Safe Reinforcement Learning for Longitudinal Control of Autonomous Vehicles: An Augmented Neural Network with Supervision using Safe Distance

Chufei Yan[1], Zhihao Cui[2], Ning Bian[3], Yulei Wang*[1]

1. School of Physics, Northeast Normal University, Changchun 130024, P. R. China
E-mail: yanchufei@nenu.edu.cn, wangyule@nenu.edu.cn

2. Clean Energy Automotive Engineering Center School of Automotive Studies, Tongji University, Shanghai 201804, P. R. China
E-mail: 2131523@tongji.edu.cn

3. Mengshi Automobile Technology Company, Wuhan, 430010, P. R. China
E-mail: biann@dfmc.com.cn

Abstract: Ensuring safety during the training process of autonomous vehicles (AVs) is a significant challenge in the field of deep reinforcement learning (DRL). In this study, we propose a new longitudinal control scheme that augments neural network (NN) with supervision using safe distance [1] to ensure the safety of AVs during training. The NN is designed by multiply Q-Networks with actions of various combinations, and a supervisor with safe distance is introduced to select one of the Q-networks that complies with the Vienna Convention [2] without collision. Additionally, the approach does not introduce any additional driver models or switching rules, thus obeying the convergence of Q-learning with probability 1. Simulation results demonstrate that the AV by our DRL enables safe vehicle following, both in terms of implementation and during training process.

Key Words: Autonomous vehicles, autonomous driving, safe reinforcement learning, longitudinal control, safe distance.

1 Introduction

In the current context of increasing vehicle ownership and a rising incidence of traffic accidents, the application of autonomous driving (AD) technology to enhance traffic safety has emerged as an increasingly vital scientific challenge [30, 31]. In most autonomous vehicles (AVs), decision-making and control are typically based on model-based methods, thus lacking of the availability of large-scale datasets, closed-loop evaluation and challenging scenarios. In comparison to modular pipelines, reinforcement learning (RL) algorithms have shown more promising results [3–5] for decision-making of AVs. However, traditional RL algorithms reply on trial-and-error rules and thus are few utilized in real experimental environments. Unpredictable exploration and bad interaction often result in unsafe behavior for AVs during the training process, which are unacceptable for vehicle engining research, development and implementation.

Safe RL is an approach that aims to ensure that the learning process adheres to certain safety constraints while an agent interacts with its environment [6]. In previous work in AVs, efforts to safe RL have centered around integrated reward function [7], deep deterministic policy gradient [8], constants driven safe RL [9], Monte Carlo tree search [10], parallel constrained policy optimization [11], Lyapunov soft actor–critic [12], adapt guide and guard RL [13], game theory-based RL [14], safety rules [15], filling action selection RL [16] , safe and reasonable RL [17] and trustworthy safe RL [18]. Apart from only guarantying result safety, the research on safe RL are often classified into two categories [6]. The first type of methods make use of soft constraints such as negative rewards to penalize the agent when it explored outside predefined boundaries. Because soft constraints represent preferences rather than absolute requirements, the safety is flexible and can be violated to some degree. Hence, we define it as approximate safety. The second type of methods design hard constraints that are strict rules and cannot be violated under any circumstances. We denote it as strict safety. Most literature in strict safety introduce driver model (model-based method) providing actions or information whenever it feels the AV unsafe. However, there is no metric or rule as to the best time for to do it. Additionally, combining driver models or model-based hard constraints with neural networks (NNs), few literature provides a strict proof for the convergence of the Q-learning theory [22].

Several publications have formalized traffic rules for safe distance [23], lane changing [24], overtaking [1] and interactions [25]. They formalized the rules using German traffic law, US state laws, or Vienna Convention on Road Traffic. For RL, The authors of [17] first utilized safe distance to supervise RL during training process but the convergence of the algorithm does not follow the Q-learning theory [22]. In comparison, the authors of [19] focused on designing a reward machine RL for integrated longitudinal and lateral control in the framework of safe distance. Although the performance of the reward machine RL has been optimized, the algorithm cannot guarantee safety during RL training process. Up to now, a sophisticated safe RL with supervision using safe distance is still worth anticipating.

To address these challenges, as shown in Fig. 1, we propose an augmented NN with supervision using safety distance, aimed at achieving Safe RL for longitudinal control of AVs. Specifically, we propose an augmented Q-network that contains all subsets of the Q-network. This design allows the multi-network architecture to perform longitudinal control tasks, at any given time, that at least one network's optimal action satisfies the safety distance than single NN

Corresponding author: Yulei Wang. This work is supported by National Natural Science Foundation (NNSF) of China under Grant 62373281 and Shanghai Municipal Science and Technology Commission 23ZR1467700.

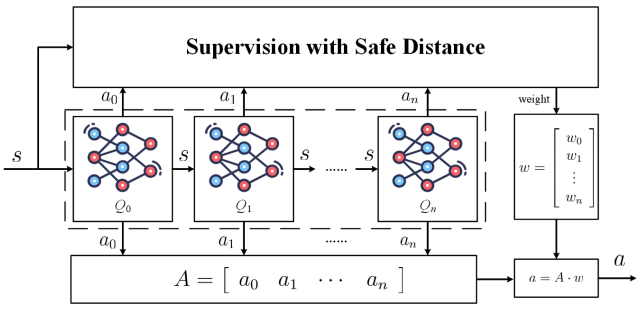

Fig. 1: The proposed scheme of the augmented NN with supervision using safe distance.

can handle. The augmented Q-network is supervised with safe distance in a layer-wise manner. The supervision can maintain a safe state as follows: When the optimal action of the upper-layer Q-network does not satisfy safety distance, the supervisor will switch to its next-layer Q-network without the optimal action of the upper-layer network. This process is carried out layer by layer until a certain layer Q-network satisfies the safety distance or reaches the bottom layer. By the augmented Q-network and the supervision with safe distance, the proposed method not only can achieve safety during training but also can satisfy the convergence of the Q-learning theory. Finally, experiments conducted in a highway environment (highway-env) demonstrate that the proposed algorithm outperforms the existing safe RL approaches with higher safe and performance scores.

The remainder of the paper is organized as follows: Section 2 reviews the preliminaries of safe RL and deep Q-network (DQN). Section 3 outlines the proposed method and provides our safe RL pseudocode. Section 4 presents the experimental details and discusses the numerical simulation results. Finally, Section 5 gives concluding remarks.

2 Preliminaries

2.1 Safe RL

Safe RL is an important research area developed from RL. It typically satisfies a Markov decision process represented by the tuple $<S, A, T, R>$, where S denotes the state space, A represents the action space, $T : S \times A \to S$ is the transition function, and $R : S \times A \to \mathbb{R}$ is the reward function. The agent employs additional criteria to evaluate how to identify an optimal policy that maximizes the expected cumulative reward while preventing large-scale negative outcomes. In Safe RL, the policy $\pi(a|s)$ represents the probability distribution over actions $a \in A$ given a state $s \in S$. At each time step, the agent, when in a specific state $s \in S$, selects an action a according to $\pi(\cdot|s)$. Subsequently, the state-action pair is inputted into a regulatory model set by external knowledge. Based on the recommendations of this regulatory model, the agent generates a compliant action a^* and executes it. The agent then receives the next state $s' \sim p(\cdot \mid s, a^*)$ and the current reward $r(s, a^*, s')$ from the environment. This process continues iteratively starting from s'. The agent's objective is to find an optimal policy π^* that maximizes the expected discounted future reward from each state $s \in S$, while adhering to the constraints established by external knowledge. In other words, the agent operates exclusively within the boundaries defined by the regulatory model, seeking to maximize long-term performance in the environment. The algorithms that strictly avoid causing harm or injury to agents, learning systems, or external entities are particularly important in the context of AD scenarios that emphasize safety.

2.2 Deep Q-Network

The traditional approach to RL employs tabular methods to store state value functions $V(s)$ or action value functions $Q(s, a)$. However, such methods face significant challenges when applied to continuous state spaces. To address this issue, value function approximation methods are commonly employed in continuous state and action spaces. Specifically, a NN $Q(s, a)$ is utilized to approximate $Q_\pi(s, a)$, where s and a represent vector representations of the state and action, respectively. In this paper, the function $Q(s, a)$ is typically implemented using forward NN.

DQN integrates value function approximation with NN techniques, employing methods such as target networks and experience replay for the training of the network [21]. The core of the DQN algorithm is the maintenance of a Q-network, which is used to inform decision-making. The action value function $Q(s, a)$ is defined so that upon reaching a state s, the algorithm traverses the entire action space to select the action that maximizes $Q(s, a)$ as follows:

$$A = \arg\max_a Q_\pi(s, a) \tag{1}$$

The DQN utilizes the Bellman equation to iteratively update the action value function $Q_\pi(s, a)$ as follows:

$$Q^*(s, a) = \sum_{s' \in S} p(s'|s, a) \left(r(s, a, s') + \gamma \max_{a' \in A} Q^*(s', a') \right) \tag{2}$$

3 Proposed Method

In this section, we introduce the augmented NN with supervision using safe distance, an adaptation of the DQN for safe RL. To begin with, we articulate the problem statement of longitudinal control for AVs on highways and the design of the associated reward mechanism. Next, we present an augmented Q-Network to ensure that our alternative action space complies with the Vienna Convention on Road Traffic at any given moment. Finally, we design a supervision strategy that determines which Q-Network should be enabled for execution at the current state.

3.1 Problem Statement and Reward Design

The experimental scenario for AVs on highways is configured using the highway-env environment. As shown in Fig. 2, we simulate a typical single-lane highway that includes an ego vehicle (EV) and two surrounding vehicles (SVs).

Each lane has a width of 4 meters, and the speed limit on the highway is set at 30 meters per second. We approximate the outline of each vehicle as a rectangle measuring 5 meters in length and 2 meters in width. We utilize relative coordinates, defining the position of the EV as the origin of the coordinate system. The initial positions of the SVs are uniformly distributed, and their initial speeds are randomly selected within the range of 23 to 25 meters per

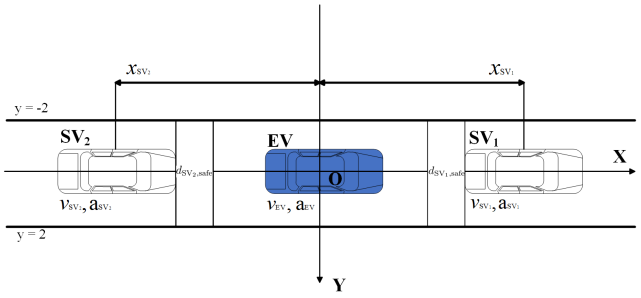

Fig. 2: Visualization of the EV and SVs in longitudinal control on a single-lane highway.

second. The driving model for the SVs is a combination of the IDM model and the MOBILE model. For each SV, the considered states include its longitudinal position x_{SV_i} and longitudinal speed v_{SV_i} with $i = 1, 2$. The initial longitudinal speed of the EV is set to 30 meters per second.

This environment's reward design is as follows: First, we divide the entire state space into four regions using a two-dimensional vector $(\mathcal{X}, \mathcal{Y})$ consisting of two Boolean values. The vector $\mathcal{X}$ is determined by Eqs. (3)-(14). When $\mathcal{X}$ is true, it follows that the EV is in a safe state. Similarly, the vector $\mathcal{Y}$ is determined by Eq.(15). When $\mathcal{Y}$ is true, it indicates that the speed of the EV is higher than that of the front SV and in the vehicle-following mode.

$$d_{\text{EV},\delta_t} = v_{\text{EV}}\delta_t \tag{3}$$

$$d_{\text{SV}_i,\delta_t} = v_{\text{SV}_i}\delta_t - \frac{1}{2}a_{\text{SV}_i,\max}\delta_t^2 \tag{4}$$

$$d_{\text{EV},SD} = \frac{v_{\text{EV}}^2}{2a_{\text{EV},\max}} \tag{5}$$

$$d_{\text{SV}_i,SD} = \frac{v_{\text{SV}_i}^2}{2a_{\text{SV}_i,\max}} \tag{6}$$

$$v_{\text{SV}_i}^* = v_{\text{SV}_i} - a_{\text{SV}_i,\max}\delta_t \tag{7}$$

$$\begin{aligned}\mathcal{A} = (d_{\text{SV}_i,\delta_t} \le d_{\text{EV},SD}) \wedge (|a_{\text{SV}_i,\max}| < |a_{\text{EV},\max}|) \\ \wedge (v_{\text{SV}_i}^* < v_{\text{EV}}) \wedge (t_{\text{EV},SD} < t_{\text{SV}_i,SD})\end{aligned} \tag{8}$$

$$\text{SV}_{i,\text{safe\_1}} = \Delta d > (d_{\text{EV},\delta_t} + d_{\text{EV},SD}) \tag{9}$$

$$\text{SV}_{i,\text{safe\_2}} = (t_{\text{SV}_i} \ge \delta_t) \wedge ((\Delta d + d_{\text{SV}_i,\delta_t}) > d_{\text{EV},SD}) \tag{10}$$

$$\text{SV}_{i,\text{safe\_3}} = \begin{cases}\mathcal{B} & \text{If Eq. (8) is true,} \\ \mathcal{C} & \text{Otherwise.}\end{cases} \tag{11}$$

$$\mathcal{B} = \Delta d > \left(\frac{(v_{\text{SV}_i}^* - v_{\text{EV}})^2}{2(|a_{\text{EV},\max}| - |a_{\text{SV}_i,\max}|)}\right) - d_{\text{SV}_i,\delta_t} + d_{\text{EV},\delta_t} \tag{12}$$

$$\mathcal{C} = \Delta d > (d_{\text{EV},SD} + d_{\text{EV},\delta_t} - d_{\text{SV}_i,SD}) \tag{13}$$

$$\mathcal{X} = \text{SV}_{i,\text{safe\_1}} \vee \text{SV}_{i,\text{safe\_2}} \vee d_{\text{SV}_i,\text{safe\_3}} \tag{14}$$

$$\mathcal{Y} = (v_{\text{EV}} > v_{\text{SV}_i}^*) \wedge (\Delta d_{\min} < \Delta d < \Delta d_{\max}) \wedge \mathcal{X} \tag{15}$$

where δ_t is the reaction time of the driver, which is set to a constant value, v_{EV} is the speed of the EV, $a_{\text{SV}_i,\max}$ is the maximal acceleration of the SV_i, $a_{\text{EV},\max}$ is the maximal acceleration of the EV, $\Delta d = x_{\text{SV}_i} - x_{\text{EV}}$ is the distance between the EV and SV_i, $t_{\text{EV},SD} = v_{\text{EV}}/a_{\text{EV},\max}$, $t_{\text{SV}_i} = v_{\text{SV}_i}/a_{\text{SV}_i,\max}$, $t_{\text{SV}_i,SD} = t_{\text{SV}_i} - \delta_t$, $\Delta d_{\min}$ and $\Delta d_{\max}$ are constants.

Note that Eq. (3) calculates the braking displacement of the lead vehicle within the reaction time, reflecting the most pessimistic displacement of the lead vehicle at the current moment. Eqs. (5) and (6) determine the stopping displacement required for both the EV and the lead vehicle, indicating the distance necessary for the vehicles to come to a complete stop. Eq. (7) characterizes the velocity of the lead vehicle after the reaction time, assuming that the lead vehicle decelerates during this period. The variable $\mathcal{A}$ represents the minimum safe distance criterion between the EV and the lead vehicle [2]. Eq. (9) assesses whether the current distance Δd to the lead vehicle satisfies the most pessimistic safe distance, specifically examining if the required braking displacement for the EV exceeds the distance when the lead vehicle's speed suddenly drops to zero. Eqs. (11)-(13) establish the minimum safe distance criteria, collectively forming a segment of the Vienna Convention on Road Traffic safety principles, in conjunction with Eqs. (9) and (10). Eq. (14) indicates that if any of the safe distance conditions is met, $\mathcal{X}$ will be set to 1. Eq. (15) states that if the current vehicle speed exceeds that of the lead vehicle after reacting and the vehicle is within the obstacle space, $\mathcal{Y}$ will be set to 1.

The reward calculation formulas under different vectors are as follows:

$$r = \begin{cases} 0, & \text{if } \mathcal{X} = 0 \\ \frac{v_{\text{EV}}}{v_{\max}}, & \text{if } (\mathcal{X} = 1) \wedge (\mathcal{Y} = 0) \\ \frac{v_{\text{EV}}}{v_{\text{SV}_i}}, & \text{if } (\mathcal{X} = 1) \wedge (\mathcal{Y} = 1) \end{cases} \tag{16}$$

where $v_{\max}$ is the maximum speed of the road, $\mathcal{Y}$ is set as the criterion for a switch between the vehicle-following and high speed driving.

3.2 Augmented Q-Network

In longitudinal control, the augmented Q-Network consists of four subnetworks. As illustrated in Fig. 3, the states of each subnetwork are the same and they are v_{EV}, a_{EV}, x_{SV_i}, v_{SV_i}, $a_{\text{SV}_i,\max}$, $\mathcal{X}$ and $\mathcal{Y}$. The total set of the actions is the target speed, i.e.,

$$a = \left\{ \text{ Faster } \quad \text{Slower } \quad \text{Idle } \right\} \tag{17}$$

where Faster $= 5$ meters per second, Slower $= -5$ meters per second and Idle $= 0$ meters per second. The four networks exhaustively explored all combinations of the three actions. The training procedure is as follows: At each step, all Q-networks simultaneously collect the state data and response the optimal action with respect to the maximum Q-value.

3.3 Supervision using Safety Distance

The supervisor is a core module of the safe RL algorithm. The module has two main functions. One of the functions is safety verification, which involves using safety distance to verify whether the state-action pair of each subnetwork is safe. Another function of the supervisor is to select which subnetwork should be activeated at each step. Specifically, the rule is as follows: when the optimal action of the upper-layer Q-network does not satisfy safety distance, the supervisor will switch to its next-layer Q-network without the optimal action of the upper-layer network. This process is carried out layer by layer until a certain layer Q-network satisfies the safety distance or reaches the bottom layer. If the

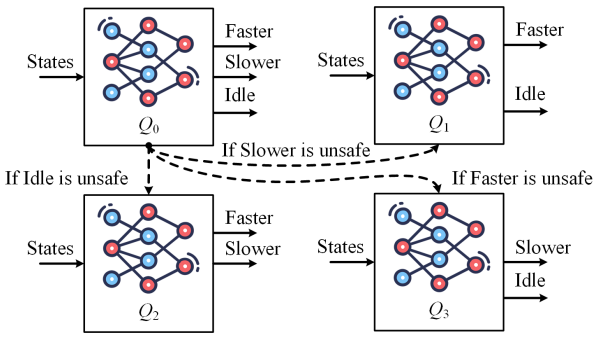

Fig. 3: Visualization of four sub-NNs of the augmented Q-Network.

action of the bottom layer Q-network is not safe, the episode ends. Algorithm 1 shows the pseudocode for with the augmented NN the supervision using safety distance.

Algorithm 1 An augmented NN with supervision using safe distance

Require: $S, A, \gamma \in [0,1], \alpha \in (0,1], u_0, A_i$ with $i \in \{0,1,2,3\}$
For all $s \in S$, $u \in U$ and $a_i \in A_i \subseteq A$
for $l = 0$ to $num\ episodes$ **do**
  Initialize $u \leftarrow u_0$ and $s \leftarrow EnvInitialState()$
  while s is not terminal **do**
    $j \leftarrow 0$ and $A_{\text{safe}} \leftarrow Supervision(s, u)$
    Choose action a_i from $\{s, u\}$ using Q_i, respectively
    if $a_0 \in A_{\text{safe}}$ **then**
      $a \leftarrow a_0$ and $j \leftarrow 0$
    else
      $j \leftarrow \arg_i \{A_i \cap \{a_0\} = \emptyset\}$
      if $a_j \in A_{\text{safe}}$ **then**
        $a \leftarrow a_j$
      else
        $a \leftarrow A \setminus \{a_0, a_j\}$
      end if
    end if
    Take action a and observe the next state s'
    Compute the reward state $u' \leftarrow \{\mathcal{X}, \mathcal{Y}\}$ in (3)-(15)
    Current reward $r \leftarrow \{\mathcal{X}, \mathcal{Y}\}$ in (16)
    if $a = a_0$ **or** $a = a_j$ **then**
      if s' is terminal **then**
        $Q_j(s,u,a) \leftarrow Q_j(s,u,a) + \alpha(r - Q_j(s,u,a))$
      else
        $R \leftarrow r + \gamma \max_{a'} Q_j(s', u', a') - Q_j(s,u,a)$
        $Q_j(s,u,a) \leftarrow Q_j(s,u,a) + \alpha R$
      end if
    end if
    Update $s \leftarrow s'$ and $u \leftarrow u'$
  end while
end for

3.4 Convergence proof

The convergence of Q-learning, specifically its ability to converge with probability 1, has been the subject of previous research [20]. Watkins [22] provided the first proof in 1989, followed by a more detailed account with Dayan [26] in 1992. In 1994, Tsitsiklis [27] applied stochastic approximation theory to demonstrate its convergence under general asynchronous structures. Building upon this work, Even-Dar et al. [28] derived more precise rates of convergence in 2003. Finally, Szepesvári and Littman [29] focused on the contraction properties of dynamic programming operators in the context of generalized Markov Decision Processes (MDPs) in their 1996 analysis of Q-learning. Note that the proposed augmented NN with supervision does not introduce any artificial prior knowledge and each subnetwork is mutually independent. Hence, as long as the data is sufficient, each subnet follows the convergence of Q-learning, which ensures that the overall augmented network is convergent as well.

Remark 3.1. To evaluate the impact of four sub-NNs of the augmented Q-Network, we conducted ablation experiments by masking specific networks. When $Q_{1,2,3}$ were masked, the model reverted to Expert Takeover RL. Masking the Q_0 caused the model to lose longitudinal control capabilities due to the supervisor lacking top-layer data.

Remark 3.2. Our architecture can be extended to multi-lane scenarios by expanding the number of sub-NNs and action space, though we omit further details due to space constraints.

4 Experimental Evaluation

4.1 Evaluation Metrics and Baselines

In this section, we will evaluate the proposed methods in a highway longitudinal control environment, assessing the performance of various algorithms under the same traffic conditions. Our experiments aim to evaluate the following metrics:

1) **Number of Collisions:** This metric measures the number of collisions occurring during the model training process over 100,000 steps.
2) **Episode Reward:** The metric measures the rewards obtained by the agent over a complete episode, aiming to evaluate the performance of the autonomous driving algorithm in longitudinal control, i. e., $r_{\text{sum}} = \sum_{i=1}^{n_{\text{total}}} r$.
3) **Number of Violating Safe Distance:** This metric is designed to measure the number of steps during the model training process over 100,000 steps that do not meet regulatory conditions. This metric is applicable for assessing the safety margin of the model during training, in relation to the level of safety.

To evaluate the performance of the proposed algorithm, we will compare it against various baseline methods as follows:

1) **Baseline 1 (Reward Machine RL or RM-RL)**: This baseline [19] utilizes a DQN algorithm combined with a reward machine, yet it does not incorporate any safe RL mechanism. The priority in this baseline is to maximize long-term rewards, where penalties for collisions result only in the termination of the current episode, thereby preventing the acquisition of subsequent step rewards.
2) **Baseline 2 (Approximate safety RL or AS-RL)**: This baseline [9] adopts an approximate safety approach, imposing penalties on all actions that may lead to unsafe behavior. It is noteworthy that this baseline employs a distinct reward computation method compared to the other baselines.
3) **Baseline 3 (Expert Takeover RL or ET-RL)**: This baseline [13] incorporates a simple expert model. When the vehicle reaches the safety distance threshold, the network switches to a simple expert model based

on the Intelligent Driver Model (IDM). The expert will cease further training and halt the vehicle, thereby ending the episode.

4) **Baseline 4 (Safe and Reliable RL or SR-RL)**: This baseline [17] employs the most common safe RL algorithm. In this baseline, if the maximum Q-value does not satisfy safety constraints, the algorithm sequentially checks whether the next highest Q-value actions are safe and outputs the safe action.

4.2 Experimental Setup

We conducted randomized experiments to evaluate each algorithm, with ten surrounding vehicles, a time limit of 40 seconds, and a policy frequency of 16 Hz. The initial positions of the surrounding vehicles were uniformly distributed, while the ego vehicle was initialized with a speed of 30 m/s. The initial speeds of the surrounding vehicles were randomly chosen within the range of 21 m/s to 24 m/s. At this speed, the initial distance between the ego vehicle and the surrounding vehicles did not satisfy the safe distance requirement, necessitating rapid deceleration from the ego vehicle. This setup aimed to increase the complexity of the environment.

The DQN algorithm was trained for 1×10^5 steps, and the median performance results were reported.

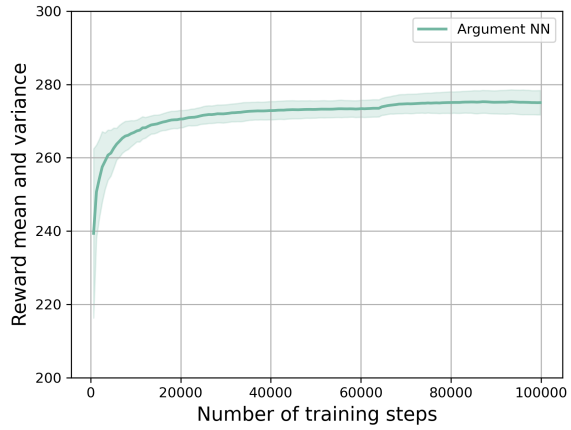

Fig. 4: Reward mean and variance of the augmented NN.

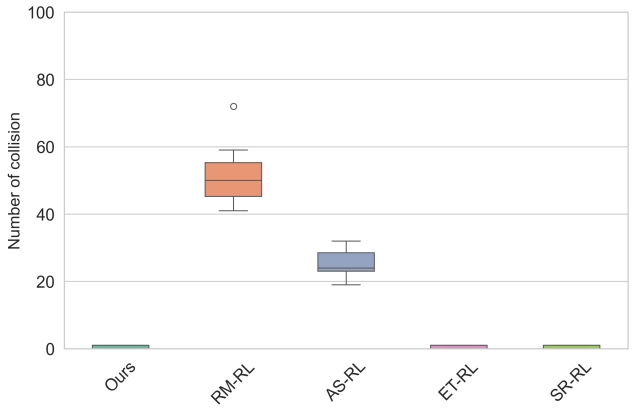

Fig. 5: Numbers of collision in different approaches.

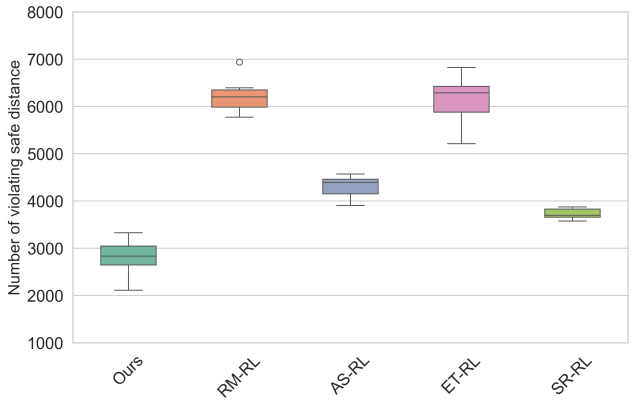

Fig. 6: Numbers of violating safe distance in different approaches.

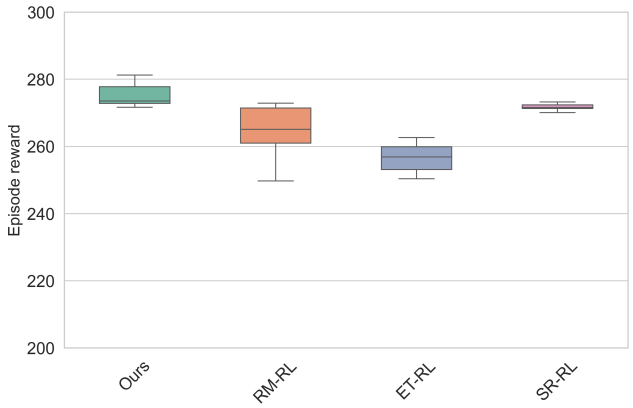

Fig. 7: Episode rewards r_{sum} in different approaches.

4.3 Experimental Results

We evaluated the performance of the ANN in a longitudinal control environment on highways, assessing each algorithm across 10 groups. The simulation results are presented in Figs. 4-7. Fig. 4 demonstrates the convergence and stability of the proposed algorithm. Figs. 5-7 compare the proposed method with other baseline methods in terms of collision number, violation of safe distance and episode reward, respectively. Note that RM-RL fails to ensure safety throughout the training process. Although AS-RL achieves performance similar to RM-RL with fewer collisions and instances of not maintaining a safe distance, it still cannot eliminate collisions entirely. ET-RL adopts an excessively conservative approach, resulting in slower convergence due to frequent switches, and presents a high number of violations of the safe distance metric. The performance of SR-RL is comparable to that of our proposed method in terms of average rewards and collision counts; however, SR-RL leads to a high number of violating safe distance and might diverge due to selecting action that violates Q-learning convergence property. In summary, the augmented NN exhibits effectiveness in ensuring safe driving with minimal compromise to vehicle-following speed and number of violating safe distance. Observe that the performance of the proposed method is similar to that of SR-RL, which may be due to the simplicity of the longitudinal control scenario.

5 Conclusion

In this paper, we developed a safe RL scheme that utilizes augmented NN with supervision using safe distance to specify the Q-value of an AV agent operating on longitudinal control and vehicle-following. Our proposed Safe RL allows for maximizing the use of data to train AVs while ensuring safety. By supervision and classification using safe distance, our safe RL significantly improves safety, performance and sample efficiency for longitudinal control of AVs.

The research on safe RL technologies for AVs is a key technology related to the practical application possibilities of RL. Numerous practical problems need to be considered to design more general end-to-end AD models that can operate in open traffic environments. In future research, we will focus on developing more comprehensive RL algorithms that can handle complex traffic scenarios, such as multilane highway and merging area.

References

[1] A. Rizaldi, J. Keinholz, M. Huber, J. Feldle, F. Immler, M. Althoff, E. Hilgendorf, and T. Nipkow, Formalizing and monitoring traffic rules for autonomous vehicles in Isabelle/HOL, *Integrated Formal Methods(IFM)*, Springer, 2017: 50–66.

[2] E. C. for Europe, Vienna Convention on Road Traffic, *http://www.unece.org/fileadmin/DAM/trans/conventn/crt1968e.pdf*, 1968.

[3] H. Peng, W. Wang, Q. An, C. Xiang, and L. Li, Path Tracking and Direct Yaw Moment Coordinated Control Based on Robust MPC With the Finite Time Horizon for Autonomous Independent-Drive Vehicles, *IEEE Trans. Veh. Technol.*, 69(6): 6053–6066, 2020.

[4] M. Yildirim, S. Mozaffari, L. McCutcheon, M. Dianati, A. Tamaddoni-Nezhad, and S. Fallah, Prediction Based Decision Making for Autonomous Highway Driving, *International Conference on Intelligent Transportation Systems(ITSC)*, IEEE, 2022: 138–145.

[5] W. Zhao, T. He, R. Chen, T. Wei, and C. Liu, Statewise Safe Reinforcement Learning: A Survey, *arXiv preprint arXiv:2302.03122*, 2023.

[6] J. García and F. Fernández, A comprehensive survey on safe reinforcement learning, *J. Mach. Learn. Res.*, 16(1): 1437–1480, 2015.

[7] S. B. Jo, P. S. Kim, and H. Y. Jeong, An Integrated Reward Function of End-to-End Deep Reinforcement Learning for the Longitudinal and Lateral Control of Autonomous Vehicles, *Vehicular Technology Conference(VTC2022-Spring)*, IEEE, 2022: 1–5.

[8] T. Tiong, I. Saad, K. T. K. Teo and H Lago, Autonomous vehicle driving path control with deep reinforcement learning, *Annual Computing and Communication Workshop and Conference(CCWC)*, IEEE, 2023: 0084–0092.

[9] F. Gao, X. Wang, Y. Fan, Z Gao and R Zhao, Constraints Driven Safe Reinforcement Learning for Autonomous Driving Decision-Making, *IEEE Access*, 2024.

[10] S. Mo, X. Pei, and C. Wu, Safe reinforcement learning for autonomous vehicle using Monte Carlo tree search, *IEEE Trans. Intell. Transp. Syst.*, 23(7): 6766-6773, 2021.

[11] L. Wen, J. Duan, S. E. Li, S Xu and H Peng, Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization, *International Conference on Intelligent Transportation Systems(ITSC)*, IEEE, 2020: 1–7.

[12] L. Zhang, R. Zhang, T. Wu, R Weng, M Han and Y Zhao., Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles, *IEEE Trans. Neural Networks Learn. Syst.*, 32(12): 5435-5444, 2021.

[13] H. Wang, Y. Cheng, H. Chu, C Zhao and B Gao, Adept Guide and Guard Reinforcement Learning for Safe Driving in Lane-Keeping Scenarios, *China Automation Congress(CAC)*, IEEE, 2023:7214–7219.

[14] R. Yang, Z. Li, B. Leng and L Xiong, Safe reinforcement learning for autonomous vehicles to make lane-change decisions: Constraint based on Incomplete Information Game Theory, *Conference on Vehicular Control and Intelligence (CVCI)*, IEEE, 2023: 1–6.

[15] J. Xu, X. Pei, and K. Lv, Decision-making for complex scenario using safe reinforcement learning, *Conference on Vehicular Control and Intelligence(CVCI)*, IEEE, 2020:1–6.

[16] F. Yang, X. Li, Q. Liu, C Liu, Z Li and Y Liu, Filling action selection reinforcement learning algorithm for safer autonomous driving in multi-traffic scenes, *Intelligent Vehicles Symposium (IV)*, IEEE, 2023: 1–7.

[17] B. Mirchevska, C. Pek, M. Werling, M Althoff and J Boedecker, High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning, *Intelligent Transportation Systems(ITSC)*, IEEE, 2018: 2156–2162.

[18] Z. Gao, S. Xu, X. Jiao, H Peng and D Yang, Trustworthy safety improvement for autonomous driving using reinforcement learning, *Transp. Res. Part C: Emerg. Technol.*, 138: 103656, 2022.

[19] Z. Cui, Y. Wang, N. Bian and H Chen, Reward Machine Reinforcement Learning for Autonomous Highway Driving: An Unified Framework for Safety and Performance, *Conference on Vehicular Control and Intelligence(CVCI)*, IEEE, 2023: 1–6.

[20] M. T. Regehr and A. Ayoub, An Elementary Proof that Q-learning Converges Almost Surely, *arXiv:2108.02827*, 2023.

[21] V. Mnih, K. Kavukcuoglu, D. Silver, A A Rusu, J Veness, M G Bellemare, et al., Human-level control through deep reinforcement learning, *Nature*, 518(7540): 529–533, 2015

[22] C. J. C. H. Watkins, Learning from delayed rewards, 1989.

[23] A. Rizaldi, F. Immler and M. Althoff, A formally verified checker of the safe distance traffic rules for autonomous vehicles, in: *NASA Formal Methods*, 175–190, 2016.

[24] C. Pek, P. Zahn and M. Althoff, Verifying the safety of lane change maneuvers of self-driving vehicles based on formalized traffic rules, *Intelligent Vehicles Symposium(IV)*, IEEE, 2017: 1477–1483.

[25] S. Maierhofer, P. Moosbrugger and M. Althoff, Formalization of intersection traffic rules in temporal logic, *Intelligent Vehicles Symposium(IV)*, IEEE, 2022: 1135–1144.

[26] C. J. C. H. Watkins and P. Dayan, Q-learning, *Machine Learning*, 8(3-4):279–292, 1992.

[27] J. N. Tsitsiklis, Asynchronous stochastic approximation and Q-learning, *Machine Learning*, 16(3):185–202, 1994.

[28] E. Even-Dar, Y. Mansour, and P. Bartlett, Learning rates for Q-learning, *Journal of Machine Learning Research*, 5(1), 2003.

[29] C. Szepesvári and M. L. Littman, Generalized Markov decision processes: Dynamic programming and reinforcement-learning algorithms, in *Proceedings of the International Conference of Machine Learning*, vol. 96, 1996.

[30] B R Kiran, L Sobh, V Talpaert, P Mannion,A A A Sallab and S Yogamani., Deep reinforcement learning for autonomous driving: A survey, *IEEE Trans. Intell. Transp. Syst.*, 23(6): 4909-4926, 2021.

[31] L. Chen, P. Wu, K. Chitta, B Jaeger, A Geiger and H Li, End-to-end autonomous driving: Challenges and frontiers, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.