World Electric Vehicle Journal

MDPI

Article

High-Accuracy, High-Efficiency, and Comfortable Car-Following Strategy Based on TD3 for Wide-to-Narrow Road Sections

Pinpin Qin [1,*], Fumao Wu [1], Shenglin Bin [1], Xing Li [1] and Fuming Ya [2]

[1] School of Mechanical Engineering, Guangxi University, Nanning 530004, China; 2111391114@st.gxu.edu.cn (F.W.); 2111391001@st.gxu.edu.cn (S.B.); 2111301027@st.gxu.edu.cn (X.L.)
[2] School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China; yfm123930@163.com
* Correspondence: qpinpin@gxu.edu.cn

Abstract: To address traffic congestion in urban expressways during the transition from wide to narrow sections, this study proposed a car-following strategy based on deep reinforcement learning. Firstly, a car-following strategy was developed based on a twin-delayed deep deterministic policy gradient (TD3) algorithm, and a multi-objective constrained reward function was designed by comprehensively considering safety, traffic efficiency, and ride comfort. Secondly, 214 car-following periods and 13 platoon-following periods were selected from the natural driving database for the strategies training and testing. Finally, the effectiveness of the proposed strategy was verified through simulation experiments of car-following and platoon-following. The results showed that compared to human-driven vehicles (HDV), the TD3 and deep deterministic policy gradient (DDPG)-based strategies enhanced traffic efficiency by over 29% and ride comfort by more than 60%. Furthermore, compared to DDPG, the relative errors between the following distance and desired safety distance using TD3 could be reduced by 1.28% and 1.37% in simulation experiments of car-following and platoon-following, respectively. This study provides a new approach to alleviate traffic congestion for wide-to-narrow road sections in urban expressways.

Keywords: car-following; twin delayed deep deterministic policy gradient (TD3); wide-to-narrow road sections; desired safety distance (DSD); traffic congestion

Citation: Qin, P.; Wu, F.; Bin, S.; Li, X.; Ya, F. High-Accuracy, High-Efficiency, and Comfortable Car-Following Strategy Based on TD3 for Wide-to-Narrow Road Sections. *World Electr. Veh. J.* **2023**, *14*, 244. https://doi.org/10.3390/wevj14090244

Academic Editor: Joeri Van Mierlo

Received: 27 June 2023
Revised: 14 August 2023
Accepted: 1 September 2023
Published: 3 September 2023

1. Introduction

In the sections of urban expressways where the road width narrows, the reduction in the number of lateral lanes leads to a decrease in the road capacity. In addition, the imperfection of drivers in adjusting their speed may result in an insufficient or excessive response to the expected value, causing frequent acceleration and deceleration [1]. Therefore, such road segments often lead to traffic congestion and reduced ride comfort. Simultaneously, the road environment becomes more complex, limiting overtaking opportunities for vehicles and thereby resulting in car-following (CF) behavior. However, adopting high-accuracy control of shorter driving distances and CF technology with strong generalization can reduce the drivers' workload, improve road traffic efficiency, and thus reduce traffic congestion [2].

CF describes the longitudinal interaction between vehicles traveling on single-lane roads with restricted overtaking [3]. CF models are mainly divided into theory-driven models and data-driven models. Theory-driven CF models rely on a few fixed parameters for theoretically modeling and deducing traffic phenomena, making it difficult to comprehensively consider the influencing factors and resulting in poor model prediction accuracy in complex traffic flows. Standard theory-driven CF models include the IDM model [4], the Newell model [5], and the cellular automaton model [6]. Data-driven CF

models mainly include deep learning CF models and deep reinforcement learning (DRL) CF models. Deep learning models rely on large-scale data and imitate pure data samples, resulting in poor generalization ability. Shared deep learning CF models mainly include the RNN CF model, LSTM CF model, and GA-BP CF model [7–9]. In contrast, DRL adopts a self-learning mode, does not require prior samples for pure imitation, and its agent learns through continuous trial and error interaction with the driving environment until the optimal strategy is obtained, demonstrating strong generalization ability [10].

Existing research shows that strategies based on DRL demonstrate superiority in solving various types of CF problems. Gao et al. researched autonomous vehicle CF decision-making based on reinforcement learning, proving that reinforcement learning systems are more adaptable to CF and have a certain degree of interpretability and stability [11]. CF strategies based on reinforcement learning mainly spend time on offline training, and after training, they can be quickly implemented in real-time on vehicles [12]. Ye et al. proposed a decision-making training framework for autonomous driving using DRL. After training, the efficiency of the autonomous vehicles increased by 7.9% compared to the vehicles controlled by the IDM model, verifying the effectiveness of the proposed model in learning driving decisions [13]. Sun et al. established a heavy vehicle adaptive cruise control strategy model based on the deep deterministic policy gradient (DDPG) algorithm, achieving adaptive cruise control objectives for heavy vehicles on roads with different curvatures, and verified the effectiveness and robustness of the model through simulation experiments [14]. Zhu et al. constructed a speed control model during CF based on the DDPG. They verified that the model-controlled vehicle is superior to human drivers and model predictive control (MPC) adaptive cruise control models in terms of safety, efficiency, and comfort, with a running speed of more than 200 times faster than the MPC algorithm during testing [15]. Shi et al. proposed a DRL-based connected autonomous vehicle collaborative control under mixed traffic flow, different penetration rates, and different behaviors of human-driven leading vehicles. The proposed model can effectively complete car tracking and energy-saving tasks [16]. Yan et al. developed a hybrid vehicle CF strategy based on DDPG and cooperative adaptive cruise control, and the proposed strategy improved vehicle tracking performance [17]. Qin et al. developed DDPG and MADDPG CF models considering longitudinal and lateral joint control, using the OpenACC dataset to train and test them under straight and curved conditions of highway free flow, and the results showed that the developed models controlled the CF effect better than human-driven vehicles (HDV) [18]. Chen et al. proposed an intelligent speed control method for autonomous vehicles in cooperative vehicle-infrastructure systems based on the DDPG, and through vertical comfort evaluation, the method was effective on rough road surfaces [19].

In summary, CF control strategies based on DRL have made significant progress, but they are mainly limited to road sections with fixed widths. Segments transitioning from wide to narrow roads are bound to encounter increasingly intricate traffic scenarios, necessitating more adept car-following control strategies. Existing research on CF strategies for road sections with variable widths, such as those found in urban expressways, has received insufficient attention, and few studies have been able to balance travel efficiency and ride comfort. To alleviate this problem, this paper develops a CF control strategy based on deep reinforcement learning. This strategy uses a self-learning method to continuously interact with the CF environment to obtain the optimal strategy. The CF periods for congested traffic with variable widths of urban expressways are used for strategies training and testing. The main contributions of this study are as follows: (1) Developing a high-accuracy, high-efficiency, and comfortable CF strategy based on the twin-delayed deep deterministic policy gradient (TD3) suitable for road sections with variable widths of urban expressways; (2) Designing a new multi-objective constraint reward function, which considers safety, travel efficiency, and ride comfort. This function employs the error between the following distance and the desired safety distance, the speed error between the following vehicle and the leading vehicle, and the jerk value, as the variables in exponential

functions. This formulation leads to rapid and stable convergence during training. Notably, in instances of collisions during the training process, a substantial penalty is imposed, prompting the agent to learn collision avoidance strategies swiftly and autonomously; (3) Conducting simulation experiments on car-following and platoon-following to verify the effectiveness of the proposed strategy.

The rest of this paper is organized as follows: Section 2 introduces the research methods and data extraction. Section 3 presents the results and discussion. The final section concludes the paper.

2. Methods and Data

2.1. CF Strategy Based on TD3

The policy function π of reinforcement learning is defined by Equation (1), which represents the probability of taking action a_t in a given environmental state s_t at the time t.

$$\pi(a_t|s_t) = p(\ A_t = a_t|S_t = s_t\ ) \tag{1}$$

Using a policy function π, the agent selects an action a_t based on the current environmental state s_t at each time step. The agent then transitions to the new state s_{t+1} according to the state transition probabilities $P(s_{t+1}|s_t, a_t)$ and receives an immediate reward r_t from the environment. This process is known as reinforcement learning, as shown in Figure 1.

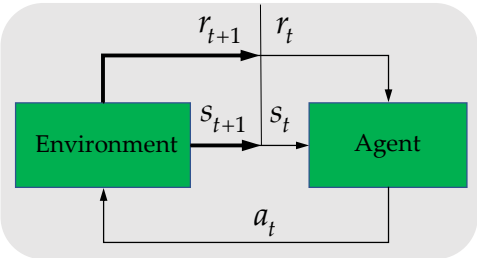

Figure 1. Reinforcement learning process.

The twin delayed deep deterministic policy gradient (TD3) is a DRL algorithm that is good at handling continuous state-action space problems. The TD3 algorithm is an improvement over the DDPG algorithm by Scott Fujimoto et al., to reduce the bias and variance introduced by function approximation in the actor–critic framework, making the model more stable [20]. In the TD3, there are a total of six neural networks. An actor network is used to fit the policy function π, while critic networks 1 and 2 are used to estimate the action-value function $Q_{i=1,2}$, with their parameters being independent. The target actor network is used to fit the target policy function π', and the target critic networks 1 and 2 are used to fit the target action-value function $Q'_{i=1,2}$, reducing the risk of overestimation.

In the TD3, the critic networks 1 and 2 are updated by minimizing the loss function $Loss(\theta_i)$ and Q' can be calculated using the temporal difference (TD) principle:

$$Loss(\theta_i) = N^{-1}\sum\left(Q' - Q_{i=1,2}(s_t, a_t)\right)^2 \tag{2}$$

$$\begin{cases} Q' = r_t + \gamma \min\limits_{i=1,2} Q'_{i=1,2}(s_{t+1}, \pi'(s_{t+1}) + \varepsilon) \\ \varepsilon \sim clip(N(0, \tilde{\sigma}), -c, c) \end{cases} \tag{3}$$

where θ_1 and θ_2 represent the parameters of critic network 1 and critic network 2, respectively, N refers to randomly selecting state transitions from the experience replay buffer as a mini-batch for training, r_t represents the reward obtained by taking a particular action in the current state, γ is the future discount reward factor, and ε represents the noise from

a Gaussian distribution with a mean of 0 and a variance of $\widetilde{\sigma}$, and this noise is confined within the range of $(-c, c)$.

The actor network is updated using the policy gradient method:

$$\nabla_{\phi} J(\phi) = N^{-1} \sum \nabla Q(s_t, a_t) \nabla_{\phi} \pi_{\phi}(s_t) \tag{4}$$

where ϕ represents the parameters of the actor network.

The target actor network and target critic networks 1 and 2 update their parameters:

$$\begin{cases} \theta'_{i=1,2} \leftarrow \tau\theta_i + (1-\tau)\theta'_i \\ \phi' \leftarrow \tau\phi + (1-\tau)\phi' \end{cases} \tag{5}$$

where θ'_1 and θ'_2 represent the parameters of the target critic network 1 and 2, respectively, τ is the soft update rate, and ϕ is the parameter of the target actor network.

After multiple experiments, it was found that using a hidden layer with 64 neurons could achieve the best balance between computational cost and accuracy. In the hidden layers of both the actor and critic networks, the ReLU activation function was employed. Additionally, a Tanh activation function was applied to the actor network's output layer to constrain the output actions' boundary values. Furthermore, all the other layers of the actor and critic networks utilized fully connected layers. For more detailed information on the network structure and parameter updates of the TD3, refer to Figure 2.

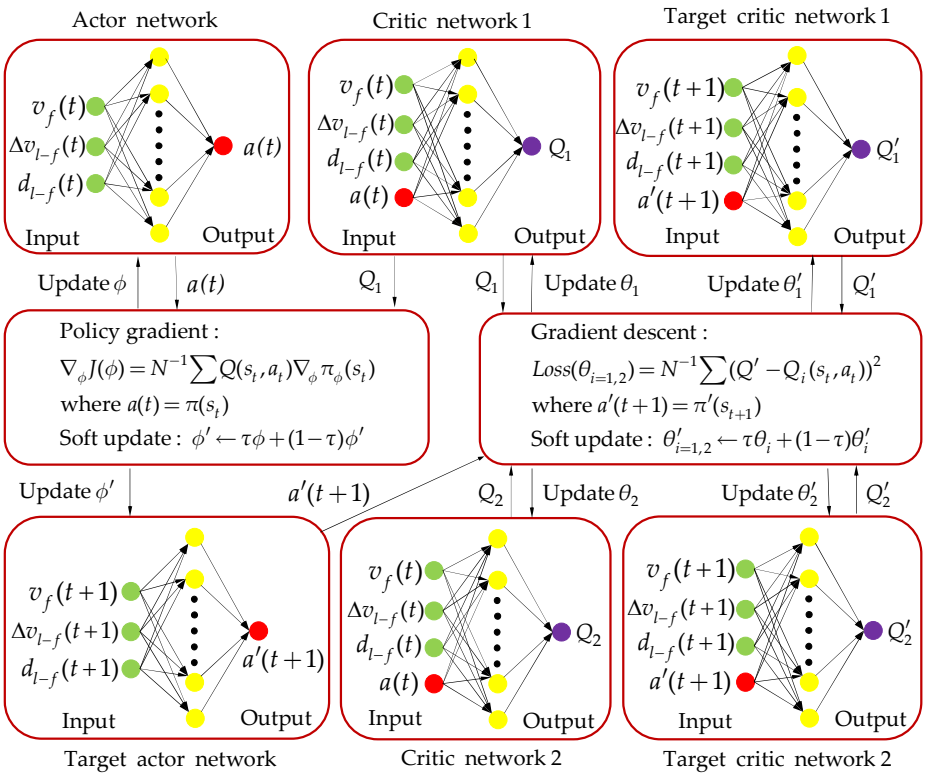

Figure 2. Network's structure and parameter update of TD3.

The process of CF can be abstracted as a reinforcement learning problem, and the partially observable Markov decision process (POMDP) can be described using the parameter tuple (s_t, a_t, r_t, s_{t+1}). At a particular time step t, the observed environment state s_t consists of the following variables: the speed of the following vehicle $v_f(t)$, the distance $\Delta d_{l-f}(t)$ to the leading vehicle, and the relative speed $\Delta v_{l-f}(t)$. The continuous action of the agent is the acceleration of the following vehicle $a_f(t) \in \left[-2m/s^2, 2m/s^2\right]$ [21]. The update of the new state observation value depends on the vehicle's kinematic Equation (6),

where $v_l(t+1)$ is the speed of the leading vehicle at the next moment and $T_s = 0.08s$ is the simulation time step.

$$\begin{cases} v_f(t+1) = v_f(t) + a_f(t) * \Delta t \\ \Delta v_{l-f}(t+1) = v_l(t+1) - v_f(t+1) \\ \Delta d_{l-f}(t+1) = \Delta d_{l-f}(t) + (\Delta v_{l-f}(t) + \Delta v_{l-f}(t+1)) * \Delta t/2 \end{cases} \quad (6)$$

The TD3 algorithm is used to construct a CF strategy in a simulation environment based on the CF trajectory data. The optimal strategy is obtained through continuous interaction between the agent and the driving environment and the supervision of the reward function signal. The training framework is shown in detail in Figure 3.

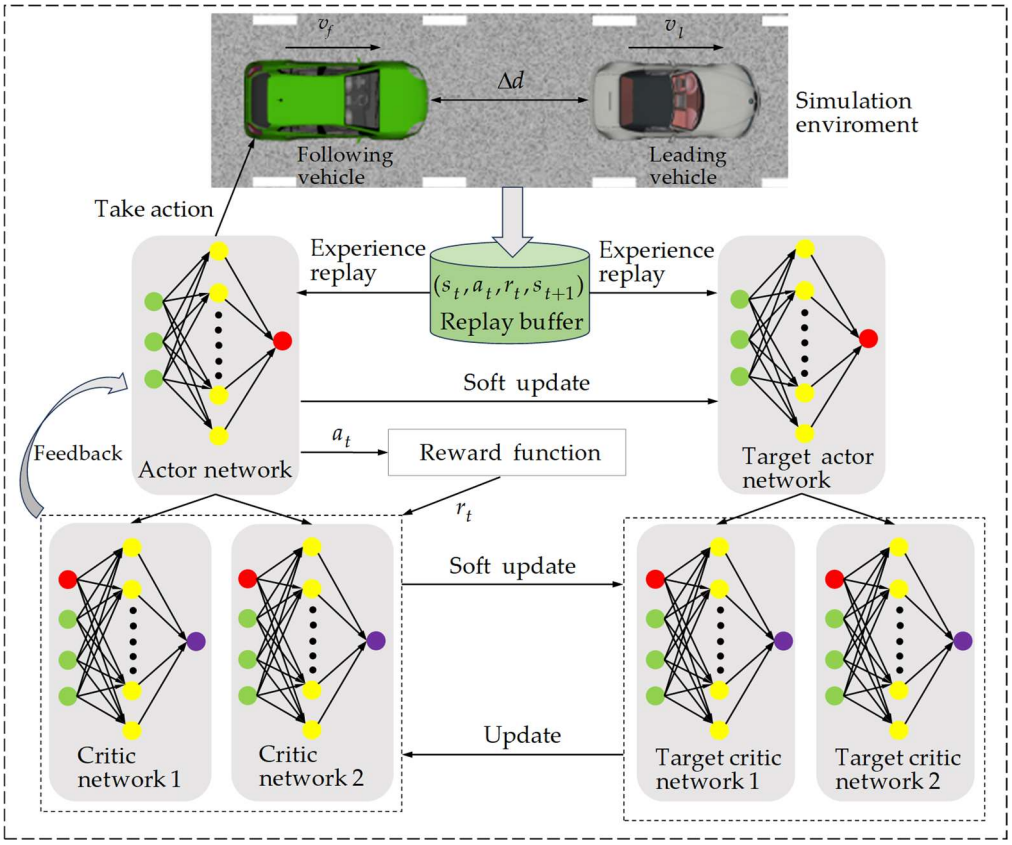

Figure 3. A training framework of a car-following strategy based on TD3.

2.2. Evaluation Metrics for Car-Following Behavior

This study introduces the generalized force model to define the desired safe distance (DSD) $d_{DSD}(t)$ [22]. The slighter the error between the following distance and the DSD, the more stable the distance control between the following and leading vehicles, and the higher the safety:

$$d_{DSD}(t) = t_r * v_f(t) + d_0 \quad (7)$$

where $t_r = 1.2s$ is the constant headway distance, and $d_0 = 2m$ is the gap between the two cars when the following speed is 0.

Time headway (THW) describes the time interval between the leading and following vehicles. Under the premise of meeting safety conditions, a minor time headway means higher traffic efficiency [23], as defined in Equation (8).

$$THW(t) = \Delta s_{l-f}(t)/v_f(t) \quad (8)$$

The jerk value represents the rate of change of acceleration during the CF process. A slighter absolute value of jerk indicates smoother acceleration and deceleration of the following vehicle, leading to increased riding comfort, as defined in Equation (9).

$$Jerk(t) = (a_f(t+1) - a_f(t))/\Delta t \tag{9}$$

2.3. Reward Function

The design goal of the reward function in this study is to control the following vehicle to travel accurately according to the DSD in the section of the urban expressway where the road width changes from wide to narrow while significantly improving road traffic efficiency and ride comfort. As the supervision signal of DRL, the quality of the reward function design directly affects the agent's ability to learn the expected strategy, thus affecting the reinforcement learning algorithm's convergence speed and final performance. Under the premise of ensuring safety, a multi-objective constrained reward function is designed by considering safety, traffic efficiency, and ride comfort. The reward function expression is designed to minimize the difference between the target and observation values.

Considering the safety of the CF behavior and road traffic efficiency, the error between the distance $d_{l-f}(t)$ and DSD $d_{DSD}(t)$ is used to design the reward function $r_1(t)$. The slighter the error between the distance and DSD, the greater the reward, as defined in Equation (10). In order to make the agent learn to avoid collisions, a penalty function $r_{1\_collision}(t)$ is designed, as shown in Equation (11). The reward function $r_2(t)$ is also designed using the speed difference between the leading and following vehicles to encourage the vehicle to maintain an appropriate speed difference, as shown in Equation (12). In consideration of ride comfort, the reward function is designed using the ratio of the jerk value to the maximum jerk value, as shown in Equation (13):

$$r_1(t) = \exp(-\omega_1 * (d_{l-f}(t) - d_{DSD}(t))^2) \tag{10}$$

$$r_{1\_collision}(t) = \begin{cases} -1 & d_{l-f}(t) < 0 \\ 0 & otherwise \end{cases} \tag{11}$$

$$r_2(t) = \begin{cases} \exp(-\omega_2 * (v_f(t) - v_l(t))^2) & 0 \le v_f(t) \le 22.22 \text{ m/s} \\ -1 & v_f(t) > 22.22 \text{ m/s} \end{cases} \tag{12}$$

$$r_3(t) = \exp(-\omega_3 * (Jerk(t)/Jerk_{\max})^2) \tag{13}$$

where $\omega_1 = 1$, $\omega_2 = 1$, and $\omega_3 = 1$ are the weight coefficients, and 22.22 m/s is the speed limit of the urban expressway.

In summary, the total reward function expression is as follows:

$$r(t) = \lambda_1 * r_1(t) + \lambda_2 * r_2(t) + \lambda_3 * r_3(t) + r_{1\_collision} \tag{14}$$

where $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.1$ are the weight coefficients of each sub-reward function.

The pseudo-code of the CF strategy based on TD3 is shown in Algorithm 1, and its corresponding hyperparameters are detailed in Table 1.

Table 1. TD3 algorithm hyperparameters.

Parameter	Symbol	Value
Sampling step (s)	T_s	0.08
Batch size	/	128
Discount factor	γ	0.91
Actor learning rate	α	3×10^{-4}
Critic learning rate	β	3×10^{-4}
Soft update rate	τ	8×10^{-3}
Replay buffer capacity	/	2×10^{6}

Algorithm 1: Car-Following Strategy Based on TD3

Initialize critic networks $Q_1(s_t, a_t)$, $Q_2(s_t, a_t)$ and actor network $\pi(s_t)$
With random parameters θ_1, θ_2, ϕ
Initialize target networks $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, $\phi' \leftarrow \phi$
Initialize replay buffer
for *episode* = 1 **to** M **do**
  Initialize random process for action exploration ε_0
  Receive initial state s
  for $t = 1, T_f$ **do**
    Choose action based on current policy and noise: $a_t \leftarrow \pi(s_t)+\varepsilon$, $\varepsilon \sim N(0, \sigma)$
    Execute action a_t, obtain the reward r_t, and enter the next state s_{t+1}
    Store the state transition sequence (s_t, a_t, r_t, s_{t+1}) in the replay buffer
    Randomly take a small batch of samples from the replay buffer:
      $a_{t+1} \leftarrow \pi'(s_{t+1}) + \varepsilon$, $\varepsilon \sim clip(N(0, \widetilde{\sigma}), -c, c)$
    Calculate based on the temporal difference:
      $Q' \leftarrow r(s_t, a_t) + \gamma \min_{i=1,2} Q_i'(s_{t+1}, a_{t+1})$
    Update critic networks
      $\nabla_{\theta_i} Loss(\theta_i) = N^{-1} \sum (Q' - Q_{i=1,2}(s_t, a_t))^2$
      $\theta_{i=1,2} \leftarrow \theta_i - \beta \nabla_{\theta_i} Loss(\theta_i)$
    if t **mod** d **then**
      Update actor network: $\phi \leftarrow \phi + \alpha \nabla_\phi J(\phi)$
      Update target networks: $\theta_{i=1,2}' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$, $\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
    end if
  end for
end for

2.4. Data

The CF trajectory data used in this study was obtained by the ubiquitous traffic eyes (UTE) team of Southeast University through high-altitude aerial photography using drones on multiple urban expressways in Nanjing, China. The speed limit on these urban expressways is 80 km per hour. The drones were set at over 200 m to cover congested and free-flow traffic conditions. Finally, the team used algorithms to extract data from the video footage [24].

The UTE team extracted six datasets, and this study selected datasets 1 and 3 as the data sources. Both datasets were collected on sections of urban expressways where lanes narrow, covering the entire evolution process from free-flow to congested traffic. Dataset 1's lane distribution is shown in Figure 4a, where the number of lanes decreased from 5 to 4, then from 4 to 3. Dataset 3's lane distribution is shown in Figure 4b, where the lanes decreased from 5 to 3 [25]. The database parameters can be found in Table 2, and the congested scenes in the video can be seen in Figure 4c.

This study extracted a total of 214 CF periods and 13 platoon-following periods from the database, which had the following characteristics: (1) The duration of each CF trajectory data was greater than 20 s; (2) To ensure that the vehicles did not change lanes or make sudden turns, the lateral position difference between the leading vehicle and the following vehicle should have been less than 1 m; (3) All vehicles were in congested traffic flow.

Table 2. Description of natural driving CF trajectory database.

Parameter	Value	
	Dataset 1	Dataset 3
Road length (m)	427	362
Duration (s)	255	545
Temporal accuracy (s)	0.01	0.01
Position accuracy (m)	0.01	0.01
Sampling frequency (Hz)	25	25

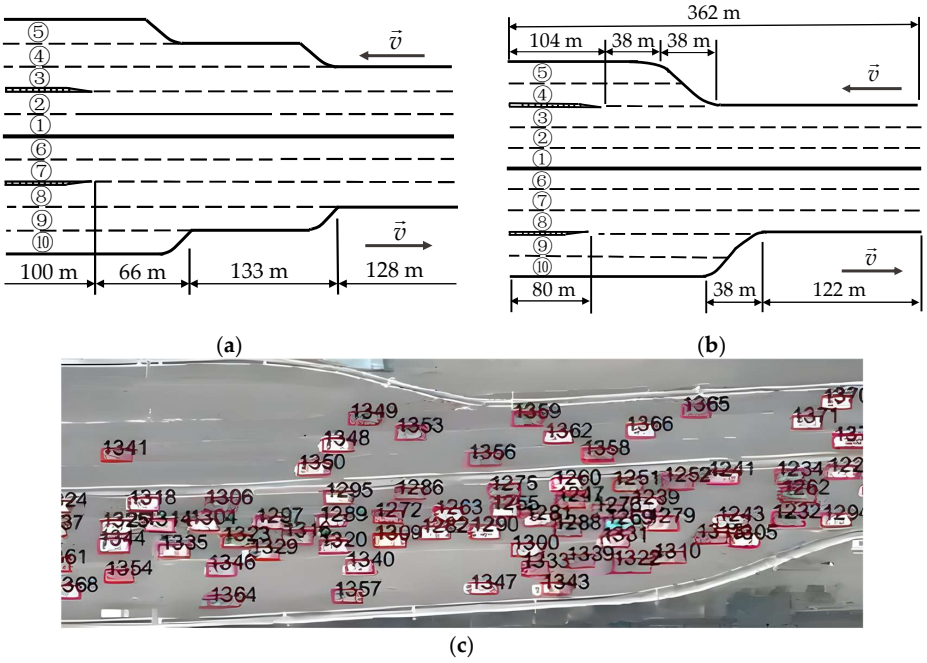

Figure 4. Lane distribution diagram and congestion scene in the video: (**a**) Lane distribution diagram of dataset 1, the numbers in the circle are the lane numbers; (**b**) Lane distribution diagram of dataset 3, the numbers in the circle are the lane numbers; (**c**) Congestion scene in the video.

3. Results and Discussion

3.1. Strategies Training

The purpose of training is to enable the agent to interact fully with the environment and obtain the optimal strategy. During the training process, a trajectory was randomly selected from 150 CF trajectories for training for each episode, with the remaining data (64 CF periods and 13 platoon-following periods) utilized as the test dataset. The training was repeated for 1800 episodes. The mean reward referred to the average reward value of all the time steps in a training episode, while the moving mean episode reward was the average reward value of a moving window of size 100. Under the supervision of the reward function signal, the agent continuously interacted with the environment through trial-and-error learning, maximizing the cumulative rewards and eventually reaching convergence. Figure 5 illustrates the training results based on the TD3 and DDPG. The TD3-based strategy began to converge after about 110 episodes and reached convergence after about 235 episodes, with a training duration of 1 h and 5 min. The DDPG-based strategy began to converge after about 168 episodes and reached convergence after about 296 episodes, with a training duration of 1 h and 35 min. Therefore, the TD3-based strategy converged faster and reduced the training time by 31.58% compared to the DDPG, effectively reducing the training costs.

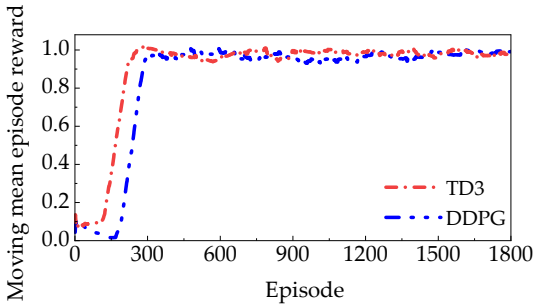

Figure 5. Training results.

3.2. Simulation Results of Car-Following Experiments

In total, 64 CF periods were tested to verify the effectiveness of the proposed strategy on CF behavior, and no collisions occurred during the entire testing process. Figure 6 illustrates that the mean rewards for the TD3 and DDPG strategies in the test results were 1.04 and 1.02, respectively, indicating that the mean reward was higher using the TD3 than the DDPG. Table 3 presents the results of all the car-following tests. Compared to the HDV and DDPG, the relative errors between the following distance and the DSD through the TD3 were reduced by 41.82% and 1.28%, respectively, suggesting that the TD3-based strategy offered the highest accuracy. The mean-time headway using the TD3 and DDPG could be reduced by 29.30% and 29.17%, respectively, compared to the HDV, and the mean absolute jerk values were reduced by 60.22% and 64.61% m/s^3, respectively. This significant reduction in the time headway and absolute jerk values for TD3 and DDPG greatly enhanced the road traffic efficiency and ride comfort. Although the average absolute jerk based on TD3 was slightly larger than the DDPG, the TD3 exhibited a distinct advantage in maintaining the error between the following distance and the desired safety distance. Moreover, it demonstrated higher traffic efficiency. The TD3-based strategy demonstrated the best performance in CF behavior.

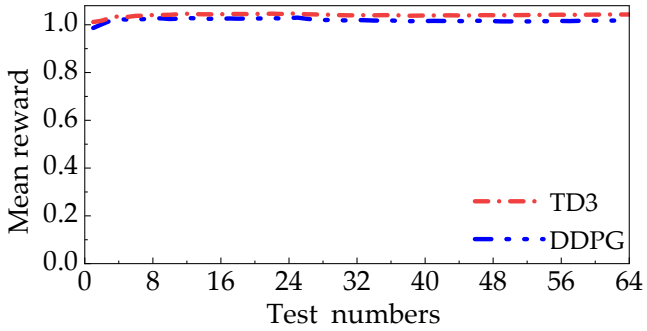

Figure 6. Test results.

Table 3. Results of all car-following simulation tests.

Comparative Indicator	TD3	DDPG	Human
Mean relative error to DSD	0.96%	2.24%	42.78%
Mean-time headway (s)	1.643	1.646	2.324
Mean absolute value of jerk (m/s^3)	0.290	0.258	0.729

Since the error between the initial distance and DSD can cause differences in CF behavior, a CF trajectory was randomly selected from the test dataset for detailed analysis and discussion under three conditions: when the initial distance was equal to, less than, or greater than the DSD.

As indicated by Table 4 and Figures 7–9, the TD3-based strategy consistently demonstrated superior control accuracy regardless of whether the initial distance was equal to, less than, or greater than the DSD. When the initial distance was equal to DSD, the following vehicle using the TD3 and DDPG could immediately drive according to the DSD. Compared with the HDV and DDPG, the relative error using the TD3 reduced by 0.73% and 15.14%, respectively. When the initial distance was less than DSD, the vehicle using the TD3 and DDPG could decelerate to reach the DSD and then drive according to DSD. Compared with HDV and DDPG, the relative error using TD3 reduces by 3.53% and 14.18%, respectively. When the initial distance is greater than the DSD, the vehicle using the TD3 and DDPG could decelerate to reach the DSD and then drive according to the DSD. Compared with the HDV and DDPG, the relative error using the TD3 reduced by 3.35% and 99.49%, respectively. In all three scenarios, the TD3-based strategy proved to be highly accurate, efficient, and comfortable in CF.

Table 4. Results of randomly selected car-following simulation tests.

Initial Condition	Data Source	Distance	Time Headway (s)			The Absolute Value of Jerk (m/s^3)		
		MREDSD *	Minimum	Mean	Maximum	Minimum	Mean	Maximum
An initial distance equal to DSD	TD3	0.30%	1.449	1.701	2.052	0	0.211	0.904
	DDPG	1.03%	1.441	1.697	2.090	0	0.218	1.075
	Human	15.44%	1.053	1.835	3.393	0.002	0.673	3.197
An initial distance less than DSD	TD3	0.88%	1.188	1.704	2.336	0.002	0.260	3.915
	DDPG	4.41%	1.188	1.647	2.343	0.002	0.215	5.787
	Human	18.59%	1.161	1.857	3.765	0	0.877	3.411
An initial distance greater than DSD	TD3	2.30%	1.539	1.837	5.303	0	0.289	2.810
	DDPG	5.58%	1.520	1.905	5.303	0.002	0.267	0.986
	Human	101.72%	2.963	3.741	5.321	0.003	0.890	2.967

* MREDSD is mean relative error to desired safety distance.

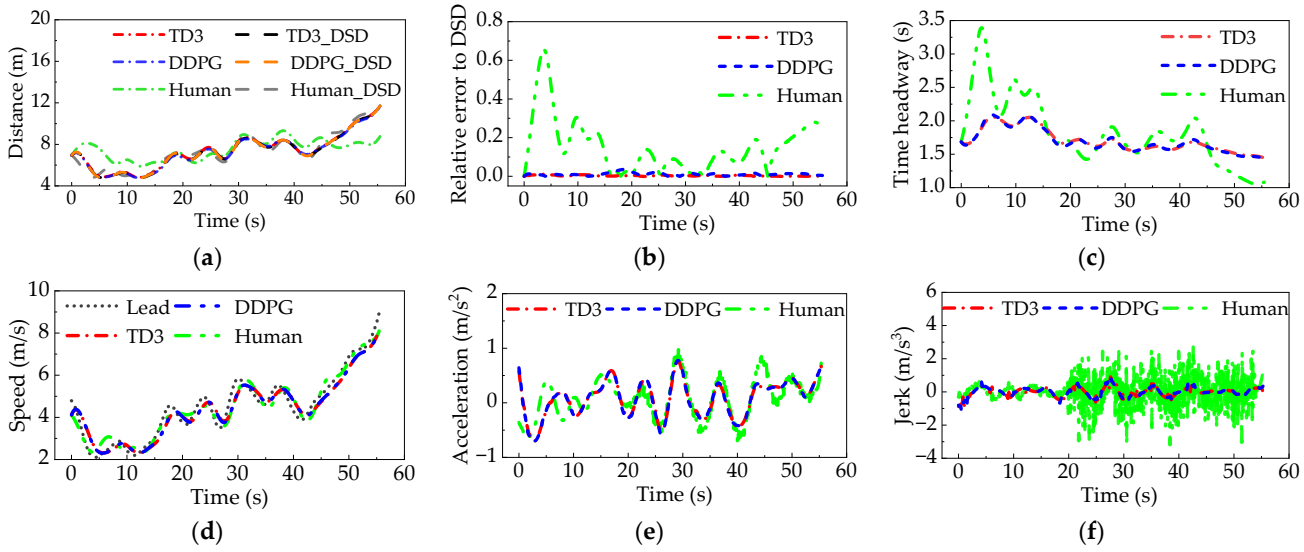

Figure 7. An initial distance equal to DSD: (**a**) Comparison of distances; (**b**) Comparison of the relative error to DSD; (**c**) Comparison of time headway; (**d**) Comparison of speed; (**e**) Comparison of acceleration; (**f**) Comparison of jerk values.

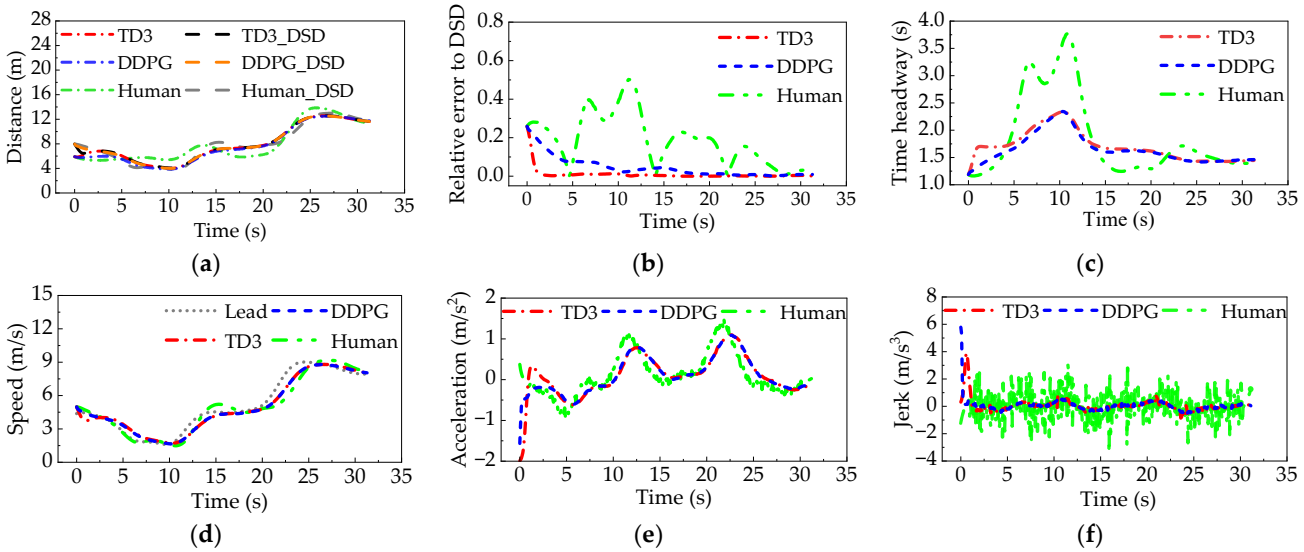

Figure 8. An initial distance less than DSD: (**a**) Comparison of distances; (**b**) Comparison of the relative error to DSD; (**c**) Comparison of time headway; (**d**) Comparison of speed; (**e**) Comparison of acceleration; (**f**) Comparison of jerk values.

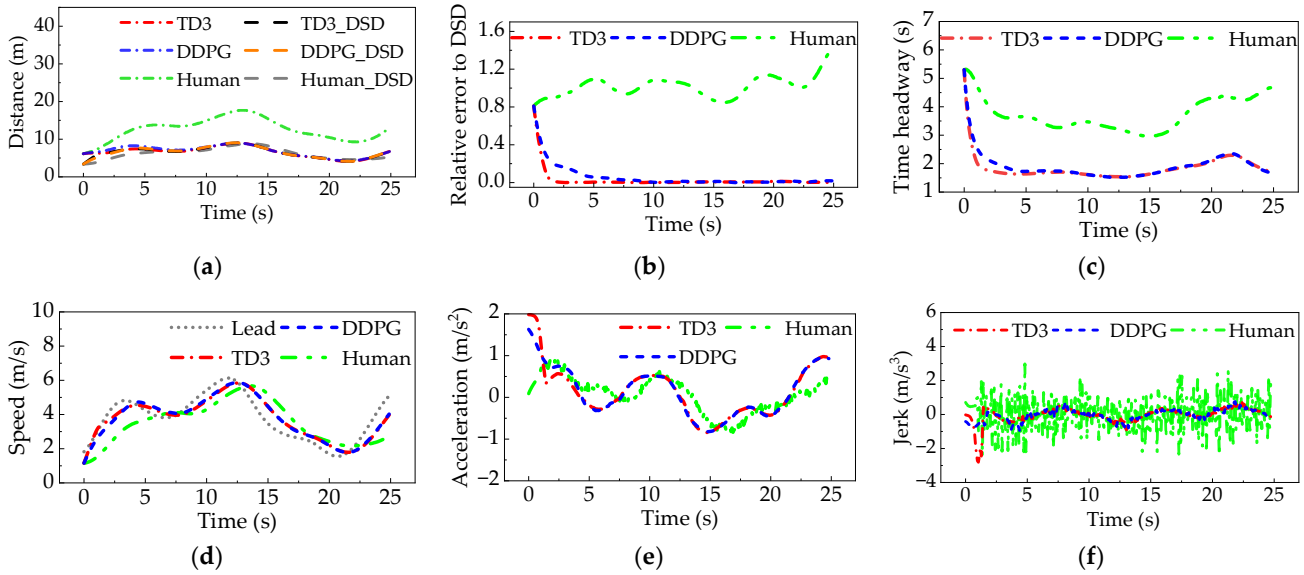

Figure 9. An initial distance greater than DSD: (**a**) Comparison of distances; (**b**) Comparison of the relative error to DSD; (**c**) Comparison of time headway; (**d**) Comparison of speed; (**e**) Comparison of acceleration; (**f**) Comparison of jerk values.

3.3. Simulation Results of Platoon-Following Experiment

In congested traffic flow, most vehicles travel in a platoon with multiple vehicles following each other. To further verify the effectiveness of the proposed strategy in this paper, thirteen platoon-following simulation experiments were conducted, each containing from five to nine vehicles. The topology structure of the platoon-following was the predecessor-following communication topology [26], as shown in Figure 10. The leading vehicle in each platoon-following was an HDV, and the initial values of the following vehicles in the simulation were derived from the trajectory data. No collisions occurred during all the platoon-following simulation experiments.

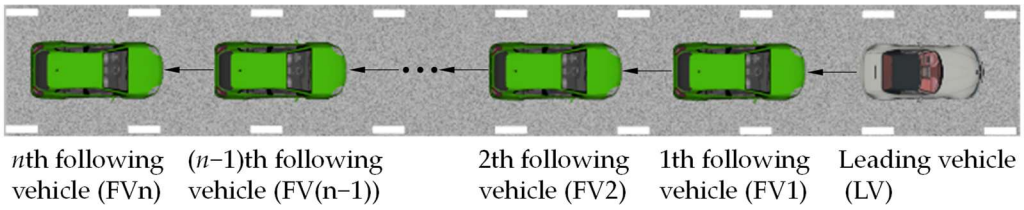

Figure 10. Predecessor-following communication topology.

Table 5 presents the results of the simulation tests for all the platoons. In these tests, the TD3-based strategy demonstrated superior control accuracy in the platoon-following simulations. When compared to the HDV and DDPG, the mean error between the following distance of the following vehicles and the DSD was reduced by 1.37% and 41.12%, respectively, when using the TD3. Furthermore, the mean-time headway of the following vehicles could be reduced by 31.59% and 31.10% when using the TD3 and DDPG, respectively, compared to the HDV, leading to a significant enhancement in the traffic efficiency. Lastly, the mean absolute jerk of the platoons could be reduced by 81.26% and 83.08% when using the TD3 and DDPG, respectively, resulting in a substantial improvement in the ride comfort.

The results of a randomly selected platoon-following experiment are presented in Table 6 and Figures 11–13. We could observe that the platoon using the TD3 could travel along the desired trajectory. The mean error between the following distance of the following vehicles and the DSD using the TD3 was 0.87%, with the highest control accuracy, and the road traffic efficiency and ride comfort were significantly improved.

Table 5. Results of platoon-following simulation experiments.

Comparative Indicator	TD3	DDPG	Human
Mean relative error to DSD	1.10%	2.47%	42.22%
Mean-time headway (s)	1.665	1.677	2.434
Mean absolute value of jerk (m/s^3)	0.181	0.155	0.780

Table 6. Results of a randomly selected platoon-following experiment.

Comparative Indicator	TD3	DDPG	Human
Mean relative error to DSD	0.87%	2.58%	36.51%
Mean-time headway (s)	1.570	1.593	2.129
Mean absolute value of jerk (m/s^3)	0.175	0.158	0.934

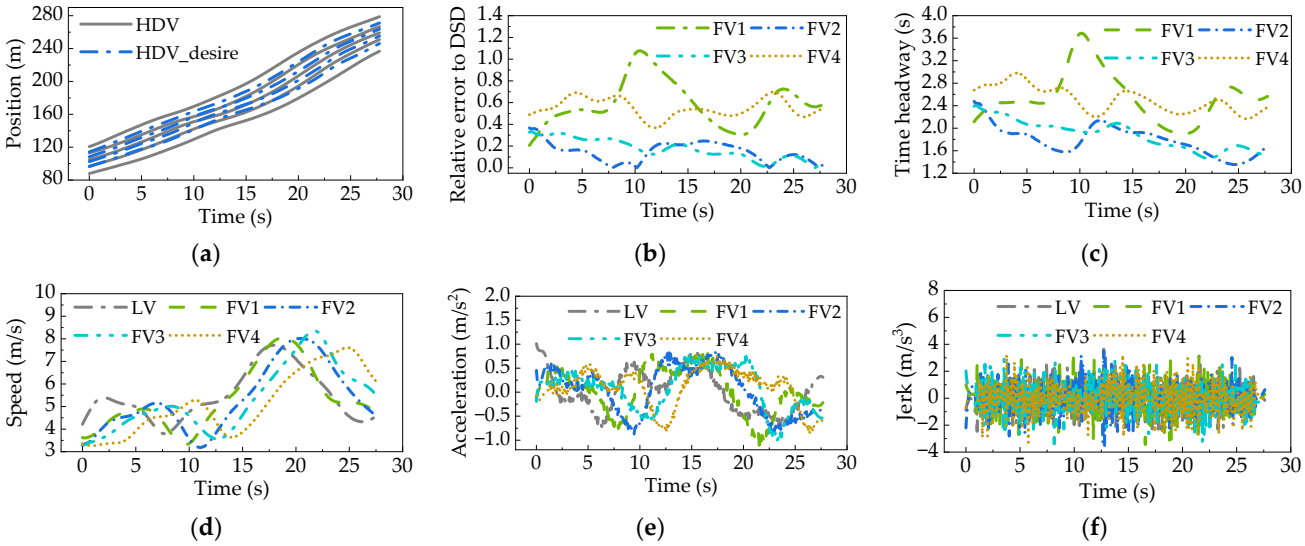

Figure 11. The result of a randomly selected platoon-following experiment based on HDV: (**a**) Driving trajectory; (**b**) The relative error to DSD; (**c**) Time headway; (**d**) Speed; (**e**) Acceleration; (**f**) Jerk values.

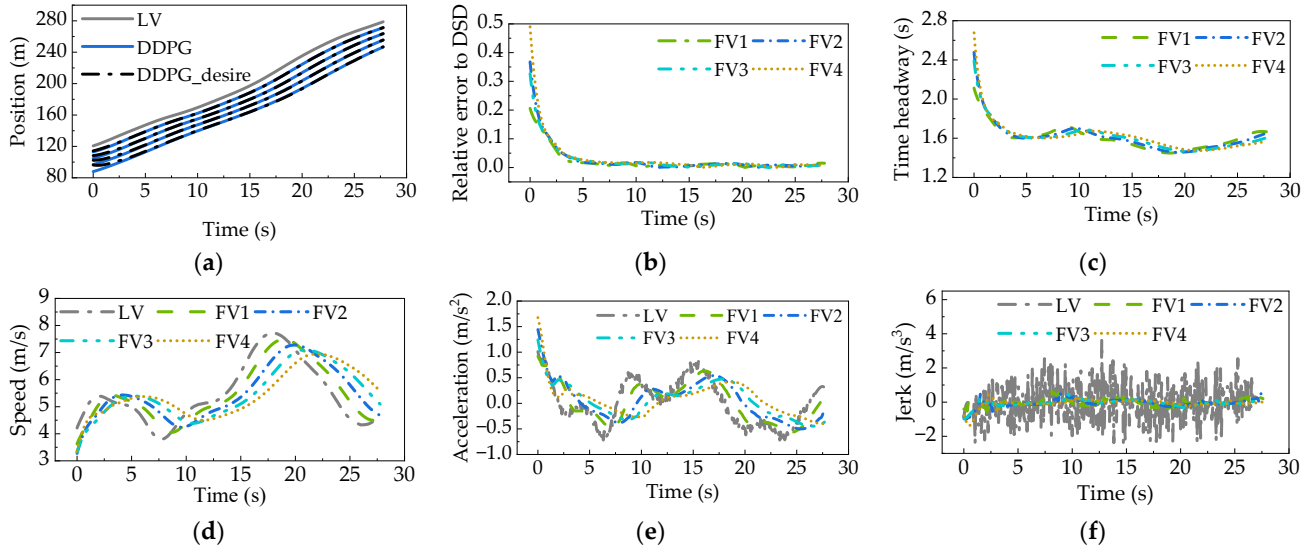

Figure 12. The result of a randomly selected platoon-following experiment based on DDPG: (**a**) Driving trajectory; (**b**) The relative error to DSD; (**c**) Time headway; (**d**) Speed; (**e**) Acceleration; (**f**) Jerk values.

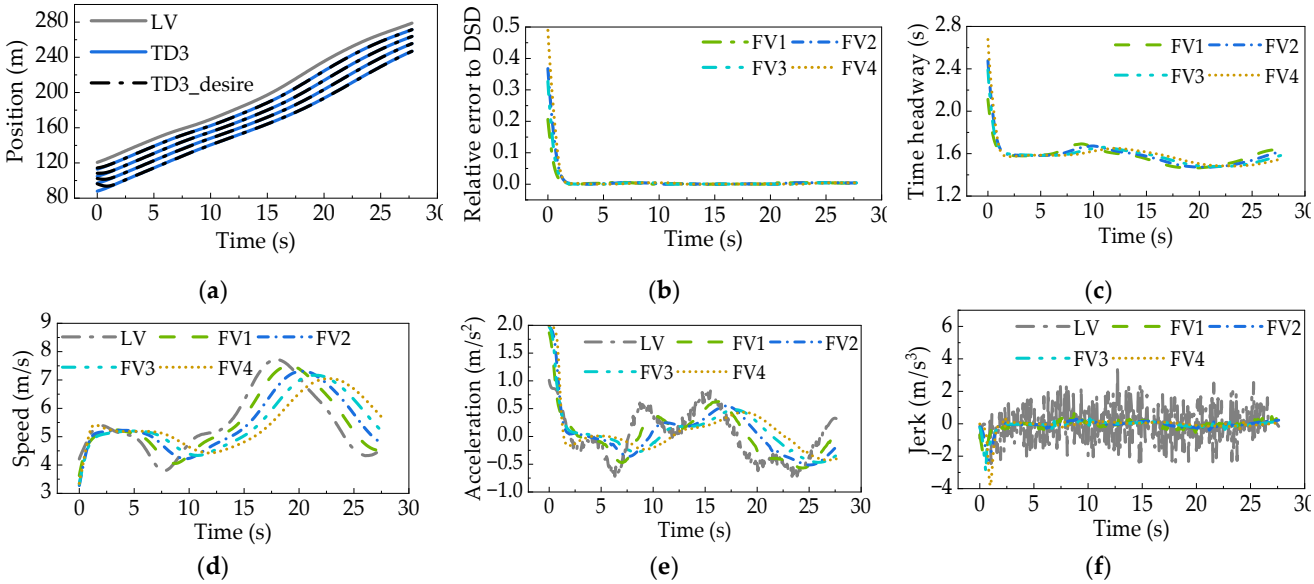

Figure 13. The result of a randomly selected platoon-following experiment based on TD3: (**a**) Driving trajectory; (**b**) The relative error to DSD; (**c**) Time headway; (**d**) Speed; (**e**) Acceleration; (**f**) Jerk values.

4. Conclusions

This study proposes a high-accuracy, high-efficiency, and comfortable CF strategy based on the TD3 for wide-to-narrow road sections in urban expressways. The results indicate that the following vehicle using the TD3, compared to the HDV and DDPG, can accurately drive according to the DSD while maintaining high traffic efficiency and ride comfort. In the test dataset of CF and platoon-following simulations, the traffic efficiency and ride comfort increased by over 29% and 60%, respectively, when using the TD3 and DDPG, compared to the HDV. The TD3-based strategy exhibited the highest control accuracy, with mean relative errors between the following distance and DSD during driving being 0.96% and 1.10%, respectively. The primary errors were due to the initial distance discrepancy with the DSD. When the initial distance equals the DSD, the vehicle using the TD3 drives according to the DSD immediately, maintaining a low jerk value and high ride comfort. If the initial distance is less or more than the DSD, the vehicle using the TD3 will decelerate or accelerate to reach the DSD, with a brief jerk value fluctuation during this transition. Once the DSD is achieved, the jerk value remains small and stable near zero, indicating that the TD3-based strategy significantly enhances the ride comfort under varying conditions.

Future research can expand on this study in several ways. Firstly, this study considers two datasets of urban expressways with wide-to-narrow road sections, and additional similar scenarios can be incorporated. Secondly, while this study focuses on longitudinal following strategies, future research could include lateral control and robustness studies of deep reinforcement learning CF strategies. Lastly, this study relies on CF trajectory data for simulation tests, and further hardware-in-the-loop or real vehicle platform tests could be conducted to validate the proposed strategy's effectiveness.

Author Contributions: Conceptualization, P.Q. and F.W.; Methodology, F.W.; Software, F.Y.; Validation, P.Q., F.W. and X.L.; Resources, P.Q.; Data curation, F.W. and S.B.; Writing—original draft preparation, F.W.; Project administration, P.Q.; Funding acquisition, P.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guangxi Science and Technology Major Project (grant numbers: GuikeAA22068061; GuikeAA22068060) and Guangxi Natural Science Foundation Project (grant number: 2019JJA160121).

Data Availability Statement: Not applicable.

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

References

1. Yeo, H.; Skabardonis, A. Understanding Stop-and-Go Traffic in View of Asymmetric Traffic Theory. In *Proceedings of the 18th International Symposium of Transportation and Traffic Theory, Hong Kong, China, 16–18 July 2009*; Hong Kong Polytech University: Hong Kong, China, 2009.
2. Goñi-Ros, B.; Schakel, W.J.; Papacharalampous, A.E.; Wang, M.; Knoop, V.L.; Sakata, I.; van Arem, B.; Hoogendoorn, S.P. Using Advanced Adaptive Cruise Control Systems to Reduce Congestion at Sags: An Evaluation Based on Microscopic Traffic Simulation. *Transp. Res. Part C-Emerg. Technol.* **2019**, *102*, 411–426. [CrossRef]
3. Saifuzzaman, M.; Zheng, Z. Incorporating Human-Factors in Car-Following Models: A Review of Recent Developments and Research Needs. *Transp. Res. Part C-Emerg. Technol.* **2014**, *48*, 379–403. [CrossRef]
4. Martin, T.; Ansgar, H.; Dirk, H. Congested Traffic States in Empirical Observations and Microscopic Simulations. *Phys. Rev. E* **2000**, *62*, 1805–1824.
5. Newell, G.F. A Simplified Car-Following Theory a Lower Order Model. *Transp. Res. Part B-Methodol.* **2002**, *36*, 195–205. [CrossRef]
6. Tian, J.; Li, G.; Treiber, M.; Jiang, R.; Jia, N.; Ma, S. Cellular Automaton Model Simulating Spatiotemporal Patterns, Phase Transitions and Concave Growth Pattern of Oscillations in Traffic Flow. *Transp. Res. Part B-Methodol.* **2016**, *93*, 560–575. [CrossRef]
7. Zhou, M.; Qu, X.; Li, X. A Recurrent Neural Network Based Microscopic Car Following Model to Predict Traffic Oscillation. *Transp. Res. Part C-Emerg. Technol.* **2017**, *84*, 245–264. [CrossRef]
8. Huang, X.; Sun, J.; Sun, J. A Car-Following Model Considering Asymmetric Driving Behavior Based on Long Short-Term Memory Neural Networks. *Transp. Res. Part C-Emerg. Technol.* **2018**, *95*, 346–362. [CrossRef]
9. Wu, C.; Li, B.; Bei, S.Y.; Zhu, Y.H.; Tian, J.; Hu, H.Z.; Tang, H.R. Research on Short-Term Driver Following Habits Based on GA-BP Neural Network. *World Electr. Veh. J.* **2022**, *13*, 171. [CrossRef]
10. Liao, Y.; Yu, G.; Chen, P.; Zhou, B.; Li, H. Modelling Personalised Car-Following Behaviour: A Memory-Based Deep Reinforcement Learning Approach. *Transp. A* **2022**, 1–29. [CrossRef]
11. Gao, H.; Shi, G.; Wang, K.; Xie, G.; Liu, Y. Research on Decision-Making of Autonomous Vehicle Following Based on Reinforcement Learning Method. *Ind. Robot.* **2019**, *46*, 444–452. [CrossRef]
12. Goerges, D. Relations between Model Predictive Control and Reinforcement Learning. *IFAC-PapersOnLine* **2017**, *50*, 4920–4928. [CrossRef]
13. Ye, Y.; Zhang, X.; Sun, J. Automated Vehicle's Behavior Decision Making Using Deep Reinforcement Learning and High-Fidelity Simulation Environment. *Transp. Res. Part C-Emerg. Technol.* **2019**, *107*, 155–170. [CrossRef]
14. Sun, M.; Zhao, W.; Song, G.; Nie, Z.; Han, X.; Liu, Y. DDPG-Based Decision-Making Strategy of Adaptive Cruising for Heavy Vehicles Considering Stability. *IEEE Access* **2020**, *8*, 59225–59246. [CrossRef]
15. Zhu, M.; Wang, Y.; Pu, Z.; Hu, J.; Wang, X.; Ke, R. Safe, Efficient, and Comfortable Velocity Control Based on Reinforcement Learning for Autonomous Driving. *Transp. Res. Part C-Emerg. Technol.* **2020**, *117*, 102662. [CrossRef]
16. Shi, H.; Zhou, Y.; Wu, K.; Wang, X.; Lin, Y.; Ran, B. Connected Automated Vehicle Cooperative Control with a Deep Reinforcement Learning Approach in a Mixed Traffic Environment. *Transp. Res. Part C-Emerg. Technol.* **2021**, *133*, 103421. [CrossRef]
17. Yan, R.; Jiang, R.; Jia, B.; Huang, J.; Yang, D. Hybrid Car-Following Strategy Based on Deep Deterministic Policy Gradient and Cooperative Adaptive Cruise Control. *IEEE Trans. Autom. Sci. Eng.* **2022**, *19*, 2816–2824. [CrossRef]
18. Qin, P.; Tan, H.; Li, H.; Wen, X. Deep Reinforcement Learning Car-Following Model Considering Longitudinal and Lateral Control. *Sustainability* **2022**, *14*, 16705. [CrossRef]
19. Chen, J.; Zhao, C.; Jiang, S.C.; Zhang, X.Y.; Li, Z.X.; Du, Y.C. Safe, Efficient, and Comfortable Autonomous Driving Based on Cooperative Vehicle Infrastructure System. *Int. J. Environ. Res. Public Health* **2023**, *20*, 893. [CrossRef]
20. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor-Critic Methods. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 January 2018.
21. Li, G.Q.; Gorges, D. Ecological Adaptive Cruise Control for Vehicles with Step-Gear Transmission Based on Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 4895–4905. [CrossRef]
22. Helbing, D.; Tilch, B. Generalized Force Model of Traffic Dynamics. *Phys. Rev. E* **1998**, *58*, 133–138. [CrossRef]
23. Zhang, G.; Wang, Y.; Wei, H.; Chen, Y. Examining Headway Distribution Models with Urban Freeway Loop Event Data. *Transp. Res. Rec. J. Transp. Res. Board* **2007**, *1999*, 141–149. [CrossRef]
24. Wan, Q.; Peng, G.; Li, Z.; Inomata, F.H.T. Spatiotemporal Trajectory Characteristic Analysis for Traffic State Transition Prediction near Expressway Merge Bottleneck. *Transp. Res. Part C-Emerg. Technol.* **2020**, *117*, 102682. [CrossRef]
25. Qin, P.; Li, H.; Li, Z.; Guan, W.; He, Y. A CNN-LSTM Car-Following Model Considering Generalization Ability. *Sensors* **2023**, *23*, 660. [CrossRef] [PubMed]
26. Wang, Z.; Bian, Y.; Shladover, S.E.; Wu, G.; Li, S.E.; Barth, M.J. A Survey on Cooperative Longitudinal Motion Control of Multiple Connected and Automated Vehicles. *IEEE Intell. Transp. Syst. Mag.* **2020**, *12*, 4–24. [CrossRef]