Nature Communications

https://doi.org/10.1038/s41467-026-69761-x

Article in Press

Breaking through safety performance stagnation in autonomous vehicles with dense learning

Received: 30 March 2025

Accepted: 4 February 2026

Cite this article as: Feng, S., Zhu, H., Sun, H. *et al.* Breaking through safety performance stagnation in autonomous vehicles with dense learning. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-69761-x

Shuo Feng, Haojie Zhu, Haowei Sun, Xintao Yan, Linxuan He, Jingxuan Yang, Guangzhen Su, Boqi Li, Shu Li, Ling Wang, Shengyin Shen & Henry X. Liu

Title

Breaking through safety performance stagnation in autonomous vehicles with dense learning

Authors

Shuo Feng[1], Haojie Zhu[2], Haowei Sun[2], Xintao Yan[2], Linxuan He[1], Jingxuan Yang[1], Guangzhen Su[1], Boqi Li[2], Shu Li[1], Ling Wang[1], Shengyin Shen[3], Henry X. Liu[2,3,4*]

Affiliations

[1]Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
[2]Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA
[3]University of Michigan Transportation Research Institute, Ann Arbor, MI, USA
[4]Mcity, University of Michigan, Ann Arbor, MI, USA
*Corresponding Author: henryliu@umich.edu

Abstract

Autonomous vehicles remain commercially limited largely due to safety performance stagnation. Existing deep learning, heavily reliant on failure data from rare safety-critical events, suffers from the seesaw effect—improvement in some scenarios causes regression in others. We introduce an innovative dense learning approach that prioritizes both informative failures and successes, informed by theoretical findings. Data is sampled proportionally to its contribution to the policy gradient and exposure frequency, excluding non-informative samples. This densifies the training dataset's information, significantly reducing learning variance without bias, enabling tasks intractable for existing methods. To validate this, we trained a safety-critical driving agent for a highly automated vehicle using mixed reality on an urban test track. Results demonstrate that our approach breaks the performance stagnation, enhancing the model's overall safety performance by one to two orders of magnitude. This marks a significant stride towards achieving human-level safety and widespread adoption for autonomous vehicles.

Introduction

The dream of autonomous vehicles (AVs) has been around for approximately a century[1]. Enormous amounts of capital, exceeding $160 billion, have been invested in this vision over the past twenty years[2]. Despite witnessing significant advancements in AV technology, no commercially available AVs have yet met the SAE Level 4 standard[3]. The pace of progress in AV development has been disappointingly slow, particularly when it comes to ensuring their safety, which has remained stagnant in recent times. The situation was further exacerbated by the unfortunate accident involving Cruise Automation in San Francisco on October 2nd, 2023[4]. This safety gap presents a significant obstacle to the deployment and commercialization of AVs, as they struggle to effectively handle a wide array of infrequent yet critical safety events, commonly referred to as the long-tail challenge for AV safety[5-6]. Existing approaches have proven inadequate in overcoming this challenge, resulting in a noticeable slowdown and even stagnation in the enhancement of AV safety performance. Consequently, the development and deployment of AVs have been severely hindered, calling for an urgent breakthrough.

We formulate the safety challenge for AVs as the Curse of Rarity (CoR)[7], which arises due to the rarity of safety-critical events in high-dimensional variable spaces. We recognize that this is a compounding effect resulting from the rarity of events combined with the high dimensionality of related variables. As the utilization of deep learning techniques is typically necessary to address the high dimensionality, the rarity of events dramatically increases the estimation variance of policy gradient, thereby impeding the ability of deep-learning models to learn. This challenge exists in different AV safety-related tasks, and we attacked it in the AV safety testing task in our prior work[8]. However, the main objective of AV testing is to evaluate the safety performance by estimating the probabilities of rare events (e.g. car collisions) given a specific AV policy, whereas the AV training task is to improve the safety performance by searching for a policy from the AV policy space that can minimize these probabilities. This fundamental distinction makes AV training considerably more challenging than AV testing. Furthermore, during the AV training process, better AV safety performance also means fewer safety-critical events, which, in turn, makes it more difficult to improve the safety performance. These issues pose significant obstacles to the development of AVs to ensure safety performance. It should be emphasized that previous advancements in AI have primarily been focused on non-safety-critical applications such as chatbots and games[9-12], where a certain level of failure is deemed acceptable. However, when it comes to safety-critical autonomous systems[13] such as AVs, an extremely high level of safety performance is demanded, resulting in the CoR challenge.

Existing approaches that attempted to tackle the CoR challenge primarily focus on learning from the data where AI systems exhibit failures[14]. For instance, Tesla had reported training their systems on datasets that represent scenarios where AVs struggle or deviate from human driver behavior[15]. However, it is important to note that these approaches lack a solid theoretical foundation, and our investigations have revealed that they can exhibit significant biases and even be misleading. Consequently, while these approaches may enhance the safety performance of AVs in specific scenarios, they run the risk of compromising performance in other safety-critical situations (Fig.1a). This phenomenon, known as the seesaw effect in different domains[16-18], hinders the improvement of overall safety performance and leads to performance stagnation. That is also a key reason why safety-critical driving situations are seemingly endless, despite years of development in AVs. Alternatively, some researchers have attempted to prevent unsafe behaviors of AVs based on rules or models such as formal methods[19-21] and constrained learning[22-24]. However, this approach faces challenges to handle the variability and complexity in high-dimensional variable spaces, because it is difficult to rely on a predefined set of parameters and assumptions for a wide spectrum of diverse driving situations[22].

We address the CoR challenge by developing an innovative dense reinforcement learning approach to overcome performance stagnation and enable AVs to continually improve their safety performance beyond the current state-of-the-art. The key idea is to remove the non-informative data to densify the information from different perspectives including the episodic data densification, state-level densification, and retrospective data densification (see Fig.1b). For the episodic data densification, we found out that the optimal episodic data distribution for training neural networks to overcome CoR should be based on their contribution to the policy gradient and their exposure frequency, as indicated in Theorem 1 in Methods. Specifically for the AV safe driving task, the training data set should include data from both avoidable crash episodes and episodes where crashes were successfully avoided, i.e., near-misses. Moreover, as each driving episode can span many time steps, we retained and reconnected the safety-critical states, which can further densify training data within each driving episode from the state-level, as indicated in

Theorem 2 in Methods. Lastly, due to the changing AV policy during training, achieving both higher precision and recall rates for identifying informative episodes and states has become difficult. To address this, we designed a learned safety metric that serves as a real-time predictive evaluation tool. It also includes a retrospective evaluation component to re-evaluate all data for the new AV policy through counterfactual simulation.

To demonstrate the efficacy of our methodology, we trained a safety-critical driving agent (referred to as SafeDriver in this study) for a SAE Level 4 (ref.[3]) AV with Autoware[25], which is widely recognized as the leading open-source automated driving system. As a safety filter system[26], SafeDriver overrides Autoware only in safety-critical situations to enhance the overall safety performance of the AV, as shown in Fig.1c. As SafeDriver is decoupled from Autoware, it is applicable for other AV models including those developed based on large neural network models. We first tested the performance of SafeDriver in multiple simulated naturalistic driving environments (NDEs), then equipped a Lincoln MKZ hybrid with Autoware, and tested the SafeDriver in a physical urban test track with mixed reality. The results obtained from both simulation and field-testing clearly demonstrate the effectiveness of our approach in training a safety-critical driving agent. By employing our methodology, the overall safety performance of the AV models can be significantly enhanced, with improvements ranging from one to two orders of magnitude.

Results

Dense learning approach for AV training

We demonstrated the seesaw effect for AVs (see Section 2.3 in Supplementary Information) and then found out that it is essentially caused by the biasedness of learning and the severe variance caused by the CoR (see Eq. (8) in Methods). Here the biasedness indicates that the expectation of the gradient estimation is different from the ground truth, so the learning process could be misled. To address this issue, we develop the dense reinforcement learning approach for AV training. The key is to remove the non-informative data, thereby reducing the learning variance while maintaining the learning unbiasedness. Removing a larger amount of non-informative data leads to a greater reduction in variance. However, accurately defining, identifying, and effectively utilizing informative data for AV training pose significant challenges. We address this challenge from different perspectives including the episodic data densification, state-level densification, and retrospective data densification.

For the episodic data densification, we first obtain the optimal training data distribution for AV training based on importance sampling theory[27] as

$$q_\pi^*(\boldsymbol{X}) \propto \|\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\|_2 \times P_\pi(\boldsymbol{X}), \quad (1)$$

where $\boldsymbol{X}$ denotes each driving episode of the training data, π denotes the AV policy, $\omega(\boldsymbol{X}) = 1$ denotes the objective event that is a rare event (for example, crash events), $\nabla_\pi P(\omega(\boldsymbol{X}) = 1)$ is the policy gradient for the probability of the objective event $\omega(\boldsymbol{X}) = 1$ given $\boldsymbol{X}$, $\|\cdot\|_2$ denotes the $l2$ norm of the vector, $P_\pi(\boldsymbol{X})$ denotes the exposure frequency of $\boldsymbol{X}$ in NDE, $q_\pi^*(\boldsymbol{X})$ denotes the optimal probability distribution of $\boldsymbol{X}$ for training the policy π, and the symbol $\propto$ means 'proportional to'. Here an episode is a segment of recorded driving data with a predetermined time duration or distance. It indicates that the optimal training data should contain both avoidable crash

events (informative failures) and near-miss events (informative successes) where $\|\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\|_2$ is non-zero, while the data of those unavoidable crash events and safe situations should be removed.

The primary obstacle in applying Eq. (1) is that $\|\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\|_2$ for each driving episode of data cannot be calculated practically, exacerbated by the changing AV policy during the training process. We note that the key to dense learning is to remove the non-informative data and keep the informative one, so Eq. (1) can be utilized as a guidance for data densification. Therefore, rather than attempting to compute the precise values of $\|\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\|_2$, we adopt an approximation through a binary classification task. In this task, $\|\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\|_2$ is approximated as one for informative data, that is, avoidable crash events and near-miss events. Specifically, we identified a crash event as avoidable if an evasive trajectory is still feasible after the vehicle state is identified as safety-critical and a non-crash event as a near-miss if the minimum relative distance between the AV and background vehicles is below a pre-determined threshold. Although these criteria are not unique, we chose these simple yet effective ones to demonstrate the effectiveness of our approach. By sampling the episodes according to $P_\pi(\boldsymbol{X})$ and rejecting these where $\|\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\|_2$ is estimated as zero, our approach realizes the training data distribution as in Eq. (1), dramatically reducing variance for rare event learning without compromising unbiasedness, as indicated in Theorem 1 in Methods.

We then conduct the state-level densification and retrospective data densification. As each driving episode may last for many time steps, we retained and reconnected the informative states to densify the data from the state-level. This is challenging in AV training, as different AV policies during the training process could have different safety-critical states and we need to identify them with both high precision and recall rates. To address this challenge, we design a learned safety metric, which obtains better performances than existing approaches (see Supplementary Fig.1 and Supplementary Video 2). Moreover, as AV policy is changed after training, a large policy gap may exist between the new AV policy and the policy that collected the data. To bridge this gap, we introduce a retrospective evaluation component to reidentify the informative episodes and states for the new AV policy through counterfactual simulation (see Section 2.1.9 and 2.11 in Supplementary Information). Furthermore, to improve the efficiency of data collection, we also utilize the intelligent testing environment (ITE) developed in our previous study[8]. Utilizing ITE allows us to accelerate the collection of informative training data by multiple orders of magnitude (see Supplementary Fig.2, Supplementary Video 3, and Section 2.7 in Supplementary Information).

Simulation results

We evaluated the effectiveness of our approach by systematic simulation analysis. To measure the overall safety performances of AVs for quantitative comparisons, we utilized the crash rate per test of AVs in NDE. As NDE is generated based on naturalistic driving data, testing results in NDE can represent the safety performance of AVs in the real world[28]. Specifically, we selected different types of base AV models, and trained SafeDriver that only takes over from the base AV models in safety-critical states (identified by the learned safety metric), resulting in the integrated AV models. We then compared the safety performances of base AV models and integrated AV models in NDE. Although we refer to the baseline models as ‘base’ AV models, they could be advanced commercial AV models with large-scale neural networks together with their own safety guards. To demonstrate the generalizability and applicability of our approach, we conducted experiments

for four types of base AV models in different driving environments including highway, roundabout, and urban environments (Fig.2). In this study, we only utilized a simple Multilayer Perceptron as backbone to demonstrate the effectiveness of our approach, so the performances could be further improved with more advanced backbones.

Figure 2a shows the results of a multi-lane highway environment. We trained a RL-based AV model using the proximal policy optimization (PPO) algorithm[28] (see Section 6.7 in Supplementary Materials) and obtained its crash rate as 1.36×10^{-2} in NDE. We further implemented the responsibility-sensitive safety (RSS) model[30] as the default safety guard. Through millions of tests of the base AV model with RSS in NDE, we obtained the crash rate as 2.71×10^{-5} crash per test. Then, we trained SafeDriver using our approach without RSS (Fig.2a, red line). The results revealed a crash rate of 3.71×10^{-6}, making an 86.3% reduction in comparison to the base AV model. When considering only avoidable crashes, our approach demonstrated a remarkable 91.7% reduction. This suggests that our approach significantly enhances the overall safety performance by approximately one order of magnitude. We also evaluated the AV model with SafeDriver and RSS, yielding a crash rate of 7.79×10^{-6}, representing a 71.3% reduction compared to the base AV model with RSS (Fig.2a, purple line). While not as optimal as the model without RSS, due to the additional constraints introduced by RSS, our approach still improves overall safety performance. This is significant considering that many AV models may already have their safety guards or constraints in place. To demonstrate the effectiveness of our approach, we also compared our approach with the provably safe RL approach[31,32] (see Section 2.2.5 in Supplementary Information).

Navigating through roundabouts poses a significant challenge for AVs due to the intricate interactions among AVs and the diverse mix of surrounding vehicles. To evaluate the effectiveness of our approach in roundabout scenarios, we expanded our simulation experiments to a real-world four-armed roundabout[33] in Germany, known for its high traffic volume and complex intersections. We generated the realistic driving behaviors of background vehicles in NDE with a Transformer-based learning approach[28], leveraging the naturalistic driving data in the RounD dataset[33] (see Section 2.6 in Supplementary Information). Using an advanced AV model with a sampling trajectory generation algorithm[34] as our base AV model, we demonstrated that our approach significantly reduces the AV's crash rate by 74.5%, along with an 89.5% reduction in avoidable crash rates (Fig.2b).

We further test the performance of SafeDriver when AVs navigate continuously through urban environments. We utilized the digital twin of the Mcity test track, and selected two AV models as base models, one is the default AV model in SUMO[35] (the intelligent driving model (IDM)[36] and the SL2015 model[35] in SUMO) and another one is Autoware[25], which is widely recognized as the leading open-source automated driving system. Results show that our approach can reduce the overall crash rate of the SUMO AV model by 98.0% and reduce the avoidable crash rate by 98.9% (Fig.2c), representing a nearly two-orders-of-magnitude improvement. We integrated the identical SafeDriver into Autoware without any additional refinement, and results indicate a reduction in the crash rate from 1.07×10^{-6} to 1.07×10^{-7}, equating to a 90.0% enhancement in safety performance. This underscores the efficacy and adaptability of our approach across various AV models.

To further investigate the contributions of our developed techniques, we conducted ablation studies on the multi-lane highway environment including 1) no episodic data densification, 2) no state-

level data densification, 3) no near-miss episodes, 4) no retrospective data densification, 5) no near-miss and retrospective data densification, 6) no trajectory resampling, and 7) no state reconnection. With the same training steps such as 780 and 1650 steps, we compared the crash rates of SafeDriver in NDE as shown in Table 1. Results demonstrate the effectiveness of all developed techniques. Specifically, both the episodic data densification and state-level data densification contribute dramatically, while the retrospective data densification together with near-miss scenarios affects the performance significantly. Please see details in Section 2.4 in Supplementary Information.

To test the scalability and generalizability of our approach, we implemented a unified SafeDriver on the nuPlan benchmark[37], recognized as the world's first extensive planning benchmark for autonomous vehicles, which contains 1200h of human driving data from 4 cities across the US and Asia with widely varying traffic patterns (Boston, Pittsburgh, Las Vegas and Singapore) (Fig.3a). The goal of SafeDriver is to utilize a learning-based planner to assist the base model in navigating through safety-critical scenarios encountered in urban driving. We selected the state-of-the-art (SOTA) planner PDM-Hybrid[38] as our base model to demonstrate the effectiveness of SafeDriver over the SOTA base model. As shown in Fig.3b, if the PDM-Hybrid model predicts a collision within 2 seconds, SafeDriver takes control; otherwise, the base model manages the ego vehicle. Results indicate that SafeDriver reduces the total number of collisions by 21.7% and decreases AV-responsible crashes by 29.2%, compared with the base model. Illustration of two examples can be found in Fig.3c-f. Please see details in Section 2.5 in Supplementary Information.

Field testing results

To demonstrate the effectiveness of our approach for real-world AVs, we outfitted a Lincoln MKZ hybrid with Autoware as the base AV model (Fig.4a) and conducted the experiments in the physical test tracks at Mcity (Fig.4b) with a mixed-reality testing platform (Fig.4c). One additional challenge for real AV training is caused by the so-called simulation-to-reality gap, where AV models developed in simulations may not seamlessly translate to real-world performance. In this work, we bridged the gap by iteratively improving the simulation models. Specifically, we utilized the same Autoware system in simulation as employed in the actual AV, mimicked the delay and latency characteristics of the real AV, and compensated for acceleration and deceleration effects induced by road slopes (Supplementary Fig.4). Results demonstrated that, with this approach, the SafeDriver effectively reduces the crash rate of the real AV in the test track from 1.44×10^{-6} to 1.42×10^{-7}, making an impressive 90.1% improvement in the AV's safety performance (see Fig.4d, e and Supplementary Fig.5). Recognizing that some crashes are unavoidable due to aggressive behaviors of background vehicles, we further assessed the crash rate of avoidable crashes, revealing a 98.8% reduction through our approach. For a more in-depth understanding, additional case studies are presented in Figure 4f.

Discussion

The dense learning approach proposed in this work is theoretically applicable to the reinforcement learning problem that aims to minimize the expectation of a rare event with an underlying distribution, although further investigations are required to validate these applications (see Problem 1 in Supplementary Information). This problem has been a long-standing challenge in multiple fields associated with safety-critical autonomous systems, such as AVs, medical robots, and aerospace systems. Despite the significant advancements in AI systems such as AlphaGo[10]

and GPT[39], their application in safety-critical domains remains difficult due to the low tolerance for issues like hallucinations[40]. The dense learning approach opens the door for leveraging AI techniques in the development of such systems. In this work, we demonstrate the effectiveness of our approach in the safe driving task of AVs and enable continuous enhancement of AVs' performance in rare safety-critical scenarios (see Supplementary Fig.8). More research is needed to extend our approach to more generic safety-critical autonomous systems. We note that how to apply the dense learning idea to supervised learning needs to be further investigated.

While AV companies have already collected a large amount of data, strategies for effectively utilizing this data to enhance AV safety performance remains unclear. Due to the rarity of safety-critical events, the information of these events is usually hidden within a vast amount of noisy data. One might think that this issue could be resolved by focusing on a small set of data related to these rare events. However, prior to this study, there was no theoretical foundation supporting this intuition, which greatly limited its effectiveness and could even lead to misleading results. Our dense learning approach addresses this challenge with a thorough theoretical analysis for defining, identifying, and effectively leveraging the set of informative data. This is particularly significant considering AVs rely on larger neural networks and require more informative data for effective training.

A limitation of our work lies in the focus primarily on moving objects and road geometry of driving environment[41], which are crucial factors influencing AV decision-making. Addressing the CoR challenges associated with additional driving environment factors (such as weather conditions) and internal AV factors[42] necessitates further exploration. We are confident in the extensibility of our approach to consider these factors by incorporating domain knowledge from relevant fields. Moreover, as demonstrated in Fig.2a, our approach seamlessly integrates with existing rules (such as RSS) or model-based approaches, offering compatibility and the potential to leverage established techniques for managing diverse driving environment factors.

We note that our approach cannot address unknown unsafe issues that are not included in any training dataset or reinforcement learning environment. For these issues, our approach needs to be integrated with other techniques. For example, one potential way is to generate such scenarios with generative methods[43], while another way is to search such scenarios with falsification techniques[14]. Our approach is complementary to these techniques. Moreover, the falsification techniques can also be utilized in the counterfactual simulation. There are also works attempting to guarantee the safety of AV for all situations, like provably safe reinforcement learning[31,32], yet they usually rely upon assumptions with the environment model[44] such as behaviors of other traffic participants and vehicle dynamics. For example, reachable set calculations are often based on max vehicle acceleration/deceleration rate, which might be impacted by road surface conditions under different weather conditions that are difficult to predict precisely.

Methods

Formulation of AV training problem

This section describes the formulation of the AV training problem. Denote the variables of the driving environment as $\boldsymbol{X} = [\boldsymbol{S}(0), \boldsymbol{A}(0), \boldsymbol{S}(1), \boldsymbol{A}(1), \cdots, \boldsymbol{S}(T)] \in \boldsymbol{\Omega}$, where $\boldsymbol{S}(k)$ denotes the states (position, speed, heading, etc.) of the AV and surrounding background vehicles or other road users at the kth time step, $\boldsymbol{A}(k)$ denotes the maneuvers of surrounding background vehicles or other road users at the kth time step, T denotes the total time steps of each driving episode, and $\boldsymbol{\Omega}$ denotes the space of the variables $\boldsymbol{X}$. The goal of AV training is to optimize the AV policy $\pi \in \Pi$ as

$$\max_{\pi \in \Pi} \mathbb{E}_{P_\pi}[f(\pi, \boldsymbol{X})], \tag{1}$$

where $f(\cdot)$ denotes the objective function of AV training and $\boldsymbol{X}$ follows an underlying joint distribution $P_\pi(\boldsymbol{X})$ in NDE. To keep the notation simple, we leave it implicit in all cases that π is a function of neural network parameters θ. For safety training, the objective could be minimizing the overall crash rate as

$$\min_{\pi \in \Pi} \mathbb{E}_{P_\pi}[P(\omega(\boldsymbol{X}) = 1)], \tag{2}$$

where $\omega(\boldsymbol{X}) = 1$ denotes the objective event (e.g., vehicle crash), and $P(\omega(\boldsymbol{X}) = 1)$ denotes the event probability of the AV policy π in the driving environment $\boldsymbol{X}$.

Curse of rarity for AV training

To solve the AV training problem, deep learning approaches have been widely applied to handle the high variability and complexity of $\boldsymbol{X}$. The key is to estimate the policy gradient at each training step for the current policy π as

$$\boldsymbol{\Psi} \overset{\text{def}}{=} \mathbb{E}_{P_\pi}[\nabla_\pi P(\omega(\boldsymbol{X}) = 1)], \tag{3}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^d$, d is the dimension of the gradient, and then the policy could be updated accordingly. For non-trivial AV policies, however, since the objective event $\omega(\boldsymbol{X}) = 1$ is a rare event, namely, $\mathbb{E}_{P_\pi}[P(\omega(\boldsymbol{X}) = 1)]$ is a near-zero value, most $P(\omega(\boldsymbol{X}_i) = 1)$ and $\nabla_\pi P(\omega(\boldsymbol{X}_i) = 1)$ are near zero. Therefore, estimating the policy gradient is essentially a rare-event estimation problem. If directly using the data collected in NDE, it is essentially a Monte Carlo estimation approach[45] as

$$\widehat{\boldsymbol{\Psi}}_{\text{MC}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\pi P(\omega(\boldsymbol{X}_i) = 1), \boldsymbol{X}_i \sim P_\pi(\boldsymbol{X}_i), \tag{4}$$

where n is the number of samples used at each training step, and $\nabla_\pi P(\omega(\boldsymbol{X}_i) = 1)$ could be estimated through the policy gradient theorem[46]. According to the properties of Monte Carlo estimation[45], however, the estimator $\widehat{\boldsymbol{\Psi}}_{\text{MC}}$ would suffer from a large variance, which severely hinders the learning effectiveness. Moreover, as $\widehat{\boldsymbol{\Psi}}_{\text{MC}}$ is usually near zero, the signal-to-noise ratio is also low, which makes the problem even worse. We call this problem the CoR[7].

Let us elaborate the CoR more rigorously. Without loss of generality, we define the set of non-informative samples as $\boldsymbol{\Phi}_{\text{non}} \subset \boldsymbol{\Omega}$ and informative samples as $\boldsymbol{\Phi}_{\text{in}} \subset \boldsymbol{\Omega}$, their indicator functions $\mathbb{I}_{\boldsymbol{\Phi}_{\text{non}}}$ and $\mathbb{I}_{\boldsymbol{\Phi}_{\text{in}}}$, and an estimator of the policy gradient $\boldsymbol{\Psi}$ that only utilizes the informative samples as

$$\widehat{\boldsymbol{\Psi}}_{\text{in}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\pi P(\omega(\boldsymbol{X}_i) = 1)\mathbb{I}_{\boldsymbol{\Phi}_{\text{in}}}, \boldsymbol{X}_i \sim P_\pi(\boldsymbol{X}_i). \tag{5}$$

Then we have the following Lemma 1, and the proof can be found in ref.[7].

Lemma 1

If $\boldsymbol{\Phi}_{\text{non}}$ and $\boldsymbol{\Phi}_{\text{in}}$ satisfy the following conditions:

(1) $\boldsymbol{\Phi}_{\text{non}} \cap \boldsymbol{\Phi}_{\text{in}} = \emptyset, \boldsymbol{\Phi}_{\text{non}} \cup \boldsymbol{\Phi}_{\text{in}} = \boldsymbol{\Omega}$;

(2) $\mathbb{E}_{P_\pi}\left[\nabla_\pi P(\omega(\boldsymbol{X}) = 1)\mathbb{I}_{\boldsymbol{\Phi}_{\text{non}}}\right] = \mathbf{0}$;

then we have the following properties:

(1) $\mathbb{E}_{P_\pi}\left[\widehat{\boldsymbol{\Psi}}_{\text{MC}}\right] = \mathbb{E}_{P_\pi}\left[\widehat{\boldsymbol{\Psi}}_{\text{in}}\right]$;

(2) $\sigma^2_{P_\pi}(\widehat{\Psi}^{(k)}_{\text{MC}}) \geq \sigma^2_{P_\pi}(\widehat{\Psi}^{(k)}_{\text{in}}), \forall k = 1, \dots d$; and

(3) $\sigma^2_{P_\pi}(\widehat{\Psi}^{(k)}_{\text{MC}}) \geq \rho^{-1}_{\boldsymbol{\Phi}_{\text{in}}}\sigma^2_{P_\pi}(\widehat{\Psi}^{(k)}_{\text{in}}), \forall k = 1, \dots d$, with the assumption

$$\mathbb{E}_{P_\pi}\left[G^2_k(\omega(\boldsymbol{X}) = 1)\mathbb{I}_{\boldsymbol{\Phi}_{\text{in}}}\right] = \mathbb{E}_{P_\pi}[G^2_k(\omega(\boldsymbol{X}) = 1)]\mathbb{E}_{P_\pi}\left[\mathbb{I}_{\boldsymbol{\Phi}_{\text{in}}}\right], \forall k = 1, \dots, d, \tag{6}$$

where $\widehat{\Psi}^{(k)}_{\text{MC}}$ and $\widehat{\Psi}^{(k)}_{\text{in}}$ are kth components of $\widehat{\boldsymbol{\Psi}}_{\text{MC}}$ and $\widehat{\boldsymbol{\Psi}}_{\text{in}}$, respectively, $\rho_{\boldsymbol{\Phi}_{\text{in}}} = \mathbb{E}_{P_\pi}\left(\mathbb{I}_{\boldsymbol{\Phi}_{in}}\right) \in [0,1]$ is the expected proportion of the informative samples $\boldsymbol{\Phi}_{\text{in}}$ in all samples with the sampling distribution $P_\pi(\boldsymbol{X})$, and $\boldsymbol{G}(\omega(\boldsymbol{X}) = 1) \stackrel{\text{def}}{=} \nabla_\pi P(\omega(\boldsymbol{X}) = 1)$ is a random vector with $\boldsymbol{G}(\omega(\boldsymbol{X}) = 1) = [G_1(\omega(\boldsymbol{X}) = 1), \dots, G_d(\omega(\boldsymbol{X}) = 1)] \in \mathbb{R}^d$, where d is the dimension of the parameters of the policy network π.

Remark 1. For AV safety training, the proportion of informative samples in all samples could be very small, in the order of 10^{-4}~10^{-6} or fewer, due to the rarity of safety-critical events in NDE. As the policy gradient $\nabla_\pi P(\omega(\boldsymbol{X}) = 1)$ is mainly determined by the parameters of neural networks, it could exhibit a stationary uncertainty that is independent of the set $\boldsymbol{\Phi}_{\text{in}}$. This is particularly true at the beginning of the learning process when the parameters are relatively random. Consequently, the assumption in Property 2 of Lemma 1 could be approximately satisfied, particularly at the beginning of the learning process. Therefore, the estimation variance of the traditional deep learning approaches based on Monte Carlo estimation could be very large.

Remark 2. Lemma 1 also indicates that if the set of informative samples $\boldsymbol{\Phi}_{\text{in}}$ could be identified, estimating the policy gradient utilizing only the informative samples has great potential to reduce the learning variance without loss of unbiasedness, thereby overcoming the CoR. However, how to define, identify, and effectively leverage informative samples for AV safety training is challenging. Prior to this work, this process is largely intuitive, for instance, many existing approaches only emphasize the crash event data or falsified cases[14], discussed as follows.

Learning from crash only

Many existing approaches used to tackle the CoR challenge primarily focus on learning from the data where AVs fail. To be more specific, the estimator of these approaches can be represented as

$$\widehat{\boldsymbol{\Psi}}_{\text{Fail}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\pi P(\omega(\boldsymbol{X}) = 1)\mathbb{I}_{\boldsymbol{\Phi}_{\text{F}}}, \boldsymbol{X}_i \sim P_\pi(\boldsymbol{X}_i), \tag{7}$$

where $\mathbb{I}(\cdot)$ is an indicator function and $\boldsymbol{\Phi}_{\text{F}}$ is the set of data samples where AV fails. However, these approaches are largely intuitive without theoretical foundation. Consequently, the definition of the set $\boldsymbol{\Phi}_{\text{F}}$ does not satisfy the conditions of $\boldsymbol{\Phi}_{\text{in}}$ in Lemma 1, resulting in a severe learning biasedness, that is,

$$\mathbb{E}_{P_\pi}\left[\widehat{\boldsymbol{\Psi}}_{\text{Fail}}\right] \neq \mathbb{E}_{P_\pi}\left[\widehat{\boldsymbol{\Psi}}_{\text{MC}}\right]. \tag{8}$$

Without the unbiasedness guaranteed, the learning process could become misleading. That is the reason why learning-from-crash-only could suffer from the seesaw effect as illustrated in Section 2.3 of Supplementary Information.

Dense learning approach for AV training

In this work, we propose the dense learning approach to overcome the CoR challenge for AV training. The key to dense learning is to define and identify the informative data samples for AV safety training, satisfying the conditions of $\boldsymbol{\Phi}_{\text{in}}$ in Lemma 1. To achieve this goal, we integrate the Lemma 1 with importance sampling theory to derive the optimal distribution of data samples. Although this optimal distribution is unavailable in practice, it provides guidance of defining and identifying the informative data for AV training. Then, we sampled the data based on the distribution, rejected the non-informative samples, and effectively leveraged the informative ones, resulting in the dense learning approach for AV training.

First, we derive the optimal sampling distribution $q_\pi^*(\boldsymbol{X})$ and the corresponding policy estimator $\widehat{\boldsymbol{\Psi}}_{\text{IS}}^*$ based on the importance sampling theory[27,45,47] as

$$\widehat{\boldsymbol{\Psi}}_{\text{IS}}^* \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{P_\pi(\boldsymbol{X}_i)}{q_\pi^*(\boldsymbol{X}_i)}\nabla_\pi P(\omega(\boldsymbol{X}_i)=1)\right], \boldsymbol{X}_i \sim q_\pi^*(\boldsymbol{X}_i), \tag{9}$$

$$q_\pi^*(\boldsymbol{X}) \propto \|\nabla_\pi P(\omega(\boldsymbol{X})=1)\|_2 P_\pi(\boldsymbol{X}), \tag{10}$$

where $\|\cdot\|_2$ denotes the $l2$ norm of the vector, and the symbol $\propto$ means 'proportional to'. Eq. (10) provides the insights that the most informative data for AV training should contain all safety-critical events, both successful and failed, with probabilities proportional to their contributions to the AVs' policy gradient as well as their exposure frequencies in the real world. The major challenge of $\widehat{\boldsymbol{\Psi}}_{\text{IS}}^*$ is that $\|\nabla_\pi P(\omega(\boldsymbol{X})=1)\|_2$ cannot be accurately calculated for each episode $\boldsymbol{X}$ in practice and is dependent of the policy π, which is changing during the training process. Some existing studies tried to calculate $\nabla_\pi P(\omega(\boldsymbol{X})=1)$ using the policy gradient theorem (see ref.[48] for example), which introduces additional constraints and severely limits the effectiveness and applicability of the approach.

To address this challenge, we propose to utilize Eq. (9-10) as a guidance to define and identify the informative samples for the dense learning approach. Therefore, instead of trying to calculate the exact values with severely limited applicability, we choose to estimate $\|\nabla_\pi P(\omega(\boldsymbol{X})=1)\|_2$ as a classification task. Specifically, we estimate $\|\nabla_\pi P(\omega(\boldsymbol{X})=1)\|_2$ as

$$\|\nabla_\pi P(\omega(\boldsymbol{X})=1)\|_2 \approx \mathbb{I}_{\boldsymbol{X}\in\boldsymbol{\Phi}} \overset{\text{def}}{=} \begin{cases} 1, & \boldsymbol{X}\in\boldsymbol{\Phi} \\ 0, & \boldsymbol{X}\notin\boldsymbol{\Phi} \end{cases}, \tag{11}$$

$$\boldsymbol{\Phi} \overset{\text{def}}{=} \{\boldsymbol{X}\in\boldsymbol{\Omega}: \mathbb{I}(\omega(\boldsymbol{X})=1)=1 \; and \; P(\omega(\boldsymbol{X})=1)<1 \; or \; \mathbb{I}(\omega(\boldsymbol{X})=1)=0 \; and \; P(\omega(\boldsymbol{X})=1)>0\},$$

where an avoidability analysis is conducted to approximate if a crash is avoidable ($P(\omega(\boldsymbol{X})=1)<1$), and a safety metric was selected to approximate if a non-crash is a near-miss ($P(\omega(\boldsymbol{X})=1)>0$). Specifically, we identified a crash event as avoidable if an evasive trajectory is still feasible after the vehicle state is being identified as safety-critical (see Section 2.8 in Supplementary Information). And we identified a non-crash event as a near-miss if the minimum relative distance between the AV and background vehicles is below a pre-determined threshold (that is, 2.5 m). A sensitivity analysis of the threshold can be found in Supplementary Fig.6.

Then, we obtain the estimator according to Eq. (9) as

$$\widehat{\boldsymbol{\Psi}}_{\text{IS}} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\nabla_\pi P(\omega(\boldsymbol{X}_i)=1), \boldsymbol{X}_i \sim \hat{q}_\pi(\boldsymbol{X}_i), \tag{12}$$

where

$$\hat{q}_\pi(\boldsymbol{X}_i) \overset{\text{def}}{=} \frac{\mathbb{I}_{\boldsymbol{X}_i\in\boldsymbol{\Phi}}}{\sum_{i=1}^{N}\mathbb{I}_{\boldsymbol{X}_i\in\boldsymbol{\Phi}}P_\pi(\boldsymbol{X}_i)}P_\pi(\boldsymbol{X}_i). \tag{13}$$

Here N denotes the total amount of data in the collected dataset.

We then have Theorem 1, and the proof can be found in Section 2.1.3 of Supplementary Information.

Theorem 1
The estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$ has the following properties:
(1) $\mathbb{E}_{\hat{q}_\pi}\left[\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}\right] = g_N^{-1}\mathbb{E}_{P_\pi}\left[\widehat{\boldsymbol{\Psi}}_{\mathrm{MC}}\right]$;
(2) $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}} = \widehat{\boldsymbol{\Psi}}_{\mathrm{MC}}$;
(3) $\Delta^2_{\hat{q}_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{IS}}\right) \le g_N \Delta^2_{P_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{MC}}\right), \forall k = 1, \dots d$;
(4) $\Delta^2_{\hat{q}_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{IS}}\right) \le g_N \rho_{\boldsymbol{\Phi}} \Delta^2_{P_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{MC}}\right), \forall k = 1, \dots d$, with the assumption

$$\mathbb{E}_{P_\pi}[G_k^2(\omega(\boldsymbol{X}) = 1)\mathbb{I}_{\boldsymbol{X}\in\boldsymbol{\Phi}}] = \mathbb{E}_{P_\pi}[G_k^2(\omega(\boldsymbol{X}) = 1)]\mathbb{E}_{P_\pi}[\mathbb{I}_{\boldsymbol{X}\in\boldsymbol{\Phi}}], \forall k = 1, \dots, d; \quad (14)$$

where $g_N \overset{\text{def}}{=} \sum_{i=1}^{N} \mathbb{I}_{\boldsymbol{X_i}\in\boldsymbol{\Phi}} P_\pi(\boldsymbol{X_i}) \in (0,1]$, $\widehat{\Psi}^{(k)}_{\mathrm{IS}}$ is the k th components of $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$, $\Delta_{\hat{q}_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{IS}}\right) \overset{\text{def}}{=} \sigma_{\hat{q}_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{IS}}\right)/\mathbb{E}_{\hat{q}_\pi}\left[\widehat{\Psi}^{(k)}_{\mathrm{IS}}\right]$ and $\Delta_{P_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{MC}}\right) \overset{\text{def}}{=} \sigma_{P_\pi}\left(\widehat{\Psi}^{(k)}_{\mathrm{MC}}\right)/\mathbb{E}_{P_\pi}\left[\widehat{\Psi}^{(k)}_{\mathrm{MC}}\right]$ are coefficients of variation, and $\rho_{\boldsymbol{\Phi}} \overset{\text{def}}{=} \mathbb{E}_{P_\pi}[\mathbb{I}_{\boldsymbol{X}\in\boldsymbol{\Phi}}] \in [0,1]$ is the expected proportion of the informative samples $\boldsymbol{\Phi}$ in all samples with the sampling distribution $P_\pi(\boldsymbol{X})$.

Remark 4. The first property of Theorem 1 ensures the unbiasedness of AV training regarding the direction of the gradient, which is critical for overcoming the seesaw effect. We note that g_N is a constant and as the learning rate is usually adjustable, such a constant will not affect the learning process.

Remark 5. The second property indicates that the new estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$ has the same calculation equation as the Monte Carlo estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{MC}}$ (see Eq. (4) and Eq. (12)), and the only difference is the data sampling distribution. This is significant as our approach could leverage all techniques that can be used for the Monte Carlo estimator (that is widely used for deep learning), without introducing any additional limitations. Therefore, our approach could leverage the advances in deep learning but with a dramatically smaller learning variance, which ensures the effectiveness and applicability of our approach.

Remark 6. The third and fourth properties are obtained as $\boldsymbol{\Phi}$ is a realization of defining and identifying the informative samples for the AV safety training task. Therefore, as indicated in Lemma 1, the coefficient of variation of $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$ is dramstically smaller than that of the Monte Carlo approach. Moreover, $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$is an approximation of $\widehat{\boldsymbol{\Psi}}^{*}_{\mathrm{IS}}$ that has the minimum variance. Both indicate that $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$ is an efficient policy gradient estimator for deep learning approaches associated with rare events. As the parameters of neural networks are usually randomly initialized, the assumption in Eq. (14) could be approximately satisfied, particularly at the beginning of the learning process, as discussed in Remark 1.

Remark 7. While the following investigation focuses on the deep reinforcement learning (DRL) tasks of AV safety training, Theorem 1 is also applicable to more generic deep learning tasks. Therefore, our approach has the great potential to address the CoR challenge in generic deep learning tasks associated with rare events.

To further integrate the estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$ with DRL approaches, we obtain the policy gradient estimator of DRL as

$$\widehat{\boldsymbol{\Psi}}_{\mathrm{DRL}} \overset{\mathrm{def}}{=} \hat{Q}_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t) \frac{\nabla \pi(\boldsymbol{A}_t|\boldsymbol{S}_t)}{\pi(\boldsymbol{A}_t|\boldsymbol{S}_t)}, \quad (\boldsymbol{S}_t, \boldsymbol{A}_t) \sim P_{\pi}(\boldsymbol{X}), \tag{15}$$

where $\boldsymbol{S}_t$ and $\boldsymbol{A}_t$ are samples of the state and action following the distribution of episodes $P_{\pi}(\boldsymbol{X})$ and under the policy π, $Q_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t)$ denotes the state-action value, $\hat{Q}_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t)$ is an unbiased estimation of $Q_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t)$, i.e., $\mathbb{E}_{\pi}\left[\hat{Q}_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t)\right] = Q_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t)$. As discussed in Remark 5, our approach is compatible with existing techniques that can be used for the Monte Carlo estimator. Therefore, we integrate the estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{IS}}$ with the dense deep reinforcement learning (D2RL) approach that was developed in our previous study for AV testing[8]. Although the AV training problem is different from the AV testing problem, the D2RL approach can still be beneficial to further reduce the learning variance by removing the non-safety-critical states and connecting the safety-critical ones. Therefore, we obtain a new estimator as

$$\widehat{\boldsymbol{\Psi}}_{\mathrm{Dense}} \overset{\mathrm{def}}{=} \hat{Q}_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t) \frac{\nabla \pi(\boldsymbol{A}_t|\boldsymbol{S}_t)}{\pi(\boldsymbol{A}_t|\boldsymbol{S}_t)} \mathbb{I}_{\boldsymbol{S}_t \in \mathbb{S}_c}, \quad (\boldsymbol{S}_t, \boldsymbol{A}_t) \sim \hat{q}_{\pi}(\boldsymbol{X}), \tag{16}$$

where $\mathbb{S}_c \overset{\mathrm{def}}{=} \left\{\boldsymbol{s} | \mathbb{E}_{\pi}\left(q_{\pi}(\boldsymbol{s}, \boldsymbol{a})\right) \neq q_{\pi}(\boldsymbol{s}, \boldsymbol{a}), \exists \boldsymbol{a}\right\}$ denotes the set of safety-critical states as defined in ref.[8]. In this study, we utilized the learned safety metric to identify the safety-critical states.

We then have Theorem 2, and the proof can be found in Section 2.1.7 of Supplementary Information.

Theorem 2

The estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{Dense}}$ has the following properties:

(1) $\mathbb{E}_{\hat{q}_{\pi}}\left[\widehat{\boldsymbol{\Psi}}_{\mathrm{Dense}}\right] = g_N^{-1} \mathbb{E}_{P_{\pi}}\left[\widehat{\boldsymbol{\Psi}}_{\mathrm{DRL}}\right]$;

(2) $\Delta^2_{\hat{q}_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) \leq g_N \Delta^2_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right), \forall k = 1, \ldots d$;

(3) $\Delta^2_{\hat{q}_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) \leq g_N \rho_{\boldsymbol{\Phi}} \Delta^2_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right), \forall k = 1, \ldots d$, with the assumption in Eq. (14);

(4) $\Delta^2_{\hat{q}_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) \leq g_N \rho_{\boldsymbol{\Phi}_D} \Delta^2_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right), \forall k = 1, \ldots d$, with the assumption

$$\mathbb{E}_{P_{\pi}}\left[\left(\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right)^2 \mathbb{I}_{\boldsymbol{X} \in \boldsymbol{\Phi}_D}\right] = \mathbb{E}_{P_{\pi}}\left[\left(\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right)^2\right] \mathbb{E}_{P_{\pi}}\left[\mathbb{I}_{\boldsymbol{X} \in \boldsymbol{\Phi}_D}\right], \forall k = 1, \ldots, d; \tag{17}$$

where $\widehat{\Psi}^{(k)}_{Dense}$ and $\widehat{\Psi}^{(k)}_{DRL}$ are kth components of $\widehat{\boldsymbol{\Psi}}_{\mathrm{Dense}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathrm{DRL}}$, respectively, $\Delta_{\hat{q}_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) \overset{\mathrm{def}}{=} \sigma_{\hat{q}_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) / \mathbb{E}_{\hat{q}_{\pi}}\left[\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right]$, $\Delta_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right) \overset{\mathrm{def}}{=} \sigma_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right) / \mathbb{E}_{P_{\pi}}\left[\widehat{\Psi}^{(k)}_{\mathrm{DRL}}\right]$, $\Delta_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) \overset{\mathrm{def}}{=} \sigma_{P_{\pi}}\left(\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right) / \mathbb{E}_{P_{\pi}}\left[\widehat{\Psi}^{(k)}_{\mathrm{Dense}}\right]$ are coefficients of variance, and $\rho_{\boldsymbol{\Phi}_D} \overset{\mathrm{def}}{=} \mathbb{E}_{P_{\pi}}\left(\mathbb{I}_{\boldsymbol{\Phi}_D}\right) \in [0,1]$ is the expected proportion of the informative states $\boldsymbol{\Phi}_D \overset{\mathrm{def}}{=} \{\boldsymbol{S}_t \in \boldsymbol{X} : \boldsymbol{S}_t \in \mathbb{S}_c \text{ and } \boldsymbol{X} \in \boldsymbol{\Phi}\}$ in all sampled states with the sampling distribution $P_{\pi}(\boldsymbol{X})$ and under the policy π.

Remark 8. Theorem 2 indicates that $\widehat{\boldsymbol{\Psi}}_{\mathrm{Dense}}$ is an unbiased and efficient policy gradient estimator of the DRL approaches, which is critical for overcoming the seesaw effect and CoR for AV training. As $\rho_{\boldsymbol{\Phi}_D}$ is smaller than $\rho_{\boldsymbol{\Phi}}$, the estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{Dense}}$ could further reduce the variance of policy gradient, compared with the estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{DRL}}$

Remark 9. As the policy gradient $\widehat{\boldsymbol{\Psi}}_{\mathrm{DRL}} \overset{\mathrm{def}}{=} \hat{Q}_{\pi}(\boldsymbol{S}_t, \boldsymbol{A}_t) \frac{\nabla \pi(\boldsymbol{A}_t|\boldsymbol{S}_t)}{\pi(\boldsymbol{A}_t|\boldsymbol{S}_t)}$ is mainly determined by the parameters of neural networks, it could exhibit a stationary uncertainty that is independent of the set $\boldsymbol{\Phi}_D$. This is particularly true at the beginning of the learning process when the parameters are relatively random. Therefore, the assumptions in (17) could be approximately satisfied. We note

that the assumptions are primarily for the theoretical analysis to be clean and are not strictly required in practice.

Additional data densification approaches

Offline resample mechanism. To realize the estimator $\widehat{\boldsymbol{\Psi}}_{\mathrm{DRL}}$, we design an offline resampling mechanism to resample and redistribute the collected episodic data according to the distribution $\hat{q}_{\pi}(\boldsymbol{X_i})$ as in Eq. (13). An avoidability analysis was developed to help approximate the distribution $\hat{q}_{\pi}(\boldsymbol{X_i})$. This mechanism is applicable for data collected through different venues such as simulations, test tracks, and public roads. It also provides compatibility to integrate with purposely generated testing environments (such as ITE) where data is collected by a different distribution, and only the offline resampling distribution needs to be modified accordingly.

Learned safety metric. Most existing safety metrics are primarily based on trajectory prediction of background road users with assumptions on their driving behaviors, which limit their effectiveness and generalization capability. This is particularly true for AV safety training, as the AV policy is changing during the training process. To address this issue, we designed a learned safety metric to identify the safety-critical states with both high precision and recall rates. We utilized the avoidability analysis approach to provide ground-truth labels for the large-scale naturalistic trajectory datasets. Specifically, all snapshots where a crash is unavoidable with evasive driving behaviors are labeled as positive, while others are labeled as negative. Since unavoidable snapshots are rare, existing training approaches suffer from the severely unbalanced data issue[49]. Instead of rebalancing the data before the training, we proposed to resample positive and negative snapshots from all the data to create balanced training data batches in each training step. In this way, the training data in each training step is rebalanced and densified. Results demonstrate that our approach enables a much better precision and recall performance than existing safety metrics (see Supplementary Fig.1).

Retrospective data densification. Ideally, as the policy π is changing during the training process, the episodic data should be re-collected according to the new policy at each training step. However, this is severely inefficient and even infeasible in practice, particularly for on-road data collection. Therefore, an off-policy mechanism is needed to fill the policy gap. In this study, we design a retrospective data densification mechanism to re-evaluate the collected data for SafeDriver through a counterfactual simulation. This mechanism could update the values of $\mathbb{I}(\omega(\boldsymbol{X}) = 1)$ and $P(\omega(\boldsymbol{X}) = 1)$ in Eq. (11), so the data resampling distribution $\hat{q}_{\pi}(\boldsymbol{X_i})$ could be updated for the new AV policy, which reduces the overall policy gap.

Counterfactual simulation. We utilized the counterfactual simulation[50] to re-evaluate the safety performance of the AV policy and collect the training data. In each simulation, the base AV model with SafeDriver is responsible for the planning task of the simulated AV. SafeDriver handles all the safety-critical states identified by the learned safety metric; otherwise, the base AV model takes control. Calibrated non-linear bicycle models are used to simulate the motion of the AV with control command (i.e., steering angle and acceleration) inputs from the planning tasks. Specifically, we employed the dynamic bicycle model when AV speed is greater than 20 m/s and the kinematic bicycle model for other cases, leveraging their modeling strengthens for different situations[51]. The trajectories of the background vehicles are replayed based on the recorded data until the recorded trajectories reach their end. For crash trajectories, if SafeDriver could avoid the original collision in the counterfactual simulation, we need to re-simulate the driving behaviors of the background

vehicles after the original collision moment to interact with SafeDriver, until the collision could be avoided completely or a new collision occurs. Specifically, we simulated the background vehicles by a predefined driving model, such as the IDM and SL2015 provided by SUMO[35]. The simulation will be terminated either upon collision or after a certain duration has elapsed.

Intelligent testing environment for efficient data collection

As $\|\nabla_\pi P(\omega(\boldsymbol{X})=1)\|_2$ is zero for most episodic data $\boldsymbol{X}$ and non-trivial AV policies π, most data collected in NDE has little information for AV training and will be rejected by the offline sampling mechanism as Eq. (13). To improve the data collection efficiency, we utilized the intelligent testing environment (ITE) that was developed in our previous study[8], in that AI-based background agents are trained to test AVs in an accelerated mode. Specifically, ITE will provide a dataset with a new distribution $\boldsymbol{X}\sim q_{\mathrm{ITE}}(\boldsymbol{X})$, each episode of data with a weight $W_{\mathrm{ITE}}(\boldsymbol{X})=P_\pi(\boldsymbol{X})/q_{\mathrm{ITE}}(\boldsymbol{X})$, which contains much more valuable safety-critical episodes. To leverage this dataset, the safety-critical episodes are resampled by the distribution proportional to their weights $W_{\mathrm{ITE}}(\boldsymbol{X})$ and then resampled according to Eq. (13). We found out that, with the dense learning approach, ITE could accelerate the collection of safety-critical data by multiple orders of magnitude, which dramatically accelerates the training process of AVs' overall safety performance (see Supplementary Fig.2). We believe that this approach opens the door for integrating AV testing and training together, accelerating both fields, which deserves further investigation.

Training settings of dense learning

We implemented the dense learning approach in highways, roundabouts, and urban environments, respectively. To keep the fidelity and efficiency of the simulations, we utilized the NDE simulation and ITE as described in Section 2.6 and 2.7 of Supplementary Information. We applied the PPO algorithm developed on the RLLIB 1.11.0 platform[52] to parallelly train SafeDriver on 400 CPU cores and 2800 GB memory high-performance computation cluster at the University of Michigan, Ann Arbor. We created a three-layer fully connected neural network, consisting of 256 neurons in each layer, to represent the SafeDriver. The neural network received input data comprising the states of up to 6 background vehicles located within 30m from the AV, where its output is the AV's acceleration and steering angle. Specifically, for the highway environment, the acceleration spanned from -4 m/s² to 2 m/s², and the steering angle varied between -10 degrees and 10 degrees; for roundabout environments, the acceleration ranged from -8 m/s² to 4.5 m/s², and the steering angle covered a range from -45 degrees to 45 degrees; and for the urban environment, the acceleration exhibited a range between -7 m/s² and 2.5 m/s², while the steering angle was within the -10 degrees to 10 degrees range. We set the learning rate as 10^{-5} and the discount factor as 0.99. In each training iteration, a total of 50,000 timesteps were used to train the neural network for 30 times. For each driving episode, the reward function is set as -1 for an AV-involved crash and 0 for others.

Field testing settings

Autonomous vehicle with SafeDriver. The vehicle under test is a Lincoln MKZ hybrid from Mcity at the University of Michigan, equipped with various sensors, including OTXS RT3003 RTK GPS, PointGrey camera, Velodyne 32 channel LiDAR, Delphi radars, Xsens MTi GPS/IMU, etc. The AV also had the Nuvo-8208GC computer and the Dataspeed drive-by-wire system installed. We applied the open-source full-stack autonomous driving system, Autoware[25], as the base AV model. Specifically, after receiving the ego vehicle's position and velocity as well as background vehicles' information, the future path is generated based on OpenPlanner 1.13[53]. We

applied the pure pursuit algorithm to convert the planned trajectory into the velocity and yaw rate and then used a proportional-integral-derivative controller provided by DataSpeed Inc. to further convert them into the vehicle by-wire control commands, that is, steering angle, throttle, and brake percentages. When the learned safety metric alerts in safety-critical situations, SafeDriver receives the normalized observation and outputs the steering angle and acceleration, which are further converted to the throttle and brake percentages.

Mixed reality testing environment. We applied the mixed reality testing environment[54] to efficiently evaluate the safety performance of AVs at Mcity, which is one of the leading testing facilities for AV development. We forwarded the states of the background vehicles from the simulation world and the signal information from the physical testing track to the AV through the Internet. Simultaneously, the states of the AV as well as proxy objects and the signal information were synchronized into the simulation world, and the behaviors of virtual background vehicles were determined according to NDE or ITE. Besides, the virtual background vehicles were rendered and blended on the front camera's view using pyrender[55]. To accelerate the testing process for evaluating the crash rates of the AV in NDE, we utilized the ITE that was developed in our previous study[8]. A framework of the mixed reality testing framework can be found in Supplementary Fig.7.

Data availability

The raw datasets that we used for modeling the naturalistic driving environment come from the Safety Pilot Model Deployment (SPMD) program[56] and the Integrated Vehicle Based Safety System (IVBSS)[57] at the University of Michigan, Ann Arbor, as well as the RounD dataset[33]. These raw datasets are subject to the data access policies and licensing terms of their respective providers and are therefore not redistributed by the authors. Access to the SPMD and IVBSS datasets can be requested from the University of Michigan Transportation Research Institute, and access to the RounD dataset is available through its original repository, subject to the corresponding usage agreements.

The processed data generated in this study, which constitute the minimum dataset necessary to interpret, verify, and extend the findings reported in this article, have been deposited in the Zenodo repository and are publicly available at: https://zenodo.org/records/12735037. Source data supporting the figures and analyses presented in this paper are provided via Zenodo at: https://zenodo.org/records/12784827.

To further illustrate the methodology and qualitative performance of the proposed approach, a collection of Supplementary Videos is provided. All Supplementary Videos, including high-resolution files, are publicly available through Zenodo at https://zenodo.org/records/14837884. These videos demonstrate the learning paradigms, the learned safety metric, the intelligent testing environment, and the performance of SafeDriver across simulated and real-world scenarios, including highway, roundabout, and Mcity test track environments, as well as case studies on the nuPlan benchmark.

Code availability

The simulation software SUMO, the NDE models, the intelligent testing environment, the automated driving system Autoware, and the RLLib platform with the implemented PPO algorithm are publicly available, as described in the text and the relevant references[8,25,35,52,58]. The source

codes for the dense learning approach for SafeDriver is available at: https://zenodo.org/records/12735669 with the DOI:10.5281/zenodo.12735668.

References

1. "Science: Radio Auto". Time Magazine. Aug 10, 1925. Retrieved 29 September 2013.
2. Hirsch, J (2022) Reality check: $160 billion can't get autonomous vehicles on road, Automotive News. Automotive News, https://www.autonews.com/mobility-report/autonomous-vehicle-reality-check-after-160-billion-spent. Accessed on February 6, 2024.
3. Society of Automotive Engineers (2021) Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, https://www.sae.org/standards/content/j3016_202104/.
4. Thadani, T. (2023) How a robotaxi crash got Cruise's self-driving cars pulled from Californian roads, Washington Post, https://www.washingtonpost.com/technology/2023/10/28/robotaxi-cruise-crash-driverless-car-san-francisco/. Accessed on Feb. 6, 2024.
5. Shladover, S.E. and Nowakowski, C. Regulatory challenges for road vehicle automation: Lessons from the California experience. *Transp. Res. Part A Policy Pract*., 122, pp.125-133 (2019).
6. Zhang, Y., Kang, B., Hooi, B., Yan, S. and Feng, J. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 10795-10816 (2023).
7. Liu, H.X. and Feng, S. Curse of rarity for autonomous vehicles. *Nat. Commun.***15**, 4808 (2024).
8. Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S. and Liu, H.X., 2023. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953), pp.620-627.
9. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. and Petersen, S. Human-level control through deep reinforcement learning. *Nature*, 518(7540), pp.529-533 (2015).
10. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. Mastering the game of go without human knowledge. *Nature* 550, 354-359 (2017).
11. Brown, N., & Sandholm, T. Superhuman AI for multiplayer poker. *Science*, 365(6456), 885-890 (2019).
12. Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., ... & Kitano, H. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896), 223-228 (2022).
13. Cummings, M. L. Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings. *AI Mag.*, 42(1), 6-15 (2021).
14. Menghi, Claudio, et al. "ARCH-COMP 2023 Category Report: Falsification." Proceedings of 10th International Workshop on Applied Verification of Continuous and Hybrid Systems. Vol. 96. 2023.
15. Karpathy, A., Tesla Inc, 2021. System and method for obtaining training data. U.S. Patent Application 17/250,825.
16. Stocker, T.F., 1998. The seesaw effect. *Science*, 282(5386), pp.61-62.

17. Tang, H., Liu, J., Zhao, M. and Gong, X., 2020, September. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. *In Proceedings of the 14th ACM Conference on Recommender Systems* (pp. 269-278).
18. Zheng, S., Zhang, Y., Zhu, Y., Xi, C., Gao, P., Zhou, X. and Chang, K.C.C., 2023. GPT-Fathom: Benchmarking Large Language Models to Decipher the Evolutionary Path towards GPT-4 and Beyond. https://arxiv.org/abs/2309.16583.
19. Seshia, S.A., Sadigh, D. and Sastry, S.S., 2022. Toward verified artificial intelligence. *Commun. ACM*, 65(7), pp.46-55.
20. Pek, C., Manzinger, S., Koschi, M. and Althoff, M., 2020. Using online verification to prevent autonomous vehicles from causing accidents. *Nature Machine Intelligence*, 2(9), pp.518-528.
21. Krasowski, H., Thumm, J., Müller, M., Schäfer, L., Wang, X. and Althoff, M., 2023. Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking. Transactions on Machine Learning Research.
22. Brunke, L., Greeff, M., Hall, A.W., Yuan, Z., Zhou, S., Panerati, J. and Schoellig, A.P., 2022. Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems, 5, pp.411-444.
23. Cao, Z., Jiang, K., Zhou, W., Xu, S., Peng, H. and Yang, D., 2023. Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning. Nature Machine Intelligence, 5(2), pp.145-158.
24. Thomas, P.S., Castro da Silva, B., Barto, A.G., Giguere, S., Brun, Y. and Brunskill, E., 2019. Preventing undesirable behavior of intelligent machines. Science, 366(6468), pp.999-1004.
25. Kato, S., Tokunaga, S., Maruyama, Y., Maeda, S., Hirabayashi, M., Kitsukawa, Y., Monrroy, A., Ando, T., Fujii, Y. and Azumi, T. Autoware on board: Enabling autonomous vehicles with embedded systems. In 2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS), pp. 287-296 (2018).
26. Hsu, K.C., Hu, H. and Fisac, J.F., 2023. The safety filter: A unified view of safety-critical control in autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems, 7.
27. Tokdar, S.T. and Kass, R.E., 2010. Importance sampling: A review. Wiley Interdisciplinary Reviews: Computational Statistics, 2(1), pp. 54-60.
28. Yan, X., Zou, Z., Feng, S., Zhu, H., Sun, H. and Liu, H.X., 2023. Learning naturalistic driving environment with statistical realism. *Nature Communications*, 14(1), p.2037.
29. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347 (2017).
30. Kochdumper N, Krasowski H, Wang X, et al. Provably safe reinforcement learning via action projection using reachability analysis and polynomial zonotopes[J]. IEEE Open Journal of Control Systems, 2023, 2: 79-92.
31. Shalev-Shwartz, S., Shammah, S. and Shashua, A., 2017. On a formal model of safe and scalable self-driving cars. https://arxiv.org/abs/1708.06374.
32. Wang X, Althoff M. Safe reinforcement learning for automated vehicles via online reachability analysis[J]. IEEE Transactions on Intelligent Vehicles, 2023.
33. Krajewski, R., Moers, T., Bock, J., Vater, L. and Eckstein, L. September. The round dataset: A drone dataset of road user trajectories at roundabouts in Germany. In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC) (pp. 1-6). (IEEE, 2020).
34. Elbanhawi, M. and Simic, M., 2014. Sampling-based robot motion planning: A review. IEEE Access, 2, pp.56-77.

35. Lopez, P. et al. Microscopic traffic simulation using sumo. International Conference on Intelligent Transportation Systems (ITSC), 2575-2582 (IEEE, 2018).
36. Treiber, M., Hennecke, A. & Helbing, D. Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E 62, 1805 (2000).
37. Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. In CVPR ADP3 workshop, 2021.
38. Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In Conference on Robot Learning (CoRL), 2023.
39. OpenAI (2023). Gpt-4 technical report. https://arxiv.org/abs/2303.08774.
40. Kandpal, N., Deng, H., Roberts, A., Wallace, E. and Raffel, C., 2023, July. Large language models struggle to learn long-tail knowledge. *In International Conference on Machine Learning* (pp. 15696-15707). PMLR.
41. Sauerbier, J., Bock, J., Weber, H., and Eckstein, L. Definition of scenarios for safety validation of automated driving functions," ATZ worldwide, vol. 121, no. 1, pp. 4245 (2019).
42. Wang, J., Zhang, L., Huang, Y., Zhao, J. and Bella, F., 2020. Safety of autonomous vehicles. Journal of advanced transportation, 2020, pp.1-13.
43. Zhu, Z., Wang, X., Zhao, W., Min, C., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., Zhang, C. and You, Y., 2024. Is sora a world simulator? a comprehensive survey on general world models and beyond. https://arxiv.org/abs/2405.03520.
44. Seshia S A, Sadigh D, Sastry S S. Toward verified artificial intelligence[J]. Communications of the ACM, 2022, 65(7): 46-55.
45. Owen, A. B. *Monte Carlo Theory, Methods and Examples*. https://artowen.su.domains/mc/ (2013).
46. Alain, G., Lamb, A., Sankar, C., Courville, A. and Bengio, Y., 2015. Variance reduction in sgd by distributed importance sampling. https://arxiv.org/abs/1511.06481.
47. Sutton, R. S., & Barto, A. G. Reinforcement Learning: An Introduction. MIT press (2018).
48. Ciosek, K. and Whiteson, S., 2017, February. Offer: Off-environment reinforcement learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).
49. Johnson, J.M. and Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. Journal of Big Data, 6(1), pp.1-54.
50. Scanlon, J.M., Kusano, K.D., Daniel, T., Alderson, C., Ogle, A. and Victor, T., 2021. Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. Accident Analysis & Prevention, 163, p.106454.
51. Johnson, J.M. and Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. Journal of Big Data, 6(1), pp.1-54.
52. Liang, Eric, et al. "RLlib: Abstractions for distributed reinforcement learning." International conference on machine learning. PMLR, 2018.
53. Darweesh, H. et al. Open source integrated planner for autonomous navigation in highly dynamic environments. J. Robot. Mechatron. 29, 668-684 (2017).
54. Feng, S., Feng, Y., Yan, X., Shen, S., Xu, S. and Liu, H.X., 2020. Safety assessment of highly automated driving systems in test tracks: A new framework. Accident Analysis & Prevention, 144, p.105664.
55. Chang AX. et al. ShapeNet: An information-rich 3d model repository. https://arxiv.org/abs/1512.03012 (2015).

56. Bezzina, D., Sayer, J. Safety pilot model deployment: Test conductor team report. (Report No. DOT HS 812 171). Washington, DC: National Highway Traffic Safety Administration (2014).
57. Sayer, J. et al. Integrated vehicle-based safety systems field operational test: final program report (No. FHWA-JPO-11-150; UMTRI-2010-36). United States. Joint Program Office for Intelligent Transportation Systems (2011).
58. Feng, S., Yan, X., Sun, H., Feng, Y., & Liu, H. X. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nat. Commun.* **12**, 1-14 (2021).
59. National Center for Statistics and Analysis. Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2018 (Report No. DOT HS 812 827). Washington, DC: National Highway Traffic Safety Administration. https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars. (2019).

Acknowledgements
This research was partially funded by the U.S. Department of Transportation (USDOT) Region 5 University Transportation Center: Center for Connected and Automated Transportation (CCAT) of the University of Michigan (69A3551747105) and the National Science Foundation (CMMI #2223517). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the official policy or position of the U.S. government.

Author contributions
S. F. and H. L. conceived and led the research program, developed the dense learning approach, and wrote the paper. S. F., H. Z., and H. S. developed the algorithms, designed the experiments, and performed the results. H. Z., H. S., L. H., and S. L. implemented the algorithms in simulation, performed the simulation tests, and prepared the simulation results. H.Z., X.Y., B.L., and S.S. implemented the algorithms in the autonomous vehicle, performed the field tests, and prepared the testing results. S. F., J. Y., G. S., and L. W. developed the theoretical analysis. All authors provided feedback during the manuscript revision and results discussions. H. L. approved the submission and accepted responsibility for the overall integrity of the paper.

Competing interests: The authors declare no competing interests.

Tables
Table 1. Evaluation results of ablation studies.

Episodic data densification	State-level data densification	Near-miss episodes	Retrospective data densification	Trajectory resampling	State reconnection	Crash rate in NDE after 780 training steps ($\times 10^{-5}$)	Crash rate in NDE after 1650 training steps ($\times 10^{-5}$)
w	w	w	w	w	w	1.07	0.65
w/o	w	w	w	w	w	2.26	2.00

w	w/o	w	w	w	w	3.60	3.31
w	w	w/o	w	w	w	1.71	1.22
w	w	w	w/o	w	w	1.93	0.90
w	w	w/o	w/o	w	w	3.56	3.27
w	w	w	w	w/o	w	2.19	0.77
w	w	w	w	w	w/o	1.34	0.94

Figure captions

Figure 1. Key challenges and our solution for AV safety training. a, Learning from all data suffers from severe variance due to the Curse of Rarity (CoR) challenge and cannot learn an effective policy. Here an effective policy means that the policy could improve the safety performance of autonomous vehicles (AVs) with a lower crash rate. Existing approaches attempted to tackle the CoR challenge primarily focus on learning from the data where AI systems exhibit failures (such as crash event data), which could mislead the training process, causing the seesaw effect. Our approach overcomes this issue by training AVs with densified data, improving AVs' overall safety performances. b, Our dense learning approach densifies the training data through three modules including episodic data densification (selection of informative driving episodes), driving state densification (retain informative states only), and retrospective data densification (re-selection of informative driving episodes through counterfactual simulation). Our approach can dramatically reduce learning variance for rare event learning without loss of unbiasedness and enable the learning of the SafeDriver model. c, During the AV driving process, the SafeDriver model will only override in safety-critical situations that are identified by a learned safety metric; in all other situations, the behaviors of the AV will be controlled by base AV models.

Figure 2. Performance evaluation of the dense learning approach in simulations. a, Performance evaluation of our dense learning approach for autonomous vehicle (AV) training in a multi-lane highway environment with a continuous driving distance of 400 meters. The red line in the map denotes the AV driving route. At each data point of our approach, we evaluated the AV's performance in naturalistic driving environment (NDE) and recollected the training data for the new AV. The shaded area represents the 90% confidence level. Our approach could dramatically reduce the overall crash rate, crash rates of different crash types, and avoidable crash rate, compared with the base AV model. Here we adopted the crash type diagram defined by the Fatality Analysis Reporting System[59]. We further investigated the evasive driving behaviors that SafeDriver learned and identified three typical ones, including emergency braking, rapid accelerating, and evasive lane changing (Supplementary Fig.2). Additional case studies are provided in Supplementary Video 4. b, Performance evaluation of our dense learning approach for AV training in roundabout scenarios. Additional case studies are detailed in Supplementary Video 5, providing comprehensive insights into our approach in roundabout scenarios. c, Performance evaluation of our dense learning approach for AV training in the simulation model of Mcity urban test track. The behaviors of background vehicles in NDE were constructed based on the large-scale naturalistic driving data from SPMD[56] and IVBSS[57] programs. Additional case studies are provided in Supplementary Video 6.

Figure 3. Performance evaluation of the dense learning approach on the nuPlan benchmark. a, Illustration of the four cities in the nuPlan benchmark where we evaluated the performance of our approach. b, Architecture of the PDM-Hybrid with SafeDriver. SafeDriver uses an attention-based state dropout encoder and generates 8-second trajectories for vehicle control. c-f, Two cases to demonstrate the effectiveness of our approach. The ego vehicle is depicted by a white rectangle.

When controlled by SafeDriver, the ego vehicle is highlighted with a red circle. Other vehicles are shown using green rectangles, while pedestrians are represented by blue rectangles. The expert trajectory is marked by an orange curve. c, When controlled by the base model, the ego vehicle proceeds straight ahead and collides with a vehicle turning right that does not yield. d, In the same scenario, SafeDriver executes a hard brake to avoid the potential collision. e, The base model directs the ego vehicle through a crosswalk, resulting in a collision with pedestrians. f, In the same scenario, when encountering pedestrians, SafeDriver executes a proactive yielding, providing enough space to avoid the crash.

Figure 4. Performance evaluation for a real-world AV at a physical test track. a, Illustration of the real autonomous vehicle (AV) under test, equipped with Autoware, Lidar, cameras, on-board computer, by-wire controller, high-definition (HD) map, and RTK (Real-Time Kinematics) GPS (Global Positioning System). b, Illustration of the Mcity test track including highways, roundabouts, intersections, urban streets, etc. c, Illustration of the mixed-reality environment combining the physical road infrastructures, proxy physical objects, and a simulation environment, where information of the real world and simulation world is synchronized. d, Safety performances of SafeDriver in the co-simulation of SUMO and Autoware at Mcity. e, Field testing results of the real AV with SafeDriver regarding the overall crash rate, crash rates of different crash types, and the avoidable crash rate. f, Cases of SafeDriver for avoiding crashes in safety-critical situations. In the first case, the SafeDriver (red vehicle) made emergency braking with right steering to avoid collisions in the situation that the background vehicle in the right lane made a reckless cut-in, while the vehicle from the opposite direction was approaching. In the second case, the SafeDriver (red vehicle) made emergency braking with left steering to avoid collisions in the situation that a background vehicle failed to yield when entering the roundabout. Additional explanations are available in Supplementary Videos 7-8.

Editor's Summary

Autonomous vehicles face the seesaw effect, where improving safety in some situations worsens it in others. This study introduces dense learning to focus on informative driving data and achieve more consistent and safer performance.

Peer review information: *Nature Communications* thanks Matthias Althoff, and Mozhgan Nasr Azadani for their contribution to the peer review of this work. A peer review file is available.