Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 9805560, 17 pages
https://doi.org/10.1155/2021/9805560

Research Article

Trajectory Optimization of CAVs in Freeway Work Zone considering Car-Following Behaviors Using Online Multiagent Reinforcement Learning

Tong Zhu,[1] **Xiaohu Li**,[2] **Wei Fan**,[3] **Changshuai Wang**,[4] **Haoxue Liu**,[5] **and Runqing Zhao**[6]

[1]*College of Transportation Engineering, Chang'an University, Xi'an, China*
[2]*China Automotive Technology & Research Center, Beijing, China*
[3]*Beijing Jingdong Century Trading Co., Ltd., Beijing, China*
[4]*School of Transportation, Southeast University, Nanjing, China*
[5]*School of Automobile, Chang'an University, Xi'an, China*
[6]*School of Aviation, UNSW Sydney, High St, Kensington, NSW 2052, Australia*

Correspondence should be addressed to Tong Zhu; zhutong@chd.edu.cn

Received 16 May 2021; Revised 12 July 2021; Accepted 12 October 2021; Published 3 November 2021

Academic Editor: Xinqiang Chen

Work zone areas are frequent congested sections considered as the freeway bottleneck. Connected and autonomous vehicle (CAV) trajectory optimization can improve the operating efficiency in bottleneck areas by harmonizing vehicles' manipulations. This study presents a joint trajectory optimization of cooperative lane changing, merging, and car-following actions for CAV control at a local merging point together with upstream points. The multiagent reinforcement learning (MARL) method is applied in this system, with one agent providing a merging advisory service at the merging point and controlling the inner-lane vehicles' headway for smooth outer-lane vehicle merging, while other agents provide lane-changing advisory services at advance lane-changing points to control how vehicles make lane changes in advance and perform corresponding headway adjustment, similar to and jointly with the merging advisory service. Uniting all agents, the coordination graph (CG) method is applied to seek the global optimum, overcoming the exponential growth problem in MARL. Using MATLAB and the VISSIM COM interface, an online simulation platform is established. The simulation results show that MARL is effective for online computation with in-timing response. More importantly, comparisons of the results obtained in various scenarios demonstrate that the proposed system obtained smoother vehicle trajectories in all controlled sections, rather than only in the merging area, indicating that it can achieve better traffic conditions in freeway work zone areas.

1. Introduction

Work zone areas are bottleneck sections affecting the operating efficiency of freeways [1]. In the United States, traffic congestion in work zone areas accounts for 10% of the total mileage driven, while in Germany, it accounts for 31% [2]. Congestion in work zone areas is mainly caused when drivers fail to change lanes in advance, which means that they must wait at merging points [3]. Due to the limitations of human driving capabilities (a long response time and limited information processing capacity) and the heterogeneity among drivers, the utilization of the benefits of transportation facilities will be seriously affected [4]. To reduce the occurrence of traffic congestion resulting from human factors, real-time vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) information sharing and communication technologies for autonomous vehicles (AVs) can be used in traffic management [5–7].

With the development of wireless communication technology, AVs are developing in the direction of

collaborative autonomous driving among multiple vehicles, also known as multivehicle cooperative driving [8, 9]. A growing number of people have begun to conduct a lot of research on autonomous vehicles by the method of data analysis. The related research mechanisms include speed harmonization and coordinated trajectory optimization among vehicles [10, 11]. Although the technologies of vehicle adaptive control can be implemented separately, when they are integrated to create a connected and autonomous vehicle (CAV) environment, traffic flow performance in terms of factors such as safety, comfort, and efficiency are expected to be dramatically improved [12]. Considering the large number of work zones for road maintenance and road reconstruction that will be necessary in the future, it is necessary to investigate approaches to improve efficiency by manipulating CAVs.

With regard to the research on car-following behavior of autonomous vehicles, the optimization efforts focus not only on the regulation of acceleration and speed to achieve the control of a single vehicle but also on the speed difference between preceding and following vehicles to achieve coordinated control between multiple vehicles [13, 14]. In addition, to improve the traffic flow operating at freeway bottlenecks, previous studies have achieved CAV control by manipulating lane changing proportion and merging timepoints [15, 16]. However, only the control of the car-following process of the leading vehicle and the following vehicle or the lane change action has been considered as the optimization objective, without considering the control of the headway in CAVs' car-following action. If the car-following action is also taken as a control object, it means that more controllable resources are available for achieving better performance. In addition, a smoother trajectory in all sections becomes desirable. Hence, the impact of the distance between vehicles during lane changing on the trajectory control problem requires further investigation [17]. To address the gap in previous research, this paper aims to propose an overall joint optimization method based on the RL for CAV control to control lane changing, merging, and car-following action in multisections upstream of the work zone area.

Recently, lot of researchers have proposed extended car-following models based on the former research results, which is of great significance in the development of car-following theory. In order to improve the safety of the vehicles in the car-following model, the research regarding the backward-looking effect is widely studied [18], which considers the optimization of the driving behavior in autonomous vehicles' following model. Zhu et al. [19] proposed a human-like autonomous car-following model based on historical driving data, where the agents in reinforcement learning from the data and find a car-following model that maps in a human-like way. This study demonstrates that the methodology of RL can offer insight into the driver behavior and can contribute to the development of car-following models. To investigate a car-following model for Chinese drivers, Zhu et al. [20] predicted the drivers' car-following behavior using the real-world data by analysing five representative car-following models. The car-following model parameters are adjusted, which contributes to reproduce the vehicle trajectory based on the existing data. Wang et al. [21] captured car-following behaviors by deep learning, where the complicated human actions are described by the intelligent algorithm, which contributes to higher simulation accuracy than the existing car-following models. However, with regard to the optimal control of CAV trajectories near a bottleneck area, such as a work zone, car-following actions have been rarely considered in the related previous studies. In addition, most of the current car-following models are solved by a complex function, and there is a lack of machine learning methods to adaptively and collaboratively adjust the driving behaviors of vehicle. Compared with the existing car-following model, the proposed new method takes into account the lane-changing behavior of vehicles in other lanes. That is, under the control of the proposed agents, the method not only takes into account the coordinated driving behaviors between the leading vehicle and the following vehicle in the main lane, but also considers how the vehicles are allowed to make lane changing behaviors from closed lanes at all times by adjusting the driving gap between the leading vehicle and the following vehicle. What's more, the newly proposed method not only has the benefit of adjusting the car-following process of the leading vehicle and the following vehicle according to the local lane conditions, but also takes into account the upstream and downstream global conditions of the lane based on the multiagent reinforcement learning algorithm.

Several problems arise when vehicles are being driven in merging areas on multilane main roads, which are urgent matters to be addressed in freeway. On the one hand, the distribution of the traffic volume may be imbalanced between multilanes, which can result in a lower merging efficiency when the closed lane traffic volume overflows. On the other hand, due to frequent lane-changing, merging, and yielding behaviors, it is typically difficult to achieve efficient operation under natural conditions. Additionally, if multiple control schemes are incorporated into a single optimization model, the model formulation will be highly complex and potentially suffering from the curse of dimensionality, which will make such a model difficult to be applied.

Thus, this study develops an online system control algorithm for trajectory optimization of CAV to intelligently solve the traffic congestion problem and optimize the driving behavior of the vehicles in multilane freeway work zone areas. The main contributions can be summarized as follows.

The operations of mandatory lane changing and discretionary lane changing near the freeway work zone areas are analyzed, and the characteristics of the car-following behavior between multiple vehicles are considered. The work zone area is divided into two regions (lane changing region and merging region); besides, a collaborative mechanism is built to solve the coordination difficulties of multivehicles in multiple areas of joint control.

The established system jointly optimizes lane changing, merging, and car-following action in the CAV environment by the cooperate mechanism, where the optimization objectives are formulated to balance the flow distribution and maximize the bottleneck outflow by controlling car-

following action at virtual lane-changing points in the lane changing region together with the merging behavior in the merging region, thus allowing efficient operation to be achieved under natural conditions by reducing the overall number of lane-changing, merging, and yielding maneuvers. For cooperate mechanism, if multiple control schemes are incorporated into a single optimization model, the model formulation will be highly complex and potentially suffering from the curse of dimensionality, which will make such a model difficult to be applied. To fulfill this objective, we introduce a multiagent reinforcement learning (MARL) method that guides different traffic controllers to work cooperatively and continuously by decomposing the global value function into local value functions easily combining the coordination graph (CG) algorithm, thereby reducing the computational complexity of the control problem and improving their performance with respect to the control objectives.

What's more, a platform based on MATLAB and VISSIM COM is developed to implement the proposed algorithm, which realizes the application of intelligent algorithms in traffic control through real-time data acquisition and control feedback.

The rest of this paper is organized as follows. Section 2 describes related work. Section 3 presents the methodology of the proposed control system. Section 4 describes a numerical study and simulation results. Section 5 presents the conclusion. Section 6 presents further discussion and plans for future research.

2. Related Work

A large number of studies reported in the literature have investigated the optimization control method of freeway bottleneck points in the CAV environment [22, 24]. These studies usually include two tasks for control. The first is to control the on-ramp (or lane), and the second is to control the main traffic flow upstream of the confluence point.

A previous study addressed CAV trajectory optimization on a single-lane freeway with one ramp merging area [25]. This study controlled the merging time-points without accounting for behavior prior to the merging area. Extending the investigations [26], Hu et al. added an early merging road point to enable vehicles to change lanes in advance, thereby improving the efficiency of traffic at the merging point. However, in this research, the trajectory optimization control of the vehicles was considered separately at the early merging road point and the merging point, resulting in global optimization not being performed. Zhang et al. [27] added multiple merging road points before the lane reduction area and applied a unified objective function for traffic flow distribution to reduce lane congestion. In the above paper, only the adjustment of the upstream lane change ratio was studied, and the regulation of the lane-changing action itself was not considered. Furthermore, in the study above, the genetic algorithm used had little capacity for self-repair during the process of solving the objective function, making it susceptible to trapping in local optima.

The real-time advantages of reinforcement learning (RL) and multiagent reinforcement learning (MARL) technology are more applicable to AV optimization control. Related prior studies have investigated the use of RL to adjust the driving action parameters of AV. Qu et al. used MARL algorithms to adjust the acceleration and speed of different vehicles [28] with the aim of mitigating traffic congestion in AV environments. Lu et al. [29] applied an RL method for ramp control to intelligently adjust the duration of the red signal on a ramp so as to reduce/increase the number of vehicles entering the ramp, thereby improving the traffic efficiency. In these investigations, reinforcement learning shows great advantages over previous methods in solving multiobjective optimization problems in a variable online environment.

As analyzed in the literature review, the main cause of traffic breakdown at a merging location is the fact that many vehicles in the outer lane make lane changes near the merging area. This high frequency of lane changing can lead to lower speeds, and the lower-speed lane-changing vehicles then create voids in the traffic stream, reducing the outflow [30]. Hence, it is necessary to propose a strategy for distributing advance lane changes, which will, in turn, distribute the effect on traffic efficiency near the merging area. The voids created by upstream lane-changing vehicles may be filled by downstream lane-changing vehicles with higher insertion speeds, thus allowing these downstream vehicles to avoid causing additional voids themselves. In addition, to alleviate congestion caused by irregular merging behavior in a lane loss section, the previous study [25] has suggested that an intelligent merging advisory mechanism upstream of the merging point on a one-lane mainline with one ramp entrance area based on V2V technology can reduce conflicts between the mainline and merging vehicles. Following this basic idea, Hu et al. [26] have proposed a combined strategy regarding early lane-changing control and cooperative merging control on a two-lane road with one ramp; however, the two control problems were optimized sequentially and separately.

Following the abovementioned ideas, the control logic presented in this paper explores the joint optimization of early lane-changing control and merging control, executed both in the upstream region and near the work zone simultaneously; specifically, both control problems are optimized collaboratively and interactively. As a result, the overall traffic efficiency can be improved. Figure 1 illustrates the trajectory optimization framework proposed in this study.

The process can be simply explained as follows:

Similar to the situation of vehicle lane changing in real life, agent 1 provides a merging advisory service to vehicles in the work zone area. When agent 1 identities that there is a sufficient merging gap in the inner lane at the merging point, it will allow vehicles in the outer lane to perform merging; otherwise, the outer-lane vehicles will stop and wait. Similar to the vehicle merging advisory process, the upstream section is divided into several segments, and the i-th agent provides a lane-changing advisory service for vehicles in the i-th segment. When there is a sufficient lane-changing gap in

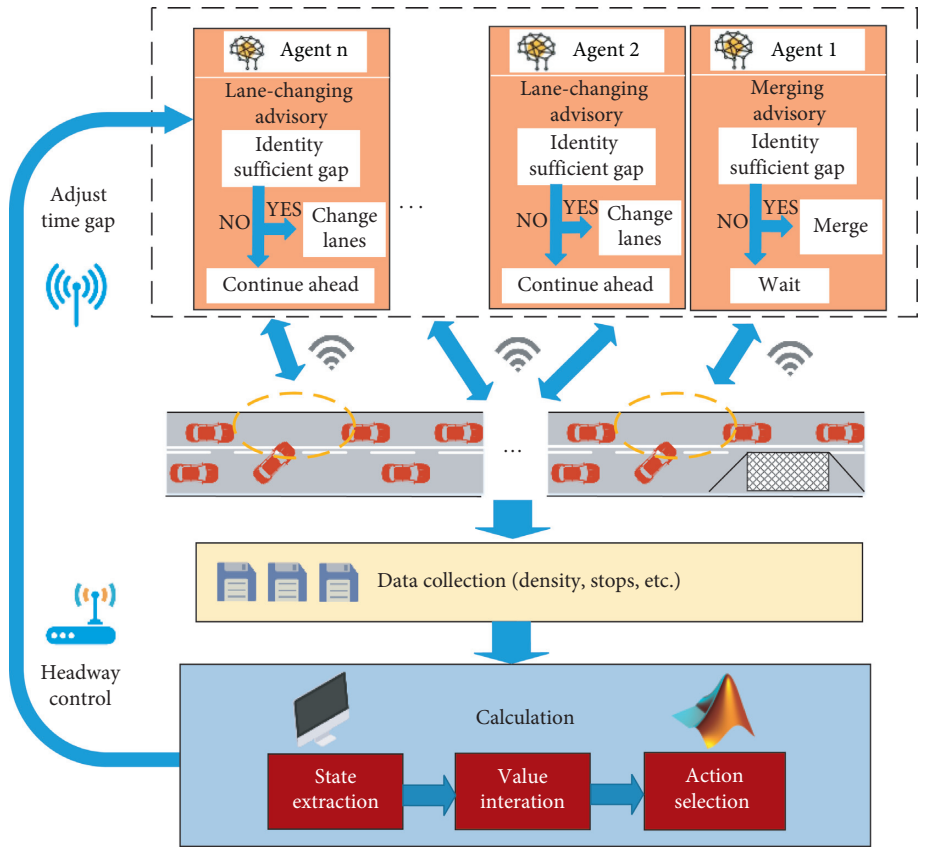

Figure 1: Variable headway execution process.

the inner lane, vehicles in the outer lane are allowed to change lanes in advance; otherwise, the vehicles continue ahead into the next segment. Each agent calculates and iterates the collected data in every time-step and selects the corresponding headway control for feedback to adjust the driving time gap of the inner-lane vehicles in the next time-step.

To simplify the entire optimization process and improve the efficiency of algorithm optimization, the controlled vehicle in work zone area is a 100% CAV environment, so as to completely eliminate the impact of the heterogeneity of human-driven vehicles. Since closed lane vehicles need to merge from the closed lane into the main road and then drive through the work zone, at this time, to reduce the agent search process and avoid conflicts caused by frequent lane changes, the vehicles controlled in this article can only change lanes from the outer lane to inner lane.

3. Methodology

3.1. Distributed Multiagent Reinforcement Learning. RL has been widely used in transportation research [31]. The research on RL is based on the theoretical framework of the Markov decision process (MDP). As a mathematical idealized form of reinforcement learning, MDP can make precise theoretical statements. The theory of MDP claims that any problem of goal-oriented learning behavior can be simplified to three signals transmitted between the agent and its environment: one signal represents the action of the agent's choice, one signal represents the environment state that is the basis for the choice action, and another signal represents the reward that the environment feedbacks to the agent. If the environment state space and the agent behavior space are limited, it is called a finite Markov decision process. The RL problem can be described in terms of a four-tuple of the form $\{S, A, \pi, r\}$, subject to the MDP. S is the state space used to describe the external environment; A is the control action; π is the state transition policy, where for the state pair $(s, s' \in S)$, $\pi^A(s, s')$ represents the probability of reaching the state s' after action A is performed in the state s, and the probabilities for all state pairs satisfy $\sum_s \pi^A(s, s') = 1$; and r: $S \times A \longrightarrow r$ is the reward function. Given a policy π, an optimal value function (Q value function) can be defined for each state-action pair (S, A) as follows:

$$Q^{\pi}(s,a) = E^{\pi}\left\{\sum_{k=0}^{\infty} \gamma^{l} R\left(s^{t+l+1}, a^{t+l+1}\right) \mid s^{t} = S, a^{t} = A\right\}. \tag{1}$$

Here, t represents the time step and l is the loss factor coefficient, which represents the infinite loss of the action value function. $\gamma \in [0, 1]$ is the discount factor. s^t is the state at time t. a^t is the selected action at time t. The ultimate goal of RL is to find the optimal policy π that can select the action that will maximize the Q value.

The agent and the environment interact in a series of discrete time-steps. At each time-step t, $t = 0, 1, 2, 3, \ldots$, the agent perceives the state of the environment and chooses the corresponding actions. After a period of time, the agent receives a digital reward, when the environment reaches a new state. Thereby, a MDP sequence is generated: $s0, a0, r1, s1, a1, r2, s2, a2, r3, \ldots$, where the policy π and reward r of the current state s transition to the next state s' depends only on the current state s and the choice action a, which is not related with the historical state and action.

A common use of RL is Q-learning, which is based on assumptions regarding the actions and focuses on the selection of the maximum value of the Q value function in each iteration; therefore, it is an offline model-free RL algorithm. The Q-learning formula is as follows:

$$Q^t\left(s^t, a^t\right) + \alpha\left[r^t\left(s^t, a^t\right) + \gamma \max_{a \in A} Q^t\left(s^{t+1}, a\right) - Q^t\left(s^t, a^t\right)\right], \tag{2}$$

where $\alpha \in [0, 1]$, which means the learning rate. With further advancement in RL research, an improved RL algorithm called SARSA has emerged [32], which iterates on the actual Q value and updates the Q value matrix based on the experience gained from the implementation of the actual policy; in other words, it is an online algorithm. In addition, the iterative processes for the value function and the action selection strategy are performed separately in Q learning, while in SARSA, the actual action is used for the value function update, meaning that the same iterative process addresses both the value function and the action selection strategy. In other words, the iterative Q value update process is performed based on the actual actions and corresponding states in SARSA, and there is no need to traverse the traffic states as in the traditional offline Q-learning algorithm; consequently, the learning efficiency is improved.

The SARSA learning formula is as follows:

$$Q^{t+1}\left(s^t, a^t\right) = Q^t\left(s^t, a^t\right) + \alpha\left[r^t\left(s^t, a^t\right) + \gamma Q^t\left(s^{t+1}, a^{t+1}\right) - Q^t\left(s^t, a^t\right)\right]. \tag{3}$$

In previous studies, RL targeting an individual object has been considered; however, the MARL method is widely used because it provides the ability to control multiple objects [33]. MARL focuses on a Markov game process to study how agents learn to perform satisfactory actions in complex environments, which provides additional advantages that cannot be gained from single-agent learning. The MARL problem can be described in terms of a tuple of the form $\{S, A_1, \ldots, A_n, \pi, r_1, \ldots, r_n\}$, where S is the external shared environment of all agents; A_i, with $i = 1, 2, \ldots, n$, are the sets of control actions of the n agents; π is the state transition policy distribution; and r_i, with $i = 1, 2, \ldots, n$, are the total reward functions of the n agents.

When solving such a MARL system using centralized MARL, a coordination mechanism for the entire multiagent system is usually taken as the learning goal. The learning task is performed by a global central learning unit, which takes the overall state of the entire multiagent system as input and generates the action assignment for each agent as output. In this way, an optimal coordination mechanism is gradually formed. The Q value update rule for centralized MARL is as follows:

$$Q^{t+1}\left(s^t, a_1^t, \ldots a_n^t\right) = Q^t\left(s^t, a_1^t, \ldots, a_n^t\right) + \alpha\left[R^t\left(s^t, a_1^t, \ldots, a_n^t\right) + \gamma \max_{a_1, \ldots a_n \in A} Q^t\left(s^{t+1}, a_1, \ldots, a_n\right) - Q^t\left(s^t, a_1^t, \ldots, a_n^t\right)\right], \tag{4}$$

where a_i means the action of the i-th agent, and a_i^t denotes the action of the i-th agent at time t.

However, such centralized learning demands that each agent consider the influence of all other agents when selecting its individual action, which requires reasoning over the joint action space of all agents; because the size of this space is exponential in the number of agents, this approach suffers from the curse of dimensionality as the number of possible state-action pairs increases [34]. In addition, the coordination mechanism corresponding to the action selection strategy is not perfect. Even if all agents use the same algorithm to learn a common optimal Q function, although they can theoretically use a greedy policy in RL to maximize their common return, the action selection strategy will break their coordination in a random manner, ultimately leading to the joint action being suboptimal. By contrast, in distributed RL, each individual agent is the main target of learning, thus ensuring the independence of every agent, and only agents that have a direct influence on each other are considered jointly. In addition, the agents learn both a response strategy with respect to the environment and a mutual coordination strategy, enabling decentralization and reducing the interrelations among agents in comparison to centralized RL. In particular, the coordination graph (CG) technique [35] used in distributed RL is an effective technique that involves decomposing the global Q value function into multiple local Q value functions; then, the update of each Q value function is performed based only on multiple agents that directly influence each other, thus allowing the problem of exponential growth to be overcome.

The CG technique is a valid means of addressing distributed coordination problems in which it is assumed that each agent's actions depend on only a subset of the other agents, and an optimal value is returned for each agent based on its contribution to the system for every possible action combination of its neighbors [36]. In summary, the main goal of the CG technique is to find every agent's optimal a relative to the fixed actions of its neighbors to maximize the Q value in state s. The variable elimination (VE) [37] is applied to solve the coordination problem and find the optimal value a. In VE, an agent first collects all value functions related to its neighbors before it is eliminated. The agents are iteratively eliminated until one agent remains, where the final value function is only determined by the remaining agent.

As shown in Figure 2, for the i-th agent, its directly related neighbors are the i-1-th and the $i+1$-th agents. The local Q value is updated in each step as follows:

$$Q(a) = Q_{i-1,i}(a_{i-1}, a_i) + Q_{i,i+1}(a_i, a_{i+1}). \tag{5}$$

The maximum $Q(a)$ can be written as follows:

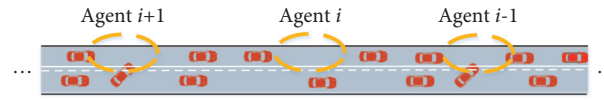

Figure 2: Neighboring agents with mutual relationships.

$$\max_{a_{i-1},a_i,a_{i+1}} Q(a) = \max_{a_{i-1},a_{i+1}} f_i(a_{i-1}, a_{i+1}) = \max_{a_{i+1}} f_{i-1}(a_{i+1}) = f_{i+1}(a_{i+1}), \tag{6}$$

where

$$\begin{aligned} f_i(a_{i-1}, a_{i+1}) &= \max_{a_i}\{Q_{i-1,i}(a_{i-1}, a_i) + Q_{i,i+1}(a_i, a_{i+1})\}, \\ f_{i-1}(a_{i+1}) &= \max_{a_{i-1}}\{f_i(a_{i-1}, a_{i+1})\}, \end{aligned} \tag{7}$$

in $f_{i+1}(a_{i+1})$, the $i+1$-th agent is the remaining agent after eliminating the $i-1$-th and the i-th agents. The maximum $Q(a)$ can be found only by adjusting the action of the $i+1$-th agent. Furthermore, the $i+1$-th agent's optimal action can be calculated as follows:

$$a_{i+1} = \arg\max\{f_{i+1}(a_{i+1})\}. \tag{8}$$

This study proposes the adoption of the CG method at multiple merging points, where each merging point corresponds to an individual agent that selects its own control action in relation to its neighbors, thus allowing a common global optimal control policy to be found. In addition, to reach the global optimum, each agent will consider the influence of the actions and states of other influential neighboring agents when updating the ontology Q value.

In CG, the Q value update rule at time $t+1$ after the i-th agent performs the corresponding action at time t is as follows:

$$Q_i^t\left(s_i^t, a_i^t, S_{N_i}^t, A_{N_i}^t\right) + \alpha\left\{r^t + \gamma Q_i^t\left(s_i^{t+1}, a_i^{t+1}, S_{N_i}^{t+1}, A_{N_i}^{t+1}\right) - Q_i^t\left(s_i^t, a_i^t, S_{N_i}^t, A_{N_i}^t\right)\right\}, \tag{9}$$

where

$$r^t = r_i^t\left(s_i^t, a_i^t, S_{N_i}^t, A_{N_i}^t\right) + \sum_j^{N_i} r_j^t\left(s_i^t, a_i^t, S_{N_j}^t, A_{N_j}^t\right). \tag{10}$$

In this formula, $S_{N_i} = \times_{j\in N^i} s_j$, $A_{N_i} = \times_{j\in N^i} a_j$, α and γ can be adjusted for different scenarios. N_i denotes the number of other agents that are adjacent to the i-th agent. S_{N_i} denotes the multidimensional state of the N_i adjacent agents. A_{N_i} denotes the action space of the N_i adjacent agents. s_i^t denotes The state of the i-th agent at time t. $S_{N_i}^t$ denotes the multidimensional state of the N_i adjacent agents at time t. $A_{N_i}^t$ denotes the action space of the N_i adjacent agents at time t. Q_i^t denotes the Q value of the i-th agent at time t. r^t denotes the sum of the rewards of all agents. $Q_i^t(s_i^t, a_i^t, S_{N_i}^t, A_{N_i}^t)$ denotes the corresponding Q value of the i-th agent at time t, and $r_i^t(s_i^t, a_i^t, S_{N_i}^t, A_{N_i}^t)$ denotes the reward value corresponding to the i-th agent at time t. Correspondingly, the multidimensional Q value of each agent under the combined influence of neighboring agents is found, and finally, all Q values are expressed as follows:

$$Q_i^t\left(s_i^t, a_i^t, S_{N_i}^t, A_{N_i}^t\right) + \sum_j^{N_i} Q_j^t\left(s_i^t, a_i^t, S_{N_j}^t, A_{N_j}^t\right). \tag{11}$$

In the process of solving the MARL problem to find the maximum Q value, the action selection strategy needs to be defined. The traditional greedy action selection strategy in RL and the abovementioned optimal action selection strategy in the CG technique are combined to solve the problem for the $i+1$-th agent. The selection mechanism is described as follows. After t steps, in time-step $t+1$, the following iterative formula gives the probability of selecting the optimal action:

$$\pi_i^{t+1}(a_{t+1}^*) = \pi_i^t(a_{t+1}^*) + \beta\left(1 - \pi_i^t(a_{t+1}^*)\right), \tag{12}$$

where $\pi_i^t(a)$ means the probability of choosing action a at time t. a_t^* means the optimal action at time t. The iterative formula for the probability of choosing another action $a \neq a_{t+1}^*$ is defined as follows:

$$\pi_i^{t+1}(a) = \pi_i^t(a) + \beta\left(0 - \pi_i^t(a)\right), \tag{13}$$

where β is the exploration rate, which satisfies $0 < \beta < 1$.

Accordingly, the final action selection policy formula for MARL based on the greedy strategy and the CG technique is as follows:

$$\pi_i^t(s,a)=\begin{cases}\pi_i^t(a_t^*), & a=\operatorname{argmax}_a\left\{Q_i^t\left(s_i^t,a_i^t,S_{N_i}^t,A_{N_i}^t\right)+\sum_j^{N_i}Q_j^t\left(s_i^t,a_i^t,S_{N_j}^t,A_{N_j}^t\right)\right\},\\ \pi_i^t(a), & a\neq\operatorname{argmax}_a\left\{Q_i^t\left(s_i^t,a_i^t,S_{N_i}^t,A_{N_i}^t\right)+\sum_j^{N_i}Q_j^t\left(s_i^t,a_i^t,S_{N_j}^t,A_{N_j}^t\right)\right\}.\end{cases}\tag{14}$$

3.2. Training Platform and Parameter Values. The training platform used in this study consists of two parts: a MARL training program based on the MATLAB compiler environment and a virtual CAV simulation environment. A schematic diagram of the MARL training platform is shown in Figure 3.

Its logical sequence is described as follows: first, in every training step, the information on speed, road density, and vehicle stops are captured in VISSIM and passed to the MARL program through the COM interface. Next, the MARL program receives the vehicle driving state and evaluates the value of the next action in terms of the reward and loss, and the results are used to update the Q value at the end of the step. In the MARL program, each node of each cube represents an agent, and each edge represents a neighbor relationship between agents; in this way, the influence of neighboring agents is taken into account in the update process for each agent, consistent with the CG method. Then, the action is sent back to VISSIM, and the vehicles perform this action in real time. Finally, as the training steps proceed, the recursive update process continues until a feasible solution is found.

To prevent vehicles from being unable to ideally change lanes and avoid merging distortion, it is necessary to modify the lane change parameter in VISSIM, which allows vehicles to automatically adjust their driving behavior in accordance with their surroundings. As shown in Figure 4, in the parameter setting of VISSIM, the maximum deceleration is reflected in the deceleration at the emergency stop position, where the greater absolute value of the maximum deceleration, the easier it is for the vehicle to change lanes. The accepted deceleration is reflected in the line parallel to the horizontal axis in Figure 4, where the greater absolute value of the accepted deceleration, the easier it is for the vehicle to change lanes. A deceleration of $-1\ \mathrm{m/s^2}$ per distance is essentially reflected in the slope of slash in Figure 4, where the larger the value, the easier it is to change lanes. The safety distance reduction factor is to reduce the expected safety distance of the vehicle during lane change, where the smaller the value, the easier it is to change lanes, and this value is set to 0.6 by default in VISSIM. Refering to the method of VISSIM internal parameter calibration [38, 39], the lane-changing parameters are calibrated according to the set traffic volume in this paper, and the specific values are shown in Table 1.

The paramount of RL is to find a balance between exploration (unknown territory) and reutilization (current knowledge). An effective algorithm should have the ability to jump out of the local search area, where the learning rate has a strong influence on this aspect. In the learning rate setting, if the value is set too small, it will slow down the learning speed. According to the literature [40], the learning rate is set to 0.1, which can reduce the oscillation of the value function. The discount factor is set to 0.9, and the exploration rate is set to 0.2 in the action selection mechanism, which ensures that all actions in certain states can be accessed to the greatest possible extent and to achieve a suitable balance between "re-exploration in the early stage" and "reutilization in the later stage." The parameter settings in the MARL program are listed in Table 1.

4. Numerical Study

4.1. Simulation Environment and Scenario. As shown in Figure 5, we implemented the CAV trajectory control in a simulated traffic scene in which the simulated road is a two-lane freeway and the work zone is located in the outer lane. Vehicles in the outer lane must change lanes to the inner lane before driving past the work zone. The section upstream of the work zone is divided into n smaller segments: segment 1 is the work zone merging area, and all the other segments are virtual merging areas where the lane-changing advisory process is executed. The advance lane-changing advisory process is similar to the merging advisory process in the work zone area.

4.2. Action Definition. The action definition for trajectory optimization is based on the headway of the vehicles in the inner lane, which will directly affect the merging/lane-changing actions of vehicles in the outer lane. The headway will be adjusted in accordance with the surrounding environment, and different headways will affect the vehicle gap available in the inner lane at the lane change point. Specifically, for merging immediately prior to the work zone (segment 1), if agent 1 identifies that there is a sufficient merging gap in the inner lane at the merging point, it will allow vehicles in the outer lane to perform merging; otherwise, the outer-lane vehicles will stop and wait. In the virtual merging areas (from segment 2 to segment n), when a vehicle intends to change lanes from the outer lane, if the target lane (inner lane) has a sufficient gap available for lane changing, then the vehicle will change lanes in advance; otherwise, the vehicle will continue to drive into the next segment.

The agent in each segment optimally controls the headways of all vehicles in the inner lane of that segment to provide sufficient lane-changing gaps for vehicles to perform advance lane-changing actions. In this paper, the headway refers to the time that a following vehicle takes to

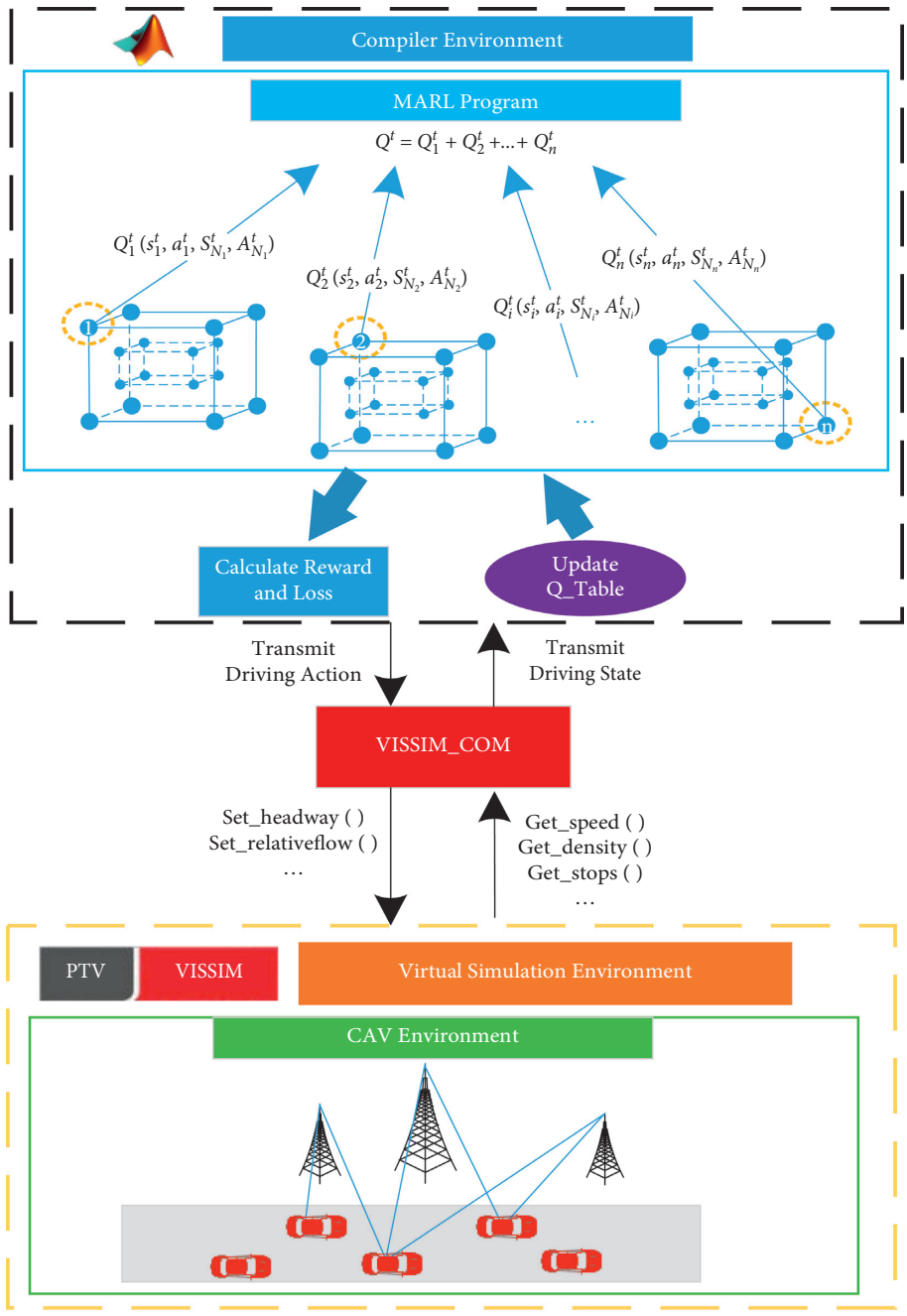

Figure 3: MARL training platform.

cover the distance between the nose of the preceding vehicle and its own nose; therefore, the headway accounts for the impact of the length of the preceding vehicle in addition to the time gap. The time considered for vehicle headway control ranges from 1 s to 3 s in discrete intervals of 0.2 s.

4.3. State Space. For the agent corresponding to the i-th segment, the state space is described in terms of two state values. When vehicles in the outer lane enter the inner lane, different state spaces will appear due to the headway implications of the actions performed and the increase in the number of vehicles in the inner lane. Hence, the first state value is the traffic volume in the inner lane of the road segment, represented by q. Since a vehicle in the outer lane will decelerate when entering the inner lane, a long duration of the merging process or the phenomenon of failed merging in the work zone area will cause the vehicles upstream in the outer lane to move at a lower speed or even pile up, and the vehicle occupancy rate of the outer lane at the lane change point has a great influence on this effect. The second state value is the occupancy rate of every segment, denoted by o. Accordingly, the state space set S of every agent can be defined as follows:

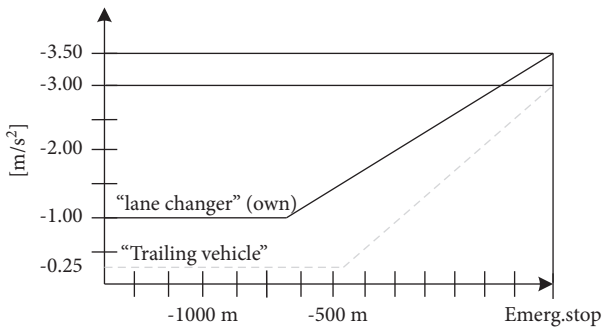

Figure 4: Mandatory lane change parameter relationship.

$$S = \{(q, o) | q \in W, o \in O\}. \tag{15}$$

In this formula, the inner-lane traffic volume value range W is (1200, 3000) veh/h, separated into discrete intervals of 200 veh/h. The occupancy value range O is $(0, 1]$, separated into discrete intervals of 0.1.

4.4. Reward Function. In RL, a reward function is defined to guide the system to produce desirable actions. For CAVs, the goal is high safety, efficiency, and driving comfort. Therefore, if CAV safety decreases, the reward value will be negative. Conversely, the reward for higher safety is positive.

As the agent on each small road segment controls the vehicles in the outer lane to perform lane-changing actions, the vehicles behind them may become congested or be forced to stop due to the resulting decrease in speed (similar to the merging process in the work zone area). Therefore, the operating efficiency of the vehicles needs to be considered. The number of vehicles stopped in the outer lane in the segment controlled by the i-th agent is used to formulate the reward R_1,

$$\begin{aligned} R_1 &= \sum_{t}^{T} \sum_{k}^{K} st_{t,k}, \\ st_{t,k} &= \begin{cases} 1, & V_{t,k} \le V_{\min}, \\ 0, & V_{t,k} \ge V_{\min}, \end{cases} \end{aligned} \tag{16}$$

where $st_{t,k}$ is a Boolean variable. When the speed of vehicle k at time t is less than a minimum vehicle speed limit $V_{\min}$, the value of $st_{t,k}$ is 1, indicating that the vehicle is in a stopped state at this time. Otherwise, the value of this variable is 0. Here, $V_{t,k}$ means the speed of vehicle k at time t. The minimum vehicle speed V_{min} is taken to be 10 km/h, which is the speed value of the stop state defined in VISSIM. In addition, it can be seen from the flow-density curve that the traffic flow reaches its maximum when the density is close to a certain critical point, and the flow decreases when the density deviates from this critical point. Therefore, maintaining the density of a road segment at the critical density is an effective control strategy for alleviating congestion. To avoid a large number of vehicles from the outer lane entering the inner lane within a small road segment, it is of great significance to maintain the road density within that segment at the critical density value. This can be achieved by considering the changes in the variable headway and the distribution of the traffic volume. For the i-th road segment, the reward R_2 is formulated in terms of the inner-lane road density in the $i+1$-th road segment and is expressed as follows:

Table 1: Parameters and values.

Parameter	Value
Maximum deceleration	-6 (m/s^2)
Accepted deceleration	-2.5 (m/s^2)
-1 m/s^2 per distance	200 (m)
Average standstill distance	1.12 (m)
Safety distance reduction factor	0.60
Additive part of safety distance	1.85
Multiplic. part of safety distance	3.02
Learning rate	0.1
Discount factor	0.9
Exploration rate	0.2

$$R_2 = \begin{cases} \dfrac{d_t}{d_{cr}} * r_{\max}, & d_t \le d_{cr}, \\ \dfrac{d_t - d_{cg}}{d_{cr} - d_{cg}} & d_{cr} \le d_t \le d_{cg}, \\ 0, & d_t \ge d_{cg}, \end{cases} \tag{17}$$

where d_t means the instantaneous density of the road; d_{cr} means the critical density; and d_{cg} means the congestion density. Through the calibration of the road parameters in VISSIM [41], the value of d_{cr} is chosen to be 26 veh/km/lane, and the value of d_{cg} is set to be 45 veh/km/lane. $r_{\max}$ denotes the maximum reward and is set to 10. The total reward for each agent, R, is defined by combining $R1$ and $R2$ as follows:

$$R = -R_1 + R_2. \tag{18}$$

To facilitate the MARL-based control, the number of upstream road segments is determined to be 2, corresponding to one advance lane-changing area in which the lane-changing advisory process is performed, in addition to the merging area, in which the merging advisory process is performed. The final control configuration is shown in Figure 6. The total length of the two-lane freeway section considered in the simulation is approximately 1.6 km. The lengths of the work zone area and each small controlled segment are 400 m. The length of the uncontrolled area before vehicles enter the first small controlled segment is 200 m, and the length of the exiting area downstream of the work zone area is 200 m. The vehicle type considered in VISSIM is 100 (cars), and the vehicle desired speed is set to 80 km/h (i.e., 22.2 m/s). Similar to the input traffic volume set for two-lane traffic considered in a previous study [25], the input traffic volume is subject to the following restrictions: (a) allow different traffic volumes to be traversed in one simulation, the input traffic volume in the simulated scene is set to be dynamic, with a range of (1200, 1800) veh/h/ln for the inner-lane traffic volume demand, where 1800

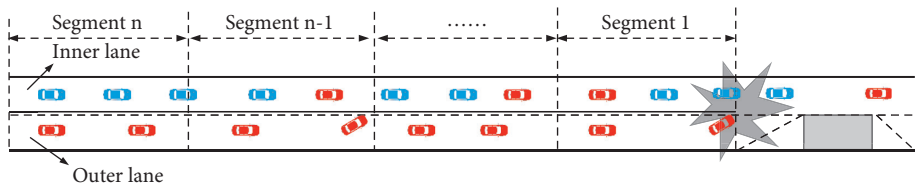

Figure 5: Vehicle driving status near a freeway work zone.

veh/h/ln is approximately the maximum single-lane demand that can be implemented before work zone area in VISSIM. The traffic volume is increased in discrete increments of 200 veh/h/ln; in other words, the initial input inner-lane traffic volume is 1200 veh/h/ln, and the traffic volume is increased by 200 at certain time intervals. (b) To prevent the collapse of control due to excessive traffic, the ratio of the traffic volume in the inner lane to that in the outer lane is set to 65 : 35. This ratio is the optimal traffic distribution determined by solving related problems in traditional research (such as ramp metering) by comparison with 80 : 20 and 50 : 50; with a ratio of 80 : 20, it would be impossible to create sufficient lane-changing gaps due to the excessive number of vehicles in the inner lane, whereas a ratio of 50 : 50 would prevent a sufficient number of vehicles from changing lanes in advance due to the large number of vehicles in the outer lane. Hence, in this study, the outer-lane input volume is varied in the range of (420, 630) veh/h in increments of 70 veh/h/ln; for example, when the input traffic volume of the inner lane is 1800 veh/h/ln, the input traffic volume of the outer lane is 630 veh/h/ln. Each small controlled segment is separated into a gap detection area and a variable headway control area, where the length of the gap detection area is set to 22.2 m/s × 0.9 s≈20 m in accordance with the principle of the priority rules in VISSIM and the default minimum headway in VISSIM, which is 0.9 s.

4.5. Scenarios for Comparison. To verify the proposed control system, we compare four scenarios, as follows:

(i) scenario 0: no optimization control
(ii) scenario 1: RL-based VSL control
(iii) scenario 2: RL-based variable headway control
(iv) scenario 3: MARL-based variable headway control

To directly demonstrate the superiority of the proposed system, we first introduce scenario 0, in which no vehicles are controlled under realistic conditions. For comparison with the traditional method of solving freeway bottleneck problems, we introduce the VSL control scenario, referred to as scenario 1 and use the RL method for VSL control. In the VSL approach, the control target is the speed of the vehicles in the inner lane in segment 1 before the work zone area. As specified in a previous study [42], the VSL area is 1 km upstream of the merging area near the work zone with a length of 500 m; the acceleration zone is established 500 m upstream from the merging point in the work zone area; and the set of action control strategies is defined as (30, 80) km/h, separated into discrete speed intervals of 10 km/h. The state and reward function definitions are similar to those mentioned above. To analyze the superiority of the logic proposed in this study in comparison with the single-agent RL method, we introduce scenario 2 in which the headway of vehicles in the inner lane is controlled only in segment 1 before the work zone area. Scenario 3 corresponds to the MARL-based control system proposed in this study in which the headway of vehicles in the inner lane is controlled in both segments 1 and 2 before the work zone area.

The settings for each simulation scenario under RL- or MARL-based control are as follows: (1) the duration of a single simulation is 1800 s; (2) the input inner-lane traffic volume is increased from 1200 veh/h in increments of 200 veh/h every 450 s; (3) the input outer-lane traffic volume is increased from 420 veh/h in increments of 70 veh/h every 450 s; (4) the total number of simulations is 500; (5) the Q value is updated every 60 s from the first simulation to the 500th simulation.; and (6) in total, the Q value is updated 15,000 times.

4.6. Visualization of Results. To illustrate the convergence and divergence properties of the RL process, the results of the iterative update process in the MARL system for the reward function and the Q value are shown in Figures 7 and 8, respectively. At early times, the Q value may fluctuate due to the instability of the reward value. For better visualization and graphical presentation, the Q value and the reward for each episode have been denoised via the moving average method to analyze the convergence of the RL process. In the moving average method, the denoising of a dataset A can be expressed as movmean(A, m), where m is the length of the moving window. The movmean(A, m) function returns an array containing the mean values of sets of m local data points, where each mean value is calculated based on a moving window consisting of m adjacent elements of A. When there are not enough elements to fill the window, the window will be automatically truncated at the endpoints, and the mean value will be calculated based on only the elements in the window. In this study, the length of the moving window is 10 [43].

After noise reduction, it can be seen from the figures illustrating the evolution of the reward function and the Q value that both the exploration and utilization of the state space are lower in the early stage of RL. At this time, only small reward values can be found, resulting in lower Q values. As the agents continue their exploration, once the number of iterations reaches approximately 10000, the agents begin to earn larger rewards; correspondingly, the Q value and the reward value are continuously rising. When the number of iterations reaches 12,000, the agents begin to

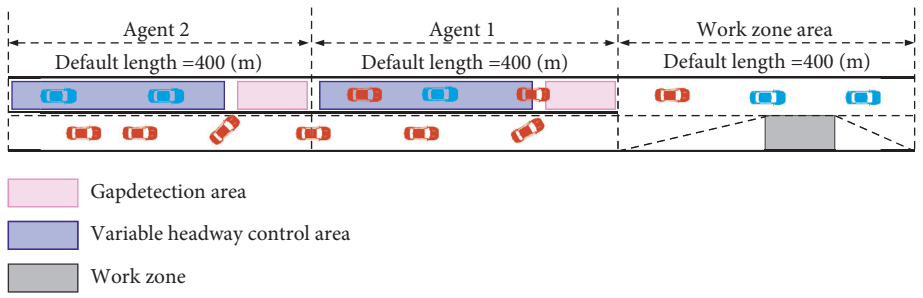

Figure 6: Configuration of the study area.

better "utilize" the values already found through exploration, and in the later iterations, the Q value and reward value reach a state of convergence.

Figure 9 presents a three-dimensional visualization of the MARL reward value, showing the rewards of the agents in each training simulation and in every simulation time-step. It can be seen that in the first 300 VISSIM simulations during the training process, the total reward was relatively high during the first 900 s of simulation time due to the low input traffic volume. After 900 s, with the increase in the input traffic volume (i.e., when the inner-lane input was 1400 veh/h and the outer-lane input was 490 veh/h), the total reward began to gradually decrease, indicating the appearance of congestion. After 300 VISSIM training simulations, when the total number of episodes reached 9000, the agents were able to explore larger rewards until the results stabilized at the maximum reward, corresponding to the observed change in the two-dimensional Q value and the two-dimensional reward value update process.

The vehicle trajectories under the four control scenarios are depicted in Figures 10–13, where the ordinate represents the vehicle position, the abscissa represents the simulation time, the magenta curves represent the trajectories of the inner-lane vehicles, the black curves represent the trajectories of the outer-lane vehicles, and the dashed red line denotes the position of the merging point.

As seen from Figure 10, in scenario 0, an accumulation of outer-lane curves begins to appear after 900 s of simulation time, indicating that the uncontrolled outer-lane vehicles become blocked or stagnant in the merging section, which manifests as congestion due to the reduction in the vehicles' speed. Because of the increase in the input traffic volume over time, the phenomenon of vehicle congestion becomes more severe after 1400 s, corresponding to the continuous decrease in the reward value observed after 900 s in the three-dimensional visualization. Under VSL control (scenario 1), the total outer-lane congestion time and the overall trend of congestion in the later period are almost the same as those in scenario 0. However, with the increase in the input traffic volume, the congestion is slightly decreased compared to scenario 0, and the overall congestion distance of the vehicles is reduced because of the variable speed of the inner-lane vehicles. Under RL-based variable headway control (scenario 2) of the vehicle trajectories, although outer-lane congestion appears earlier than in scenario 1, the overall congestion is alleviated. With the increase in the traffic volume, the congestion distance is significantly reduced, and the ability to relieve the congestion is obviously better than that in scenario 1, as seen from the relatively slight bending of the inner-lane vehicle trajectories. However, after 1600 s, this scenario results in congestion in the inner lane that spreads upstream, indicating that the single-agent adjustment of headway in the inner lane cannot satisfy the merging needs for a large number of outer-lane vehicles. Under MARL-based variable headway control (scenario 3), the distance between the adjacent inner-lane curves sometimes increases, reflecting the variable headway of the inner-lane vehicles. Due to the MARL-based joint control of the lane-changing process and headway both upstream and in the work zone area, compared to scenario 2, the alleviation of congestion is significantly improved, and there is almost no congestion in the outer lane, indicating that outer-lane vehicles are able to merge smoothly without compromising the efficiency of the inner-lane vehicles. Although minor congestion is evident at approximately 1600 s, this congestion disappears immediately thereafter, corresponding to the high reward values observed in the three-dimensional visualization at approximately 1600 s after 300 VISSIM training simulations.

4.7. Quantitative Results. In this paper, the global average delay time of the vehicles, the average number of stops, the average speed, the total travel time, and the average stop delay are chosen as evaluation indicators for vehicle operation, which can be used to comprehensively analyze the phenomenon of vehicle congestion due to the advance lane-changing and merging processes. The values of these indicators in the four investigated scenarios are compared in Table 2.

Regarding the average delay, travel time, and average speed, when the inner-lane vehicles under VSL control (scenario 1) make speed adjustments to allow more outer-lane vehicles to merge, the trajectories in Figure 11(a) become slightly more curved than the trajectories in Figure 10(a); however, the global vehicle indicators in scenario 1 outperform those in scenario 0, with the average delay decreasing by 8.6% and the average speed increasing by 6.4%. This indicates that the RL-based VSL control strategy considered in this study can improve the overall operating efficiency of all vehicles by slightly reducing the speed of vehicles in the main lane, consistent with the results of

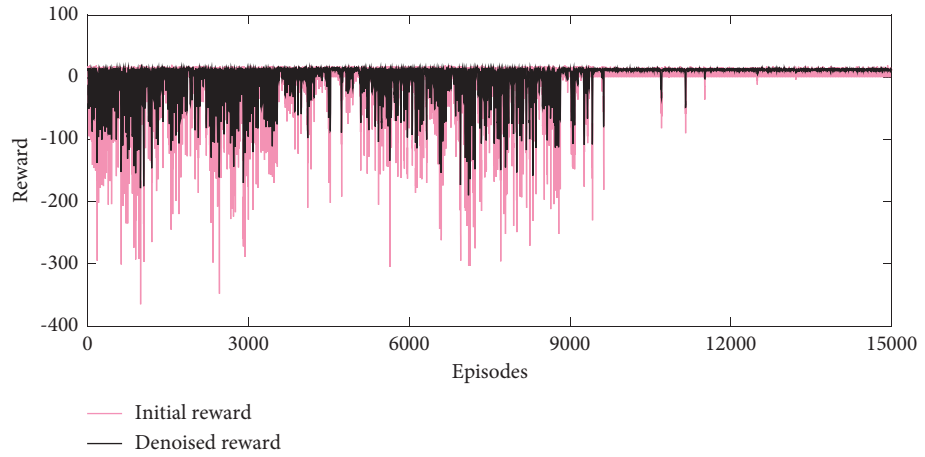

Figure 7: Multiagent reward value update process.

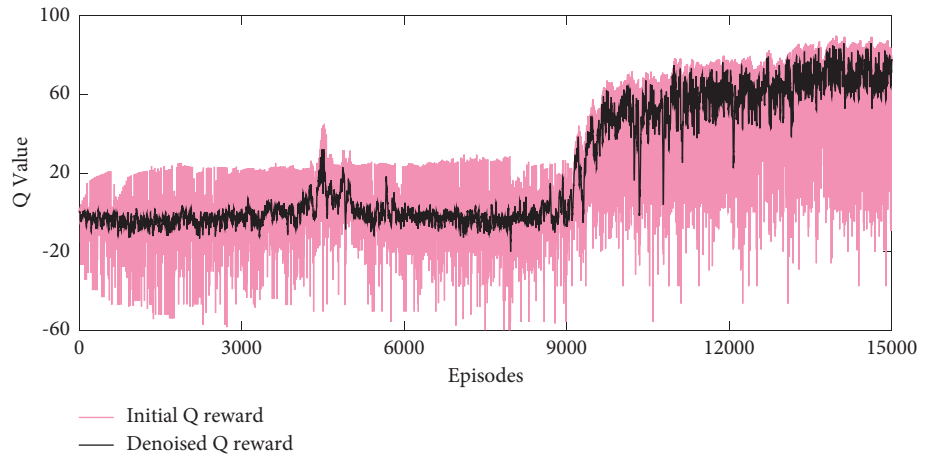

Figure 8: Multiagent Q value update process.

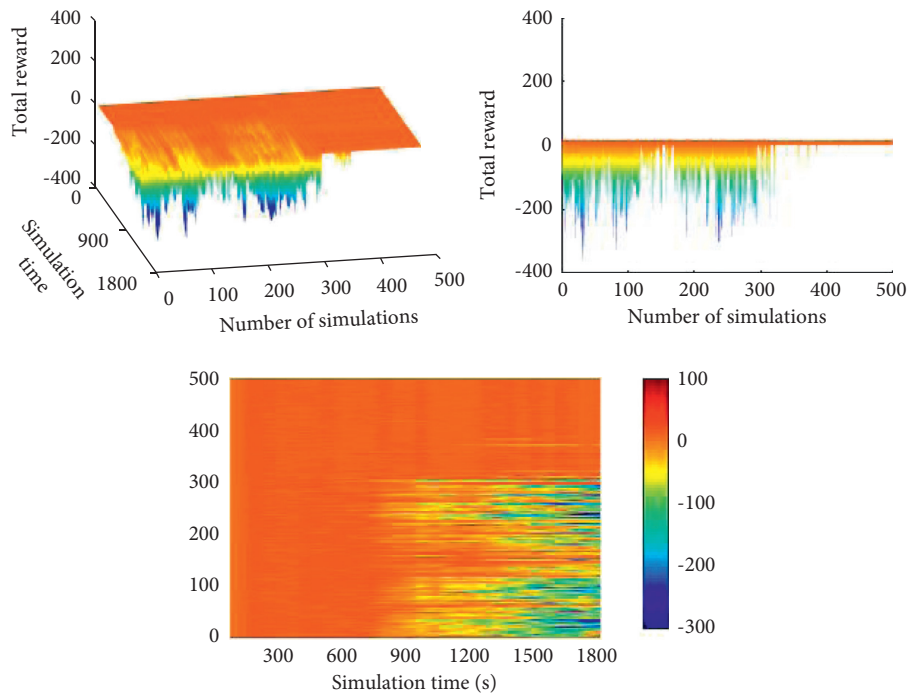

Figure 9: Three-dimensional visualization of the multiagent reward value.

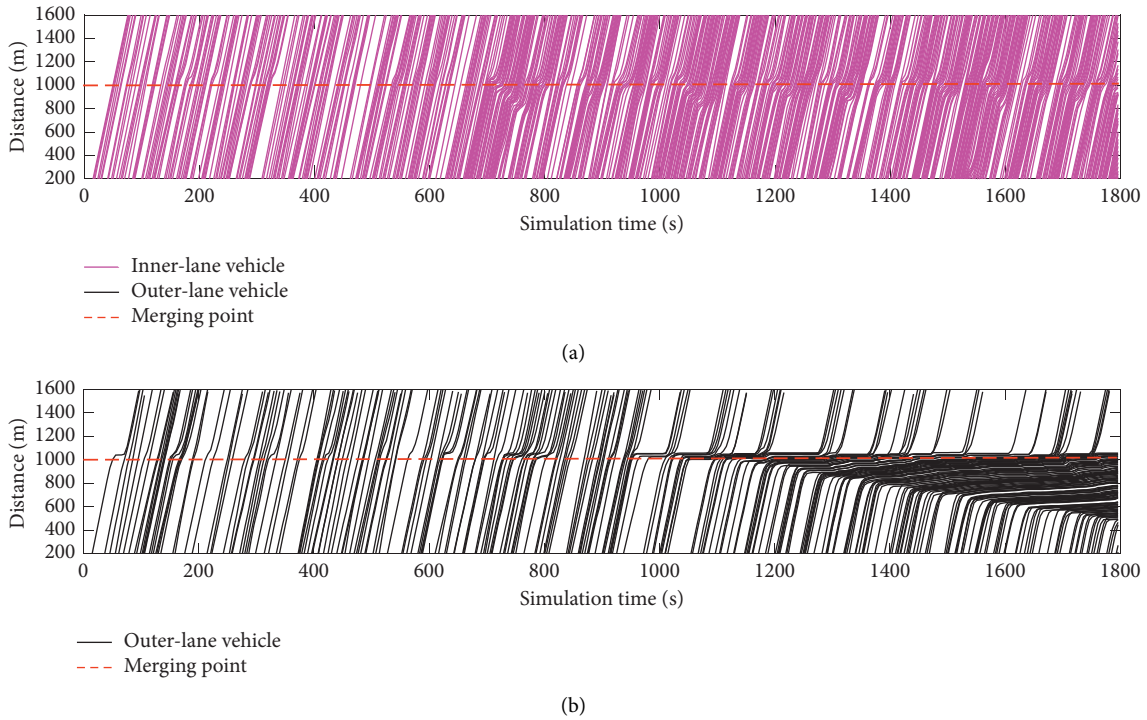

Figure 10: Scenario 0 vehicle trajectories. (a) Inner-lane vehicle trajectories. (b) Outer-lane vehicle trajectories.

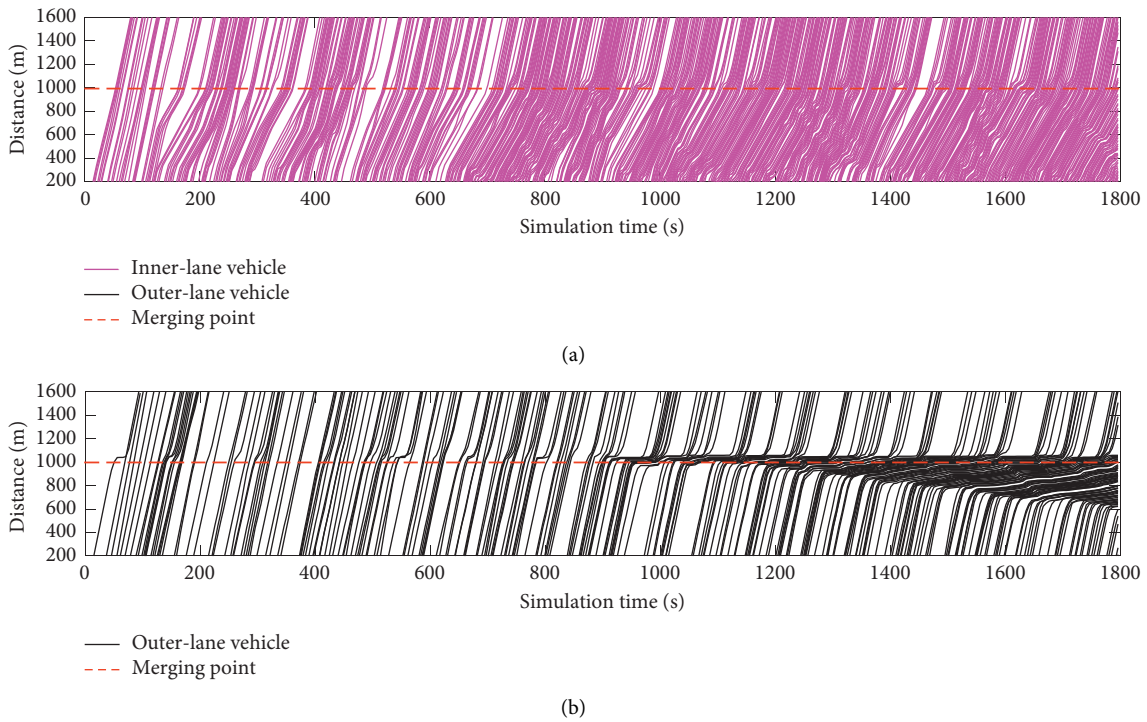

Figure 11: Scenario 1 vehicle trajectories. (a) Inner-lane vehicle trajectories. (b) Outer-lane vehicle trajectories.

previous research on the VSL approach. In turn, scenario 2 consistently outperforms scenario 1, with the average delay time decreasing by 4.4% and the travel time decreasing by 6.4%. This is attributed to the significantly smoother nature of the outer-lane vehicle trajectories in Figure 12(b) compared with those in scenario 1, as depicted in Figure 11(b); by contrast, the inner-lane vehicle trajectories in scenarios 1 and 2 are roughly the same. More importantly, compared with scenario 2, scenario 3 shows an even more prominent optimization effect with the average delay

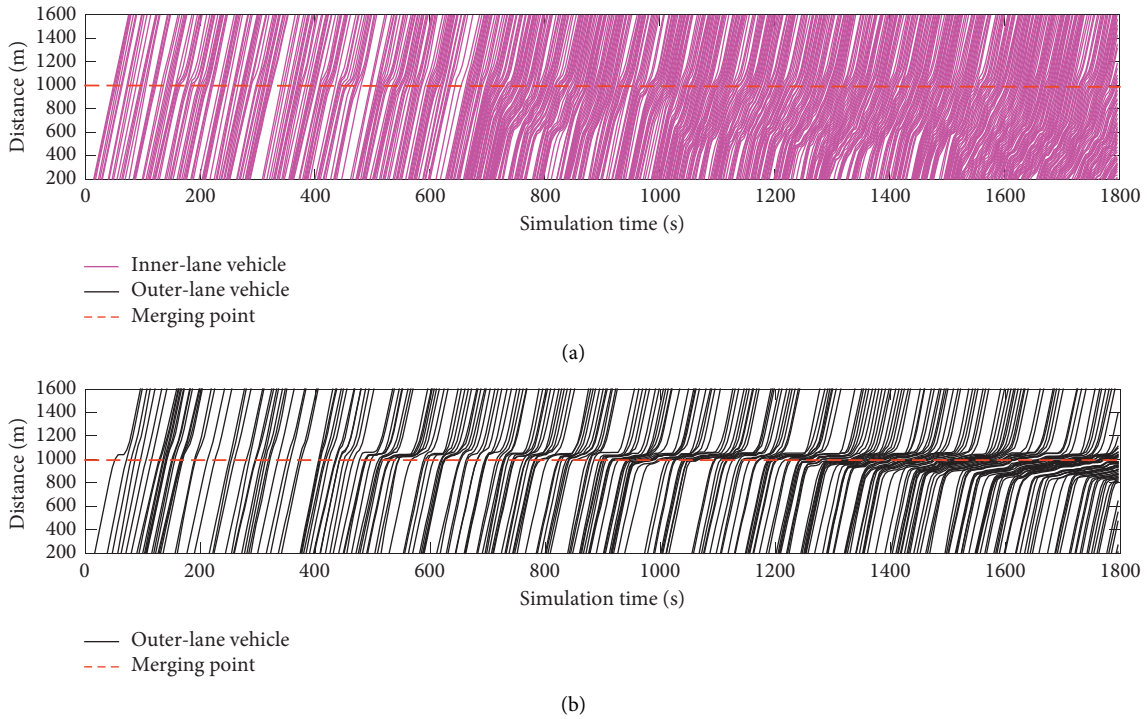

Figure 12: Scenario 2 vehicle trajectories. (a) Inner-lane vehicle trajectories. (b) Outer-lane vehicle trajectories.

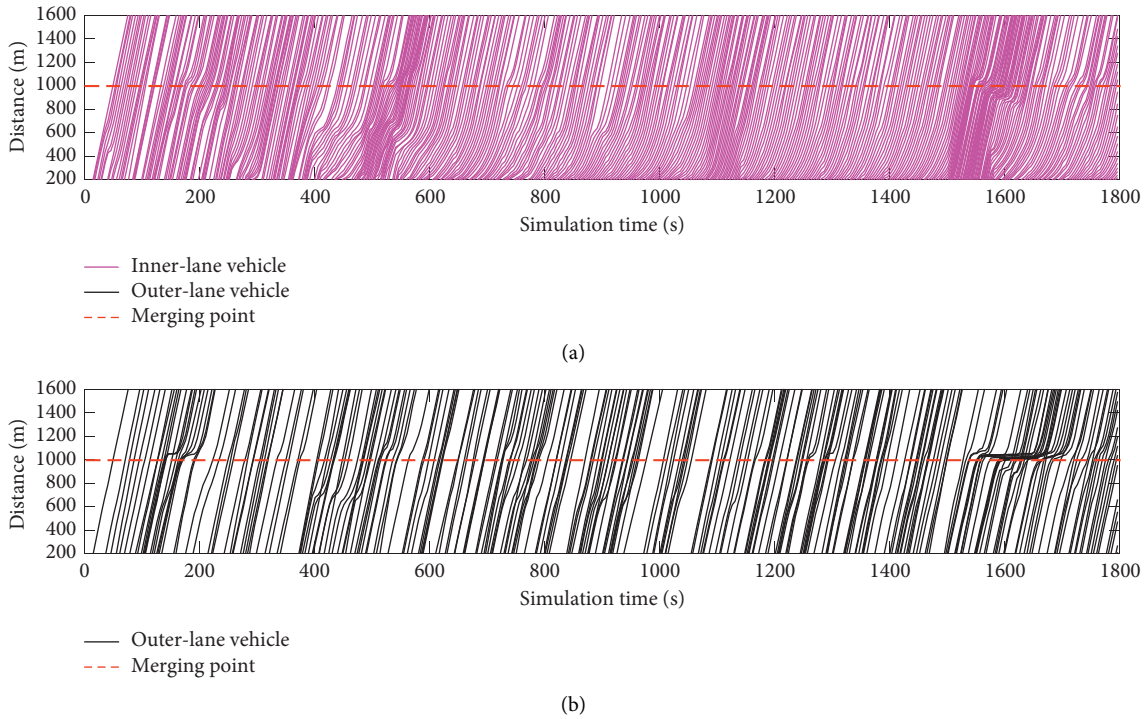

Figure 13: Scenario 3 vehicle trajectories. (a) Inner-lane vehicle trajectories. (b) Outer-lane vehicle trajectories.

Table 2: Network performance.

	Average delay (s)	Travel time (h)	Average speed (km/h)	Average stops	Stop delay (s)
Scenario 0	63.412	36.234	40.030	6.661	9.035
Scenario 1 (% diff. vs. scenario 0)	57.979 (−8.6)	35.160 (−2.9)	42.591 (6.4)	5.602 (−15.9)	6.235 (−3.1)
Scenario 2 (% diff. vs. scenario 1)	55.451 (−4.4)	33.218 (−6.4)	43.876 (3.0)	4.994 (−10.9)	5.404 (−13.3)
Scenario 3 (% diff. vs. scenario 2)	50.327 (−9.2)	26.128 (−21.3)	48.697 (10.9)	3.533 (−29.3)	3.923 (−27.4)

decreasing by 9.2%, the travel time decreasing by 21.3%, and the average speed increasing by 10.9%; these findings are consistent with the fact that almost no accumulation of trajectory curves is observed in Figure 13. These results indicate that the proposed MARL-based control logic enables a considerable efficiency improvement in comparison to the other scenarios.

The average number of stops and the stop delay are reduced by 15.9% and 3.1%, respectively, in the RL-based VSL control scenario (scenario 1) compared to scenario 0; by 10.9% and 13.3%, respectively, in the RL-based variable headway control scenario (scenario 2) compared to scenario 1; and by 29.3% and 27.4%, respectively, in the MARL-based variable headway control scenario (scenario 3) compared to scenario 2. In summary, the significant optimization results for the average number of stops and the stop delay can be attributed to the establishment of the RL reward function. Overall, the MARL-based control leads to better traffic conditions in a freeway work zone area.

5. Conclusions

This paper proposed an online system based on MARL for variable headway control in a freeway work zone area in a CAV environment and present a joint trajectory optimization combining lane changing, merging, and car-following actions for CAV control at a local merging point together with upstream points with the same formula using MARL algorithms.

The operation of mandatory lane changing and discretionary lane changing near the freeway work zone areas is analyzed. Furthermore, the work zone area is divided into two regions (lane changing region and merging region); besides, a collaborative mechanism is built to solve the coordination difficulties of multivehicles in multiple areas of joint control. The flow distribution in lane changing region is balanced, which allows efficient operation to be achieved under natural conditions by reducing the overall number of lane-changing, merging, and yielding maneuvers. The problem of merging behavior has been solved in the merging region, which provides more opportunities to find time gaps for advance lane-changing and merging to ensure the efficient operation of vehicles.

Moreover, the proposed approach improves the overall vehicle operating state and alleviates traffic congestion compared with traditional VSL control for active traffic management and other scenarios, resulting in smoother trajectories and greatly improving several selected evaluation indicators. In addition, based on the CG method, the best actions to achieve the maximum Q value are selected. In the CG method, the global Q value function is decomposed into local Q functions. When each agent updates the ontology Q value, it also considers the influence of the actions and states of other neighboring agents in its local group.

A simulation platform based on the VISSIM COM interface and MATLAB has been applied to perform online calculations and create visualizations of real-time trajectory optimization for each vehicle. The convergence process during training was investigated to elucidate the system training process, and the algorithm was found to reach convergence after the number of training episodes reached 12000. Based on the selected evaluation indicators, the proposed MARL-based variable headway control scenario was compared with the uncontrolled scenario, the VSL control scenario, and the RL-based variable headway control scenario, with respect to which the average delay was found to decrease by 20.6%, 13.2%, and 9.2%, respectively; the average number of stops decreased by 46.9%, 36.6%, and 29.3%, respectively; the average speed increased by 21.7%, 14.3%, and 10.9%, respectively; the total travel time was reduced by 27.9%, 25.7%, and 21.3%, respectively; and the average stop delay was reduced by 56.6%, 37.1%, and 27.4%. Visualization of results shows that the trajectory curves drawn for the vehicles in both the inner and outer lanes were also smoother. This indicates that the joint adjustment of headway and lane-changing actions in the CAV environment can effectively improve traffic operations.

These simulation results also verify that the MARL is an efficient technique for online optimization of the trajectories in the CAV environment due to its advantage in decomposing the value function and avoiding the computational difficulty involved in the joint optimization model.

6. Future Works

Although this research has yielded promising outcomes, some open issues still remain to be discussed in future works.

On the one hand, the scene selected for simulation is a one-way two-lane road; therefore, this research does not consider the case of 3 or more lanes, which may allow the agents to more evenly distribute the vehicles changing from the outer to the inner lanes in the upstream region in accordance with a policy that can make full use of the freeway environment. Hence, in future works, research on optimization control can be performed with slight modifications to consider a freeway bottleneck area with more than two lanes. For this purpose, it will mainly be necessary to add a gap detection area to control the process of middle-lane vehicles changing to the innermost lane.

On the other hand, the vehicle driving environment mentioned in this paper is a full CAV environment, where the case of mixed fleets of the connected vehicles and manually driven vehicles was not taken into consideration, and the compatibility between the connected vehicles and manually driven vehicles was also not addressed in this research. The different proportions of CAV will have an impact on the proposed control effect. Therefore, it is necessary to consider the condition that not all vehicles are connected. The agents in the proposed method will recognize the manually driven vehicles when facing different proportions of the CAV environment. Like a full CAV environment, CAV will also maintain the headway calculated by the agents with the manually driven vehicles and make corresponding lane changes and car-following behaviors. This will be divided into the following two cases for consideration: in the case of a high proportion of CAV, the connected vehicles drive normally under the control of the agents, where the manually driven vehicles will have less

impact on the system. In the case of a low proportion of CAV or the proportion of CAV is approximately the same as the proportion of the manually driven vehicles, CAVs will assume to accept corresponding vehicle platoon formation controls by the agents to reduce communication barriers with manually driven vehicles, where the mentioned control will only be slightly affected in the process of participation with a high proportion of the manually driven vehicles. Hence, in future works, we will introduce manual vehicles into the CAV environment at a certain proportion to consider the simultaneous operation of connected vehicles and manual vehicles.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by In-Depth Accident Study for Improved Injury Assessment Tool and its Coupling with Driver Behaviors for Precise Injury Prevention (Project no. 2019YFE0108000).

References

[1] F. La Torre, L. Domenichini, and A. Nocentini, "Effects of stationary work zones on motorway crashes," *Safety Science*, vol. 92, pp. 148–159, 2017.

[2] Y. Chung, "Assessment of non-recurrent traffic congestion caused by freeway work zones and its statistical analysis with unobserved heterogeneity," *Transport Policy*, vol. 18, no. 4, pp. 587–594, 2011.

[3] A. Ghasemzadeh and M. M. Ahmed, "A tree-based ordered probit approach to identify factors affecting work zone weather-related crashes severity in north carolina using the highway safety information system dataset," in *Proceedings of the Transportation Research Board 96th Annual Meeting*, Washington, DC, USA, January 2017.

[4] X. Qu, J. Zhang, and S. Wang, "On the stochastic fundamental diagram for freeway traffic: model development, analytical properties, validation, and extensive applications," *Transportation Research Part B: Methodological*, vol. 104, pp. 256–271, 2017.

[5] B. Zhu, Y. Jiang, J. Zhao, R. He, N. Bian, and W. Deng, "Typical-driving-style-oriented personalized adaptive cruise control design based on human driving data," *Transportation Research Part C: Emerging Technologies*, vol. 100, pp. 274–288, 2019.

[6] B. Wu, J. Zhang, T. L. Yip, and C. Guedes Soares, "A quantitative decision-making model for emergency response to oil spill from ships," *Maritime Policy & Management*, vol. 48, no. 3, pp. 299–315, 2021.

[7] J. Wang, S. Gong, S. Peeta, and L. Lu, "A real-time deployable model predictive control-based cooperative platooning approach for connected and autonomous vehicles," *Transportation Research Part B: Methodological*, vol. 128, pp. 271–301, 2019.

[8] L. Vasconcelos, L. Neto, S. Santos, A. B. Silva, and Á. Seco, "Calibration of the Gipps car-following model using trajectory data," *Transportation Research Procedia*, vol. 3, pp. 952–961, 2014.

[9] N. Virdi, H. Grzybowska, S. T. Waller, and V. Dixit, "A safety assessment of mixed fleets with connected and autonomous vehicles using the surrogate safety assessment module," *Accident Analysis & Prevention*, vol. 131, pp. 95–111, 2019.

[10] V. Milanés, S. E. Shladover, J. Spring et al., "Cooperative adaptive cruise control in real traffic situations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 296–305, 2013.

[11] X. Chen, Z. Li, Y. Yang et al., "High-resolution vehicle trajectory extraction and denoising from aerial videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3190–3202, 2020.

[12] H. S. Mahmassani, "50th anniversary invited article-autonomous vehicles and connected vehicle systems: flow and operations considerations," *Transportation Science*, vol. 50, no. 4, pp. 1140–1162, 2016.

[13] Y. Ye, X. Zhang, and J. Sun, "Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 155–170, 2019.

[14] H. Yang and H. Rakha, "Feedback control speed harmonization algorithm: methodology and preliminary testing," *Transportation Research Part C: Emerging Technologies*, vol. 81, pp. 209–226, 2017.

[15] W. Jiekang, W. Zhijiang, M. Xiaoming, W. Fan, T. Huiling, and C. Lingming, "Risk early warning method for distribution system with sources-networks-loads-vehicles based on fuzzy C-mean clustering," *Electric Power Systems Research*, vol. 180, p. 106059, 2020.

[16] H. Yao, J. Cui, X. Li, Y. Wang, and S. An, "A trajectory smoothing method at signalized intersection based on individualized variable speed limits with location optimization," *Transportation Research Part D: Transport and Environment*, vol. 62, pp. 456–473, 2018.

[17] Y. Zhou, M. E. Cholette, A. Bhaskar et al., "Optimal vehicle trajectory planning with control constraints and recursive implementation for automated on-ramp merging," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3409–3420, 2018.

[18] D. Ma, Y. Han, F. Qu, and S. Jin, "Modeling and analysis of car-following behavior considering backward-looking effect," *Chinese Physics B*, vol. 30, no. 3, Article ID 034501, 2021.

[19] M. Zhu, X. Wang, and Y. Wang, "Human-like autonomous car-following model with deep reinforcement learning," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 348–368, 2018.

[20] M. Zhu, X. Wang, A. Tarko, and S. e. Fang, "Modeling car-following behavior on urban expressways in Shanghai: a naturalistic driving study," *Transportation Research Part C: Emerging Technologies*, vol. 93, pp. 425–445, 2018.

[21] X. Wang, R. Jiang, L. Li, Y. Lin, X. Zheng, and F. Y. Wang, "Capturing car-following behaviors by deep learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 910–920, 2019.

[22] Y. Xie, H. Zhang, N. H. Gartner, and T. Arsava, "Collaborative merging strategy for freeway ramp operations in a connected and autonomous vehicles environment," *Journal of Intelligent Transportation Systems*, vol. 21, no. 2, pp. 136–147, 2017.

[23] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[24] B. Hellinga and M. Mandelzys, "Impact of driver compliance on the safety and operational impacts of freeway variable speed limit systems," *Journal of Transportation Engineering*, vol. 137, no. 4, pp. 260–268, 2011.

[25] C. Letter and L. Elefteriadou, "Efficient control of fully automated connected vehicles at freeway merge segments," *Transportation Research Part C: Emerging Technologies*, vol. 80, pp. 190–205, 2017.

[26] X. Hu and J. Sun, "Trajectory optimization of connected and autonomous vehicles at a multilane freeway merging area," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 111–125, 2019.

[27] C. Zhang, N. R. Sabar, E. Chung, A. Bhaskar, and X. Guo, "Optimisation of lane-changing advisory at the motorway lane drop bottleneck," *Transportation Research Part C: Emerging Technologies*, vol. 106, pp. 303–316, 2019.

[28] X. Qu, Y. Yu, M. Zhou, C.-T. Lin, and X. Wang, "Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: a reinforcement learning based approach," *Applied Energy*, vol. 257, Article ID 114030, 2020.

[29] C. Lu, H. Chen, and S. Grant-Muller, "Indirect reinforcement learning for incident-responsive ramp control," *Procedia-Social and Behavioral Sciences*, vol. 111, pp. 1112–1122, 2014.

[30] J. A. Laval and C. F. Daganzo, "Lane-changing in traffic streams," *Transportation Research Part B: Methodological*, vol. 40, no. 3, pp. 251–264, 2006.

[31] J. Jin and X. Ma, "A multi-objective agent-based control approach with application in intelligent traffic signal system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3900–3912, 2019.

[32] S. Tiwari and N. Sharma, "Q-Learning approach for minutiae extraction from fingerprint image," *Procedia Technology*, vol. 6, pp. 82–89, 2012.

[33] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[34] B.-N. Wang, Y. Gao, Z.-Q. Chen, J.-Y. Xie, and S.-F. Chen, "A two-layered multi-agent reinforcement learning model and algorithm," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1366–1376, 2007.

[35] C. Guestrin, M. Lagoudakis, and R. Parr, "Coordinated reinforcement learning," in *Proceedings of the ICML 2002*, vol. 2, pp. 227–234, New South Wales, Australia, July 2002.

[36] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[37] C. Yu, X. Wang, X. Xu et al., "Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 2, pp. 735–748, 2019.

[38] Z. Lu, T. Fu, L. Fu, S. Shiravi, and C. Jiang, "A video-based approach to calibrating car-following parameters in VISSIM for urban traffic," *International journal of transportation science and technology*, vol. 5, no. 1, pp. 1–9, 2016.

[39] S. M. P. Siddharth and G. Ramadurai, "Calibration of VISSIM for Indian heterogeneous traffic conditions," *Procedia-Social and Behavioral Sciences*, vol. 104, pp. 380–389, 2013.

[40] Z. Qu, Z. Pan, Y. Chen, X. Wang, and H. Li, "A distributed control method for urban networks using multi-agent reinforcement learning based on regional mixed strategy Nash-equilibrium," *IEEE Access*, vol. 8, pp. 19750–19766, 2020.

[41] C. Wang, J. Zhang, L. Xu, L. Li, and B. Ran, "A new solution for freeway congestion: cooperative speed limit control using distributed reinforcement learning," *IEEE Access*, vol. 7, pp. 41947–41957, 2019.

[42] B. Khondaker and L. Kattan, "Variable speed limit: a microscopic analysis in a connected vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 146–159, 2015.

[43] L. Ruimin, Y. Zhen, and L. Bin, "Optimal control and simulation of hard shoulder running on highways," *Journal of System Simulation*, vol. 30, no. 3, p. 1036, 2018.