Accident Analysis and Prevention 144 (2020) 105664

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Safety assessment of highly automated driving systems in test tracks: A new framework

Shuo Feng[a], Yiheng Feng[b,*], Xintao Yan[a], Shengyin Shen[b], Shaobing Xu[c], Henry X. Liu[a,b]

[a] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, 48109, United States
[b] University of Michigan Transportation Research Institute (UMTRI), Ann Arbor, MI, 48109, United States
[c] Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, 48109, United States

ARTICLE INFO

Keywords:
Safety assessment
Highly automated driving systems
Test track
Testing scenario library
Augmented reality

ABSTRACT

Among the three major safety assessment methods (i.e., simulation, test track, and on-road test) for highly automated driving systems (ADS), test tracks provide high fidelity and a safe and controllable testing environment. However, due to the lack of realistic background traffic, scenarios that can be tested in test tracks are usually static and limited. To address this limitation, a new safety assessment framework is proposed in this paper, which integrates an augmented reality (AR) testing platform and a testing scenario library generation (TSLG) method. The AR testing platform generates simulated background traffic in test tracks, which interact with subject ADS under test, to create a realistic traffic environment. The TSLG method can systematically generate a set of critical scenarios under each operational design domain (ODD) and the critical scenarios generated from the TSLG method can be imported into the AR testing platform. The proposed framework has been implemented in the Mcity test track at the University of Michigan with a Level 4 ADS. Field test results show that the proposed framework can accurately and efficiently evaluate the safety performance of highly ADS in a cost-effective fashion. In the cut-in case study, the proposed framework is estimated to accelerate the assessment process by 9.87×10^4 times comparing to the on-road test approach.

1. Introduction

Safety assessment is a critical step in the development and deployment of highly automated driving systems (ADS). The assessment procedures for human-driven vehicles, such as Federal Motor Vehicle Safety Standards (FMVSS) (Federal Motor Vehicle Safety Standards, 1999) and ISO 26,262 (Road Vehicles – Functional Safety) (ISO 26262, 2011), are not enough to evaluate ADS comprehensively. A high-level (i.e., Level 3 or higher) ADS should be able to perform driving tasks including perceiving the environment, planning the route, and executing planned trajectories without human drivers. Besides the vehicle safety features defined in current standards, it is more important to evaluate how an ADS interacts with the roadway environment as well as other traffic participants in a safe and intelligent way. Due to the high complexity of the naturalistic driving environment, there exist millions of testing scenarios, which make this problem extremely challenging and time-consuming. As implied in the 2018 "Automated Vehicle 3.0" document (Preparing for the Future of Transportation, 2018) and by the National Highway Traffic Safety Administration (NHTSA) (Pilot Program for Collaborative Research on Motor Vehicles with High or Full Driving Automation, 2018), vehicle-level performance-based standards and testing procedures for highly ADS do not exist today and are needed before any commercial ADS can be deployed on public roads.

Currently, assessment of ADS is mainly performed in simulation, on test tracks, and public roads (Thorn et al., 2018). Simulation is useful for developing prototype models and testing of ADS at early stages. Simulation is cost-efficient, but it is well known that modeling the exact vehicle dynamics and road environment is challenging. Public roads provide the most realistic testing environment, but mistakes an ADS make on a public road can be expensive, dangerous, and even fatal (Liu and Feng, 2018). At least four fatal crashes have been reported in the past few years involving automatic driving functions (Favarò et al., 2017). Testing on public roads is also inefficient. ADS would have to drive hundreds of millions of miles to validate safety performance (Kalra and Paddock, 2016). Comparing with simulation and public roads, test tracks, in which subject ADS can be evaluated in a realistic environment, have their unique advantages. First, testing ADS on physical roadways and infrastructure (e.g., traffic signals and signs) mitigates the fidelity issues in simulation. Second, the testing

* Corresponding author.
E-mail address: yhfeng@umich.edu (Y. Feng).

https://doi.org/10.1016/j.aap.2020.105664
Received 10 February 2020; Received in revised form 18 May 2020; Accepted 27 June 2020
Available online 10 July 2020

environment is more controlled and therefore, safer than public roads. Third, corner testing scenarios that rarely happen on public roads (e.g., red-light running) can be "created" repeatedly in test tracks. Therefore, the test efficiency can be improved. In the past few years, dozens of closed-facilities are constructed around the world, among which Mcity at the University of Michigan is one of the representatives (Anon, 2020).

Although the controlled environment of a test track brings benefits such as improved safety and efficiency, it has two major limitations. To effectively assess the safety performance of an ADS, besides the roadway environment that a test track can provide, a dynamic traffic environment is also critical. However, a test track merely provides empty roadways, and introducing real vehicles as background traffic is not only costly but also difficult to coordinate and control (Feng et al., 2018). Without interactions with other road users, both types and number of scenarios that can be tested are limited. To address this limitation, hardware-in-the-loop (HIL) simulation approaches incorporate certain levels of physical hardware into the simulation with simulated background vehicles (BVs) (Thorn et al., 2018). One example is the vehicle-in-the-loop simulation, where a vehicle is placed on a roller test bench, such as a chassis-dynamometer, to allow physical actuations, while the vehicle remains static. However, the virtual road environment sometimes cannot reflect accurate vehicle performances. For example, vehicle's braking performances can be affected by road surface and weather conditions, which is critical in crash or near-crash scenarios.

Second, even with BVs introduced in the test track, how to generate a library of testing scenarios systematically, e.g., the maneuvers of BVs, remains a big challenge. On public roads, testing scenarios are not generated purposely but encountered naturally by the ADS under different routes, traffic patterns, and time of day. This is the exact reason why the efficiency of on-road test is very low because most of the encountered scenarios are not critical and less challenging. The major difficulty of generating scenarios in test tracks lies on how to guarantee assessment accuracy (e.g., the same level of performance as on-road test) with higher efficiency (e.g., fewer required testing miles than road test). Most existing scenario generation approaches only focus on the textual or graphic description of scenario categories (Feng et al., 2020a). However, even for one scenario category under one operational design domain (ODD), e.g., cut-in, there are millions of specific scenarios with different parameters, e.g., different cut-in ranges, range rates, and cut-in angles of the BV. How to determine a set of critical parameters given an ODD is the key to the problem. Yet, to the best of our knowledge, there is no generic way of generating and identifying such a set of critical parameters. Some studies (Hunger, 2017) enumerate and test all possible combinations of parameters. They may work under low dimensional cases where the number of parameters is limited. However, under higher dimensions, the number of combinations explodes exponentially.

This paper proposes a new framework for assessing the safety performance of highly ADS in test tracks, to address the abovementioned two problems. For the first problem, an augmented reality (AR) testing platform that combines a physical test track and a simulation environment is constructed. Movements of the real ADS in the physical test track are synchronized in the simulation platform, and information of virtual BVs generated in simulation is fed back to the real ADS. The real ADS in the physical test track can interact with the virtual BVs as if in a realistic traffic environment. Since the real vehicle is tested on real roadways, exact vehicle dynamics and road environment is ensured, which are critical for evaluating the safety performance of the ADS accurately. Meanwhile, comparing with introducing real BVs, simulated BVs can be easily controlled and manipulated in generating different scenarios with less cost and safety concerns.

For the second problem, the testing scenario library generation (TSLG) method, which was first proposed in our previous studies (Feng et al., 2020a, 2020b), is applied to systematically generate safety-critical scenarios in test tracks. A new definition of scenario criticality is adopted and a critical scenario searching method is developed based on importance sampling theory and optimization techniques. Different from (Feng et al., 2020a, 2020b), where human driving models were used as the surrogate model (SM) and only simulations were conducted, this paper leverages a Level 4 ADS model as the SM to generate critical scenarios, and tests a real ADS vehicle in the Mcity test track.

The major contributions of this paper are summarized as follows:

(1) Propose a new framework for the safety assessment of ADS in test tracks, which integrates the augmented reality testing platform and the testing scenario library generation method.
(2) Provide further validation of the theoretical findings in (Feng et al., 2020a) via test tracks. Field test results show that, by using the high-level ADS model as the SM to generate the scenario library, the ADS can be evaluated accurately with a further reduced number of required tests.
(3) Provide an exemplar demonstration of testing ADS from simulations to test tracks using mixed reality (see (Fremont et al., 2020) for another example), which is a promising direction for testing the ADS with the integration of different testing venues.

The rest of this paper is organized as follows. Section 2 introduces the overall framework. Section 3 describes the TSLG method. The implementation of the testing platform is presented in Section 4. The field tests and the results are analyzed in Section 5. Finally, Section 6 concludes the paper.

2. Overview of the ADS assessment framework

An overview of the proposed assessment framework is illustrated in Fig. 1. The framework includes four major steps, i.e., scenario description, metric design, library generation, and ADS evaluation. The scenario description step chooses the decision variables of scenarios given an ODD. Then the metric design step defines a quantitative index (e.g., accident rate) to represent safety. In the library generation step, a set of critical scenarios (i.e., the library) is obtained. The core idea is to first define scenario criticality and then search for scenarios with higher criticality values. The scenario criticality is measured by both maneuver challenge and exposure frequency. An SM is constructed to estimate the maneuver challenge of each scenario. In the ADS evaluation step, testing scenarios are sampled from the generated library, and then the sampled scenarios are imported to the AR platform and tested with a real ADS vehicle. Finally, the performance index value is calculated based on the testing results. The final output of the testing framework is an estimated index value for a given ODD, e.g., the accident rate in cut-in scenarios.

2.1. Scenario description

The goal of the scenario description is to determine the decision variables of the scenario considering ODD. In this paper, the terms scene, scenario, and ODD defined in (Preparing for the Future of Transportation, 2018; Ulbrich et al., 2015) are adopted. A scene describes a snapshot of the environment, including both static (i.e., all geo-spatially stationary elements) and dynamic elements (i.e., moving or have the ability to move). A scenario describes the temporal development of a sequence of scenes. An ODD is defined by where (such as what roadway types and speeds) and when (under what conditions, such as day/night, weather limits, *etc.*) an ADS is designed to operate. The ODD essentially provides static parameters and a feasible set of dynamic parameters. Let θ and x denote static parameters (e.g., road environment) and dynamic parameters (e.g., background vehicle maneuvers) respectively. Note in our problem formulation, θ is given and x is the decision variables. Then a specific scenario in a given ODD can be represented as (x, θ), $x \in \mathrm{X}$, where X is the set of all feasible

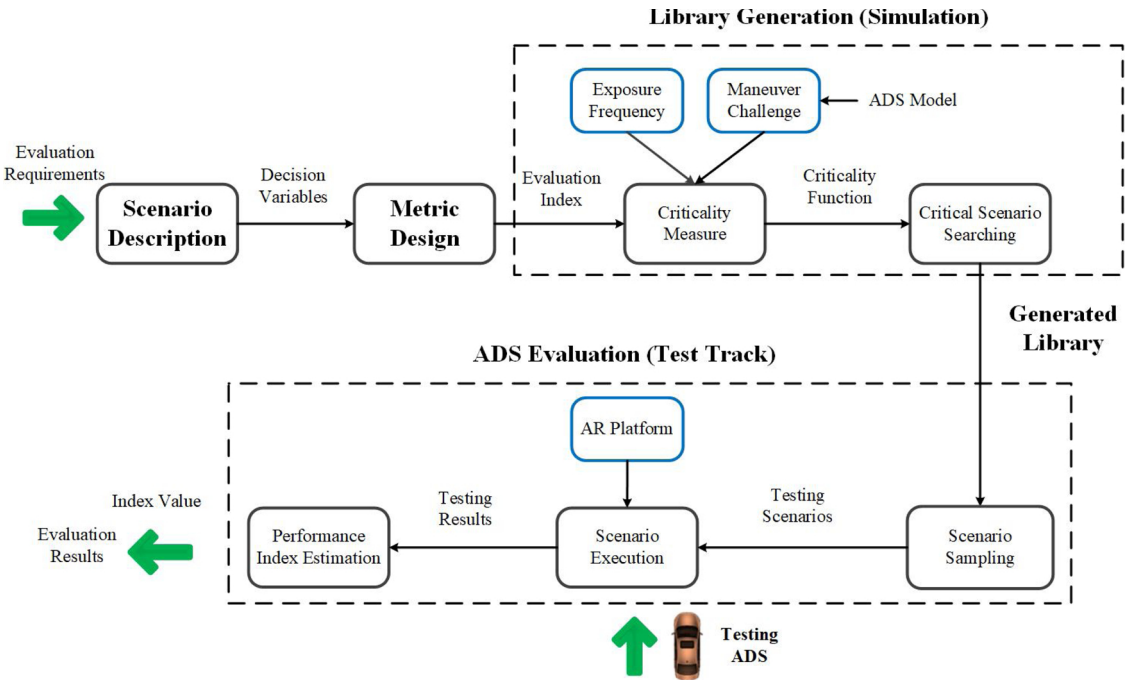

Fig. 1. The proposed framework for the safety assessment of highly ADS.

scenarios. Our goal is to find a subset of the feasible set, i.e., $\Phi \subset X$, which represents critical scenarios (i.e., the library). Taking the simplified cut-in scenario (Zhao et al., 2016) as an example, the decision variables can be considered as

$$x = (R, \dot{R}), \tag{1}$$

where R and $\dot{R}$ denote the relative distance (range) and the relative speed (range rate) between the cut-in vehicle and the subject vehicle (SV) at the cut-in moment. The static parameters θ include roadway types, weather, the initial speed of the cut-in BV, among others. The scenario description is generic for different scenario categories under different ODDs, including but not limited to, free driving, following, lane changing, overtaking, leave lane, cut-through, slow traffic, stop and go, lane violation, wrong-way driver, obstacle avoidance, and pedestrian crossing. A systematic way of determining scenario categories can be found in (Thorn et al., 2018).

2.2. Metric design

To quantitatively measure a metric, a performance index needs to be designed. For safety measures, the performance index is usually defined as the accident rate on public roads, as in many existing studies. Let A denote the accident event when ADS is driving on public roads, then the accident rate is denoted as $(A|\theta)$. The on-road test approach is essentially an estimation of $P(A|\theta)$ in the naturalistic driving environments. In the cut-in case, if an ADS experiences n cut-in scenarios on public roads, and has m accident events, then the accident rate is estimated as

$$P(A|\theta) \approx \frac{m}{n}. \tag{2}$$

Besides the accident rate, the method is also applicable for accident surrogates, such as conflict and injury. As long as the safety surrogate can be defined as an event A, the rate of the safety surrogate can be calculated by $P(A|\theta)$. The focus of the proposed framework is to estimate the performance index accurately and efficiently.

2.3. Library generation

Library generation is the most critical step in the framework. The goal is to find a subset of scenarios in the whole scenario space X (i.e., $\Phi \subset X$) which are critical, to construct the library. First, a new definition of criticality, i.e., $V(x)$, is adopted from (Feng et al., 2020a) as the combination of maneuver challenge and exposure frequency. For safety assessment, the maneuver challenge measures the dangerous level of a scenario, while the exposure frequency denotes the probability of the scenario occurring on public roads. An SM is constructed to estimate the maneuver challenge of each scenario. Different models can serve as SMs, such as human driving models (Feng et al., 2020a, 2020b) and high-level ADS path planning models. The new definition is fundamentally different from most existing studies, which usually overvalue infrequent scenarios (Ma and Peng, 1999; Zhao et al., 2016, 2018). Based on the criticality definition, different searching methods can be designed to find the set of critical scenarios with criticality values above a threshold. In our case study, an optimization-based searching method is applied. Finally, the library is constructed which consists of the critical scenarios and their associated criticality values, i.e., $V(x)$, $x \in \Phi$. More details on the library generation will be provided in the next section.

2.4. ADS evaluation

With the generated library, real ADS can be tested and evaluated using the AR platform as shown in Fig. 2. The AR platform generates background traffic in microscopic simulation to augment the functionality of the physical test track. Movements of test ADS in the real world are synchronized within simulation, and information of simulated background traffic is fed back to the test ADS through a real-time wireless communication network. The real test ADS can interact with simulated background traffic as if in a realistic traffic environment. As a result, testing scenarios that require interactions with other vehicles can be executed. Comparing with involving real BVs, simulated BVs can be easily controlled and manipulated in generating different scenarios with less cost and safety concerns.

To build the AR platform, a microscopic traffic simulator is needed to construct the simulation world. VISSIM (PTV, 2013) is chosen due to

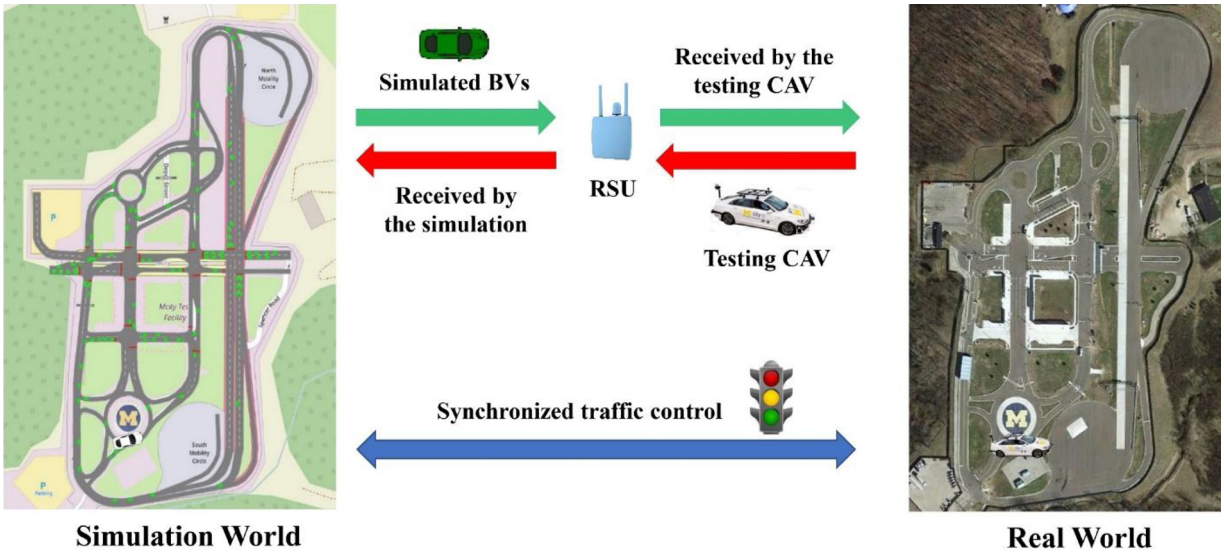

Fig. 2. The Augmented Reality Testing Platform.

its flexibility in traffic network construction and various user APIs (e.g., DriverModel.dll and COM interface) for customized vehicle generation and control. Wireless communications between the simulation world and the real world are established by the roadside units (RSU) installed in the test track via Dedicated Short-Range Communications (DSRC). Vehicle information is encoded to SAE J2735 complied Basic Safety Messages (BSMs) which contain vehicle statuses such as location, speed, acceleration, and heading. Traffic signal information is encoded to SAE J2735 complied Signal Phasing and Timing (SPaT) messages and also broadcast from RSUs via DSRC. Traffic signals in the two worlds are synchronized so that different vehicles can respond to the same signal indication. A data collection and management system is built to manage testing data and evaluate the performance of the ADS. More details of the AR platform can be found in (Feng et al., 2018).

Three steps are included to test the ADS in the AR platform, as scenario sampling, scenario execution, and performance index estimation. First, a series of specific scenarios are sampled from the scenario library. The ε -greedy sampling policy is applied to balance exploitation and exploration of the generated library as

$$q(x) = \begin{cases} (1-\varepsilon)\dfrac{V(x)}{W}, & x \in \Phi \\ \dfrac{\varepsilon}{N(X)-N(\Phi)}, & x \notin \Phi \end{cases}, \tag{3}$$

where $N(X)$ denotes the total number of feasible scenarios, and W is a normalization factor as

$$\mathrm{W} = \sum_{x\in\Phi} V(x). \tag{4}$$

After the testing scenarios are sampled, they are imported into the AR platform. A real ADS is tested with simulated vehicles whose trajectories are specified as the sampled scenarios, and the accident events are recorded. Finally, the performance index can be estimated as

$$\begin{aligned} P(A|\theta) &= \sum_{x\in X} P(A|\theta,x)P(x|\theta), \\ &= \sum_{x\in X} \frac{P(A|\theta,x)P(x|\theta)}{q(x)} q(x), \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \frac{P(A|\theta,x_i)P(x_i|\theta)}{q(x_i)}, \ x_i \sim q(x), \end{aligned} \tag{5}$$

where n denotes the total number of tests, $x_i \sim q(x)$ denotes that testing scenarios are sampled according to the distribution $q(x)$ (see Eq. (3)), and $P(A|\theta,x_i)$ denotes the performance of the ADS in the specific scenario (θ, x_i), which can be obtained by the testing results (e.g., crash or not). The last equivalence is derived by the Monte Carlo method with importance sampling. More details of the theoretical analysis of Eq. (5) can be found in (Feng et al., 2020a).

The framework provides an important insight that the ε -greedy sampling policy and criticality values of the testing scenarios essentially construct an importance function, i.e., $q(x)$. We proved in (Feng et al., 2020a) that the estimation of $P(A|\theta)$ is unbiased which indicates that the proposed method is accurate. To determine the total number of required tests (i.e., efficiency), it is critical to analyze the variance of the importance function. According to Monte Carlo theory (Owen, 2013), the estimation variance of Eq. (5) can be expressed as $Var = \sigma^2/n$, where

$$\sigma^2 = \sum_{x\in X} \frac{(P(A|x,\theta)P(x|\theta))^2}{q(x)} - P(A|\theta)^2. \tag{6}$$

The estimation accuracy can be measured by relative half-width given a confidence level (Ross, 2017). With the confidence level at $100(1-\alpha)\%$, the relative half-width is defined as

$$l_r = \frac{\Phi^{-1}(1-\alpha/2)}{P(A|\theta)}\sqrt{Var} = \frac{\Phi^{-1}(1-\alpha/2)}{P(A|\theta)}\frac{\sigma}{\sqrt{n}}, \tag{7}$$

where Φ^{-1} denotes the inverse cumulative distribution function of standard normal distribution $\mathcal{N}(0,1)$. Then, for a required half-width β, i.e., $l_r \le \beta$, the required number of tests is derived as

$$n \ge \left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{P(A|\theta)\beta}\right)^2 \sigma^2. \tag{8}$$

As the right term is determined by σ^2, Eq. (8) indicates that the required number of tests is fewer if the generated library has a smaller σ^2. We showed in (Feng et al., 2020a) that, by choosing a proper exploration rate ε in the sampling policy, the estimation variance can be minimized. In other words, the estimation efficiency is improved.

Note that the proposed framework is generic for assessing any types of ADS performances in test tracks. We will introduce a specific TSLG method applying optimization methods in the following section.

3. Testing scenario library generation

Most existing studies focus on the textual or graphic description of scenario categories (Thorn et al., 2018; Li et al., 2016, 2019). However, even for one scenario category, e.g., cut-in on a two-lane highway of

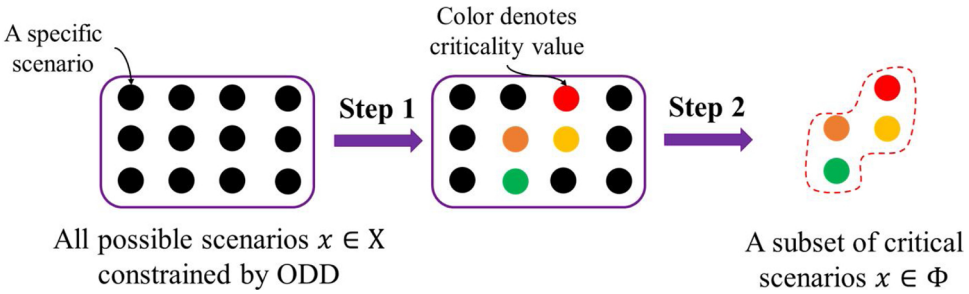

Fig. 3. Conceptual illustration of the two steps for library generation given an ODD.

45 mph speed limit during day time with the good weather condition, there could be millions of specific scenarios with different cut-in parameters. As discussed in the previous section, a specific scenario is denoted as (x, θ), $x \in X$, and the key to the problem is to find a set of critical scenarios, i.e., $\Phi \subset X$. Fig. 3 illustrates the two steps to find the library. In step 1, the criticality values of all scenarios are calculated, while in step 2, a searching method is designed to identify the set of critical scenarios. The two steps are generic to different library generation methods. In this paper, we introduce a new optimization-based searching method as follows.

The criticality is defined as the combination of maneuver challenge and exposure frequency as in (Feng et al., 2020a):

$$V(x) \stackrel{\text{def}}{=} P(S|x,\theta)P(x|\theta), \tag{9}$$

where S denotes the accident event with a surrogate model (SM) of ADS. The SM is designed to represent generic features of different ADS, e.g., keep a safe distance and interact with surrounding vehicles. It is similar to human drivers, where different drivers have different driving habits, but generic features exist among all drivers. An ideal SM should be calibrated from actual ADS driving data. Due to very limited publicly available real-world ADS driving data, calibrated human driving models were used as the SM in (Feng et al., 2020a, 2020b), which is a natural baseline of ADS, as discussed in (Bojarski et al., 2016; Zhang and Cho, 2016; Li et al., 2018). As pointed out in (Feng et al., 2020c), however, the dissimilarity between the SM and ADS usually exists and reduces the evaluation efficiency. If a high-level ADS planning model, which can better capture the features of the real ADS vehicle under test (i.e., smaller dissimilarity), is available, then it should be used as the SM. In our case study, such a model is obtained from the ADS vehicle developers and applied as the SM.

The term $P(S|x,\theta)$ measures the probability the SM encounters the accident event in a specific scenario (x, θ), named as maneuver challenge, which can be obtained by simulation of the SM. The term $P(x|\theta)$ measures the probability of the scenario occurring on public roads, named as exposure frequency, which can be obtained by analyzing naturalistic driving data (NDD). The definition is consistent with our intuition that both dangerous and frequent scenarios are more critical for safety assessment. If a scenario never happens in the real world, it is not necessary to be evaluated. On the other hand, if a scenario cannot challenge the ADS, it is also not critical.

Based on the criticality definition, an optimization-based method is designed to search for critical scenarios. The basic idea is to conduct a local search to efficiently find a portion of critical scenarios (i.e., local optimal solutions), and then expand from these scenarios to construct the library by the seed-fill method (Nosal, 2008), as shown in Fig. 4. To provide searching directions, an auxiliary objective function is constructed by a combination of the surrogate safety measure and surrogate exposure measure. In the literature, different surrogate safety measures are applied, such as time-to-collision (TTC), the criticality metric from PEGASUS (Junietz et al., 2018), and Responsibility-Sensitive Safety (RSS) (Shalev-Shwartz et al., 2017). In our study, the minimal normalized positive enhanced time-to-collision (nmpETTC) is applied as the surrogate measure for maneuver challenge. To estimate the exposure frequency, surrogate exposure measure is designed as the distance between the scenario and a high exposure zone in NDD. Note that the optimization-based method is not unique. The brute force method can also work for very simple cases. For high-dimensional scenarios, a reinforcement learning-based searching method can be more effective (Feng et al., 2020b).

It is worth noting that the scenarios outside the library still have a small probability of being sampled and tested with the ε -greedy sampling policy (see Eq. (3)), because the generated library may not cover all critical scenarios. However, scenarios in the library have much higher probabilities of being sampled, according to their criticality values. This is a typical strategy to balance exploration and exploitation.

The theoretical foundation of the proposed TSLG method is the important sampling theory (Owen, 2013), which was first introduced into the field of ADS assessment in (Zhao et al., 2016). The proposed framework applies this theory and finds that generating a library is equivalent to constructing an importance function. Therefore, finding an optimal library is equivalent to constructing an optimal importance function. Unfortunately, the optimal importance function is impossible to obtain unless the exact ADS model is known. That is the reason why the SM is applied as an alternative.

4. Implementation at Mcity

The proposed framework is implemented at the Mcity test track at the University of Michigan (Anon, 2020), and a Level 4 Lincoln MKZ hybrid automated vehicle (Xu et al., 2018; Xu and Peng, 2019) is used as the ADS under test. The Mcity test track is the world's first purpose-built full-fidelity proving ground for assessing the performance of highly ADS (Anon, 2020). It occupies 32 acres and includes about five lane miles of roadways including a highway segment, arterial roads, intersections, and traffic signals. The MKZ is equipped with various sensors, by-wire control, and a communication system, and a self-driving system has been deployed on the vehicle. Fig. 5 illustrates the implementation framework, including the TSLG platform, the AR testing platform, the Mcity test facility, and the ADS under test. To build the AR testing platform for Mcity, the Mcity's road network and infrastructure components (e.g., traffic signals) are built and calibrated in VISSIM, using a high-fidelity survey map. The TSLG platform is responsible for generating scenario libraries under different ODDs based on NDD and SM. After testing scenarios are sampled from the library, the AR testing platform generates virtual BVs to replicate the scenarios in VISSIM and feeds the simulated vehicle information to the real test ADS in real-time through the RSUs installed in Mcity. Note that the main purpose of this study is to evaluate the safety performance of the ADS accurately so that a real ADS vehicle is used to avoid difficulties in modeling exact vehicle dynamics and road environment in simulation. The BVs are mainly used to create the testing scenarios so that the ADS vehicle can respond accordingly. As a result, using simulated vehicles as BVs will not affect the testing results. There are two additional advantages of using this mixed reality setting: 1) simulated BVs are easy to manipulate to exactly replicate the testing scenario, while real BVs are

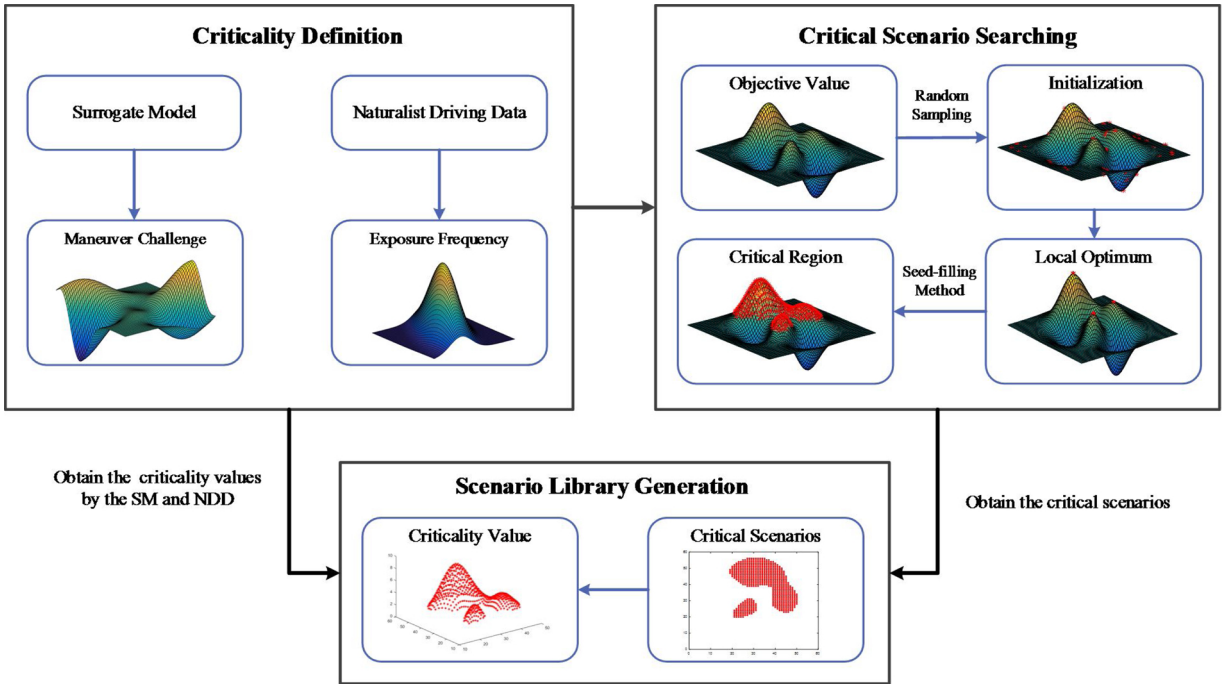

Fig. 4. Illustration of the entire library generation process.

much more difficult to control and coordinate. 2) Crashes are inevitable during testing, especially under critical scenarios, which may damage the ADS vehicle if real BVs are used. With simulated BVs, even a crash happens, no real damage will occur.

The data management module is used to record and process the testing data. The testing process stops when required relative half-width (Eq. 8) is reached.

5. Case study

To validate the proposed framework, the cut-in scenario is evaluated. NDD from the Safety Pilot Model Deployment (SPMD) project (Sayer et al., 2011) is utilized to measure the exposure frequency. The ADS under test is the Lincoln MKZ, introduced in the previous section. The Lincoln vehicle uses the Gipps car-following model (Gipps, 1981) as its high-level planning model. As a result, the Gipps model is applied as the SM, as discussed in Section 3. The objective is to evaluate the

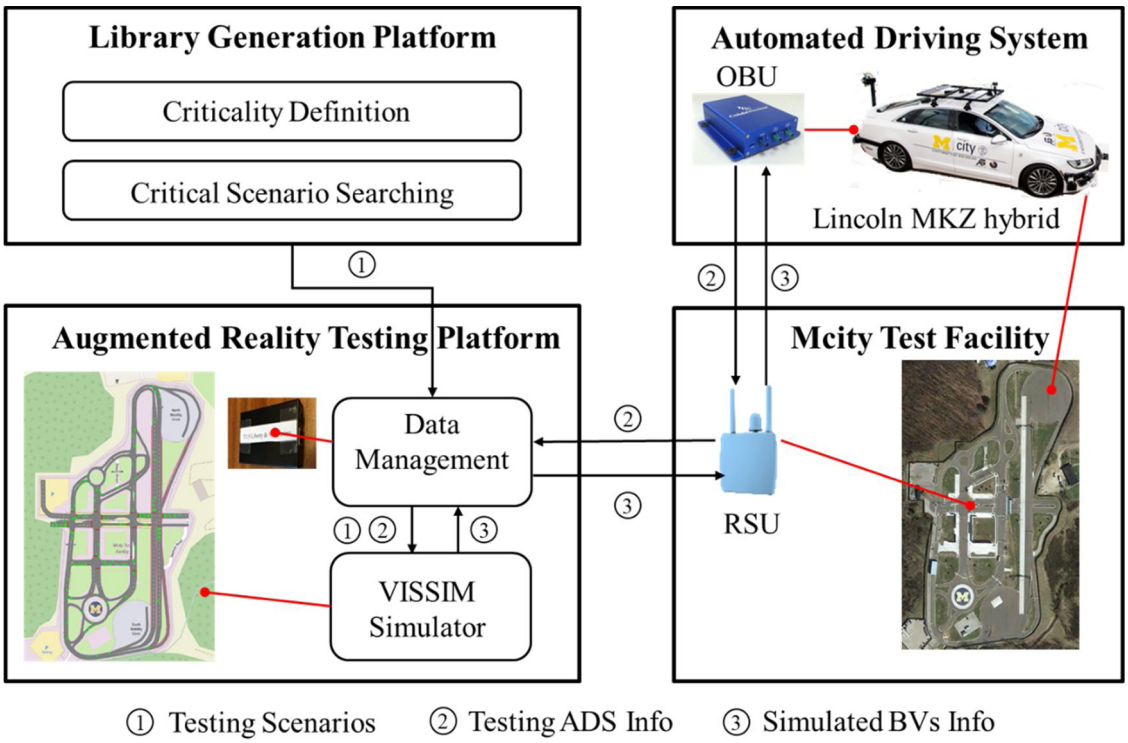

Fig. 5. Implementation of the proposed framework at Mcity test track.

Fig. 6. Bird's-eye view of the cut-in test scenario at Mcity.

safety performance in terms of accident rate of the Lincoln vehicle in cut-in scenarios. An accident is identified if the relative distance between the BV and the ADS is zero or less than zero (i.e., overlapped). In reality, the cut-in vehicle may choose different lateral speeds to perform cut-in and adjust its speed after cut-in, which introduces more parameters (dimensions). To better illustrate and visualize the case study, we simplify the case and consider only two dimensions, i.e., range and range rate at the cut-in moment, as illustrated in Eq. (1). The cut-in vehicle will maintain a constant speed after the cut-in maneuver. This is also a common practice in many other studies (e.g., Zhao et al., 2016). As illustrated in Fig. 6, the Lincoln MKZ starts from the scenario starting point at the south of the highway segment at Mcity, adjusts its speed to the expected value, and reaches the "cut-in" region where the simulated BV performs a cut-in with the pre-determined range and range rate. Different values of the range and range rate are sampled from the generated library. By analyzing the trajectories of the two vehicles, the accident event can be identified.

In the following, we will go through all the major steps of the assessment framework using the cut-in scenario as an example.

5.1. Naturalistic driving data analysis

The SPMD database is one of the largest NDD databases that record naturalistic driving behaviors over 34.9 million miles from 2842 equipped vehicles in Ann Arbor, Michigan. In the database, there are 98 sedans equipped with the data acquisition system and Mobileye cameras, which enable measuring and recording the position and speed data between the host vehicle and preceding vehicles at a frequency of 10 Hz. The following query criteria are designed to extract all cut-in events: (a) the vehicles' speeds at the cut-in moment belong to (2 m/s, 40 m/s); (b) the range at the cut-in moment belongs to (0.1 m, 90 m). A total number of 414,770 qualified cut-in events are successfully obtained. Without loss of generalization, the range and range rate are discretized by 2 m and 0.4 m/s respectively. Fig. 7(a) shows the exposure frequency of the cut-in scenarios.

5.2. Surrogate model construction

The key idea of constructing the SM is to include prior knowledge of the ADS under test to measure the maneuver challenge. For the ADS vehicle in the experiment (i.e., Lincoln MKZ), the Gipps model (Gipps, 1981) is adopted as the high-level car-following planning model. Although the exact behaviors of the Lincoln vehicle are much more complicated and intractable for prediction, the Gipps car-following model is a natural choice of the SM. The values of parameters and constraints are also obtained from the vehicle developers (Xu et al., 2018; Xu and Peng, 2019). Specifically, the SM is described as

$$v_m(t+\tau) = \min\left\{ v_m(t) + 2.5a_m\tau\left(1 - \frac{v_m(t)}{V_m}\right)\left(0.025 + \frac{v_m(t)}{V_m}\right)^{\frac{1}{2}}, b_m\tau + \sqrt{b_m^2\tau^2 - b_m[2[x_{m-1}(t) - s_{m-1} - x_m(t)] - v_m(t)\tau - v_{m-1}^2(t)/b]} \right\} \tag{10}$$

where x, v denote the position and speed, t denotes the current time, τ denotes the time interval, m denotes the following vehicle, $m-1$ denotes the leading vehicle, and a_m, V_m, b_m, s_{m-1}, b are static parameters. The values of the static parameters can be found in Table 1. The constraints of acceleration and speed of the model are included as

$$a_{min} \le u_n \le a_{max}$$

$$v_{min} \le v_n \le v_{max}$$

Fig. 7(b) shows the maneuver challenge from the SM simulation in all cut-in scenarios. The yellow region denotes the SM has accident events in these scenarios, i.e., $P(S|x,\theta) = 1$. The blue region denotes that the SM has no accident event in these scenarios, i.e., $P(S|x,\theta) = 0$. Note that the SM, in this case, is deterministic. It is a reasonable setting because, for most ADS, safety-related behaviors are desired to be stable.

5.3. Library generation

As shown in Eq. (9), the criticality of scenarios can be measured based on the exposure frequency and maneuver challenge. Fig. 7(c) shows the normalized criticality values of all cut-in scenarios. The lighter colors denote the higher criticality values. The set of scenarios with criticality values greater than a threshold (i.e., 0) are included in the library. To obtain the subset, the optimization-based searching method described in Fig. 4 is applied. In this case, 57 critical scenarios are obtained, which is about 1.67 % of all feasible scenarios. With the generated library, ε -greedy sampling policy in Eq. (3) is applied to sample testing scenarios. The value of ε is selected as 0.1.

5.4. ADS evaluation

Based on the sampled testing scenarios, the Lincoln vehicle is tested and evaluated in the AR platform at Mcity. In one test, a BV is generated by the simulation and performs the cut-in maneuver in front of the Lincoln vehicle with sampled range and range rate. Detail trajectories of the Lincoln MKZ and the cut-in vehicle are recorded with the onboard data acquisition system at the frequency of 20 Hz. A screenshot of the field test video is shown in Fig. 8, including the augmented view of the Lincoln vehicle (white and black: Lincoln; red: cut-in vehicle), view of the vehicle's front camera, scenario parameters (range, and speeds), and the simulation view (yellow: Lincoln; red: cut-in vehicle). After all the tests are done, the accident rate of the Lincoln MKZ is calculated by Eq. (5), and the relative half-width is calculated by Eq. (7). To provide

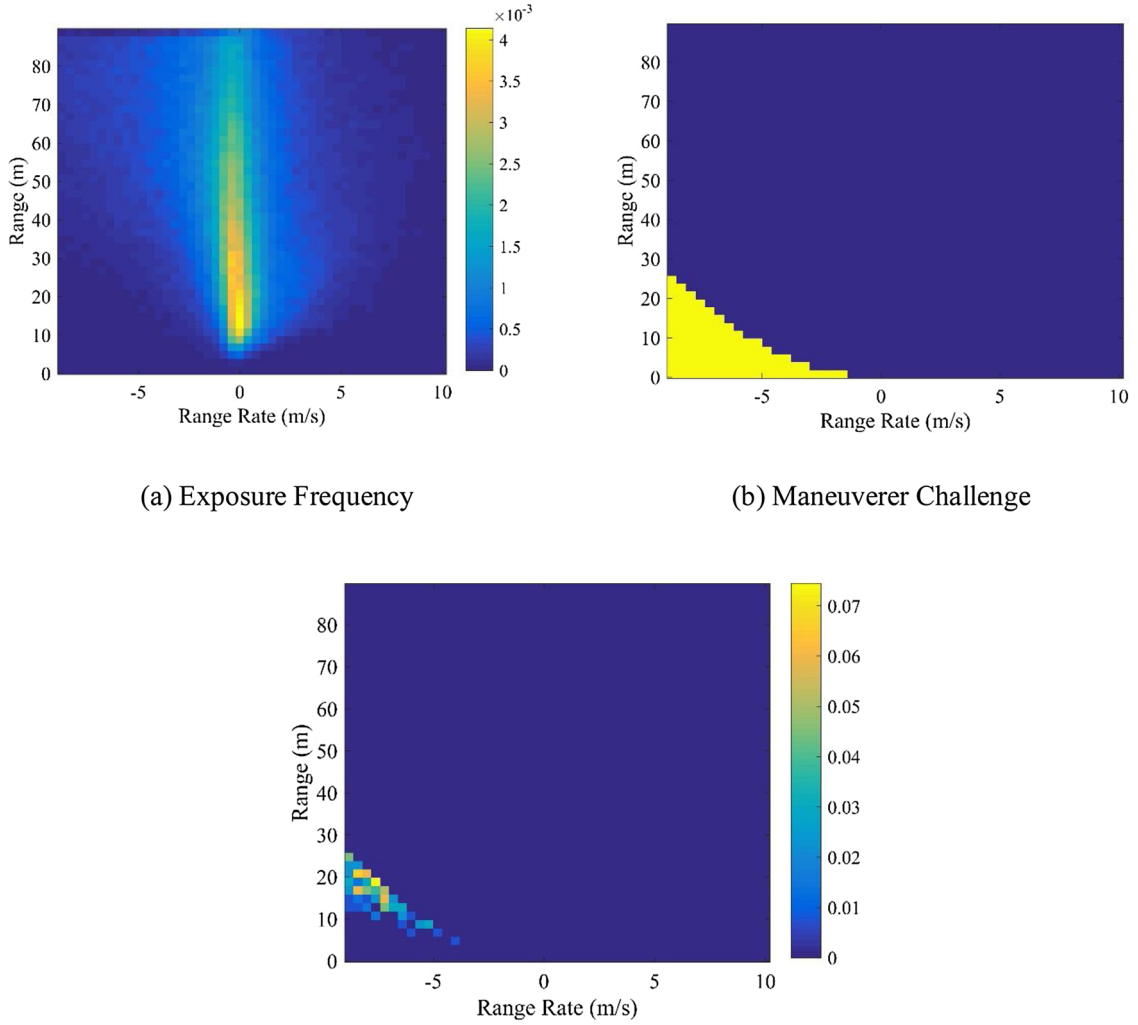

Fig. 7. Results of the TSLG method.

Table 1
Values of the parameters in the case study.

Parameter	Value
a_m	$2\,m/s^2$
V_m	$12\,m/s$
b_m	$-2.5\,m/s^2$
s_{m-1}	$10\,m$
b	$-2.5\,m/s^2$
a_{min}	$-2\,m/s^2$
a_{max}	$2\,m/s^2$
v_{min}	$0\,m/s$
v_{max}	$40\,m/s$
τ	$0.25\,s$

ground truth of the accident rate and a baseline for efficiency comparison, the on-road test approach is simulated, where testing scenarios are sampled from the exposure frequency in Fig. 7(a) obtained from NDD. The method is denoted as the NDD evaluation.

Fig. 9 shows the evaluation results comparing the proposed method and NDD evaluation. The bottom x -axis denotes the number of tests from NDD evaluation, while the top x-axis denotes the number of tests from the proposed method. Fig. 9(a) shows the proposed method can obtain the same accident rate as the NDD evaluation, i.e., accuracy. Note that the maximum deceleration of the ADS under test is set to a moderate value (about $-2\,m/s^2$) and the emergency brake function is disabled for experimental convenience, e.g., safety driver's comfort. Therefore, the accident rate of the ADS is about 2×10^{-5}, which is higher than normal human drivers. The main purpose of the field test is to validate the proposed safety assessment framework, not estimating a true cut-in accident rate. Fig. 9(b) shows the proposed method can obtain the same precision level by a much smaller number of tests. In this case, for the relative half-width $\beta = 0.2$, the required number of tests for the proposed method and NDD evaluation are 31 and 3×10^6 respectively. The results of NDD evaluation are consistent with (Kalra and Paddock, 2016), where the ADS would have to drive hundreds of millions of miles to validate safety. The proposed method significantly reduces the number of tests and accelerates the safety assessment process by 9.87×10^4 times. If one cut-in scenario occurs every certain number of miles on the public road, we can claim that one testing mile of the critical scenarios is equivalent to 9.87×10^4 miles of the public road test. The improved efficiency can significantly reduce both the time and cost of the ADS validation and verification (V&V) process.

6. Conclusions

In this paper, a new framework is proposed integrating the AR

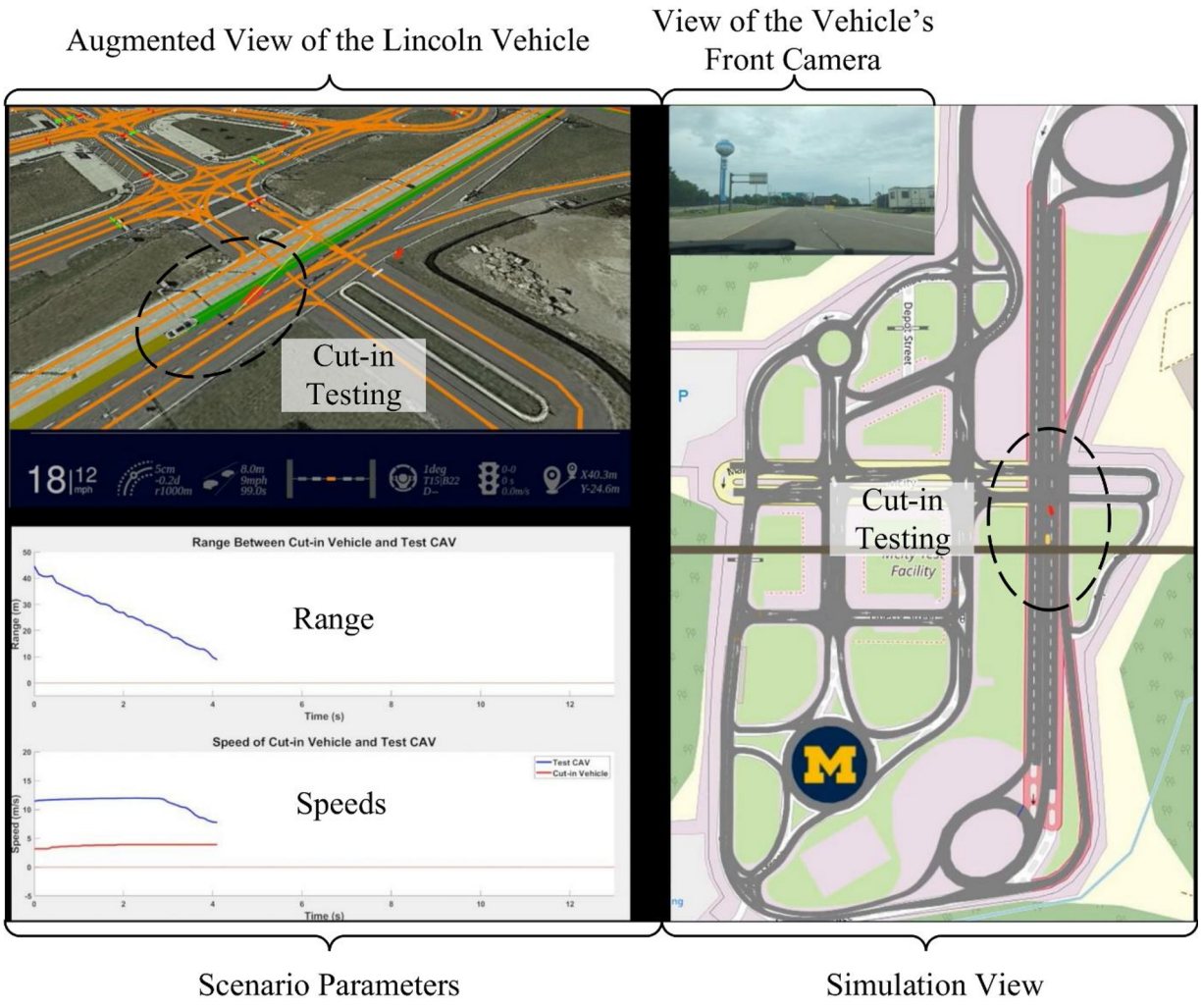

Fig. 8. A Screenshot of the field test video. (https://traffic.engin.umich.edu/research/automated-vehicle-system-testing-and-evaluation).

testing platform and the TSLG method for the safety assessment of highly ADS in test tracks. The AR testing platform, which combines the virtual and real worlds, provides a cost-effective method to generate background traffic in closed test tracks. The TSLG method, which identifies critical scenarios and constructs a scenario library, accelerates the evaluation process without losing accuracy. The framework is implemented and tested at the Mcity test track with a Level 4 ADS vehicle. Field test results show that the proposed framework can assess the safety of highly ADS accurately and efficiently. The safety assessment process of the cut-in scenario is accelerated by 9.87×10^4 times comparing with the NDD evaluation.

In the future study, a sensor simulation component will be added to

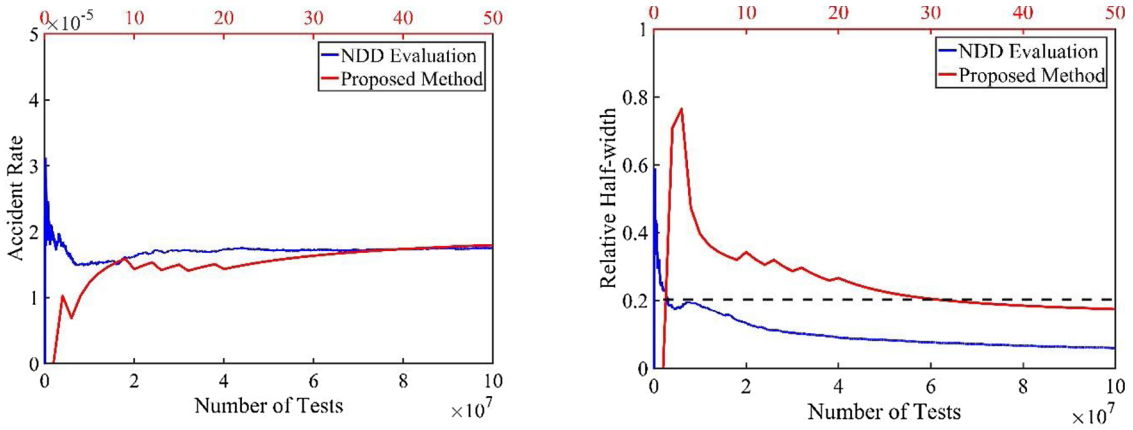

Fig. 9. Results of the accident rate estimation (a) and relative half-width (b) with the increasing number of tests.

the AR testing platform to test the perception module of the ADS. Moreover, a more complete list of scenario categories is required, so the library under different ODDs can be generated for comprehensive safety assessment. To validate the framework for high-dimensional scenarios, the highway driving scenarios will be studied, where both lateral and longitudinal maneuvers of the ADS are modeled, and multiple background vehicles may interact with the ADS. Furthermore, the generated safety-critical scenarios can be utilized to train the CAV models purposely. By training better prediction models or planning strategies (e.g., evasive maneuvers), the safety performance of the CAV can be further improved.

CRediT authorship contribution statement

Shuo Feng: Methodology, Formal analysis, Validation, Investigation, Data curation, Writing - original draft. **Yiheng Feng:** Conceptualization, Software, Investigation, Data curation, Funding acquisition, Writing - review & editing. **Xintao Yan:** Data curation, Software, Validation, Visualization. **Shengyin Shen:** Data curation, Software, Validation, Visualization. **Shaobing Xu:** Data curation, Software. **Henry X. Liu:** Conceptualization, Funding acquisition, Investigation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially funded by the US Department of Transportation (USDOT) Region 5 University Transportation Center: Center for Connected and Automated Transportation (CCAT), and Mcity of the University of Michigan. The views and opinions expressed in this paper are those of the authors.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105664.

References

Anon https://mcity.umich.edu. Accessed July 24, 2019.
Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al., 2016. End to end learning for self-driving cars. arXiv preprint arXiv 1604.07316.
Favarò, F.M., Nader, N., Eurich, S.O., Tripp, M., Varadaraju, N., 2017. Examining accident reports involving autonomous vehicles in California. PLoS One 12 (9), e0184952.
Federal Motor Vehicle Safety Standards, 1999. National Highway Traffic Safety Administration. Department of Transportation, United States.
Feng, S., Feng, Y., Sun, H., Zhang, Y., Liu, H.X., 2003c. 2020. Testing Scenario Library Generation for Connected and Automated Vehicles: An Adaptive Framework. arXiv preprint arXiv 03712.
Feng, Y., Yu, C., Xu, S., Liu, H.X., Peng, H., 2018. An augmented reality environment for connected and automated vehicle testing and evaluation. In: Presented at 29th IEEE Intelligent Vehicle Symposium. Changshu, China.
Feng, S., Feng, Y., Yu, C., Zhang, Y., Liu, H.X., 2020a. Testing scenario library generation for connected and automated vehicles, part I: methodology. IEEE Trans. Intell. Transp. Syst. https://doi.org/10.1109/TITS.2020.2972211.
Feng, S., Feng, Y., Sun, H., Bao, S., Zhang, Y., Liu, H.X., 2020b. Testing scenario library generation for connected and automated vehicles, part II: case studies. IEEE Trans. Intell. Transp. Syst. https://doi.org/10.1109/TITS.2020.2988309.
Fremont, D.J., Kim, E., Pant, Y.V., Seshia, S.A., Acharya, A., Bruso, X., Wells, P., Lemke, S., Yu, Q., Mehta, S., 2020. Formal scenario-based testing of autonomous vehicles: from simulation to the real world. arXiv preprint arXiv 2003 07739.
Gipps, P.G., 1981. A behavioural car-following model for computer simulation. Transp. Res. Part B Methodol. 15 (2), 105–111.
Hunger, H., 2017. Test Specifications for Highly Automated Driving Functions: Highway Pilot. [Online]. Available:. Tech. Rep.. https://www.pegasusprojekt.de.
ISO 26262, 2011. Road Vehicles - Functional Safety. Final Draft (FDIS), Geneva.
Junietz, P., Bonakdar, F., Klamann, B., Winner, H., 2018. November. Criticality metric for the safety validation of automated driving using model predictive trajectory optimization. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE. pp. 60–65.
Kalra, N., Paddock, S.M., 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transp. Res. Part A Policy Pract. 94, 182–193.
Li, L., Huang, W.L., Liu, Y., Zheng, N.N., Wang, F.Y., 2016. Intelligence testing for autonomous vehicles: a new approach. IEEE Trans. Intell. Veh. 1 (2), 158–166.
Li, L., Lin, Y.L., Zheng, N.N., Wang, F.Y., Liu, Y., Cao, D., Wang, K., Huang, W.L., 2018. Artificial intelligence test: a case study of intelligent vehicles. Artif. Intell. Rev. 50 (3), 441–465.
Li, L., Wang, X., Wang, K., Lin, Y., Xin, J., Chen, L., Xu, L., Tian, B., Ai, Y., Wang, J., Cao, D., 2019. Parallel Testing of Vehicle Intelligence Via Virtual-real Interaction.
Liu, X.H., Feng, Y.Real, 2018. World Meets Virtual World: Augmented Reality Makes Driverless Vehicle Testing Faster, Safer, and Cheaper. Accessed July 24, 2019. https://mcity.umich.edu/wp-content/uploads/2018/11/mcity-whitepaper-augmented-reality.pdf.
Ma, W.H., Peng, H., 1999. A worst-case evaluation method for dynamic systems. J. Dyn. Syst. Meas. Control 121 (2), 191–199.
Nosal, E.M., 2008. October. Flood-fill algorithms used for passive acoustic detection and tracking. In: 2008 New Trends for Environmental Monitoring Using Passive Systems. IEEE. pp. 1–5.
Owen, A.B., 2013. Monte Carlo Theory, Methods and Examples.
Pilot Program for Collaborative Research on Motor Vehicles with High or Full Driving Automation, 2018. NHTSA-2018-0092. National Highway Traffic Safety Administration. United States, Department of Transportation.
Preparing for the Future of Transportation, 2018. - Automated Vehicle 3.0. U. S. Department of Transportation.
PTV, 2013. VISSIM 6.0 User Manual. Karlsruhe, Germany.
Ross, S.M., 2017. Introductory Statistics. Academic Press.
Sayer, J.R., Bogard, S.E., Buonarosa, M.L., LeBlanc, D.J., Funkhouser, D.S., Bao, S., Blankespoor, A.D., Winkler, C.B., 2011. Integrated Vehicle-based Safety Systems Light-vehicle Field Operational Test Key Findings Report.
Shalev-Shwartz, S., Shammah, S., Shashua, A., 2017. On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv 1708.06374.
Thorn, E., Kimmel, S.C., Chaka, M., Hamilton, B.A., 2018. A Framework for Automated Driving System Testable Cases and Scenarios (No. DOT HS 812 623). Department of Transportation. National Highway Traffic Safety Administration, United States.
Ulbrich, S., Menzel, T., Reschka, A., Schuldt, F., Maurer, M., 2015. September. Defining and substantiating the terms scene, situation, and scenario for automated driving. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE. pp. 982–988.
Xu, S., Peng, H., 2019. Design, analysis, and experiments of preview path tracking control for autonomous vehicles. IEEE Trans. Intell. Transp. Syst.
Xu, S., Peng, H., Song, Z., Chen, K., Tang, Y., 2018. June. Accurate and smooth speed control for an Autonomous vehicle. In: 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE. pp. 1976–1982.
Zhang, J., Cho, K., 2016. Query-efficient imitation learning for end-to-end autonomous driving. arXiv preprint arXiv 1605.06450.
Zhao, D., Lam, H., Peng, H., Bao, S., LeBlanc, D.J., Nobukawa, K., Pan, C.S., 2016. Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques. IEEE Trans. Intell. Transp. Syst. 18 (3), 595–607.
Zhao, D., Huang, X., Peng, H., Lam, H., LeBlanc, D.J., 2018. Accelerated evaluation of automated vehicles in car-following maneuvers. IEEE Trans. Intell. Transp. Syst. 19 (3), 733–744.