2021 IEEE Intelligent Vehicles Symposium (IV)
July 11-17, 2021. Nagoya, Japan

Automatic Overtaking on Two-way Roads with Vehicle Interactions Based on Proximal Policy Optimization

Xiaochang Chen[1], Jieqiang Wei[2], Xiaoqiang Ren[1], Karl H. Johansson[3], Xiaofan Wang[1]

***Abstract*— Overtaking the lead vehicle on two-way roads in the presence of several oncoming vehicles is a complex task for autonomous vehicles. In this paper, we formulate the overtaking behavior of an ego vehicle based on a deep reinforcement learning (DRL) method. First, a two-way urban road is created, wherein the ego vehicle aims to reach the destination safely and efficiently while considering multiple traffic participants. We use different intelligent driver model (IDM) parameters to account for different drivers' habits. Furthermore, we introduce different responses of other vehicles when the ego vehicle takes overtaking maneuver. Then, a hierarchical control framework is proposed to manage vehicles on the road, which supervises vehicle behaviors at the high layer and controls the motion at the lower layer. The DRL method named Proximal Policy Optimization is applied to derive the high-level decision-making policies. A self-attention mechanism is further introduced to improve the performance of our algorithm. Finally, the overtaking maneuvers of the ego vehicle in different training timesteps are analyzed and how the responses of other vehicles affect the ego one's overtaking behavior is investigated. Simulation results show that our approach can achieve good performance to deal with the two-way road autonomous overtaking task. Supplementary video is available at `https://youtu.be/jPEGjM7cBuk`.**

I. INTRODUCTION

Autonomous driving (AD) has received unprecedented interest from the public since the DAPRA Grand Challenge [1]. In recent years, autonomous vehicles are very capable in localization [2], lane-keeping [3], obstacle avoidance, and braking [4]. Despite AD has made remarkable achievements, there are still many challenges underexplored. One of the big challenges is interacting with various vehicles on the road [5]. Autonomous vehicles tend to be overly defensive when encountering complex interactive scenarios (like multi-way intersection, double merge [6], overtaking, and other dense traffic), which is known as $the\ freezing\ robot\ problem$ [7]. Their conservative behaviors cause traffic jams and may furthermore bring accidents such as rear-end collisions [8]. In fact, the actions taken by an autonomous vehicle influence the behaviors of surrounding human-driver vehicles and vice-versa [9] [10]. In general, the ego vehicle can rely on some interactive behaviors with other vehicles on the road to achieve the desired task.

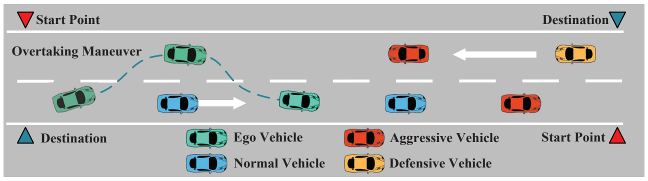


Fig. 1. The schematic diagram of the overtaking problem. We use green to represent the ego vehicle, yellow defensive vehicles, blue normal vehicles, and red aggressive vehicles.

We consider the following scenario which is depicted in Fig. 1. One ego vehicle is driving behind several lead vehicles in the same lane, while some oncoming vehicles are moving in the adjusted lane in the opposite direction. The ego vehicle must autonomously make decisions to stay within the origin lane or overtake the slow-moving lead vehicle. We set several driving styles for other vehicles to reproduce the different driving habits of human drivers. In addition, the overtaking maneuver operated by the ego vehicle not only affects the vehicles in the same lane but influences the actions of the oncoming vehicles as well. The surrounding vehicles will take different responses when the ego vehicle is overtaking. Notice that, the driver styles of other vehicles are unknown to the ego vehicle. Although overtaking maneuver in this scene is very dangerous due to lack of on-road gaps, limited visibility and non-cooperative behaviors from other drivers, etc. Safely overtaking the lead vehicle with several oncoming vehicles on the two-way road is necessary since this can be very critical in some emergencies. In summary, our problem can be succinctly stated as: *Successfully execute safe and efficient overtaking maneuvers on two-way roads with behavioral interactions.*

In this paper, we concentrate on the autonomous overtaking problem in the above scenario. We adopt a hierarchical control framework for overtaking decision strategy. The high-level manages the longitudinal and lateral behaviors of the vehicle, and the low-level focuses on governing vehicle velocity and acceleration. We utilize the ability of Deep Reinforcement Learning (DRL) to learn complex policies from data to implicitly obtain the joint behavioral interactions between the vehicles in the environment [7]. The Proximal Policy Optimization (PPO) algorithm is applied to learn the overtaking maneuver while considering the different reactions of other vehicles. A thought of self-attention is introduced to improve the performance of the algorithm. Our main contributions can be summarized as follows:

1: School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China, (`e-mails: cxchang, xqren, xfwang@shu.edu.cn`).

2: Ericsson AB, Stockholm, Sweden (`e-mail: jieqiang.wei@gmail.com`).

3: Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, (`e-mail: kallej@kth.se`).

The work of X. Chen, X. Ren and X. Wang is supported by National Key R&D Program of China No. 2018AAA0102804. The work of K. H. Johansson is supported in part by the Knut and Alice Wallenberg Foundation and the Swedish Research Council.

2021 IEEE Intelligent Vehicles Symposium (IV) | 978-1-7281-5394-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IV48863.2021.9575954

- We produce a challenging scenario for autonomous driving on a two-way road that other vehicles have different driving styles. To account for the interactions between vehicles, the intelligent driver model (IDM) is modified to include different responses from other human drivers to the ego vehicle's overtaking maneuver.
- We employ a policy gradient-based DRL approach named PPO with a self-attention mechanism to learn driving policies. In addition, a well-designed reward function is proposed to balance between velocity and security by overtaking maneuver effectively. The extensive simulations show that our method can successfully accomplish the driving task with the appropriate overtaking behaviors.

The rest of the paper is organized as follows: Section II provides an overview of related work. The detailed problem setup and the PPO algorithm with a self-attention layer are introduced in Section III. The performance of our algorithm is shown by several numerical examples in Section IV. After that, Section V concludes this paper.

II. RELATED WORK

Traditional planning and control methods may be difficult to apply in the two-way road scene with multiple traffic participants, due to the complexity of the environment and the uncertainty of other vehicles' behaviors. Classical proportion integral differential (PID) controller and linear quadratic regulator (LQR), struggle when dealing with complex tasks like overtaking behaviors [11]. A lane-changing fuzzy controller is designed in [12] to perform the overtaking maneuver, where some prior knowledge about the autonomous driving system is needed. Model predictive control (MPC) algorithms have been applied to the predictions of dynamic environments like lane changing [13], but those methods are computationally heavy during online implementation. In [14], authors present a stochastic control formulation to minimize the probability of collisions when overtaking in two-way roads. The authors in [15] implement offline and online phases to identify the intention of the lead vehicle to decide whether or not it is safe to overtake in a two-way road. These approaches are difficult to model interactions between vehicles in a highly dynamic environment, which may lead to unpredicted failures when encountering unknown states.

For these reasons, researchers try to find solutions using machine learning approaches. DRL shows the advantages of the adaptability to learning complex policies in high dimensional environments [16]. Compared with the existing literature that relies on machine learning approaches, this paper makes contributions two-fold. First, most of the existing literature only assumes that the other vehicles share the same driving style [11], [17], [18]. On the contrary, our paper utilizes different IDM parameters to reveal various driving styles. Second, notice that the interactions between the ego vehicle and the others are not considered in the aforementioned literature. The cooperative behaviors are taken into account in [19], [20]. However, the non-cooperative or conflict behaviors are not considered between the ego vehicle and others. In order to show these various behaviors between vehicles, we apply different responses of other vehicles to the overtaking maneuver of the ego vehicle. Of late, attention-based DRL methods to make decisions in complex intersection scene has raised. The authors in [21] use Deep Q-Network (DQN) with social attention to make the optimal sequence decisions successfully in the interactions scene. In [22], spatial attention and temporal attention with the Deep Deterministic Policy Gradient (DDPG) framework are applied to make lane change behaviors. In our study, the PPO algorithm is applied to learn appropriate overtaking maneuvers from scratch. We follow a similar trend of learning combined with an attention mechanism to deal with the problem of overtaking on the two-way road while considering the cooperative or non-cooperative behaviors from the other vehicles to the overtaking behavior of the ego vehicle.

III. METHODOLOGY

We intend to teach autonomous vehicles to make proper high-level decisions on a two-way road (Fig. 1) by coping with a continuous state space and a discrete action space. The ego vehicle needs to utilize the opposite lane to overtake a slow-moving lead vehicle and make a trade-off between safety or effectiveness.

A. Hierarchical control framework

Autonomous driving should consider the diverse behaviors of other vehicles on the road. The movement of all vehicles in the scene are mastered by a hierarchical control framework, which is shown in Fig. 2. A perception module processes sensor data and provides the most relevant information about the environment. The high-level decision controller utilizes this information and generates control commands which will be executed by the low-level controllers. The low-level controller manages the longitudinal and yaw acceleration of the vehicle. Output values from the low-level controllers are then taken by vehicle actuators for vehicle movement. We suppose these modules are coupled and work collaboratively, thus we focus on the high-level decision-making controller in our study. The details of the low-level controllers can be obtained from [23].

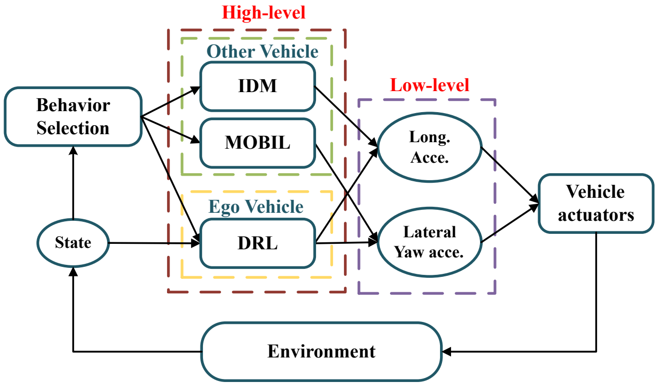


Fig. 2. The hierarchical control structure for vehicles.

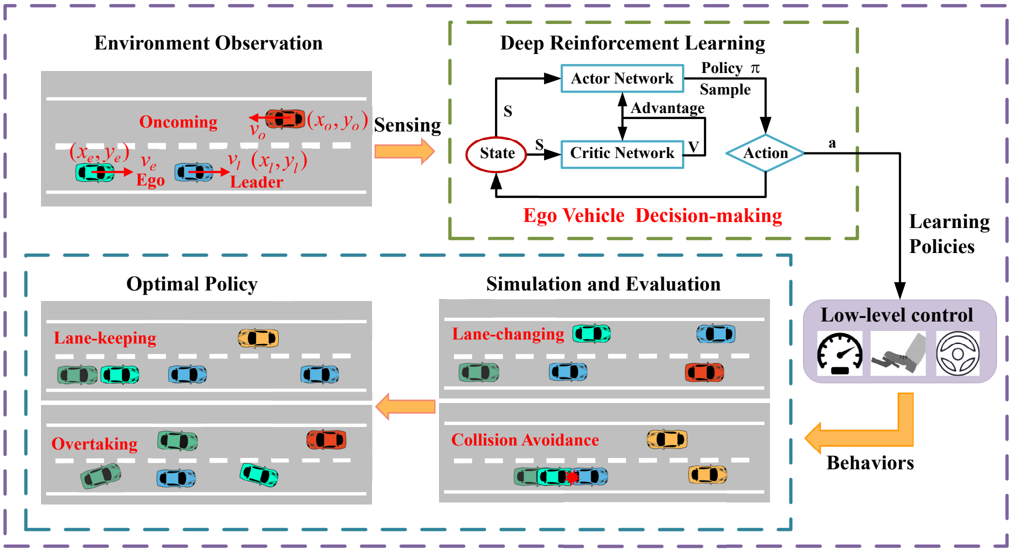


Fig. 3. Illustration of the DRL-based decision-making process for autonomous vehicles.

The process of learning the decision and control of the ego vehicle is illustrated in Fig. 3. The ego vehicle is driving in a two-way road environment and aiming to run through a particular driving scenario. First, the DRL method PPO is derived for the ego vehicle to learn the high-level decision-making policies. Then, the policy output is executed by the low-level controller to act out different driving behaviors.

Other vehicles in our driving scene follow realistic behaviors that dictate how they accelerate and steer on the road. IDM [24] and Minimizing Overall Braking Induced by Lane Change (MOBIL) [25] for lane-keeping and lane-changing, respectively. IDM describes the change of vehicles from free flow to congested flow in a unified way. The velocity difference between adjacent vehicles and the distance between vehicles are considered.

B. Driver Behaviors

From the naturalistic human drivers' study [26], how a driver reacts to the stimulus from surrounding vehicles can be affected by the individual experience and history of driving. Based on the fact that different drivers may behave differently in identical traffic conditions and the variables of vehicle models are directly related to driving behaviors, we determine the parameters of IDM and MOBIL for three classes of drivers to represent “defensive”, “normal", and “aggressive" driving styles. The corresponding parameters in our simulation are listed in Table I.

In addition, the behaviors of the ego vehicle can affect other drivers' behaviors and result in complex interactions. We make a few modifications of IDM to include what other vehicles will do in response to the overtaking maneuver of the ego vehicle that the defensive vehicles brake to facilitate overtaking, the aggressive vehicles accelerate to prevent overtaking and the normal vehicles take no action. The details of the interactive behaviors in our simulations are described as follows:

- ***Aggressive vehicle.*** Assume the direct leader with an aggressive driver. When it finds that the following ego vehicle within the rear $[50, 100]m$ has a steering angle about $5°$, the leader will execute an acceleration of $2m/s^2$ to increase the distance from the rear and urge the ego vehicle to give up overtaking. When the ego vehicle with a steering angle exceeds $10°$ within $50m$ behind the leader, it can be considered that the ego vehicle is executing overtake maneuver. The aggressive vehicle will produce an acceleration of $3m/s^2$ to make it more difficult for the ego vehicle to overtake. If the aggressive vehicle is an oncoming vehicle driving in the opposite lane, when it finds that the steering angle of the ego vehicle exceeds $5°$ within $[150, 250]m$ or $[80, 150]m$ ahead, it can be judged that the ego vehicle has overtaking intention. Therefore, the oncoming aggressive vehicle will generate accelerations of $2m/s^2$ and $3m/s^2$ respectively to prevent the ego vehicle from overtaking. When the steering angle of the ego vehicle exceeds $10°$ in the front $[40, 80]m$ horizon, the oncoming aggressive vehicle will produce an acceleration of $1m/s^2$, which will make the overtaking condition of the ego vehicle more severe.
- ***Defensive vehicle.*** Cautious drivers drive defensively and are more likely to produce cooperative behaviors when interacting with other vehicles. We assume there will be some behavioral changes for a defensive driver as well. If the defensive vehicle is the direct lead vehicle, the acceleration of $-2m/s^2$ or $-3m/s^2$ will be generated respectively when the steering angle of the ego vehicle exceeds $5°$ or $10°$ within the range of $[50, 100]m$ or $[0, 50]m$. It means the cautious driver will slow down and create an opportunity for cooperation. When it is an oncoming vehicle, the acceleration will be $-2m/s^2$ or $-3m/s^2$ if the ego vehicle's turning angle exceeds $5°$ in the range of $[150, 250]m$ or $[80, 150]m$.

When the ego vehicle's turning angle exceeds 10° in the range of $[0, 80]m$, the acceleration will be $-4m/s^2$, leaving a longer safety distance for the ego vehicle to overtake.

- ***Normal vehicle.*** The behaviors change of this driving style vehicles is based on the relevant parameters of IDM, we do not set additional rules.

Notice that, our modification of IDM only focuses on the steering behavior of the ego vehicle, and the principal behavior changes depend on the relevant parameters of IDM. In our simulations, vehicles with different driving styles are randomly generated, which increases the complexity of the traffic scene. What's more, we assume that when the ego vehicle gives up overtaking, the other vehicles will gradually recover to their own normal states. In addition, for the sake of more intuitively reflecting the overtaking maneuver of the ego vehicle, we have banned the automatic lane-changing maneuvers of other vehicles.

C. Vehicle Model

To reflect the real vehicle characteristics and simplify the movement of the vehicle, a modified kinematic bicycle model [24] is assumed for the vehicles on the road.

D. Reinforcement Learning

In this paper, we use a policy gradient based RL algorithm. The policy gradient algorithms usually maximize the following objective via gradient ascent:

$$\bigtriangledown_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta(\tau)}[\sum_{t=1}^{T} \bigtriangledown_\theta log\pi_\theta(a_t|s_t) \sum_{t=1}^{T} r(a_t, s_t)], \quad (1)$$

where τ is a trajectory, $\pi_\theta(\tau)$ is the likelihood to execute the trajectory under the current policy π_θ. The formula $\pi_\theta(a_t|s_t)$ is the probability to select action a_t at state s_t, and $r(a_t, s_t)$ is the reward value after executing that action.

1) Proximal Policy Optimization: As one of the most popular reinforcement learning algorithms based on policy gradient, PPO has the ability to accomplish complex discrete or continuous control tasks. To simplify the problem and improve learning efficiency, we use PPO with discrete action space to deal with high-level decision-making tasks. In PPO, a clipped surrogate objective or actor has the following form:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(\rho_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)],$$

where $\rho_t(\theta) = \pi_\theta(a_t|s_t)/(\pi_{\theta_{old}}(a_t|s_t))$ is the probability ratio under the new and old policy parameter. The expression $\hat{A}_t$ is an advantage estimator within timesteps T. The formula $clip(\cdot)$ is a clip function and ϵ a hyperparameter. The loss function of the critic has the following form:

$$L(\phi) = -\sum_{t=1}^{T}(\sum_{t'>t} \gamma^{\ t'-t} r_{t'} - V_\phi(s_t))^2 = -\sum_{t=1}^{T}(\hat{A}_t)^2, \quad (2)$$

where the actor aims to maximize $L^{CLIP}(\theta)$ and the goal of critic is to minimize $\hat{A}_t$. The actor will adopt the new policy based on the old policy according to advantage estimates $\hat{A}_t$. The clip item prevents excessive policy updates.

TABLE I: Parameters of different driving styles with IDM and MOBIL in simulations.

IDM Parameters	Normal	Defensive	Aggressive
Desired velocity v_0 (m/s)	18	15	21
Desired time gap T (s)	1.5	2.0	1.0
Desired jam distance s_0 (m)	10	15	5
Maximum acceleration a (m/s^2)	3	2	4
Maximum deceleration b (m/s^2)	-4	-2	-6
Acceleration exponent δ	4	4	4
MOBIL Parameters	Normal	Defensive	Aggressive
Politeness factor p	0.5	1.0	0.0
Acceleration threshold a_{th} (m/s^2)	0.1	0.2	0.0
Safe braking $b_{safe}(m/s^2)$	2.0	1.0	3.0

2) Attention mechanism: In human driving, the position, velocity, and heading of surrounding vehicles have different contributions to the decision-making of the ego vehicle. Therefore, the ego vehicle should pay different attention to surrounding vehicles that are related to the driving tasks. We add a self-attention structure in the critic network to model this form. The scaled dot-product attention mechanism used in this paper can be written as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V, \quad (3)$$

where the details of the attention mechanism are illustrated in Fig. 4. The process of calculating attention is to use a Q (query), calculate its similarity with each K (key), and sum the calculated results with all V (value) weighted. The features of scene observation are split into the ego vehicle features and the other features. And each of both is embedded with Multilayer Perceptron (MLP). The query Q is computed with a linear layer where the features are only from the ego-embedding. The key K and value V are computed with a single linear layer, of which the inputs are combined with ego-embedding and other-embedding. The similarities between query Q and key K are assessed by scaled dot-product $(QK^T)/(\sqrt{d_k})$. Normalizing these similarities with a softmax function, we obtain the attention matrix. The final attention output is described as formula (3). We use two heads of self-attention to capture the dependencies between the ego vehicle and the others. The attention outputs, after a linear layer, are then combined with the ego embedding as in residual networks.

3) Neutral Network structure: We employ an actor-critic structure network that is trained using the PPO algorithm. The detailed architecture of our network is shown in Fig. 4. Some hyperparameters of our algorithm are shown as: $\gamma = 0.92$, $learning\ rate = 5 \times 10^{-5}$, $\epsilon = 0.2$, $\lambda = 0.85$.

E. The Ego Vehicle Observations and Actions

Observations. For autonomous vehicles, we can obtain numerous environmental information through the developed perception system, some of which may not be directly related to the decision-making task. Therefore, we only extract the most relevant state representation information, in which the state variables can be described by the continuous positions, orientation angles, and velocity of the vehicle. We set an

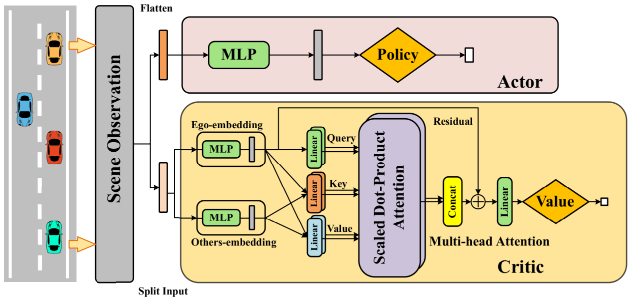


Fig. 4. Network architecture. ***MLP***: 2 layers, ReLU activations, each layer output is 1×256. ***Policy***: 1 layer for categorical distribution and layer output is 1×256. ***Value***: 3 layers, [ReLU, ReLU, Linear] activations, the first two layers output are 1×256. ***Linear***: 1 layer, Linear activations, each layer output is 1×256.

$V \times F$ array to represent a series of V vehicles with characteristic size F. Then, the joint states of road vehicles which we call *kinematics observation* are shown as :

$$s = (p_i, x_i, y_i, v_i^x, v_i^y, cos\varphi_i, sin\varphi_i)_{i \in [0,N]}, \tag{4}$$

where s_0 is the ego vehicle's state and $s_i, 1 \leq i \leq N$ are states of the other N vehicles. Each item of s_i is described as follows. Feature p_i represents whether the surrounding i-th vehicle can be observed by the ego vehicle, and p_0 is set as 1 for all the time. The features (x_i, y_i), (v_i^x, v_i^y), and $(cos\varphi_i, sin\varphi_i)$ are the global position, velocity, and heading of each vehicle. In our simulations, we transform all absolute features of the others into features relative to the ego vehicle.

Several other types of observations like occupancy-grid observation [19], [21] or image-input observation [22] are also used in autonomous driving scene. Compared with other observations, *kinematics observation* has less dimension. In addition, the features measured by the kinematics observation method can be obtained directly by low-cost sensors, which have high robustness in a harsh environment.

Actions. The ego vehicle makes high-level decision-making behaviors in discrete action space. For longitudinal cruise control, accelerate or brake behaviors can be chosen from a finite set of actions $A = \{Slower, Keep\ lane, Faster\}$. The lateral actions are listed in $B = \{Left\ lane-change, Keep\ lane, Right\ lane-change\}$.

F. Reward Function

The design of the reward function is of great significance to guide the agent to learn the driving policies. The reward function R is designed to balance safety, efficiency, and the completion of driving tasks. Specifically, our reward function contains several terms to encourage different driving behaviors. The ego vehicle obtains a penalty $R_{collision}$ when collisions have occurred. Positive rewards R_r and R_o are used to encourage the ego vehicle to keep the lane and drive in the opposite lane, respectively. Relatively high velocity can improve traffic efficiency, thus the ego vehicle will obtain the maximize velocity rewards $R_{velocity}$ when it drives at full velocity, which will linearly map to zero for minimum velocity. We encourage the ego vehicle to overtake by giving the reward $R_{overtaking}$ if it overtakes the lead vehicle successfully. The more vehicles that the ego vehicle overtakes, the better reward obtained. Besides, when the ego vehicle successfully arrives at the destination without collisions, we give a terminal reward $R_{terminal}$. The instantaneous reward R_{all} in autonomous driving can be expressed as follows:

TABLE II: Terms in the reward function.

Reward Term	Reward function
Collision penalty	$R_{collision} = \gamma_1 R_c, \quad R_c \in \{0,1\}$
Lane keeping bonus	$R_{lane} = \begin{cases} \gamma_2 R_r, & R_r \in \{0,1\} \\ \gamma_3 R_o, & R_o \in \{0,1\} \end{cases}$
Velocity bonus	$R_{velocity} = \gamma_4 \frac{v_{ego} - v_{min}}{v_{max} - v_{min}}$
Overtaking bonus	$R_{overtaking} = \gamma_5 (L - n_{front})$
Terminal bonus	$R_{terminal} = \gamma_6 R_t, \quad R_t \in \{0,1\}$

$$R_{all} = R_{collision} + R_{lane} + R_{velocity} + R_{overtaking} + R_{terminal}.$$

The specific reward functions are shown in Table II. Here L and n_{front} denote the total number of vehicles on the right lane and the number of real-time vehicles drives in front of the ego vehicle in the same direction at a given episode, respectively. The parameter γ_i is the weight of each term. Then we normalize the reward values to the range (0,1) and search for different weighting parameters to find the combination that generates the best result. In our simulations, the detail coefficients of each term are $\gamma_1 = -1.5$, $\gamma_2 = 0.21$, $\gamma_3 = 0.42$, $\gamma_4 = 1.6$, $\gamma_5 = 0.2$, $\gamma_6 = 0.2$.

IV. SIMULATION RESULTS

A. Simulation Environment

We train the ego vehicle in a two-way road scenario as shown in Fig. 5. It is implemented by using an open highway traffic simulation framework *highway-env* [27]. Without loss of generality, the parameters of our environment are settled as follows. We create a two-way road with a length of $1000m$. There are 3 lead vehicles in front of the ego vehicle and 4 oncoming vehicles in the opposite lane. The driving style of each vehicle is randomly generated, with a proportion of $1/3$. The initial velocity of the ego vehicle is randomly selected from the range $[25, 30]m/s$ and its maximum velocity is $30m/s$. All vehicles have the length = $5m$ and width = $2m$. The parameters setting of other vehicles are shown in Section III. The simulation frequency is $15Hz$ and the duration time of one episode is $38s$. Each vehicle has a random initial position on the road. For longitudinal acceleration and lateral steering angle, we allow values in the range $[-6, 6]m/s^2$ and $[-\pi/4, \pi/4]rad$. The episode is terminated either when the ego vehicle collides with another vehicle or reaches the destination, or when the time duration exceeds a preset threshold.

B. Performance Evaluation

In our experiments, the performance of our algorithm is compared with the other two benchmark algorithms, namely the PPO baseline (with the same network structure yet no attention layer) and DQN. All these algorithms have

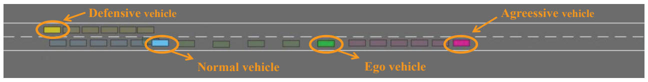
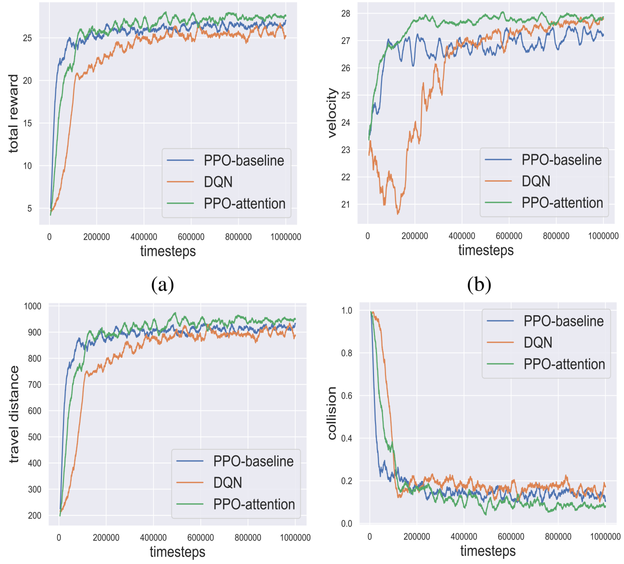


Fig. 5. Bird view of the overtaking scenario in *highway-env*. The ego vehicle is marked with a green rectangle and its trajectory is shown behind. The aggressive vehicles are shown in purple, the normal vehicles in blue, and the defensive vehicles in yellow.

been trained a million timesteps. The following metrics are utilized to quantitatively evaluate the performance of the three methods, and then the results are plotted in Fig. 6.

- ***Total accumulated rewards.*** The total reward represents the cumulative return along the vehicle trajectory and can be used to evaluate the quality of the policy.
- ***Average velocity.*** This is the average velocity of the ego vehicle in one episode.
- ***Collision-free travel distance.*** The maximum driving distance of the ego vehicle when it is collision-free or reaching the destination.
- ***Collision rate.*** The collision rate is the percentage of collisions in the last 600 episodes executed in the environment.



(a) (b)

(c) (d)

Fig. 6. Performance of each metric during training. (a) Total accumulated rewards (higher the better), (b) Average velocity (sooner the better), (c) Collision-free travel distance (longer the better), (d) Collision rate (lower the better). We show the mean learning curve using an average of over 20 random seeds with a 95% confidence interval.

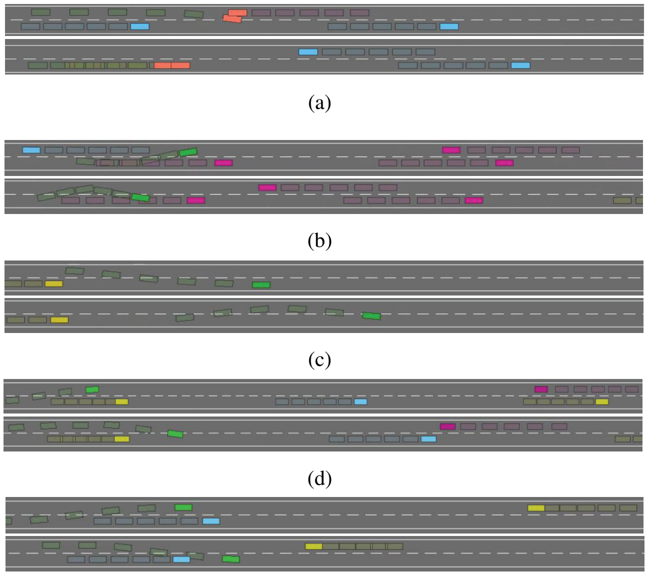
It is obvious in Fig. 6 that the DQN method has a slower convergence rate compared with the PPO based methods (i.e., ours and the PPO baseline). To better illustrate the performance of the three methods, we present the averaged ones of the last 5000 episodes (around 2×10^5 timesteps) in Table III. One sees that our method has the largest total reward, fastest velocity, longest distance, and lowest collision rate. Though the DQN method only has a slightly lower velocity than our method, its velocity curve oscillates much wider than ours, as depicted in Fig.6 (b).

TABLE III: The average performance of the three methods in training.

Metrics	Our method	PPO baseline	DQN
Average total reward	27.41	26.49	25.74
Average velocity (m/s)	27.84	27.26	27.67
Average distance (m)	944.13	916.81	902.51
Average collision rate	8.1%	13.6%	16.7%

C. Driving performance during training

To analyze the learning process at different timesteps of our method, we save a series of checkpoints during training. Fig. 7 shows the visualization of some typical vehicle behaviors on different timesteps at some saved checkpoints.



(a)

(b)

(c)

(d)

(e)

Fig. 7. Scenario shots of representative training cases. (a) Collision with another vehicle, (b) The behavior of failing to overtake and returning to the original lane, (c) Swinging driving behavior of the ego vehicle, (d) Overtaking with an aggressive oncoming vehicle, (e) Overtaking with a defensive oncoming vehicle.

Fig. 7 (a) shows the scenarios with typical vehicle collision behaviors. Most collisions occur when the ego vehicle is trying to overtake or being forced to change a lane. The ego vehicle has a rear-end collision with a direct lead vehicle (the bottom one of Fig.7 (a)) or collision with an oncoming vehicle at the opposite lane (the top one of Fig.7 (a)). One case is the high-velocity ego vehicle fails to follow the leader after overtaking since the ego vehicle does not learn to brake ahead of time. The reason is that the action space is explored randomly and the ego vehicle does not know what good actions should be taken.

Fig. 7 (b) displays an unsuccessful overtaking maneuver of the ego vehicle. Since the oncoming vehicle is close to the ego vehicle, there is not enough safe overtaking gap in the opposite lane. After trying to steer, the ego vehicle gives up overtaking and drives back to the original lane with the behavior of braking. This behavior depicts that in the process of training, the learning ego vehicle gradually behaves like

TABLE IV: Performance comparison during test.

Algorithms	Average total reward	Success rate
Our method	28.26	93.3%
PPO baseline	27.35	88.5%
DQN	25.82	84.3%

a novice driver.

Fig. 7 (c) indicates the swing driving behaviors of the ego vehicle on the road after passing all vehicles safely. Whereas the ego vehicle has learned overtaking behavior, it does not learn the best maneuver when driving close to the destination.

Fig. 7 (d) demonstrates a safe overtaking. The aggressive vehicle accelerates when the ego vehicle prepares to overtake, which will shorten the safe overtaking time and may force the ego vehicle to give up overtaking.

Fig. 7 (e) reveals another successful overtaking scene. The defensive vehicle brakes earlier and leave more safety distance for the overtaking maneuver of the ego vehicle. It shows that the cooperative behavior of the defensive vehicle contributes to the successful overtaking of the ego vehicle.

Furthermore, we use the trained model to run 1000 timesteps in the two-way road to test the adaptability of our method. The most concerning measures are the average reward and success rate. As shown in Table IV, although all methods can learn overtaking behavior, our method has a higher total reward and success rate than other two methods.

TABLE V: ***Experiment 1.*** Other drivers on the two-way road do not exhibit responses to the overtaking maneuver of the ego vehicle.

Metrics	Our method	PPO baseline	DQN
Average total reward	28.01	27.24	26.15
Average velocity (m/s)	27.34	26.41	27.33
Average distance (m)	953.83	934.50	921.22
Average collision rate	7.2%	8.6%	12.1%

D. *Comparative Experiments*

To further demonstrate the influence of different driving styles of other vehicles and their reactions to the overtaking maneuver of the ego vehicle, several experiments are conducted. Tables V and VI contain results from experiments 1 and 2 respectively. All the performances are the averaged one of the last 5000 episodes during the training. We then test the adaptability of our trained model in experiment 3.

Experiment 1. We compare the performance of different methods when the other vehicles will not react to the overtaking maneuver of the ego vehicle. Due to the complexity reduction of the dynamic scenario, the ego vehicle only needs to regard other vehicles as moving obstacles with different velocities. As depicted in Table V, our method arguably has the best performance among all benchmarks when all three methods perform better than in the interactive scenarios (see Table III).

TABLE VI: ***Experiment 2.*** The performance of our method during training in three different scenarios, where the proportion of aggressive drivers on the road is 0 (Defensive), 1 (Aggressive), $1/3$ (Mixed).

Metrics	Defensive	Mixed	Aggressive
Average total reward	28.90	27.41	25.79
Average velocity (m/s)	29.73	27.84	26.92
Average distance (m)	975.43	944.13	909.62
Average collision rate	3.5%	8.1%	11.7%

TABLE VII: ***Experiment 3.*** The generalization ability of our training model in similar but different environments, where the proportion of aggressive drivers on the road is 0 (Defensive), 1 (Aggressive), $1/3$ (Mixed).

Metrics	Without Responses			With Responses	
	Agg.	Mixed	Def.	Agg.	Def.
Average total reward	27.94	28.04	28.34	25.54	28.67
Success rate(%)	91.3%	92.8%	93.5%	87.9%	95.0%

Experiment 2. We create three driving scenarios, in which the other drivers are all aggressive style or defensive style, or generate random driving style by $1/3$. It can be seen in Table VI that the braking maneuver (cooperative behavior) of the defensive drivers is conducive to the overtaking maneuver of the ego vehicle. The ego vehicle can safely overtake at a high velocity and have a high total reward, thus it performs more aggressively. On the contrary, when the other drivers are all aggressive, safe overtaking behavior is the most challenging for the ego vehicle. Due to the acceleration maneuver (non-cooperative behavior) of the aggressive drivers, the ego vehicle tends to adopt more conservative policies (with low total reward and slow velocity) and has a poor success rate when overtaking. Generally, the different reactions of the other vehicles will influence the overtaking maneuver of the ego vehicle significantly. The cooperative behaviors between traffic participants benefit to safe driving on the road.

Experiment 3. In order to certify the robustness and generalization capability of our method, we use the model trained by our method to test in similar scenarios, where the other vehicles behave differently. Notice that our algorithm is trained when the aggressive driver is randomly selected with a probability of $1/3$. We perform 1000 timesteps tests on each setting for the other vehicles on the road. Experiment results are shown in Table VII (The test results of mixed scenario considering the response of the other vehicles are listed in Table IV). Our model has good performance in various scenarios. In the scenario where other vehicles are all defensive and interactive, the ego vehicle has the best performance. On the contrary, when the other vehicles are all aggressive and interactive, the performance is worst, whereas the success rate is still acceptable. In general, the performance of our model decreases slightly with the increasing complexity of the environment.

V. CONCLUSIONS & FUTURE WORK

We present a two-way road scenario with multiple traffic participants for overtaking tasks. We consider the different driving styles of the other vehicles and their reactions to the overtaking behavior of the ego vehicle. A DRL technique named PPO with self-attention is used to deal with the decision-making problem and the resulting policies successfully learn overtaking maneuver in this scenario. The performance of our algorithm is illustrated by numerous simulations. We suppose that the combination of attention mechanism improves the efficiency and safety of decision-making, which makes the behaviors of the ego vehicle more human-like.

In this paper, only the interactions between different vehicles are concerned. When driving in a city-like situation, we need to consider traffic lights, pedestrians, and road signals, which is a more challenging problem worthy of further study. Furthermore, due to the limited control accuracy of the discrete action space, we cannot guarantee collision-free during training and execution. This situation will be harmful to real-world driving systems, thus our future work includes safe reinforcement learning methods.

REFERENCES

[1] M. Buehler, K. Iagnemma, and S. Singh, *The 2005 DARPA grand challenge: the great robot race*. Springer, 2007, vol. 36.

[2] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[3] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 2017, no. 19, pp. 70–76, 2017.

[4] P. Falcone, H. Eric Tseng, F. Borrelli, J. Asgari, and D. Hrovat, "Mpc-based yaw and lateral stabilisation via active front steering and braking," *Vehicle System Dynamics*, vol. 46, no. S1, pp. 611–628, 2008.

[5] M. Zhou, J. Luo, J. Villela, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. Fadakar, Z. Chen *et al.*, "Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving," *arXiv e-prints*, pp. arXiv–2010, 2020.

[6] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[7] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 797–803.

[8] P. Lin, "Why can't google cars avoid rear-end accidents," *Forbes, 2015*, 2015.

[9] D. Sadigh, N. Landolfi, S. S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state," *Autonomous Robots*, vol. 42, no. 7, pp. 1405–1426, 2018.

[10] P. Trautman, J. Ma, R. M. Murray, and A. Krause, "Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 335–356, 2015.

[11] X. Li, X. Qiu, J. Wang, and Y. Shen, "A deep reinforcement learning based approach for autonomous overtaking," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2020, pp. 1–5.

[12] J. E. Naranjo, C. Gonzalez, R. Garcia, and T. De Pedro, "Lane-change fuzzy control in autonomous vehicles for the overtaking maneuver," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 438–450, 2008.

[13] G. Cesari, G. Schildbach, A. Carvalho, and F. Borrelli, "Scenario model predictive control for lane change assistance and autonomous driving on highways," *IEEE Intelligent transportation systems magazine*, vol. 9, no. 3, pp. 23–35, 2017.

[14] A. Raghavan, J. Wei, J. S. Baras, and K. H. Johansson, "Stochastic control formulation of the car overtake problem," *IFAC-PapersOnLine*, vol. 51, no. 9, pp. 124–129, 2018.

[15] V. S. Chipade, Q. Shen, L. Huang, N. Ozay, S. Z. Yong, and D. Panagou, "Safe autonomous overtaking with intention estimation," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 2050–2057.

[16] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *arXiv preprint arXiv:2002.00444*, 2020.

[17] L. Yu, X. Shao, and X. Yan, "Autonomous overtaking decision making of driverless bus based on deep q-learning method," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2017, pp. 2267–2272.

[18] M. Kaushik and K. M. Krishna, "Learning driving behaviors for automated cars in unstructured environments," in *European Conference on Computer Vision*. Springer, 2018, pp. 583–599.

[19] D. M. Saxena, S. Bae, A. Nakhaei, K. Fujimura, and M. Likhachev, "Driving in dense traffic with model-free reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5385–5392.

[20] S. Bae, D. Saxena, A. Nakhaei, C. Choi, K. Fujimura, and S. Moura, "Cooperation-aware lane change maneuver in dense traffic based on model predictive control with recurrent neural network," in *2020 American Control Conference (ACC)*, 2020, pp. 1209–1216.

[21] E. Leurent and J. Mercat, "Social attention for autonomous decision-making in dense traffic," *arXiv preprint arXiv:1911.12250*, 2019.

[22] Y. Chen, C. Dong, P. Palanisamy, P. Mudalige, K. Muelling, and J. M. Dolan, "Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2019, pp. 1326–1334.

[23] J. Liao, T. Liu, X. Tang, X. Mu, B. Huang, and D. Cao, "Decision-making strategy on highway for autonomous vehicles using deep reinforcement learning," *IEEE Access*, vol. 8, pp. 177 804–177 814, 2020.

[24] J. Kong, M. Pfeiffer, G. Schildbach, and F. Borrelli, "Kinematic and dynamic vehicle models for autonomous driving control design," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 1094–1099.

[25] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.

[26] R. Chen, K. D. Kusano, and H. C. Gabler, "Driver behavior during overtaking maneuvers from the 100-car naturalistic driving study," *Traffic injury prevention*, vol. 16, no. sup2, pp. S176–S181, 2015.

[27] E. Leurent, "An environment for autonomous driving decision-making," 2018.