RESEARCH ARTICLE

Proximal policy optimization learning based control of congested freeway traffic

Shurong Mo1 | Nailong Wu1,2 | Jie Qi1,2 | Anqi Pan1 | Zhiguang Feng3 | Huaicheng Yan4 | Yueying Wang5

1College of Information Science and Technology, Donghua University, Shanghai, China
2Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai, China
3College of Automation, Harbin Engineering University, Harbin, China
4School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
5School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China

Correspondence
Nailong Wu, College of Information Science and Technology, Donghua University, Shanghai 201620, China.
Email: nathan_wu@dhu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 62173084, 52101346, 22Z032003162, 62303108

Abstract
In this paper, a delay compensation feedback controller based on reinforcement learning is proposed to adjust the time interval of the adaptive cruise control (ACC) vehicle agents in the traffic congestion by introducing the proximal policy optimization (PPO) scheme. The high-speed traffic flow is characterized by a two-by-two Aw Rasle Zhang nonlinear first-order partial differential equations (PDEs). Unlike the backstepping delay compensation control, the PPO controller proposed in this paper consists of the current traffic flow velocity, the current traffic flow density and the previous one step control input. Since the system dynamics of the traffic flow are difficult to be expressed mathematically, the control gains of the three feedback can be determined via learning from the interaction between the PPO and the digital simulator of the traffic system. The performance of Lyapunov control, backstepping control and PPO control are compared with numerical simulation. The results demonstrate that PPO control is superior to Lyapunov control in terms of the convergence rate and control efforts for the traffic system without delay. As for the traffic system with unstable input delay value, the performance of PPO controller is also equivalent to that of backstepping controller. Besides, PPO is more robust than backstepping controller when the parameter is sensitive to Gaussian noise.

KEYWORDS
adaptive control, adaptive cruise control, input delay, proximal policy optimization, traffic flow

1 INTRODUCTION

Traffic congestion is one of the most persistent problems of society today. Therefore, how to alleviate traffic congestion has always been a hot research topic in the field of traffic control, which has attracted widespread attentions.\(^1\) On the freeway, when the density of traffic flow reaches a high level, the traffic flow will decrease, and the density and velocity of traffic flow will oscillate, which is called the traffic wave,\(^2\) that is, the stop-and-go wave. Traffic waves not only cause traffic jams and reduce driving comfort,\(^3,4\) but also lead to increased fuel consumption.\(^5,6\) To analyze this phenomenon and study the spatiotemporal distribution characteristics of the traffic flow, macroscopic traffic modeling is used, which is particularly well suited to control design since it describes the overall spatiotemporal dynamics of the traffic flow on a freeway segment. The state variables of the macroscopic model are continuous in time and space, and they can be
measured and be regulated via modern traffic management system. Lighthill, Whitham, and Richards\textsuperscript{7,8} proposed the LWR traffic flow model, which represents a conservation law of traffic density and describes the formation and propagation of traffic density waves on the road. However, the oscillatory, unstable behaviors observed in the stop-and-go traffic require a higher-order model rather than the static LWR model, for example, the Payne-Whitham (PW) model, the Aw-Rascle-Zhang (ARZ) model. The PW model incorporates a term that estimates the state of the traffic flow ahead to represent the driver’s estimation of the traffic condition ahead and a relaxation term that represents the time required for the current traffic flow speed to reach the equilibrium speed. However, there is a serious problem: the PW model is a quasi-fluid model, and the difference between the traffic flow model and the fluid model is that there is no conservation of momentum in the traffic flow model. Its information propagation speed is greater than the speed of the traffic flow, which indicates that the front car is affected by the rear car, which violates the anisotropy of the traffic flow. The second-order ARZ model proposed by Aw, Rascle, and Zhang\textsuperscript{9,10} adds another partial differential equation (PDE) for the velocity state, which leads to a nonlinear coupled hyperbolic PDE system. In the literature,\textsuperscript{11} the authors investigated the extent to which the second-order ARZ traffic models could improve the prediction accuracy compared with the first-order LWR model.

The utilization of second-order PDE traffic flow models becomes imperative as they adeptly capture the dynamics of congested traffic conditions, such as stop-and-go traffic.\textsuperscript{12,13} boundary control designs have been recently developed for such systems.\textsuperscript{12-17} In the literature,\textsuperscript{13} the authors developed the full-state feedback boundary control to reduce two-lane traffic congestion in a freeway segment. In the literature,\textsuperscript{16} boundary feedback control of the inhomogeneous ARZ model was proposed using the infinite-dimensional backstepping method, and the boundary feedback control law of the on-ramp flow was designed to stabilize the traffic flow in the upstream of the freeway. The boundary control usually relies on static road infrastructure to regulate traffic flow, such as ramp metering and varying speed limit. With the development of adaptive cruise control (ACC) technology, in-domain control has been applied to many traffic systems.\textsuperscript{18-20} By controlling the time gap of ACC vehicles in the literature,\textsuperscript{18} the oscillation of traffic flow on the road was decreased, and the maximum capacity of the freeway could be used as much as possible to stabilize the density and velocity of the traffic flow and to alleviate traffic congestion. A Lyapunov state feedback controller was designed in the literature.\textsuperscript{19} By linearizing and diagonalizing the ARZ model, an analysis was conducted on the characteristics of unstable mixed traffic flow, and a stable target system was designed. By comparing the two systems, the state feedback controller was obtained. The Lyapunov stability analysis method was applied to prove the exponential stability of the closed-loop system in the $C^1$ norm. This controller eliminated the traffic flow fluctuations by controlling the time gap of ACC vehicles and achieved significant optimization of the traffic flow system in terms of the fuel consumption, total travel time, and travel comfort.

The input delay of the traffic system can be derived from the transmission of information and the reaction time of a driver,\textsuperscript{21} which is another reason for the traffic waves. The study in the literature\textsuperscript{22} illustrates the contradictory relationship between the delay robustness and the convergence rate of the hyperbolic PDE system, and the controller is also designed to balance the convergence rate and delay robustness of the target system by using the backstepping method. In the literature,\textsuperscript{21} a backstepping compensation of the in-domain controller was proposed for a same traffic model as that in the paper, which used a transport equation to represent the input delay, and the $L_2$ norm exponential stability was proven. If the $L_2$ norm of the traffic flow model converges over time, it implies that the traffic flow system is stable under certain conditions. Stability is a crucial property in the study of traffic flow models, as practical transportation systems aim for stable traffic rather than congestion or instability. Based on the linearized ARZ model, the Backstepping method was applied to establish an explicit feedback delay compensation controller comprising traffic velocity, traffic density, and historical states of the actuator. The original closed-loop system was transformed into a stable target system through the Backstepping transformation. The controller was obtained by solving the kernel function of Backstepping.

The Lyapunov controllers and the Backstepping controller have achieved certain theoretical achievements. The Lyapunov control stabilizes the delay-free system and the Backstepping control shows excellent performance in delay compensation. However, there are still some unsolved problems, such as Lyapunov control is not robust to time delay and the Backstepping control is not robust to parameter perturbation. Of late, reinforcement learning (RL)-based methods have been introduced to solve these problems. The RL has been used in traffic.\textsuperscript{24-27} In the literature,\textsuperscript{24} the author used RL to compare macro and micro traffic controls and verified through simulation software that the macro traffic control could achieve better results than the micro traffic control. In the literature,\textsuperscript{27} a ramp metering control law based on the RL-based algorithm was proposed and compared with the boundary controller designed
by the Backstepping method and Lyapunov method, indicating the potential of this RL algorithm in the field of traffic control.

In this paper, we develop a RL-based controller which is trained by the proximal policy optimization (PPO) for a traffic system composed of $2 \times 2$ PDEs with input delay and show the controller is robust to parameter disturbance. PPO$^2$ is chosen over other policy gradient methods, such as trust policy optimization$^2$ or deep deterministic policy gradient (DDPG) algorithm,$^3$ because PPO is more robust to the step size and shows higher performance and lower computational complexity. PPO uses an objective function that places a limit on the size of policy update, enabling it to learn policies that work well with smaller batch sizes, requiring less data overall. PPO controls the size of the policy update to prevent it from moving too far from the previous policy, leading to less variability in training and more stable results. Unlike the boundary ramp metering control in the literature,$^2$ where the PPO trains boundary inputs, that is, one value per step, in this study, the time-gap of the ACC-equipped vehicle is used as manipulation values which are in-domain inputs. It is difficult for the PPO to train so many inputs of all the in-domain points. Inspired by the Backstepping controller but much simpler than that of the Backstepping segmented controller, we take a weighted sum of three feedbacks, the feedback from the current two states and the feedback from the previous step input. The gains of the controller are trained through the interaction between the PPO and the numerical traffic system simulator.

We design numerical simulation experiments to compare the performance of the Lyapunov control, the Backstepping control and the PPO control. For a delay-free system, the PPO control has better convergence speed and less control effort than the Lyapunov control. For an input delay system, the Lyapunov control cannot stabilize the system, while the PPO and the Backstepping control show similar control performance. When the delay value is mismatched, that is, the delay applied in control is different from the actual delay value, the performance of the PPO control is comparable to that of the Backstepping control. However, the PPO control can effectively stabilize the system in the presence of parameter disturbance, thus demonstrating its robustness against such disturbances.

The contributions of this paper can be summarized as follows. First, we proposed a PPO-based controller via training three normal distributed control gains to stabilize a delayed traffic flow system containing ACC-equipment vehicles. Second, although the feedback form of the PPO controller is much simpler than that of the Backstepping controller, the performance of the PPO controller in compensating for input delay is comparable to that of the Backstepping delay compensator.$^2$ Third, the proposed PPO control is robust to parameter disturbance, in which case the Backstepping method fails.

The rest of this paper is arranged as follows. Section 2 introduces the traffic model. Section 3 proposes the controller based on the proximal policy optimization. Section 4 includes numerical simulation experiments and comparative analysis. Section 5 summarizes the paper.

## 2 | MACROSCOPIC TRAFFIC MODEL

### 2.1 | The ARZ model

An ARZ-type traffic model in a highway stretch $L$ proposed in the literature$^{23}$ is utilized to describe the traffic flow dynamics consisting of both adaptive cruise control-equipped (ACC-equipped) and manual vehicles. The percentage of ACC-equipped vehicles with respect to the total number of vehicles is denoted by $\alpha \in [0, 1]$. The state variables of the model are the traffic density $\rho(x, t)$ and the traffic speed $v(x, t)$, both defined in domain $(x, t) \in [0, L] \times \mathbb{R}^+$ where $t$ is time, and $x$ is the spatial variable at the concerned highway. The traffic flow model with state variables traffic density $\rho(x, t)$ and traffic velocity $v(x, t)$ in domain $x \in [0, L], t \in \mathbb{R}^+$ is expressed as:

\[
\rho_t(x, t) = -\rho_x(x, t)v(x, t) - \rho(x, t)v_x(x, t),
\]

\[
v_t(x, t) = -\rho(x, t)\frac{\partial V_{\text{mix}}(\rho(x, t), h_{\text{acc}}(x, t - D))}{\partial \rho}v_x(x, t) + \frac{V_{\text{mix}}(\rho(x, t), h_{\text{acc}}(x, t - D)) - v(x, t)}{\tau_{\text{mix}}} - v(x, t)v_x(x, t),
\]

where $\tau_{\text{mix}}$ is the time delay in the mixing process.
and the boundary conditions are as follow:

\[ v(0, t) = q_{in} / \rho(0, t), \]  
\[ v(L, t) = V_{mix}(\rho(L, t), h_{acc}(L, t - D)) - v(L, t) / \tau_{mix}, \]  

where

\[ \tau_{mix}(\alpha) = \frac{1}{\alpha \tau_{acc} + (1 - \alpha) \tau_{m}}, \quad 0 \leq \alpha \leq 1, \]  

is time constant for a mixture traffic reflecting how quickly vehicles adjust their velocity to the desired value, which depends on both time constant \( \tau_{acc} \) of ACC vehicles and time constant \( \tau_{m} \) of manual vehicles. \( \tau_{mix} \) also depends on \( \alpha \), the percentage of ACC vehicles to total vehicles. Parameter \( q_{in} > 0 \) is a constant external inflow. The traffic flow observes the mass conservation law, which is described by Equation (1) and momentum equation derived from the velocity dynamics of the ARZ model, which is described by Equation (2). To stabilize the traffic flow so as to mitigate the stop-to-go wave in the congested regime, we will design a control input actuating on the ACC-equipped vehicle by manipulation of the time-gap of ACC-equipped vehicles from their leading car, denoted by \( h_{acc}(x, t) > 0 \). The ACC-equipped vehicles are uniformly distributed. This controller aims to control the ACC-equipped vehicles at specific time and space intervals throughout the entire targeted road section. The ACC-equipped vehicles on the road section upload their positioning data \( x \) at time point \( t \) which is then used by the controller to calculate the time gap \( h_{acc}(x, t) \) that the vehicle maintains. The penetration rate \( \alpha \) of ACC-equipped vehicles is determined by the number of ACC-equipped vehicles linked to the controller and the total number of vehicles traveling on the respective road section. The total vehicle count and input flow rate \( q_{in} \) at the entrance can be determined by accessing the entry and exit counts of the road section. Input delays sometimes occur due to information transmission or driver’s response to control instruction. Hence, we consider time delay \( D > 0 \) in the model and design control which can compensate the input delay.

Consider the mixed traffic containing both manual and ACC-equipped vehicles, which gives the following equilibrium speed profile:

\[ V_{mix}(\rho, h_{acc}) = \frac{1}{h_{mix}(h_{acc})} \left( \frac{1}{\rho} - l \right), \quad \rho > \rho_{min}, \]  

with mixed time gap \( h_{mix} \) defined as follows

\[ h_{mix}(h_{acc}) = \frac{\alpha + (1 - \alpha) \frac{\tau_{acc}}{\tau_{m}}}{\alpha + (1 - \alpha) \frac{\tau_{acc}}{\tau_{m}} h_{acc}}, \]  

where parameter \( l > 0 \) denotes the average effective vehicle length, and \( h_{m} > 0 \) is the time gap of manual vehicles. Obviously, \( h_{mix} = h_{m} \) when \( \alpha = 0 \); \( h_{mix} = h_{acc} \) when \( \alpha = 1 \). Figure 1 shows the relationship of mixed time gap \( h_{mix} \) and the percentage of ACC vehicles to total vehicles \( \alpha \) in both cases of \( h_{acc} > h_{m} \) and \( h_{acc}' < h_{m} \).

### 2.2 | Properties of the model

The effect of the penetration rate of ACC-equipped vehicles is incorporated via the mixed relaxation time (5) and the mixed time gap (7), with \( h_{min} < \min\{h_{m}, h_{acc}\} \leq h_{mix} \leq \max\{h_{m}, h_{acc}\} < h_{max} \), where \( h_{min} \) and \( h_{max} \) denote the minimal and the maximal value of \( h \), respectively, for all \( \alpha \in [0, 1] \).

Figure 2 shows the relationship of mixed traffic flow \( Q \) and the traffic density \( \rho \), where \( \rho_{c} > 0 \) is the critical value of the traffic density, that is, when the density is greater than it, the traffic becomes congested.
The traffic flow $Q_{h_{\text{mix}}}$ is defined as

$$Q_{h_{\text{mix}}} (\rho) = \begin{cases} v_f \rho, & 0 \leq \rho \leq \rho_c, \\ \frac{1-\rho}{h_{\text{mix}}}, & \rho_c < \rho \leq \frac{1}{l}. \end{cases}$$

(8)

where $\rho_c = 1/(l + h_{\text{mix}} v_f)$ and $v_f$ is free-flow speed. Since we know $V_{\text{mix}}(\rho, h_{\text{acc}}) = Q_{h_{\text{mix}}} / \rho$, $0 < V_{\text{mix}}(\rho, h_{\text{acc}}) < v_f$ for all $\alpha \in [0, 1]$ and $\rho_{\text{min}} < \rho < (1/l)$ is guaranteed by $\max\{h_{\text{acc}}, h_{m}\} \leq h_{\text{max}}$ and $\min\{h_{\text{acc}}, h_{m}\} \geq h_{\text{min}}$. Hence, Equation (6) defines a reasonable fundamental diagram for mixed traffic in congested conditions. Using an analysis method similar to the one employed in the literature, one can find that system (1)–(4) is anisotropic.

Given an inflow $q_{\text{in}}$ and a constant time gap input $h_{\text{acc}}$, we find the equilibrium of system (1)–(4) as follows, which is the same as that in the literature,$^{19}$

$$\bar{v} = \frac{l}{(1/q_{\text{in}}) - \bar{h}_{\text{mix}}}, \quad \bar{\rho} = \frac{1}{l + \bar{h}_{\text{mix}} \bar{v}}.$$  

(9)

with mixed time gap

$$\bar{h}_{\text{mix}} = \frac{\alpha + (1-\alpha) \frac{h_{\text{acc}}}{\tau_m}}{\alpha + (1-\alpha) \frac{h_{\text{acc}}}{\tau_m} \bar{h}_{\text{acc}}}.$$  

(10)
The open-loop system is unstable to the equilibrium (9). The Lyapunov feedback control proposed in the literature\textsuperscript{19} can stabilize the linearized system without delay, and the Backstepping control designed in the literature\textsuperscript{23} can stabilize the linearized system with input delay. In the next section, we will design a RL-based controller under the states feedback and historical input feedback framework, where its control policy is trained by the PPO.

3 | PROPOSED METHOD

3.1 | Interactions between controller and environment

PPO is a model-free, policy-gradient method, whose hyperparameters are robust to multiple tasks. Also, the PPO algorithm exhibits high performance and low computational complexity\textsuperscript{28} compared to other policy gradient methods. The explicit knowledge of the traffic model is not required for PPO-based control approaches. In the paper, we simulate traffic dynamics by discretizing the nonlinear model by the finite difference method and use this numerical simulator as the environment with which the PPO interacts. The solution $\rho(x, t)$ and $v(x, t)$ to (1)–(4) is discretized in domain $[0, L] \times [0, T]$. The discretization space step is denoted by $\Delta x = L/M$ and time step is denoted by $\Delta t = T/N$, which are chosen such that the Courant–Friedrichs–Lewy condition is met, namely $\Delta t \geq \max |\lambda_{1,2}|$, where $\lambda_{1,2}$ are the characteristic speed of the model (1)–(4). Hence, the discretized states and input at time $t$ can be written as:

$$\rho^t = [\rho(x_0, t), \rho(x_1, t), ..., \rho(x_m, t)],$$

$$v^t = [v(x_0, t), v(x_1, t), ..., v(x_m, t)],$$

$$h^{t-1}_{\text{acc}} = [h_{\text{acc}}(x_0, t - 1), h_{\text{acc}}(x_1, t - 1), ..., h_{\text{acc}}(x_m, t - 1)],$$

where $x_i = \Delta x \cdot i$ for $i = 0, ..., M$ and $t = \Delta t \cdot j$ for $i = 0, ..., N$.

To avoid too many outputs of the policy network which may cause the training of the weights to diverge, we employ the PPO to learn the control gains of the feedback, rather than all the inputs distributed in domain. The control form is inspired by the Backstepping delay compensator in the literature,\textsuperscript{23} which consists of three components:

$$h^t_{\text{acc}} = -\eta^t_1(h^{t-1}_{\text{acc}} - \bar{h}_{\text{acc}}) - \eta^t_2(v^t - \bar{v}) + \eta^t_3(\rho^t - \bar{\rho}).$$

The control includes the feedback of the current velocity and density states and the previous one step input, which is much simpler in form than that of the Backstepping control.\textsuperscript{23}

Consider an infinite-horizon discounted Markov decision process, defined by the tuple $(S, A, P, R)$ representing the dynamics of the traffic flow.

$S$: Space of states of the environment, which are the input of the policy network. Let $s^t \in S$ be a state at time $t$ defined as

$$s^t = [\rho^t, v^t, h^{t-1}_{\text{acc}}].$$

$A$: Space of actions. At each time $t$, the control policy choose an action $a^t \in A$, which is the input of the environment. Define $a^t$ as follows:

$$a^t = [\eta^t_1, \eta^t_2, \eta^t_3].$$

Policy: Let $\pi(a|s)$ denote a policy function $\pi : S \times A \rightarrow [0, 1]$, which represents the probability of taking an action $a \in A$ given state $s$. In the control learning, we suppose the policy functions are the three normal distributions of three gains, respectively:

$$\eta^t_1 \sim \mathcal{N}(\mu^t_1, \sigma^t_1).$$
\[ \eta_2' \sim \mathcal{N}(\mu_2', \sigma_2'), \quad \mu_2' = \mathfrak{f}_{DNN}(\theta); \]
\[ \eta_3' \sim \mathcal{N}(\mu_3', \sigma_3'), \quad \mu_3' = \mathfrak{f}_{DNN}(\theta); \]

where \( \mu_i' \) and \( \sigma_i' \) are the mean and the variance for \( \eta_i' \), respectively. The policy-based method applied, we defined a policy network \( f_{DNN} \) to approximate the policy functions:

\[ [\mu', \sigma'] = f_{DNN}(a'|s'; \theta), \quad \mu' = [\mu_1', \mu_2', \mu_3'], \quad \sigma' = [\sigma_1', \sigma_2', \sigma_3'], \]

The outputs of the network are the means and variances of the three gains. In practice, the action is drawn from sampling and then the control input is calculated according to (14).

\textbf{P}(s'|s', a'): The state-transition probability. It defines the probability that the state \( s' \) transforms to \( s'^{+1} \) under action \( a' \). The randomness of the state transition is from the environment. In the paper, the environment is a simulator of the traffic model which can be written as:

\[ s'^{+1} = f_{traffic}(s', h_{acc} - d), \]

with \( d = D/\Delta t \) representing the input delay. If \( D > 0 \), the control input will not work until \( d \) units of time \( \Delta t \). The function \( f_{traffic} \) represents the numerical simulated dynamics for the temporal evolution of the traffic model (1)-(4). Although the traffic dynamics is deterministic, we still generally express the state transition process by a Markov decision process probability

\[ s'^{+1} \sim \text{P}(s'|s', a'). \]

The relation between \( h_{acc}^{-d} \) and \( a' \) can be found in (14), (16), and (20).

### 3.2 Policy evaluation

\( R(s') \): The total discounted return from time \( t \) to \( t + j \) can be expressed as:

\[ R(s') = r' + \gamma r'^{+1} + \ldots + \gamma^j r'^{+j} = \sum_{j=0}^{j} \gamma^j r'^{+j}, \]

where \( \gamma \in (0, 1] \) is the discount factor that encodes the importance of future rewards, and \( r' \) is the reward depending on \( s' \) as follows:

\[ r(s') = - \sum_{i=0}^{N} \left[ \frac{p(x_i, t) - \bar{p}}{\bar{p}} \right]^2 + \left[ \frac{v(x_i, t) - \bar{v}}{\bar{v}} \right]^2. \]

Thus, given \( s' \), the return \( R(s') \) depends on actions \( a', a'^{+1}, a'^{+2}, \ldots \) and states \( s', s'^{+1}, s'^{+2}, \ldots \)

\textbf{Value functions:} First we define the action-value function for policy \( \pi \) given the current state \( s' \) and action \( a' \)

\[ Q_{\pi}(s', a') = \mathbb{E}_{a'}[R(s')|s', a'], \]

which represents the expected total discounted return on actions \( a', a'^{+1}, a'^{+2}, \ldots \) and states \( s', s'^{+1}, s'^{+2}, \ldots \) The probability of actions obeys the policy function and the randomness of the states come from the state-transition probability. Accordingly, we define the optimal action-value function as
Then, the state-value function is

\[ V_\pi(s^t) = \mathbb{E}\{ Q_\pi(s^t, a^t) \}. \]  

which is the expected value of action-value function on all actions in \( \mathcal{A} \). The advantage function, reflecting how advantageous the action \( a^t \) is compared to the expectation of all actions in \( \mathcal{A} \), is defined as:

\[ A_\pi(s^t, a^t) = Q_\pi(s^t, a^t) - V_\pi(s^t), \]  

3.2.1 | Actor–critic method

The Actor–Critic method combines the policy learning and value learning, where two neural networks are used to approximate the policy function and the state-value function, respectively.

3.3 | Proximal policy optimization with clip

First, we consider the policy network (actor network), denoted by \( \pi_\theta(a^t|s^t) \), where \( \theta \) indicates the parameters of the neural network which is updated by using the PPO method. The PPO allows for continuous states and action space, which trains a continuous-valued stochastic control policy. There are two policy networks running in the learning process simultaneously, one is training network, denoted by \( \pi_\theta \); the other is interaction network denoted by \( \pi_\theta^{old} \). The interaction network interacts with the environment so as to collect the data of a group of actions and the corresponding states. These data are necessary for the training network to update its parameters. After collecting enough data, network \( \pi_\theta^{old} \) updates its parameters according to \( \pi_\theta(a^t|s^t) \).

The policy gradient is estimated via optimizing:

\[ \theta' = \arg \max_{\theta} J(\theta), \]  

where the objective function\(^{36} \) is

\[ J(\theta) = \mathbb{E}[\min\{ g'(\theta), \text{clip}(g'(\theta), 1 - \epsilon, 1 + \epsilon)\} A_\pi(s^t, a^t)]. \]  

with

\[ g'(\theta) = \frac{\pi_\theta(s^t|a^t)}{\pi_\theta^{old}(s^t|a^t)}, \]  

\[ \text{clip}(g'(\theta), 1 - \epsilon, 1 + \epsilon) = \begin{cases} 
1 - \epsilon, & g'(\theta) < 1 - \epsilon, \\
1 - \epsilon \leq g'(\theta) \leq 1 + \epsilon, & g'(\theta), \\
1 + \epsilon, & g'(\theta) > 1 + \epsilon. 
\end{cases} \]  

The first term of (31) inside the min is \( g' \). The second term, \( \text{clip}(g'(\theta), 1 - \epsilon, 1 + \epsilon) \)\(^{30} \) clips the probability ratio, which removes the incentive for moving \( g' \) outside of the interval \([1 - \epsilon, 1 + \epsilon]\) with the hyperparameter \( \epsilon \). The clipped objective (31) improves the convergence of the algorithm. The importance sampling is applied to increase sample efficiency, in which the expectation of \( \pi_\theta \) is computed by sampling the old policy \( \pi_\theta^{old} \).

To train the policy network, the value function \( V \) is required to supervise the update of the policy parameters. As the value function is unknown, we use a critic network to approximate to the state-value function as follows:

\[ V_\pi(s^t) = g_{\text{DNN}}(s^t, \omega), \]
where $\omega$ is the trainable parameter of the network, which is trained by minimizing the loss function defined as:

$$
L(\omega) = \frac{1}{\epsilon} \sum_{i=1}^{\epsilon} (R(s^i) - V_{\pi}(s^i))^2.
$$

A deep neural network (DNN) $g_{DNN}(s', \omega)$ is used as the critic.

The parameters of the two networks are updated by Adam optimizer. Adam optimizer performs well when dealing with unstable objective functions. In this paper, the objective functions involve both control and traffic congestion, which are usually very complex and high-dimensional, and Adam optimizer can solve these problems relatively well. Besides, the adaptive property of Adam optimizer can prevent gradient explosion or vanishing problems during the optimization process, thereby improving the training effect and efficiency. This is particularly important in this paper because traffic problems involve many complex and uncertain factors, which may lead to gradient explosion or vanishing problems during the training process.

Figure 3 illustrates the learning process of the PPO-based control. At each time $t = 0, 1, ..., M$, input state $s^t$ into the policy network $\pi_{\theta_{\text{old}}}$, which results in the distributions for the three control gains. Sampling the control gains from the distributions, namely action $a^t$, the control $h_{\text{acc}}^{t-d}$ is calculated from equation (14). The delayed input $h_{\text{acc}}^{t-d}$ is fed to the environment, namely, the discretized PDE model, which results in the states $s^{t+1}$. Iterating the process $E = \Delta T/\Delta t$ times, all the $M$ states and actions are collected and then input into the policy network $\pi_{\theta}$ for training its parameters. After update the parameters of the network $\pi_{\theta}$ using the PPO gradient estimation, the new parameter of network $\pi_{\theta}$ is transmitted to network $\pi_{\theta_{\text{old}}}$. The implementation of the PPO-based controller is summarized in Algorithm 1.

4 | SIMULATION RESULTS

4.1 | Traffic simulation

The traffic model presented in Section 2 with the same parameter values as in Reference 19 was used in the simulations, as Table 1 shows.
Algorithm 1. Proximal policy optimization procedure

1: Initialize parameters for states, actors, networks and $N = T/\Delta t$. 
2: for $i = 0, 1, 2, 3, ...$ do 
3: for $j = 0$ to $N$ do 
4: Set $t' = i \cdot N + j$. 
5: Update $s'$ to policy $\pi_{\theta_{ol}}$ and sampling to get $a'$. 
6: Calculate $h_{acc}'$ from (14) and push $h_{acc}'$ to the delay buffer. 
7: Get $h_{acc}' - d$ from the delay buffer and get state data from traffic system $s_{t+1} = f_{\text{traffic}}(s', h_{acc}' - d)$. 
8: end for 
9: Collect set of trajectories driven by policy $\pi_{\theta_{ol}}$. 
10: Compute total discounted reward $R$. 
11: Compute advantage estimates, $A_\pi$ from critic $V_\pi$. 
12: Update the actor, $\theta$ by Adam optimizer. 
13: Update the critic, $\omega$ by Adam optimizer. 
14: $\theta_{ol} \leftarrow \theta$ 
15: end for

| TABLE 1 Parameters of system. |
|-------------------------------|
| $L = 1000$ m                  |
| $l = 5$ m                     |
| $q_{in} = 1200$ veh/h         |
| $\alpha = 0.15$              |
| $\tau_{acc} = 2$ s           |
| $\tilde{h}_{acc} = 1.5$ s    |
| $\tau_m = 60$ s              |
| $h_m = 1$ s                  |

The chosen parameters are reasonable for a traffic model. Let $h_m$ be lower than $1.2$ s to reflect that drivers follow a preceding vehicle at smaller time gaps in the congested traffic compared to the case of light traffic conditions. The operating point of the traffic system, namely, the steady-state value of the mixed time gap according to (10), is set to $\tilde{h}_{acc} = 1.5$ s, which is a little larger than that of manual driving in heavy traffic. Traffic flow undergoes a transition from smooth to stop-and-go waves before eventually becoming completely blocked. Thus, the aim of this paper is to control traffic flow during the stop-and-go wave phase, to increase the maximum vehicle capacity of the road section, optimize the driving experience, and lower energy consumption. When the traffic density of the target road section reaches a certain level, the traffic flow system will create stop-and-go waves, which can cause the driver to become increasingly anxious due to frequent acceleration and deceleration that increase the chances of accidents and fuel consumption. To mitigate these effects, this paper seeks to stabilize the stop-and-go waves which display a characteristic sine wave. Thus, the model only considers a fixed input flow rate $q_{in} = 1200$ veh/h.

The initial condition is set to $\rho(\mathbf{x}, 0) = 10 \cos(8\pi \mathbf{x}/L)$ and $v(\mathbf{x}, 0) = q_{in}/\rho(\mathbf{x}, 0)$ which deviates from the equilibrium value and shows a sinusoidal shape to simulate stop-and-go initial traffic conditions.

### 4.2 Network structure

The policy-actor network and the value-critic network both have eight layers. For the two networks, the first layer is input layer, and the other layers are hidden layers. The input layers have 600 neurons, the second hidden layers have 1024 neurons, and the rear hidden layers contain 512 neurons. The outputs of the actor network are mean $\mu'$ and variance $\sigma'$ of the normal distribution that action $a'$ obeys. The outputs of the critic network are the state-defined in (15).

To ensure that the variance and mean are positive, the activation function of the last layer adopts the sigmoid function, and the activation function of the last critic network adopts the hyperbolic tangent function.

To avoid the variance being too small, we set the offset value to $10^{-6}$. The learning rate is 0.001. The system (1)–(4) is discretized with time step $\Delta t = 0.1$ s and spatial step $\Delta x = 5$ m. The update time is chosen as $T = 10$ s, and thus $M = 200$ and batch size $\epsilon = 100$ equals to $N = 100$. Each batch is trained 150 times to update the parameter. The numerical equipment is carried on CPU Intel(R) Core(TM)i9-7900X and GPU TITAN Xp.
4.3 Comparison results and analysis

In this subsection, the Lyapunov controller proposed in the literature\(^{19}\) and the Backstepping controller developed in Reference \(^{23}\) are used for comparative analysis with the proposed controller. All simulations are performed on nonlinear PDE systems. To check the robustness of the system (1)–(4) under different controllers, we consider the two situations where a Gaussian noise is added to parameter \(\alpha\) and mismatched delay value, respectively.

4.3.1 Learning process and reward

The convergence of the reward of the PPO controller and the Backstepping controller under different delay \(D\) is simulated, and Figure 4 shows the robustness of the PPO controller to delay \(D\), and the reward convergence curve is relatively consistent. In Figure 2, the control performance of the Backstepping controller varies greatly under different delay \(D\), but convergence is still guaranteed. The delay in the traffic flow system is mainly caused by measurement and vehicle response times. Typically, traffic flow data measurements take approximately 2 s. However, vehicle response times are unpredictable, and the controllers have good performance for high delay. Therefore, a longer response time 2 s will ensure that the controller performs well under more realistic conditions when simulating mixed traffic flow.

The learning curve of the PPO controller for the system with 4-s delay is shown in Figure 5A. The rewards \(r\) reflecting learning performance for five different situations is shown in Figure 5B. To avoid random perturbations due to sampling when the reward converges the maximum, we set a stop rule for the PPO: the algorithm stops as the reward is equal to or greater than \(-0.1\). As shown in Figure 5B, the system under the open-loop control cannot converge, while the reward of the PPO control converges quickly. Moreover, the PPO is robust to input delay because the reward curves of the PPO control for the system with delay and without delay almost coincide.

Figure 6A shows the relationship between the setting time \(t_s\) and \(\alpha\) for the system velocity to enter the 2\% error band. The penetration rate \(\alpha\) represents the proportion of ACC vehicles to the total number of vehicles, and the controller’s control variable is the time gap \(h_{acc}\) of ACC vehicles. When \(\alpha\) is 0, the system is an uncontrollable open-loop system. As \(\alpha\) increases, the number of ACC vehicles also increases. During the driving process, ACC vehicles use space more efficiently when decelerating and accelerating, thereby reaching the desired speed faster, and optimizing the flow and speed of the entire traffic flow. However, when the proportion of ACC vehicles reaches a certain level, the instability of such mixed traffic flow becomes more significant due to the significantly slower reaction speed of human-driven vehicles. In this case, the controller’s performance will decrease with the proportion of ACC vehicles. In the model of this paper, the mixed time gap \(h_{mix}\) is a function of \(\alpha\) by Equation (7), which is equivalent to limiting the controller output \(h_{acc}\). The smaller the \(\alpha\), the stronger the limiting effect. Figure 6B shows the convergence of rewards under different \(\alpha\), and an increase in \(\alpha\) leads to slower reward convergence.

**Figure 4** (A) The reward convergence for the proximal policy optimization controller with different delays. (B) The reward convergence for the Backstepping controller with different delays.
4.3.2 State evolution

The states evolution and the control input of the traffic system are shown in Figures 7 and 8. The simulation results of the traffic system are shown in Figures 7 and 8. Figure 7 illustrates the evolution of density \( \rho(x, t) \), velocity \( v(x, t) \) and control input \( h_{\text{acc}}(x, t) \) without delay under the PPO control and Lyapunov control. Comparing Figure 7A and D, it is observed that the PPO controller results in the state \( \rho \) converging much more quickly than the Lyapunov controller does, but as shown in Figure 7B, E, the state \( v \) converges a little faster under the Lyapunov controller than under the PPO controller. The reason for this phenomenon is that the Lyapunov control gives control priority to velocity, which suppresses density fluctuations by first eliminating velocity fluctuations. However, the PPO control allows the velocity and density of the traffic flow to converge simultaneously guided by the value function. It is observed from Figure 7C, F that the control effort of the PPO control is less than that of the Lyapunov control.

Figure 8 shows the evolution of the system subject to a 4-s input delay under the Lyapunov control, the Backstepping control and the PPO control. The results showed that the Lyapunov control cannot stabilize the 4-s input delayed system. The PPO control stabilizes the system with a 4-s input delay, whose converging rate is comparable to the Backstepping delay-compensator.

The closed-loop system with the Backstepping controller exhibits exponential stability in \( L_2 \) norm in the literature.\textsuperscript{23} The PPO control form \( (14) \) is inspired by the Backstepping controller, and the reward is defined as Equation \( (25) \) which consists of \( L_2 \) norm of states \( \rho \) and \( v \). In order to illustrate the state evolution more clearly, we plot the \( L_2 \) norm of the states under the PPO and the Backstepping controllers in Figure 9, with definition of the \( L_2 \) norm:

\[
||\bar{\rho}||_{L_2} = \sqrt{\frac{\sum_{i=0}^{n}(\rho(x_i, t) - \bar{\rho})^2}{n}},
\]

\[
||\bar{v}||_{L_2} = \sqrt{\frac{\sum_{i=0}^{n}(v(x_i, t) - \bar{v})^2}{n}},
\]
FIGURE 7  (A)–(C) The states and control input of a delay free system under the Lyapunov controller; (D)–(F) The state and control input of a delay free system under the proximal policy optimization control.

It can be observed from Figure 9 that the PPO control shows better convergence performance of the density compared to the Backstepping control, but worse convergence performance of the velocity. The reason is also that the target system of the Backstepping method is inspired by the closed-loop system with the Lyapunov control which gives control priority to velocity over density.

4.3.3 Performance evaluation

We apply three indexes to evaluate the performance of the closed-loop system under the different controllers. These indexes are total travel time (TTT), fuel consumption and driving comfort defined as follows:\(^2\)

\[
J_{TTT} = \int_0^T \int_0^L \rho(x, t) dx dt, \tag{38}
\]

\[
J_{\text{fuel}} = \int_0^T \int_0^L \xi(x, t) \cdot \rho(x, t) dx dt, \tag{39}
\]

\[
J_{\text{comfort}} = \int_0^T \int_0^L \left( a(x, t)^2 + a_t(x, t)^2 \right) \rho(x, t) dx dt, \tag{40}
\]

where

\[
\xi(x, t) = \max\{0, b_0 + b_1 v(x, t) + b_2 v(x, t) + b_3 v(x, t) a(x, t) \}, \tag{41}
\]

\[
a(x, t) = v_t(x, t) + v(x, t) v_x(x, t), \tag{42}
\]
FIGURE 8  (A)–(C) The states and control input of a system with a 4-s input delay under the Lyapunov control; (D)–(F) The states and control input of a system with a 4-s input delay under the Backstepping control; (G)–(I) The states and control input of a system with a 4-s input delay under the proximal policy optimization control.

FIGURE 9  The $L_2$ norm of the states errors between the actual value and the desired value for a system with a 4-s input delay under the proximal policy optimization and the Backstepping control, respectively.
TABLE 2  Performance.

| Performance indices | Open-loop | Backstepping | Proximal policy optimization |
|---------------------|-----------|--------------|-----------------------------|
| $J_{TTT}$           | $3.4 \times 10^7$ | $3.2 \times 10^7$ | 5.97% | 3.2 $\times 10^7$ | 5.97% |
| $J_{fuel}$          | $8.6 \times 10^5$ | $8.1 \times 10^5$ | 5.80% | 8.1 $\times 10^5$ | 5.80% |
| $J_{comfort}$       | $9.8 \times 10^5$ | $1.3 \times 10^5$ | 87.1% | $1.5 \times 10^5$ | 84.3% |

FIGURE 10  (A)–(C) The states and control input of a system with a 4-s input delay which is subject to parameter disturbance, that is, $\alpha \sim \mathcal{N}(0.15, 0.15^2)$ under the Backstepping control; (D)–(F) The states and control input of a system with a 4-s input delay which is subject to parameter disturbance, that is, $\alpha \sim \mathcal{N}(0.15, 0.15^2)$ under the proximal policy optimization control.

and the same values of the parameters are chosen as those in: $^2 b_0 = 25 \cdot 10^{-3}, b_1 = 24.5 \cdot 10^{-6}, b_3 = 32.5 \cdot 10^{-6}, b_4 = 125 \cdot 10^{-6}$ and $T = 300$ s.

As shown in the Table 2, all indexes are improved compared to the open-loop system. Only the data in the first 300 s are used to compute the indexes, considered that the open-loop system is unstable. Although the PPO control has a slightly lower index of comfort than the Backstepping control, it is comparable to the backstepping in other two indexes.

4.3.4  Robustness comparison via simulation

In the simulation, the pretrained PPO controller is used, which is trained using the data drawn from the model with $\alpha = 0.15$ and delay $D = 4$ s. We perform two kinds of numerical experiments. First, we compare the robustness of the PPO control and the Backstepping control to the parameter disturbance by adding a Gaussian noise $\mathcal{N}(0, 0.15^2)$ to $\alpha$. Second, simulated experiments with mismatched delay values are performed, that is, the delay value adopted by the controller is different from the actual delay.

Figure 10 shows the evolution of the states considering the parameter disturbance to $\alpha$ for a system with 4-s input delay under the Backstepping control and the PPO control. The Backstepping controller fails stabilize the system with
FIGURE 11 (A) The $L_2$ norm of the states errors between the actual value and the desired value of the traffic system with 4-s delay which is subject to parameter disturbance delay where the adopted delay 4-s in control is greater than the actual delay 3 s. (B) The $L_2$ norm of the states errors between the actual value and the desired value of the traffic system with under-matched delay where the adopted delay 4-s in control is less than the actual delay 5 s.

parameter disturbances, whereas the PPO controller stabilizes the system such that the state converges to the desired value, although there exist slight oscillations due to the Gaussian noise added on parameter $a$.

Figure 11 shows the evolution of the state considering unmatched delay under the Backstepping control and the pretrained PPO control, respectively. The over-matched case of the adopted delay 4-s in control greater than the actual delay 3 s is shown in Figure 11A, in which the PPO control shows better convergence performance for the $L_2$ norm of the density compared to the backstepping control, but worse convergence performance for the $L_2$ norm of the velocity. The under-matched situation of the adopted delay 4-s in control less than the actual delay 5 s is shown in Figure 11B, which illustrates the PPO controller has slight better convergence performance than the Backstepping controller for the $L_2$ norm of density, but worse convergence performance for the $L_2$ norm of velocity.

5 | CONCLUSION

In this work, a PPO reinforcement learning control method is developed for stabilizing a freeway traffic flow system with input delay. The traffic flow is mixed with manual and ACC-equipped vehicles. By manipulation of the time gap of the
ACC-equipped vehicles, the fluctuation of the traffic flow is suppressed, which alleviates congestion. The existing control methods heavily depend on model accuracy, while the PPO-based control is model free method by training the controller through interactions with the environment that can be real system simulator from data, although in the paper, we use the ARZ model to simulate the traffic environment.

The simulation experiments illustrate that the convergence performance of the PPO control is better than that of the Lyapunov control for the delay free system. The PPO control is comparable to the Backstepping control on the delay compensation for both common delay systems and mismatched delay systems although the PPO has much more simple controller form than the Backstepping. Besides, the PPO control is robust to parameter disturbance, while the Backstepping control cannot stabilized the system subject to parameter disturbances.

Applying the PPO control method to the real traffic system by pre-training the networks using the real data collected from the intelligent traffic system is a promising research direction. The PPO learning algorithm can also be improved by modifying the reward function, resetting the state space, or changing the network parameters optimization method.

AUTHOR CONTRIBUTIONS
All authors contributed to the study conception and design. The data collection, analysis and visualization were performed by Shurong Mo and Nailong Wu. The first draft of the manuscript was written by Shurong Mo. All authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENT
This work was partly supported by National Natural Science Foundation of China (62173084, 52101346, 22Z32003162 and 62303108). Besides, many thanks to the support from the Fundamental Research Funds for the Central Universities of Donghua University.

CONFLICT OF INTEREST STATEMENT
The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID
Nailong Wu https://orcid.org/0000-0003-4869-6361
Jie Qi https://orcid.org/0000-0001-6565-7984
Huaicheng Yan https://orcid.org/0000-0001-5496-1809

REFERENCES
1. Dadashova B, Li X, Turner S, Koeneman P. Multivariate time series analysis of traffic congestion measures in urban areas as they relate to socioeconomic indicators. Socioecon Plann Sci. 2021;75:100877.
2. Treiber M, Kesting A. Traffic Flow Dynamics: Data, Models and Simulation. Vol 67. Springer-Verlag; 2014:54.
3. Li G, Lai W, Sui X, et al. Influence of traffic congestion on driver behavior in post-congestion driving. Accid Anal Prev. 2020;141:105508.
4. Zhou J, Zhu F. Modeling the fundamental diagram of mixed human-driven and connected automated vehicles. Transport Res Pt C Emerg Technol. 2020;115:102614.
5. Wen Y, Zhang S, Zhang J, et al. Mapping dynamic road emissions for a megacity by using open-access traffic congestion index data. Appl Energy. 2020;260:114357.
6. Shen T, Hong Y, Thompson MM, Liu J, Huo X, Wu L. How does parking availability interplay with the land use and affect traffic congestion in urban areas? The case study of Xi’an. China Sustainable Cities and Society. 2020;57:102126.
7. Lighthill MJ, Whitham GB. On kinematic waves II. A theory of traffic flow on long crowded roads. Proc R Soc Lond Ser A Math Phys Sci. 1955;229(1178):317-345.
8. Richards P.I. Shock waves on the highway. Oper Res. 1956;4(1):42-51.
9. Aw A, Rascle M. Resurrection of “second order” models of traffic flow. SIAM J Appl Math. 2000;60(3):916-938.
10. Zhang HM. A non-equilibrium traffic model devoid of gas-like behavior. Transport Res Pt B Methodol. 2002;36(3):275-290.
11. Fan S, Seibold B. Data-fitted first-order traffic models and their second-order generalizations: comparison by trajectory and sensor data. Transport Res Record. 2013;2391(1):32-43.
12. Kolb O, Göttlich S, Goatin P. Capacity drop and traffic control for a second order traffic model. Netw Heterogeneous Media. 2017;12(4):663.
13. Yu H, Krstic M. Output feedback control of two-lane traffic congestion. Automatica. 2021;125:109379.
14. Belletti F, Huo M, Litrico X, Bayen AM. Prediction of traffic convective instability with spectral analysis of the Aw-Rascle-Zhang model. Phys Lett A. 2015;379(38):2319-2330.

15. Karafyllis I, Bekaris-Liberis N, Papageorgiou M. Feedback control of nonlinear hyperbolic PDE systems inspired by traffic flow models. IEEE Trans Automat Control. 2019;64(9):3647-3662.

16. Yu H, Krstic M. Traffic congestion control for Aw-Rascle-Zhang model. Automatica. 2019;100:38-51.

17. Zhang L, Prieur C. Necessary and sufficient conditions on the exponential stability of positive hyperbolic systems. IEEE Trans Automat Control. 2017;62(7):3610-3617.

18. Diakaki C, Papageorgiou M, Papamichail I, Nikolos I. Overview and analysis of vehicle automation and communication systems from a motorway traffic management perspective. Transport Res Pt A Policy Pract. 2015;75:147-165.

19. Bekaris-Liberis N, Delis AI. PDE-based feedback control of freeway traffic flow via time-gap manipulation of ACC-equipped vehicles. IEEE Trans Control Syst Technol. 2021;29(1):461-469.

20. Zheng Y, Wang J, Li K. Smoothing traffic flow via control of autonomous vehicles. IEEE Internet Things J. 2020;7(5):3882-3896.

21. Burger M, Göttlich S, Jung T. Derivation of second order traffic flow models with time delays. Netw Heterogeneous Media. 2019;14(2):265.

22. Auriol J, Aarsnes UJF, Martin P, Di Meglio F. Delay-robust control design for two heterodirectional linear coupled hyperbolic PDEs. IEEE Trans Automat Control. 2018;63(10):3551-3557.

23. Qi J, Mo S, Krstic M. Delay-compensated distributed PDE control of traffic with connected/automated vehicles. IEEE Trans Autom Control. 2022;68(4):2229-2244. doi:10.1109/TAC.2022.3174032

24. Wei H, Liu X, Mashayekhy L, Decker K. Mixed-autonomy traffic control with proximal policy optimization. 2019 IEEE Vehicular Networking Conference (VNC). IEEE; 2019:1-8.

25. Rasheed F, Yau KA, Noor RM, Wu C, Low Y. Deep reinforcement learning for traffic signal control: a review. IEEE Access. 2020;8:208016-208044.

26. Wang C, Ran B, Yang H, Zhang J, Qu X. A novel approach to estimate freeway traffic state: parallel computing and improved kalman filter. IEEE Intell Transp Syst Mag. 2018;10(2):180-193.

27. Yu H, Park S, Bayen A, Moura S, Krstic M. Reinforcement learning versus PDE Backstepping and PI control for congested freeway traffic. IEEE Trans Control Syst Technol. 2021;30(4):1595-1611.

28. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Truly proximal policy optimization. Uncertainty in Artificial Intelligence. PMLR; 2020:113-122.

29. Shahid AA, Piga D, Braghin F, Roveda L. Continuous control actions learning and adaptation for robotic manipulation through reinforcement learning. Autonom Robot. 2022;46(3):483-498.

30. Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. ArXiv. 2017:abs/1707.06347.

31. Kingma DP, Ba J. Adam: a method for stochastic optimization. CoRR. 2014;abs/1412.6980.

32. Delis AI, Nikolos IK, Papageorgiou M. Macroscopic traffic flow modeling with adaptive cruise control: development and numerical solution. Comput Math Appl. 2015;70(8):1921-1947.

33. Fan S, Herty M, Seibold B. Comparative model accuracy of a data-fitted generalized aw-rascle-zhang model. Netw Heterogeneous Media. 2014;9(2):239-268.

34. Ngoduy D. Instability of cooperative adaptive cruise control traffic flow: a macroscopic approach. Commun Nonlinear Sci Num Simul. 2013;18(10):2838-2851.

35. Song X, Dong Z, Ding F, Lu W. Data-iteration cooperative adaptive cruise control of mixed heterogeneous vehicle platoons with unknown dynamic characteristics. Asia J Control. 2023;25(4):2776-2789. doi:10.1002/asjc.2975

How to cite this article: Mo S, Wu N, Qi J, et al. Proximal policy optimization learning based control of congested freeway traffic. Optim Control Appl Meth. 2024;45(2):719-736. doi: 10.1002/oca.3068