Traffic Signal Dynamic Optimization Algorithm Based on Deep Reinforcement Learning

Lijun Chang, Dan Wei
School of Transportation, Jilin University of Architecture and Technology, Changchun 130114, Jilin, China
Corresponding author: Lijun Chang (e-mail: clj_654321@163.com)

ABSTRACT With the acceleration of urbanization, traffic congestion at multiple intersections has become a core bottleneck restricting urban operational efficiency. To address this, this paper proposes a traffic signal dynamic optimization algorithm, the Cross-Attention Mechanism and Dueling Double DQN (CAM-D3QN). This method utilizes a novel crisscross attention module to dynamically model spatial dependencies between intersections and incorporates the Dueling Double DQN architecture for robust Q-value estimation. Validated on CityFlow using Grid-4×4 and Hangzhou-real networks, CAM-D3QN significantly outperforms the state-of-the-art baseline, GPLight, achieving relative improvements of approximately 10.5% in average vehicle delay, 11.3% in average queue length, 3.3% in throughput, alongside notable reductions in stops (12.1%) and fuel consumption (7.2%). Ablation experiments further demonstrate that removing the cross-attention module increases queue length by 30.6% in sudden congestion scenarios. The proposed method achieves superior performance to the baseline across four typical traffic scenarios on the Hangzhou-real road network, demonstrating its generalization capabilities. By leveraging the coordinated optimization of dynamic spatial perception and robust value assessment, this paper provides an effective solution for efficient and robust coordinated traffic signal control. The framework can be combined with traffic states acquired from radar, roadside sensors or wireless communication units in intelligent transportation systems.

INDEX TERMS traffic signal control, deep reinforcement learning, cross-attention mechanism, spatiotemporal coordination, wireless traffic sensing

I. INTRODUCTION

With the acceleration of urbanization, traffic congestion [1,2] has become a core bottleneck restricting the efficiency of urban operations, particularly hindering the time-sensitive processes of industries . Traditional traffic signal control systems, such as fixed timing or rule-based adaptive systems, are difficult to cope with the high dynamics, nonlinearity and spatiotemporal heterogeneity of traffic flow, especially in multi-intersection areas, which can easily cause congestion propagation and imbalance in resource allocation [3,4]. In recent years, deep reinforcement learning [5,6] has provided a new data-driven paradigm for signal control with its end-to-end perception-decision-making capabilities.

To improve regional coordination capabilities, recent research has focused on the design of dedicated DRL architectures. FRAP (It is a invariant tosymmetric operations like Flip and Rotation and considers AllPhase confgurations) [7] from "Learning Phase Competition for Traffic Signal Control" introduces "phase-lane" interaction embedding to explicitly model the internal competition relationship of the phase, significantly outperforming the basic DQN. MixLight [8] models the road network as a graph structure and uses a graph attention network to aggregate information about adjacent intersections to achieve spatial dependency perception. Additionally, the majority still use basic Double DQN or standard DQN, which are ineffective at separating action advantage from state value. As a result, policy robustness is impacted by the ongoing issue of Q-value overestimation [9,10]. In conclusion, the two main problems of robustly evaluating action value and accurately modeling dynamic spatial dependencies have not yet been addressed by current research.

This paper suggests a traffic signal dynamic optimization algorithm (CAM (Cross Attention Mechanism) - D3QN (Dueling Double Deep Q-Network)) that combines CAM and D3QN in order to overcome these limitations. Using the target intersection state as the query and the adjacent intersection states as the key and value, this paper first designs a cross-attention module. It overcomes the limitations of self-attention or static graph networks in expressing dynamic correlations by dynamically calculating cross-intersection attention weights, which accurately capture the spatiotemporal heterogeneity of traffic flow and dynamic correlations in the spatial dimension. This allows for differentiated modeling of complex spatial dependencies. Second, this paper presents the Dueling Double DQN architecture, which includes a Double Q-learning mechanism and explicitly decouples the state-value function and action-advantage function. In non-steady-state traffic environments, this effectively suppresses Q-value overestimation and enhances the strategy's evaluation stability and generalization ability. Furthermore, by combining prioritized experience replay with a multi-agent parameter sharing mechanism, efficient and scalable regional collaborative control is achieved on the CityFlow platform. The contribution of this paper lies in introducing CAM into multi-intersection signal control, enabling dynamic spatial awareness of the spatiotemporal heterogeneity of traffic systems; and systematically integrating D3QN to ensure the robustness of action value assessment. Together, these two methods form the core of the proposed dynamic optimization algorithm, effectively supporting adaptive signal timing optimization for complex traffic scenarios, which also translates to substantial operational efficiency gains for time-critical urban freight, particularly benefiting sectors like manufacturing by ensuring faster and more predictable transit of goods.

II. METHOD

A. PROBLEM MODELING

This paper formalizes the regional traffic signal control problem [11,12] as a partially observable Markov decision process (POMDP), as practical deployments often require limited observations (such as detector data or video recognition results) and traffic systems are highly dynamic and stochastic. This POMDP can be described by a five-tuple ⟨S,A,P,R,Ω⟩, where the state space S is not fully observable, and the agent can only obtain partial information through the observation space Ω. To simplify the modeling and adapt it to the deep reinforcement learning framework, a historical observation splicing or state reconstruction strategy is used to approximate the problem as a fully observable Markov decision process.

Specifically, consider a regional road network with N signalized intersections [13,14]. At each discrete decision time t, the system state s_t is the aggregate of the local traffic states of each intersection:

$$s_t = \{x_i^t\}_{i=1}^N, \quad x_i^t \in \mathbb{R}^L \tag{1}$$

In Formula 1, x_i^t represents the L-dimensional feature vector of the i-th intersection, which includes the following key traffic metrics: the number of vehicles (or detector counts) in each entrance lane; the average waiting time in each entrance lane (accumulated since the vehicle reached the stop line); the one-hot encoding of the currently active phase; the historical phase sequence, lane saturation, etc.

The state space contains the detailed traffic characteristics of each intersection, such as the number of vehicles in each entrance lane, the average waiting time of vehicles, the current active phase code, the historical phase sequence and lane saturation. The action space is defined as that each intersection independently selects actions from its preset set of feasible phases. The reward function is designed as a negative measure of the overall regional traffic efficiency, specifically the negative value of the weighted sum of the total queue length and the cumulative waiting time. The two weight coefficients are adjusted to balance their impact on agent decision-making.

The action space A is defined as the Cartesian product of predefined phases independently selected at each intersection:

$$a_t = (a_1^t, a_2^t, \ldots, a_N^t) \in A, \quad a_i^t \in \{1, 2, \ldots, K\} \tag{2}$$

In Equation 2, K represents the number of feasible phases at each intersection, and action a_i^t represents the phase number assigned to intersection i at time t.

The reward function $r_t = R(s_t, a_t)$ is designed as a negative measure of the overall traffic efficiency of the region to guide the agent to minimize congestion:

$$r_t = -\left(w_1 \cdot \sum_{i=1}^N Q_i^t + w_2 \cdot \sum_{i=1}^N W_i^t\right) \tag{3}$$

In Formula 3, Q_i^t is the total queue length at intersection i at time t, defined as the sum of the number of vehicles queuing in all entrance lanes; W_i^t is the cumulative waiting time at intersection i, which is the sum of the waiting times of all queued vehicles since their arrival; and $w_1, w_2 \geq 0$ is an adjustable weight coefficient used to optimize the balance between queue length and waiting time (referred to as $w_1 = w_2 = 0$ in this article).

Traditional GNN methods usually rely on predefined static graph structure to model the connection between intersections. This static modeling method is difficult to adapt to the highly dynamic, non local and heterogeneous spatio-temporal dependence in traffic flow. The cam module proposed in this paper takes the target intersection status as the query, and the adjacent intersection status as the key/value. By dynamically calculating the intersection attention weight, it realizes the real-time response and differential modeling to the change of traffic status. This mechanism can adaptively focus on the adjacent intersection that has the most impact on the target decision at the current time, so as to more accurately capture the non local dynamic dependence in the traffic flow. Especially in the scenario of sudden congestion or strong directional flow, cam can quickly identify the key congestion sources and achieve forward-looking coordinated control, while the static GNN is limited by the fixed topology, which is difficult to achieve such dynamic and goal-directed information aggregation. Therefore, cam is not only more flexible in structure, but also improves the essential modeling of the essential characteristics of dynamic traffic system in function.

Different from the fixed topology based models such as graph attention network (GAT), the spatial dependencies in the traffic network are dynamic, non local and highly heterogeneous in nature. They depend not only on the physical connection of the road network, but also on the real-time changing traffic flow state. Gat and its variants usually operate on a predefined static graph, and its attention calculation is limited by a fixed set of neighbors, which is difficult to effectively capture the dynamic association driven by real-time traffic conditions and across multiple intersections. The cam module proposed in this paper takes the real-time status of the target intersection as the query, and the status of other intersections in the network as the key and value. By calculating the cross attention weight, a goal oriented, data-driven dynamic dependency modeling is realized. This mechanism allows each intersection to adaptively focus on other intersections (regardless of their physical distance or preset topological relationship) that have the most influence on its decision-making according to the current global traffic situation, so as to be more consistent with the inherent characteristics of the dynamic traffic system in essence. Therefore, cam is not a simple replacement for gat, but a more expressive and adaptive architecture innovation for dynamic graph structure modeling.

B. CAM-D3QN ALGORITHM ARCHITECTURE

Improving overall control performance in regional traffic signal coordinated control requires robust action value assessment and efficient modeling of the dynamic spatial dependencies between several intersections. This paper proposes the CAM-D3QN algorithm, which combines CAM with the Dueling Double DQN [15,16] architecture. By using differentiated attention interactions between the target intersection and nearby intersections, this algorithm improves state representation capabilities. In order to suppress Q-value overestimation and achieve effective and reliable signal timing optimization in intricate road networks, it also makes use of the Dueling architecture and Double Q-learning mechanism. Figure 1 depicts the CAM-D3QN architecture.

The CAM-D3QN algorithm's overall network architecture is depicted in Figure 1. For effective perception and reliable decision-making at several intersections, its fundamental design depends on the natural fusion of CAM and Dueling Double DQN. Each intersection's raw state vector, which includes important traffic metrics like the number of cars in the lane, waiting time, and current phase, is sent to the input layer. After that, these state vectors are fed into a cross-attention module, where the key and value are the states of nearby intersections and the query is the state of the target intersection. Differentiated fusion of information from nearby intersections is made possible by dynamically calculating attention weights. In order to overcome the limitations of static graph convolution or traditional self-attention in representing dynamic relationships, this mechanism adaptively modifies the amount of attention given to each adjacent intersection based on the traffic conditions at the time. This allows the mechanism to accurately capture the spatial dependencies between intersections.

C. MULTI-AGENT COLLABORATIVE TRAINING MECHANISM

During the execution phase, each agent i computes its local Q-value function $Q_i(s_t, a_i; \theta)$ and independently selects an action $a_i^t = \arg\max_{a_i} Q_i(s_t, a_i; \theta)$ based solely on its local observation x_i^t and the adjacent intersection states $\{x_j^t\}_{j \in N(i)}$ ($N(i)$ represents the first-order neighboring set of intersection i) through a shared CAM-D3QN network. This design ensures scalability and low communication overhead for real-world deployments. During the training phase, the system uses the global state $s_t = \{x_i^t\}_{i=1}^N$ as input to jointly generate actions $a_t = (a_1^t, \ldots, a_N^t)$ for all agents. These actions are executed in the CityFlow simulation environment to obtain a global reward r_t and the next state s_{t+1}. Experience tuples (s_t, a_t, r_t, s_{t+1}) are stored in a shared experience replay pool with a capacity of $|D| = 10^5$.

To improve sample utilization efficiency and accelerate convergence, this paper adopts the Prioritized Experience Replay (PER) mechanism[17-18]. PER[19-20] prioritizes experiences based on the absolute value of the temporal difference error and samples them according to the priority. The priority of each experience $e_t = (s_t, a_t, r_t, s_{t+1})$ is defined as:

$$p_t = |\delta_t| + \epsilon \tag{4}$$

In Equation 4, ϵ is a small constant (here, $\epsilon = 10^{-6}$) to avoid zero priority. δ_t is the timing differential error:

$$\delta_t = r_t + \gamma Q_{\text{target}}\left(s_{t+1}, \arg\max_{a'} Q(s_{t+1}, a'; \theta); \theta^-\right) - Q(s_t, a_t; \theta) \tag{5}$$

In Formula 5, θ is the primary network parameter, θ^- is the target network parameter, and γ is the discount factor.

The empirical sampling probability is:

$$P(e_t) = \frac{p_t^\alpha}{\sum_{e_k \in D} p_k^\alpha} \tag{6}$$

In formula 6, α controls the priority strength ($\alpha = 0.6$ in this paper). To eliminate the bias introduced by PER, an importance sampling weight is applied to the loss function during training:

$$w_t = \left(\frac{1}{|D| \cdot P(e_t)}\right)^\beta \tag{7}$$

In Equation 7, $\beta \in [0, 1]$ is linearly increased from an initial value of 0.4 to 1.0 during training.

Network parameters are updated by minimizing the weighted temporal difference loss:

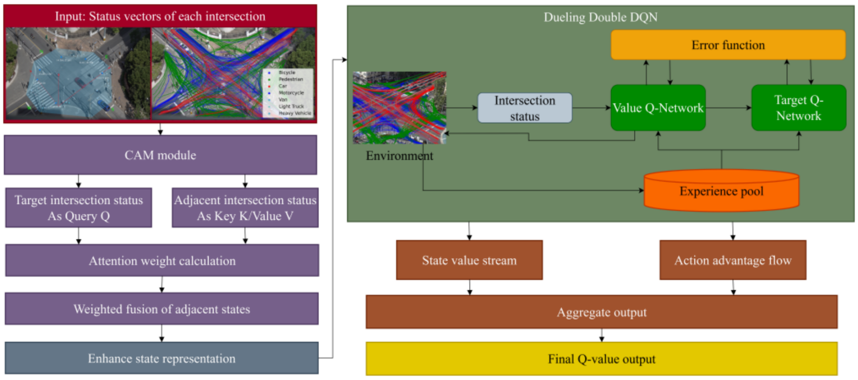

FIGURE 1. CAM-D3QN Architecture

$$L(\theta) = \mathbb{E}_{e_t \sim D}\left[w_t \cdot \delta_t^2\right] \tag{8}$$

This multi-agent collaborative training mechanism maintains decentralized execution while effectively balancing collaboration, sample efficiency, and scalability through parameter sharing, a global experience pool, and the PER mechanism. This ensures stable training of CAM-D3QN in complex regional road networks.

D. SIGNAL CONTROL DECISION PROCESS

The proposed CAM-D3QN algorithm achieves automatic optimization of regional traffic signal timing strategies through end-to-end interactive learning. Given a road network topology $G = (V, E)$ (where V represents the intersection set, $|V| = N$ and E represents the connection relationship), historical or simulated traffic flow data, and a set of hyperparameters (including the learning rate η, discount factor γ, target network soft update coefficient τ, PER parameters α and β), the algorithm outputs a trained and converged policy network with parameters θ that can map traffic states to optimal phased actions in real time.

The optimization process follows the standard deep Q-learning paradigm and consists of five core stages:

(1) Initialization: The main network parameters θ and the target network parameters θ^- are randomly initialized, and the experience replay pool is empty;

(2) Interactive sampling: The agent interacts with the CityFlow simulation environment, executes the ϵ-greedy strategy, collects experience tuples and stores them in the experience replay pool;

(3) CAM encoding [21-22]: For each state s_t, an enhanced state representation $\tilde{s}_t$ is generated through the cross-attention module, explicitly modeling the dynamic dependencies between intersections;

(4) D3QN update: Batch data is sampled from the experience replay pool according to the PER mechanism, the weighted loss is calculated, and θ is updated through the Adam optimizer;

(5) Target network soft update: The target network parameters are smoothly synchronized in a low-frequency manner:

$$\theta^- \leftarrow \tau\theta + (1-\tau)\theta^- \tag{9}$$

In Equation 9, $\tau \ll 1$ (here, this paper uses $\tau = 0.005$) ensures training stability.

The pseudocode for the CAM-D3QN algorithm is as follows:

Input: Road network topology G, traffic flow data, hyperparameters $\eta, \gamma, \tau, \alpha, \beta, \epsilon$
Output: Trained policy network parameters θ
Initialize main network parameters θ
Initialize target network parameters $\theta^- \leftarrow \theta$
Initialize replay buffer $D \leftarrow \emptyset$
Set $\beta \leftarrow \beta_{\text{init}}$
for episode = 1 to M do
Reset environment, obtain initial state s_0
$\epsilon \leftarrow \epsilon_{\text{init}}$
for $t = 0$ to T do
$\tilde{s}_t \leftarrow \text{CAM}(s_t)$
With probability ϵ: $a_t \sim \text{Uniform}(A)$
Else: $a_t \leftarrow \arg\max_a Q(\tilde{s}_t, a; \theta)$
Execute a_t, observe r_t, s_{t+1}
Store transition (s_t, a_t, r_t, s_{t+1}) in D
if $t \bmod C_{\text{train}} = 0$ and $|D| > B$ then
Sample batch $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^{B}$ from D via PER
Compute TD errors δ_i and importance weights w_i
Update θ by minimizing $L(\theta) = \frac{1}{B}\sum_{i=1}^{B} w_i\delta_i^2$
Update $\beta \leftarrow \min(1.0, \beta + \Delta\beta)$
end if
if $t \bmod C_{\text{target}} = 0$ then
$\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$
end if
$\epsilon \leftarrow \epsilon \cdot \rho$
end for
end for
return θ

III. EXPERIMENTAL SETUP

A. SIMULATION PLATFORM AND ROAD NETWORK CONFIGURATION

This paper uses CityFlow as the traffic simulation platform [23,24]. CityFlow [25] is an open-source, high-fidelity, microscopic traffic simulator that supports large-scale road networks. It has parallel computing capabilities, a flexible signal control interface (supporting cycle-by-cycle action delivery), and seamless integration with deep reinforcement learning frameworks (PyTorch, TensorFlow). It has been widely used in cutting-edge research in the field of intelligent transportation.

Experiments were conducted on two representative road networks:

(1) Grid-4×4 is a 4×4 uniform grid network consisting of 16 standard four-phase intersections with a uniform section length of 200 meters, representing an idealized urban core area structure;

(2) Hangzhou-real is a 25-intersection sub-area road network constructed based on real traffic data from Hangzhou, China, which includes non-uniform intersection spacing, heterogeneous intersection types (such as T-shaped and Y-shaped), and irregular topological connections.

Traffic flow scenarios cover four typical modes: morning peak (high inflow, strong directionality), evening peak (high outflow, concentrated return), flat peak (medium flow, balanced distribution), and sudden congestion (local road accidents causing sudden changes in flow), to comprehensively evaluate the robustness and adaptability of the algorithm under different dynamic intensities.

The road network types are shown in Figure 2.

The two road networks chosen for this paper depict complicated real-world situations and idealized ones, respectively. Every intersection in the Grid-4×4 network structure has the same phase configuration and road segment lengths, making it highly symmetrical and evenly connected. This makes it appropriate for confirming the convergence properties and coordination upper bound of the algorithm in a controlled setting. However, there is a lot of topological heterogeneity in the Hangzhou-real road network, which was developed using data from real cities. Strong directional and spatiotemporal traffic flow, an uneven connectivity distribution, a large number of non-four-way intersections, and uneven intersection spacing are all examples of this.

B. COMPARATIVE ALGORITHMS

To comprehensively evaluate the performance of proposed algorithm, this paper compared it with existing mainstream algorithms. The comparative algorithm information is shown in Table 1.

TABLE 1. Comparative Algorithm Information

Algorithm	Category	Features and Status
Fixed-Time	Traditional baseline	The most basic comparison baseline
MaxPressure	Traditional adaptive	A strong model-based benchmark with high efficiency
Rainbow DQN	General purpose DRL	A comprehensive DQN architecture with strong performance
PPO	General purpose DRL	A universal policy gradient algorithm with stability and reliability
FRAP	Specialized DRL	Designing a network architecture for TSC (Traffic Signal Control), considering phase competition
CoLight	Specialized DRL	Using GNN (Graph Neural Network) to model coordination between intersections
MPLight	Specialized DRL	Introducing the concept of pressure, emphasizing fairness and interpretability
GPLight	Specialized DRL	Considering TSC as a graph sorting problem, handling large-scale networks

C. HYPERPARAMETER SETTINGS

To ensure experimental fairness and reproducibility, this paper strictly standardizes the training and network hyperparameters of all compared algorithms. The core configuration of CAM-D3QN is fixed in the ablation experiments. Key hyperparameters were determined through grid search and validation set tuning. The specific settings are shown in Table 2.

TABLE 2. Hyperparameter Configuration

Hyperparameter	Value	Instructions
Learning rate	0.001	Adam optimizer initial learning rate
Discount factor	0.95	Future reward attenuation coefficient
Replay buffer size	100,000	Maximum capacity of shared experience pool
Batch size	32	Sample size for each training session
Target update frequency	1,000 steps	Soft update interval (hard update)
Target network soft update coefficient	0.005	Used for $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$
ϵ-Greedy initial value	1	Initial exploration rate
ϵ-Greedy decay rate	0.999	Multiplicative decay per step: $\epsilon \leftarrow \epsilon \cdot \rho$
PER priority index (α)	0.6	Control priority intensity
Initial value of PER importance sampling index (β)	0.4	Importance weight compensation initial value
CAM embedding dimension	64	Query/Key/Value vector dimension
Dueling network hidden layer	[128, 64]	Fully connected layer structure

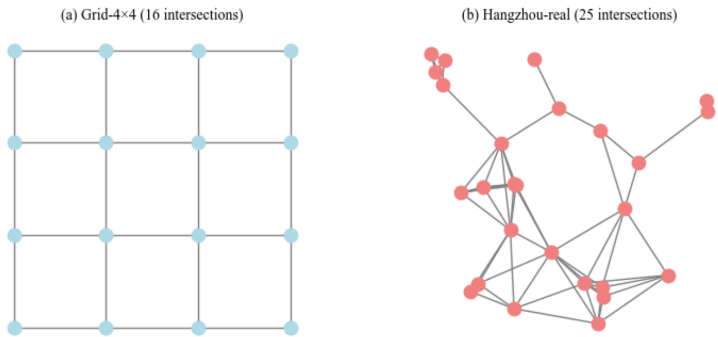

FIGURE 2. Road Network Types. Figure 2(a) Grid-4×4; Figure 2(b) Hangzhou-Real

D. EVALUATION METRICS

To evaluate the performance of the CAM-D3QN algorithm in regional traffic signal control, this paper uses five widely recognized traffic efficiency and environmental impact metrics. All metrics are calculated based on vehicle trajectory data output by CityFlow at the end of the simulation cycle.

Average Vehicle Delay (AVD) is defined as the average difference between the actual travel time and the free-flow travel time of all vehicles passing through the road network within the intersection area, reflecting the impact of signal control on traffic efficiency:

$$\text{AVD} = \frac{1}{M}\sum_{v=1}^{M}\left(t_v^{\text{actual}} - t_v^{\text{free}}\right) \tag{10}$$

In Equation 10, M is the total number of vehicles that successfully exit the network, t_v^{actual} is the total time between vehicle v entering the network and leaving, and t_v^{free} is its free flow time.

The average queue length (AQL) is defined as the spatiotemporal average of the maximum queue lengths across all entry lanes during the simulation period and is used to measure the severity of local congestion:

$$\text{AQL} = \frac{1}{T \cdot D}\sum_{t=1}^{T}\sum_{d=1}^{D} q_{d,t} \tag{11}$$

In Equation 11, T is the total number of simulation steps, D is the total number of controlled entrance lanes in the network, and $q_{d,t}$ is the number of vehicles queued in lane d at time t (according to CityFlow standards: speed < 0.1 m/s and upstream of the stop line).

Throughput is defined as the total number of vehicles that successfully exit the network per unit simulation time, reflecting the overall service capacity of the network:

$$\text{Throughput} = \frac{M}{T_{\text{sim}}} \tag{12}$$

In Equation 12, T_{sim} is the total simulation duration.

Stop Count is defined as the total number of times all vehicles within the road network decrease from a speed above a threshold to a speed below a threshold. It represents driving smoothness and comfort:

$$\text{StopCount} = \sum_{v=1}^{M}\sum_{t=1}^{T-1}\mathbb{I}\left(v_{v,t} > v_{\text{th}} \wedge v_{v,t+1} \le v_{\text{th}}\right) \tag{13}$$

In Equation 13, $v_{v,t}$ is the speed of vehicle v at time t, and $v_{\text{th}} = 0.1$ m/s is the stop threshold.

Fuel Consumption: Based on the widely adopted Comprehensive Modal Emissions Model (CMEM), fuel consumption is estimated (in milliliters) using the vehicle's instantaneous speed and acceleration:

$$F_v = \sum_{t=1}^{T_v}\left(a_0 + a_1 v_{v,t} + a_2 v_{v,t}^2 + a_3 a_{v,t} v_{v,t}\right)\Delta t \tag{14}$$

In Equation 14, T_v is the total number of steps taken by vehicle v in the road network, $a_{v,t}$ is its acceleration, $\Delta t = 1$ s is the simulation step length, and the coefficients $a_0 = 2.275 \times 10^{-2}$, $a_1 = 1.238 \times 10^{-3}$, $a_2 = 5.495 \times 10^{-5}$, $a_3 = 6.389 \times 10^{-4}$ are standard parameters for gasoline vehicles. Total fuel consumption is the sum of all vehicles.

IV. RESULTS AND ANALYSIS

A. PERFORMANCE COMPARISON

Table 3 shows the average performance of each algorithm under four traffic scenarios (morning rush hour, evening rush hour, off-peak, and sudden congestion) in a Grid-4×4 road network (Data are the averages of the four scenarios).

TABLE 3. Comprehensive Performance Comparison of Algorithms in a Grid-4×4 Road Network

Algorithm	AVD (s)	AQL (veh)	Throughput (veh/h)	Stop Count ($\times 10^3$)	Fuel Consumption (L)
Fixed-Time	48.7	12.3	2,840	9.8	42.6
MaxPressure	41.2	10.1	3,010	8.5	38.9
Rainbow DQN	38.5	9.4	3,120	7.9	36.7
PPO	36.8	8.9	3,180	7.4	35.2
FRAP	34.1	8.2	3,250	6.8	33.5
CoLight	32.6	7.6	3,310	6.3	32.1
MPLight	31.9	7.4	3,340	6.1	31.6
GPLight	30.5	7.1	3,380	5.8	30.4
GAT-DRL	29.8	6.9	3,405	5.6	29.5
CAM-D3QN	27.3	6.3	3,490	5.1	28.2

Note: All values are averages of 10 random seed runs across four traffic scenarios (morning rush hour, evening rush hour, off-peak, and sudden congestion).

In every metric, the suggested CAM-D3QN performs better than the comparison techniques. It achieves a 10.5% reduction in AVD, an 11.3% reduction in AQL, a 3.3% increase in throughput, and a 12.1% and 7.2% reduction in stops and fuel consumption, respectively, when compared to the next-best method, GPLight.

This benefit results from two main mechanisms: In order to precisely identify the main causes of congestion and plan phases ahead of time, the cross-attention module first dynamically models the state dependencies between the target intersection and nearby intersections. This effectively reduces queue spread and, consequently, AQL and AVD. Second, strategy evaluation is strengthened by the Dueling Double DQN architecture, which suppresses Q-value over-estimation and decouples state value from action advantage. To evaluate the learning efficiency and policy stability of each algorithm, Figure 3 shows the training convergence curves on a Grid-4×4 road network.

In addition to reaching a stable plateau rather quickly (about 125 epochs), CAM-D3QN outperforms current techniques with a final average cumulative reward of -45. This benefit results from its two novel mechanisms: First, the cross-attention module greatly enhances the semantic relevance and collaborative perception of state representations, which speeds up policy learning. It does this by allowing each intersection to adaptively focus on the adjacent intersection states that are most important for the current decision through goal-oriented dynamic information aggregation. By suppressing value over-estimation through Double Q-learning and successfully decoupling state value and action advantage, the Dueling Double DQN architecture guarantees action selection reliability and prevents training oscillations. In contrast, although FRAP introduces phase competition priors, it lacks modeling of interactions between intersections.

B. ABLATION EXPERIMENTS

To verify the effectiveness of the core components of CAM-D3QN, this paper conducted ablation experiments on a Grid-4×4 road network, removing the cross-attention module (w/o CAM), the Dueling structure (without Dueling), and the Double Q-learning mechanism (w/o Double). Using average queue length (AQL) as the evaluation metric, the contribution of each module to the coordinated control performance was quantitatively analyzed under four typical traffic scenarios: morning rush hour, evening rush hour, off-peak, and sudden congestion.

The ablation results are shown in Table 4.

TABLE 4. Ablation Results (AQL (veh))

Algorithm	Morning rush hour	Evening rush hour	Off-peak hours	Sudden congestion
Full (CAM-D3QN)	6.1	5.9	4.3	7.2
w/o CAM	7.8	7.5	5.6	9.4
w/o Dueling	6.9	6.7	5.0	8.3
w/o Double	7.2	7.0	5.2	8.6

Note: Data are averages of 10 random seed runs; lower is better.

The core value of CAM in dynamic traffic conditions is highlighted by ablation experiments that demonstrate that removing the crisscross attention module (without (w/o) CAM) significantly improves the AQL in all scenarios, especially in the sudden congestion scenario, where the AQL increases from 7.2 veh to 9.4 veh (+30.6%). To accomplish differentiated, goal-oriented modeling of spatial dependencies, CAM uses a target intersection as a query and nearby intersections as keys or values. On arterial roads, congestion upstream can swiftly spread downstream in highly directional situations, like the morning rush hour. In order to reduce traffic congestion, CAM can automatically give upstream intersections higher attention weights and extend green lights ahead of time. The AQL rose in each of the four scenarios following the removal of the Dueling structure (w/o Dueling), indicating the structure's indispensable contribution to boosting policy robustness. The Dueling network can more precisely discern between the relative merits of particular actions and the overall quality of the current state by explicitly separating the state value function from the action advantage function.

The AQL performance deteriorates marginally more than w/o Dueling after the Double Q-learning mechanism is removed (w/o Double). The four scenarios' respective AQL performances are 7.2 veh, 7.0 veh, 5.2 veh, and 8.6 veh, suggesting that the main bottleneck compromising policy stability is Q-value overestimation. Because of its maximization operation, the standard DQN introduces a systematic positive bias that causes overestimation of action values and a propensity for the agent to select phases that seem to have high rewards but are actually suboptimal. This bias might only lead to a slight decrease in efficiency in situations with low to medium traffic volumes, like off-peak traffic. However, overestimation can intensify policy oscillations in high-variance environments, like morning rush hour or sudden congestion. For instance, mistaking short queues for long ones can result in frequent phase switching, making vehicle starts and stops more difficult, and intensifying queue accumulation. By separating action selection from value assessment, double Q-learning successfully reduces this bias. In situations of sudden congestion, the AQL without Double Q-learning was 8.6 veh, which was higher than the full model's 7.2 veh, indicating its susceptibility

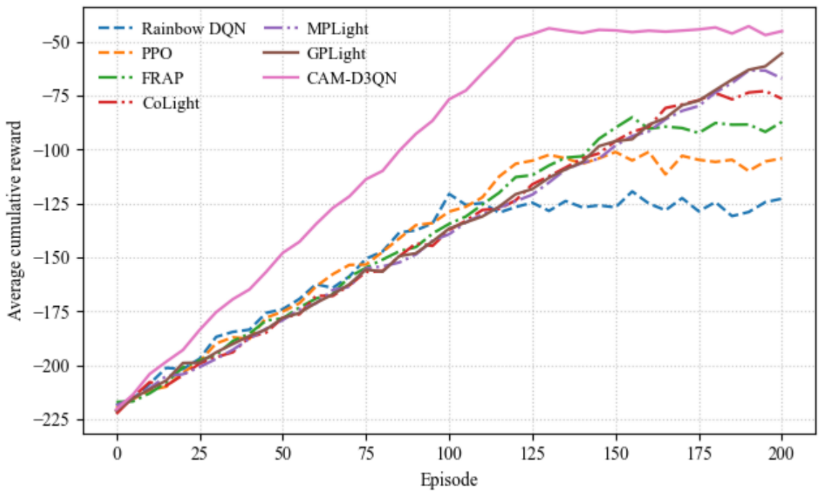

FIGURE 3. Training Convergence of Different Algorithms

to uncertainty. Therefore, the foundation for guaranteeing the dependable implementation of DRL strategies in actual traffic is a strong value assessment mechanism.

C. ATTENTION WEIGHT VISUALIZATION

Figure 4 displays a heatmap of the attention weights applied to eight neighboring intersections (including diagonal intersections) for a central intersection (Intersection 9) in a Grid-4×4 network to confirm that CAM can dynamically capture the inter-intersection traffic state dependencies. The model's dynamic attention to important neighbors is reflected in the weights, which are determined in real time during morning rush hour peak queuing conditions.

Figure 4's heatmap demonstrates how CAM can dynamically distribute attention resources according to traffic conditions in real time. The weight distribution also closely complies with the physical laws of traffic flow. The city's main roads show a strong directional commuter flow (from north to south) during the morning rush hour. Due to the constant flow of traffic, the upstream intersection (5) north of the target intersection (9) is currently extremely crowded. Compared to its neighbors (like the east-side intersection 10, which has an attention weight of 0.15), it has a higher attention weight (0.23). By releasing green light time ahead of time to alleviate lines and stop congestion from spreading to the target intersection, CAM-D3QN is able to prioritize the phase coordination of upstream high-load intersections thanks to this differentiated attention. Notably, because they are off the main road, the diagonal neighbors (intersections 4 and 6) have much lower weights, suggesting that CAM captures functional traffic associations in addition to topological distance. Furthermore, the model's comprehension of the causal logic that "downstream status has limited impact on current decisions" is demonstrated by the actively suppressed (0.11) weight of the downstream intersection (13) as well. This data-driven dynamic perception mechanism far exceeds the expressive power of static graph convolution or self-attention and is the key to achieving precise coordinated control.

CAM's attention weights also make the model interpretable, providing an intuitive basis for traffic engineers to understand and validate strategies. The high-weighted areas in Figure 4 closely correspond to actual congestion sources, demonstrating that the model's decisions are not black boxes but are based on reasonable traffic state dependencies. For example, in a sudden congestion scenario, if an accident occurs on a side branch, CAM can quickly focus attention on the accident point and its upstream intersection, coordinating signals to restrict traffic flow while ensuring access to the main road. More importantly, the

CAM weights can serve as a diagnostic tool: if a particular control performance is poor, the attention distribution can be retrospectively analyzed to determine whether it was due to neglect of critical neighbors, thereby guiding model improvements or manual intervention.

D. Q-VALUE ESTIMATION BIAS

Figure 5 compares the deviation distributions between Q-value estimates and Monte Carlo true returns during training for the standard DQN and D3QN, demonstrating the mechanistic benefits of the D3QN in suppressing overestimation and providing a quantitative assessment of the effect of Q-value estimation bias on policy performance.

The linear fit slope (1.28) is greater than 1, indicating a systematic positive bias, and the Q-value estimates (red dots) of the standard DQN are typically above the ideal line (Q=G), as seen in Figure 5(a). A persistent overestimation of action values in high-variance states is the result of the maximization operation and value assessment in DQN using the same network. This bias can lead to overly sensitive policies in high-noise situations, like morning rush hour or sudden congestion, by making the agent misjudge some suboptimal phases as having high long-term rewards in traffic signal control. For instance, cutting off a green light too

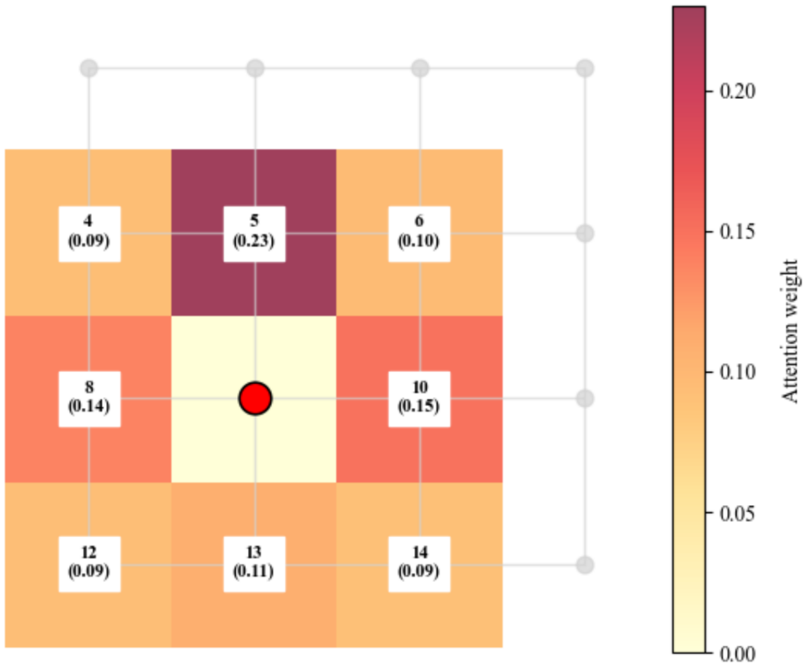

FIGURE 4. Cross-Attention Weight Heatmap

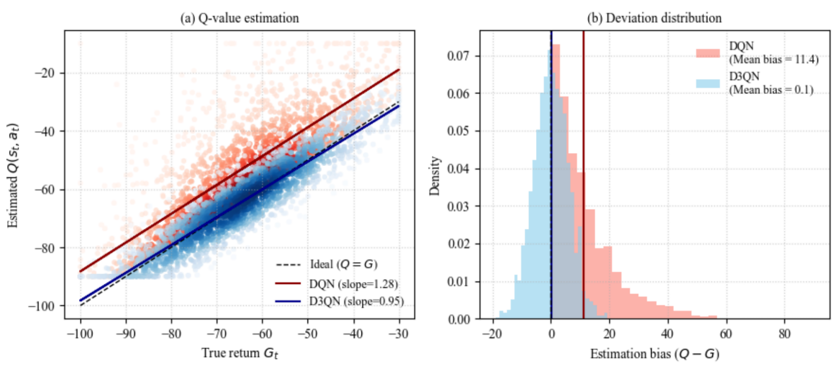

FIGURE 5. Q-Value Estimation Bias Analysis. Figure 5(a) Q-Value Estimation; Figure 5(b) Bias Distribution; Note: G represents the actual reward

soon because there isn't any traffic in a particular direction for a moment can make lines worse later. The DQN bias is further confirmed to be right-skewed by the bias distribution in Figure 5(b), which shows that extreme overestimations are common. In addition to impairing policy performance, this erratic evaluation process can cause training oscillations, which significantly reduces the dependability of real deployments.

D3QN's Q-value estimations (blue dots) exhibit nearly no systematic bias and are widely dispersed around the ideal line, with a fitted slope of 1 (0.95). Its dual debiasing process is the source of this benefit: The primary network chooses actions, and the target network assesses their worth. This is the first way that Double Q-learning separates action selection from value appraisal. This essentially suppresses overestimation by severing the link between the evaluation network and the maximization operation. Second, by separating state value from action advantage, the Dueling architecture helps the network eliminate value misjudgments brought on by slight variations in the advantage function in low-variance conditions. This allows the network to more precisely simulate the overall quality of the state and the relative merits of actions. This implies that D3QN can evaluate the long-term advantages of prolonging the current phase as opposed to switching phases in traffic control more realistically. This can enhance traffic continuity and avoid needless phase switching, particularly in off-peak or smooth traffic situations. The high reliability of its assessment is demonstrated by the deviation distribution, which is roughly symmetrical and concentrated close to zero. The CAM-D3QN strategy's primary assurance of sustaining high performance and stability in intricate traffic situations is its strong value function learning capability.

E. ROBUSTNESS UNDER DIFFERENT TRAFFIC INTENSITIES

Figure 6 shows the average AVD performance of each method when the traffic intensity changes from 500 to 3000 veh/h.

Fundamental variations and inherent robustness in the algorithms' responses to congestion pressures are revealed through analysis of the average vehicle delays of different algorithms at varying traffic intensities. Deep reinforcement learning-based techniques typically perform better than conventional techniques in the low-to-medium-volume range (500–1800 veh/h), highlighting the benefits of DRL in utilizing the network's potential and adjusting to changing traffic flows. The nonlinear and dynamically coupled nature of traffic flow, however, dramatically increases as traffic reaches high loads (>2000 veh/h) and even oversaturation (>2500 veh/h), and the performance difference between algorithms continues to be substantial.

Because of their entirely static strategies, traditional fixed-time methods cannot adjust to any dynamic changes. MaxPressure's decision-making approach, which is based on instantaneous queue pressure, is constrained even though it can react locally dynamically. Its performance quickly deteriorates in high traffic situations because it is unable to anticipate and stop the chain propagation of congestion within the road network. Although some learning capabilities are present in general DRL algorithms like Rainbow DQN and PPO, their network structures are not tailored to the spatial dependencies of traffic networks, which leads to inadequate state representation capabilities and limits the quality of their decision-making in highly complex scenarios.

Dedicated DRL architectures show more promise. By simulating phase competition, FRAP increases efficiency at a single intersection but falls short in terms of inter-intersection coordination. Better performance is achieved by CoLight, MPLight, and GPLight by explicitly taking spatial structure into account through the use of graph neural networks, pressure sensing, or graph sorting mechanisms.

In all traffic ranges, but especially in oversaturated situations (>2500 veh/h), the CAM-D3QN suggested in this paper maintains the lowest AVD. Its core architecture makes two contributions that contribute to this exceptional performance: CAM employs the target intersection state as a query and dynamically determines its association weights with neighboring intersections. This reduces the spread of lines and improves efficiency through spatial coordination by enabling proactive coordination and precise identification of the main causes of congestion. The Dueling Double DQN architecture successfully lowers Q-value overestimation by separating state value and action advantage and putting in place a Double Q-learning mechanism. This ensures reliable decision-making in high-dimensional, nonlinear state spaces from the perspective of robust value assessment and policy evaluation. To better understand and manage the global dynamics of complex road networks, CAM-D3QN must integrate these two approaches to achieve sustained, stable, and superior control performance across the entire traffic flow range, particularly in high-pressure scenarios.

F. GENERALIZATION CAPABILITY IN REAL-WORLD ROAD NETWORKS

Figure 7 shows a comprehensive performance comparison of various methods in four typical traffic scenarios on the Hangzhou real-world road network (25 intersections, irregular topology).

With a highly directional commuter flow during the morning rush hour, CAM-D3QN has the best average traffic flow (AVD) at 31.6 seconds, which is 8.1% slower than GPLight. In order to guarantee effective commuter corridors and stop lines from spilling onto branch roads, its CAM precisely locates the sources of upstream congestion on arterial roads and dynamically coordinates signals. Conventional techniques like MaxPressure are less predictive and often react to local pressure, which causes excessive phase switching and worsens delays.

Return traffic is more scattered during the evening peak, which increases the need for coordinated control. Due to D3QN's reliable evaluation of the value of multidirectional traffic, CAM-D3QN maintains optimal performance (AVD reaches 32.8s) without temporarily overestimating the efficiency of a single direction. The dynamic shift in traffic center of gravity during the evening peak is difficult for methods like CoLight, which rely on static adjacency relationships, to adjust to, leading to comparatively poor performance. The difference between algorithms gets smaller during balanced and low-peak traffic, but CAM-D3QN is still in the lead with an AQL of 5.3 vehicles. The system becomes less sensitive to phase changes during this time, and the low deviation advantage of D3QN results in a smoother distribution of green light, which minimizes needless starts and stops. Exploration noise or overestimation still causes slightly suboptimal switching in algorithms like FRAP, PPO, and others.

With an AVD of 50.4s, an AQL of 12.1 veh, a throughput of 2970 veh/h, a stop count of 7.6×103, and a fuel consumption of 38.2L, CAM-D3QN showed excellent resilience in the face of abrupt congestion brought on by a localized accident. Its CAM module focused on the intersections around the accident scene and quickly coordinated the implementation of a "interception-relief" strategy. D3QN, however, ensured that overestimation would not lead to poor recovery timing judgment under high uncertainty, thereby preventing the spread of congestion and demonstrating its utility in emergency situations.

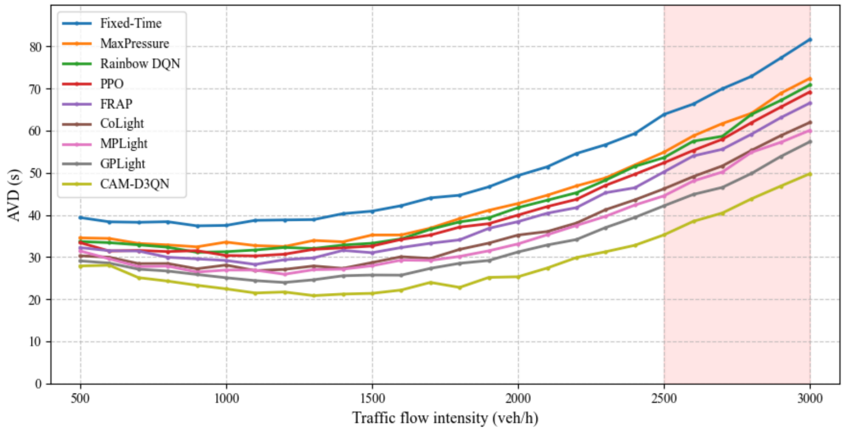

FIGURE 6. Robustness at Different Traffic Intensities

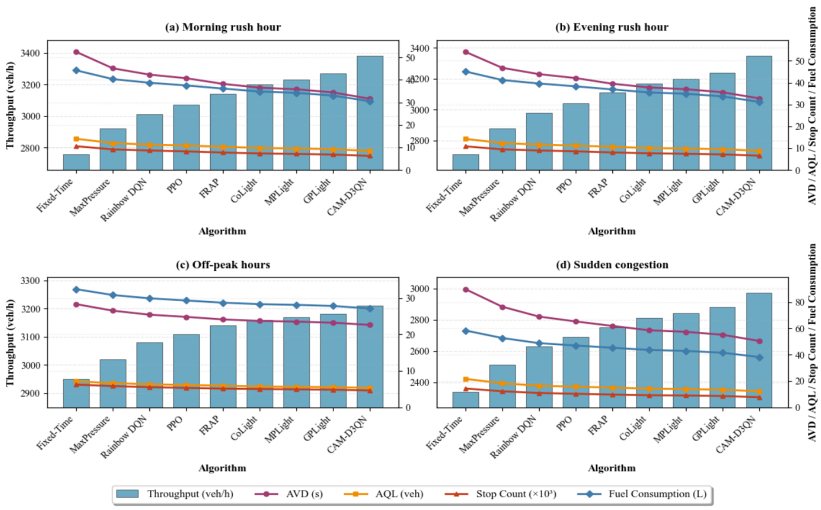

FIGURE 7. Performance Comparison on a Real-World Hangzhou Road Network. Figure 7(a) Morning Rush Hour; Figure 7(b) Evening Rush Hour; Figure 7(c) Off-Peak Hours; Figure 7(d) Sudden Congestion

The experimental results show that cam-d3qn has excellent and stable performance in four typical preset traffic scenarios on Hangzhou real road network, which proves that the method has good adaptability and robustness in the face of known scenarios with different flow characteristics and direction patterns. Although these typical scenarios can represent many real situations, to comprehensively evaluate the generalization ability of the algorithm for completely unknown, volatile or highly random traffic demand patterns, it is still necessary to introduce a broader and more random test set for verification in future research. The results of this study provide strong evidence for the effective application of this method in structured traffic mode.

V. CONCLUSION

This paper addresses the two primary problems of over-optimal Q-value estimation and inadequate spatial dependency modeling in coordinated traffic signal control at multiple intersections by presenting the CAM-D3QN algorithm, a technological advancement critical for improving urban freight movement efficiency, thereby directly supporting the high-volume, time-critical material flow required by the manufacturing sector. This method creatively integrates CAM to dynamically aggregate data from nearby intersections using the target intersection state as a query. This allows to accurately identify and proactively coordinate the primary sources of congestion by overcoming the limitations of static graph convolution or traditional self-attention in expressing dynamic associations. By integrating the Dueling Double DQN architecture, the Q-value function is divided into a state value stream and an action advantage stream. When paired with the Double Q-learning mechanism, this effectively suppresses systematic positive biases in value function estimation, preventing performance fluctuations from suboptimal phase switching and guaranteeing the robustness and reliability of policy evaluation in high-dimensional, nonlinear state spaces. In order to accomplish goal-oriented differentiated perception of neighborhood states, this study suggests a dynamic spatial dependency modeling approach combined with CAM. By methodically integrating the Dueling Double DQN architecture, the approach enhances policy stability and action value assessment accuracy. The approach provides a data-driven solution for regional traffic signal coordinated control that combines high performance, strong robustness, and good generalization capabilities. Experiments on a Grid-4×4 network and a real-world Hangzhou road network show that the method outperforms current mainstream algorithms on multiple traffic efficiency and environmental metrics.

AUTHOR CONTRIBUTIONS

Conceptualization – Lijun Chang and Dan Wei; methodology – Lijun Chang and Dan Wei; investigation – Lijun Chang and Dan Wei; writing-original draft preparation – Lijun Chang and Dan Wei. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

Science and Technology Research Project of Jilin Provincial Education Department: Research on Key Technologies of Intersection Signal Timing Based on the "Prediction-Optimization-Evaluation" Closed Loop (JJKH20261688KJ).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] D. R. Aleko and S. Djahel, “An efficient adaptive traffic light control system for urban road traffic congestion re-duction in smart cities,” Information, vol. 11, no. 2, pp. 119-119, 2020, doi: 10.3390/info11020119.
[2] L. F. P. De Oliveira, L. T. Manera, and P. D. G. Da Luz, “Development of a smart traffic light control system with real-time monitoring,” IEEE Internet of Things Journal, vol. 8, no. 5, pp. 3384-3393, 2020, doi: 10.1109/JIOT.2020.3022392.
[3] A. Atta, S. Abbas, M. A. Khan, G. Ahmed, and U. Farooq, “An adaptive approach: Smart traffic congestion control system,” Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 9, pp. 1012-1019, 2020, doi: 10.1016/j.jksuci.2018.10.011.
[4] S. S. S. M. Qadri and E. Oner, “State-of-art review of traffic signal control methods: challenges and opportunities,” European transport research review, vol. 12, no. 1, pp. 55-55, 2020, doi: 10.1186/s12544-020-00439-1.
[5] F. S. P. Borges, A. P. Fonseca, and R. C. Garcia, “Deep reinforcement learning model to mitigate congestion in re-al-time traffic light networks,” Infrastructures, vol. 6, no. 10, pp. 138-138, 2021, doi: 10.3390/infrastructures6100138.
[6] T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang, et al., “Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks,” IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 8243-8256, 2020, doi: 10.1109/TVT.2020.2997896.
[7] R. Zhao, H. Hu, Y. Li, Y. Fan, F. Gao, Z. Gao, et al., “Sequence Decision Transformer for Adaptive Traffic Signal Control,” Sensors, vol. 24, no. 19, pp. 6202-6202, 2024, doi: 10.3390/s24196202.
[8] M. Yang, Y. Wang, Y. Yu, and M. Zhou, “MixLight: Mixed-agent cooperative reinforcement learning for traffic light control,” IEEE Transactions on Industrial Informatics, vol. 20, no. 2, pp. 2653-2661, 2023, doi: 10.1109/TII.2023.3296910.
[9] Q. U. Jamil, K. D. Kallu, M. J. Khan, M. Safdar, A. Zafar, M. U. Ali, et al., “Urban traffic signal control optimization through Deep Q Learning and double Deep Q Learning: a novel approach for efficient traffic management,” Multimedia Tools and Applications, vol. 84, no. 22, pp. 24933-24956, 2025, doi: 10.1007/s11042-024-20060-x.
[10] S. M. M. R. Swapno, S. M. N. Nobel, P. Meena, V. P. Meena, A. T. Azar, Z. Haider, et al., “A reinforcement learning ap-proach for reducing traffic congestion using deep Q learning,” Scientific Reports, vol. 14, no. 1, pp. 30452-30452, 2024, doi: 10.1038/s41598-024-75638-0.
[11] M. Eom and B. I. Kim, “The traffic signal control problem for intersections: a review,” European transport research review, vol. 12, no. 1, pp. 50-50, 2020, doi: 10.1186/s12544-020-00440-8.
[12] H. Wei, G. Zheng, V. Gayah, and Z. Li, “Recent advances in reinforcement learning for traffic signal control: A sur-vey of models and evaluation,” ACM SIGKDD explorations newsletter, vol. 22, no. 2, pp. 12-18, 2021, doi: 10.1145/3447556.3447565.
[13] W. H. Lee and C. Y. Chiu, “Design and implementation of a smart traffic signal control system for smart city applica-tions,” Sensors, vol. 20, no. 2, pp. 508-508, 2020, doi: 10.3390/s20020508.
[14] X. Wang, L. Ke, Z. Qiao, and X. Chai, “Large-scale traffic signal control using a novel multiagent reinforcement learning,” IEEE transactions on cybernetics, vol. 51, no. 1, pp. 174-187, 2020, doi: 10.1109/TCYB.2020.3015811.
[15] Z. Fu, J. Zhang, F. Tao, and B. Ji, “Traffic signal phase control at urban isolated intersections: an adaptive strategy utilizing the improved D3QN algorithm,” Measurement Science and Technology, vol. 36, no. 1, pp. 016203-016203, 2024, doi: 10.1088/1361-6501/ad8212.
[16] L. Li-Juan, G. M. Bai, and H. R. Karimi, “Advanced Traffic Signal Control System Using Deep Double Q-Learning with Pedestrian Factors,” Journal of Artificial Intelligence and Soft Computing Research, vol. 15, no. 3, pp. 239-255, 2025, doi: 10.2478/jaiscr-2025-0012.
[17] S. Li, P. Shi, A. Yang, H. Qi, and X. Dong, “Dual-Priority Delayed Deep Double Q-Network (DPD3QN): A Dueling Dou-ble Deep Q-Network with Dual-Priority Experience Replay for Autonomous Driving Behavior Deci-sion-Making,” Algorithms, vol. 18, no. 5, pp. 291-291, 2025, doi: 10.3390/a18050291.
[18] B. Wang, Z. He, J. Sheng, and Y. Chen, “Deep reinforcement learning for traffic light timing optimization,” Processes, vol. 10, no. 11, pp. 2458-2458, 2022, doi: 10.3390/pr10112458.
[19] Y. Dong, H. Huang, G. Zhang, and J. Jin, “Adaptive Transit Signal Priority Control for Traffic Safety and Efficiency Optimization: A Multi-Objective Deep Reinforcement Learning Framework,” Mathematics, vol. 12, no. 24, pp. 3994-3994, 2024, doi: 10.3390/math12243994.
[20] Z. Yang, Y. Kong, and C. H. Hsia, “Derlight: A deep reinforcement learning traffic light control algorithm with dual experience replay,” Journal of Internet Technology, vol. 25, no. 1, pp. 79-86, 2024, doi: 10.53106/160792642024012501007.
[21] N. Triki, M. Karray, and M. Ksantini, “A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles,” Applied Sciences, vol. 13, no. 8, pp. 4793-4793, 2023, doi: 10.3390/app13084793.
[22] Z. Wang, H. Zhu, M. He, Y. Zhou, X. Luo, N. Zhang, et al., “GAN and multi-agent DRL based decentralized traffic light signal control,” IEEE Transactions on Vehicular Technology, vol. 71, no. 2, pp. 1333-1348, 2021, doi: 10.1109/TVT.2021.3134329.
[23] H. Mei, X. Lei, L. Da, B. Shi, and H. Wei, “Libsignal: An open library for traffic signal control,” Machine Learning, vol. 113, no. 8, pp. 5235-5271, 2024, doi: 10.1007/s10994-023-06412-y.
[24] P. Michailidis, I. Michailidis, C. R. Lazaridis, and E. Kosmatopoulos, “Traffic Signal Control via Reinforcement Learning: A Review on Applications and Innovations,” Infrastructures, vol. 10, no. 5, pp. 114-114, 2025, doi: 10.3390/infrastructures10050114.
[25] Q. Wu, J. Wu, J. Shen, Yong, B, and Q. Zhou, “An edge based multi-agent auto communication method for traffic light control,” Sensors, vol. 20, no. 15, pp. 4291-4291, 2020, doi: 10.3390/s20154291.