Transportmetrica A: Transport Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ttra21

Cooperative lane-changing in mixed traffic: a deep reinforcement learning approach

Xue Yao, Zhaocheng Du, Zhanbo Sun, Simeon C. Calvert & Ang ji

To cite this article: Xue Yao, Zhaocheng Du, Zhanbo Sun, Simeon C. Calvert & Ang ji (18 Apr 2024): Cooperative lane-changing in mixed traffic: a deep reinforcement learning approach, Transportmetrica A: Transport Science, DOI: 10.1080/23249935.2024.2343048

To link to this article: https://doi.org/10.1080/23249935.2024.2343048

TRANSPORTMETRICA A: TRANSPORT SCIENCE
https://doi.org/10.1080/23249935.2024.2343048

Cooperative lane-changing in mixed traffic: a deep reinforcement learning approach

Xue Yao[a], Zhaocheng Du[b], Zhanbo Sun[c,d], Simeon C. Calvert[a] and Ang ji [c,d]

[a]Department of Transport & Planning, Delft University of Technology, Delft, The Netherlands; [b]Department of Civil Engineering, McGill University, Montreal, Canada; [c]School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China; [d] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Chengdu, China

ABSTRACT

Deep Reinforcement Learning (DRL) has made remarkable progress in autonomous vehicle decision-making and execution control to improve traffic performance. This paper introduces a DRL-based mechanism for cooperative lane changing in mixed traffic (CLCMT) for connected and automated vehicles (CAVs). The uncertainty of human-driven vehicles (HVs) and the microscopic interactions between HVs and CAVs are explicitly modelled, and different leader-follower compositions are considered in CLCMT, which provides a high-fidelity DRL learning environment. A feedback module is established to enable interactions between the decision-making layer and the manoeuvre control layer. Simulation results show that the increase in CAV penetration leads to safer, more comfort, and eco-friendly lane-changing behaviours. A CAV-CAV lane-changing scenario can enhance safety by 24.5%–35.8%, improve comfort by 8%–9%, and reduce fuel consumption and emissions by 5.2%–12.9%. The proposed CLCMT promises advantages in the lateral decision-making and motion control of CAVs.

ARTICLE HISTORY
Received 12 June 2023
Accepted 8 April 2024

KEYWORDS
Connected automated vehicles (CAVs); cooperative lane-changing for mixed traffic (CLCMT); deep reinforcement learning (DRL); feedback mechanism

1. Introduction

The negative consequences of traffic, such as congestion, safety risks, and pollution, have substantial economic and social impacts, which are coming under increased scrutiny. With the aid of vehicle-to-everything (V2X) communication, connected and automated vehicle (CAV) technologies offer a promising solution to solve these problems (Calvert and van Arem 2020; A. Ji and Levinson 2020; Jin et al. 2023; Ren et al. 2017; Sun et al. 2023), as contrary to Human-driven vehicles (HVs), CAV driving manoeuvres can be designed and controlled for certain purposes (Dong et al. 2021). For instance, a proposed cooperative decision-making for mixed traffic (CDMMT) mechanism at ramp-merging sections can mitigate traffic conflicts, smooth acceleration/deceleration, and further increase throughput (Sun, Huang, and Zhang 2020). Certainly, before the Society of Automotive Engineers (SAE) level 4 or 5 automated vehicles become widespread, numerous challenges must be

CONTACT Zhanbo Sun zhanbo.sun@home.swjtu.edu.cn School of transportation and logistic, Southwest Jiaotong University, Chengdu, 610031, China

tackled such as reliable sensor perception, and motion planning. Efforts have been made to investigate intelligent machinery fault diagnosis (X. Li et al. 2022, 2023). Motion planning is a persistent challenge due to the unpredictable behaviours of the surrounding vehicles (sometimes mixed with both CAVs and HVs). Particularly in lane-changing behaviour, this task becomes more complex as both longitudinal and lateral behaviours need to be considered (A. Ji, Ramezani, and Levinson 2023a; Paul et al. 2021; Shi et al. 2019). This underscores the need to design robust motion planning strategies for autonomous vehicles to handle complex lane-changing scenarios.

One potential solution to this issue involves dividing motion planning into two distinct layers: decision-making and manoeuvre execution (Mirchevska et al. 2018). The decision-making layer is tasked with making high-level decisions, such as lane maintaining or changing, while the manoeuvre execution layer follows the planning and execute detailed movements (Sun, Huang, and Zhang 2020). Various methods, including game-theoretical approaches (A. Ji, Ramezani, and Levinson 2023b; D. Li and Pan 2022; Yu, Tseng, and Langari 2018), analytic hierarchy processes (Deng and Feng 2019), multilane cellular automaton models (Pan et al. 2021), and safety potential field theory (L. Li et al. 2020), have been implemented to model the decision-making process of lane-changing. For instance, cooperative controllers have been thoughtfully designed to address the total cost associated with merging manoeuvres (M. Wang et al. 2015). For manoeuvre execution, techniques such as quantal response equilibrium (Arbis and Dixit 2019) and fifth-degree polynomial curves (Shi et al. 2019) are commonly utilised. Some studies have also addressed lane-changing as an optimal control problem and used model predictive control (MPC) to tackle it (Hou et al. 2023; J. Ji et al. 2016; Rasekhipour et al. 2016). While these methods serve as a valuable theoretical foundation for addressing motion planning during lane-changing, they may encounter certain challenges related to scalability and adaptability in complex scenarios. These underscore the compelling need for alternative approaches to effectively address the dynamic lane-changing.

Recently, deep reinforcement learning (DRL) has achieved prominent success in various challenging areas, such as gaming and robotic control (Ha et al. 2023). The DRL agent enhances its policy through direct interaction with the environment, making it ideal for CAV motion planning where obtaining an accurate system model is difficult. Numerous studies have demonstrated the potential of DRL-based models to carry out lane-changing manoeuvres, following comfortable and safe-oriented trajectories (Lin, Li, and Jabari 2019; Sun et al. 2024; G. Wang et al. 2021; Xu et al. 2020). Some studies have simplified the problem and enhanced learning efficiency by discretizing the control space. For instance, Double Q-learning (DQN) has been used to learn vehicle speed control, considering three control actions: acceleration, deceleration, and maintenance (J. Li et al. 2020). Similar techniques have been employed in other studies, where the control space included lane-keeping and lane-changing (Bouadi et al. 2022; Shi et al. 2019). The Asynchronous Actor-Critic (A3C) method has been used to learn vehicle control policies, with a control action space comprising 32 discrete values (Jaritz et al. 2018). Although discretizing the action space boosts learning efficiency, it compromises control accuracy. To address this problem, one study used the Deep Deterministic Policy Gradient (DDPG) to describe lane-changing behaviour with continuous action in a model-free dynamic driving environment (P. Wang, Li, and Chan 2019). In a decision-making training and learning framework based on DRL, the average speed of lane-changing behaviour was improved by approximately 2.4%. To this end,

DRL has shown advantages in handling lane-changing issues of a single CAV in a pure CAV environment.

When it comes to mixed traffic, modelling lane-changing of CAVs becomes more complex as HVs present in the surrounding traffic (Sun, Huang, and Zhang 2020). The behaviour of HVs is highly stochastic, uncertain, and beyond direct control (Z.-C. Li, Huang, and Lam 2012). They may exhibit unpredictable behaviours during CAV lane-changing, such as adopting hostile movements to obstruct lane-changing of CAVs. There is a need to develop cooperative strategies for CAVs to handle HV uncertainty, reducing the risk of collisions. Additionally, multi-CAV control strategy deserves attention as it allows several CAVs to collaborate in dealing with HV uncertainty. This can provide more opportunities to effectively handle HV uncertainty, which helps to improve safety as well as traffic efficiency.

To bridge these research gaps, in this paper, a cooperative lane-changing in mixed traffic (CLCMT) mechanism based on deep reinforcement learning (DRL) is developed to facilitate optimal lane-changing strategies. Two DRL algorithms named Deep Deterministic Policy Gradient (DDPG) and Twin Delayed Deep Deterministic Policy Gradient algorithm (TD3) are employed to deal with lane-changing problems with continuous control. The novel contributions of the paper include: (i) The establishment of a feedback module to facilitate integration between the decision-making layer and the manoeuvre control layer. The utilities of all potential motion executions are pre-calculated, which guides the optimal decision-making; (ii) The uncertainty of HVs and the microscopic interactions between HVs and CAVs are explicitly considered in CLCMT. This provides a high-fidelity environment to facilitate DRL agent learning strategies that can deal with complex situations; (iii) Scenarios encompassing various common lane-changing situations in mixed traffic are evaluated, which enables a holistic analysis of lane-changing behaviour.

The remainder of this paper is structured as follows. Section 2 describes the methodology for the proposed CLCMT, encompassing problem description, algorithm preparation, and DRL modelling. Section 3 presents experiments and numerical results, including the training environment settings, results analysis and discussions. Section 4 summarises the key findings of this study and proposes directions for future research.

2. Methodology

In this section, we first introduce the cooperative lane-changing problem in mixed traffic (CLCMT). Then the DRL-based CLCMT mechanism is well-described, including MDP formulation, basic theory of DRL algorithms, the architecture of CLCMT, and a feedback module.

2.1. Problem description

Lane-change decision-making can be considered on strategic and tactical levels. Strategic decisions are motivated by long-term objectives, such as travel efficiency or route planning, and initiating lane-changing behaviour. On the other hand, tactical decision-making is employed when the target vehicle already intends to change lanes and must know when and how to change (Shi et al. 2019). We focus on decision-making on a tactical level where the target vehicle has already decided to change lanes, but should still decide which target lane and gap to choose. This decision is important as it guides the target vehicle to execute

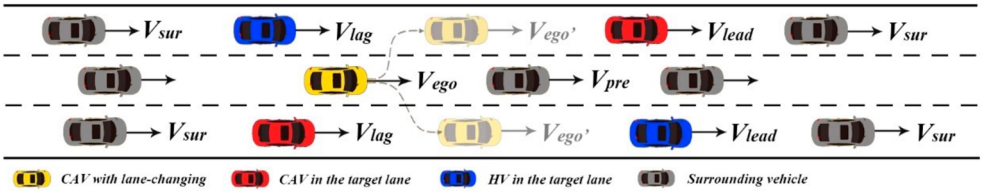


Figure 1. The cooperative lane-changing problem in mixed traffic: an illustrative example.

a lane change. However, due to the uncertainty in mixed traffic flow, e.g. HV obstructs lane-changing, this guidance may become inferior. Instead, if possible lane change execution can be captured in advance and the lane change decision can be optimised based on this information, potential failure of lane change can be reduced. Here comes our CLCMT mechanism. Specially, we design a feedback module to pre-calculate the utilities of potential manoeuvre execution and guide optimal decision-making. This subsection briefly describes the problem that the CLCMT mechanism needs to solve, details of the mechanism will be introduced in the following subsection.

In a typical three-lane highway scenario shown in Figure 1, the target vehicle (V_{ego} who has the intention to change lanes) in the middle lane can choose either the left lane or the right to change. V_{pre} represents the leading vehicle in the current lane. The red and blue vehicles are potential new leaders (V_{lead}) and followers (V_{lag}) in the target lane. V_{sur} represents leaders and followers of V_{lead} and V_{lag}. Vehicles coloured in yellow and red represent CAVs, and those coloured in blue and grey denote HVs. CAVs have the capability to strictly comply with the control strategy to accommodate the dynamic traffic environment, while HVs do not. Surrounding vehicles (V_{pre} and V_{sur}) are considered as a part of the training environment in the CLCMT mechanism. The fundamental task of the CLCMT is to learn a strategy to select the appropriate target lane and gap and then execute the motion safely, efficiently, comfortably, and environmentally friendly.

As illustrated in Figure 2, each lane-changing manoeuvre involves three vehicles, vehicle V_{ego} in the current lane, the new leader V_{lead} and the new follower V_{lag} in the target lane. In mixed traffic, there are four potential leader-follower combinations: CAV-CAV, CAV-HV, HV-CAV, and HV-HV, referred as Case 1 – Case 4. V_{ego}, V_{lead}, and V_{lag} are potential controlled objects in CLCMT. Motions of CAVs follow the cooperative manoeuvre control (CMC). If one

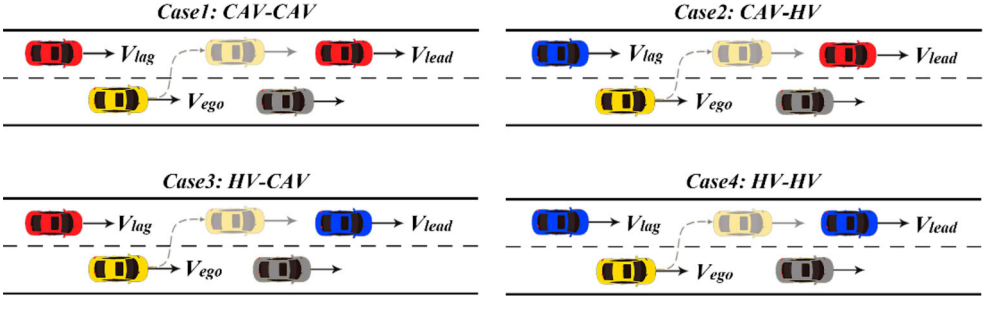


Figure 2. Compositions of leader-follower types.

or both of V_{lead} and V_{lag} are HVs, they are given cooperative manoeuvre control recommendations (CMCR), such as suggested speed and acceleration. Assumed that the HV(s) of them adopt the CMCR with a probability of p.

2.2. DRL-based CLCMT mechanism

2.2.1. MDP formulation

Lane-change manoeuvre control is formulated as a Markov Decision Process (MDP) in this paper, which is defined by 5-tuple (S, A, P, R, λ). S represents the set of system states, A the set of actions, P the state transition probability, R the reward function, and λ the discount factor. Based on the principles of MDP, the state space, action space, and reward function of the proposed DRL-based CLCMT are provided as follows.

State Space: The state of agents includes the position and speed of the target vehicle $O_1(x_{ego}, y_{ego}, v_{xego}, v_{yego})$, the leading vehicle $l_1(x_{lead}, y_{lead}, v_{lead})$ and the following vehicle $l_2(x_{lag}, y_{lag}, v_{lag})$. Additionally, it includes the surrounding vehicles $s_1(x_{sur1}, y_{sur1}, v_{sur1})$ and $s_2(x_{sur2}, y_{sur2}, v_{sur2})$, perceived as environmental data. The state space of DRL policy in the four scenarios can be represented as:

$$s = (O_1(x_{ego}, y_{ego}, v_{xego}, v_{yego}), l_1(x_{lead}, y_{lead}, v_{lead}), l_2(x_{lag}, y_{lag}, v_{lag}), s_1(x_{sur1}, y_{sur1}, v_{sur1}), s_2(x_{sur2}, y_{sur2}, v_{sur2})) \in \mathcal{S} \tag{1}$$

Action Space: The action space of each object refers to the available range of acceleration and deceleration that an object can undertake. For Case 1, the actions during lane-changing involve the actions of three vehicles, as shown below:

$$a_1 = (a_{egox}, a_{egoy}, a_{lead}, a_{lag}) \in \mathcal{A}_1 \tag{2}$$

Here, $a_{egox}, a_{lead}, a_{lag}$ represent the longitudinal acceleration of the target vehicle, new leader, and new follower, while a_{egoy} represents the lateral acceleration of the target vehicle. Action space for Cases 2 to 4 can be expressed as in Equation (2) if the HV(s) comply with CMCR. Otherwise, the control spaces are represented as follows:

$$a_2 = (a_{egox}, a_{egoy}, a_{lead}) \in \mathcal{A}_2 \tag{3}$$

$$a_3 = (a_{egox}, a_{egoy}, a_{lag}) \in \mathcal{A}_3 \tag{4}$$

$$a_4 = (a_egox, a_egoy) \in \mathcal{A}_4 \tag{5}$$

Reward Function: Precise control is crucial in driving, as any deviations can have serious consequences. Learning driving behaviours without pre-existing knowledge can be difficult, which emphasises the significance of formulating rational reward functions. In this study, we incorporate elements of safety, comfort, fuel consumption, and emissions into the reward function.

The safety-related reward is given in Equation (6a). As the DRL agent's actions are driven by the pursuit of rewards at each step, the first component of Equation (6a) is designed to encourage the DRL agent to keep moving forward by offering reasonable rewards. The second component is designed to punish collision. When the agent fails to meet security

conditions, i.e. $d_{tar} \leq l_{veh}$, a large negative reward will be given. d_{tar} represents the current distance between V_{lead} and V_{ego}, calculated by the positional difference between two vehicles and a minimum distance d_0, see Equation (6b).

$$\mathcal{R}_s = \begin{cases} \alpha \sum_{i=1}^{I} \Delta x_{ego} + \beta, & \text{otherwise} \\ -c, & d_{tar} \leq l_{veh} \end{cases} \tag{6a}$$

$$d_{tar} = x_{lead} - x_{ego} + (v_{lead} - v_{ego})t + \frac{1}{2}(a_{lead} - a_{ego})t^2 + d_0 \tag{6b}$$

here, Δx_{ego} represents differences of x_{ego} between two timestamps, l_{veh} stands for the length of V_{ego}; t is the time at which the lane-change manoeuvre occurs; α, β, κ, and c are coefficients.

Smooth transitions during lane-changing can provide comfortable experience for CAV users. As such, a comfort reward function $\mathcal{R}_c$ is designed to penalise abrupt jerks and extensive yaws, as shown below.

$$\varphi = \frac{\mathrm{d}a}{\mathrm{d}t} \tag{7a}$$

$$\theta = \arctan \frac{v_{egoy}(t)}{v_{egox}(t)} - \arctan \frac{v_{egoy}(t-1)}{v_{egox}(t-1)} \tag{7b}$$

$$\mathcal{R}_c = -b_1|\varphi| - b_2|\theta| \tag{7c}$$

where φ stands for the acceleration/deceleration changing rate of controlled vehicle(s). θ indicates the yaw changing rate, calculated by the differences between yaws of two adjacent timestamps. b_1 and b_2 are coefficients.

According to Nie and Li (2013), the fuel consumption F and emissions E_{CO} can be calculated using Equations (8) and (9), respectively. The total fuel emission for a certain length of the trip (l_t) is shown in Equation (10). From the literature, E_{CO} is found to be approximately equal to three times F. Thus, the fuel consumption and emissions can be calculated using $4T_F$. For details on the models as well as their coefficients and corresponding default values please refer to Nie and Li (2013). Then the reward function for assessing fuel consumption and emissions can be expressed as Equation (11). κ serves as a modifier or adjustment coefficient. The optimal lane-changing strategy is trained by considering the fuel emissions of all vehicles present in the scenarios.

$$F(v,a) = \frac{f}{v} = \frac{\phi}{\lambda}\left(\sum_{i=0}^{3} \alpha_i v^{i-1} + \beta a\right) \tag{8}$$

$$E_{CO_2}(v,a) = \frac{e_{CO_2}}{v} = \gamma_1 F(v,a) + \frac{\gamma_0}{v} \tag{9}$$

$$T_F = l_t F(v_1, 0) + \phi_a \sigma_1 + \phi_a \sigma_2 + \phi_0 \sigma_3 \tag{10}$$

$$\mathcal{R}_f = -\kappa T_F \tag{11}$$

The positional deviation reward is designed to effectively guide the DRL agent (V_{ego}) to promote correct lane-changing directions and be alignment with the centreline of the target

lane. Lane-changing finishing within 0.5 m from the centreline of the target lane is considered an effective strategy. Before reaching this range, the DRL agent receives linearly increasing rewards for continuous lateral movement in the correct direction, as described in the second part of Equation (12). After achieving this effective range, the closer to the centreline, the more rewards can be obtained. Thus, a U-shaped function (the right half of the u-function) is designed to encourage the agent to explore superior strategies, illustrated in the first part of Equation (12). ϱ, δ, ζ, and ω are constants, serving as tuning parameters.

$$\mathcal{R}_l = \begin{cases} \omega|\Delta d_{lat}|, & \text{otherwise} \\ \varrho(|\Delta d_{lat}| - \theta)^2 + \zeta, & |\Delta d_{lat}| \leq 0.5 \end{cases} \tag{12}$$

The overall reward, $\mathcal{R}$, is the aggregate of all previously mentioned rewards, as depicted in Equation (13). The determination of all coefficients within the reward is achieved through refined sensitivity analysis during pre-training experiments. Note that we use a simple sum to indicate that all types of rewards are taken into account, while their weights significantly change in the iterative refinement process.

$$\mathcal{R} = \mathcal{R}_s + \mathcal{R}_c + \mathcal{R}_f + \mathcal{R}_l \tag{13}$$

2.2.2. DRL algorithms

The formulated MDP in CLCMT consists of continuous state and action spaces. This is difficult to solve by classical RL algorithms, such as *Q* learning, due to their poor scalability features (Sutton and Barto 2018). By leveraging the generalisation and fitting capability of DNNs, DRL algorithms have shown good performance when dealing with this challenge.

Deep Deterministic Policy Gradient (DDPG) is an advanced variant of the Deterministic Policy Gradient (DPG) model, employed in continuous control tasks, including autonomous vehicles (Liao et al. 2022). It introduces two components including:

Critic Q-function: $Q(s, a \mid \theta^Q)$, which evaluates state-action pairs and updates parameters using Temporal Difference (TD) loss (Equation (14)). It emphasises long-term rewards and stability during training.

$$L_k = \frac{1}{N} \sum \left(y_i - Q(s_i, a_i \mid \theta^Q) \right)^2 \tag{14}$$

Actor policy function: $\mu(s \mid \theta^\mu)$, which translates states into actions and updates its parameters via the policy gradient algorithm (Equation (15)).

$$\nabla_{\theta^\mu} \mathcal{J} \approx \frac{1}{N} \sum_{i=1}^{N} \left[\nabla_a Q(s_i, \mu(s_i) \mid \theta^Q) \nabla_{\theta^\mu} \mu(s_i \mid \theta^\mu) \right] \tag{15}$$

DDPG employs two additional target networks, i.e. Critic network Q' and Actor network μ', to enhance training stability (Equations (16) and (17)). These networks gradually update their parameters using the hyperparameter τ to improve learning stability (Nguyen, Nguyen, and Nahavandi 2020). However, DDPG sometimes faces challenges, such as overestimation issues and sensitivity to hyperparameters (Fujimoto, Hoof, and Meger 2018;

Zhou et al. 2021).

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \tag{16}$$

$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \tag{17}$$

To address the limitations of DDPG, a cutting-edge algorithm known as Twin-Delayed DDPG (TD3) is developed (Fujimoto, Hoof, and Meger 2018). The network structure of TD3 is demonstrated in Figure 3. In TD3, two target critic networks compute two Q values, and the smaller one replaces the Bellman equation as the TD target (Equation (18)).

$$y_i = r_i + \gamma \min_{j=1,2} Q_j'\left(s', a' \mid \theta_j^{Q'}\right) \tag{18}$$

This approach provides unbiased Q value estimates for Actor-provided actions, effectively mitigating overestimation. The Actor network updates less frequently than the Critic network to reduce error accumulation (Zhou et al. 2021). Additionally, truncated normal distribution noise is introduced to the target action to balance bias and variance, preventing overfitting (Equation (19)).

$$\tilde{a} \leftarrow \mu'(s' \mid \theta^{\mu'}) + \epsilon, \epsilon \sim \text{clip}(N(0,\sigma), -c, c), \quad c > 0 \tag{19}$$

2.2.3. Architecture of DRL-based CLCMT

Based on the MDP formulation and basic knowledge of DRL algorithms, the DRL-based CLCMT mechanism is constructed and shown in Figure 4. The process can be outlined in

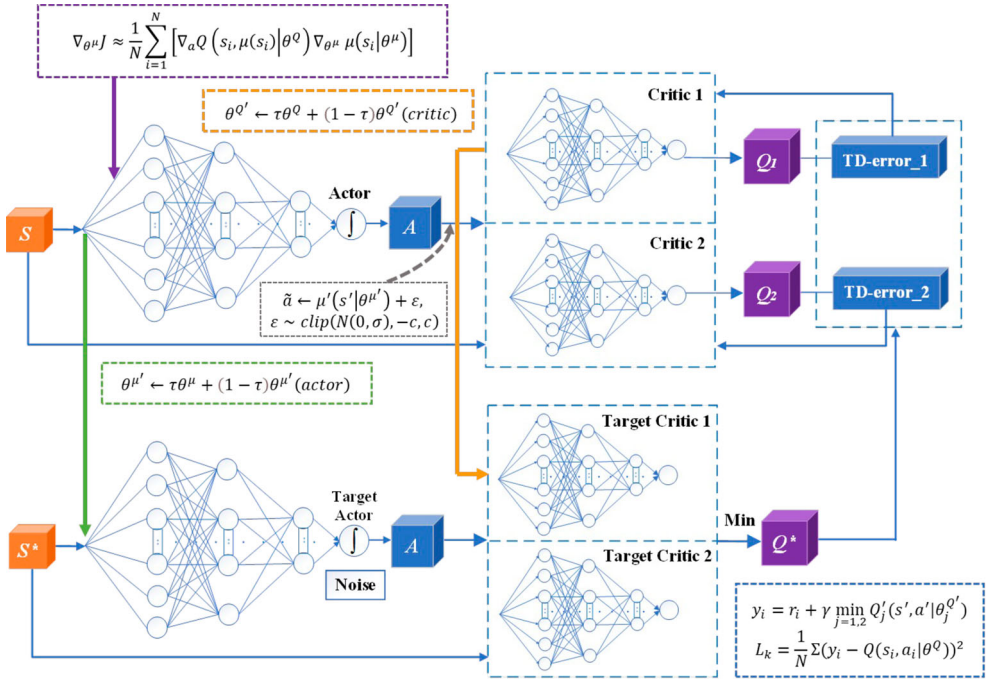


Figure 3. Network structure of TD3.

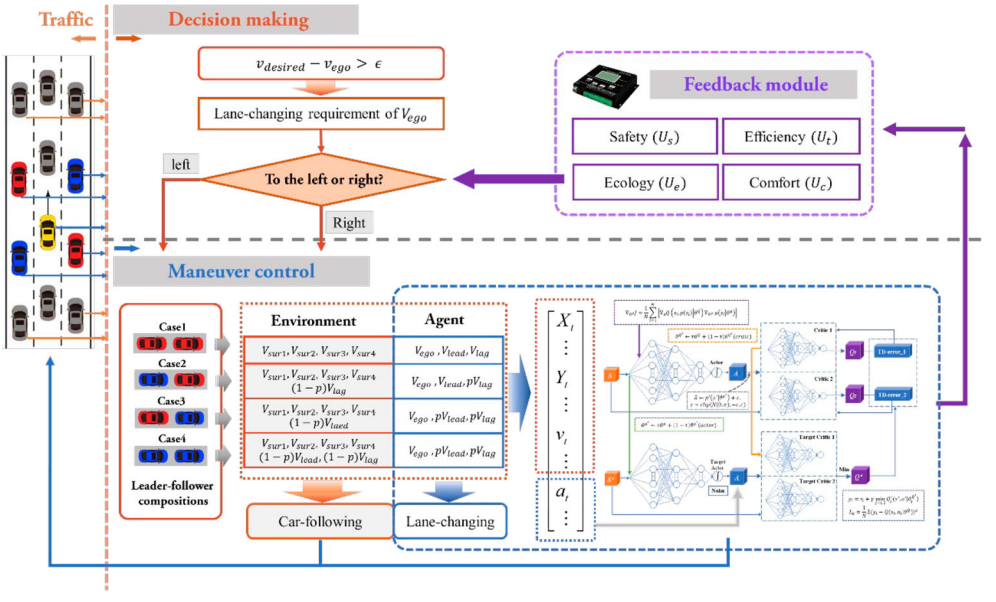


Figure 4. Architecture DRL-based lane-changing mechanism: an example of TD3.

several steps. First, the current speed and desired speed of the target vehicle are acquired. If the former is less than the latter by a threshold of ϵ, a request for lane-changing is triggered. Next, the DRL agent receives traffic states from environment, including leader-follower compositions as well as their gaps in adjacent lanes. These states serve as potential lane-changing scenarios. Then, detailed manoeuvre control actions behind each potential scenario are executed. This is achieved by DRL-based policies to learn actions including two-dimensional accelerations of V_{ego} (and longitudinal acceleration of V_{lead} and V_{lag} if they are controlled). Based on various lane-changing strategies learned by DRL algorithms, the feedback module computes the utilities of each cooperative lane-changing strategy under different scenarios. According to pre-calculated utilities (including safety, efficiency, comfort, and ecology) and a personalised evaluation function, the feedback module recommends the optimal lane-changing strategy for the decision-making layer where the lane-changing strategy is determined.

Specifically, the DRL agent learns the two-dimensional accelerations of the controlled vehicles by DRL-based policies and executes actions and updates position and speed by the following kinematic models:

$$x(t_{i+1}) = x(t_i) + v_x(t_i)\Delta t + \frac{1}{2} a_x (\Delta t)^2 \tag{20}$$

$$y(t_{i+1}) = y(t_i) + v_y(t_i)\Delta t + \frac{1}{2} a_y (\Delta t)^2 \tag{21}$$

$$v_x(t_{i+1}) = v_x(t_i) + a_x \Delta t \tag{22}$$

$$v_y(t_{i+1}) = v_y(t_i) + a_y \Delta t \tag{23}$$

where x, y are the positions, v_x and v_y are longitudinal and lateral speed, respectively. a_x and a_y are longitudinal and lateral acceleration, respectively. $\Delta t = t_{i+1} - t_i$, denotes the time step.

Uncontrollable vehicles (HVs) follow the speed and position updating rules based on the car-following model. The well-known Intelligent Driver Model (IDM) is regarded as a complete and accident-free model that can provide plausible behaviours in almost all single-lane traffic situations (Sun et al. 2021), which was adopted in the experiment; see Equation (24):

$$a(t_{i+1}) = a_1 \left[1 - \left(\frac{v(t_i)}{v_0} \right)^{\delta} \right] - \left(\frac{s^*(v(t_i), \Delta v)}{s_0} \right)^2 \tag{24}$$

$$s^*\left(v(t_i), \Delta v\right) = s_0 + \max\left(0, v(t_{i+1})\, T + \frac{v(t_i), \Delta v}{2\sqrt{a_1\, b_1}}\right) \tag{25}$$

here a_1 is the maximum acceleration/deceleration of the follower, δ is the acceleration index, v_0 is the desired speed and s_0 is the minimum gap. $s^*(v(t_i), \Delta v)$ means the desired gap, which is a function of $v(t_i)$ and Δv, as shown in Equation (25). T is the safety time gap and b_1 is the comfortable deceleration.

2.2.4. The feedback mechanism

As an important part of the proposed CLCMT mechanism, the feedback mechanism calculates utilities of rewards obtained from DRL model training. The time required to complete the lane-changing process, measured by the number of time steps, is utilised to evaluate the efficiency of lane-changing strategy ($\mathcal{U}_t$). The crash rate, calculated as the proportion of collisions among the total number of training episodes after convergence, serves as an indicator of safety level ($\mathcal{U}_s$). The designed comfort reward $\mathcal{R}_c$ is employed to assess the comfort level, denoted as $\mathcal{U}_c$. Additionally, the reward $\mathcal{R}_f$ which is adopted to quantify fuel consumption and emissions during the lane-changing process, serves as an evaluation of the ecology utility ($\mathcal{U}_e$).

Utilities of all potential lane-changing scenarios under a certain traffic state are calculated, serving as a data source for decision-making. An evaluation function is introduced to facilitate optimal lane-changing strategy, see Equation (26). This function can be customised to yield specific lane-changing strategies with two advantages. First, it allows for personalised relationships among the four types of utilities, such as being linear or nonlinear. Second, it can be easily used to design user-oriented decision-making strategies by adjusting the coefficients associated with different utilities. For instance, an increase in the weight assigned to $\mathcal{U}_s$ signifies a safety-centric strategy, while a larger coefficient for $\mathcal{U}_e$ indicates an environmentally conscious approach. After such evaluation, the feedback module recommends the user-oriented optimal lane-changing strategy to the decision-making layer where the lane-changing direction and gap are determined. This decision serves as a command to execute lane-changing in the manoeuvre control layer.

$$f = (\mathcal{U}_t, \mathcal{U}_s, \mathcal{U}_c, \mathcal{U}_e) \tag{26}$$

3. Experiments and numerical results

In this section, we first introduce experiment preparation, including training environmental settings, parameter tuning and model pre-training. Then the DRL models are trained in various traffic scenarios. Based on the training results, discussions are conducted on the effectiveness of the proposed CLCMT mechanism.

3.1. Experiment preparation

3.1.1. Training environmental settings

For the convenience of customisation to align precisely with research requirements and flexibility to conduct experiments, we develop a gym-based learning platform using Python, including various training environments and dynamic visualisation process [1]. According to different leader-follower compositions in mixed traffic, various lane-changing scenarios were considered. As shown in Figure 5, a 300 m road with two lanes was constructed. Each lane is 3.75 m wide. The y-axis and x-axis represent the longitudinal and lateral movements of driving, respectively. Each vehicle's length is denoted by l_{veh} and the spacing (distance between the leader and the follower) is represented by L. The position of each vehicle is illustrated by the (x, y) coordinates at time t. Assume that all vehicles, except for the target vehicle, remain close to the centreline of their lanes.

The parameters for the training environment were set as follows: a time step of 0.1 s, with the target vehicle's initial coordinates set at (60, 1.875). The gap L in the initial traffic state was chosen based on typical spacings found in highway traffic flow. Based on empirical experiences, the original spacing L_{ori} is set to 30 m with the initial speed of the traffic flow being 15 m/s. Uncontrolled vehicles, i.e. HVs, update their state by following IDM car-following model with corresponding parameters listed in Table 1. The probability p of a human-driven vehicle (HV) being willing to cooperate was set to 0.5.

As for DRL algorithms, the detailed structure of Actor and Critic in TD3 is shown in Figure 6, each comprising a series of fully connected layers with specific configurations. Specifically, the Actor network has three layers with a state_dim of 16 and neurons of 256. ReLU is used as an activation following the fully connected layer which is a linear function, and a tanh activation function is applied to the output in the Actor network to ensure the

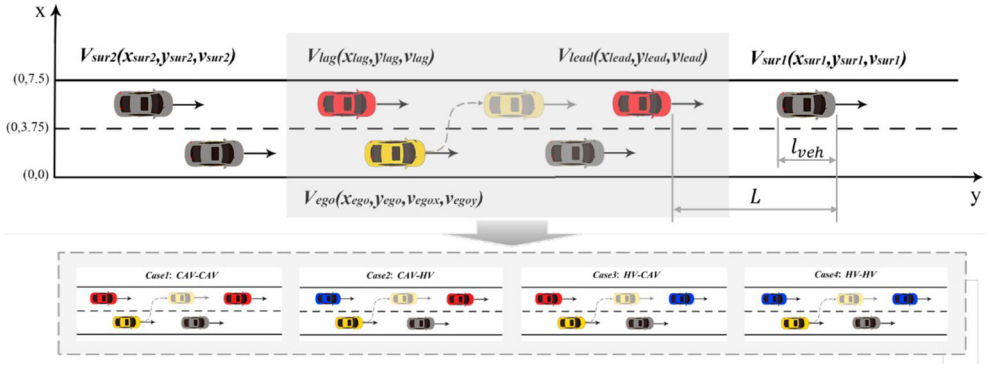
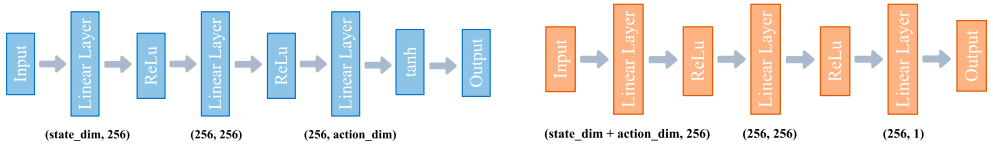


Figure 5. Lane-changing manoeuvre control scenarios.

Table 1. Parameters of car-following model.

Name	Description	Unit	Value
a_0	Maximum acceleration/deceleration	m/s^2	3
v_0	Desired speed	m/s	20
s_0	Minimum gap	m	2
δ	Acceleration index	–	4
T	Safety time gap	s	1
b_0	Comfortable deceleration	m/s^2	1.5



(a) Detailed structure of Actor. (b) Detailed structure of Critic.

Figure 6. Detailed structure of Actor and Critic in TD3. (a) Detailed structure of Actor and (b) Detailed structure of Critic.

actions are bound within the appropriate range. The Critic network, responsible for evaluating the action-value pairs, follows a similar structure but incorporates action inputs in its architecture, as typically required in actor-critic methods. For some hyperparameters, the DRL models were trained with a batch size of 256, and the length of the burn-in period of reply buffer was 25,000. The evaluation frequency was 5000 steps, meaning that we evaluated the DRL policy every 5000 steps. The reward discounted factor was 0.99 to balance immediate and future rewards. Both the exploration noise and smoothing regularisation noise were set to 0.5, aiming to improve robustness. The maximum timestamp was set to be 3×10^5 to guarantee convergence. To minimise the impact of perceptual information errors and to improve the robustness of learning strategies, we also added some noise when resetting the environment to the original states. The reset noises were uniformly distributed within the range of $[0, 1]$, $[0, 0.5]$, and $[0, 2]$ for x, y and v, respectively. To avoid unreasonable actions and facilitate optimal lane-changing strategies during the training process, the action space of DRL was set as $[-3, 3]$ for longitudinal direction and $[-1, 1]$ for lateral direction.

3.1.2. Pre-training and parameter tuning

The overall reward function represents a multi-objective optimisation problem, as it combines four distinct objectives. In our CLCMT problem, these objectives correspond to various aspects of driving behaviour, such as collision avoidance, forward progress, lane change rewards, etc. Determination of weights for these different types of rewards is a critical aspect of fine-tuning DRL agents. We have approached this by considering the relative importance of each objective in achieving desired driving behaviours. Trial-and-error experimentation and hyperparameter tuning were conducted to determine parameters, which helps to ensure that the DRL agent optimally balances the proposed objectives, leading to effective and safe cooperative lane-changing behaviour. Table 2 displays the ablation studies on key parameters of reward functions. Note that the collision penalty c is critical for the convergence of DRL algorithm. Results show that the coefficients of comfort

Table 2. Ablation studies on key parameters.

c	b_1	b_2	ω	Convergence	Crash rate	Trajectories
100,000	20	100	2000	Yes	2.40%	(a)
50,000	20	100	2000	No	43.25%	(b)
10,000	20	100	2000	No	89.67%	(c)
100,000	10	50	2000	Yes	4.23%	(d)
100,000	30	150	2000	Yes	5.01%	(e)
100,000	20	100	1000	Yes	4.82%	(f)
100,000	20	100	3000	Yes	3.94%	(g)

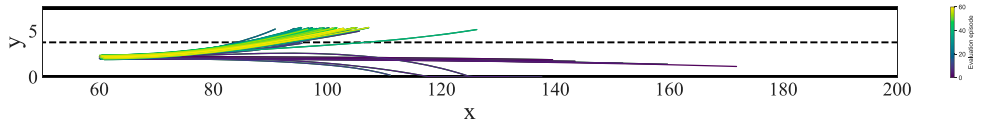
and lane changes have little effect on the convergence of algorithms and the safety of the learned strategies. However, as shown in Figure 7, these parameters affect the agent's exploration process and the final strategies. For example, if the comfort coefficients are too large, the agent tends to extend the lane-changing process to ensure comfort, thus learning an inefficient strategy, see Figure 7(e). Conversely, a large lane-changing coefficient can result in unreasonable lane-changing trajectories, as displayed in Figure 7(g). Among these fine-tuning experiments, we found the optimal combination of parameters with reasonable trajectories and a small crash rate, as shown in Table 2 and Figure 7(a). Finally, the optimal combination of parameters is found, which is shown in Table 3.

To further ensure the agent learned reasonable actions during training, the distribution of average acceleration and velocity in each episode are shown in Figure 8. Note that the acceleration is mostly distributed in (1.4, 2.6), with small numbers near 0, see Figure 8(a). There are a few negative values, which may caused by failed attempts during training. Velocity roughly performs as a normal distribution with a mean of 13 m/s, as shown in Figure 8(b). Some examples of trajectories learned by the DRL agents are shown in Figure 9, which illustrates that both the accelerations and velocities of the target vehicle display reasonable actions.

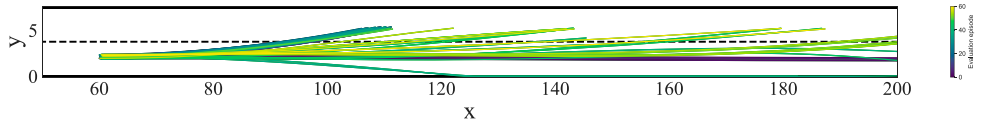
The training of the four cases was conducted on a laptop equipped with an AMD R7-5800H 3.2 GHz processor and NVIDIA 3060 GPU. The pre-training process is illustrated in Figure 10. The colour from blue to yellow represents the increase in training episodes. Note that the target vehicle performs smooth lane-changing trajectories based on the proposed DRL-based CLCMT mechanism. Although there were inappropriate lane-changing direction attempts at the beginning, these errors were corrected during the learning process. After learning a correct lane-changing direction, it continued to optimise the trajectories, such as seeking a shorter trajectory with less time or a more comfort strategy. Finally, optimal lane-changing trajectories were explored, as the yellow lines show in Figure 10. Figure 10(a,b) represent trajectories learned by TD3 and DDPG, respectively. Observed that both the DRL algorithms explored good lane-changing trajectories at the end. Notably, TD3 made fewer errors during the exploration, indicating its advantage in solving CLCMT problem.

3.2. Results and discussions

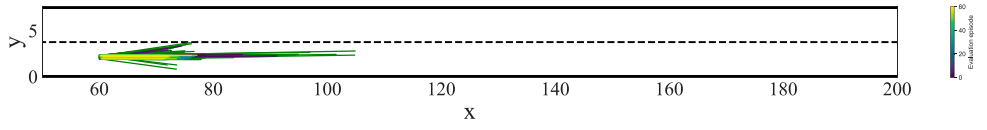
Based on the aforementioned settings, the DRL model was trained in the established lane-changing scenarios. In this subsection, we analyse these training results and discuss various utilities calculated in the feedback module.



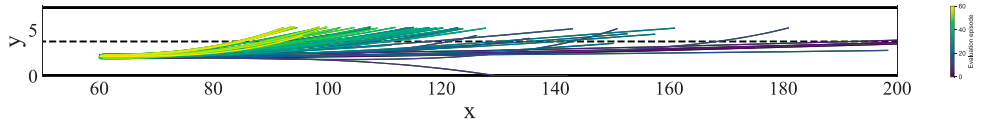
(a) $c = 100000, b_1 = 20, b_2 = 100, \omega = 2000$



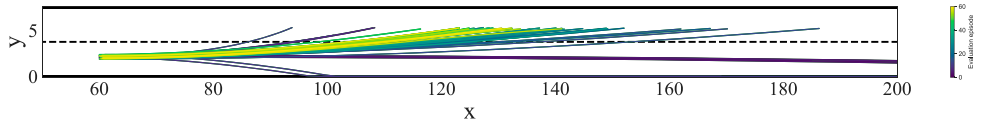
(b) $c = 50000, b_1 = 20, b_2 = 100, \omega = 2000$



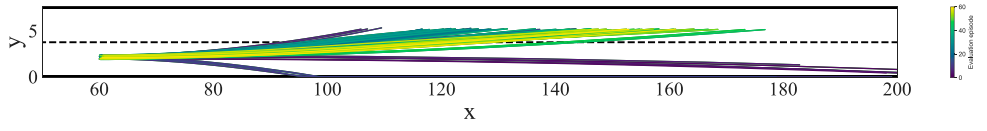
(c) $c = 10000, b_1 = 20, b_2 = 100, \omega = 2000$



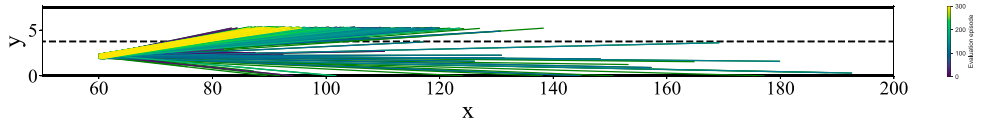
(d) $c = 100000, b_1 = 10, b_2 = 50, \omega = 2000$



(e) $c = 100000, b_1 = 30, b_2 = 150, \omega = 2000$



(f) $c = 100000, b_1 = 20, b_2 = 100, \omega = 1000$

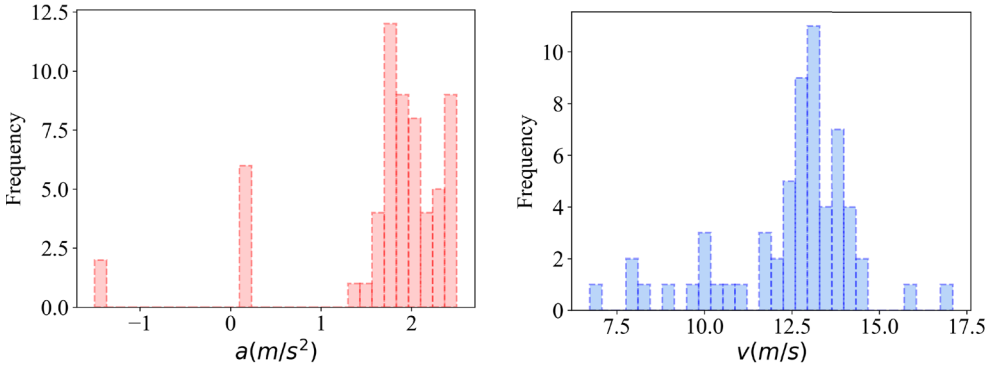


(g) $c = 100000, b_1 = 20, b_2 = 100, \omega = 3000$

Figure 7. Training exploration trajectories of target vehicle (An example of Case 1: CAV-CAV). (a) $c = 100{,}000, b_1 = 20, b_2 = 100, \omega = 2000$. (b) $c = 50{,}000, b_1 = 20, b_2 = 100, \omega = 2000$. (c) $c = 10{,}000, b_1 = 20, b_2 = 100, \omega = 2000$. (d) $c = 100{,}000, b_1 = 10, b_2 = 50, \omega = 2000$. (e) $c = 100{,}000, b_1 = 30, b_2 = 150, \omega = 2000$. (f) $c = 100{,}000, b_1 = 20, b_2 = 100, \omega = 1000$ and (g) $c = 100{,}000, b_1 = 20, b_2 = 100, \omega = 3000$.

Table 3. Reward coefficients of RL.

Name	Description	Unit	Value
c	Collision penalty	–	−100, 000
α	Coefficient of function $\mathcal{R}_s$	1/m	960
β	Coefficient of function $\mathcal{R}_s$	–	22
l_{veh}	The length of vehicle	m	5
b_1	Coefficient of jerk	s^3/m	20
b_2	Coefficient of yaw	–	100
κ	Adjustment coefficient of $\mathcal{R}_f$	m/gram	800
ϱ	U-curve coefficient of $\mathcal{R}_l$	1/m	2046
δ	U-curve coefficient of $\mathcal{R}_l$	m	1
ζ	U-curve coefficient of $\mathcal{R}_l$	–	692
ω	U-curve coefficient of $\mathcal{R}_l$	1/m	2000



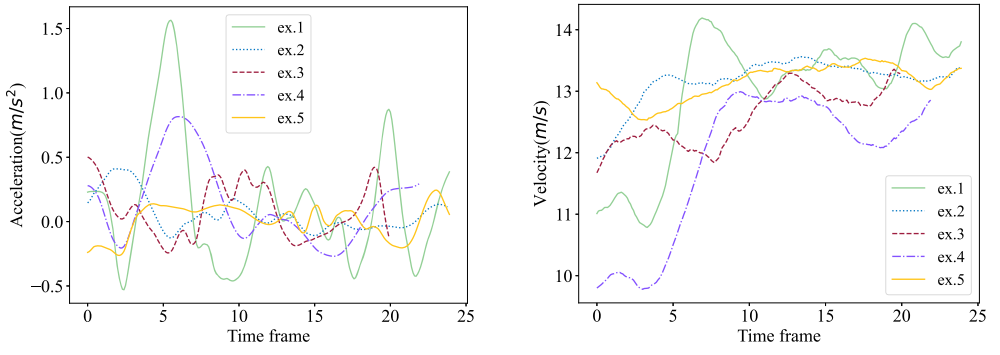
(a) Acceleration (b) Velocity

Figure 8. Distribution of evaluated acceleration and velocity during training. (a) Acceleration and (b) Velocity.

3.2.1. Training results of DRL-based lane-changing execution

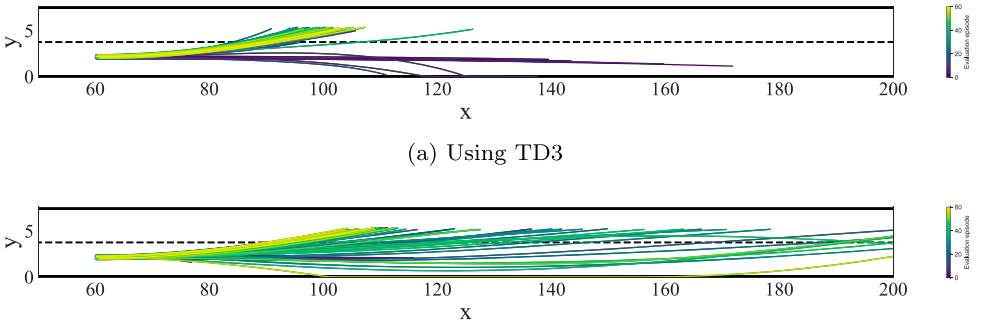
In the field of DRL, the total reward accumulated during each episode and the convergence time serve as important measures to evaluate the performance of the DRL policy. We preset 5000 steps as an evaluation episode thus the total rewards are the average of each 5000 steps, which we refer to as the 'average total reward'. A higher average total reward indicates a better policy. Figure 11 illustrates the average total rewards obtained by TD3 and DDPG under the traffic state $L_{ori} = 30$. x and y axis in Figure 11(a,b) indicate rewards and training steps of DRL algorithms. We observe that TD3 obtains larger average total rewards with a smaller convergence time, consequently, TD3 is adopted as the DRL algorithm in CLCMT to conduct the following experiments.

Specifically, in Figure 11(a), TD3 gained stable rewards at approximately 30,000 generations in Case 1. Through further exploration with fluctuated rewards, a lane-changing strategy with higher total rewards was discovered at around 100,000 generations and converged at this level. Case 2 also exhibited fast convergence, occurring at approximately 30,000 generations without additional explorations. After a longer exploration, the agent eventually converged at a total reward that was comparable to Case 2. In contrast, agents in Case 4 made more explorations and thus used the longest time to converge. However, it finally failed to exhibit advantages over other strategies. Overall, Case 1 demonstrated



(a) Accelerations of target vehicle (b) Velocity of target vehicle

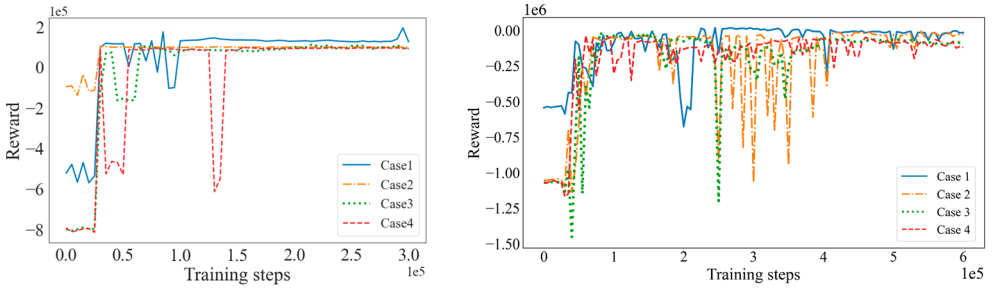
Figure 9. Examples of acceleration and velocity during training. (a) Accelerations of target vehicle and (b) Velocity of target vehicle.



(a) Using TD3

(b) Using DDPG

Figure 10. Training exploration trajectories of the target vehicle (Example of Case 1: CAV-CAV). (a) Using TD3 and (b) Using DDPG.

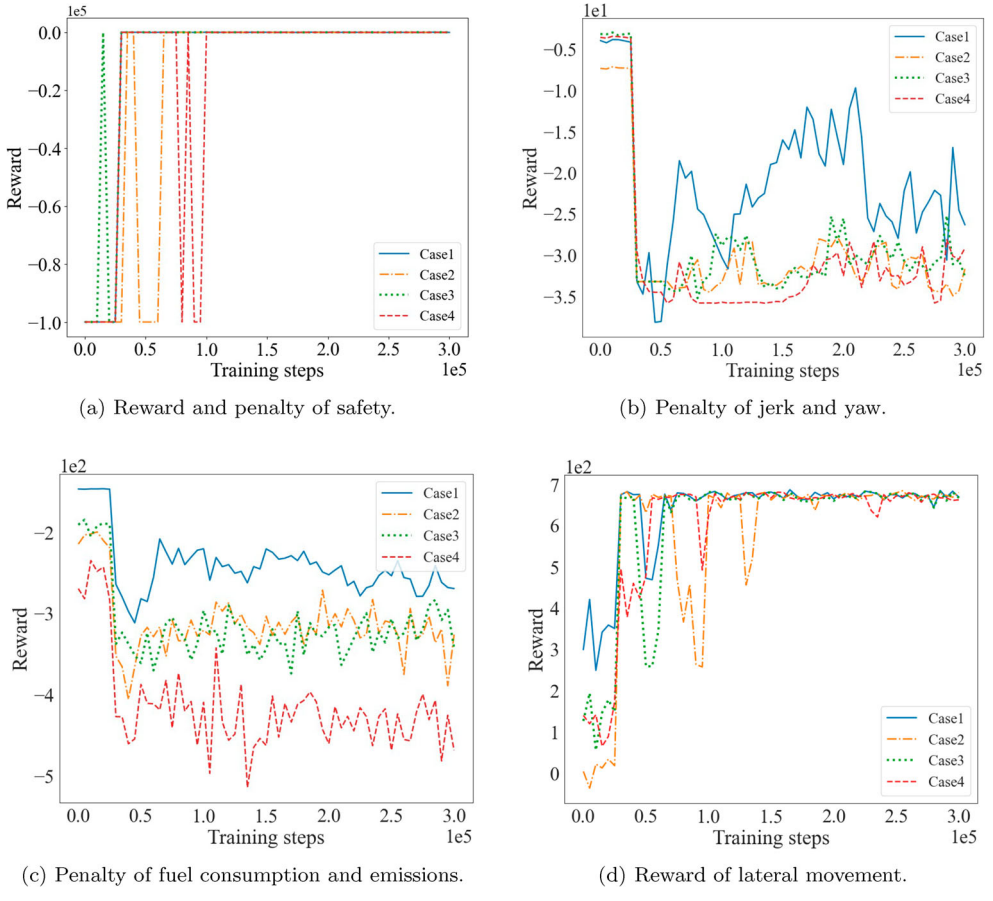


(a) TD3 (b) DDPG

Figure 11. Average total rewards during training. (a) TD3 and (b) DDPG.

the highest total reward among the four cases, indicating that the cooperative behaviour of more agents helps to explore better lane-changing strategies.

Figure 12 depicts the individual reward components of the TD3-based lane-changing model for the four leader-follower cases. Cases 1–4 are represented by blue solid lines,



(a) Reward and penalty of safety.

(b) Penalty of jerk and yaw.

(c) Penalty of fuel consumption and emissions.

(d) Reward of lateral movement.

Figure 12. Rewards of TD3-based lane-changing control. (a) Reward and penalty of safety. (b) Penalty of jerk and yaw. (c) Penalty of fuel consumption and emissions and (d) Reward of lateral movement.

orange dashed lines, green dotted lines, and red dotted lines, respectively. Figure 12(a) illustrates safety reward, in which the extremely low rewards indicate collisions. Observed that Case 2–4 get more collisions than Case 1 in the learning process. This may be due to more surrounding HVs which provide higher uncertainty. Comfort rewards are shown in Figure 12(b). Note that Case 1 has the lowest jerk and yaw penalty among the four cases, the other three cases obtain similar rewards towards comfort. Figure 12(c) illustrates the reward of fuel consumption and emissions, in which Case 1 has the smallest penalty, while Case 4 has the largest penalty. Additionally, the rewards for lateral movement are displayed in Figure 12(d). After some exploration, the agent successfully learns the lane-changing strategies in each case, thus all cases have similar rewards after convergence.

3.2.2. Utilities of lane-changing manoeuvres

After finishing the training process, the utilities of the four lane-changing cases are calculated in the feedback module. The results are shown in Table 4. Case 1 has a crash rate of only 0.439%, while Case 4 exhibits a significantly higher crash rate of 6.927%. This indicates that increasing the penetration of CAVs leads to safer, more comfortable, and eco-friendly

Table 4. Utilities of lane-changing in various scenarios.

Leader-follower compositions	Efficiency ($\mathcal{U}_t/s$)	Safety ($\mathcal{U}_s/\%$)	Comfort ($\mathcal{U}_c$)	Ecology ($\mathcal{U}_e$)
Case 1 (CAV-CAV)	3.587	0.439	−21.257	−248.967
Case 2 (CAV-HV)	3.342	5.902	−31.560	−316.802
Case 3 (HV-CAV)	3.056	4.878	−30.503	−324.692
Case 4 (HV-HV)	2.575	6.927	−31.626	−416.916

lane-changing behaviours. Specifically, Case 2 exhibits a higher crash rate than Case 3, indicating that the new follower in the target lane plays a more important role compared to the new leader vehicle. Higher CAV penetration also shows advantages in comfort and ecology, as more controlled vehicles can make wider adjustments of movements to cooperate with each other as well as with surrounding HVs. Meanwhile, this increased cooperation extends the exploration process, thus needing a longer time compared to other cases. Evidence can be found in Table 4, where Case 4 completes lane-changing with around 2.575 s, whereas Case 1 has a longer time with 3.587 s. This indicates that when HVs are involved in the leader-follower composition, the target vehicle tends to prioritise quicker lane-changing at the expense of safety, comfort, and fuel efficiency.

3.2.3. Utilities of manoeuvre control under various initial conditions

To further explore the effectiveness of the DRL-based manoeuvre control strategy in the four cases, we conduct further experiments with different original traffic states. Due to space constraints, we present the results for two more instances where the original leader-follower gap (L_{ori}) is varied as 20 m and 40 m. The results are illustrated in Figure 13. Each boxplot denotes the upper quartile, lower quartile, and median of the total rewards. The whiskers extend to the most extreme data points that are considered outliers, while any outliers are indicated by hollow circles. The colours pink, blue, and green correspond to the original gaps of 20 m, 30 m, and 40 m, respectively.

As depicted in Figure 13(a1)–(a4), smaller original spacing results in a longer duration for the target vehicle to execute the lane-changing manoeuvre in each leader-follower combination. This smaller spacing also leads to higher crash rates, as shown in Figure 13(b1)–(b4). Notably, this type of traffic state provides the most comfortable experience due to the extended lane-changing time, as illustrated in Figure 13(c1)–(c4). In contrast, a leader-follower composition with a larger original spacing, say 40 m, offers a safer and more efficient lane-changing strategy while compromising some comfort utility, as observed from Figure 13(a1)–(a4), (b1)–(b4), and (c1)–(c4). As demonstrated in Figure 13(d1)–(d4), fuel consumption and emissions are not unilaterally influenced by changes in the original spacing. The most eco-friendly driving strategy is achieved when $L_{ori} = 30$.

3.2.4. Decision-making based on the feedback mechanism

To facilitate an effective comparison of the results in each case, we normalise the utility of each indicator. The percentage difference in the normalised values (DPN) is then utilised to evaluate the lane-changing utility. The results of normalised results for different cases are illustrated in Table 5.

Notably, when considering the same original status, Case 1 exhibits DPN values of 24.5%–35.8% in safety, 8%–9% in comfort, and 5.2%–12.9% in ecology compared to the

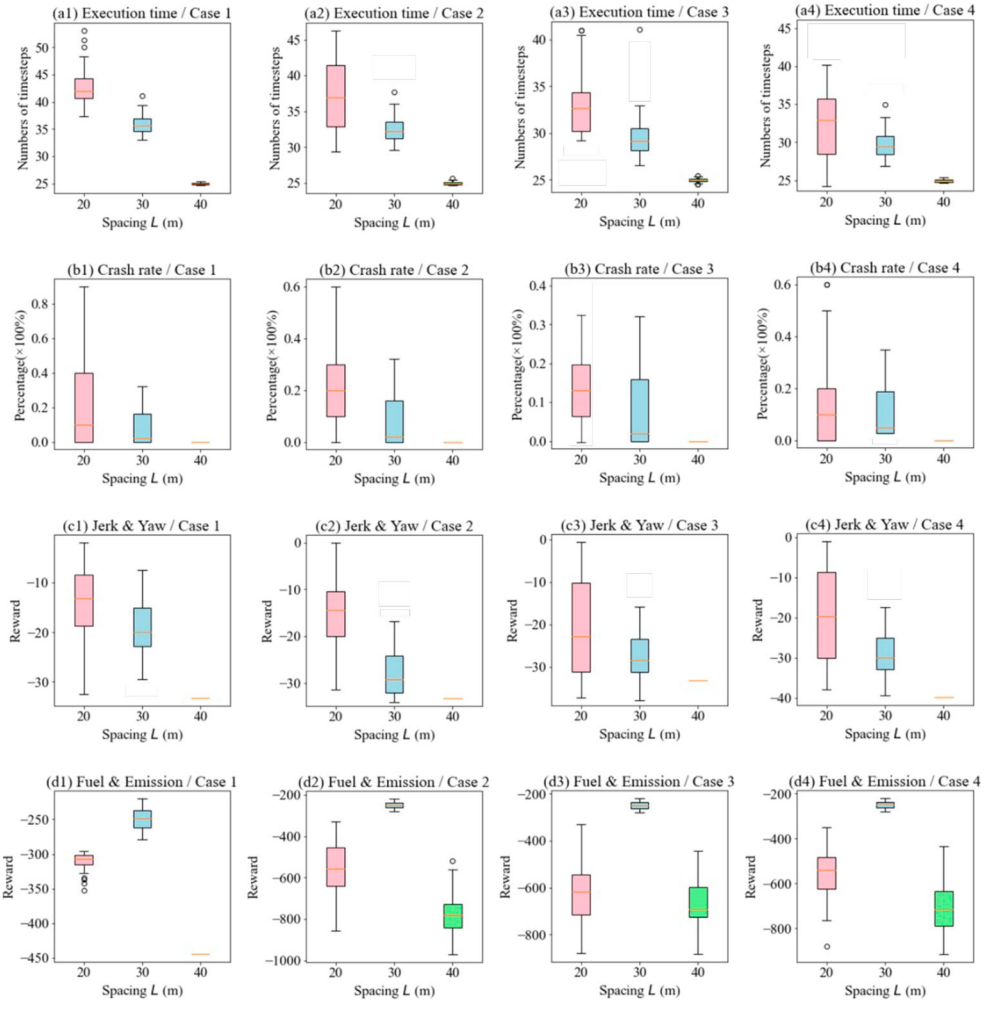


Figure 13. Utilities of lane-changing under various original states.

Table 5. Normalized utilities of different cases (e.g. $L_{ori} = 30$).

Leader-follower compositions	Efficiency ($\mathcal{U}_t$)	Safety ($\mathcal{U}_s$)	Comfort ($\mathcal{U}_c$)	Ecology ($\mathcal{U}_e$)
Case 1 (CAV-CAV)	0.286	0.024	0.185	0.190
Case 2 (CAV-HV)	0.266	0.325	0.275	0.242
Case 3 (HV-CAV)	0.243	0.269	0.265	0.248
Case 4 (HV-HV)	0.205	0.382	0.275	0.319

other three cases. Increasing the penetration of HVs leads to decreased safety and comfort for the target vehicle and increased fuel consumption and emissions during lane-changing. If the four utilities follow a linear function with even weights, the feedback module favours recommending the adjacent lane with a CAV-CAV composition, while the HV-HV composition is considered less favourable. Following different user-oriented evaluation function, the feedback module recalculates the utilities and provides recommendations on personalised lane-changing strategies.

4. Concluding remarks

This paper proposed a cooperative lane-changing in mixed traffic (CLCMT) mechanism based on the TD3 method to facilitate optimal lane-changing strategies. A feedback module was designed to enable interactions between the decision-making layer and the manoeuvre control layer, in which the utilities of potential motion execution can be pre-calculated to provide guidance for decision-making. The uncertainty in HV decisions was considered in the mechanism. Various leader-follower compositions in mixed traffic, including CAV-CAV, CAV-HV, HV-CAV, and HV-HV, were included, covering a wide range of typical lane-changing scenarios. Training results show that the TD3-based CLCMT approach enabled the target vehicle to learn smooth lane-changing trajectories with enough safety. The evaluation regarding safety, efficiency, comfort, and ecology in the feedback module contributed to recommending the optimal lane-changing strategy. The findings of this research highlighted the advantages of the proposed CLCMT in terms of lateral decision-making and motion control for CAVs. The key findings are summarised below.

(i) With similar initial gap settings in traffic, increasing the penetration of CAVs results in safer, more comfortable, and environmentally friendly lane-changing. Opting for a CAV-CAV lane-changing strategy can enhance safety by 24.5%–35.8%, improve comfort by 8%–9%, and reduce fuel consumption and emissions by 5.2%–12.9% when compared to other lane-changing options.
(ii) Larger spacing between vehicles also leads to higher lane-changing utilities, aligning well with real-life experiences.
(iii) The most eco-friendly lane-changing manoeuvres are achieved when the initial spacing between vehicles is 30 m, indicating that the relationship between leader-follower spacing and ecology utilities does not exhibit a monotonic pattern.

Though the proposed CLCMT strategy has provided valuable insights into dealing with complex lane-changing scenarios, we acknowledge its limitations. The movements of vehicles are controlled by longitudinal and lateral accelerations, leading to a decoupling of horizontal and vertical driving behaviour. Note that the lateral movement of a vehicle is influenced by various factors such as steering angle, tyre dynamics, road conditions, and external forces, refined vehicle dynamics control should be considered to facilitate high-fidelity driving trajectories. Future research may involve this enhancement in control precision to assist lane-changing vehicles in acquiring more effective execution strategies. Moreover, exploring multi-agent actor-critic approaches to consider cooperative-competitive driving behaviours of multiple vehicles also holds promise (Parada et al. 2023). Additionally, heterogeneity of driving behaviour, such as low- and high-compliance of HVs, can be considered to provide more realistic learning environments. This helps to learn strategies that can deal with more complex situations.

Note

1. Example can be found by https://www.youtube.com/watch?v = gZIwcZZR1P0)

Acknowledgements

The authors confirm their contribution to the paper as follows: study conception and design: Xue Yao, and Zhanbo Sun; data collection: Xue Yao, Zhao Chengdu; analysis and interpretation of results: Xue Yao, Zhao Chengdu; draft manuscript preparation: Xue Yao, Zhanbo Sun, Simeon C. Calvert, Zhao Chengdu, and Ang Ji. All authors reviewed the results and approved the final version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work is supported by the National Natural Science Foundation of China via grant 52072316 and 52302418, the Fundamental Research Funds for the Central Universities via grant 2682023CX047, and the Postdoctoral International Exchange Program via grant YJ20220311.

ORCID

Ang ji http://orcid.org/0000-0002-7943-7461

References

Arbis, D., and V. V. Dixit. 2019. "Game Theoretic Model for Lane Changing: Incorporating Conflict Risks." *Accident Analysis & Prevention* 125:158–164. https://doi.org/10.1016/j.aap.2019.02.007.

Bouadi, M., B. Jia, R. Jiang, X. Li, and Z. Gao. 2022. "Optimizing Sensitivity Parameters of Automated Driving Vehicles in An Open Heterogeneous Traffic Flow System." *Transportmetrica A: Transport Science* 18 (3): 762–806. https://doi.org/10.1080/23249935.2021.1896592.

Calvert, S., and B. van Arem. 2020. "A Generic Multi-level Framework for Microscopic Traffic Simulation with Automated Vehicles in Mixed Traffic." *Transportation Research Part C: Emerging Technologies* 110:291–311. https://doi.org/10.1016/j.trc.2019.11.019.

Deng, J.-H., and H.-H. Feng. 2019. "A Multilane Cellular Automaton Multi-attribute Lane-changing Decision Model." *Physica A: Statistical Mechanics and Its Applications* 529:121545. https://doi.org/10.1016/j.physa.2019.121545.

Dong, C., H. Wang, Y. Li, X. Shi, D. Ni, and W. Wang. 2021. "Application of Machine Learning Algorithms in Lane-changing Model for Intelligent Vehicles Exiting to Off-ramp." *Transportmetrica A: Transport Science* 17 (1): 124–150. https://doi.org/10.1080/23249935.2020.1746861.

Fujimoto, S., H. Hoof, and D. Meger. 2018. "Addressing Function Approximation Error in Actor-Critic Methods." In *International Conference on Machine Learning*, 1587–1596. PMLR.

Ha, P., S. Chen, J. Dong, and S. Labi. 2023. "Leveraging Vehicle Connectivity and Autonomy for Highway Bottleneck Congestion Mitigation Using Reinforcement Learning." *Transportmetrica A: Transport Science* 19 (1): 1–26. https://doi.org/10.1080/23249935.2023.2215338.

Hou, K., F. Zheng, X. Liu, and Z. Fan. 2023. "Cooperative Vehicle Platoon Control Considering Longitudinal and Lane-changing Dynamics." *Transportmetrica A: Transport Science* 20 (3): 1–29. https://doi.org/10.1080/23249935.2023.2182143.

Jaritz, M., R. De Charette, M. Toromanoff, E. Perot, and F. Nashashibi. 2018. "End-to-End Race Driving with Deep Reinforcement Learning." In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2070–2075. IEEE.

Ji, J., A. Khajepour, W. W. Melek, and Y. Huang. 2016. "Path Planning and Tracking for Vehicle Collision Avoidance Based on Model Predictive Control with Multiconstraints." *IEEE Transactions on Vehicular Technology* 66 (2): 952–964. https://doi.org/10.1109/TVT.2016.2555853.

Ji, A., and D. Levinson. 2020. "A Review of Game Theory Models of Lane Changing." *Transportmetrica A: Transport Science* 16 (3): 1628–1647. https://doi.org/10.1080/23249935.2020.1770368.

Ji, A., M. Ramezani, and D. Levinson. 2023a. "Joint Modelling of Longitudinal and Lateral Dynamics in Lane-changing Maneuvers." *Transportmetrica B: Transport Dynamics* 11 (1): 996–1025.
Ji, A., M. Ramezani, and D. Levinson. 2023b. "Pricing Lane Changes." *Transportation Research Part C: Emerging Technologies* 149:104062. https://doi.org/10.1016/j.trc.2023.104062.
Jin, J., H. Huang, Y. Li, G. Zhang, Y. Dong, B. Zhou, and H. Xue. 2023. "Variable Speed Limit Modelling to Improve Traffic Safety and Efficiency of Mixed Traffic Flow by a Two-stage Framework." *Transportmetrica A: Transport Science* 19 (1): 1–25. https://doi.org/10.1080/23249935.2023.2253476.
Li, L., J. Gan, K. Zhou, X. Qu, and B. Ran. 2020. "A Novel Lane-changing Model of Connected and Automated Vehicles: Using the Safety Potential Field Theory." *Physica A: Statistical Mechanics and Its Applications* 559:125039. https://doi.org/10.1016/j.physa.2020.125039.
Li, Z.-C., H.-J. Huang, and W. H. Lam. 2012. "Modelling Heterogeneous Drivers' Responses to Route Guidance and Parking Information Systems in Stochastic and Time-dependent Networks." *Transportmetrica* 8 (2): 105–129. https://doi.org/10.1080/18128600903568570.
Li, D., and H. Pan. 2022. "Two-lane Two-way Overtaking Decision Model with Driving Style Awareness Based on a Game-theoretic Framework." *Transportmetrica A: Transport Science* 19 (3): 1–26. https://doi.org/10.1080/23249935.2022.2076755.
Li, X., Y. Xu, N. Li, B. Yang, and Y. Lei. 2022. "Remaining Useful Life Prediction with Partial Sensor Malfunctions Using Deep Adversarial Networks." *IEEE/CAA Journal of Automatica Sinica* 10 (1): 121–134. https://doi.org/10.1109/JAS.2022.105935.
Li, J., L. Yao, X. Xu, B. Cheng, and J. Ren. 2020. "Deep Reinforcement Learning for Pedestrian Collision Avoidance and Human-machine Cooperative Driving." *Information Sciences* 532:110–124. https://doi.org/10.1016/j.ins.2020.03.105.
Li, X., S. Yu, Y. Lei, N. Li, and B. Yang. 2023. "Intelligent Machinery Fault Diagnosis with Event-Based Camera." *IEEE Transactions on Industrial Informatics* 20 (1): 380–389. https://doi.org/10.1109/TII.2023.3262854.
Liao, Y., G. Yu, P. Chen, B. Zhou, and H. Li. 2022. "Modelling Personalised Car-following Behaviour: A Memory-based Deep Reinforcement Learning Approach." *Transportmetrica A: Transport Science* 20 (1): 1–29. https://doi.org/10.1080/23249935.2022.2035846.
Lin, D., L. Li, and S. E. Jabari. 2019. "Pay to Change Lanes: A Cooperative Lane-changing Strategy for Connected/automated Driving." *Transportation Research Part C: Emerging Technologies* 105:550–564. https://doi.org/10.1016/j.trc.2019.06.006.
Mirchevska, B., C. Pek, M. Werling, M. Althoff, and J. Boedecker. 2018. "High-Level Decision Making for Safe and Reasonable Autonomous Lane Changing Using Reinforcement Learning." In *2018 IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2156–2162. IEEE.
Nguyen, T. T., N. D. Nguyen, and S. Nahavandi. 2020. "Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications." *IEEE Transactions on Cybernetics* 50 (9): 3826–3839. https://doi.org/10.1109/TCYB.6221036.
Nie, Y. M., and Q. Li. 2013. "An Eco-routing Model Considering Microscopic Vehicle Operating Conditions." *Transportation Research Part B: Methodological* 55: 154–170. https://doi.org/10.1016/j.trb.2013.06.004.
Pan, T., W. H. Lam, A. Sumalee, and R. Zhong. 2021. "Multiclass Multilane Model for Freeway Traffic Mixed with Connected Automated Vehicles and Regular Human-piloted Vehicles." *Transportmetrica A: Transport Science* 17 (1): 5–33. https://doi.org/10.1080/23249935.2019.1573858.
Parada, L., E. Candela, L. Marques, and P. Angeloudis. 2023. "Safe and Efficient Manoeuvring for Emergency Vehicles in Autonomous Traffic Using Multi-agent Proximal Policy Optimisation." *Transportmetrica A: Transport Science* 18 (1): 1–29. https://doi.org/10.1080/23249935.2023.2246586.
Paul, G., N. Raju, S. Arkatkar, and S. Easa. 2021. "Can Segregating Vehicles in Mixed-traffic Stream Improve Safety and Throughput? Implications Using Simulation." *Transportmetrica A: Transport Science* 17 (4): 1002–1026. https://doi.org/10.1080/23249935.2020.1826595.
Rasekhipour, Y., A. Khajepour, S.-K. Chen, and B. Litkouhi. 2016. "A Potential Field-based Model Predictive Path-planning Controller for Autonomous Road Vehicles." *IEEE Transactions on Intelligent Transportation Systems* 18 (5): 1255–1267. https://doi.org/10.1109/TITS.2016.2604240.

Ren, J., Y. Chen, L. Xin, J. Shi, and H. Mahama. 2017. "Detecting and Locating of Traffic Incidents in a Road Segment Based on Lane-changing Characteristics." *Transportmetrica A: Transport Science* 13 (10): 853–873. https://doi.org/10.1080/23249935.2017.1348400.

Shi, T., P. Wang, X. Cheng, C.-Y. Chan, and D. Huang. 2019. "Driving Decision and Control for Automated Lane Change Behavior Based on Deep Reinforcement Learning." In *2019 IEEE 22nd International Conference on Intelligent Transportation Systems (ITSC)*, 2895–2900. IEEE.

Sun, Z., J. Huang, A. Ji, R. Zhao, and G. Zheng. "Cooperative Merging for Connected Automated Vehicles in Mixed Traffic: A Multi-Agent Reinforcement Learning Approach." *Available at SSRN 4564695*.

Sun, Z., T. Huang, and P. Zhang. 2020. "Cooperative Decision-making for Mixed Traffic: A Ramp Merging Example." *Transportation Research Part C: Emerging Technologies* 120:102764. https://doi.org/10.1016/j.trc.2020.102764.

Sun, Z., R. Liu, H. Hu, D. Liu, and Z. Yan. 2023. "Cyberattacks on Connected Automated Vehicles: A Traffic Impact Analysis." *IET Intelligent Transport Systems* 17 (2): 295–311. https://doi.org/10.1049/itr2.v17.2.

Sun, Z., X. Yao, Z. Qin, P. Zhang, and Z. Yang. 2021. "Modeling Car-following Heterogeneities by Considering Leader–follower Compositions and Driving Style Differences." *Transportation Research Record* 2675 (11): 851–864. https://doi.org/10.1177/03611981211020006.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press.

Wang, M., S. P. Hoogendoorn, W. Daamen, B. van Arem, and R. Happee. 2015. "Game Theoretic Approach for Predictive Lane-changing and Car-following Control." *Transportation Research Part C: Emerging Technologies* 58:73–92. https://doi.org/10.1016/j.trc.2015.07.009.

Wang, G., J. Hu, Z. Li, and L. Li. 2021. "Harmonious Lane Changing Via Deep Reinforcement Learning." *IEEE Transactions on Intelligent Transportation Systems* 23 (5): 4642–4650. https://doi.org/10.1109/TITS.2020.3047129.

Wang, P., H. Li, and C.-Y. Chan. 2019. "Continuous Control for Automated Lane Change Behavior Based on Deep Deterministic Policy Gradient Algorithm." In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1454–1460. IEEE.

Xu, H., Y. Zhang, C. G. Cassandras, L. Li, and S. Feng. 2020. "A Bi-level Cooperative Driving Strategy Allowing Lane Changes." *Transportation Research Part C: Emerging Technologies* 120:102773. https://doi.org/10.1016/j.trc.2020.102773.

Yu, H., H. E. Tseng, and R. Langari. 2018. "A Human-like Game Theory-based Controller for Automatic Lane Changing." *Transportation Research Part C: Emerging Technologies* 88:140–158. https://doi.org/10.1016/j.trc.2018.01.016.

Zhou, J., S. Xue, Y. Xue, Y. Liao, J. Liu, and W. Zhao. 2021. "A Novel Energy Management Strategy of Hybrid Electric Vehicle Via An Improved Td3 Deep Reinforcement Learning." *Energy* 224:120118. https://doi.org/10.1016/j.energy.2021.120118.