Engineering 21 (2023) 75–91

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence and Autonomous Driving—Article

Toward Human-in-the-Loop AI: Enhancing Deep Reinforcement Learning via Real-Time Human Guidance for Autonomous Driving

Jingda Wu, Zhiyu Huang, Zhongxu Hu, Chen Lv *

School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 9 October 2021
Revised 4 April 2022
Accepted 10 May 2022
Available online 20 July 2022

Keywords:
Human-in-the-loop AI
Deep reinforcement learning
Human guidance
Autonomous driving

ABSTRACT

Due to its limited intelligence and abilities, machine learning is currently unable to handle various situations thus cannot completely replace humans in real-world applications. Because humans exhibit robustness and adaptability in complex scenarios, it is crucial to introduce humans into the training loop of artificial intelligence (AI), leveraging human intelligence to further advance machine learning algorithms. In this study, a real-time human-guidance-based (Hug)-deep reinforcement learning (DRL) method is developed for policy training in an end-to-end autonomous driving case. With our newly designed mechanism for control transfer between humans and automation, humans are able to intervene and correct the agent's unreasonable actions in real time when necessary during the model training process. Based on this human-in-the-loop guidance mechanism, an improved actor-critic architecture with modified policy and value networks is developed. The fast convergence of the proposed Hug-DRL allows real-time human guidance actions to be fused into the agent's training loop, further improving the efficiency and performance of DRL. The developed method is validated by human-in-the-loop experiments with 40 subjects and compared with other state-of-the-art learning approaches. The results suggest that the proposed method can effectively enhance the training efficiency and performance of the DRL algorithm under human guidance without imposing specific requirements on participants' expertise or experience.

1. Introduction

The development of autonomous vehicles (AVs) has gained increasing attention from both academia and industry in recent years [1]. As a promising application domain, autonomous driving has been boosted by ever-growing artificial intelligence (AI) technologies [2]. From the advances made in environment perception and sensor fusion to the successes achieved in human-like decision and planning [3], we have witnessed great innovations being developed and applied in AVs [4]. As an alternative option to the conventional modular solution that divides the driving system into connected modules such as perception, localization, planning, and control, end-to-end autonomous driving has become promising. It now serves as a critical test-bed for developing the perception and decision-making capabilities of AI and AVs.

Imitation learning (IL) and deep reinforcement learning (DRL) are two main branches of learning-based approaches, especially in the fields of end-to-end autonomous driving. IL aims to mimic human drivers to reproduce demonstration control actions in given states. Thanks to its intuitive and easy-to-use characteristics, IL has been applied in AV control strategies in many specific cases, including rural [5] and urban driving scenarios [6]. However, two major inherent issues of IL have been exposed in practical applications. The first issue is the distributional shift, that is, imitation errors accumulated over time lead to deviations from the training distribution, resulting in failures in control [7]. Various methods, including dataset aggregation (DAgger) IL [8], generative adversarial IL (GAIL) [9], and their derivative methods, have been proposed to mitigate this problem. The other issue is the limitation of asymptotic performance. Since IL behavior is derived from the imitation source (i.e., the experts who provide the demonstrations), the performance of the learned policies is limited and is unlikely to surpass that of the experts. DRL, which is another data-driven self-optimization-based algorithm, shows great potential for mitigating the aforementioned issues [10–12]. Constructed by

* Corresponding author.
E-mail address: lyuchen@ntu.edu.sg (C. Lv).

https://doi.org/10.1016/j.eng.2022.05.017

exploration–exploitation and trial-and-error mechanisms, DRL algorithms are able to autonomously search for feasible control actions and optimize a policy [13]. During the early stage of DRL development, some model-free algorithms, such as deep Q-learning (DQL) and deep deterministic policy gradient (DDPG) [14], were popular in driving policy learning for AVs [15]. More recently, actor-critic DRL algorithms with more complex network structures have been developed and have achieved better control performance in autonomous driving [16]. In particular, state-of-the-art algorithms including soft actor-critic (SAC) [17] and twin-delayed DDPG (TD3) [18] have been successfully implemented in AVs in many challenging scenarios, such as complex urban driving and high-speed drifting conditions [19].

Although many achievements have been made in DRL methods, challenges remain. The major challenge is the computational or learning efficiency. In most situations, the efficiency of the interactions between the agent and environment is very low, and the model training consumes a remarkable amount of computational resources and time [20]. The learning efficiency can be even worse when the reward signal generated by the environment is sparse. Thus, reward-shaping methods have been proposed to improve learning efficiency in a reward-sparse environment [21]. Another challenge is that DRL methods (particularly with training from scratch) exhibit limited capabilities in scene understanding under complex environments, which inevitably deteriorates their learning performance and generalization capability. Therefore, in AV applications, DRL-enabled strategies are still unable to surpass and replace human drivers in handling various situations due to the limited intelligence and ability of these strategies [22,23]. In addition, certain emerging methods have reconsidered human characteristics and attempted to learn from commonsense knowledge and neuro-symbolics [24] to improve machine intelligence. As humans exhibit robustness and high adaptability in context understanding and knowledge-based reasoning, it is promising to introduce human guidance into the training loop of data-driven approaches, thereby leveraging human intelligence to further advance learning-based methods for AVs.

Human intelligence can be reflected in several aspects of DRL training, including human assessment, human demonstration, and human intervention. Some researchers have made great efforts to introduce human assessments into DRL training and have in fact succeeded in related applications, such as simulation games [25] and robotic action control [26]. However, these methods struggle to handle many other more complex application scenarios in which explicit assessments are unavailable. Instead, humans' direct control over and guidance for agents could be more efficient for algorithm training, which gives rise to the architecture of incorporating DRL with learning from demonstration (LfD) [27] and learning from intervention (LfI) [28]. Within these two frameworks, behavior cloning (BC) [29] and inverse reinforcement learning [30] have been integrated with representative algorithms, such as DQL [31,32] and DDPG [27]. Associated implementations in robotics were subsequently reported, demonstrating improved performance compared with the original reinforcement learning [33]. However, these methods are still far from mature. They either directly replace the output actions of DRL by using human actions or use supervised learning (SL) with human demonstrations to pre-train the DRL agent, while the learning algorithm architecture remains unchanged.

Recently, attempts have been made to modify the structure of DRL. By redefining policy functions and adding BC objectives, the new DRL schemes are able to effectively accelerate the training process of DRL by leveraging offline human experience [34,35]. However, for offline human-guidance-based (Hug)-DRLs, it is difficult to design a threshold beforehand for human intervention due to the involvement of many non-quantitative factors. Instead, the rapid scene-understanding and decision-making abilities of humans in complex situations can be presented via real-time human–environment interactions and further help improve the performance of DRL agents. Therefore, compared with offline human guidance, real-time Hug schemes would more efficiently train a DRL agent.

Nevertheless, there are still two main issues with the existing DRL methods under real-time human guidance. First, long-term supervision and guidance are exhausting for human participants [36]. To adapt to a human driver's physical reactions in the real world, the procedure of an existing DRL algorithm must be slowed down in a virtual environment [37]. The resulting extensive training process decreases learning and computational efficiency and leads to negative subjective feelings among humans [32]. Second, existing DRL methods with human guidance usually require expert-level demonstrations to ensure the quality of the data collected and achieve an ideal improvement in performance. However, costly manpower and a shortage of professionals in real-world large-scale applications limit the usage of this type of method [38]. Therefore, the capability of existing approaches—particularly their data-processing efficiency—should be further improved to ensure that Hug-DRL algorithms are feasible in practice. In addition, more explorations should be conducted to lower the requirements for human participants in Hug-DRL algorithms.

To fill the abovementioned research gap and further advance the DRL method, the present work develops a human-in-the-loop DRL framework that effectively leverages human intelligence in real time during model training. A real-time Hug-DRL method is developed and successfully applied to agent training in autonomous driving scenarios. Under the proposed architecture, we propose a dynamic learning process leveraging human experience with the aim of optimizing the learning efficiency and performance of an off-policy DRL agent. In every single learning step, an evaluation module weights the human guidance actions and the DRL agent's actions according to their respective utilities. The high-level architecture of the proposed method is illustrated in Fig. 1, and the concept behind this prototype is extensively applicable beyond the specific scenario of this study. The detailed algorithms, experimental results, and methodology adopted are reported below.

2. Enhanced DRL algorithm with real-time human guidance

In typical applications of DRL, such as autonomous driving, the control of the DRL agent can be formulated as a Markov decision process (MDP), which is represented by a tuple $\mathcal{M}$, including the state space $\mathcal{S} \in \mathbb{R}^n$, action space $\mathcal{A} \in \mathbb{R}^m$ (where n and m are the dimensional of the state space and action space, respectively; $\mathbb{R}$ is the real number set), transition model $\mathcal{T}: \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and reward function $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, as follows:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}) \tag{1}$$

At a given time step t, the agent executes an action $a_t \in \mathcal{A}$ in a given state $s_t \in \mathcal{S}$ and receives a reward signal $r_t = \mathcal{R}(s_t, a_t)$. Then, the environment transitions into a next-step state $s_{t+1} \in \mathcal{S}$ according to the environmental dynamics $\mathcal{T}(\cdot|s_t, a_t)$. In the autonomous driving scenario, the transition probability model $\mathcal{T}$ for environmental dynamics is difficult to formulate. Thus, we adopt model-free reinforcement learning, which does not require the transition dynamics to be modeled, to solve this problem.

In this work, a state-of-the-art off-policy actor-citric method—namely, TD3—is used to construct the high-level architecture, as shown in Fig. S1 in Appendix A. The TD3 algorithm chooses a deterministic action through policy network μ, adjusting its action–selection policy under the guidance of value network Q. The value network approximates the value of the specific state and action

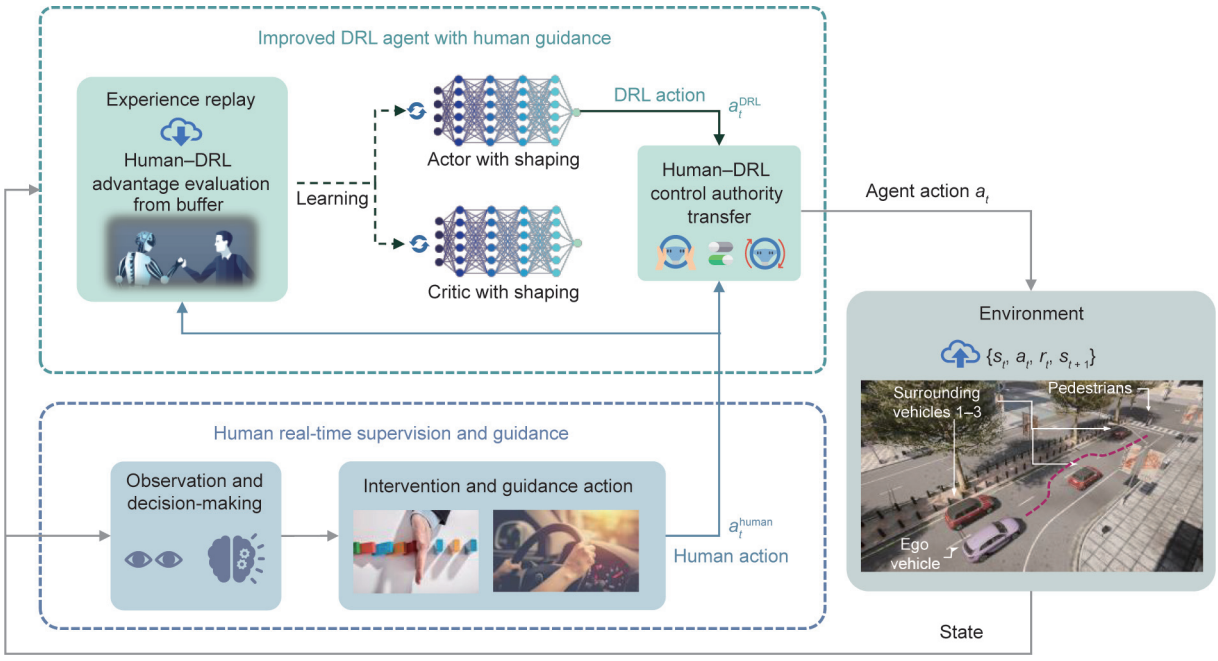

Fig. 1. The high-level architecture of the proposed Hug-DRL method with real-time human guidance. By introducing human guidance actions into both real-time manipulation and the offline learning process, the training performance is expected to be significantly improved. t: the time step; a_t^{DRL}: the DRL policy's action; a_t^{human}: the human guidance action; a_t: the eventual action to interact with the environment; s_t: the current state variable; s_{t+1}: the next-step state variable; r_t: the reward signal.

based on the Bellman iterative equation. Next, TD3 sets two value networks, Q_1 and Q_2, to mitigate the overestimation issue. To smooth the learning process, target networks μ', Q_1', and Q_2' are adopted. The overall structure used is shown in Fig. S2 in Appendix A.

To realize the human-in-the-loop framework within the reinforcement learning algorithm, we combine LfD and LfI into a uniform architecture in which humans can decide when to intervene and override the original policy action and provide their real-time actions as demonstrations. Thus, an online switch mechanism between agent exploration and human control is designed. Let $\mathcal{H}(s_t) \in \mathbb{R}^n$ denote a human's policy. The human intervention guidance is formulated as a random event $I(s_t)$ with the observation of the human driver to the current states. Then, agent action a_t can be expressed as follows:

$$a_t = I(s_t) \cdot a_t^{\text{human}} + [1 - I(s_t)] \cdot a_t^{\text{DRL}} \tag{2a}$$

$$a_t^{\text{DRL}} = \text{clip}\left(\mu(s_t|\Theta^{\mu}) + \text{clip}(\epsilon, -c, c), a_{\text{low}}, a_{\text{high}}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma) \tag{2b}$$

where $a_t^{\text{human}} \in \mathcal{H}$ is the guidance action given by a human; a_t^{DRL} is the action given by the policy network; $I(s_t)$ is equal to 0 when there is no human guidance or 1 when human action occurs; Θ^{μ} denotes the parameters of the policy network; a_{low} and a_{high} is the lower and upper bounds of the action space, respectively; ϵ is the noise subject to a Gaussian distribution with a standard deviation of σ; and c is the clipped noise boundary. The purpose of adding Gaussian noise is to incentivize explorations in the deterministic policy. The mechanism designed by Eq. (2a) fully transfers the driving control authority to the human participant whenever he or she feels it is necessary to intervene in an episode during agent training.

The value network approximates the value function, which is obtained from the expectation of future reward as follows:

$$Q^{\pi}(s,a) = \mathop{\mathbb{E}}_{s\sim\mathcal{T}, a\sim\pi(\cdot|s)}\left[\sum_{i=0}^{\infty} \gamma^i \cdot r_i\right] \tag{3}$$

where γ is the discount factor to evaluate the importance of future rewards; $\mathbb{E}[\cdot]$ denotes the mathematical expectation; i denotes the index of counted time step. Let $Q(s,a)$ be the simplified representation for $Q^{\pi}(s,a)$. The superscript regarding the policy π is omitted unless specified.

To solve the above expectation, the Bellman iteration is employed, and the expected iterative target of value function y at step t can be calculated as follows:

$$y_t = r_t + \gamma \min_{j=1,2} Q_j'\left(s_{t+1}, \mu'\left(s_{t+1}|\Theta^{\mu'}\right)\middle|\Theta^{Q_j'}\right) \tag{4}$$

where $\Theta^{\mu'}$ denotes the parameters of the target policy network; $\Theta^{Q_j'}$ denotes the parameters of the target value networks; j denotes the index of two value networks Q_1 and Q_2.

The two value networks Q_1 and Q_2 with the same structure aim to address the overestimation issue through clipped functionality. In addition, target policy network μ'—rather than policy network μ—is used to smooth policy updates. Then, the loss function of the value networks in TD3 is expressed as follows:

$$\mathcal{L}^{Q_j}\left(\Theta^{Q_j}\right) = \mathop{\mathbb{E}}_{\left(s_t, a_t, r_t, s_{t+1}\right)\sim\mathcal{D}}\left[\left|\left|y_t - Q_j\left(s_t, a_t|\Theta^{Q_j}\right)\right|\right|^2\right] \tag{5}$$

where $\mathbb{E}$ denotes the expected value; Θ^{Q_j} denotes the parameters of the value networks; and $\mathcal{D}$ denotes the experience replay buffer, which consists of the current state, the action, the reward, and the state of the next step.

The policy network that determines the control action is intended to optimize the value of the value network—that is, to improve the control performance in the designated autonomous driving scenario in this study. Thus, the loss function of the policy network in the TD3 algorithm is designed as follows:

$$\mathcal{L}^{\mu}(\Theta^{\mu}) = -\mathbb{E}\left[Q_1\left(s_t, a_t^{\text{DRL}}\right)\right] = -\mathop{\mathbb{E}}_{s_t\sim\mathcal{D}}\left[Q_1(s_t, \mu(s_t|\Theta^{\mu}))\right] \tag{6}$$

Eq. (6) indicates that the expectation for the policy is to maximize the value of the value network, which corresponds to minimizing the loss function of the policy network. The unbiased estimation of a_t^{DRL} is equal to that of $\mu(s_t|\Theta^{\mu})$, since the noise in Eq. (2b) is of a zero-mean distribution.

When human guidance a_t^{human} occurs, the loss function of the TD3 algorithm should be revised accordingly to incorporate it with human experience. Thus, the value network in Eq. (5) can be rewritten as follows:

$$\mathcal{L}^{Q_j}\left(\Theta^{Q_j}\right)=\underset{\left(s_t,a_t,r_t,s_{t+1}\right)\sim\mathcal{D}}{\mathbb{E}}\left[\left(\mathcal{Y}_t-Q_j\left(s_t,a_t^{\text{human}}|\Theta^{Q_j}\right)\right)^2\right] \tag{7}$$

In fact, the mechanism shown in Eq. (7) modified from Eq. (4) is sufficient for establishing a Hug-DRL scheme [34,35], which has been validated and reported in existing studies [32]. However, merely modifying the value network without updating the loss function of the policy network would affect the prospective performance of human guidance, as stated in Refs. [34,35], because the value network is updated based on $\{s_t,a_t^{\text{human}}\}$, whereas the policy network still relies on $\{s_t,\mu(s_t|\Theta^{\mu})\}$. This would lead to inconsistency in the updating direction of the actor and critic networks. The detailed rationale behind this phenomenon will be analyzed in detail in Section 6.

To address the abovementioned inconsistency issue, we modify the loss function of the policy network shown in Eq. (6) by adding a human guidance term I:

$$\mathcal{L}^{\mu}(\Theta^{\mu})=\underset{[s_t,a_t,I(s_t)]\sim\mathcal{D}}{\mathbb{E}}\left\{-Q_1(s_t,a_t)+I(s_t)\cdot\omega_I\cdot\left[a_t-\mu(s_t|\Theta^{\mu})\right]^2\right\} \tag{8}$$

where ω_I is a factor for adjusting the weight of the human supervision loss; a_t^{DRL} in Eq. (6) can then be simply replaced with a_t, which covers both human actions and DRL policy actions. In this way, the updated direction is aligned with $\{s_t,a_t^{\text{human}}\}$ when human guidance occurs. Although this generic human-guided framework has recently been proposed in some state-of-the-art methods, there are several drawbacks in their settings; thus, further investigation and refinement are needed. For example, the conversion between the original objective and the human guidance term is conducted rigidly, and the weighting factor of the human guidance term is manually set and fixed [29,34]. However, one concern is that the weighting factor ω_I is crucial for the overall learning performance of a DRL algorithm, as it determines the degree of reliance of the learning process on human guidance. Thus, it is reasonable to design an adaptive assignment mechanism for factor ω_I that is associated with the trustworthiness of human actions. To do this, we introduce the Q-advantage as an appropriate evaluation metric, and the proposed weighting factor can be modified as follows:

$$\omega_I=\lambda^k\cdot\{\max\left[\exp\left(Q_1(s_t,a_t)-Q_1(s_t,\mu(s_t|\Theta^{\mu}))\right),1\right]-1\} \tag{9}$$

where λ is a hyperparameter that is slightly smaller than 1, and k is the index of the learning episode. The temporal decay factor λ^k indicates that the trustworthiness of human guidance decreases when the policy function gradually matures. The clip function ensures that the policy function only learns from "good" human guidance actions, and the exponential function amplifies the advantages brought by those "good" human guidance actions.

Intuitively, the adaptive weighting factor proposed above adjusts the trustworthiness of the human experience by quantitatively evaluating the potential advantages of the human's actions compared with those of the original policy. This mechanism forms the dynamic loss function of the policy network instead of a fixed learning mechanism with manually tuned weighting factors, as has been reported in existing methods [34]. Since the factor aptly distinguishes among the varying performances of different human guidance actions, the requirements for the quality of human demonstration—that is, humans' proficiency and skills—can be eased. Moreover, although the weighting mechanism involves differentiable information with respect to both the critic and actor networks, the calculation of the weighting vector does not participate in the gradient back-propagation updating of the neural networks. Therefore, it will not disturb the network training process. To the best of our knowledge, this is the first time that an updating mechanism that is adaptive to trustworthiness in human experience has been proposed in LfD/LfI-based reinforcement learning approaches. We will demonstrate its effectiveness and advantages over state-of-the-art techniques in Section 5.

Based on Eq. (9), the batch gradient of the policy network can be given by

$$\begin{aligned}\nabla_{\Theta^{\mu}}\mathcal{L}(\Theta^{\mu})=\frac{1}{N}\sum_{t=1}^{N}\Big\{&\left(-\nabla_a Q_1(s,a)|_{s=s_t,a=\mu(s_t)}\nabla_{\Theta^{\mu}}\mu(s)|_{s=s_t}\right)\\&+\left(\nabla_{\Theta^{\mu}}\left(\omega_I\cdot\|a-\mu(s)\|^2\right)|_{s=s_t,a=a_t}\right)\cdot I(s_t)\Big\}\end{aligned} \tag{10}$$

where N is the batch size sample from experience replay buffer $\mathcal{D}$.

Although the proposed objective function of the policy network looks similar to the control authority transfer mechanism of real-time human guidance shown in Eq. (2), the principles of these two stages—namely, real-time human intervention and off-policy learning—are different in the proposed method. More specifically, for real-time human intervention, the rigid control transfer illustrated by Eq. (2) enables the human's full takeover when human action occurs. For off-policy learning, we assign weighted trustworthiness to human guidance without fully discarding the agent's autonomous learning, as shown in Eqs. (8)–(10), allowing the learning process to be more robust.

Lastly, the originally stored tuple of the experience replay buffer is changed, and the human guidance component is then included as follows:

$$\mathcal{D}=\{s_t,a_t,r_t,s_{t+1},I(s_t)\} \tag{11}$$

In this way, a refactored DRL algorithm with real-time human guidance is obtained. The hyperparameters used and the algorithm procedure are provided in Table S1 and Note S1 in Appendix A, respectively.

3. Experimental design

3.1. Experiment overview

To investigate the feasibility and effectiveness of the proposed improved DRL with human guidance, a series of experiments with 40 human participants was conducted in designed autonomous driving scenarios on a human-in-the-loop driving simulator. In particular, the experimental descriptions were shown in Fig. 2, and the six scenarios utilized were provided in Fig. 3; one was for the training process of the proposed method (associated with Experiments A–E) (Table 1), and the other five were designed for testing and evaluating the performance of the designed algorithm, as illustrated in Experiment F (Table 1). The training scenario considered a challenging driving task—namely, continuous lane changing and overtaking, where the reward from the environment encouraged non-collision and smooth driving behaviors. To successfully complete the designed tasks, in all scenarios, the ego vehicle was required to start from the spawn position, stay on the road, avoid collision with any obstacles, and eventually reach the finishing line. If the ego vehicle collided with the road boundary or other traffic participants, the episode was immediately terminated and a new episode was started with new spawned vehicles to continue the training process. The types, positions, and speeds of surrounding objects varied in the testing scenarios to improve the training performance of the policies under various situations with higher requirements.

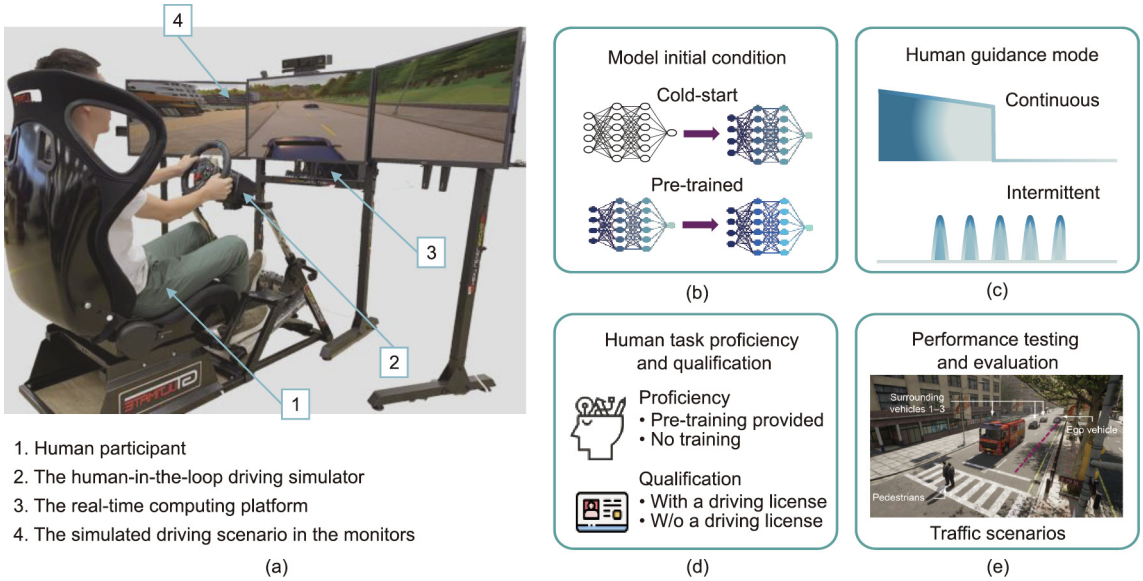

Fig. 2. Experimental setup. (a) The experimental platform used in this study was a human-in-the-loop driving simulator. Key components included a steering wheel, a real-time computation platform, three monitors, and simulated driving scenarios. (b) There were two different initial conditions of the DRL agent during training: "cold-start" and "pre-trained." The cold-start condition was used in the initial training of the DRL agent, and the pre-trained policy condition was used for evaluating the fine-tuning performance of the DRL agent. (c) Two different modes of human intervention and guidance—namely, the continuous and intermittent modes—were studied in the experiments. (d) The human task proficiency and driving qualifications were selected as two human factors studied in this work. Their impacts on the training performance of the proposed Hug-DRL method were analyzed through experiments. (e) Various driving scenarios were designed in the experiments to test the control performance of the autonomous driving policies obtained by different DRL methods. W/o: without.

To validate the improvement in the training performance, Experiment A was conducted by comparing the proposed method with other Hug-DRL approaches. First, we implemented all related baseline DRL algorithms with the same form of real-time human guidance for convenience during the comparison. More specifically, the three baseline DRL approaches were: intervention-aided DRL (IA-RL), with a fixed weighting factor ω_I for human guidance in the policy function of DRL [29,35]; human intervention DRL (HI-RL), with a shaped value function but without modification of the policy function [32]; and the vanilla DRL method (the standard TD3 algorithm without human guidance). All policy networks in these methods were pre-initialized by SL to enable faster convergence. Details on the implementation of the abovementioned approaches are provided in Section 4.

To investigate the effects of different human factors on the DRL training, we conducted Experiments B and C to respectively address two key elements: the human intervention mode and the task proficiency. Experiment B was conducted to explore how different intervention modes—that is, continuous and intermittent modes, as illustrated in Fig. 2(c)—affected the DRL training performance. The continuous mode requires more frequent human supervision and intervention than the intermittent mode, and it allows human participants to disengage from the supervision loop for a while. The contrast was expected to reveal the impacts of human participation frequency on learning efficiency and subjective human fatigue. Subjects with higher proficiency or qualifications regarding a specific task are usually expected to generate better demonstrations. Experiment C was designed to investigate this expectation and to study the correlations between human task proficiency/qualifications and DRL performance improvement, as shown in Fig. 2(d).

Despite the pre-initialization, the three experiments still started with a train-from-scratch DRL agent, denoted as "cold-start for initial training" in Fig. 2(b). However, in real-world applications such as automated driving, even if the DRL agent has been sufficiently trained beforehand, an online fine-tuning process is needed to further improve and ensure policy performance after deployment. Thus, Experiment D was designed to explore the varying effects and performance of the policies pre-trained under different algorithms throughout the fine-tuning process, as denoted by "pre-trained for fine-tuning" in Fig. 2(b). Here, "pre-trained" refers to the well-trained DRL policy rather than to the pre-initialization conducted by SL.

We also carried out an ablation study in Experiment E to investigate the effect of pre-initialization and reward shaping on DRL performance.

The abovementioned experimental arrangements (Experiments A–E) were intended to demonstrate the superiority of the proposed method over other state-of-the-art Hug-DRLs with respect to training efficiency and performance improvement. However, it is also necessary to test the performance of different policies in autonomous driving tasks under various scenarios. In addition, as imitation learning holds a great advantage in training efficiency due to non-interactive data generation, it would be interesting to compare the performances of the IL and DRL paradigms in testing. Thus, in Experiment F, we compared the driving policies obtained from the proposed Hug-DRL, the selected DRL baseline methods, and the IL methods (i.e., BC and DAgger), as illustrated in Fig. 2(e). Different performance metrics under autonomous driving, including the task-completion rate and vehicle dynamic states, were evaluated. Table 1 provides an overview of all the experiments involved in the comparison. The statistical results are presented as the mean (M) and the standard deviation (SD). The experimental results are reported below, and the detailed methodology and experimental setup can be found in Section 4 and in Appendix A.

3.2. Experimental scenarios

The human-in-the-loop driving simulator shown in Fig. 2(a) was the experimental platform used for a range of experiments in this study. Technical details and the specifications of the hardware and software are reported in Note S2 and Table S2 in Appendix A.

In total, six scenarios indexed from 0 to 5 were utilized in this study. The visualized scenarios are reported in Fig. 3. The ego

Fig. 3. Schematic diagram of the scenarios for training and testing the autonomous driving agent. (a) Scenario 0, it serves as a simple situation with all surrounding vehicles being set as stationary states. In this scenario, which was utilized only for the training stage, two pedestrians were spawned at random positions in some episodes. (b) Scenario 1, it was used to test the steady driving performance of the agent on the freeway, with the removal of all surrounding traffic participants. It was used to evaluate the anti-overfitting performance of the generated driving policy. (c–f) Scenarios 2–5, they were used to test the adaptiveness of the obtained policy in unseen situations shielded from the training stage. Moving pedestrians, motorcycles, and buses were added into the traffic scenarios. Since the interactive relationships between the ego vehicle and the traffic participants were changed, the expected trajectories of the ego vehicle were expected to differ from those in the training process. These driving conditions were set to evaluate the scene-understanding ability and the adaptiveness and robustness of the autonomous driving agent.

vehicle (i.e., the autonomous driving agent to be trained) and the surrounding vehicles and pedestrians were all spawned in a two-lane road with a width of 7 m. Scenario 0 was used for DRL training; the relative velocity between the ego vehicle and the three surrounding vehicles ($v_{ego} - v_1$) was set to 5 m·s^{-1}, and two pedestrians with random departure points in specific areas were set to cross the street. Scenarios 1–5 were used to evaluate the robustness and adaptiveness of the learned policies under different methods. More specifically, in Scenario 1, all surrounding traffic participants were removed to examine whether the obtained policies could achieve steady driving performance on a freeway. In Scenario 2, we changed the positions of all obstacle vehicles and pedestrians, and we set the relative velocity between the ego vehicle and obstacle vehicles ($v_{ego} - v_2$) to 3 m·s^{-1}, in order to generate a representative lane-change task under urban conditions for the ego vehicle. In Scenario 3, the coordinates of the surrounding vehicles were further changed to form an urban lane-keeping scenario. For Scenario 4, the relative velocities between the ego vehicle and the three obstacle vehicles were changed to $(v_{ego} - v_3) = 2$ m·s^{-1}, $(v_{ego} - v_4) = 4$ m·s^{-1}, and $(v_{ego} - v_5) = 3$ m·s^{-1}, respectively, and pedestrians were removed to simulate a highway driving scenario. In Scenario 5, we added pedestrians with different characteristics and inserted various vehicle types, including motorcycles and buses, into the traffic scenario. In all scenarios, we were able to adjust random seeds during the training and testing sessions, which would lead to reproducible comparisons across different policies.

3.3. Experimental design

3.3.1. The initial training condition

Two initial conditions were used for the model training:

Cold-start for initial training. The initial condition of training starting from scratch was denoted as "cold-start." Under this condition, the DRL agent had no prior knowledge about the environment, except for the pre-initialized training.

Pre-trained for fine-tuning. Under this condition, the initial training with the cold-start was completed by the agent under the standard DRL algorithm, and the agent was generally capable of executing the expected tasks. However, the behavior of the agent could still be undesirable in some situations; thus, the parameters of the algorithms were fine-tuned during this phase to further improve the agent's performance.

3.3.2. Human intervention activation and termination

During the experiments, the participants were not required to intervene in the DRL training at any specific time. Instead, they were required to initiate the intervention by operating the steering wheel and providing guidance to the agent whenever they felt it

Table 1
Illustration of the six experiments.

Experiment	Method	Proficient human participant	Qualified human participant	Pre-initializing trick	Reward shaping scheme	Model initial condition	Training/ testing
A	Hug-DRL	Both	Both	Y	0	Cold-start	Training
	IA-RL	Both	Both	Y	0	Cold-start	Training
	HI-RL	Both	Both	Y	0	Cold-start	Training
	Vanilla-DRL	N/A	N/A	Y	0	Cold-start	Training
B	Hug-DRL	Y	Y	Y	1	Cold-start	Training
		N	Y	Y	1	Cold-start	Training
C	Hug-DRL	Y	Y	Y	1	Cold-start	Training
		Y	N	Y	1	Cold-start	Training
D	Hug-DRL	Y	Y	N/A	0	Pre-trained	Training
	IA-RL	Y	Y	N/A	0	Pre-trained	Training
	HI-RL	Y	Y	N/A	0	Pre-trained	Training
E	Hug-DRL	Y	Y	Y	0	Cold-start	Training
		Y	Y	N	0	Cold-start	Training
		Y	Y	Y	0	Cold-start	Training
		Y	Y	Y	1	Cold-start	Training
		Y	Y	Y	2	Cold-start	Training
F	Hug-DRL	N/A	N/A	N/A	N/A	N/A	Testing
	IA-RL	N/A	N/A	N/A	N/A	N/A	Testing
	HI-RL	N/A	N/A	N/A	N/A	N/A	Testing
	Vanilla-DRL	N/A	N/A	N/A	N/A	N/A	Testing
	BC-IL	N/A	N/A	N/A	N/A	N/A	Testing
	DAgger-IL	N/A	N/A	N/A	N/A	N/A	Testing

IA-RL: intervention-aided DRL; HI-RL: human intervention DRL; 0: no shaping; 1, 2: two different reward-shaping techniques and detailed descriptions of the reward-shaping techniques are provided in Section 4; Y: yes; N: no; N/A: not applicable.

was necessary. The goal of their guidance tasks was to keep the agent on the road and try to avoid any collision with the road boundary or other surrounding obstacle vehicles. Once the human participants felt that the agent was heading in the correct direction and behaving reasonably, the participants could disengage. The detailed activation and termination mechanisms set in the experiments are explained below.

Intervention activation. If a steering angle of the handwheel exceeding five degrees was detected, then the human intervention signal was activated and the entire control authority was transferred to the human.

Intervention termination. If variation in the steering angle of the handwheel was undetected after 0.2 s, then the human intervention was terminated and full control authority was transferred back to the DRL agent.

3.3.3. The two human guidance modes

Two human guidance modes were used:

Intermittent guidance. In this mode, the participants were required to provide guidance intermittently. The entire training for a DRL agent in the designated scenario comprised 500 episodes, and human interventions were dispersed throughout the entire training process. More specifically, the participants were allowed to participate in only 30 episodes per 100 episodes, and the participants determined whether to intervene and when to provide guidance. For the rest of the time, the monitors were shut off to disengage the participants from the driving scenarios.

Continuous guidance. In this mode, the participants were required to continuously observe the driving scenario and provide guidance when they felt it was needed throughout the entire training session.

3.3.4. Human subjects' proficiency and qualifications

Human task proficiency was considered in this study. The proficiency of the participants was defined as follows:

Proficient subjects. Before the experiment, the participants were first asked to naturally operate the steering wheel in a traffic scenario on the driving simulator for 30 min to become proficient in the experimental scenario and device operation.

Non-proficient subjects. The participants were not asked to engage in the training session before participating in the experiment.

In addition to proficiency, driving qualifications were considered.

Qualified subjects. Participants with a valid driving license were considered to be qualified subjects.

Unqualified subjects. Participants without a valid driving license were regarded as unqualified subjects.

3.3.5. Experimental tasks

In this work, multiple experimental tasks were designed.

Experiment A. The purpose of this experiment was to test the performance of the proposed Hug-DRL method and compare its performance with that of the selected baseline approaches. In total, ten participants with a valid driving license were included in this experiment. Before the experiment, the participants were asked to complete a 30 min training session on the driving simulator to become proficient in the experimental scenario and device operation. During the experiment, each participant was asked to provide intermittent guidance for the proposed Hug-DRL method and baseline methods—that is, IA-RL and HI-RL. However, the participants were not informed about the different algorithms used in the tests. In addition, the vanilla-DRL method was used to conduct agent training ten times without human guidance. The initial condition of the training was set as cold-start, and the driving scenario was set as the abovementioned Scenario 0. In addition, each participant was required to complete a questionnaire after their tests to provide their subjective opinion on the workload level, which was rated on a scale from 1 (very low) to 5 (very high).

Experiment B. The purpose of this experiment was to study the impact of the human guidance modes on the agent's performance improvement for the proposed Hug-DRL method. The same ten participants recruited in Experiment A were included in this experiment. Before the Experiment B, the participants were asked to complete an additional 30 min training session on the driving simulator to become proficient in the experimental scenario and

device operation. During the experiment, each participant was asked to provide continuous guidance to the driving agent for the proposed Hug-DRL method. The initial condition of the training was set as cold-start, and the driving scenario was set as the abovementioned Scenario 0. In addition, each participant was required to complete a questionnaire after their tests to provide their subjective opinion on the workload level, which was rated on a scale from 1 (very low) to 5 (very high).

Experiment C. The purpose of this experiment was to study the impact of human proficiency and driving qualifications on the performance improvement of the proposed Hug-DRL method. Ten new subjects were recruited to participate in this experiment. Among them, five subjects holding valid driving licenses were considered to be qualified participants, and the other five participants without a driving license were considered to be unqualified participants. The participants were not provided with a training session before participating in the agent training experiment. During the experiment, each participant was asked to provide continuous guidance to the driving agent for the proposed Hug-DRL method. The initial condition of the training was set as cold-start, and the driving scenario was set as the abovementioned Scenario 0.

Experiment D. The purpose of this experiment was to study the online fine-tuning ability of the proposed Hug-DRL method and compare its fine-tuning ability to that of the selected baseline methods. In this experiment, the initial condition of the training was set as fine-tuning rather than cold-start. Fifteen new participants were recruited for this experiment. Before the experiment, the participants were provided with a 10 min training session to become acclimated to the environment and the devices. The entire fine-tuning phase comprised 30 episodes in total. During the experiment, the subjects were allowed to intervene in the agent training only in the first ten episodes, providing guidance when needed. For the next 20 episodes, the participants were disengaged from the tasks. However, the agent's actions were continually recorded to assess its performance. Each participant was asked to engage in this experiment under the proposed Hug-DRL method and the baseline methods—that is, IA-RL and HI-RL. Before the experiment, the participants were not informed about the different algorithms used in the tests. The driving scenario of this experiment was set to Scenario 0.

Experiment E. The purpose of this experiment was to test the impacts of the adopted pre-initialized training and the reward-shaping techniques on training performance. In ablation Group 1, five participants were required to complete the task in Experiment A, and the Hug-DRL agent used was not pre-trained by SL. The results were compared with those of the pre-trained Hug-DRL obtained in the training process. A similar setup was used in ablation Group 2, and the adopted Hug-DRL agents were equipped with three different types of reward schemes: no reward shaping, reward-shaping Route 1, and reward-shaping Route 2. In each subgroup experiment, five participants were asked to complete the task of Experiment A. The details of the different reward-shaping schemes are explained later in Eqs. (21) and (22).

Experiment F. The purpose of this experiment was to test and compare the performance of the autonomous driving agent trained by different methods under various scenarios. We first completed the training process of two IL-based policies—that is, BC and DAgger. The human participants were asked to operate the steering wheel, controlling the IL agent to complete the same overtaking maneuvers as the DRL agents (collision avoidance with surrounding traffic participants). For BC, the agent was fully controlled by the human participants, and there was no agent to interact with the humans through the transfer of control authority. Gaussian noise was injected into the agent's actions for the purpose of data augmentation. The collected data were used for offline SL to imitate human driving behaviors. For DAgger, the agent learned to improve its control capability from the human guidance. In one episode, whenever a human participant felt the need to intervene, he or she obtained partial control authority, and only his or her guidance actions were recorded to train the DAgger agent in real time. Since the agent was refined through the training episodes, DAgger was expected to collect more data and obtain a more robust policy than BC. The tested methods included Hug-DRL, IA-RL, HI-RL, vanilla-DRL, Dagger, and BC. The driving scenarios used in this experiment included the designed Scenarios 1–5.

3.4. Baseline algorithms

The following five baseline algorithms were compared:

Baseline A: IA-RL. In this method, human guidance was introduced into the agent-training process. The human actions directly replaced the output actions of the DRL, and the loss function of the policy network was modified to fully adapt to human actions when guidance occurred. In addition, the algorithm penalized the DRL agent in human-intervened events, which prevented the agent from getting trapped in catastrophic states. This method was derived from and named after existing work reported in Refs. [29,32] and was further modified in the present work to adapt to the off-policy actor-critic DRL algorithms. The detailed algorithm for this approach can be found in Note S3 in Appendix A, and the hyperparameters are listed in Tables S1 and S3 in Appendix A.

Baseline B: HI-RL. In this method, human guidance was introduced into the agent-training process; however, human actions were used to directly replace the output actions of the DRL agent without modifying the architecture of the neural networks. As a result, human actions affected only the update of the value network. In addition, the algorithm penalized the DRL agent in human-intervened events, which prevented the agent from getting trapped in catastrophic states. This baseline approach, which was derived from and named after the work reported in Ref. [32], was further modified to adapt the actor-critic DRL algorithm in our work. The detailed algorithm can be found in Note S4 in Appendix A, and the hyperparameters are listed in Tables S1 and S3.

Baseline C: vanilla-DRL. This standard DRL method (the TD3 algorithm) was used as a baseline approach in this work. The detailed algorithm can be found in Note S5 in Appendix A, and the hyperparameters are listed in Tables S1 and S3.

Baseline D: BC. BC with data augmentation was also adopted as a baseline method. In this study, a deep neural network with the BC method was used to develop an autonomous driving policy for comparison with other DRL-based approaches. The detailed mechanism of this method is introduced in Fig. S3 in Appendix A, and the detailed procedures of data collection and model training under BC are introduced in Note S6 in Appendix A. The hyperparameters and the network architecture are listed in Tables S1 and S3, respectively.

Baseline E: DAgger. This is an IL method with real-time Hug. Under this approach, human participants serve as experts to supervise and provide necessary guidance to an actor agent that learns from human demonstrations and improves its performance through training. The detailed mechanism of DAgger is illustrated in Fig. S4 in Appendix A. The detailed procedures of data collection and model training are introduced in Note S6 in Appendix A. The hyperparameters and the network architecture are listed in Tables S1 and S3, respectively.

4. Implementation and human-in-the-loop testing under autonomous driving tasks

4.1. Algorithm implementation for autonomous driving

The proposed Hug-DRL method is developed based on TD3 with the introduction of real-time human guidance. For the DRL

algorithm, appropriate selections of the state and action space, as well as the elaborated reward function design, are significant for efficient model training and performance achievement. In this work, the target tasks for the autonomous driving agent are set as completing lane changing and overtaking under various designed scenarios. To better demonstrate the feasibility, effectiveness, and superiority of the proposed method, a challenging end-to-end paradigm is selected as the autonomous driving configuration for the proof of concept. More specifically, non-omniscient state information is provided to the policy, and the state representation is selected for semantic images of the driving scene through a single channel representing the category of 45 × 80 pixels:

$$s_t = \left\{\mathcal{P}_{ij} | \mathcal{P} \in [0,1]\right\}_{45\times 80} \tag{12}$$

where $\mathcal{P}_{ij}$ is the channel value of pixel $i \times j$ normalized into [0, 1]. The semantic images are obtained from the sensing information provided by the simulator. A typical state variable is provided in Fig. S5 in Appendix A.

The steering angle of the handwheel is selected as the one-dimensional (1D) action variable, and the action space can be expressed as follows:

$$a_t = \{\alpha_t | \alpha \in [0,1]\} \tag{13}$$

where α is the steering wheel angle normalized into [0, 1], where the range [0, 0.5) denotes the left-turn command and (0.5, 1.0] denotes the right-turn command. The extreme rotation angle of the steering wheel is set to ±135°.

The reward function should consider the requirements of real-world vehicle applications, including driving safety and smoothness. The basic reward function is designed as a weighted sum of the metrics, which is given by

$$r_t = \tau_1 c_{\text{side},t} + \tau_2 c_{\text{front},t} + \tau_3 c_{\text{smo},t} + \tau_4 c_{\text{fail},t} \tag{14}$$

where $c_{\text{side},t}$ denotes the cost of avoiding a collision with the roadside boundary; $c_{\text{front},t}$ is the cost of collision avoidance with an obstacle vehicle to the front; $c_{\text{smo},t}$ is the cost of maintaining vehicle smoothness; $c_{\text{fail},t}$ is the cost of a failure that terminates the episode; τ_1–τ_4 are the weights of each metric.

The cost of a roadside collision is defined by a two-norm expression as follows:

$$c_{\text{side},t} = -\left|\left|1 - f_{\text{sig}}\left(\min\left[d_{\text{left},t}, d_{\text{right},t}\right]\right)\right|\right|_2 \tag{15}$$

where d_{left} and d_{right} are the distances to the left and right roadside boundaries, respectively; f_{sig} is the sigmoid-like normalization function transforming the physical value into [0, 1].

The cost of avoiding an obstacle to the front is defined by a two-norm expression:

$$c_{\text{front},t} = \begin{cases} -\left|\left|1 - f_{\text{sig}}(d_{\text{front}})\right|\right|_2, & \text{if a front obstacle exists} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

where d_{front} is the distance to the front-obstacle vehicle in the current lane.

The cost of maintaining smoothness is

$$c_{\text{smo},t} = -\left(\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} + (\alpha_t - 0.5)\right) \tag{17}$$

The cost of failure can be expressed as follows:

$$c_{\text{fail},t} = \begin{cases} -1 & \text{if fail} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

The above reward signals stipulate practical constraints. However, the feedback is still sparse and does not boost exploration behaviors, which means that the DRL could easily become trapped in the local optima. The reward-shaping technique is an effective tool to prevent this issue. Reward shaping transforms the original rewards by constructing an additional function with the aim of improving performance. We describe the three kinds of reward-shaping methods utilized in this paper and conduct an ablation study to explore their utilities in Experiment E.

First, human-intervention penalty-based reward shaping is introduced. The shaping function $\mathcal{F}^1$ is based on a typical intervention penalty function $\mathcal{F}: \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, written as follows:

$$\mathcal{F}_t^1(s_{t-1}, s_t) = -10 \cdot \{[I(s_t) = 1] \wedge [I(s_{t-1}) = 0]\} \tag{19}$$

Recall that human interventions aim to correct the DRL agent's behavior and avoid catastrophic states. Hence, this equation suggests that a penalty signal is added to the original reward when a human decides to intervene at a specific state. To pursue high cumulative rewards, the DRL agent should avoid human intervention by decreasing visits to harmful states. The intervention penalty is triggered only at the first time step when a human intervention event occurs. The rationale behind this is that, once human manipulation begins, the intervention usually lasts for at least several time steps, but only the first intervention time step can be confirmed as a participant-judged "harmful" state/behavior.

Another form of reward shaping relies on a potential function, which is well known for its straightforward and efficient implementation [39]. A typical potential-based reward-shaping function $\mathcal{F}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ can be written as follows:

$$\mathcal{F}(s_t, a_t, s_{t+1}) = \gamma\phi(s_{t+1}) - \phi(s_t) \qquad \forall s_t \in \mathcal{S} \tag{20}$$

where $\phi: \mathcal{S} \to \mathcal{A}$ is a value function, which ideally should be equal to $\mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s,a)]$. Since the accurate values of Q are intractable before training convergence, prior knowledge regarding the task requirement becomes a heuristic function ϕ to incentivize the DRL's exploration. Accordingly, the function $\mathcal{F}^2$ is defined to be associated with the longitudinal distance from the spawn point, which can be calculated as follows:

$$\mathcal{F}_t^2 = P_{y,t}(s_t, a_t) - P_{y,\text{spawn}} \tag{21}$$

where $P_{y,t}$ and $P_{y,\text{spawn}}$ are the current and initial positions of the agent in the longitudinal direction, respectively. This indicates that the agent is encouraged to move forward and explore further, keeping itself away from the spawn position.

The last reward-shaping method is a state-of-the-art technique named Never Give Up (NGU) [40]. Its main idea is also to encourage exploration and prevent frequent visits of previously observed state values.

$$\mathcal{F}_t^3 = r_t^{\text{episode}} \cdot \min\left\{\max\left\{1 + \frac{\|f(s_{t+1}|\psi) - f(s_{t+1})\| - \mathbb{E}[f(s_{t+1}|\psi)]}{\sigma[f(s_{t+1}|\psi)]}, 1\right\}, L\right\} \tag{22}$$

where $f(\cdot|\psi)$ and $f(\cdot)$ are embedded neural networks with fixed weights ψ and adjustable weights, respectively; The norm $\|\cdot\|$ is used to calculate the similarity between the embedded state feature; σ denotes the SD operation; and L is a regularization hyperparameter. The overall idea of employing $f(\cdot)$ is to assign higher additional rewards to unvisited states, particularly during the training process (see Ref. [40] for details). r_t^{episode} also encourages exploration in unvisited states, particularly during the current episode. The utilized hyperparameters are provided in Table S3.

Thus, the overall reward function can be obtained by adding the terms $\mathcal{F}_t^1$, $\mathcal{F}_t^2$, and $\mathcal{F}_t^3$ to the original function r_t. Finally, the termination of an episode with successful task completion occurs when the last obstacle vehicle is passed and the finishing line is reached without any collisions. With the above steps, the detailed

implementation of the standard DRL in the designed driving scenario is completed.

For the proposed Hug-DRL, real-time human guidance is achieved by operating the steering wheel in the experiments. Therefore, the steering angle of the handwheel is used as the human intervention signal, and a threshold filtering unexpected disturbance is required. Here, the event of human intervention and guidance is

$$I(s_t)=\begin{cases}1, & \text{if} \quad \left(\dfrac{\mathrm{d}\alpha_t}{\mathrm{d}t}>\varepsilon_1\right)\cap \text{not}\ \ q\\ 0, & \text{otherwise}\end{cases} \tag{23}$$

where ε_1 is the threshold, set as 0.02; q denotes the detection mechanism of human intervention termination, which is defined as follows:

$$q=\prod_{t}^{t+t_{\mathrm{N}}}\left(\frac{\mathrm{d}\alpha_t}{\mathrm{d}t}<\varepsilon_2\right) \tag{24}$$

where ε_2 is the threshold, set to 0.01; t_{N} is the time-step threshold for determining the intervention termination, and it is set to 0.2 s, as mentioned above.

For the proposed Hug-DRL method, when human participants engage in or disengage from the training process, the control authority of the agent is transferred between the human and the DRL algorithm in real time. The detailed mechanism of control transfer is illustrated in Eq. (2).

4.2. Participants in the human-in-the-loop tests

In total, 40 participants (26 males and 14 females) ranging in age from 21 to 34 (M_{age} = 27.43, $\mathrm{SD}_{\mathrm{age}}$ = 3.02) (where M_{age} is mean value of age and $\mathrm{SD}_{\mathrm{age}}$ is standard deviation of age) were recruited for the experiments. The study protocol and consent form were approved by the Nanyang Technological University Institutional Review Board, protocol number IRB-2018-11-025. All research was performed according to relevant guidelines/regulations. Informed consent was obtained from all participants. All participants had no previous knowledge of the research topic and had never previously experienced real-time intervention or guidance during model training in a driving scenario. Before the experiments, the participants were informed that the DRL agent would receive their guidance and improve its performance over the course of the training process.

4.3. Statistical analysis approach

4.3.1. Statistical methods

A statistical analysis of the experimental data was conducted for the designed experiments in MATLAB (R2020a, MathWorks, USA) using the Statistics and Machine Learning Toolbox and in Microsoft Excel. The data generally obeyed a normal distribution; thus, the difference in the mean values between two groups was determined using paired t-tests (with the threshold level α = 0.05), and the difference for multiple groups was determined using one-way analysis of variance (ANOVA). To investigate the statistical significance of the difference between groups in Fig. 4, non-parametric tests, including the Mann–Whitney U-test and the Kruskal–Wallis test, were adopted with the threshold selection of α = 0.05.

4.3.2. Definition of evaluation metrics

The following metrics were adopted in this study to evaluate the agent's performance. The reward, which reflected the agent's performance, was chosen as the first metric. For both the step reward and the episodic reward, the mean and SD values were calculated and used when evaluating and comparing the agent's performance across different methods and different conditions throughout the paper. The length of the episode, which was obtained by calculating the number of steps in one episode, was also selected as an evaluation metric to reflect the current performance and learning ability of the agent. Another adopted metric was the intervention rate, which reflected the frequency of human intervention and guidance. The intervention rate could be represented in two ways: count by episode and count by step. The former was calculated based on the total number of steps guided by a human in a specific episodic interval, and the latter was calculated based on the number of episodes in which a human intervened. The success rate was defined as the percentage of successful episodes within the total number of episodes throughout the testing process. The vehicle dynamic states, including the lateral acceleration and the yaw rate, were selected to evaluate the dynamic performance and stability of the agent vehicle.

5. Results

5.1. The improved training performance of the proposed Hug-DRL method

The results shown in Fig. 4, which were obtained from Experiment A, validate the performance improvement brought by the proposed Hug-DRL method compared with the other state-of-the-art Hug algorithms—namely, IA-RL and HI-RL—and compared with vanilla-DRL without human guidance (a pure TD3 algorithm). During the experiments, the time-step reward and duration of each episode were recorded and assessed for each participant in order to evaluate the training performance throughout an entire training session under each method. Both the episodic reward and the length of the episode were evaluated, as reflected in Figs. 4(a) and (b). The results indicated that the Hug-DRL method was advantageous over all other baseline methods with respect to asymptotic rewards and training efficiency. The statistical results shown in Fig. 4(c) demonstrate that the average reward obtained with the proposed method during the entire training process was the highest (M_{r} = −0.649, SD_{r} = 0.036) (where M_{r} and SD_{r} are the mean value and standard deviation of the average reward, respectively), followed by that obtained with the HI-RL method (M_{r} = −0.813, SD_{r} = 0.434), the IA-RL method (M_{r} = −0.954, SD_{r} = 0.456), and then the vanilla-DRL method (M_{r} = −1.139, SD_{r} = 0.567). In addition, the differences between the methods were tested according to the one-way ANOVA presented in Table S4 in Appendix A. The length of the episode, which accurately described task-completion ability, was also compared for the three methods. Based on the results shown in Fig. 4(d), the mean value of the proposed method (M_{l} = 93.1, SD_{l} = 2.4) (where M_{l} and SD_{l} are the mean value and standard deviation of the length of the episode, respectively) was advantageous over that of the IA-RL method (M_{l} = 83.2, SD_{l} = 12.7), the HI-RL method (M_{l} = 75.8, SD_{l} = 5.5), and the vanilla-DRL method (M_{l} = 44.3, SD_{l} = 16.8). Their differences were statistically significant, with $F(4, 36)$ = 36.91, as reflected by the ANOVA presented in Table S5 in Appendix A. In terms of asymptotic rewards, compared with vanilla-DRL, the performance improvements under Hug-DRL, IA-RL, and HI-RL were 34.4%, 10.1%, and 20.9%, respectively. To evaluate the computational efficiency, we took the asymptotic performance as the evaluation parameter and compared the proposed Hug-DRL with other baseline methods. More specifically, to reach the same asymptotic average reward achieved by IA-RL, Hug-DRL only needed 171 episodes, improving the efficiency by 192.4%. Furthermore, Hug-DRL was able to reach the same asymptotic length realized by IA-RL in 276 episodes, improving the efficiency by 81.1%. In comparison with vanilla-DRL, the improvements provided by Hug-DRL were 276.0% and 963.8%, in terms of the asymptotic average reward and asymptotic length of the episode, respectively. Taken

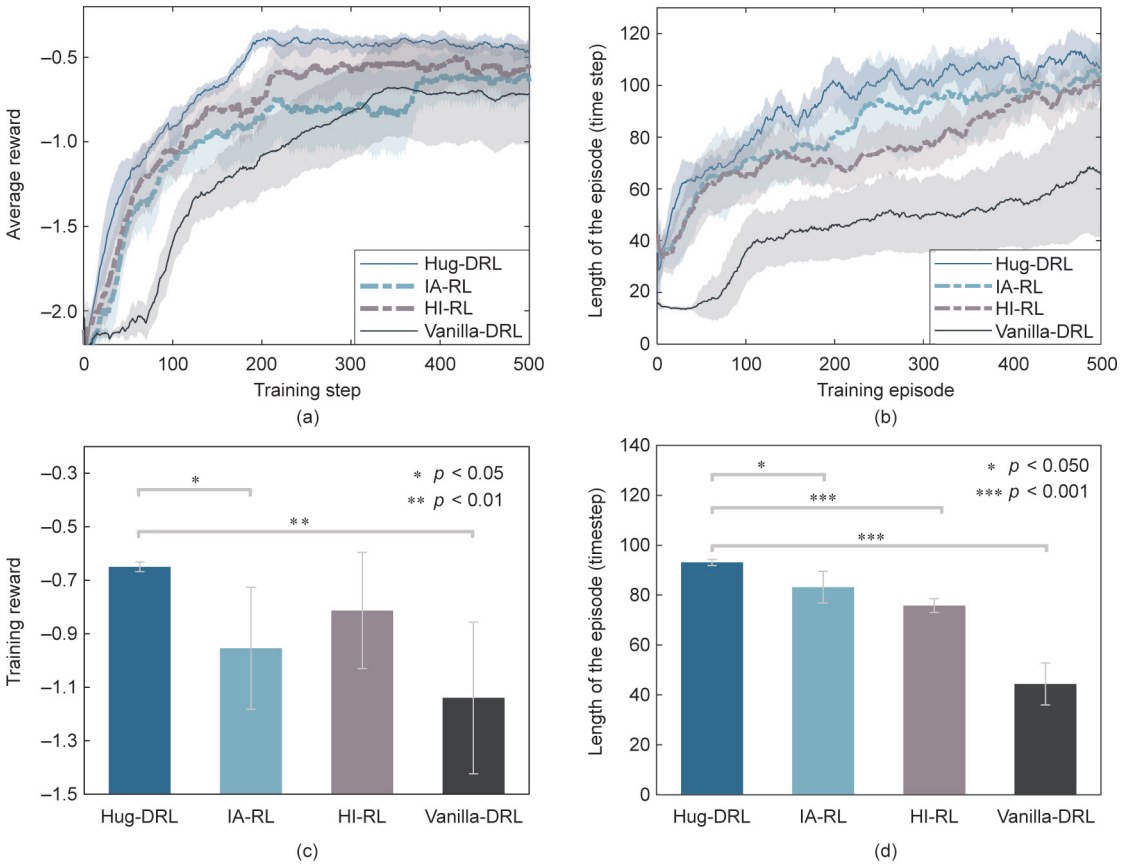

Fig. 4. Results of the initial training performance under four different methods. (a) Results of the episodic training reward under different methods, the mean and SD values of the episodic training reward were calculated based on the values of the obtained rewards per episode across all subjects under each method; (b) results of the episodic length under the three methods, the mean and SD values of the episodic step length were calculated based on the values of the episodic length achieved per episode across all subjects under each method; (c) results of the average reward during an entire training session under different methods, the statistical values of the training reward were calculated based on the average value of the obtained rewards during the overall training process across all subjects under each method; (d) results of the average episodic length during the entire training session under different methods, the statistical values of the episodic length were calculated based on the average value of the achieved episodic length during the overall training process across all subjects under each method. *p*: the value indicating the probabilistic significance in the *t*-test.

together, these results demonstrate the effectiveness of human guidance in improving DRL performance.

5.2. The effects of different human guidance modes on training performance

We conducted two groups of tests, requiring each human subject to participate in the DRL training using intermittent and continuous intervention modes (refer to Section 4 for a detailed explanation). Example data on the episodic rewards throughout the training session for the continuous and intermittent guidance modes obtained from a representative participant are shown in Figs. 5(a) and (b). The results show that both the continuous and intermittent modes led to a consistently increasing trend for the episodic reward during training. Although the episodic reward increased earlier in the former mode, as the human intervened more frequently in the initial training phase, the final rewards achieved were at the same level for both modes. The human intervention rates during the entire training session for the continuous and intermittent guidance modes were further investigated, as shown in Figs. 5(c) and (d). The mean values of the intervention rates (count by step) across participants for the continuous and intermittent modes were M_i = 25.0%, SD_i = 8.3%, and M_i = 14.9%, SD_i = 2.8%, respectively (where M_i and SD_i are the mean value and standard deviation of the mean value of the intervention rate, respectively). Moreover, we split one training process into three separate sections—namely, the human-guided section, the non-guided section, and the overall section—and the achieved rewards were examined for each section in detail for the two intervention modes separately. As illustrated in Fig. 5(e), within the human intervention sections, the mean values of the training rewards for the continuous and intermittent modes were M_r = −0.03, SD_r = 0.41, and M_r = 0.07, SD_r = 0.25, respectively, but no significant difference was found between the two (p = 0.85). Similarly, for the non-intervention sections, although the average reward of the continuous mode (M_r = −0.26, SD_r = 0.18) was higher than that of the intermittent mode (M_r = −0.42, SD_r = 0.14), no significant difference was found (p = 0.064). These results indicated that, in terms of the final DRL performance improvement, there was no significant difference between the continuous and intermittent modes of human guidance. However, from the perspective of human workload, the intermittent mode was advantageous over the continuous mode, according to our subjective survey administered to participants (Fig. S6 and Table S6 in Appendix A).

5.3. Effects of human proficiency/qualifications on training performance

Task proficiency or qualifications are other human factors that may have affected DRL training performance under human guidance. Experiment C was conducted to examine the correlations between the improvement of DRL performance and task proficiency/qualifications. As shown in Figs. 5(f) and (g), the agent training rewards achieved by proficient/non-proficient and qualified/

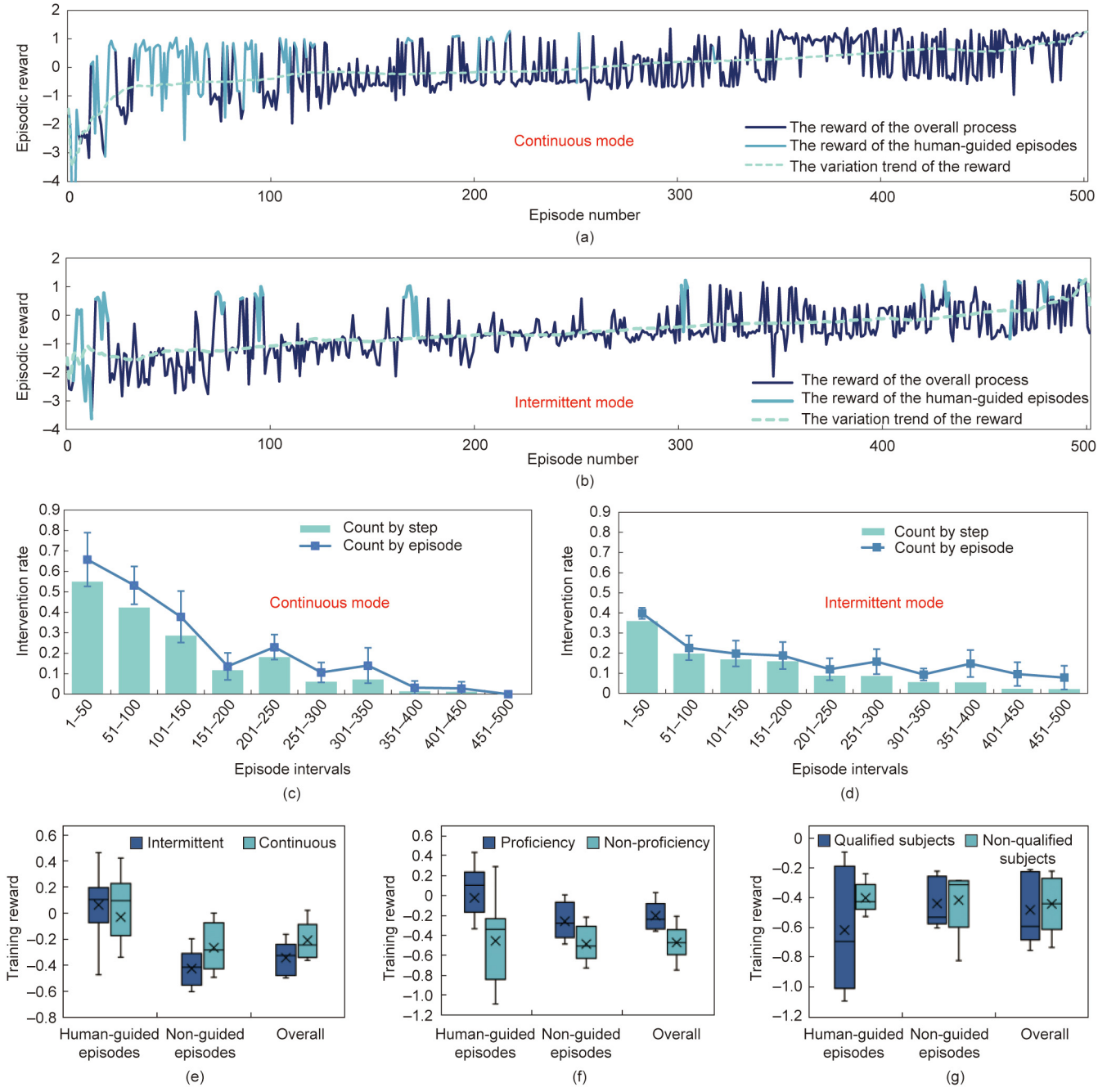

Fig. 5. Results of the impacts of human factors on DRL training performance. (a) Example data of the episodic rewards over the entire training session for the continuous guidance mode obtained by a representative subject. The human-guided episodes were mainly distributed in the first half of the training process, and the guidance actions were relatively continuous. (b) Example data of the episodic reward over the entire training session for the intermittent guidance mode obtained by a representative subject. The human-guided episodes were sparsely distributed throughout the entire training session. (c) The human intervention rates during the entire training sessions for the continuous guidance mode. Here, two indicators—namely, "count by step" and "count by episode"—were adopted to evaluate the human intervention rate. The former was calculated based on the total number of steps guided by a human in a specific episodic interval, whereas the latter was calculated based on the number of episodes intervened in by a human. (d) Human intervention rates during the entire training session for the intermittent guidance mode. (e) Box plots of the training rewards achieved under the intermittent and continuous guidance modes. Under each mode, the training rewards were further analyzed separately based on the human-guided episodes, non-guided episodes, and the entire process. (f) Box plots of the training rewards achieved under the guidance provided by proficient and non-proficient participants. (g) Box plots of the training rewards achieved under the guidance provided by qualified and unqualified participants.

unqualified participants were illustrated and compared. In the intervention sections, proficient participants guided the DRL agent to gain a higher reward ($M_r = -0.03$, $SD_r = 0.41$) than non-proficient participants ($M_r = -0.46$, $SD_r = 0.42$). For the non-intervention sections, the values of the average rewards under the guidance of proficient and non-proficient subjects were $M_r = -0.26$, $SD_r = 0.18$, and $M_r = -0.49$, $SD_r = 0.18$, respectively. In the overall training sessions, although there was a slight difference between the two groups with respect to the training reward (i.e., $M_r = -0.21$, $SD_r = 0.14$ for the proficient group and $M_r = -0.48$, $SD_r = 0.17$ for the non-proficient group), no significant difference was found between the two based on a within-group comparison ($p = 0.11$). Tables S7 and S8 in Appendix A present a non-parametric ANOVA of the performance resulting from the standard DRL method and from proficient/non-proficient participants of the proposed Hug-DRL method. In addition, no significant difference was found between the results of

qualified and unqualified participants. These comparison results indicate that the proposed real-time human guidance-based method has no specific requirement for task proficiency, experience, or qualifications of the participating human subjects.

5.4. The improved online fine-tuning performance of the Hug-DRL

As validated by the above exploration, the proposed real-time human guidance approach was capable of effectively improving DRL performance under the initial condition of a "cold-start." Subsequently, it was very interesting to conduct Experiment D to explore the online fine-tuning ability of the proposed method, which further improved the agent's performance. The online training performance is demonstrated in Fig. 6. As shown in the representative examples in Fig. 6(a), in the experiments, the participants were asked to provide guidance whenever they felt it was necessary within the first ten training episodes of the fine-tuning phase, helping the agent to further improve the driving policy online.

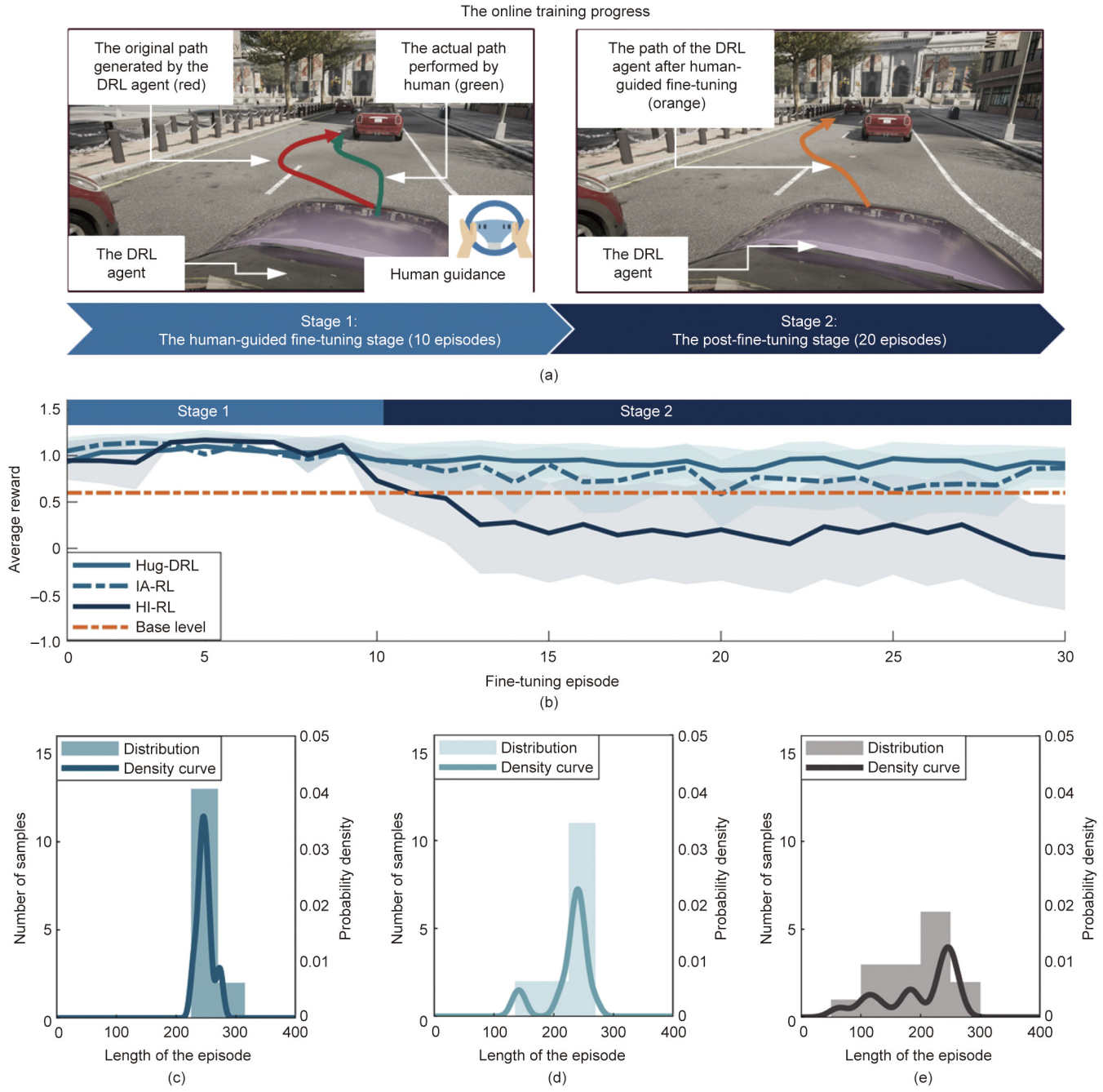

Fig. 6. Results of the online training performance of the DRL agent under the proposed method. (a) Schematic diagram of the agent performance during the online training progress under the proposed Hug-DRL method. The entire online training progress was divided into two stages: Stage 1, a ten-episode human-guided fine-tuning stage; and Stage 2, a 20-episode non-guided post-fine-tuning stage. During fine-tuning, some undesirable actions of the agent were further optimized by human guidance. As a result, the performance of the DRL agent was further improved, which was reflected by the generated smooth path in the post-fine-tuning stage. (b) The results of the episodic reward during the online training process under the proposed and two baseline approaches. Before fine-tuning, the DRL agent was pre-trained in the training Scenario 0, and the average reward achieved after the pre-training session was set as the base level for comparison in the fine-tuning stage. (c) Distribution of the episodic length obtained under the proposed Hug-DRL method across participants during the post-fine-tuning stage. (d) Distribution of the episodic duration obtained under the baseline IA-RL method across participants during the post-fine-tuning stage. (e) Distribution of the episodic duration obtained under the baseline HI-RL method across participants during the post-fine-tuning stage.

Afterward, the DRL agent continued the remaining 20 episodes until the end of the online training session. In this experiment, the proposed Hug-DRL method was compared with the other two Hug approaches—namely, IA-RL and HI-RL. Based on the performance shown in Fig. 6(b), in the fine-tuning stage, the proposed method and the baseline methods achieved similar episodic rewards (proposed method: M_r = 1.02, SD_r = 0.36; IA-RL: M_r = 1.06, SD_r = 0.08; HI-RL: M_r = 1.03, SD_r = 0.10). However, in the session after human-guided fine-tuning, the average reward of the proposed method (M_r = 0.92, SD_r = 0.35) was higher than that of IA-RL (M_r = 0.76, SD_r = 0.50) and much higher than that of HI-RL (M_r = 0.19, SD_r = 1.01). Moreover, the results shown in Figs. 6(c) and (e) show that the distribution of the episodic length obtained after fine-tuning under the proposed Hug-DRL method was more concentrated than that under the two baseline methods. The mechanism for the better performance of Hug-DRL and IA-RL compared with that of HI-RL after fine-tuning was also analyzed, as illustrated in Fig. S7 in Appendix A. In short, although the evaluation curve of the value network was updated by the human guidance action during fine-tuning, the policy network of HI-RL fell into the local optima trap during the post-fine-tuning stage, failing to converge to the global optima (Figs. S7(a)–(c)). Hug-DRL and IA-RL could successfully solve this issue (Figs. S7(d)–(f)), and Hug-DRL achieved a better performance than IA-RL. Overall, the above results indicate that the proposed method has a higher ability to fine-tune the DRL agent online than the other state-of-the-art Hug-DRL methods. More detailed illustrations regarding this observation are provided in Section 6.

5.5. Testing the autonomous driving policy trained by Hug-DRL under various scenarios

To construct and optimize the configuration of the DRL-based policy, an ablation test was carried out in Experiment E to analyze the significance of the pre-initialization and reward-shaping techniques. According to the results shown in Fig. S8(a) in Appendix A, we confirmed that the removal of the pre-initialization process led to deterioration in the training performance of the DRL agent (length of episode: M_l = 93.1, SD_l = 2.44 for the pre-initialization scheme, M_l = 84.8, SD_l = 4.8 for the no-initialization scheme, $p < 0.001$). We also found that different reward-shaping mechanisms had varying effects on performance, based on the results in Figs. S8(b)–(f).

Finally, to further validate feasibility and effectiveness, in Experiment F, the trained model for the proposed method was tested in various autonomous driving scenarios (introduced in Fig. 3 in detail) and compared with five other baseline methods: IA-RL, HI-RL, vanilla-DRL, BC (Fig. S3), and DAgger (Fig. S4). Various testing scenarios were designed to examine the abilities of the learned policy, including environmental understanding and generalization.

The success rate of task completion and the vehicle dynamic states (i.e., the yaw rate and lateral acceleration) were selected as evaluation parameters to assess the control performance of the autonomous driving agent. The heat map provided in Fig. 7(a) shows that the agent trained by Hug-DRL successfully completed tasks in all untrained scenarios, while agents under all baseline methods could complete only parts of the testing scenarios. More specifically, the success rates of the baseline methods were 84.6% for vanilla-DRL and DAgger, 76.9% for HI-RL, 73.1% for BC, and 65.3% for IA-RL. In addition, the yaw rate and lateral acceleration of the agent for each method under Scenario 1 were recorded and assessed, as shown in Fig. 7(b). Hug-DRL led to the smoothest driving behavior, with an acceleration of 0.37 $m{\cdot}s^{-2}$, and HI-RL resulted in the most unstable driving behavior (1.85 $m{\cdot}s^{-2}$). The performances of the other baseline methods were roughly similar.

In addition to performing the above investigations, it was of interest to explore the decision-making mechanism of Hug-DRL. One representative example of a testing scenario with a trained Hug-DRL agent is shown in Fig. 7(c), which provides a schematic diagram of the scenario, the lateral position of the ego vehicle over time, the values given the current state and action, and the action of the agent. As shown in Fig. 7(c), approaching two motorcycles would cause a two-fold decrease in the Q value in the current state if the current action were maintained, indicating a higher potential risk. Correspondingly, the ego agent would change its action to avoid the objects and drive slightly to the left. Subsequently, the collision risk with the front bus increased, as reflected by the remarkably decreased Q value, and the DRL agent promptly decided to change lanes. These results show the effects of varying surrounding traffic participants on the decision-making process of the DRL agent, and the intention and reasonable actions of the agent are reflected in the results of the value evaluation function.

6. Discussion

The existing training process of DRL-based policy is very time-consuming and demands many computing resources, especially when dealing with complex tasks with high-dimensional data for scene representation. To address these limitations and further improve DRL algorithms by leveraging human intelligence, a novel human-in-the-loop DRL framework with human real-time guidance is proposed and investigated from different perspectives in this study. In addition to the proposed Hug-DRL approach, two baseline methods with different real-time human guidance mechanisms are implemented and compared, along with non-human-involved algorithms. As reflected by the results shown in Fig. 3, all human-involved DRL methods were found to be advantageous over the vanilla-DRL method in terms of training efficiency and reward achieved, demonstrating the necessity and significance of real-time human supervision and guidance in the initial training stage.

The reason why the introduction of real-time human guidance can effectively improve DRL performance should be discussed. For actor-critic DRL algorithms, actions are determined by the policy function, where the update optimizes the value function, as expressed in Eq. (6). Thus, the updating rate of the policy network is constrained by the convergence rate of the value function, which relies on a relatively low-efficiency exploration mechanism. In contrast, from the perspective of human beings, who hold prior knowledge and a better understanding of the situation and the required task, this learning is clumsy, because the agent has to experience numerous failures during explorations before gradually reaching feasible solutions. This constitutes the "cold-start" problem. However, in all human-involved DRL methods, random and unreasonable actions are replaced with appropriate human guidance actions. Consequently, more reasonable combinations of states and actions are being fed to the value network, effectively improving the distribution of the value function and its convergence toward the optimal point in a shorter time. Therefore, the updating of the value network becomes more efficient, accelerating the entire training process.

With regard to the three human-involved DRL approaches, the proposed Hug-DRL approach achieves the best training efficacy and asymptotic performance; IA-RL performs second best, and HI-RL performs the worst. The underlying reason for these results is the human guidance term of Hug-DRL and IA-RL (Eq. (8)). More specifically, in addition to the action replacement scheme in HI-RL, the human guidance term directly encourages the policy network to output human-like actions, which accelerates the value function's evaluation of acceptable policies. The subsequent

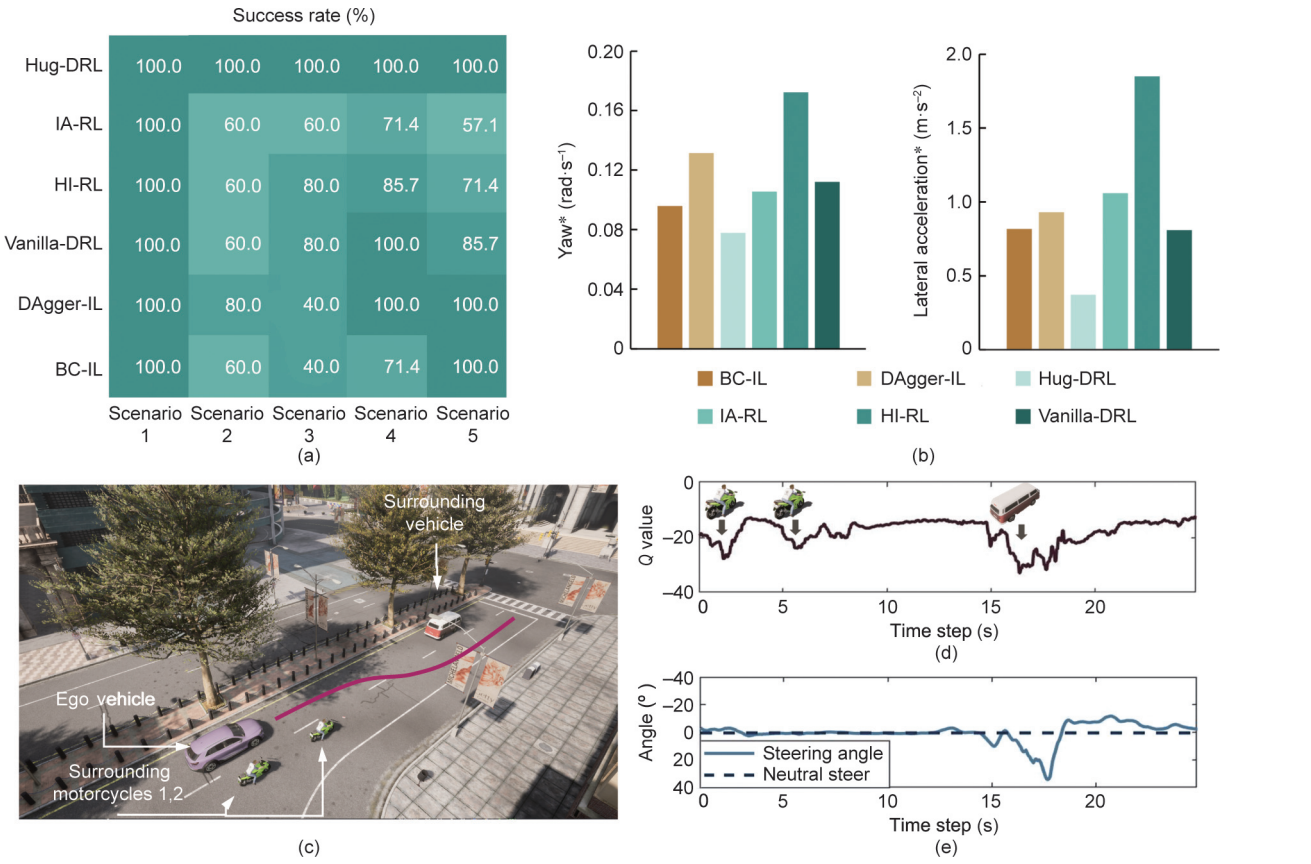

Fig. 7. Results of the agent's performance under various driving scenarios. The agent's policy was trained by the six methods, separately. The five scenarios—that is, Scenarios 1–5—were unavailable in the training process and were only used for the performance testing. (a) Success rates of the agent trained by different methods across the five testing scenarios, where the ego vehicle was spawned in different positions to calculate the success rate in one scenario; (b) plots of the mean of the agent's indicators under different scenarios, where two indicators (the mean of the absolute value of the yaw rate and the mean of the absolute value of the lateral acceleration) were recorded as indicators; (c) illustration of a representative testing scenario with an agent that was trained beforehand using the Hug-DRL; in the testing scenario, the agent was required to surpass two motorcycles and a bus successively; (d) averaged Q value of the DRL agent in the testing scenario (c), the value declined when the DRL agent approached the surrounding obstacles; (e) variation of the control action—that is, the steering wheel angle of the DRL agent in the testing scenario (c); negative values represent left steering and positive values correspond to right steering actions.

problem becomes how to balance human guidance and the policy gradient-based updating principle. The competing methods either shield the gradient term whenever humans provide guidance or pre-set a fixed ratio between two terms. These methods fail to consider the effect of different human participants and the ever-improving ability of the DRL agent. In the proposed Hug-DRL method, the weighting assignment mechanism adaptively adjusts the dynamic trustworthiness of the DRL policy against different human guidance in the training process. In comparison with the stiff conversion mechanism of the IA-RL baseline method, Hug-DRL leverages human experience more reasonably and scores higher, as shown in Fig. 4.

In addition to demonstrating performance improvement during the training-from-scratch process, Hug-DRL proved beneficial with respect to its online fine-tuning ability. For learning-based approaches, including DRL, even if the models are well trained, their performance is compromised in real-world implementations due to unpredictable and uncertain environments. Thus, an online fine-tuning process after deployment is of great importance for DRL applications in the real world. In this study, we evaluated the fine-tuning performance of all three methods—that is, Hug-DRL, IA-RL, and HI-RL—involving human guidance. As shown in the subplots of Figs. 6(b)–(e), the performance improvement of HI-RL vanished throughout the fine-tuning. However, our approach successfully maintained the improved performance throughout the post-fine-tuning phase, indicating its higher ability. This phenomenon may be explained by the consistency of the updates between the policy and value networks under human guidance. For the HI-RL model that receives human guidance, its policy network is updated according to the objective function with $\{s, \mu(s|\Theta^{\mu})\}$ in Eq. (6). However, the value network is constructed according to $\{s, a^{\text{human}}\}$, as expressed by Eq. (7). In general, a human guidance action generates a higher true value, but the action is not correctly evaluated by the value network before fine-tuning. As online fine-tuning progresses, the value network realizes the deficiency and gradually updates its output. However, the policy function sometimes struggles to catch up with the pace of the policy network's update. As a result, even if the policy network has already converged toward a local optimum in the initial training phase, the change of a single point on the value function distribution that benefited from human guidance does not optimize the gradient descent-based policy function. Accordingly, the policy still updates the function around the original local optima and thus fails to further improve itself in the expected direction. The inconsistency between the policy and value networks can be observed from the results shown in Fig. S7. Notably, this inconsistency problem rarely occurs in the training-from-scratch process due to the high adaptivity of the value network.

To solve the inconsistency issue described above, modified policy functions were proposed in Hug-DRL and IA-RL. By dragging the policy's outputs, the effect of the policy-gradient-based update was weakened in the human-guided steps, which avoided the issue of

the local optima trap. Thereafter, the policy could continue the noise-based exploration and gradient-based update in a space closer to the global optima. Theoretically, the inconsistency issue that occurred in HI-RL could be addressed by Hug-DRL and IA-RL. However, we found from the experimental results that IA-RL failed to achieve the expected competitive performance, mainly due to the different forms of human guidance. In general, the reinforcement learning agent achieves an asymptotic performance by means of large-scale batch training with the experience replay buffer. However, fine-tuning is essentially a learning process with small-scale samples. Thus, it is very difficult for IA-RL to find an appropriate learning rate in this situation, resulting in an unstable fine-tuning performance. The weighting factor in the proposed Hug-DRL can automatically adjust the learning rate and mitigate this issue, hence achieving the best performance, as shown in Fig. 6.

In addition to the training performance discussed above, the ability and superiority of the proposed method were validated in testing scenarios in comparison with other baseline approaches. More specifically, we tested the effectiveness, adaptiveness, and robustness of the proposed Hug-DRL method under various driving tasks and compared the method with all related DRL baseline methods, as well as BC and DAgger. The results regarding the success rate across various testing scenarios, as shown in Fig. 7(a), reflect the adaptiveness of these methods. The proposed Hug-DRL achieved the best performance of all methods across all testing scenarios. The success rates of the IL approaches were significantly affected by variations in the testing conditions, while the DRL methods maintained their performance and thus demonstrated better adaptiveness. Meanwhile, DAgger outperformed BC; its performance was similar to that of vanilla-DRL but lagged behind that of Hug-DRL. In terms of success rate, IA-RL and HI-RL performed worse than vanilla-DRL; this result differed from the previously observed results in the training process. A feasible explanation is that undesirable actions by human beings interrupted the original training distribution of the DRL and accordingly deteriorated the robustness. Similarly, according to the results shown in Fig. 7(b), the average yaw rate and lateral acceleration of IA-RL and HI-RL were higher than those of vanilla-DRL, indicating their worse performance in motion smoothness. Hug-DRL achieved the highest performance, which demonstrates that, beyond accelerating the training process, the proposed human guidance mechanism can achieve an effective and robust control performance during the testing process.

The proposed Hug-DRL method was also investigated from the perspective of human factors. Real-time human guidance has proven effective for enhancing DRL performance; however, long-term supervision may also have negative effects, such as fatigue, on human participants. Fortunately, the results shown in Fig. 5(e) demonstrate that the intermittent guidance mode did not significantly deteriorate performance improvement compared with the continuous mode. In addition, the participants' subjective feelings on task workload under intermittent guidance were satisfactory, according to the survey results shown in Fig. S6. These results suggest that, within the proposed human-in-the-loop DRL framework, human participants do not necessarily remain in the control loop constantly to supervise agent training. Intermittent guidance is a good option that generates satisfactory results for both agent training performance and human subjective feelings.

We were also curious about whether the proposed Hug-DRL method relied heavily on participants' proficiency, skills, experience, or qualifications with respect to a specific task. As the DRL performance improvement results illustrate in Fig. 5(d), there was no significant difference between the proficient and non-proficient participant groups. This observation can be reasonably explained by the mechanism of the proposed algorithm. Assume that a standard DRL agent is in a specific state, and noise-based exploration can be effective only within a certain area close to the current state. Thus, the distribution is modified progressively and slowly based on the gradient update of the neural networks, which are far from convergent. However, in the designed Hug-DRL method, human guidance actions can facilitate the update of the distribution to be much more efficient. Thereafter, even if the guidance actions input from non-proficient participants are undesirable, the explorations leveraging human guidance are still more efficient than those in the standard DRL method. Video S1 in Appendix A provides a representative example of the exploration processes under the Hug-DRL and standard DRL methods, further illustrating the above opinion. Similar results can also be found in Figs. 5(f) and (g), where there are no significant differences between the two participant groups with and without a driving license with respect to the achieved reward. These findings provide us with more confidence that the proposed Hug-DRL method poses no high requirements for the quality of data associated with humans' experience, proficiency, or task qualifications.

7. Conclusions

In this study, a real-time Hug-DRL method was developed for policy training in an end-to-end autonomous driving case. An improved actor-critic architecture with a modified policy and value networks was developed. Humans could intervene and correct the agent's unreasonable actions of DRL in real time during the training process. The developed method was validated by human-in-the-loop experiments with 40 subjects and was compared with other state-of-the-art learning approaches.

The experimental results suggest that the proposed Hug-DRL is advantageous over existing methods in terms of learning efficiency and testing performance. The proposed method can effectively improve the agent's training performance in both the initial training and online fine-tuning stages. Intermittent human guidance can be a good option to generate satisfactory results for DRL performance improvement; at the same time, it exerts no substantial burden on human workload. In particular, the proposed method largely reduces the requirements on the human side. Participating subjects do not need to be experts with a mastery of skilled knowledge or experience in specific areas. As long as they are able to perform normally with common sense, the DRL can be well trained and effectively improved, even if humans' actions are undesirable. These factors make the proposed approach very promising in future real-world applications. The high-level framework, the methodology employed, and the algorithms developed in this work have great potential to be expanded to a wide range of AI and human–AI interaction applications.

Acknowledgments

This work was supported in part by the SUG-NAP Grant of Nanyang Technological University and the A*STAR Grant (W1925d0046), Singapore.

Author contributions

Jingda Wu developed the algorithms, designed and performed experiments, processed and analyzed data, interpreted results, and wrote the paper. Zhiyu Huang performed experiments, analyzed data, interpreted results, and wrote the paper. Zhongxu Hu designed and developed the experimental platforms and analyzed data. Chen Lv supervised the project, designed the system concept, methodology and experiments, analyzed data, interpreted results, and led the writing of the paper. All authors reviewed the manuscript.

Compliance with ethics guidelines

Jingda Wu, Zhiyu Huang, Zhongxu Hu, and Chen Lv declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.05.017.

References

[1] Stilgoe J. Self-driving cars will take a while to get right. Nat Mach Intell 2019;1(5):202–3.
[2] Mo X, Huang Z, Xing Y, Lv C. Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. IEEE Trans Intell Transp Syst. In press.
[3] Huang Z, Wu J, Lv C. Efficient deep reinforcement learning with imitative expert priors for autonomous driving. IEEE Trans Neural Netw Learn Syst. In press.
[4] Feng S, Yan X, Sun H, Feng Y, Liu HX. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nat Commun 2021;12:748.
[5] Codevilla F, Müller M, López A, Koltun V, Dosovitskiy A. End-to-end driving via conditional imitation learning. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA); 2018 May 21–25; Brisbane, QLD, Australia. IEEE; 2018. p. 4693–700.
[6] Huang Z, Wu J, Lv C. Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning. IEEE Trans Intell Transp Syst. In press.
[7] Codevilla F, Santana E, López AM, Gaidon A. Exploring the limitations of behavior cloning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. IEEE; 2019. p. 9329–38.
[8] Ross S, Gordon GJ, Bagnell JA. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS); 2011 Apr 11–13; Fort Lauderdale, FL, USA. PMLR; 2011. p. 627–35.
[9] Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016); 2016 Dec 5–10; Barcelona, Spain. NIPS; 2016. p. 1–9.
[10] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529(7587):484–9.
[11] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550(7676):354–9.
[12] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 2018;362(6419):1140–4.
[13] Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed. Cambridge: MIT press; 2018.
[14] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[15] Wolf P, Hubschneider C, Weber M, Bauer A, Härtl J, Dürr F, et al. Learning how to drive in a real world simulation with deep Q-Networks. In: Proceedings of 2017 IEEE Intelligent Vehicles Symposium (IV); 2017 Jun 11–14; Los Angeles, CA, USA. IEEE; 2017. p. 244–50.
[16] Sallab AE, Abdou M, Perot E, Yogamani S. Deep reinforcement learning framework for autonomous driving. Electron Imaging 2017;29:70–6.
[17] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. PMLR; 2018. p. 1861–70.
[18] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. PMLR; 2018. p. 1587–96.
[19] Cai P, Mei X, Tai L, Sun Y, Liu M. High-speed autonomous drifting with deep reinforcement learning. IEEE Robot Autom Lett 2020;5(2):1247–54.
[20] Neftci EO, Averbeck BB. Reinforcement learning in artificial and biological systems. Nat Mach Intell 2019;1(3):133–43.
[21] Harutyunyan A, Dabney W, Mesnard T, Azar MG, Piot B, Heess N, et al. Hindsight credit assignment. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 9–14; Vancouver, BC, Canada. NeurIPS; 2019. p. 12498–507.
[22] Huang Z, Lv C, Xing Y, Wu J. Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sens J 2021;21(10):11781–90.
[23] Lv C, Cao D, Zhao Y, Auger DJ, Sullman M, Wang H, et al. Analysis of autopilot disengagements occurring during autonomous vehicle testing. IEEE/CAA J Autom Sin 2018;5(1):58–68.
[24] Mao J, Gan C, Kohli P, Tenenbaum JB, Wu J. The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. In: Proceedings of the 7th International Conference on Learning Representations (ICLR); 2019 May 6–9; New Orleans, LA, USA. ICLR; 2019. p. 1–28.
[25] Knox WB, Stone P. Reinforcement learning from human reward: discounting in episodic tasks. In: Proceedings of 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication; 2012 Sep 9–13; Paris, France. IEEE; 2012. p. 878–85.
[26] MacGlashan J, Ho MK, Loftin R, Peng B, Wang G, Roberts DL, et al. Interactive learning from policy-dependent human feedback. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia. PMLR; 2017. p. 2285–94.
[27] Vecerik M, Hester T, Scholz J, Wang F, Pietquin O, Piot B, et al. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. 2017. arXiv:1707.08817.
[28] Rajeswaran A, Kumar V, Gupta A, Vezzani G, Schulman J, Todorov E, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In: Proceedings of Robotics: Science and Systems; 2018 Jun 26–30; Pittsburgh, PA, USA. RSS; 2018. p. 1–9.
[29] Ibarz B, Leike J, Pohlen T, Irving G, Legg S, Amodei D. Reward learning from human preferences and demonstrations in Atari. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS); 2018 Dec 3–8; Montreal, QC, Canada. NeurIPS; 2018. p. 8011–23.
[30] Ziebart BD, Maas A, Bagnell JA, Dey AK. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence; 2008 Jul 13–17; Chicago, IL, USA. AAAI Press; 2008. p. 1433–8.
[31] Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, et al. Deep Q-learning from demonstrations. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA. AAAI Press; 2018. p. 3223–30.
[32] Saunders W, Sastry G, Stuhlmüller A, Evans O. Trial without error: towards safe reinforcement learning via human intervention. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems; 2018 Jul 10–15; Stockholm, Sweden. AAMAS; 2018. p. 2067–9.
[33] Krening S, Harrison B, Feigh KM, Isbell CL, Riedl M, Thomaz A. Learning from explanations using sentiment and advice in RL. IEEE Trans Cogn Dev Syst 2017;9(1):44–55.
[34] Nair A, McGrew B, Andrychowicz M, Zaremba W, Abbeel P. Overcoming exploration in reinforcement learning with demonstrations. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA); 2018 May 21–25; Brisbane, QLD, Australia. IEEE; 2018. p. 6292–9.
[35] Wang F, Zhou B, Chen K, Fan T, Zhang X, Li J, et al. Intervention aided reinforcement learning for safe and practical policy optimization in navigation. In: Proceedings of the 2nd Conference on Robot Learning; 2018 Oct 29–31; Zürich, Switzerland. PMLR; 2018. p. 410–21.
[36] Littman ML. Reinforcement learning improves behaviour from evaluative feedback. Nature 2015;521(7553):445–51.
[37] Droździel P, Tarkowski S, Rybicka I, Wrona R. Drivers' reaction time research in the conditions in the real traffic. Open Eng 2020;10(1):35–47.
[38] Hu Z, Zhang Y, Xing Y, Zhao Y, Cao D, Lv C. Toward human-centered automated driving: a novel spatiotemporal vision transformer-enabled head tracker. IEEE Veh Technol Mag. In press.
[39] Machado MC, Bellemare MG, Bowling M. Count-based exploration with the successor representation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York City, NY, USA. AAAI Press; 2020. p. 5125–33.
[40] Badia AP, Sprechmann P, Vitvitskyi A, Guo D, Piot B, Kapturowski S, et al. Never give up: learning directed exploration strategies. In: Proceedings of the 8th International Conference on Learning Representations (ICLR 2020); 2020 Apr 26–May 1; Addis Ababa, Ethiopia. ICLR; 2020. p. 1–26.