Int. J. Appl. Math. Comput. Sci., 2021, Vol. 31, No. 3, 461–470
DOI: 10.34768/amcs-2021-0031

AN AUTOMATED DRIVING STRATEGY GENERATING METHOD BASED ON WGAIL–DDPG

MINGHENG ZHANG [a,b,*], XING WAN [b], LONGHUI GANG [c], XINFEI LV [b], ZENGWEN WU [b], ZHAOYANG LIU [b]

[a]State Key Laboratory of Structure Analysis for Industrial Equipment
Dalian University of Technology
Liaoning, Dalian 116024, China
e-mail: zhangmh@dlut.edu.cn

[b]School of Automotive Engineering
Dalian University of Technology
Liaoning, Dalian 116024, China

[c]School of Navigation
Dalian Maritime University
Liaoning, Dalian 116026, China

Reliability, efficiency and generalization are basic evaluation criteria for a vehicle automated driving system. This paper proposes an automated driving decision-making method based on the Wasserstein generative adversarial imitation learning–deep deterministic policy gradient (WGAIL–DDPG(λ)). Here the exact reward function is designed based on the requirements of a vehicle's driving performance, i.e., safety, dynamic and ride comfort performance. The model's training efficiency is improved through the proposed imitation learning strategy, and a gain regulator is designed to smooth the transition from imitation to reinforcement phases. Test results show that the proposed decision-making model can generate actions quickly and accurately according to the surrounding environment. Meanwhile, the imitation learning strategy based on expert experience and the gain regulator can effectively improve the training efficiency for the reinforcement learning model. Additionally, an extended test also proves its good adaptability for different driving conditions.

Keywords: automated driving system, deep learning, deep reinforcement learning, imitation learning, deep deterministic policy gradient.

1. Introduction

In recent years, with the rapid growth of vehicle numbers, safety and efficiency have become an urgent traffic problem that needs to be solved. Automated driving is regarded as an effective way to do so. According to information processing, the design of automated driving systems (ADSs) is divided into three steps: environment perception, decision planning, and motion control (Ziegler *et al.*, 2014). Decision planning is one of the key issues for automated driving applications.

At present, there are three solutions for ADS decision-making: rule based (Xiong *et al.*, 2015; Chen *et al.*, 2017), deep learning (DL) based (Pomerleau, 1998; Xu *et al.*, 2017), and the deep reinforcement learning (DRL), based decision-making method (Xia and Li, 2017). Rule-based solutions cannot enumerate all possibilities and emergencies (Bai *et al.*, 2019), while DRL-based solutions have many merits such as self-learning, self-reinforcement and good scenario adaptability. Thus, DRL has been considered the main solution to the problems of ADS decision-making (Zhu *et al.*, 2018).

For better performance in the field of unmanned control, Chang *et al.* (2019) proposed a mobile edge computing-based vehicular cloud of the cooperative adaptive driving approach to avoid shock-waves

*Corresponding author

efficiently in platoon driving. Hedjar and Bounkhel (2019) presented a real-time obstacle avoidance algorithm for multiple autonomous surface vehicles based on constrained convex optimization. In terms of DRL, reliability, efficiency and generalization are essential for an effective model design. Gao *et al*. (2019) proposed a vehicle decision-making model that performed well in simple traffic scenarios; Zong *et al*. (2017) established a driving decision model based on the deep deterministic policy gradient (DDPG) to solve the problems in complex scenarios. To tackle the efficiency problem, Anderson *et al*. (2015) used pre-training tricks to improve the model's training efficiency.

Imitation learning aims at imitating the distribution of expert data and has the advantage of making the agent learn basic skills quickly by using this prior knowledge. For automated driving research, it can make the controlled vehicle learn basic driving rules based on the driver's experience. Dossa *et al*. (2020) proposed a hybrid automated driving model based on the reinforcement learning method with an imitation learning strategy introduced. Test results show that the hybrid model can be trained efficiently based on the driver's experience imitation. Zou *et al*. (2020) proposed a deep deterministic policy gradient–imitation learning (DDPG–IL) algorithm which introduces a dual experience pool to store expert data and common data, and has a faster convergence speed and better performance than the ordinary DDPG algorithm.

Thus, it can be seen that, in order to improve the training efficiency of RL, related research has paid more attention to model pre-training or an experience pool improvement and has made significant progress with this issue. However, most of the approaches made improvements by using existing experience or an experience pool directly to generate an initial training strategy for RL. To some extent, this strategy cannot make full use of the relevant experience information from the perspective of optimization, so appropriate research needs to be further made so as to improve the RL training efficiency caused by the blind exploration of the agent.

Based on the above analysis, this paper proposed a WGAIL–DDPG(λ) model for ADS decision-making based on DRL, in which the reward function is specifically designed based on the requirements of vehicle driving performance. The training efficiency of the proposed DDPG model is improved by introducing imitation learning tricks. Additionally, a gain regulator is used to smooth the transition process from the imitation to the reinforcement learning phase.

The main contributions of this paper are as follows: firstly, since the essence of DRL is to enumerate all possible actions and evaluate their effects accordingly, the search space of a model's training is too large. How to improve the model training efficiency is a general and important problem to be solved. Based on the expert experience, this research reduces the search space effectively through the proposed Wasserstein generative adversarial imitation learning (WGAIL) module. Meanwhile, with the introduction of imitation learning, how to transit the training phase from imitation to reinforcement learning is a key problem that needs to be considered. In this paper, a gain regulator is designed to solve this problem. Finally, the reward function designed based on the requirements of vehicle performance has a positive effect on the training process of reinforcement learning.

2. Methodology

2.1. ADS decision-making model design.

2.1.1. Model inputs and outputs. For the DRL-based ADS decision-making model, the model's input vector (state) and output vector (action) should be determined in advance. According to the purpose of this research, vehicle control can be achieved by integrating lateral and longitudinal control. For vehicle control, the function of longitudinal control is mainly responsible for the vehicle's acceleration, deceleration and braking, and lateral control for the vehicle's steering. Therefore, herein the ADS model's output is defined as a vector combined with variables of the brake pedal travel, the accelerator pedal travel and the steering angle. Meanwhile, considering the shortages of using an image as a model's input vector in computational complexity, to simplify the ADS model's construction, herein we define the model's input vector as $[b, d, V_x, V_y, V_z, \theta]$. Detailed information about the variables can be found in Table 1.

2.1.2. Generative adversarial imitation learning algorithm. To solve the problem of insufficient generalization of expert data in imitation learning, Ho and Ermon (2016) proposed the generative adversarial imitation learning (GAIL) model based on generative adversarial networks (GANs) and the inverse reinforcement learning algorithm. Its main idea is training a data generator that can imitate the distribution of the expert policy with random noisy data. For the GAIL model, GAN is composed of two modules: a generator (G) and a discriminator (D), as shown in Fig. 1. Here G is used to generate a sample similar to the expert data distribution, and the sample is regarded as a fake one. D is used to distinguish a true sample from fake ones. In the training process of the GAIL model, G and D are optimized alternately in a zero-sum game. With the generator upgrade, the discriminator cannot distinguish fake samples from the input ones, i.e., the correct rate for any sample is 50%. At this time, the model parameters will no longer change and the model tends to be stable.

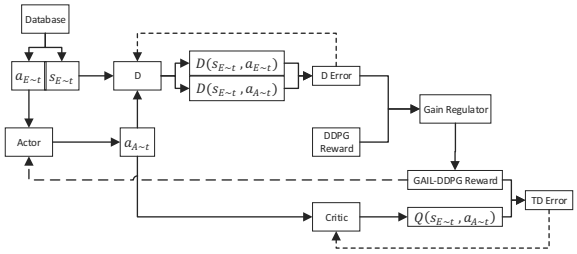


Fig. 1. Framework diagram of GAIL.

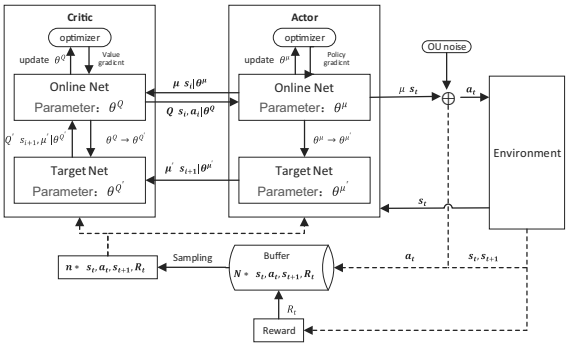


Fig. 2. DDPG-based automated driving decision model.

2.1.3. Reinforcement learning model. In this research, the proposed DDPG-based ADS decision-making model is displayed in Fig. 2. The actor-online module is used to generate an action under the current circumstance, while the critic-online module is used to evaluate the actions generated by the actor. Considering that the learning process is unstable with a single network, the DDPG divides the Actor Net and the Critic Net into two sub-networks, respectively, i.e., the Online Net and the Target Net. Both of them have same structure, but different parameters. The Online Net uses the latest parameters and updates the Target Net during certain training steps. The difference in network parameters cuts off the correlation between the Online Net and the Target Net, and this strategy makes the model learning process more stable.

For the Actor Net, which is used to generate the corresponding decision based on the current agent state, the network parameters are optimized with the Critic Net output. The loss function is

$$\nabla_{\theta_Q} J = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - Q\left(s_t, a_t \mid \theta^Q\right)\right)^2. \quad (1)$$

For the Critic Net, which is used to evaluate the output strategy from the Actor Net, the loss function is

$$\nabla_{\theta\mu} J = \frac{1}{N}\nabla_a Q\left(s_t, a_t \mid \theta^Q\right)\Big|_{s=s_t, a=\mu_{(s_t)}} \times \nabla_{\theta\mu}\mu\left(s \mid \theta^\mu\right) \mid s_t, \quad (2)$$

where y_t denotes

$$y_t = r_t + \gamma Q'\left(s_{t+1}, \mu'(s_{t+1} \mid \theta^{u'}) \mid \theta^{Q'}\right), \quad (3)$$

n stands for the number of samples per sample, r_t denotes the reward value at the current moment, γ denotes the discount factor, θ^u denotes the parameter of the Online Net in the Actor Net, $\theta^{u'}$ denotes the parameter of the Target Net in the Actor Net, θ^Q denotes the parameter of the Online Net in the Critic Net, $\theta^{Q'}$ denotes is the parameter of the Target Net in the Critic Net.

2.1.4. WGAIL–DDPG(λ) model. Due to the disadvantages of a large search space and a low efficiency of trial-and-error procedure, the reinforcement learning algorithm usually cannot yield high learning efficiency in the early stage of model training. Therefore, in this research, an imitation learning module WGAIL is used to pre-train the reinforcement learning module DDPG. The framework of the proposed WGAIL–DDPG model is shown in Fig. 3. Herein the DDPG is used as a Generator for automated driving decisions. Additionally, the gradient descent direction for DDPG updating depends on the discriminator score and the reward function.

The influence of the imitation and reinforcement module on the DDPG updating process is adjusted by a Gain regulator. Therefore, from the point of view of the Gain regulator, the proposed WGAIL–DDPG(λ) model training process can be divided into three phases: the imitation learning phase, the transition phase and reinforcement learning phase. The imitation learning strategy from expert data is to make the agent have a primary decision-making function to avoid excessive blind attempts. The gain regulator is designed to realize gradual transition from the imitation to the reinforcement learning phase, so as to ensure the reinforcement learning module can inherit the results of the imitation learning module and make full use of the expert experience effectively. Furthermore, the reward function is used to evaluate the performance of the reinforcement learning process. Based on the analysis of driving characteristics, the reward function is designed in terms of three aspects: safety, dynamic and ride comfort performance.

2.2. Experiment data description. For vehicle driving, the main external environment factors that affect a driver's decision-making include vehicular, environmental and road factors (Gu *et al*., 2020). Figure 4 shows a schematic diagram of the external environment for vehicle driving. Here, *b* reflects the lateral offset of the vehicle from the center line, *d* reflects the relative distance between the vehicles, *W* is the width of the lane, V_x is the longitudinal speed and V_y is the lateral speed of the target

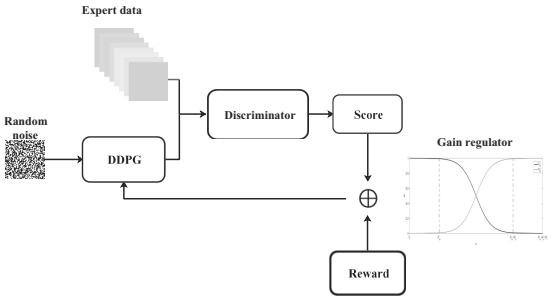


Fig. 3. Framework of the WGAIL–DDPG(λ) model.

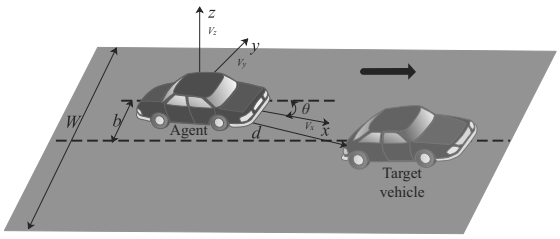


Fig. 4. Schema of the external environment for vehicle driving.

vehicle, and θ denotes an angle from the central-line to driving direction.

Based on the above analysis and the purpose of this paper, the list of acquired experimental data is shown in Table 1.

2.3. Reward function design. Reinforcement learning is a process where intelligent agents achieve the maximum reward during the interaction with the environment.

Safety is the primary requirement for intelligent vehicles, and the other requirements such as traffic efficiency and ride comfort should also be considered on this basis. Therefore, from the essential requirements analysis of the automated driving decision-making system, it should be designed meeting the driving safety requirement first and then the factors of traffic efficiency and ride comfort. Based on this opinion, in this research, safety, traffic efficiency and ride comfort are considered in reward function design. According to safety, a smaller lateral speed and offset, a greater distance from other vehicles or obstacles are conducive to the safety of vehicles. For dynamic performance, larger longitudinal velocity is beneficial to the improvement of transportation efficiency. For the ride comfort, the lower the vertical speed V_z, the better the ride comfort performance. Based on this analysis, the reward function R_D for the DDPG model and R_G for WGAIL–DDPG model can be designed respectively as

$$R_D = \begin{cases} C_S^T V_S + C_D V_D + C_R V_R, & |b| \le \frac{W}{2}, \\ -100, & |b| > \frac{W}{2}, \\ -\text{dmg}, & d < 0, \end{cases} \tag{4}$$

$$R_G = \begin{cases} (1-\lambda)(C_S^T V_S + C_D V_D \\ \quad + C_R V_R) + \lambda S_i, & |b| \le \frac{W}{2}, \\ -100, & |b| > \frac{W}{2}, \\ -\text{dmg}, & d < 0. \end{cases} \tag{5}$$

where $C_S = [c_1, c_2]^T$, $C_D = c_3$, and $C_R = c_4$ denote the weight coefficients of safety, dynamic performance and ride comfort, respectively, which are used to characterize the different proportion in the reward function. Also,

$$V_S = \left[-|b| V_x \sin\theta, \text{Sgn}\left(V_d - V_x \cos\theta\right)(200 - d)\right]^T,$$

$$V_D = V_x \cos\theta,$$

$$V_R = -|V_z|$$

denote the relevant vectors of safety, dynamics and ride comfort, respectively, while V_d is the target vehicle speed; λ is the hyper-parameter of reinforcement learning, which is used to adjust the weight of the reward function and the gradient descent direction. Additionally, S_i is the score of the Discriminator; 'dmg' is the degree of damage to the agent when a collision occurs, the symbolic function 'Sgn' is the enumeration constant, and its value is -1, 0 or 1. Here 0 means that there is no target vehicle around the agent, 1 means that the target vehicle is in front of the agent, and -1 denotes that the target vehicle is behind the agent.

2.4. Gain regulator design. To implement a gradual transition from the imitation learning to the reinforcement learning phase, a gain regulator is designed to achieve the dynamic adjustment between the two training phases. The main idea of the regulator design is as follows: in the early training phase, the model's input mainly depends on the imitation learning module; then the training process gradually shifts to the reinforcement learning module. For the WGAIL–DDPG(λ) model, within the imitation phase, the score of the discriminator plays a major role in generator action optimization, while, for reinforcement learning phase, the reward function is important for the agent action optimization. Therefore, λ should have characteristics of gradual attenuation with the training process and a low attenuation rate in the early transition phase.

Based on this analysis, three types of gain regulators are discussed; linear attenuation, exponential attenuation,

Table 1. Experimental data acquisition.

	Symbol	Unit	Description
Agent information	V_x	km/h	longitudinal velocity
	V_y	km/h	lateral velocity
	V_z	km/h	vertical velocity
Environment information	θ	rad	heading angle of the agent (angle deviation)
	b	m	distance from the agent to the road center line (lateral deviation)
	d	m	distance from the agent to the target vehicle (safe distance)
	dmg	–	current damage of the agent (the higher the value, the higher the damage)

and a 1-sigmoid attenuation model, which are shown in Fig. 5. Here, the expressions for λ are

$$\lambda_1 = N_0 - \alpha n, \tag{6}$$

$$\lambda_2 = N_0 e^{-\alpha n}, \tag{7}$$

$$\lambda_3 = N_0 - \frac{1}{1 + e^{-\alpha(n-c)}}, \tag{8}$$

where N_0 denotes the initial value of the attenuation gain regulator and α denotes the attenuation index. Consequently, the gain regulator decreases with the training rounds n.

From Fig. 5, some conclusions can be drawn as follows:

1. For a linear attenuation gain regulator with constant rate α, the transition process may be unstable with a large α. In this case, the expert data are not fully utilized. On the other hand, for a small α, the model training rounds in the transition phase increased correspondingly, which makes the overall training efficiency decrease.

2. For an exponential attenuation gain regulator, which has a variable attenuation rate, the attenuation degree of λ gradually decreases with the training rounds. In the early transition phase, with a large attenuation rate, the reinforcement learning decision may deviate from the expert experience.

3. For a 1-sigmoid attenuation gain regulator, the attenuation degree of λ in the early transition phase is low, which can satisfy the requirements of this research.

For the 1-sigmoid regulator, there are three hyper-parameters: N_0, α and the symmetry axis c. In this research, $N_0 = 1$, $c = N_e/2, N_e$ denotes the total number of training rounds of the transition phase and α is an undetermined coefficient. The influence of α on the gain regulator is shown in Fig.6.

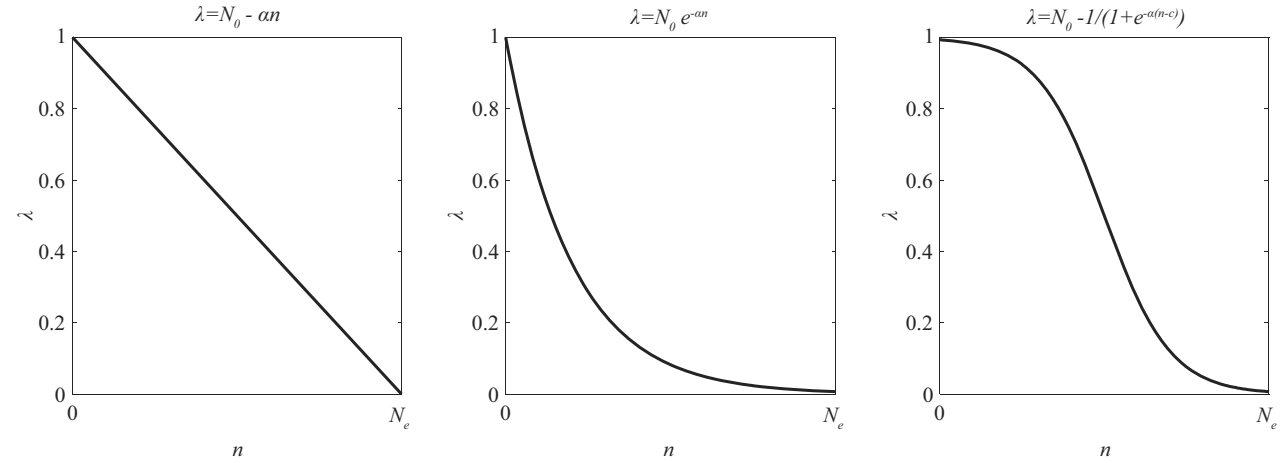
From Fig. 6 it can be seen that the gain regulator has the characteristic of a step change and the maximum of the first order derivative increases with α. For this research, the gain regulator should have the characteristic of a low attenuation rate and no breakpoints, so $\alpha = 0.2$ is selected. Based on this analysis, for the entire model, the transition process of the gain regulator is shown in Fig. 7.

Here, n denotes the model training rounds, while w, D_w and R_w denote the coefficient weight, discriminator weight and the reward function weight in the DDPG model, respectively. N_m, N_e, and N_r denote the training times of the imitation learning phase, the transition phase and the reinforcement learning phase, respectively. Within stage $[0, N_m]$, $\lambda = 1$ indicates that the discriminator score plays major role for the optimization of the generator action and the reward function does not work; within $[N_m, N_m + N_e]$, the reward function starts to work with the training proceeding, and the supervisory role of the discriminator gradually decreases. Within the reinforcement learning phase, $\lambda = 0$ indicates that the distribution of expert data will not affect the agent, so that the latter can explore more advanced decision-making strategies.

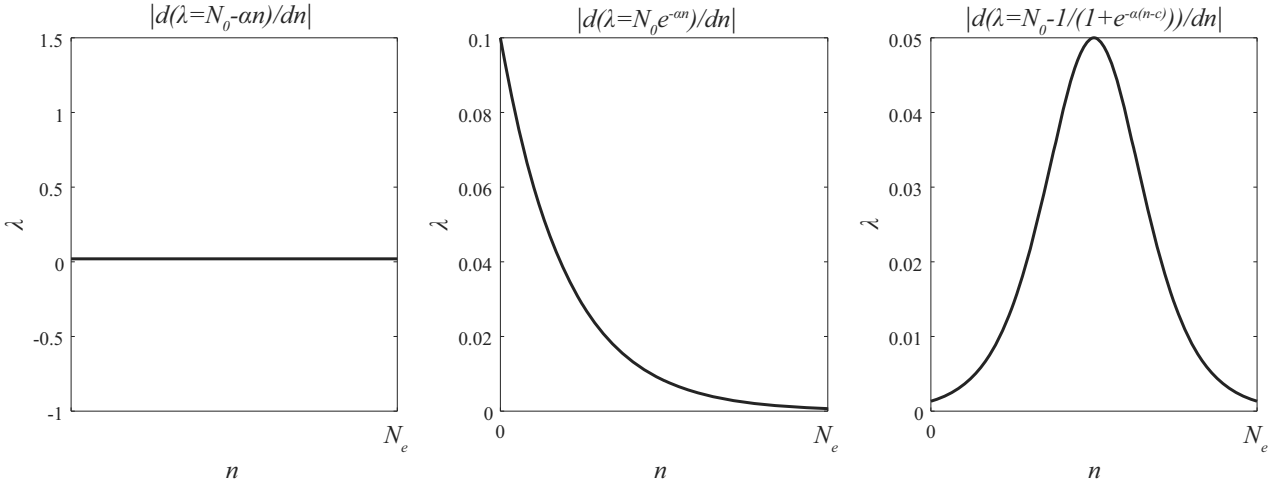
3. Testing and result analysis

In terms of this research target, the following experiments and analyses are carried out, including model performance, adaptability and efficiency.

3.1. Experimental scenarios. TORCS (The Open Racing Car Simulator) is an open-source automated driving simulator that can create a physical separation between the game engine and the drivers. Users can obtain the vehicle state and environment information without



(a) three kinds of gain regulators



(b) absolute value of the first derivative of the different gain regulators

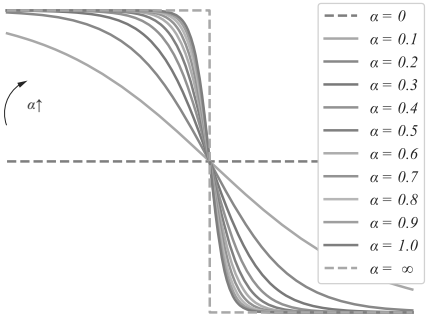
Fig. 5. Characteristic of different gain regulators.

having to understand the internal program structure. It can efficiently improve the ADS algorithm development. TORCS provides a series of test scenarios that are created from a natural driving environment. In this research, the scenery for data acquisition and training is determined as the CG-1 track (shown in Fig. 8). Details about CG-1 are as follows it is: 2057.56 meters long, 15 meters wide and has a variety of line shapes such as long straight lines and curves with different curvatures.

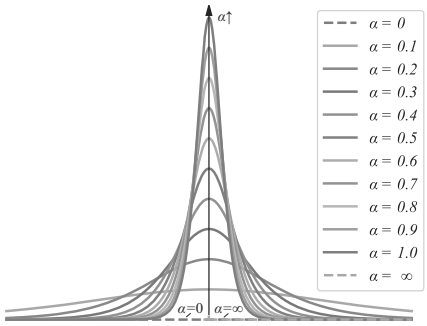
3.2. Model performance. In this research, the reward function design is based on three aspects; the relevant tests are conducted and some conclusions can be drawn.

1. *Stability performance.* Stability is one of the basic properties that evaluate the vehicle's handling performance. Here, the normalized distance $b/(W/2)$ is used to evaluate the stability performance of the proposed ADS model. The closer the value of $b/(W/2)$ to 0, the better the tracking stability of the control system. From Fig. 9, it can be seen that the normalized distance changes in the range of $[-0.3, 0.3]$, which indicates that the proposed model can control the vehicle driving well.

2. *Safety performance.* Here, the safety distance d between vehicles is used for safety evaluation. The smaller the value of d, the greater the risk of a collision. From the test results shown in Fig. 10, it can be seen that the safe distance is maintained at more than 10 meters under stable driving conditions, which can ensure the vehicle's safe driving. When disturbed by other vehicles, as shown in the circle mark, the proposed ADS can



(a) gain regulator with different α



(b) first order derivative of gain regulators

Fig. 6. Characteristic of regulators with different α.

make accurate adjustments according to the changes in the environment, so as to avoid collision accidents.

3.3. Model's adaptability. Adaptability denotes the ability of an algorithm to adapt to new samples or environments. To verify the proposed WGAIL–DDPG model's performance, three types of tracks are used for testing the model's adaptability. The composition of road alignment of each track and comparisons are shown in Fig. 11.

The test tracks are classified according to the driving difficulty level. The component of the line shape for the test track includes Straight, Simple, Right-angle, U-turn, S-turn and Acute Shape. For instance, the Alpine track includes six line shapes and the CG-2 track includes three line shapes (shown in Fig. 11). From this, the driving difficulty of the three test tracks is Alpine $>$ CG-1 $>$ CG-2.

It should be noted (Table 2). that the proposed model can control the vehicle to drive safely and stably on the CG-2 track. For the track with a more complex road alignment (Alpine), the model can work well without any specific training, although there are many unknown road conditions. However, for the more complex Alpine track, a minor collision occurred during the second test lap. The reason is that the track CG-1 used for model training does not include U-turns with a comparable difficulty level to the Alpine track, which makes the agent unable to handle this untrained situation. For this problem, it can be solved by increasing the line shapes of training tracks.

3.4. Learning efficiency. The relationship between the cumulative return and the number of training rounds is used for evaluating the efficiency of the proposed model. Therefore, in order to test the model's efficiency under complicated driving conditions, multiple vehicles are used for enhancing the complicity of the environment. To make a comparison, the DDPG-based and WGAIL–DDPG(λ)-based models are tested separately and the comparison results are shown in Fig. 12. From the results of efficiency comparison, some conclusions can be drawn as follows.

Phase-1, $n \in [1, 100]$. From the slope of the cumulative return, it can be seen that, within the early training phase, the efficiency of WGAIL–DDPG(λ) is significantly higher than that of the DDPG. The reason is that the imitation learning phase can make the agent achieve the expert experience quickly. In addition, it also further verifies the effectiveness of the introduction of the imitation learning strategy.

Further analysis shows that when n is almost equal to 40, the cumulative return value of WGAIL–DDPG(λ)

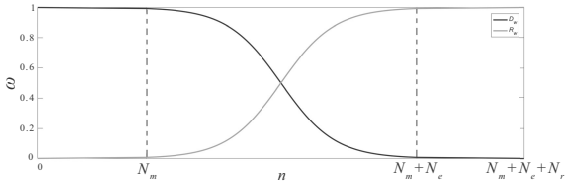
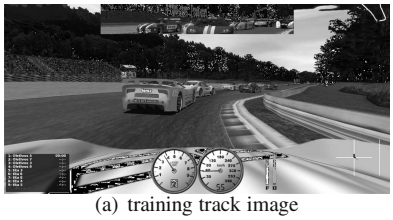


Fig. 7. Transition process from imitation to reinforcement learning.



(a) training track image

(b) CG-1 track map

Fig. 8. CG-1 track.

Table 2. Test results of the model's adaptability.

Track	First lap	Second lap	Third lap	Forth lap	Fifth lap
CG-1	✔	✔	✔	✔	✔
CG-2	✔	✔	✔	✔	✔
Alpine	✔	✗	✔	✔	✔

Note: ✗means a collision, ✔means no collisions.

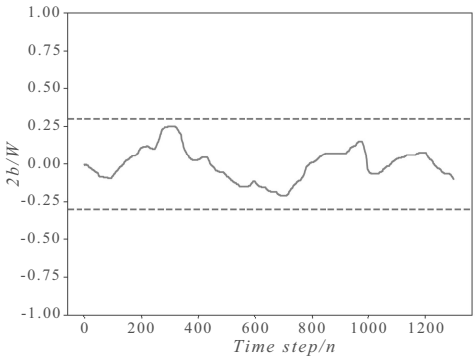
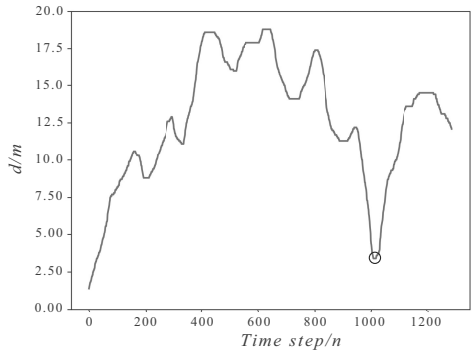
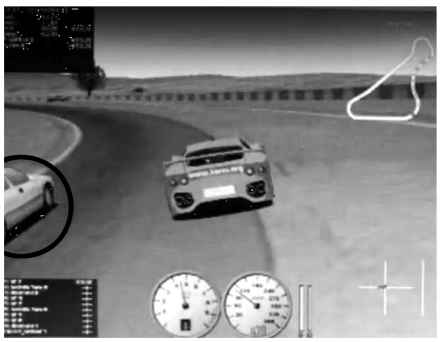


Fig. 9. Stability performance test.



(a) distance between the agent and the target vehicle



(b) diagram when the agent is interfered by the target vehicle

Fig. 10. Safety performance.

reaches a level of 9,000 quickly, which indicates that the agent had simpler driving strategies such as lane-following, while the DDPG is still in the trial-and-error phase.

Furthermore, the imitation learning phase can avoid the blind attempt of the agent in the initial training phase, which greatly improves the training efficiency of reinforcement learning.

Phase-2, $n \in [100, 550]$. In the subsequent phase of the training, the cumulative return obtained by WGAIL–DDPG(λ) is significantly higher than that of the DDPG. The reason is that, after learning the primary driving strategies, the agent uses the gain regulator to realize a transition from the imitation phase to the reinforcement learning phase, so that the agent can explore more advanced driving strategies.

Further analysis shows that when the number of training rounds is about 140, the agent of WGAIL–DDPG(λ) is stable at the cumulative return of 19,000. This indicates that the agent has a basic behavior of avoiding vehicles and driving stably along the lane. By comparison, the agent of the DDPG achieves the same goal often around 480 training rounds. Additionally, the learning speed of WGAIL–DDPG(λ) is approximately 3.4 times faster than that of the DDPG.

Thus, a smooth transition from the imitation learning phase to the reinforcement learning phase can be achieved through the designed gain regulator. Additionally, it also can allow the agent to further explore advanced strategies after basic driving strategies learning from imitation.

4. Conclusions

With the rapid development of vehicle intelligence, automated driving has become the focus in the research field of vehicle engineering. For the entire vehicle automated driving process, a decision-making strategy based on the dynamic environment is a core problem to be solved for ADS development. Focusing on this problem, this paper proposed a WGAIL–DDPG(λ) model based on DRL to solve vehicle automated driving decision-making problems.

To improve the model's training efficiency that is important for deep learning application, this research reduces the search space through an imitation learning

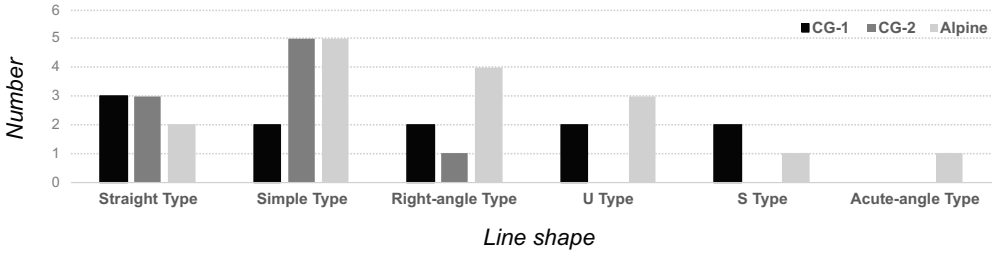


Fig. 11. Comparisons of different tracks.

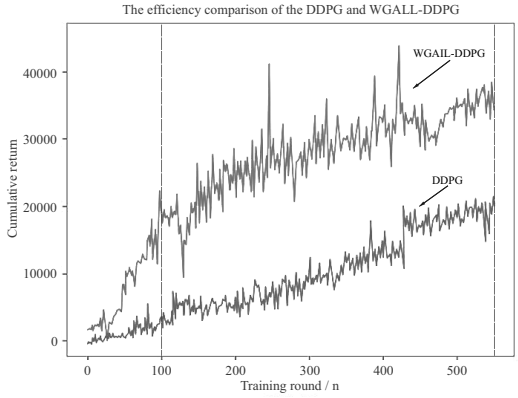


Fig. 12. Comparisons of the model's learning efficiency.

strategy. Meanwhile, to ensure the efficiency of the transition process from the imitation learning module to the reinforce learning module, a gain regulator is designed to balance the relationship between them. For the reward function, the basic problems to be considered for deep learning, it was designed based on the performance requirements including the vehicle's safety, dynamic and ride comfort performance. Test results show that the designed reward function can effectively ensure the reliable output of the DDPG decision model. The proposed imitation learning strategy based on expert experience and the designed gain regulator can effectively improve training efficiency for the reinforcement learning module.

In addition, it should be pointed out that, due to the difficulty of real driving environment tests, this research only verified the proposed model on the environment under a simulation platform. Its reliability and advantages need to be further verified with actual experiment data. In the future, the proposed model should be improved by merging more performance requirements in actual driving conditions.

Acknowledgment

This project is supported by the National Natural Science Foundation of China (grants no. 51675077 and 52171345) and the China Postdoctoral Science Foundation (2017T100178, 2015M581329).

References

Anderson, C.W., Lee, M. and Elliott, D.L. (2015). Faster reinforcement learning after pretraining deep networks to predict state dynamics, *2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland*, pp. 1–7.

Bai, Z., Shangguan, W., Cai, B. and Chai, L. (2019). Deep reinforcement learning based high-level driving behavior decision-making model in heterogeneous traffic, *2019 Chinese Control Conference (CCC), Guangzhou, China*, pp. 8600–8605.

Chang, B.-J., Hwang, R.-H., Tsai, Y.-L., Yu, B.-H. and Liang, Y.-H. (2019). Cooperative adaptive driving for platooning autonomous self driving based on edge computing, *International Journal of Applied Mathematics and Computer Science* **29**(2): 213–225, DOI: 10.2478/amcs-2019-0016.

Chen, X., Tian, G., Miao, Y. and Gong, J.-w. (2017). Driving rule acquisition and decision algorithm to unmanned vehicle in urban traffic, *Transactions of Beijing Institute of Technology* **37**(5): 491–496.

Dossa, R.F., Lian, X., Nomoto, H., Matsubara, T. and Uehara, K. (2020). Hybrid of reinforcement and imitation learning for human-like agents, *IEICE Transactions on Information and Systems* **103**(9): 1960–1970.

Gao, H., Shi, G., Wang, K., Xie, G. and Liu, Y. (2019). Research on decision-making of autonomous vehicle following based on reinforcement learning method, *Industrial Robot: The International Journal of Robotics Research and Application* **46**(3): 444–452.

Gu, X., Han, Y. and Yu, J. (2020). Vehicle lane-changing decision model based on decision mechanism and support vector machine, *Journal of Harbin Institute of Technology* **52**(07): 111–121.

Hedjar, R. and Bounkhel, M. (2019). An automatic collision avoidance algorithm for multiple marine surface vehicles, *International Journal of Applied Mathematics and Computer Science* **29**(4): 759–768, DOI: 10.2478/amcs-2019-0056.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning, *Advances in Neural Information Processing Systems* **29**: 4572–4580.

Pomerleau, D. (1998). ALVINN: An autonomous land vehicle in a neural network, *in* D.S. Touretzky (Ed), *Advances in Neural Information Processing Systems*, Morgan Kaufmann Publishers, Burlington.

Xia, W. and Li, H. (2017). Training method of automatic driving strategy based on deep reinforcement learning, *Journal of Integration Technology* **6**(3): 29–40.

Xiong, G.-m., Li, Y. and Wang, S.-y. (2015). A behavior prediction and control method based on FSM for intelligent vehicles in an intersection, *Transactions of Beijing Institute of Technology* **35**(1): 7.

Xu, H., Gao, Y., Yu, F. and Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA*, pp. 2174–2182.

Zhu, M., Wang, X. and Wang, Y. (2018). Human-like autonomous car-following model with deep reinforcement learning, *Transportation Research C: Emerging Technologies* **97**: 348–368.

Ziegler, J., Bender, P., Schreiber, M., Lategahn, H., Strauss, T., Stiller, C., Dang, T., Franke, U., Appenrodt, N., Keller, C.G., Kaus, E., Herrtwich, R.G., Rabe, C., Pfeiffer, D., Lindner, F., Stein, F., Erbs, F., Enzweiler, M., Knoppel, C., Hipp, J., Haueis, M., Trepte, M., Brenk, C., Tamke, A., Ghanaat, M., Braun, M., Joos, A., Fritz, H., Mock, H., Hein, M. and Zeeb, E. (2014). Making Bertha drive—An autonomous journey on a historic route, *IEEE Intelligent Transportation Systems Magazine* **6**(2): 8–20.

Zong, X., Xu, G., Yu, G., Su, H. and Hu, C. (2017). Obstacle avoidance for self-driving vehicle with reinforcement learning, *SAE International Journal of Passenger Cars—Electronic and Electrical Systems* **11**(07-11-01-0003): 30–39.

Zou, Q., Xiong, K. and Hou, Y. (2020). An end-to-end learning of driving strategies based on DDPG and imitation learning, *2020 Chinese Control And Decision Conference (CCDC), Hefei, China*, pp. 3190–3195.

Mingheng Zhang received his MS and PhD degrees from Jilin University in 2004 and 2007, respectively. He has published more than 40 papers. Now he is a full associate professor at the Dalian University of Technology, School of Automotive Engineering. His research interests include intelligent vehicles, safety assistant driving, and image processing.

Xing Wan obtained his BS degree in vehicle engineering from the Shandong University of Technology in Zibo, China, in 2019. He is now studying for a Master's degree in the School of Automotive Engineering at the Dalian University of Technology. His main research interests include reinforcement learning and imitation learning.

Longhui Gang received her MS and PhD degrees in transportation engineering from Jilin University in 2004 and 2007, respectively. Now she is a full associate professor at Dalian Maritime University, School of Navigation. Her research interests include traffic safety and traffic information proceeding.

Xinfei Lv received his BS degree in vehicle engineering from the Shandong University of Science and Technology, Qingdao, China, in 2020. He is now studying for his MS degree at the School of Automotive Engineering of the Dalian University of Technology. His main research interests include reinforcement learning and automatic driving.

Zengwen Wu received his BS degree at the Shandong University of Science and Technology, Tsingtao, China, in 2019. He is now studying for his MS degree in vehicle engineering at the School of Automotive Engineering of the Dalian University of Technology. His main research interests include ADAS and man-machine driving.

Zhaoyang Liu received his BS degree in automobile service engineering from Liaocheng University, China, in 2020. He is now studying for his MS degree at the School of Automotive Engineering of the Dalian University of Technology. His main research interests include reinforcement learning and ADAS.

Received: 22 January 2021
Revised: 9 June 2021
Accepted: 12 July 2021