798, Apr. 2023, doi: [10.1111/mice.12924](https://doi.org/10.1111/mice.12924).
- [32] G. Abu-Lebdeh and R. F. Benekohal, "Development of traffic control and queue management procedures for oversaturated arterials," *Transp. Res. Rec.*, vol. 1603, no. 1, pp. 119–127, Jan. 1997, doi: [10.3141/1603-16](https://doi.org/10.3141/1603-16).
- [33] E. B. Lieberman, J. Chang, and E. Shenk Prassas, "Formulation of real-time control policy for oversaturated arterials," *Transp. Res. Rec.*, vol. 1727, no. 1, pp. 77–88, Jan. 2000, doi: [10.3141/1727-10](https://doi.org/10.3141/1727-10).
- [34] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.
- [35] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, Jan. 2018, pp. 1861–1870. [Online]. Available: <https://proceedings.mlr.press/v80/haarnoja18b.html>
- [36] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.
- [37] F. Mao, Z. Li, and L. Li, "A comparison of deep reinforcement learning models for isolated traffic signal control," *IEEE Intell. Transp. Syst. Mag.*, vol. 15, no. 1, pp. 160–180, Jan. 2023, doi: [10.1109/MITS.2022.3144797](https://doi.org/10.1109/MITS.2022.3144797).
- [38] *Transport and Maritime Affairs*, 2nd ed., Ministry Land, Korea Highway Capacity Manual (KHCM), Sejong, South Korea, 2013.
- [39] K. I. Ahmed, "Modeling drivers' acceleration and lane changing behavior," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1999.
- [40] H. Summala, "Brake reaction times and driver behavior analysis," *Transp. Human Factors*, vol. 2, no. 3, pp. 217–226, Sep. 2000.
- [41] M. Green, "How long does it take to stop? methodological analysis of driver perception-brake times," *Transp. Human Factors*, vol. 2, no. 3, pp. 195–216, Sep. 2000.
- [42] E. Liang, Z. Su, C. Fang, and R. Zhong, "OAM: An option-action reinforcement learning framework for universal multi-intersection control," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 4, pp. 4550–4558.
- [43] J. Jung, I. Kim, and J. Yoon, "EcoMRL: Deep reinforcement learning-based traffic signal control for urban air quality," *Int. J. Sustain. Transp.*, vol. 2024, pp. 1–10, Jan. 2024.
- [44] S. Jiang, X. Chen, X. Lin, and M. Li, "Intersection control based on fine-grained phases for connected and automated vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 163, Jun. 2024, Art. no. 104619.
- [45] Z. Sun, Z. Wang, X. Qi, D. Wang, Y. Li, and H. Lu, "A two-stage robust optimal traffic signal control with reversible lane for isolated intersections," *IET Intell. Transp. Syst.*, vol. 18, no. 4, pp. 723–742, Apr. 2024.
- [46] Z. Zeinaly, M. Sojoodi, and S. Bolouki, "A resilient intelligent traffic signal control scheme for accident scenario at intersections via deep reinforcement learning," *Sustainability*, vol. 15, no. 2, p. 1329, Jan. 2023.
- [47] A. Alruwaili and K. Xie, "Modeling the influence of connected vehicles on driving behaviors and safety outcomes in highway crash scenarios across varied weather conditions: A multigroup structural equation modeling analysis using a driving simulator experiment," *Accident Anal. Prevention*, vol. 199, May 2024, Art. no. 107514.
- [48] T. Nokes, B. Baxter, H. Scammell, D. Naberezhnykh, and L. Provvedi, "Cost analysis of V2I deployment," Ricardo Energy Environ., London, U.K., Tech. Rep., 2020, no. 5.
- [49] D. Gopalakrishna, P. Carlson, P. Sweatman, D. Raghunathan, L. Brown, and N. U. Serulle, "Impacts of automated vehicles on highway infrastructure," Federal Highway Admin., McLean, VA, USA, Tech. Rep. FHWA-HRT-21-015, 2021. [Online]. Available: <https://fhwa.dot.gov/publications/research/infrastructure/21015/21015.pdf>
- [50] Y. Han, M. Wang, and L. Leclercq, "Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation," *Commun. Transp. Res.*, vol. 3, Dec. 2023, Art. no. 100104.
- [51] H. Zhao, C. Dong, J. Cao, and Q. Chen, "A survey on deep reinforcement learning approaches for traffic signal control," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108100.
- [52] M. Kim, Y. Han, and Y. Kim, "Phase-free traffic signal control for balanced flow in sensor-limited environments," *Sci. Rep.*, vol. 15, no. 1, Feb. 2025, Art. no. 5834.
- [53] Z. Chen, A. L. Stuart, Y. Guo, Y. Zhang, and X. Li, "Distributional equity impacts of automated vehicles: A disaggregated approach," *Transp. Res. C, Emerg. Technol.*, vol. 167, Oct. 2024, Art. no. 104828.
- [54] L. Y. Koh and K. F. Yuen, "Public acceptance of autonomous vehicles: Examining the joint influence of perceived vehicle performance and intelligent in-vehicle interaction quality," *Transp. Res. A, Policy Pract.*, vol. 178, Dec. 2023, Art. no. 103864.



JIWON KANG received the M.S. degree in transportation engineering from Chung-Ang University. Her research interests include machine learning and traffic operation.



YONGJUN CHOI is currently pursuing the master's degree with the Department of Civil Engineering, Chung-Ang University. His research interests include machine learning and traffic operation.



KEEMIN SOHN was born in Seoul, South Korea, in 1968. He received the Ph.D. degree in transportation planning from the Department of Civil Engineering, Seoul National University (SNU), in 2003.

He is currently a Professor with the Department of Urban Engineering and the Department of Smart City, Chung-Ang University. His research interests include applications of artificial intelligence to transportation engineering and planning.

• • •