Continuous Decision-Making in Lane Changing and Overtaking Maneuvers for Unmanned Vehicles: A Risk-Aware Reinforcement Learning Approach With Task Decomposition

Sifan Wu, Daxin Tian, *Senior Member, IEEE*, Xuting Duan, Jianshan Zhou, Dezong Zhao, *Senior Member, IEEE*, and Dongpu Cao, *Senior Member, IEEE*

***Abstract*—Reinforcement learning methods have shown the ability to solve challenging scenarios in unmanned systems. However, solving long-time decision-making sequences in a highly complex environment, such as continuous lane change and overtaking in dense scenarios, remains challenging. Although existing unmanned vehicle systems have made considerable progress, minimizing driving risk is the first consideration. Risk-aware reinforcement learning is crucial for addressing potential driving risks. However, the variability of the risks posed by several risk sources is not considered by existing reinforcement learning algorithms applied in unmanned vehicles. Based on the above analysis, this study proposes a risk-aware reinforcement learning method with driving task decomposition to minimize the risk of various sources. Specifically, risk potential fields are constructed and combined with reinforcement learning to decompose the driving task. The proposed reinforcement learning framework uses different risk-branching networks to learn the driving task. Furthermore, a low-risk episodic sampling augmentation method for different risk branches is proposed to solve the shortage of high-quality samples and further improve sampling efficiency. Also, an intervention training strategy is employed wherein the artificial potential field (APF) is combined with reinforcement learning to speed up training and further ensure safety. Finally, the complete intervention risk classification twin delayed deep deterministic policy gradient-task decompose (IDRCTD3-TD) algorithm is proposed. Two scenarios with different difficulties are designed to validate the superiority of this framework. Results show that the proposed framework has remarkable improvements in performance.**

***Index Terms*—Reinforcement learning, unmanned systems, decision-making, driving risk, task decomposition.**

I. Introduction

WITH the advancement of technology, artificial intelligence has been widely employed in academia and industry, where unmanned systems have received more attention as the derivative of artificial intelligence [1], [2]. Unmanned systems can replace humans to conduct civilian and military missions, which is strategically significant for constructing intelligent transportation systems and developing army equipment. In contrast to manned systems, unmanned systems are mission-orientated, which eliminates the need for unmanned systems to consider the passenger's needs (comfort, space constraints, impact resistance, etc.). Many of the technologies cannot be applied to manned platforms because the physiological limits of the human body constrain them. However, they can be applied to unmanned platforms. Among these, an overtaking mission is a typical unmanned system mission characterized by the need to complete tasks without slowing down to ensure maneuverability and time efficiency. For example, unmanned rescue vehicles overtake slow-moving vehicles to reach the emergency area, and unmanned vehicles overtake obstacles to navigate to the target positions and complete the mission, which is more focused on mission execution efficiency than on providing a comfortable driving experience [3]. In addition, the risk is another factor to consider with unmanned vehicles; a risk is typically an event that is not fatal to the system at the current moment but could lead to severe consequences. In this factor, the loss is not just a number in a simulator, but the damage caused to the entire unmanned system; unlike a fatal event such as a collision, an at-risk unmanned vehicle is still safe but may encounter a fatal event shortly and the probability of a fatal event and loss is more significant as the risk increases [4]. Therefore, being aware of the risks of events and taking appropriate actions to prevent fatal events is an effective way to avoid serious accidents [5].

In the whole unmanned system, unmanned vehicles as an essential part of the unmanned system have a significant prospect; unmanned vehicles are intelligent vehicles that do not have a human driving mechanism and can autonomously perform tasks such as transport, logistics, patrol, rescue, combat, surveillance and so on. In the civilian field, unmanned vehicles have

Manuscript received 19 February 2024; revised 1 March 2024; accepted 10 March 2024. Date of publication 25 March 2024; date of current version 12 June 2024. This work was supported by the National Natural Science Foundation of China under Grant U20A20155, Grant 62173012, and Grant U22A2046. *(Corresponding authors: Daxin Tian; Xuting Duan.)*

Sifan Wu and Jianshan Zhou are with the Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: wusifan9070@163.com; jianshanzhou@foxmail.com).

Daxin Tian and Xuting Duan are with the Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China, and also with the Zhongguancun Laboratory, Beijing 100081, China (e-mail: dtian@buaa.edu.cn; duanxuting@buaa.edu.cn).

Dezong Zhao is with the James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K. (e-mail: dezong.zhao@glasgow.ac.uk).

Dongpu Cao is with the School of Vehicle and Mobility, Tsinghua University, Beijing 100190, China (e-mail: dp_cao2016@163.com).

become the core of future intelligent transport and smart city construction. In the military, unmanned vehicles have become a new generation of army equipment competing for military powers and a new generation of army equipment. Therefore, the accelerated development of mission-orientated unmanned vehicles is a promising trend for the future [6].

A. Related Works

Unmanned vehicles sense the environment and respond instantaneously to surrounding vehicles by making appropriate decisions to find a collision-free path [7], [8]. However, decision-making in a complex environment, more specifically, the act of making the optimal decision that is conducive to avoiding obstacles in a complex driving environment has always been an issue for research [9], [10]. Existing methods for solving path-planning for collision avoidance decision-making are mainly classified into traditional and learned methods. In the traditional methods, for instance, Li et al. used the Cartesian frame to solve the trajectory planning problem on curvy roads. The proposed method decomposes the complex collision and motion constraints into the subproblems to find the optima in the continuous solution space [11]. Tu et al. proposed a hybrid A* algorithm using a new search strategy and combining interpolation and nonlinear optimization to improve the performance [12]. Mashayekhi et al. introduced a hybrid rapidly exploring random tree (RRT) method based on bidirectional and unidirectional searches. Then, a merging idea was applied to the search tree to optimize the performance of hybrid RRT [13]. Vinayak et al. developed a bezier curve-path search method based on a boundary condition search to determine appropriate control points [14]. However, the above traditional solutions frequently are limited by their high time complexity, so methods can occasionally result in paths that collide with and cannot ensure real-time planning. The artificial potential field (APF) is a typical technique in path-planning for collision avoidance decision-making [15]. Wang et al. incorporated a safe obstacle-avoidance model into APF to guarantee the safety of driving [16]. Wu et al. proposed a method for lane-changing decision-making incorporating APF, which combined the properties of human drivers' lane-changing cognition and proposed a spatially varying moving potential field to address the driver's focus transformation process during the lane-changing process [17]. Wang et al. defined the potential field as the "risk field" and used it to describe the risk from the surrounding environment [18]. Li et al. addressed the problem of APF easily falling into local optima and unreachable targets by introducing dynamic distance factors and using the invasive weeds algorithm to solve these problems, respectively [19]. Nevertheless, most current APF-based techniques typically are easily trapped in local optimal solutions in complex or dynamic environments, thus failing to generate collision-free paths [20].

Among the learning-based methods, two types of supervised learning and reinforcement learning are now becoming the prevailing methods for decision-making. In the supervised learning research area, Xi et al. used mixed integer quadratic problem based optimization to generate motion trajectories as the training data and devised a hierarchical supervised learning model, which employed the support vector machine and multi-layer perception as the decision module to solve the obstacle-avoidance problem [21]. Teng et al. proposed a semantic bird's eye view model with imitation learning to present an interpretation of the surrounding environment, and their model then fused with the pure-pursuit algorithm to output the control command to successfully avoid obstacles for decison-making [22]. However, supervised learning methods require a large number of human-labeled samples, which can be labor-intensive and time-consuming. The absence of some collision or near-collision data impedes learning, leading to supervised learning methods always having unsatisfactory practical applications [23]. Deep reinforcement learning algorithms are increasingly being used in unmanned vehicles as artificial intelligence technology advances [24]. By interacting with the simulated environment and generating the data without supervised information [25], reinforcement learning can effectively reduce the use of human resources and time consumption [26]. With much fewer constraints than rule-based approaches, reinforcement learning-based decision-making methods are acceptable for most cases since they require fewer constraints to be defined than rule-based methods. Yang et al. proposed a reinforcement learning decision-making model for lane changing based on uncertainty estimation to quantify the reliability of the strategy and identify unknown scenarios [27]. Tang et al. developed a reinforcement learning-based decision-making method that takes into account different driving strategies [28]. Xu et al. proposed a safe reinforcement learning algorithm to ensure safety, which combined reinforcement learning algorithms with APF and trajectory tracking methods to output the actions by weighting. Unfortunately, the method only performed well in low-obstacle scenarios and was not validated for high-complexity performance [29]. Recently, risk awareness has become an essential topic for driving decisions and is critical to making safe decision [30]. However, risk awareness is rarely introduced into reinforcement learning to train driving strategies in unmanned systems. When reinforcement learning is applied to driving decisions, the system's specificity requires that the strategies be risk-aware to handle emerging dangers. While this process can be achieved by encoding risk awareness into rewards, it is a heuristic process in which empirical knowledge of reinforcement learning is required, and learning about the heterogeneity between different risks in this process requires further consideration. Li et al. proposed a novel decision-making method based on risk assessment, which used a risk-based probabilistic assessment model and applied it to the reinforcement learning driving strategy to minimize driving risk. However, this method essentially incorporates risk into the reward function. It is only deployed to control the steer in a discrete action space [31]. Marwan et al. developed a reinforcement learning-based decision-making method that combined reinforcement learning methods with responsibility-sensitive safety (RSS) models and used RSS models to design reward functions and assess driving risk. The method was validated using the Car Learning to Act (CARLA) simulator. However, the application scenario of the method needs to be more homogeneous, considering only scenarios with two vehicles as obstacles [32]. Wu et al. added risk-aware rewards to the reinforcement learning-based decision-making system via reward shaping to avoid collisions [33]. Li et al. used an end-to-end model based on lightweight transformers and integrated driving

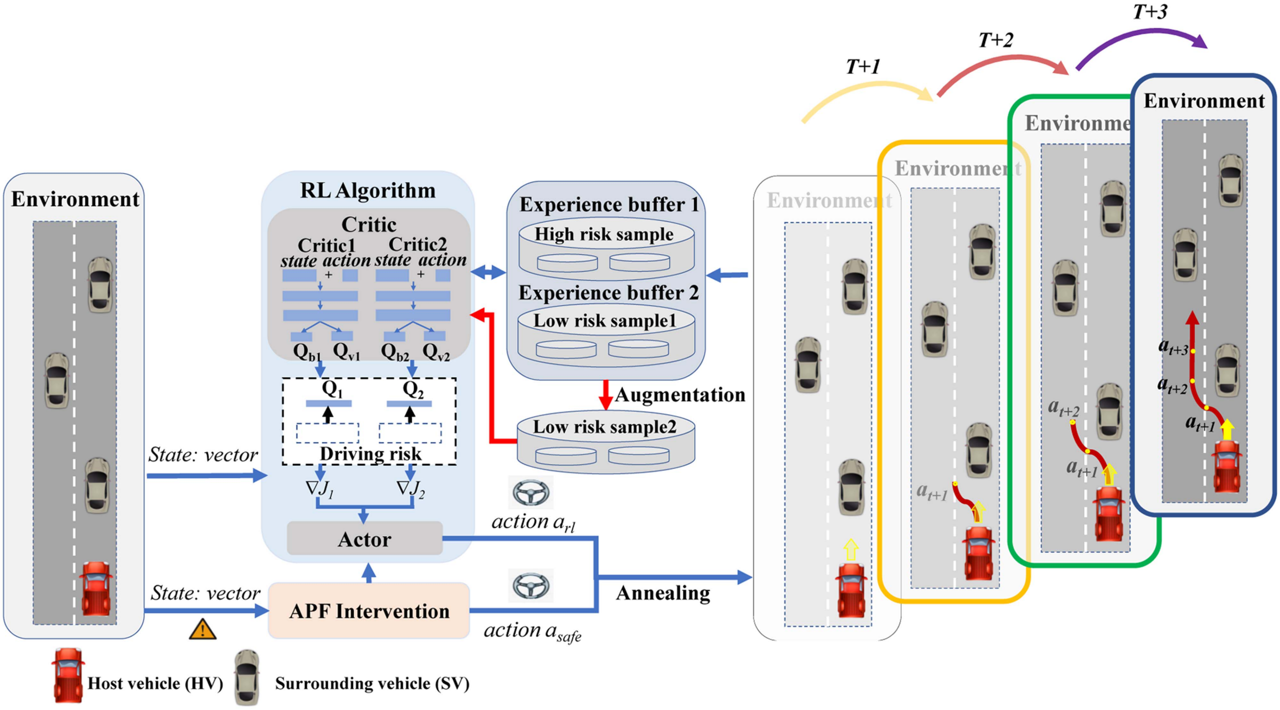


Fig. 1. Solution framework of our proposed methods.

strategies with minimal expected risk into a reinforcement learning architecture for safe lane-changing decisions [34]. Although considering the driving risk, the abovementioned reinforcement learning methods cannot distinguish between risk variations resulting from different risk sources, which leads to the fact using a single network model to learn heterogeneous risks may not ensure the convergence and stability of risk learning. Therefore, designing risk-aware reinforcement learning models, especially in risk-aware decision-making and considering heterogeneous risks, is crucial to strategy and still requires more effort. On the other hand, inefficient sample efficiency is also a common problem with reinforcement learning, which learns optimal decision-making by interacting with the environment. However, this process requires numerous samples to support reinforcement learning to explore the entire state space, which causes the learning efficiency to be inefficient. To solve this problem, some methods generate the priority of the samples by calculating the TD-error to select the samples with higher priority more frequently [35], [36]. These methods increase time complexity of the algorithm by adding operations to change priorities to the experience samples and scanning the experience buffer for experience samples with high priorities. Moreover, numerous experience samples of agents interacting with the environment have different characteristics, and different classification criteria are set to classify the samples. The different classification criteria to classify the experiences are also an effective way to improve the efficiency of experience utilization [37], [38]. However, these methods have relatively few high-quality samples at the beginning of training, which results in these high-quality samples being learned many times, which may impact the stability of the algorithm.

B. *Motivation and Contributions*

Reinforcement learning methods hold a greater potential for unmanned vehicles than traditional and supervised learning techniques since they are more similar to human decision-making behavior and do not call for elaborate definitions of restrictions or expensive human efforts to acquire data. However, developing a risk framework to consider the heterogeneity of the risks in reinforcement learning remains a great challenge. On the other side, the reinforcement learning needs to interact with the environment to find the strategy and requires high number of training samples that is sample inefficient, so how to improve sample utilisation is a further consideration to be considering. To this end, a deep reinforcement learning-based driving risk decomposition framework is proposed. Our method enables vehicles to be better aware of risks from different sources and minimize risks by considering the heterogeneity of risks, which makes the decision-making system more reasonable and better adapted to potential risks in different complex environments. In addition, combining reinforcement learning with driving risk awareness not only improves the ability of unmanned vehicles to solve complex environments but also provides interpretability for reinforcement learning to solve decision-making problems, which is essential for achieving safer and more efficient decision-making systems. Therefore, our method has practical significance for the development and application of decision-making systems for unmanned vehicles.

Fig. 1 shows the proposed framework, the framework mainly comprises four parts, including driving risk decomposition architecture, low-risk episodic sampling augmentation mechanism, intervention training strategy, and environment.

The driving risk decomposition architecture receives information about the environment and performs a task decomposition of driving risks to output decision behavior. Then, the low-risk episodic sampling augmentation mechanism classifies episodic samples from different branches by samples obtained from the environment and augments the experience buffer with low-risk episodic sampling augmentation methods. The intervention training strategy consists of two modules: the APF intervention module is used to intervene in the decision-making behavior of the vehicle in the dangerous zone, and the risk state estimation module uses the current state information and combines actions output from the reinforcement learning and the APF intervenor to predict the risk of the future state in the environment respectively, and ultimately selects the optimal decision-making to the environment. In addition, an annealing strategy is used to assist in training the decision-making model. Therefore, we develop a closed-loop optimization mechanism to optimize the decision-making model by interacting with the environment. The contributions of this study can be summarized as follows:

1) A deep reinforcement learning decision-making method based on driving risk fields is developed, and a novel risk task decomposition framework to reduce learning difficulty by establishing different risk branches is proposed.
2) To further increase the utilisation of samples, a low-risk episodic sampling augmentation method are applied to the experience buffer in different risk branches to solve the shortage of low-risk samples.
3) A new intervention training strategy is proposed that uses the APF to improve the performance of reinforcement learning in dangerous environments and adopts a risk evaluation to avoid the algorithm falling into the local optimum. We design two scenarios with different difficulties, and the superiority of the proposed algorithm is analyzed.

The rest of this study is organized as follows. Section II introduces the preliminaries and reinforcement learning-based description of the lane-changing task. Then, a DRL-based decision-making framework is proposed, and the framework details are discussed in Section III. Next, the experiments and results are presented and discussed in Section IV. Finally, the conclusion and future work of this study are described.

II. Preliminaries and Problem Formulation

A. Markov Decision Process

The Markov Decision Process (MDP) is a theoretical framework for implementing goal-based decision-making tasks through interactive learning. The agent is an individual who performs the learning and implements the decision-making behavior. All objects interacting with the agent are collectively called the environment. The agent interacts with the environment, and the environment responds accordingly to the action chosen, presenting the agent with new states and generating rewards [39]. MDP is usually represented as a 5-tuple $\langle S, A, P, R, \gamma \rangle$, where S is the set of environment states, A is the set of actions, P is the state transition probability, and R is the reward function. In this study, the lane change task is based on the information about the environment the ego-vehicle observes within a certain range. First, the vehicle chooses the action $a_t \in A$ based on the observed state $s_t \in S$ at the current time step t. Then, it acts on the environment. Finally, the unmanned vehicle can receive the reward $r_{t+1} \in R$ and arrive at the new state $s_{t+1} \in S$. The problem can be described as the finite-horizon MDP problem, and the process can be written as the finite-horizon MDP trajectory $\{s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_t, a_t, r_{t+1}\}$. The objective is to maximize the cumulative reward by the adopted policy π as follows:

$$G_t = \sum_{i=0}^{\infty} \gamma^i \cdot r(s_{i+t}, a_{i+t}). \tag{1}$$

The state-value function $v_\pi(s)$ of an MDP is the expected return starting from state s, and then the Bellman expectation equation for $v_\pi(s)$ is defined as follows:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s\right], \tag{2}$$

where γ is called the discount factor and γ is less than 1.

The action-value function $q_\pi(s, a)$ is the expected return starting from state s, taking action a, and then the Bellman expectation equation for $q_\pi(s, a)$ is defined as follows:

$$q_\pi(s; a) = \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a\right]. \tag{3}$$

The iterative goal of MDP is to find an optimal value function that achieves the best possible performance in MDP:

$$v_*(s) = \max_\pi v_\pi(s), \tag{4}$$

$$q_*(s, a) = \max_\pi q_\pi(s; a). \tag{5}$$

The optimal strategy is generated by iterating through the Bellman equation:

$$\pi(s) = \arg\max_{a \in A} q_\pi(s; a). \tag{6}$$

B. Twin Delayed Deep Deterministic Policy Gradient (TD3)

In this study, we adopt Twin Delayed Deep Deterministic Policy Gradient (TD3) [40] as the fundamental model in our framework. Among the reinforcement learning algorithms that can be used to solve the continuous action space include asynchronous policy algorithms such as Deep Deterministic Policy Gradient (DDPG) [41], Soft Actor-Critic (SAC) [42] and TD3, and synchronous policy algorithms such as Trust Region Policy Optimization (TRPO) [43], and Proximal Policy Optimization (PPO) [44]. The synchronous policy algorithm could be more inefficient in sampling, requiring new samples to be taken at each gradient update step. On the contrary, the asynchronous policy algorithm uses an experience buffer to store samples. During the gradient update, the samples are randomly sampled from the experience buffer, which is reused to improve efficiency. Among them, TD3 is the most popular asynchronous policy reinforcement learning algorithm, which has been successfully applied to unmanned systems [45], [46], [47], [48]. The TD3 algorithm can be seen as an upgraded version of DDPG, an actor-critic framework for deterministic output, which inherits the strengths of the deep Q-networks (DQN) [49] and deterministic policy gradient (DPG) [50].

The DDPG algorithm uses an actor-critic framework and borrows the concept of target networks from DQN, using the dual network structure with parameters θ^{π}, θ^{Q} and target parameters $\theta^{\pi'}$, $\theta^{Q'}$ to achieve more stable results.

$$\left\{Actor(\theta^{\pi}), Actor\ target(\theta^{\pi'})\right\} \in Actor, \tag{7}$$

$$\left\{Critic(\theta^{Q}), Critic\ target(\theta^{Q'})\right\} \in Critic. \tag{8}$$

1) Update of Actor network

$$L_{actor} = \underset{s_{t+1}\sim p(\cdot|s_t,a_t)}{\mathbb{E}}[Q\left(s_t, \pi\left(s_t \mid \theta^{\pi}\right)\right) \mid \theta^{Q}\big) \nabla_{\theta^{\pi}} \pi\left(s_t \mid \theta^{\pi}\right)], \tag{9}$$

where $Q(\cdot)$ is the optimal value function and $\pi(s_t \mid \theta^{\pi})$ is a Q-optimal policy.

2) Update of Critic network

$$L_{critic} = \underset{s_{t+1}\sim p(\cdot|s_t,a_t)}{\mathbb{E}}[\left(y_t - Q\left(s_t, a_t \mid \theta^{Q}\right)\right)^2], \tag{10}$$

where $y_t = r_t + \gamma \cdot Q'(s_{t+1}, \pi(s_{t+1} \mid \theta^{\pi'}) \mid \theta^{Q'})$ is the target value and $y_t - Q(s_t, a_t \mid \theta^{Q})$ is TD-error.

The actor target and critic target network are updated differently to DQN, using a soft update for the parameters:

$$\theta^{Q'} \leftarrow \tau \cdot \theta^{Q} + (1-\tau) \cdot \theta^{Q'}, \tag{11}$$

$$\theta^{\pi'} \leftarrow \tau \cdot \theta^{\pi} + (1-\tau) \cdot \theta^{\pi'}, \tag{12}$$

where τ is a soft update factor between 0 and 1.

More algorithm details can be observed in the study [41].

The TD3 algorithm retains the overall framework of the DDPG algorithm and makes improvements that have helped improve the algorithm's performance.

- TD3 uses two critic networks to update Q values to solve the problem of over-estimation of Q values.

$$y_t = r_t + \gamma \cdot \min(Q_1'(s_{t+1}, \pi(s_{t+1} \mid \theta^{\pi'}) \mid \theta_1^{Q'}), \\ Q_2'(s_{t+1}, \pi(s_{t+1} \mid \theta^{\pi'}) \mid \theta_2^{Q'})). \tag{13}$$

- A delayed update strategy is proposed to prevent training instability. The action network is updated every certain number of steps d, while the critic network continues to use the single-step update strategy.
- A normal distribution noise $(\varepsilon \sim N(0,\sigma))$ with a certain range $(-c, c)$ is added to the target action, which makes the policy less likely to exploit actions with high Q-value estimates.

$$y_t = r_t + \gamma \cdot \min(Q_1'(s_{t+1}, \pi(s_{t+1} \mid \theta^{\pi}) \mid \theta_1^{Q'}), \\ Q_2'(s_{t+1}, \pi(s_{+1} \mid \theta^{\pi'}) \mid \theta_2^{Q'})) + \text{clip}(\varepsilon, -c, c)). \tag{14}$$

III. Proposed Framework

In this section, road and vehicle risk potential fields are first developed based on different attributes. Then an actor-critic framework for risk task decomposition is proposed. Next, the risk samples are classified according to the critical differences and a data augmentation idea are applied to low risk experience buffer. Also, An new intervention training strategy is proposed. Finally, the complete intervention risk classification TD3-task decompose (IDRCTD3-TD) algorithm is proposed based on the above components.

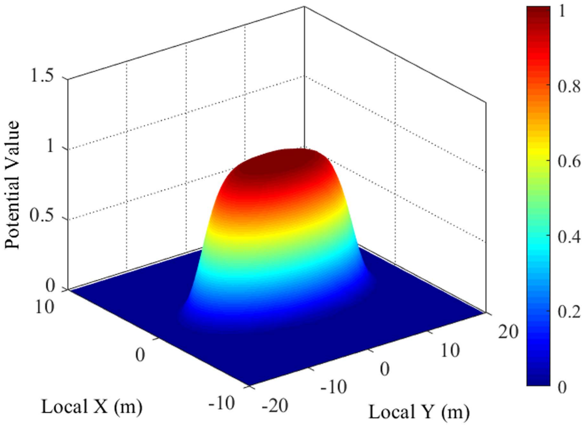


Fig. 2. Potential risk diagram of the vehicle.

A. Classification of the Driving Risk Potential Field

In this study, driving risk is subdivided into road boundary risk and vehicle risk, which is the critical concern for regularizing roads to address the lane change problem for unmanned vehicles. Therefore, the corresponding risk potential fields are constructed according to the risk sources.

1) Vehicle Potential Field: According to the sensors information, the ego vehicle can detect the position and speed of the surrounding vehicles (SVs) relative to the host vehicle (HV). Fig. 2 shows the results of the risk field distribution, obstacles in the potential field are represented using a squared-negative-exponential form [51], and we incorporate virtual distances to rectify real-space distances, aligning the risk field strength's area of influence with the genuine hazardous conditions, specifically, longitudinal relative velocities are introduced to adjust longitudinal distances. Considering the above information, a vehicle risk potential field E_v is built, as follows:

$$E_v = A_v \cdot M_v \cdot e^{-\left(\frac{(x-u_x)^2}{\sigma_x^2} + e^{-D_y \beta_y v_y} \cdot \frac{(y-u_y)^2}{\sigma_y^2}\right)^k}, \tag{15}$$

$$D_y = \begin{cases} 1, & \text{if } y_{ego} < y_{sv}, \\ -1, & \text{otherwise }, \end{cases} \tag{16}$$

where A_v denotes the field intensity coefficient of the vehicle, M_v denotes the equivalent mass [18], which is directly proportional to the vehicle velocity, (u_x, u_y) denotes the coordinates of the risk source, β_y denote the speed factors in longtitudinal diretion. σ_x and σ_y are the shape function of the obstacle. D_y denote relative location factor in longtitudinal diretion and v_y denote relative speed in longtitudinal direction, where $v_y = v_y^{ego} - v_y^{sv}$.

In additional, the pose of the vehicle is also considered to reflect the variation of the risk field in different directions.

$$\begin{cases} x' = (x - u_x) \cdot \cos\varphi - (y - u_y) \cdot \sin\varphi + u_x, \\ y' = (x - u_x) \cdot \sin\varphi + (y - u_y) \cdot \cos\varphi + u_y, \\ E_v' = E_v, \end{cases} \tag{17}$$

where φ denotes the heading angle.

2) Road Boundary Potential Field: The potential risk arises at the road boundary and characterizes the risk to vehicles because vehicles close to the road boundary limit the driving

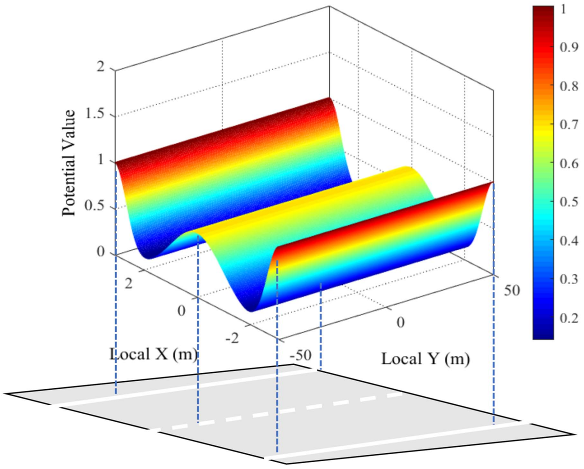


Fig. 3. Potential risk diagram of the road boundary.

area. Road boundary potential fields E_r are as follows:

$$E_b = A_b \cdot e^{-\frac{(l_{pos}-l_{lane})^2}{2{\sigma_r}^2}}, \tag{18}$$

where A_b denotes the field intensity coefficient of the road boundary, l_{pos} denotes the lateral position of HV, and l_{lane} represents the position of the lane boundary. σ_b determines the risk distribution range of the road boundary. As σ_b becomes larger, the impact of risk distribution becomes wider.

3) Lane-Marking Potential Field: Lane marking is less dangerous than the road boundary, and it should be ensured that the vehicle runs on the center line of each lane as much as possible [52]. The lane-marking potential field E_l is as follows:

$$E_l = A_l \cdot e^{-\frac{(l_{center}-l_{pos})^2}{2\sigma_l^2}}, \tag{19}$$

where A_l denotes the field intensity coefficient of the lane marking, l_{center} denotes the position of lane marking. Similarly, σ_l also determines the risk distribution range of the lane marking. In this study, E_l and E_b are collectively referred to as the road risk field E_r and the results of the risk field distribution for E_r shown in Fig. 3.

B. MDP Model

According to the description in Section II, a lane change process can be described as an MDP decision process, the components of which are described below.

1) State: The information about the SVs and information about the HV are considered inputs to the state, which are specified as follows

$$SV_i = [\mathbb{I}_i, \Delta X_{sv_i}, \Delta Y_{sv_i}, \Delta V_{sv_i}, \Delta V_{sv_i}, yaw_{sv_i}], \tag{20}$$

where $\mathbb{I}_i$ denotes the lane id where SV$_i$ is located. ΔX_{sv_i} and ΔY_{sv_i} denote the relative lateral distance and the longitudinal distance between the SV$_i$ and the HV, respectively, ΔV_{sv_i} and ΔV_{sv_i} denote the relative lateral speed, and the longitudinal speed between SV$_i$ and HV, respectively, and yaw_{sv_i} denotes the yaw angle of the SV$_i$. The state information of HV is described as follows:

$$HV = [\mathbb{I}_{ego}, X_{ego}, V_{ego}^x, V_{ego}^y, yaw_{ego}], \tag{21}$$

where $\mathbb{I}_{ego}$ denotes the lane id where the HV is located, X_{ego} denotes the lateral position of HV, V_{ego}^x and V_{ego}^y denote the lateral and longitudinal speeds of HV, and yaw_{ego} denotes the yaw angle of HV.

2) Action: This study is concerned with the consideration of the angle of the steering wheel as the action. For longitudinal behaviors, an advanced driving assistance system or an intelligent driver model (IDM) [53] can control the target speed [31]. The rotation angle of the steering wheel can be expressed as:

$$a_t \in [-\lambda \cdot \pi, \lambda \cdot \pi], \tag{22}$$

where negative values represent a left-turn command, and positive values represent a right-turn command, and λ is the scaling factor that limits the steering range [36], λ is $\frac{1}{12}$.

Since model-free reinforcement learning algorithms sometimes can lead to unstable actions that can make passengers feel uncomfortable, an exponential smoothing strategy is applied to smooth the planned path [31], [34], denoted as follows:

$$a_t^{final} = a_{t-1} + \omega \cdot (a_t - a_{t-1}), \tag{23}$$

where a_t^{final} denotes smoothed action, ω denotes the smoothing factor, a_t and a_{t-1} denote the actions generated by the reinforcement learning at time t and $t-1$, respectively.

3) Reward: In the problem solved in this study, our goal is to find a strategy that maximizes the cumulative reward. However, the definition of the risk potential field is to find a strategy that minimizes the risk. Therefore, a negative sign is added to the risk value to transform the minimization risk problem into the maximization reward problem. The maximum risk value is added in the risk field to the risk description to represent driving safety because a positive reward is intrinsically more consistent with human intuition [54]. The reward is specifically defined as follows:

$$R = \underbrace{|\max E_v - E_v|}_{R_v} + \underbrace{|\max E_b - E_b| + |\max E_l - E_l|}_{R_r} + R_{exit}, \tag{24}$$

where $\max E_v$, $\max E_b$ and $\max E_l$ represent the maximum value defined by the vehicle risk potential field, road boundary potential field and lane-marking potential field. $\max E_v$, $\max E_b$ both are 1, and $\max E_l$ are 0.6. R_{exit} is added to the total reward to encourage vehicles to drive as free of collisions as possible at each timestep t. $R_{exit} = 0.1$ means the vehicle has no collision and stays within the road boundary, and $R_{exit} = -1$ means that the vehicle collides with other SVs or the road boundary. The values 0.1 and -1 in R_{exit} are commonly used in previous studies [31].

C. Risk Task Decomposition Architecture

Driving risk is classified into road and vehicle risks according to the classification in Section A. However, the direct translation of risk values into reward functions inevitably results in the homogenization of all risks and the inability to identify their categories. When a collision occurs, for instance, the agent cannot determine which aspect of the risk is causing its collision behavior. Inspired by [55], [56], a risk task decomposition framework is proposed and deployed on the TD3 algorithm, note that other actor-critic algorithms can also apply this framework. The framework of TD3-TD is shown in Fig. 4, the different

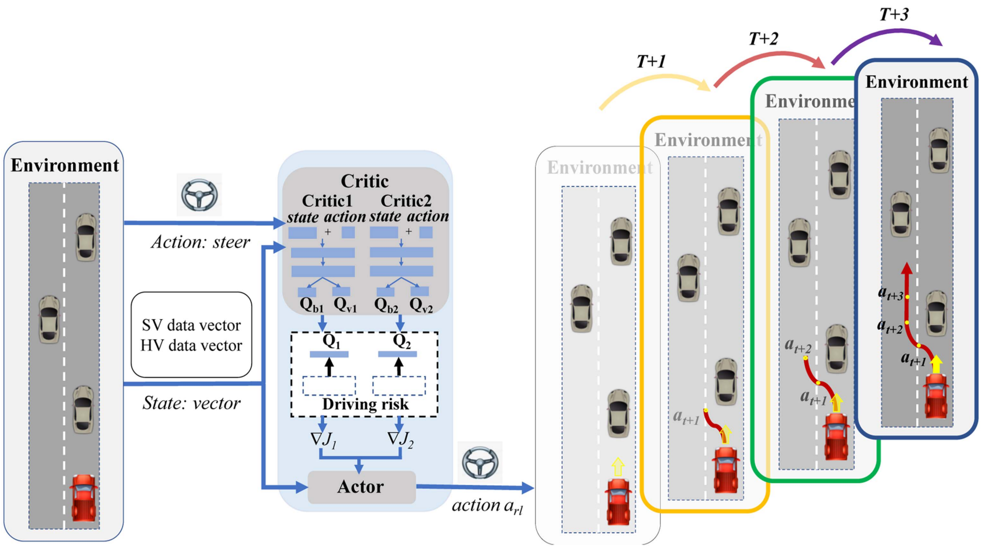


Fig. 4. Risk decomposition architecture for TD3.

branches in the risk task decomposition architecture get the state information of the HV and the SVs from the environment, respectively, by decomposing the driving task to learn the different risk branches in parallel in the critic network, and finally output the decision-making behavior to the environment through the actor network. The risk rewards for different branches are described as follows:

$$r_t^{risk}\left(s_t, a_t, s_{t+1}\right) = \sum_{b=r,v} r_t^b\left(s_t, a_t, s_{t+1}\right), \tag{25}$$

$$r_t^b\left(s_t, a_t, s_{t+1}\right) = R_t^b\left(s_t, a_t, s_{t+1}\right) + R_t^{exit}\left(s_t, a_t, s_{t+1}\right), \tag{26}$$

where unlike aggregated reward functions that consider the single update of R_{exit} at each timestep, in the proposed framework, R_{exit} needs to be considered for updates in both the vehicle risk and road risk branches, so that R_{exit} reflects the importance in different branches to ensure optimality.

As two risk branches are considered in the proposed framework, the critic network uses two branches to learn Q values from the risk task separately. Critic networks based on optimal value functions can jointly learn policies for different risk branches, and the loss function is calculated as follows:

$$\begin{aligned} L_{critic} = & \underset{s_{t+1}\sim p(\cdot|s_t,a_t)}{\mathbb{E}} \sum_{b=r,v} \left[\left(y_t^b - Q_1^b\left(s_t, a_t \mid \theta_1^{Q^b}\right)\right)^2 \right. \\ & \left. + \left(y_t^b - Q_2^b\left(s_t, a_t \mid \theta_2^{Q^b}\right)\right)^2 \right], \end{aligned} \tag{27}$$

$$\begin{aligned} y_t^b = r_t^b + \gamma \cdot \min(& Q_1^{b'}(s_{t+1}, \pi(s_{t+1} \mid \theta^{\pi'}) \mid \theta_1^{Q^{b'}}), \\ & Q_2^{b'}(s_{t+1}, \pi(s_{t+1} \mid \theta^{\pi'}) \mid \theta_2^{Q^{b'}})). \end{aligned} \tag{28}$$

In the critic network, each sub-critic network corresponds to the corresponding risk branch, and risk preferences are optimized using different risk branches, so the linear combination of risk branches calculates the final Q value:

$$Q\left(s_t, a_t; \theta^Q\right) = \sum_{b=r,v} Q_b\left(s_t, a_t; \theta^{Q^b}\right). \tag{29}$$

According to (29), the loss function of the actor network is rewritten as follows:

$$\begin{aligned} L_{actor} &= \sum_{b=r,v} \nabla J_b(\theta^\pi) \\ &= \underset{s_{t+1}\sim p(\cdot|s_t,a_t)}{\mathbb{E}} \sum_{b=r,v} Q_b\left(s_t, \pi\left(s_t\right) \mid \theta_1^{Q^b}\right) \nabla_{\theta^\pi} \pi\left(s_t \mid \theta^\pi\right). \end{aligned} \tag{30}$$

D. Low-Risk Episodic Sampling Augmentation Method in Different Risk Branches

Some experience samples facilitate model learning during training more effectively than others. However, spending more time on low-quality samples, the equal probability selection of experience samples can lengthen the algorithm's training time. Previous research has shown that samples of high-return rewards positively affect learning [37], [57], so we would like to take full advantage of these experiences and integrate them into the proposed framework.

At the first level, risk samples are classified according to the cumulative reward in each episode, and a certain number m_ε of low-risk samples is sampled according to a certain percentage p. By continuously replaying the low-risk samples sampled from the high-reward episode trajectories during training, the strategy is continuously improved toward the "low-risk" driving strategy while improving the learning efficiency of the model. Specifically, two experience buffer, D_l and D_h, are established to classify the samples according to the cumulative rewards of the episodes. Initially, samples from the current episode are collected into the temporary cache D_{tc}. After the current episode ended, the cumulative rewards of the current episode are calculated, and the samples are classified by determining the level of risk of the samples according to the classification conditions set. Subsequently, the maximum value of the episode cumulative rewards is updated according to the cumulative rewards of the current episode. The classification condition and update conditions can be written as:

$$\underbrace{(R > R^{\max} * k)}_{condition1} \cup \underbrace{(R_v > R_v^{\max} * k_v)}_{condition2} \cup \underbrace{(R_r > R_r^{\max} * k_r)}_{condition3}, \tag{31}$$

$$\underbrace{(R > R^{\max})}_{condition1} \cup \underbrace{(R_v > R_v^{\max})}_{condition2} \cup \underbrace{(R_r > R_r^{\max})}_{condition3}, \tag{32}$$

where R, R_v, R_r denote the cumulative reward for the current episode, the cumulative reward of the vehicle risk branch for the current episode and the cumulative reward of the road risk branch for the current episode, respectively, and $R^{\max}$, $R_v^{\max}$, $R_r^{\max}$ denote the maximum cumulative reward in the past episodes, the maximum cumulative reward for the vehicle risk branch for the past episodes and the maximum cumulative reward for the road risk branch for the past episodes, respectively. k, k_v, k_r denote the percentage of risk classification, the percentage of vehicle risk classification, and the percentage of road risk classification, respectively.

After the first-level classifying, we notice that the number of samples in the low-risk experience buffer is significantly less than the ones in the high-risk experience buffer, which indicates a significant shortage of high-quality experiences. Thereby, inspired by existing studies [58], this study proposes a method for low-risk episodic sampling augmentation with reinforcement learning in unmanned vehicles. We define a one-to-one mapping $\sigma : D_l \rightarrow D_l$ that maps the samples of different risk branches in the low-risk experience buffer to the new sample experience buffer.

$$e^{(\sigma)}(s_t, a_t, r_t^v, s_{t+1})_{D_{lv}} = A(\sigma) \cdot e(s_t, a_t, r_t^v, s_{t+1}))_{D_{lv}}{}^T, \tag{33}$$

$$e^{(\sigma)}(s_t, a_t, r_t^v, s_{t+1})_{D_{lr}} = A(\sigma) \cdot e(s_t, a_t, r_t^r, s_{t+1}))_{D_{lr}}{}^T, \tag{34}$$

where $e(s_t, a_t, r_t^r, s_{t+1}))_{D_{lv}}$ and $e(s_t, a_t, r_t^r, s_{t+1}))_{D_{lr}}$ denote the cumulative trajectory samples of the whole episode in different risk branches, $e^{(\sigma)}(s_t, a_t, r_t^r, s_{t+1})_{D_{lr}}$ and $e^{(\sigma)}(s_t, a_t, r_t^v, s_{t+1})_{D_{lv}}$ denote the transformed trajectory samples of the whole episode in different risk branches, respectively. $A(\sigma)$ denotes permutation vector. Here, σ is a symmetry that exist one-to-one mapping and $e^{(\sigma)}(s_t, a_t, r_t^v, s_{t+1})_{D_{lv}}$ and $e^{(\sigma)}(s_t, a_t, r_t^r, s_{t+1})_{D_{lr}}$ as shown below:

$$\begin{cases} SV_i^{(\sigma)} = [1 - \mathbb{I}_i, -\Delta X_{sv_i}, \Delta Y_{sv_i}, -\Delta V_{sv_i}, -yaw_{sv_i}], \\ HV^{(\sigma)} = [1 - \mathbb{I}_{ego}, -X_{ego}, -V_{ego}^x, V_{ego}^y, -yaw_{ego}], \\ a^{(\sigma)} = -a, \\ r_v^{(\sigma)} = r_v, \\ r_r^{(\sigma)} = r_r, \end{cases} \tag{35}$$

where $SV_i^{(\sigma)}$, $HV^{(\sigma)}$, $HV^{(\sigma)}$, $a^{(\sigma)}$, $r_v^{(\sigma)}$ and $r_r^{(\sigma)}$ denote the transformed vectors from SV_i, HV, a, r_v and r_r. Note that $A(\sigma)$ is one of the permutation rules, and other permutation rules can be used to complete the mapping, e.g., translations or rotations.

Reward invariance: σ is a form of symmetry in which maps the episodic samples $e(s_t, a_t, r_t^v, s_{t+1}))_{D_{lv}}$ and $e(s_t, a_t, r_t^r, s_{t+1}))_{D_{lr}}$ to new episodic samples $e^{(\sigma)}(s_t, a_t, r_t^v, s_{t+1})_{D_{lv}}$ and $e^{(\sigma)}(s_t, a_t, r_t^r, s_{t+1})_{D_{lr}}$. The symmetry $\sigma : D_l \rightarrow D_l$ only change position of the trajectory in space, the order of the rewards and the values of the rewards do not change with symmetry.

Mapping decomposable invariance: mapping $\sigma : D_l \rightarrow D_l$ is a one-to-one mapping for the low risk sample experience buffer. According to *reward invariance* and (25), $e^{(\sigma)}(s_t, a_t, r_t, s_{t+1})_{D_l}$ can be decomposed into $e^{(\sigma)}(s_t, a_t, r_t^v s_{t+1})_{D_{lv}}$ and $e^{(\sigma)}(s_t, a_t, r_t^r, s_{t+1})_{D_{lr}}$. Mapping $\sigma : D_l \rightarrow D_l$ can naturally be applied to different risk branches in low risk sample experience buffer.

We apply the symmetry $\sigma : D_l \rightarrow D_l$ to episodic samples from the low-risk experience buffer D_l and put them into D_l.

$$\mathbf{D}_{lr} \leftarrow \mathbf{D}_{lr} \cup \left\{ e_i \left(s_t^{(\sigma)}, a_t^{(\sigma)}, r_t^r, s_{t+1}^{(\sigma)} \right)_{i=1,2,3\ldots n} \right\}, \tag{36}$$

$$\mathbf{D}_{lv} \leftarrow \mathbf{D}_{lv} \cup \left\{ e_i \left(s_t^{(\sigma)}, a_t^{(\sigma)}, r_t^v, s_{t+1}^{(\sigma)} \right)_{i=1,2,3\ldots n} \right\}, \tag{37}$$

$$\mathbf{D}_l \leftarrow \mathbf{D}_{lr} \cup \mathbf{D}_{lv}. \tag{38}$$

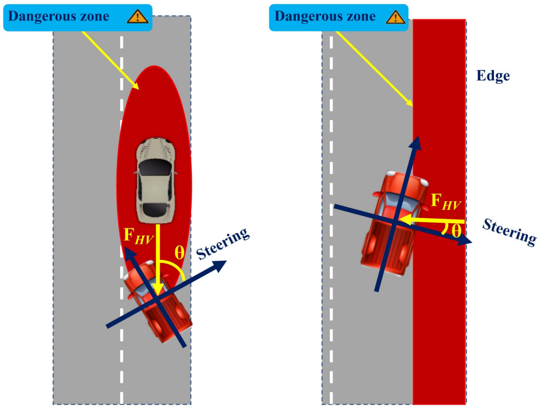
E. Intervention Training Strategy

In this study, a intervention training strategy is proposed to further speed up the training and improve overall performance. Firstly, the repulsive potential field force decomposition strategy is applied to the intervener controller [29]. According to the description of the potential risk field, the field force F is repulsive, and the larger the field intensity E, the larger the field force F, which means that the field intensity is proportional to the field force. Therefore, the electric field force formula can describe this relationship. According to the electric field force formula: $F = E \cdot q$, in which E can be interpreted as the electric field intensity generated by the vehicle or road boundary, q is the quantity of electric charge carried by the object. Note that the road boundary is only used to describe the risk of the road because the risk of HV close to the road boundary is the only concern. In this study, we consider all vehicles and the road boundary as obstacles, q is related to the motion state of HV, and consequently, the field force $F_{HV(\boldsymbol{o})}$ formed by the road boundary potential field and the vehicle potential field is described as follows:

$$F_{HV(\boldsymbol{o})} = \begin{cases} e^{v} \cdot |E_v| & \text{if field generated by the vehicle,} \\ e^{v_x} \cdot |E_b| & \text{other,} \end{cases} \tag{39}$$

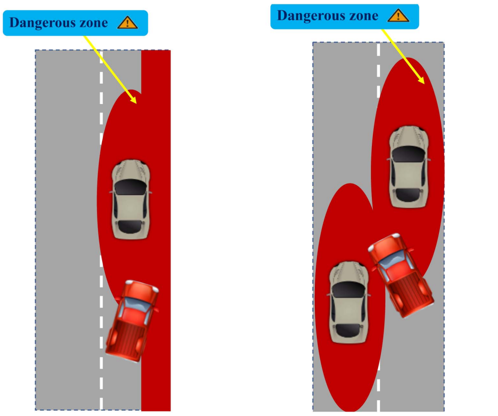
where $F_{HV(\boldsymbol{o})}$ denotes the electric field force on the HV by the obstacle object $\boldsymbol{o}$ and v, v_x denote the speed of HV and the lateral speed of HV, respectively.

The effect of the intervention for different types of obstacles is shown in Fig. 5. The red zone is dangerous, the yellow lines represent the repulsive force, and the blue coordinate decompose the HV decision behaviour. In Fig. 5(a), the dangerous zone of the vehicle is described using an ellipse, where $\min(t_{ttc}, d_{gap})$ is used for the long axis and the minimum safety distance R_s for the short axis. Safe headway distance $d_{gap}^{safe} = v_{ego} \cdot t_{thw}^{safe}$ and safe time-to-collision (TTC) t_{ttc}^{safe}= 1.5 s are used to limit the occurrence of longitudinal risks in vehicles [59]. The intervention controller module will activate the steering if the HV does not meet the minimum safety requirements. Similarly, as



(a) Repulsive force on intervention of the vehicles. (b) Repulsive force on intervention of the road boundary.

Fig. 5. Repulsive force and intervention controller module.



(a) Repulsive force on intervention of the vehicle and the road boundary. (b) Repulsive force on intervention of the vehicle and the vehicle.

Fig. 6. Example case where HV could lead to a collision.

shown in Fig. 5(b), the dangerous zone of the road boundary is equal to $[\max(\frac{l_{lane}}{2}, (l_{lane} - \alpha \cdot v_{egox} \cdot \Delta t)), l_{lane}]$, where α is step length and v_{egox} is the lateral speed of the HV, $\triangle t$ is the scaling factor. As shown in Fig. 5, the angle θ formed between the HV and the obstacle by decomposition of the repulsive forces applied to x (steering) and y (throttle) axis of the coordinates, respectively. Steering axis components are shown in the following equation:

$$F_{rep\_steer} = -\sum_{\boldsymbol{o}} \left\| F_{HV(\boldsymbol{o})} \right\| \cdot \cos\theta_i. \tag{40}$$

The intervention controller controls the final steering, as shown in the following equation:

$$a_{in} = \eta \cdot F_{req\_steer}, \tag{41}$$

where η represents a custom weighting factor, and η is 0.1. a_{in} denotes the steering of the intervention controller.

Algorithm 1: Risk Evaluation ε-Annealing Strategy.

```
Generate random $p \in (0,1)$.
Get the next time predicted risk value $\hat{D}^{in}_{t+\triangle t}$ using the intervention controller.
Get the the next time predicted risk value $\hat{D}^{rl}_{t+\triangle t}$ using the reinforcement learning.
if $\hat{D}^{in}_{t+\triangle t} < \hat{D}^{rl}_{t+\triangle t}$ then
    if $p < \varepsilon$ then
        $a = a_{in}$
    else
        $a = a_{rl}$
    end if
else
    $a = a_{in}$
end if
```

However, the intervention controller can lead to local minima or cause the vehicle to driving towards a more dangerous zone, as shown in Fig. 6. When the HV enters the dangerous zone, the HV is affected by the repulsive forces, which may causes the vehicle to fall into local minima due to the fact that the repulsive forces are in opposite directions. So this study proposes a method that adopts a risk evaluation to avoid the algorithm falling into the local optimum and uses the intervention controller as the assistance controller to improve the performance of reinforcement learning. Specifically, we assume that the vehicle drive at a constant speed and constant steering for $\triangle t$, received an action $a_t = [a^{HV}_t, a^{SV_i}_t], i \in \{1, 2, 3 \ldots, N\}$ and the next motion state of the vehicle is described by the kinematics equations:

$$\begin{cases} x_{t+\Delta t} = x_t + \Delta t \cdot v_t \cdot \cos(\varphi_t), \\ y_{t+\Delta t} = y_t + \Delta t \cdot v_t \cdot \sin(\varphi_t), \\ v_{t+\Delta t} = v_t + \Delta t \cdot a_t, \\ \varphi_{t+\Delta t} = \varphi_t + \Delta t \cdot v_t \cdot \frac{\tan(\varphi_t)}{L}, \end{cases} \tag{42}$$

where $\triangle t$ is the decision period and L is the distance between shafts.

When the HV enters the dangerous zone, risk value is compared as the switching strategy, reinforcement learning and intervention controller output the corresponding action a_{rl} and a_{in} according to the state s_t, we choose the action by the predicted risk values $\hat{D}^{in}_{t+\triangle t}$ and $\hat{D}^{rl}_{t+\triangle t}$ at the next time $(t + \triangle t)$. It is worth noting that the intervention controller can only guarantee the single-step optimum, which often makes the intervenor not work. In this study, we adopt the intervention controller as an assistance controller to improve the performance of reinforcement learning. Our strategy is shown in the following Algorithm 1.

As the number of episodes increases, the times of actions generated by the intervention controller decreases, and finally, the intervention controller is discarded. Since IDRCTD3-TD is an off-policy reinforcement learning method, where the algorithm updates the policy through the experience buffer, the algorithm can benefit from learning samples generated by the intervention controller and more easily explore favorable experiences to speed up the training.

Algorithm 2 depicts the whole training process of the framework based on the descriptions of the above subsection.

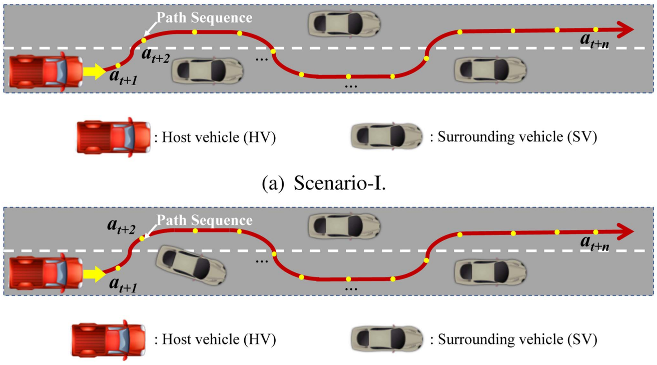
IV. Simulation and Experiments Result

Given the high-risk factor of actual vehicles and legal restrictions, scenario-based virtual testing has the advantages of high

Algorithm 2: IDRCTD3-TD.

```
Initialization:
Initialize actor network $\pi(s \mid \theta^{\pi})$ and critic network $Q_1^r(s,a \mid \theta_1^{Q^r}), Q_1^v(s,a \mid \theta_1^{Q^v}), Q_2^r(s,a \mid \theta_2^{Q^r}), Q_2^v(s,a \mid \theta_2^{Q^v})$.
Initialize actor target network $\theta^{\pi'} \leftarrow \theta^{\pi}$ and critic target network $\theta_1^{Q^{r'}} \leftarrow \theta_1^{Q^r}, \theta_1^{Q^{v'}} \leftarrow \theta_1^{Q^v}, \theta_2^{Q^{r'}} \leftarrow \theta_2^{Q^r}, \theta_2^{Q^{v'}} \leftarrow \theta_2^{Q^v}$.
Implementation:
for $i = 1$ to $m$ do
    for $t = 1$ to $n$ do
        Select action $a$ according to $\pi(s_t \mid \theta^{\pi}) + noise$.
        if the intervener condition is satisfied then
            Execute action $a_t = a_t^{in}$ and get next state $s_{t+1}$ and reward $r_t^v$ and $r_t^r$.
        else
            Execute action $a_t = a_t^{rl}$ and get next state $s_{t+1}$ and reward $r_t^v$ and $r_t^r$.
        end if
        Store transition $(s_t, a_t, s_{t+1}, r_t^v)$ and $(s_t, a_t, s_{t+1}, r_t^r)$ to temporary cache $D_{tc}$.
        Sample $m_\varepsilon$ transition $(s_t, a_t, s_{t+1}, r_t^v)$ and $(s_t, a_t, s_{t+1}, r_t^r)$ from experience buffer $D_h$ with probability $\varepsilon$ and $m_{1-\varepsilon}$ transition $(s_t, a_t, s_{t+1}, r_t^v)$ and $(s_t, a_t, s_{t+1}, r_t^r)$ from experience buffer $D_l$ with probability $1 - \varepsilon$.
        Update critic network loss by (27).
        set $y_i$ by (29).
        if $t \,\%\, d = 0$ then
            Update actor network loss by (30).
            Update actor and critic target network:
            $\theta^{\pi'} \leftarrow \tau\theta^{\pi} + (1-\tau)\theta^{\pi'}$
            $\theta_i^{Q^{v'}} \leftarrow \tau\theta_i^{Q^v} + (1-\tau)\theta_i^{Q^{v'}}$
            $\theta_i^{Q^{r'}} \leftarrow \tau\theta_i^{Q^r} + (1-\tau)\theta_i^{Q^{r'}}$
        end if
    end for
    Sampling classification and augmentation:
    if (31) is satisfied then
        Store current episode sample to experience buffer $D_l$.
        Mapping experience buffer $D_{lv}$ and $D_{lr}$ to experience buffer $D_{lv}^{(\sigma)}$ and $D_{lr}^{(\sigma)}$.
        $D_l \leftarrow D_l \cup (D_{lv}^{(\sigma)} \cup D_{lr}^{(\sigma)})$.
    else
        Store current episode sample to experience buffer $D_h$.
    end if
    if (32) is satisfied then
        Update $R^{\max} = R$, $R_v^{\max} = R_v$, $R_r^{\max} = R_r$.
    end if
    Empty the Temporary cache $D_{tc}$.
end for
```

environment reproduction, high testing efficiency, so this study builds the scenarios based on the CARLA [60] virtual simulation platform, where the computer configuration, and the environment configuration, are described as follows: Ubuntu18.04, Python3.7, Pytorch1.10.0, the CPU is i7-11700f, RAM capacity is 16 GB and the GPU is RTX3060Ti.



(a) Scenario-I.

(b) Scenario-II.

Fig. 7. Diagram for two scenarios.

TABLE I
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND TRADITIONAL METHOD

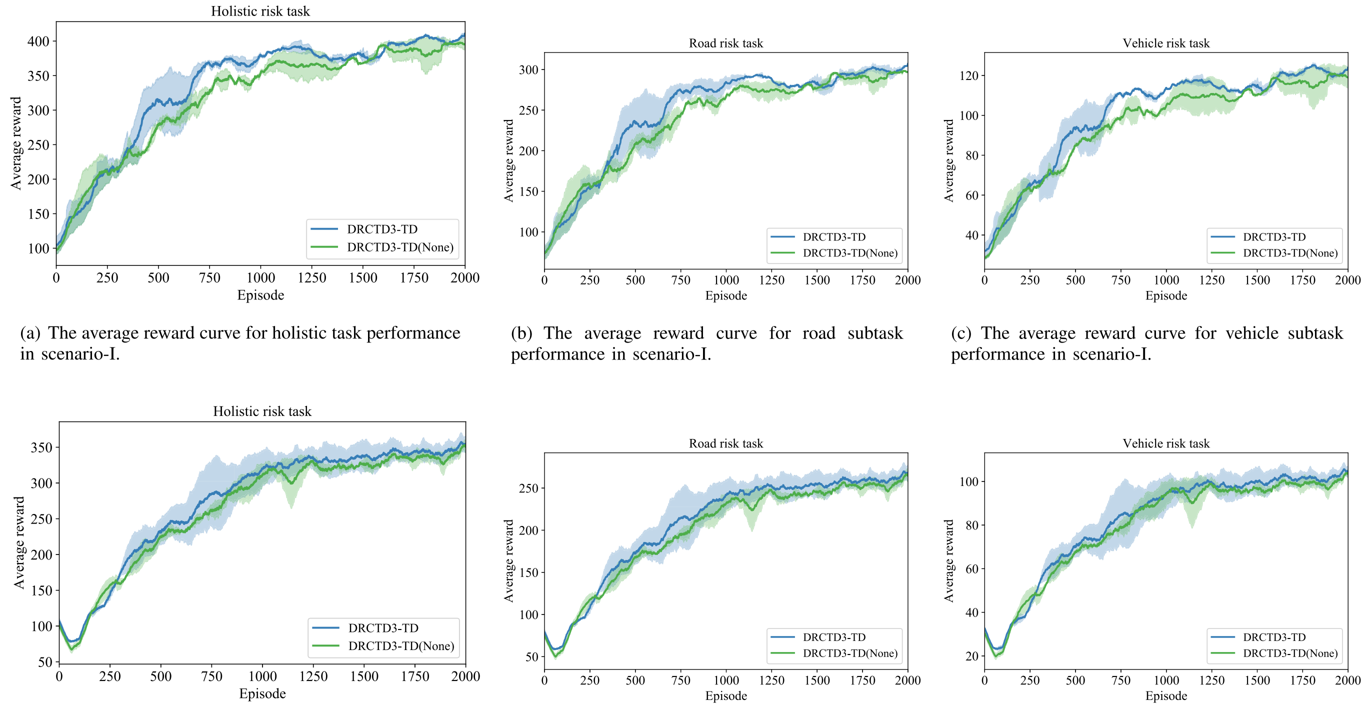
Scenario	Performance	DRCTD3-TD(None)	DRCTD3-TD
scenario-I	Average return	430.76±2.38	435.25±1.58
	Collision rates(%)	6.80±1.85	4.20±0.83
	Surviving distance(m)	385.64±1.52	387.82±1.46
	Finish rates(%)	95.22	95.76
scenario-II	Average return	380.35±3.82	395.67±1.92
	Collision rates(%)	12.40±3.01	8.80±3.29
	Surviving distance(m)	370.76±3.37	376.54±2.78
	Finish rates(%)	91.55	92.97

A. *Scenario Description*

In the training phase, 14–20 vehicles are placed on a straight road 405 m long, The same scenario in CARLA can be observed in [31], [34]. All SVs are in dynamic driving and are set to be on the CARLA autopilot. Therefore, the HV should overtake obstacles in the environment to ensure safe driving, and in the environment, the initial speed of HV is 0 and the target speed is set 12 m/s–16 m/s, SVs are initialized with 5 m/s–8 m/s. Furthermore, we design the scenarios with different difficulties to evaluate the performance of the algorithm during the evaluation phase, the corresponding evaluation scenarios are depicted in Fig. 7. To demonstrate the effectiveness of the algorithm, algorithm is evaluated in 100 episodes, where HV drives approximately 40 km and makes approximately 1,000 lane changes in each scenario, the placement of the vehicle is initialized randomly, determining both its position and lane selection through Gaussian-based sampling [31].

Scenario-I: In the evaluation scenarios, to demonstrate the robustness of the method, 16–23 vehicles are placed in the dense flow. The scenario is challenging in the dense flow, where the mean barrier spacings are only 19 m.

Scenario-II: In this scenario, all SVs has a possibility of p ($p = 0.3$) to lane change, and this behaviour sometimes may



(a) The average reward curve for holistic task performance in scenario-I.

(b) The average reward curve for road subtask performance in scenario-I.

(c) The average reward curve for vehicle subtask performance in scenario-I.

(d) The average reward curve for holistic task performance in scenario-II.

(e) The average reward curve for road subtask performance in scenario-II.

(f) The average reward curve for vehicle subtask performance in scenario-II.

Fig. 8. Average reward curve over the last 100 episodes in scenario-I and scenario-II.

TABLE II
RL Parameter Value

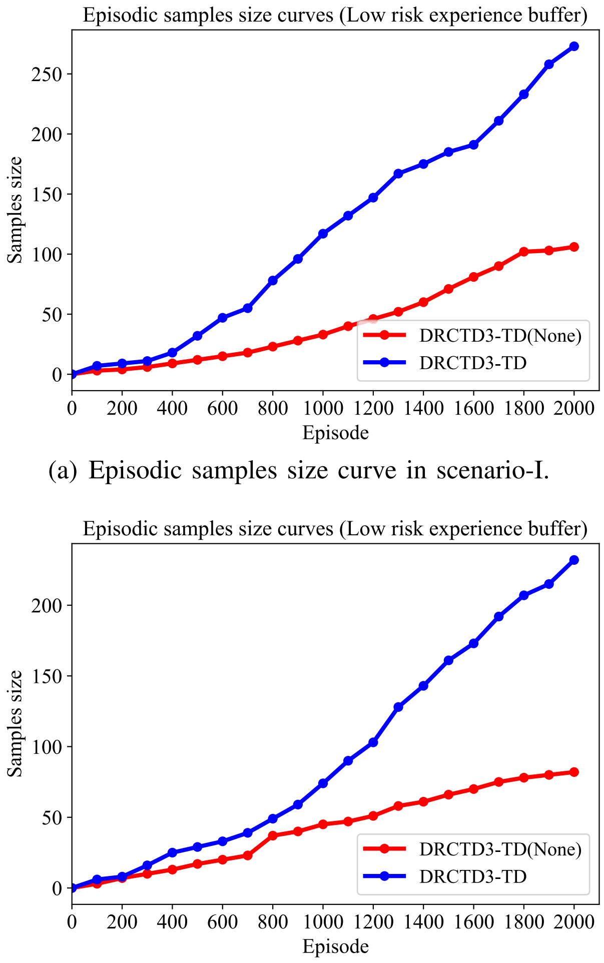
Parameter	Value
Critic learning rate	0.0003
Actor learning rate	0.0003
Hidden layers	2
Hidden layer 1 nodes	256
Hidden layer 2 nodes	256
Soft update coefficient	0.05
Total replay memory size	100000
High risk replay memory size	95000
Low risk replay memory size	5000
Mini batch size	64
Sample probability from D_h	0.9
The percentage of risk classification	0.85
Discount factor	0.99
Soft update factor	0.005
Delayed update step	2

force the HV to make the emergency lane change and overtake. All the other settings are the same as the scenario-I.

B. *Effects of Low-Risk Episodic Sampling Augmentation With the Experience Buffer*

This section explores the impact of low-risk episodic sampling augmentation on the training performance of the algorithm. DRCTD3-TD(None) refers to the fact that no low-risk sample augmentation method is used in the algorithm. We run five experiments in the two scenarios and conduct sensitivity analyses. The corresponding training results are shown in Fig. 8. From Fig. 8(a), (b), and (c), it can be seen that DRCTD3-TD reaches convergence at approximately 1000 episodes in the scenario-I, while DRCTD3-TD(None) reaches convergence at approximately 1200 episodes, indicating a significant improvement in learning speed. According to Fig. 9(a), it is essential to note that the number of episodic samples does not count the enhanced samples, ensuring the comparison's fairness. The results show that the number of samples of low-risk samples in DRCTD3-TD compared to DRCTD3-TD(None) is significantly higher, and finally, the number of samples in DRCTD3-TD is 273, while the number of samples in DRCTD3-TD(None) is 106. Based on the above analysis results, we observe that the performance of different risk-branching networks in DRCTD3-TD is also improved with the increase of high-quality samples, which may be due to the presence of diversified high-quality samples in the experience buffer of DRCTD3-TD during the training process, which gives the different risk branch networks a chance to sample more high-quality samples during the sampling process, thereby improving the overall performance.

We also run scenario-II five times for each algorithm and plot the learning curves in Fig. 8(d), (e), and (f). Similar results are shown in scenario-II. In this scenario, we can see that as the difficulty of the scenario increases, the algorithm decreases slightly in terms of performance and learning speed, DRCTD3-TD reaches convergence at approximately 1200 episodes, while DRCTD3-TD(None) reaches convergence at approximately 1300 episodes. Nevertheless, DRCTD3-TD still outperforms the DRCTD3-TD(None) algorithm in terms of performance for different risk-branching networks and the number of episodic samples in the low-risk experience buffer shows the general trends as in Fig. 9(a) which the number of samples for



(a) Episodic samples size curve in scenario-I.

(b) Episodic samples size curve in scenario-II.

Fig. 9. Episodic samples size curve.

DRCTD3-TD and DRCTD3-TD(None) is 232 and 82, respectively, which somewhat validates our previous conclusion that sampling more high-quality samples can help improve the performance of each risk network.

Additionally, we record and analyze several performance metrics, including the average returns, collision rates, survival distances (driving distances before collision in each episode) and finish rates (percentage of surviving distance in the total driving distance). The performance comparison corresponding to the evaluated episodes is shown in Table I. The experimental results demonstrate that DRCTD3-TD achieves superior performance to DRCTD3-TD(None) in two scenarios. Specifically, the average return when using DRCTD3-TD(None) and DRCTD3-TD are 430.76 and 435.25, respectively. The collision rate declines from 6.80% to 4.20% when using DRCTD3-TD(None) and DRCTD3-TD. The results of surviving distance and the finish rates show similar trends. The surviving distance increases from 385.64 to 387.82, and the finish rates increases from 95.22 to 95.76 when using DRCTD3-TD(None) and DRCTD3-TD.

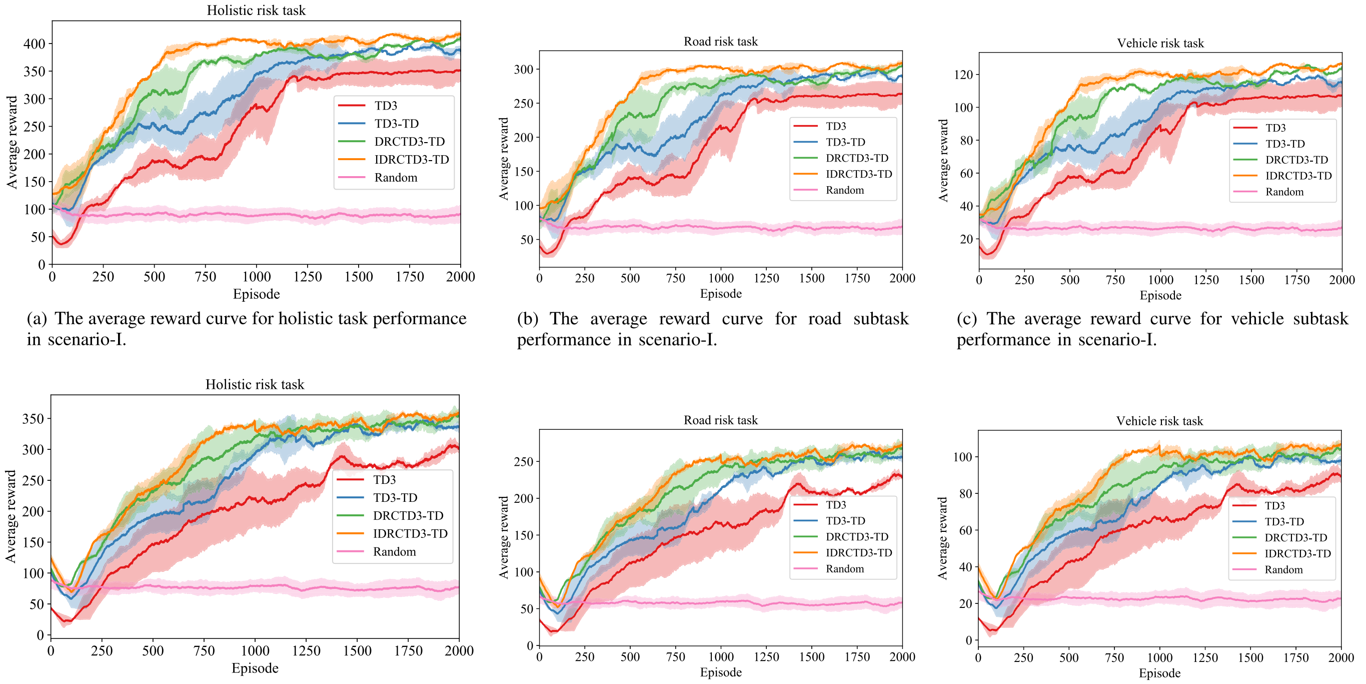
Similarly, we run Scenario II five times for two methods. The average return when using DRCTD3-TD and DRCTD3-TD(None) are 395.67 and 380.35, respectively. Correspondingly, the collision rate of the DRCTD3-TD declines from 12.40% to 8.80% when using DRCTD3-TD and DRCTD3-TD(None). In addition, the surviving distance increases from 370.76 to 376.54, and the finish rates increases from 91.55 to 92.97 when using DRCTD3-TD and DRCTD3-TD(None). In general, low-risk episodic sampling augmentation method expedites the training process, facilitates faster convergence, and leads to improved overall performance.

C. *Quantitative Results*

The IDRCTD3-TD algorithm is compared with different ablation algorithms, and the performance of the IDRCTD3-TD algorithm is quantitatively analyzed to demonstrate its effectiveness. Consequently, the hyperparameters of all algorithms are kept consistent, as shown in Table II.

- TD3: the vanilla TD3 algorithm, which served as the base framework for our algorithm.
- TD3-TD: TD3 algorithm with risk task decomposition, the algorithm uses different risk branches to learn different risk tasks.
- DRCTD3-TD: a low-risk episodic sampling augmentation method is used in TD3-TD.
- Random strategy: to reflect the difficulty of the experimental scenario [31].

The algorithm's performance is first compared when training in scenario-I and scenario-II are shown in Fig. 10, which shows the average reward curve over the last 100 episodes. Shaded areas show the standard deviation over five random seeds, from Fig. 10(a) can be observed that the TD3-TD algorithm with risk task decomposition shows the most significant performance and convergence speed compared to the vanilla TD3, which takes more time to learn the lane change strategy in the beginning phase. In contrast, the TD3-TD with risk task decomposition learns the corresponding risk task using different branches, allowing each branch to understand the individual risk task better and learn the minimized risk strategy for each branch. The TD3-TD considers the risk task in different branches during the Q-value update process, therefore learning the permutation strategy faster. The results show that the TD3-TD with risk task decomposition performs better when it comes to tackling the problem of unmanned vehicle decision making. Then, DRCTD3-TD adopts the low-risk episodic sampling augmentation mechanism, which provides more high-quality samples in the experience buffer and helps the algorithm samples diversified low-risk trajectories. By learning diversified high-quality samples, the algorithm converges faster and maintains a high reward relative to TD3-TD. The overall performance of the IDRCTD3-TD is significantly better than other comparative algorithms, and the reward curve converges faster than other algorithms, which shows that the intervention training strategy can further accelerate the algorithm's convergence. This is mainly reflected in beginning of the training phase, IDRCTD3-TD has yet to learn an effective strategy, the intervention controller can guide the reinforcement learning method to train effectively and



(a) The average reward curve for holistic task performance in scenario-I.

(b) The average reward curve for road subtask performance in scenario-I.

(c) The average reward curve for vehicle subtask performance in scenario-I.

(d) The average reward curve for holistic task performance in scenario-II.

(e) The average reward curve for road subtask performance in scenario-II.

(f) The average reward curve for vehicle subtask performance in scenario-II.

Fig. 10. Average reward curve over the last 100 episodes in scenario-I and scenario-II.

TABLE III
Performance Comparison of Different Algorithms Under Different Scenarios

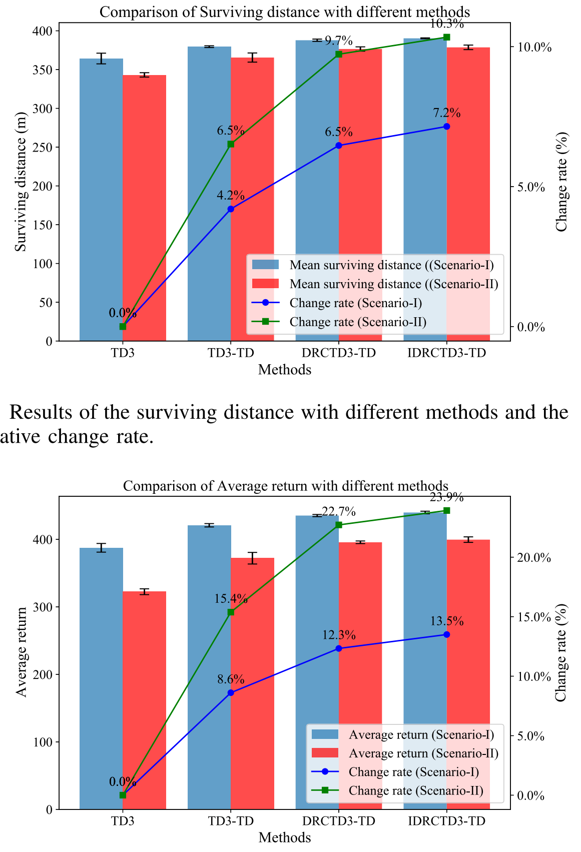
Scenario	Method	Collision rates(%)	Average return	△(%)	Surviving distance	△(%)	Finish rates(%)
Scenario-I	TD3	23.40±3.00	387.49±6.40	-	364.25±6.87	-	89.84
	TD3-TD	13.40±1.80	420.86±2.51	8.61 ↑	379.54±1.17	4.20 ↑	93.71
	DRCTD3-TD	4.20±0.83	435.25±1.58	12.33 ↑	387.82±1.46	6.47 ↑	95.76
	IDRCTD3-TD	**3.50±1.12**	**439.79±1.95**	**13.50 ↑**	**390.30±0.60**	**7.15 ↑**	**96.37**
Scenario-II	TD3	33.00±1.81	322.45±4.31	-	343.14±2.81	-	84.73
	TD3-TD	18.80±4.98	372.05±8.57	15.38 ↑	365.51±5.92	6.52 ↑	90.25
	DRCTD3-TD	8.80±3.29	395.67±1.92	22.71 ↑	376.54±2.78	9.73 ↑	92.97
	IDRCTD3-TD	**7.20±3.68**	**399.58±4.09**	**23.92 ↑**	**378.63±2.93**	**10.34 ↑**	**93.49**

Δ : The relative change rate for average return.

improve the algorithm's convergence speed. During the training process, IDRCTD3-TD can explore the favorable decision experience more easily and have the optimal strategy with the help of the intervention controller. Ultimately, IDRCTD3-TD speeds up convergence, and improves overall performance. To further illustrate the algorithm's performance, Fig. 10(b) and (c) show the risk task reward for each risk branch separately. It is found that the TD3-TD significantly outperforms the vanilla TD3 regarding reward in both risk branches while maintaining consistency with the overall reward curve. This finding shows that the risk task decomposition framework effectively improves the performance of the different risk branches, thereby enhancing the algorithm's overall performance. Accordingly, the DRCTD3-TD and the IDRCTD3-TD further improve the overall strategy by the low-risk episodic sampling augmentation and the intervening training strategy in dangerous environments, effectively improving the strategy for each risk branch.

We design a more challenging scenario-II to further demonstrate the algorithm's effectiveness. The average reward of the proposed method when training in scenario-II is shown in Fig. 10(d), (e), and (f), the general trends are similar to their performances in scenario-I, even though the increase in the scenario's difficulty leads to a slight decrease in the reward of the algorithms and the convergence rate.

In addition, our trained model is tested 100 times in five experiments conducted in two challenging scenarios. Table III and Fig. 11 show that the IDRCTD3-TD algorithm outperforms the performance of the other algorithms in two challenging scenarios. The IDRCTD3-TD algorithm, DRCTD3-TD algorithm and the TD3-TD algorithm have only average collision rates of 3.50%, 4.20% and 13.40%, compared to 23.40% for the vanilla TD3 algorithm, and the average returns of IDRCTD3-TD, DRCTD3-TD, and TD3-TD are 439.79, 435.25 and 420.86, respectively, the improvements of 13.50%, 12.33%



(a) Results of the surviving distance with different methods and the relative change rate.

(b) Results of the average return with different methods and the relative change rate.

Fig. 11. Results comparision with different methods.

and 8.61% compared to the vanilla TD3 algorithm. Similarly, the performance of our proposed methods on surviving distance shows similar trends, the surviving distance increase from 364.25 when using vanilla TD3, to 379.54, 387.82, and 390.30 when using the TD3-TD, DRCTD3-TD, and IDRCTD3-TD and the finish rates increases from 89.84% when using vanilla TD3, to 93.71%, 95.76%, and 96.37% when using the TD3-TD, DRCTD3-TD, and IDRCTD3-TD in scenario-I.

The task becomes more complex as the scenario complexity increases, leading to more visible effects in scenario-II. For instance, in contrast to the vanilla TD3 algorithm, the average collision rates decrease from 33.00% to 18.80%, 8.80%, and 7.20 when using the TD3-TD, DRCTD3-TD, and IDRCTD3-TD. Besides, TD3-TD, RCTD3-TD, and IDRCTD3-TD also exhibit significantly enhanced average returns, surviving distances, and finish rates compared to the vanilla TD3. The performance of the proposed algorithm is higher than that of the vanilla TD3 algorithm and can effectively improve safety, which is consistent with the results analyzed above.

Furthermore, Fig. 12 shows the results of the four algorithms in two different scenarios more visually. From Fig. 12(a) and (d), the effect of the risk task decomposition framework is prominent, as can be seen in the risk task decomposition framework outperforming TD3 with a large margin in terms of average return. Such results are also reflected in the average reward of specific subtasks. It is found that, especially for the road risk task, the performance improvement achieved by the algorithm of the risk task decomposition framework is evident, indicating that the agent can drive in the center of the lane and reduce collisions, maintaining safe driving. Correspondingly, performance improvements are achieved in each subtask based on the low-risk episodic sampling augmentation method and intervention training strategy.

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND TRADITIONAL METHOD

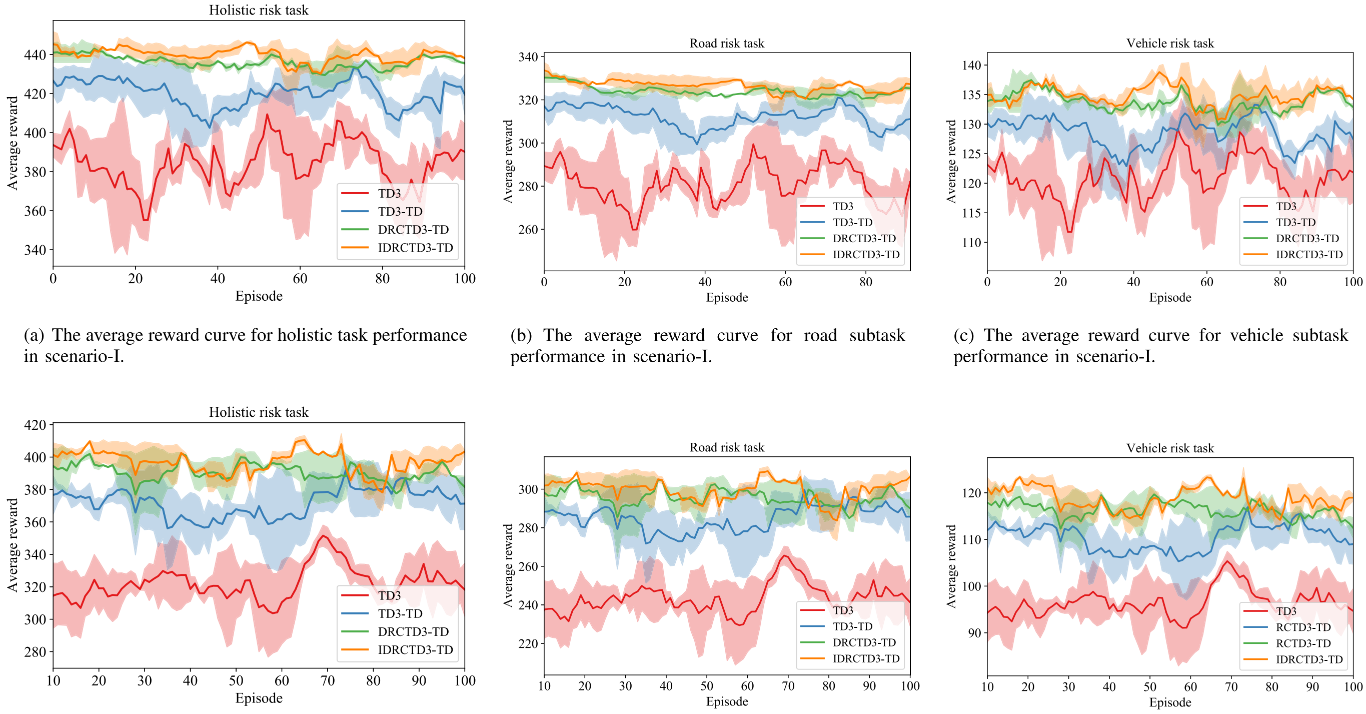
Scenario	**Performance**	APF	IDRCTD3-TD
scenario-I	Average return	428.70±4.38	439.79±1.95
	Collision rates(%)	13.40±4.11	3.40±1.12
	Surviving distance(m)	380.62±5.48	390.30±0.60
	Finish rates(%)	93.98	96.37
	Computing time(Max value)(ms)	0.10(0.19)	0.53(1.22)
scenario-II	Average return	305.61±3.73	399.58±4.09
	Collision rates(%)	74.60±3.78	7.20±3.68
	Surviving distance(m)	296.37±4.62	378.63±2.93
	Finish rates(%)	73.18	93.49
	Computing time(Max value)(ms)	0.10(0.19)	0.54(1.29)

D. Performance Comparsion Between the Proposed Method and Traditional Method

To illustrate the algorithm's effectiveness further, we compare the proposed algorithm with the APF algorithm, and the comparison results are shown in Table IV. In scenario-I, because of the relatively simple structure of the APF method, the computation time is only 0.1 ms within the decision period, compared to 0.53 ms for the proposed method. However, it is affected by the superposition of the potential field, which leads to the average collision rate of 13.40%, with the completion rate of 93.98%, which is still a gap compared to the proposed algorithm. Moreover, in Scenario-II, the fact that SVs have lane-changing behavior, which makes the scenario highly dynamic and stochastic, causing APF method failures and collisions in most cases, which shows that the APF method is poorly adapted to dynamic environments and is unable to address such scenario. In summary, combining the results above, the proposed algorithm still performs best and can satisfy the real-time requirements.

E. Analysis of Failure Scenarios

Although the proposed IDRCTD3-TD algorithm improves driving safety in different scenarios, there are still some failure cases, as shown in Fig. 13. By analyzing the failure cases, we find that failure case I is when HV performs obstacle avoidance in scenario-I. Because of the proximity of the SVs to the current and target lanes, the HV cannot react quickly enough to prevent a collision. Similarly, in failure case II, when the SV cuts in, the HV cannot react quickly enough to change lanes, primarily because only the HV's lateral behavior is considered in this study. Simultaneously, the adaptive braking behavior is also significant in avoiding collisions for the HV. Therefore, it is



(a) The average reward curve for holistic task performance in scenario-I.

(b) The average reward curve for road subtask performance in scenario-I.

(c) The average reward curve for vehicle subtask performance in scenario-I.

(d) The average reward curve for holistic task performance in scenario-II.

(e) The average reward curve for road subtask performance in scenario-II.

(f) The average reward curve for vehicle subtask performance in in scenario-II.

Fig. 12. Average reward curve for the trained models in scenario-I and scenario-II.

necessary to include longitudinal behavior to deal with similar scenarios in future research.

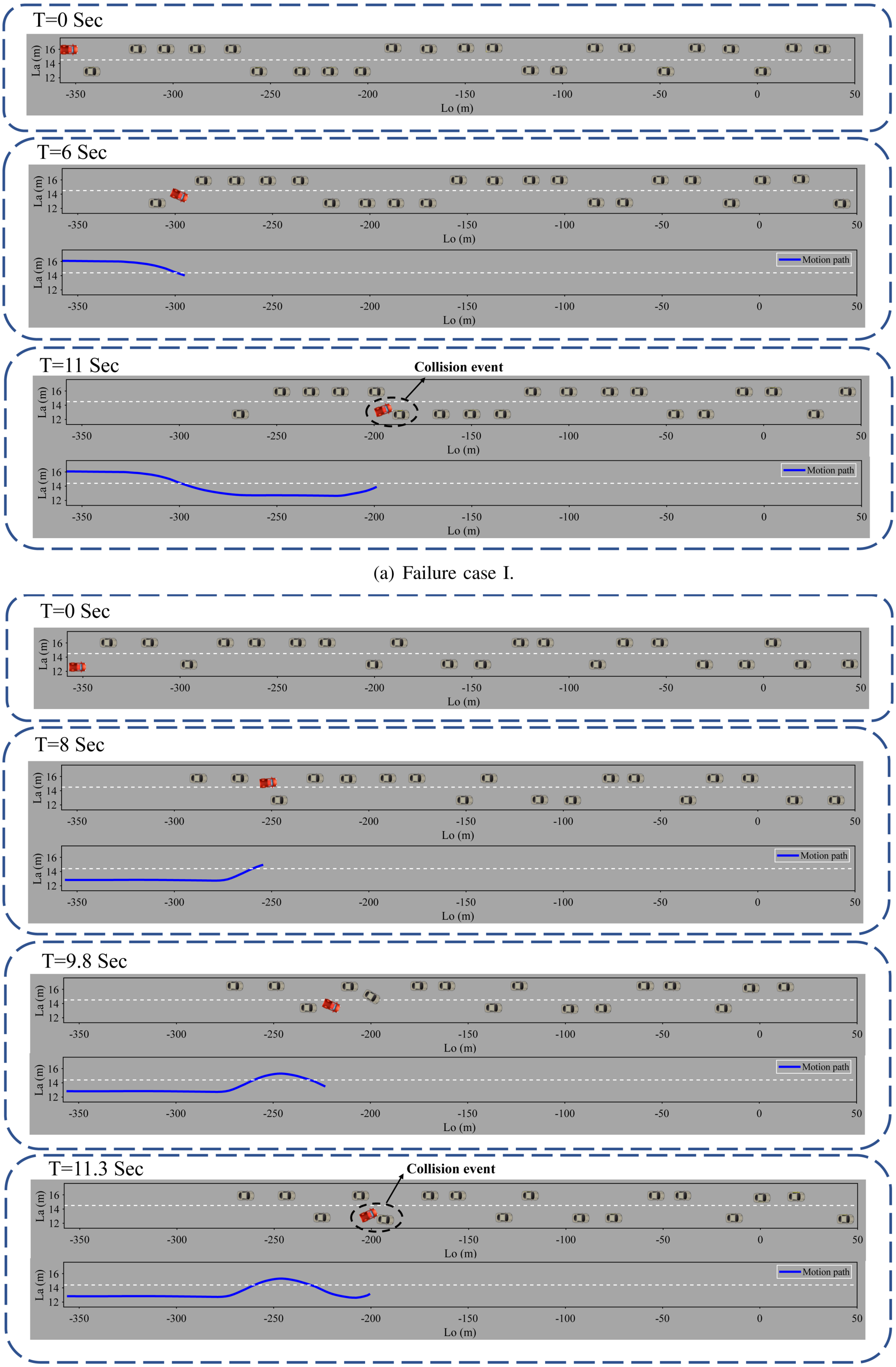
Another important reason leading to these failures is that the behavioral intention of the SVs is highly uncertain, and it is difficult to recognize the driving intention of the SVs just by risk potential fields combined with reinforcement learning, so it is not possible to make effective decisions to avoid collisions in advance. To improve the safety performance of the framework, it is necessary to incorporate the driving intentions of the SVs into the framework to predict the driving intentions.

The third causation of these failures may be that the HV could generate a more aggressive driving style caused by an imbalance in the setting of the reward function. Considering that the driving style affects the decision-making behavior of the vehicle, the future is dedicated to correcting for the effects of different driving styles on decision-making behavior.

V. Conclusion and Future Work

This study proposes a deep reinforcement learning decision-making framework for unmanned vehicles in lane change and overtaking scenarios. The framework addresses continuous lane-changing and overtaking decision-making problems to minimize driving risks and further reduces ineffective exploration, our framework avoids taking ineffective risky and dangerous behavior, resulting in faster and acceptable performance in an accurate simulation environment, which is critical for unmanned tasks with safety and efficiency as considerations. First, the risk potential functions are constructed based on road and vehicle risks. Then, the two categories of risk tasks are decomposed to learn the different risk branches separately to minimize the cumulative risk of each branch. Besides, a low-risk episodic sampling augmentation method is designed, using which the proposed framework can sample more low-risk samples to improve sample utilization. Furthermore, an intervention training strategy incorporating APF is designed to improve the performance of reinforcement learning. The experiment results with two difficulties scenarios demonstrate that the proposed algorithm not only improves the performance of different risk branches but also generates correct lane-changing decision-making with the proposed framework, thus reducing the driving risk and preventing HV collisions. Compared with the vanilla TD3 and the traditional method, the proposed algorithm shows more significant performance in collision rates and average returns and the proposed algorithm has more potential to be deployed in unmanned vehicles.

For future work, we focus on solving the problem of incorporating longitudinal behavior into the algorithm to guide the vehicle in a cooperative approach to improve performance further, and to apply the framework to scenarios outside of this study, the formulation of the MDP will need to be further adapted, or an MDP covering multiple scenarios will need to be constructed to accommodate new scenarios. In addition, the intentions prediction of SVs and consideration of driving style preferences will assist in improving the safety performance of the framework.



(a) Failure case I.

(b) Failure case II.

Fig. 13. Some failure cases of the proposed IDRCTD3-TD algorithm.

REFERENCES

[1] L. Chen, Y. He, Q. Wang, W. Pan, and Z. Ming, "Joint optimization of sensing, decision-making and motion-controlling for autonomous vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 4642–4654, May 2022.

[2] X. Xia et al., "An automated driving systems data acquisition and analytics platform," *Transp. Res. Part C: Emerg. Technol.*, vol. 151, 2023, Art. no. 104120.

[3] L. Zhang, R. Zhang, T. Wu, R. Weng, M. Han, and Y. Zhao, "Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5435–5444, Dec. 2021.

[4] F. Guo, S. G. Klauer, J. M. Hankey, and T. A. Dingus, "Near crashes as crash surrogate for naturalistic driving studies," *Transp. Res. Rec.*, vol. 2147, no. 1, pp. 66–74, 2010.

[5] X. Xia, R. Xu, and J. Ma, "Secure cooperative localization for connected automated vehicles based on consensus," *IEEE Sensors J.*, vol. 23, no. 20, pp. 25061–25074, Oct. 2023.

[6] L. Chen et al., "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1046–1056, Feb. 2023.

[7] Z. Zhang, L. Zhang, J. Deng, M. Wang, Z. Wang, and D. Cao, "An enabling trajectory planning scheme for lane change collision avoidance on highways," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 147–158, Jan. 2023.

[8] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D LiDAR," *IEEE Trans. Intell. Veh.*, vol. 8, no. 8, pp. 4069–4080, Aug. 2023.

[9] J. Lu, L. Han, Q. Wei, X. Wang, X. Dai, and F.-Y. Wang, "Event-triggered deep reinforcement learning using parallel control: A case study in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 2821–2831, Apr. 2023.

[10] X. Han et al., "Foundation intelligence for smart infrastructure services in transportation 5.0," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 39–47, Jan. 2024.

[11] B. Li, Y. Ouyang, L. Li, and Y. Zhang, "Autonomous driving on curvy roads without reliance on frenet frame: A cartesian-based trajectory planning method," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15729–15741, Sep. 2022.

[12] K. Tu, S. Yang, H. Zhang, and Z. Wang, "Hybrid A* based motion planning for autonomous vehicles in unstructured environment," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2019, pp. 1–4.

[13] R. Mashayekhi, M. Y. I. Idris, M. H. Anisi, and I. Ahmedy, "Hybrid RRT: A semi-dual-tree RRT-based motion planner," *IEEE Access*, vol. 8, pp. 18658–18668, 2020.

[14] A. Vinayak, M. A. Zakaria, K. Baarath, and A. P. A. Majeed, "A novel bezier curve control point search algorithm for autonomous navigation using N-order polynomial search with boundary conditions," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 3884–3889.

[15] W. Chi, C. Wang, J. Wang, and M. Q. H. Meng, "Risk-DTRRT-based optimal motion planning algorithm for mobile robots," *IEEE Trans. Automat. Sci. Eng.*, vol. 16, no. 3, pp. 1271–1288, Jul. 2019.

[16] P. Wang, S. Gao, L. Li, B. Sun, and S. Cheng, "Obstacle avoidance path planning design for autonomous driving vehicles based on an improved artificial potential field algorithm," *Energies*, vol. 12, no. 12, 2019, Art. no. 2342.

[17] P. Wu, F. Gao, and K. Li, "Humanlike decision and motion planning for expressway lane changing based on artificial potential field," *IEEE Access*, vol. 10, pp. 4359–4373, 2022.

[18] J. Wang, J. Wu, X. Zheng, D. Ni, and K. Li, "Driving safety field theory modeling and its application in pre-collision warning system," *Transp. Res. Part C: Emerg. Technol.*, vol. 72, pp. 306–324, 2016.

[19] Y. Li, W. Yang, X. Zhang, X. Kang, and M. Li, "Research on automatic driving trajectory planning and tracking control based on improvement of the artificial potential field method," *Sustainability*, vol. 14, no. 19, 2022, Art. no. 12131.

[20] Z. Qi, T. Wang, J. Chen, D. Narang, Y. Wang, and H. Yang, "Learning-based path planning and predictive control for autonomous vehicles with low-cost positioning," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1093–1104, Feb. 2023.

[21] C. Xi, T. Shi, Y. Wu, and L. Sun, "Efficient motion planning for automated lane change based on imitation learning and mixed-integer optimization," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–6.

[22] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical interpretable imitation learning for end-to-end autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 673–683, Jan. 2023.

[23] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 740–759, Feb. 2022.

[24] X. He, H. Yang, Z. Hu, and C. Lv, "Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 184–193, Jan. 2023.

[25] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," 2017, *arXiv:1708.05866*.

[26] S. Alighanbari and N. L. Azad, "Deep reinforcement learning with NMPC assistance NASH switching for urban autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2604–2615, Mar. 2023.

[27] K. Yang, X. Tang, S. Qiu, S. Jin, Z. Wei, and H. Wang, "Towards robust decision-making for autonomous driving on highway," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11251–11263, Sep. 2023.

[28] X. Tang, B. Huang, T. Liu, and X. Lin, "Highway decision-making and motion planning for autonomous driving via soft actor-critic," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 4706–4717, May 2022.

[29] X. Xiong, J. Wang, F. Zhang, and K. Li, "Combining deep reinforcement learning and safety based control for autonomous driving," 2016, *arXiv:1612.00147*.

[30] X. Huang, A. Jasour, M. Deyo, A. Hofmann, and B. C. Williams, "Hybrid risk-aware conditional planning with applications in autonomous vehicles," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 3608–3614.

[31] G. Li, Y. Yang, S. Li, X. Qu, N. Lyu, and S. E. Li, "Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness," *Transp. Res. Part C: Emerg. Technol.*, vol. 134, 2022, Art. no. 103452.

[32] M. A. Hebaish, A. Hussein, and A. El Mougy, "Towards safe and efficient modular path planning using twin delayed DDPG," in *Proc. IEEE 95th Veh. Technol. Conf.*, 2022, pp. 1–7.

[33] L. Wu, Z. Zhang, S. Haesaert, Z. Ma, and Z. Sun, "Risk-aware reward shaping of reinforcement learning agents for autonomous driving," in *Proc. IEEE 49th Annu. Conf. Ind. Electron. Soc.*, 2023, pp. 1–6.

[34] G. Li et al., "Lane change strategies for autonomous vehicles: A deep reinforcement learning approach based on transformer," *IEEE Trans. Intell. Veh.*, vol. 8, no. 3, pp. 2197–2211, Mar. 2023.

[35] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2015, *arXiv:1511.05952*.

[36] J. Wu, Z. Huang, W. Huang, and C. Lv, "Prioritized experience-based reinforcement learning with human guidance for autonomous driving," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 855–869, Jan. 2024.

[37] K. K. Tseng, H. Yang, H. Wang, K. L. Yung, and R. F. Y. Lin, "Autonomous driving for natural paths using an improved deep reinforcement learning algorithm," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 6, pp. 5118–5128, Dec. 2022.

[38] D. Yang, X. Qin, X. Xu, C. Li, and G. Wei, "Sample efficient reinforcement learning method via high efficient episodic memory," *IEEE Access*, vol. 8, pp. 129274–129284, 2020.

[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[40] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[41] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[42] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[43] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.

[44] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[45] R. B. Grando, J. C. de Jesus, V. A. Kich, A. H. Kolling, and P. L. J. DrewsJr., "Double critic deep reinforcement learning for mapless 3D navigation of unmanned aerial vehicles," *J. Intell. Robot. Syst.*, vol. 104, no. 2, 2022, Art. no. 29.

[46] V. R. F. Miranda, A. A. Neto, G. M. Freitas, and L. A. Mozelli, "Generalization in deep reinforcement learning for robotic navigation by reward shaping," *IEEE Trans. Ind. Electron.*, vol. 71, no. 6, pp. 6013–6020, Jun. 2024.

[47] Z. Chu, B. Sun, D. Zhu, M. Zhang, and C. Luo, "Motion control of unmanned underwater vehicles via deep imitation reinforcement learning algorithm," *IET Intell. Transport Syst.*, vol. 14, no. 7, pp. 764–774, 2020.

[48] C. Wei, Y. Li, Y. Ouyang, and Z. Ji, “Deep reinforcement learning with heuristic corrections for UGV navigation,” *J. Intell. Robot. Syst.*, vol. 109, no. 1, 2023, Art. no. 18.

[49] V. Mnih et al., “Playing Atari with deep reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst. Workshop.*, 2013, pp. 1–9.

[50] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 387–395.

[51] H. D. Nguyen, M. Choi, and K. Han, “Risk-informed decision-making and control strategies for autonomous vehicles in emergency situations,” *Accident Anal. Prevention*, vol. 193, 2023, Art. no. 107305.

[52] H. Liu, K. Chen, Y. Li, Z. Huang, J. Duan, and J. Ma, “Integrated behavior planning and motion control for autonomous vehicles with traffic rules compliance,” in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2023, pp. 1–7.

[53] M. Liebner, M. Baumann, F. Klanner, and C. Stiller, “Driver intent inference at urban intersections using the intelligent driver model,” in *Proc. IEEE Intell. Veh. Symp.*, 2012, pp. 1162–1167.

[54] J. Duan, S. Eben Li, Y. Guan, Q. Sun, and B. Cheng, “Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data,” *IET Intell. Transport Syst.*, vol. 14, no. 5, pp. 297–305, 2020.

[55] H. Van Seijen, M. Fatemi, J. Romoff, R. Laroche, T. Barnes, and J. Tsang, “Hybrid reward architecture for reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5392–5402.

[56] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. D. Velez, “Explainable reinforcement learning via reward decomposition,” in *Proc. IJCAI/ECAI Workshop Explainable Artif. Intell.*, 2019, pp. 47–53.

[57] J. Guan et al., “A discrete soft actor-critic decision-making strategy with sample filter for freeway autonomous driving,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2593–2598, Feb. 2023.

[58] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, “Reinforcement learning with augmented data,” in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 19884–19895.

[59] A. R. A. Van der Horst, “A time-based analysis of road user behaviour in normal and critical encounters,” Ph.D. dissertation, TU Delft, Delft, The Netherlands, 1990.

[60] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.

Sifan Wu is currently working toward the Ph.D. degree with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His research interests include autonomous driving, decision-making, reinforcement learning, artificial intelligence, and intelligent transportation systems.

Daxin Tian (Senior Member, IEEE) received the Ph.D. degree in computer application technology from Jilin University, Changchun, China, in 2007. He is currently a Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His research interests include intelligent transportation systems, autonomous connected vehicles, swarm intelligent, and mobile computing. He was the recipient of the Changjiang Scholars Program (Young Scholar) of Ministry of Education of China in 2017, National Science Fund for Distinguished Young Scholars in 2018, and Distinguished Young Investigator of China Frontiers of Engineering in 2018. He was the Technical Program Committee member/Chair/Co-Chair for several international conferences, including EAI 2018, ICTIS 2019, IEEE ICUS 2019, IEEE HMWC 2020, and GRAPH-HOC 2020.

Xuting Duan received the Ph.D. degree from the School of Transportation Science and Engineering, Beihang University, Beijing, China. He is currently an Associate Professor with the School of Transportation Science and Engineering, Beihang University. His research interests include connected and autonomous vehicles, vehicular ad-hoc networks, and vehicular localization.

Jianshan Zhou received the B.Sc., M.Sc., and Ph.D. degrees in traffic information engineering and control from Beihang University, Beijing, China, in 2013, 2016, and 2020, respectively. He is currently an Associate Professor with the School of Transportation Science and Engineering, Beihang University. From 2017 to 2018, he was a Visiting Research Fellow with the School of Informatics and Engineering, University of Sussex, Brighton, U.K. He is an author or coauthor of more than 30 international scientific publications. His research interests include the modeling and optimization of vehicular communication networks and air–ground cooperative networks, the analysis and control of connected autonomous vehicles, and intelligent transportation systems. He is/was the Technical Program Session Chair with the IEEE EDGE 2020, IEEE ICUS 2022, ICAUS 2022, a TPC member of the IEEE VTC2021-Fall track, and a Youth Editorial Board Member of the *Unmanned Systems Technology*. He was the recipient of the First Prize in the Science and Technology Award from the China Intelligent Transportation Systems Association in 2017, First Prize in the Innovation and Development Award from the China Association of Productivity Promotion Centers in 2020, and Excellent Doctoral Dissertation Award from Beihang University in 2021.

Dezong Zhao (Senior Member, IEEE) received the B.Eng. and M.S. degrees in control science and engineering from Shandong University, Jinan, China, in 2003 and 2006, respectively, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2010. He is currently a Senior Lecturer of autonomous systems with the School of Engineering, University of Glasgow, Glasgow, U.K. His research interests include connected and autonomous vehicles, machine learning, and control engineering.

Dongpu Cao (Senior Member, IEEE) received the Ph.D. degree from Concordia University, Montreal, QC, Canada, in 2008. He is currently the Canada Research Chair of driver cognition and automated driving and an Associate Professor and the Director of Waterloo Cognitive Autonomous Driving (CogDrive) Lab, University of Waterloo, Waterloo, ON, Canada. His research interests include driver cognition, automated driving, and cognitive autonomous driving. He has contributed more than 200 papers and three books. He was the recipient of the SAE Arch T. Colwell Merit Award in 2012, IEEE VTS 2020 Best Vehicular Electronics Paper Award, and three Best Paper Awards from the ASME and IEEE conferences. Prof. Cao is the Deputy Editor-in-Chief for *IET Intelligent Transport Systems Journal*, and an Associate Editor for IEEE Transactions on Vehicular Technology, IEEE Transactions on Intelligent Transportation Systems, IEEE/ASME Transactions on Mechatronics, IEEE Transactions on Industrial Electronics, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Computational Social Systems, and ASME Journal of Dynamic Systems, Measurement and Control. He was the Guest Editor of *Vehicle System Dynamics*, IEEE Transactions on SMC: Systems, and IEEE Internet of Things Journal.