CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

UMTRI-2023-6
March 2023

Connected and Automated Vehicle (CAV) Testing Scenario Design and Implementation Using Naturalistic Driving Data and Augmented Reality

Dr. Yiheng Feng
Dr. Shan Bao
Dr. Henry Liu

UMTRI

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Report No. UMTRI-2023-6

March 2023

Project Start Date: October, 2017

Project End Date: March, 2019

Connected and Automated Vehicle (CAV) Testing Scenario Design and Implementation Using Naturalistic Driving Data and Augmented Reality

by

Yiheng Feng, Assistant Research Scientist

Shan Bao, Associate Professor

Henry Liu, Professor

University of Michigan

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3551747105 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Feng, Y., Bao, S., & Liu, H.X. (2019). Connected and Automated Vehicle (CAV) Testing Scenario Design and Implementation Using Naturalistic Driving Data and Augmented Reality. Final Report. UMTRI-2023-6.
DOI:10.7302/7023

Contacts

For more information:

Dr. Yiheng Feng
University of Michigan
2901 Baxter Rd, Ann Arbor, MI, 48109
Phone: (734) 936-1052
Email: yhfeng@umich.edu

Dr. Shan Bao
University of Michigan
2901 Baxter Rd, Ann Arbor, MI, 48109
Phone: (734) 936-1127
Email: shanbao@umich.edu

Dr. Henry X. Liu
University of Michigan
2350 Hayward, Ann Arbor, MI, 48109
Phone: (734) 647-4796
Email: henryliu@umich.edu

Center for Connected and Automated Transportation
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48152
umtri-ccat@umich.edu
ccat.umtri.umich.edu
(734) 763-2498

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Technical Report Documentation Page

1. Report No. UMTRI-2023-6	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Connected and Automated Vehicle (CAV) Testing Scenario Design and Implementation using Naturalistic Driving Data and Augmented Reality DOI:10.7302/7023		**5. Report Date** March 2023 **6. Performing Organization Code**
7. Author(s) Feng, Yiheng, Ph.D., https://orcid.org/0000-0001-5656-3222 Bao, Shan, Ph.D., https://orcid.org/0000-0002-0768-5538 Liu, Henry, Ph.D., https://orcid.org/0000-0002-3685-9920		**8. Performing Organization Report No.**
9. Performing Organization Name and Address UMTRI 2901 Baxter Road Ann Arbor, MI 48109		**10. Work Unit No.** **11. Contract or Grant No.** Contract No. 69A3551747105
12. Sponsoring Agency Name and Address Center for Connected and Automated Transportation University of Michigan Transportation Research Institute 2901 Baxter Road Ann Arbor, MI 48109		**13. Type of Report and Period Covered** Final Report October 2017 – March 2019 **14. Sponsoring Agency Code**

15. Supplementary Notes
Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.

16. Abstract
Testing and evaluation is a critical step in the development and deployment of connected and automated vehicle (CAV) technology. Testing standards for human-driven vehicles, such as Federal Motor Vehicle Safety Standards (FMVSS), were established a long time ago. However, current standards cannot be applied to CAVs, because they often assume the presence of a human driver, who conducts the driving tasks. It is very important to develop test procedures and identify applicable test scenarios (user cases) for CAVS to evaluate the "intelligence" of the vehicle. The intelligence level indicates whether a CAV can drive safely and efficiently without human intervention. The newly released Automated Driving Systems Guideline 2 has made it very clear that the new automated driving systems need validation methods and to be tested by incorporating behavior competencies. In this research, a unified framework is designed to solve the entire test scenario library generation (TSLG) problem, where a novel method is proposed for the library generation question. Theoretical analysis provides justifications of the proposed method regarding both evaluation accuracy and efficiency. Specifically, the proposed method obtains unbiased index estimation of performance metrics (i.e., accuracy) with a fewer number of required tests (i.e., efficiency). The three case studies verify the proposed methodology and the results show that the evaluation process can be accelerated by 10^3 times compared with the NDD evaluation method, with the same accuracy.

17. Key Words Connected vehicles, automated vehicles, augmented reality, acceptance testing and evaluation	**18. Distribution Statement** No restrictions.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 33	**22. Price**

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

Table of Contents

List of Figures

Project Summary

Testing and evaluation is a critical step in the development and deployment of connected and automated vehicle (CAV) technology. Testing standards for human-driven vehicles, such as Federal Motor Vehicle Safety Standards (FMVSS), were established a long time ago. However, current standards cannot be applied to CAVs, because they often assume the presence of a human driver, who conducts the driving tasks. It is very important to develop test procedures and identify applicable test scenarios (user cases) for CAVS to evaluate the "intelligence" of the vehicle. The intelligence level indicates whether a CAV can drive safely and efficiently without human intervention. The newly released Automated Driving Systems Guideline 2 has made it very clear that the new automated driving systems need validation methods and to be tested by incorporating behavior competencies. In this research, a unified framework is designed to solve the entire test scenario library generation (TSLG) problem, where a novel method is proposed for the library generation question. Theoretical analysis provides justifications of the proposed method regarding both evaluation accuracy and efficiency. Specifically, the proposed method obtains unbiased index estimation of performance metrics (i.e., accuracy) with a fewer number of required tests (i.e., efficiency). The three case studies verify the proposed methodology and the results show that the evaluation process can be accelerated by 10^3 times compared with the NDD evaluation method, with the same accuracy.

1. Introduction

Testing and evaluation is a critical step in the development and deployment of connected and automated vehicles (CAVs). Testing procedures for human-driven vehicles, such as Federal Motor Vehicle Safety Standards (FMVSS), have been established for a long time. However, current standards only regulate automobile safety-related components, systems, and design features, because all driving tasks are performed by human drivers. For CAVs, it is essential to evaluate the "intelligence" of the vehicle [1], similar to a driver's license test, which indicates whether a CAV can operate safely and efficiently without human intervention.

Currently, CAV testing and evaluation is mainly conducted via the following steps: simulation test, closed facility test, and public road test. Simulation test is a cost-effective method, but it is difficult to model exact vehicle dynamics and road environment. Public road test is the most realistic method, but has the following problems: First, at the current stage of CAV technology, safety is still a significant issue. At least four fatal crashes have been reported in the past two years involving automatic driving functions [2]. Second, testing on public roads is extremely inefficient. A CAV would have to drive hundreds of millions of miles, sometimes hundreds of billions of miles to validate both safety and reliability at the level of human driven vehicles [3]. The underlying reason is that most scenarios on public roads are not challenging enough to evaluate the performances of a CAV. Only a small portion of the scenarios are critical, which are rare events on public roads. For instance, if we want to evaluate the safety performance (e.g., accident rate) of a CAV by analyzing its reaction to red light running vehicles at signalized intersections, it may require the CAV to pass thousands or even millions of intersections to accumulate enough accident events, which becomes intractable.

Closed facility test, which can test real CAVs in a controlled environment, has its unique advantages over the other two methods. First, testing real CAVs resolves the problem of modeling exact vehicle dynamics in simulation. Second, the closed facility test provides a more controlled and therefore safer environment for CAV testing than the public road test. Third, the closed facility test has potential to greatly improve the testing efficiency, i.e., obtain the evaluation results with the same accuracy by fewer number of tests.

The key to exploiting the advantages of closed facility test is to generate testing scenario libraries. A testing scenario library is defined as a set of critical scenarios that can comprehensively evaluate certain pre-defined performance metrics. Each scenario in the library has its testing value, which quantitatively measures the criticality of the scenario. After the library is generated, CAVs can be tested in closed facilities by sampling scenarios from the library. Scenarios with smaller testing values are sampled with smaller probabilities. Since the library includes more critical scenarios, the CAV evaluation can be performed much more efficiently than that of public road test. To efficiently and effectively evaluate different CAVs in closed facilities, the testing scenario library generation (TSLG) problem needs to be solved. In this project, we propose four

research questions to describe the TSLG and CAV evaluation process. A novel framework is proposed to solve the problem by utilizing naturalistic driving data (NDD) and augmented reality testing environment.

2. Testing Scenario Library Generation

The TSLG problem can be described as: how to generate a testing scenario library for one scenario type (e.g., car-following), which can be used to accurately and efficiently evaluate different CAVs with a pre-defined performance metric (e.g., safety).

The TSLG problem can be disassembled into four research questions:

(1) How to describe a testing scenario and formulate the decision variables? (Scenario Description)

(2) What are the performance metrics for CAV evaluation? (Metric Design)

(3) How to generate a testing scenario library for a specific performance metric? (Library Generation)

(4) How to use the generated library to evaluate CAVs? (CAV Evaluation)

The first question focuses on the description of testing scenarios and decision variable formulation. A scenario describes the temporal development between a sequence of scenes, which include snapshots of the environment (e.g., background vehicles, road information, and environment conditions) [4]. Decision variables denote what requires to be changed in testing scenarios. Most existing studies construct the decision variables by listing all possible influencing factors, which is infeasible when the testing scenarios are complex. To reduce the complexity, Li et al. [5] described testing scenarios as a temporal-spatial combination of assigned tasks, so the decision variables are formulated as the temporal-spatial locations of assigned tasks. Zhou et al. [6] described testing scenarios by several basic scenarios and a set of transition rules. The PEGASUS project [7] proposed a three-level framework to describe testing scenarios, i.e., functional level, logical level, and concrete level. If parameters of the top two levels are pre-determined, then the decision variables include only the parameters of the concrete level. However, all these methods do not consider the operational design domain (ODD) [8] of testing CAVs. Yet testing scenarios outside the ODD are meaningless for CAV evaluation.

The second question aims to design performance metrics for CAV evaluation. Most current studies only focus on safety, which is usually assessed by indices, e.g., the disengagement rate or the accident rate on public roads [9][10]. Although safety is the foundation of all CAV applications, a safe but over-conservative CAV may fail in simple driving tasks. Therefore, functionality, which represents the vehicle's ability to complete driving tasks, should also be

included in the evaluation process. Furthermore, mobility and rider's comfort can be considered as higher level requirements. Although critical scenarios for different performance metrics may differ, the framework of solving the TSLG problem should remain the same.

The third and the key question is how to generate a testing scenario library for a specific performance metric. The most straightforward method is to design a "test matrix" based on expert knowledge, which is similar to the validation of human-driven vehicles [11][12][13]. However, this method relies heavily on the external input, and the accident typology of CAVs may not be reflected in the predefined test matrix. Improvements were made to generate testing scenarios based on particular CAV models. The worst-case scenario evaluation method (WCSE) was proposed to generate testing scenarios with model-based optimization methods [14]. The critical step of WCSE is to model the exact CAV dynamics and driving behaviors, which is not realistic for implementation. To resolve this problem, some black-box model-based methods were proposed. An adaptive searching method was proposed to generate testing scenarios based on a specific black-box CAV model [15]. However, the "black-box" model method requires to conduct real vehicle testing for each step of scenario searching, which is time-consuming and expensive. Moreover, the generated scenarios can only be applied to a specific CAV, which are not suitable for other CAVs. All these methods can only provide some representative scenarios, which cannot comprehensively evaluate CAVs without a testing scenario library. The PEGASUS project [7] proposed an exhaustive method to construct a testing scenario library, which suffers from computational complexity for high-dimensional scenarios.

The fourth question focuses on CAV evaluation with the generated library. For safety evaluation, most existing methods estimate the accident rate of a CAV using a scenario library from Naturalistic Driving Data (NDD), such as naturalistic field operational tests [16] and crude Monte Carlo method [17][18]. However, this method is proved inefficient and intractable for even low-dimensional scenarios [3]. The evaluation efficiency of low-dimensional scenarios was significantly improved by the accelerated evaluation (AE) method proposed by Zhao et al. [10]. The importance sampling technologies were first applied into the CAV evaluation problem. The major idea is to construct an importance function, which attaches more importance to critical scenarios. However, each step of searching the importance function is based on one test run of a real CAV. Thus it is time-consuming and expensive to construct the importance function for high-dimensional scenarios. As a result, under high-dimensional car-following scenarios, the AE method degrades to a white-box method with the assumption of knowing exact CAV models [19], which is usually impossible for real applications. Moreover, the generated scenarios can only be applied to a specific CAV, which is not generic.

Notwithstanding the related studies, all existing methods have limitations in either scenario types that can be handled (e.g., low-dimensional scenarios only), CAV models (e.g., a specific CAV only), or performance metrics (e.g., safety evaluation only). To the best of our knowledge, there

is no existing study that integrates all parts of the TSLG problem together and generates libraries for different scenario types, CAV types, and performance metrics. In this project, a unified framework is designed to solve the entire TSLG problem, where a novel method is proposed for the library generation question.

2.1 Scenario Description

The terms scene and scenario defined in [4] are adopted. A scene describes a snapshot of the environment including the scenery and dynamic elements. A scenario describes the temporal development between several scenes in a sequence of scenes. The scenery includes all geo-spatially stationary elements, which entails metric, semantic, and topological information about roads and all their components like lanes, lane markings, road surfaces, or the roads' domain types. The dynamic elements are moving or have the ability to move, e.g., pedestrians and vehicles. Slightly different with the definitions in [4], a scene denotes ground truth of the environment (objective) in this paper, instead of observations (subjective). Therefore, the scene representation is considered to be static.

Testing scenarios should be consistent with the operational design domain (ODD) of testing CAVs. The ODD describes the specific conditions under which a given CAV is intended to function [8]. To define the capability boundaries, the following information is required at a minimum in the ODD: roadway types, geographic area, speed range, and environmental conditions. Therefore, most of the scenery and part of dynamic elements have been specified in the ODD. The determination of remaining parts of scenarios is the critical step to generate testing scenarios. If the remaining parts are denoted as a vector of decision variables x, e.g., acceleration profiles of background vehicles, a testing scenario is generated with each realization of x. If the ODD is defined following some specific structures, e.g., the three-level structure in the PEGASUS project [7], then the vector can be formulated in the simplified way. For less specified ODD, the vector should include temporal variables of dynamic elements and spatial variables of scenery, e.g., trajectories of all traffic participants and spatial development of road parameters

2.2 Metric Design

Performance metrics define what aspects a CAV needs to be evaluated. Most existing studies focus only on safety evaluation, which is essential but insufficient for a commercialized CAV. In this project, we define the performance metrics to reflect people's incremental expectations towards CAVs, including safety, functionality, mobility, and rider's comfort, as shown in Figure 1.

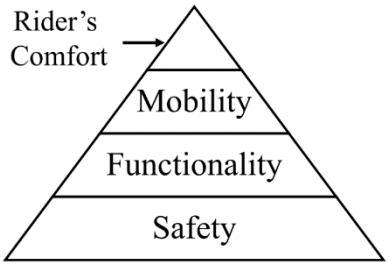


Figure 1 Illustration of the incremental performance metrics

[Description: This figure shows a pyramid with four layers of performance metrics. From the top to bottom are rider's comfort, mobility, functionality, and safety]

Safety is the foundation of all CAV applications, which is usually assessed by the accident rate during the test without human intervention or the disengagement rate [9][10]. Taking the commonly used scenario, i.e., cut-in scenario, for an example, a background vehicle (BV) changes its lane in front of a CAV in the adjacent lane with pre-determined parameters, i.e., cut-in distance and speed difference. Whether an accident (e.g., conflict or crash) may happen or not depends on the CAV's response to the BV's maneuvers. After a certain number of tests with varying parameters, the accident rate of the CAV could be estimated, which is used to indicate the safety performance in the lane-change scenario.

The second level of the performance metric is functionality, which is defined by whether a CAV can complete a given task in a specific scenario. Considering a scenario that a CAV needs to make a lane change to the right and exit the highway within a certain distance, several BVs are driving on the right lane following pre-determined parameters (e.g., initial distance to the CAV, acceleration profiles). If the CAV is very conservative and keeps a long safety distance with surrounding vehicles, it may fail to complete the lane-change task before the freeway exit. In the case, the vehicle may pass the safety evaluation but fail in the functionality evaluation. Similar to safety evaluation, the functionality of a CAV can be evaluated by the failure rates of the CAV in completing certain driving tasks with different environment settings and BVs' trajectories.

Both safety and functionality are critical for CAV evaluation at the current technology maturity level. Unless a CAV can safely complete all driving tasks without human interventions, it may not be accepted by the general public. For higher level requirements, mobility and rider's comfort should also be considered into the evaluation scope. Mobility is utilized to measure the travel efficiency in completing a series of driving tasks, while rider's comfort measures the physical and psychological feeling of passengers.

2.3 Library Generation

To generate the testing scenario library, the criticality of scenarios is defined, and the searching method is designed for efficiently searching critical scenarios. An illustration of the entire framework is shown in Figure 2. The proposed definition provides theoretical foundation to construct the optimal importance function and indicates that both maneuver challenge and exposure frequency are critical for CAV evaluations, which is fundamentally different from most existing studies.

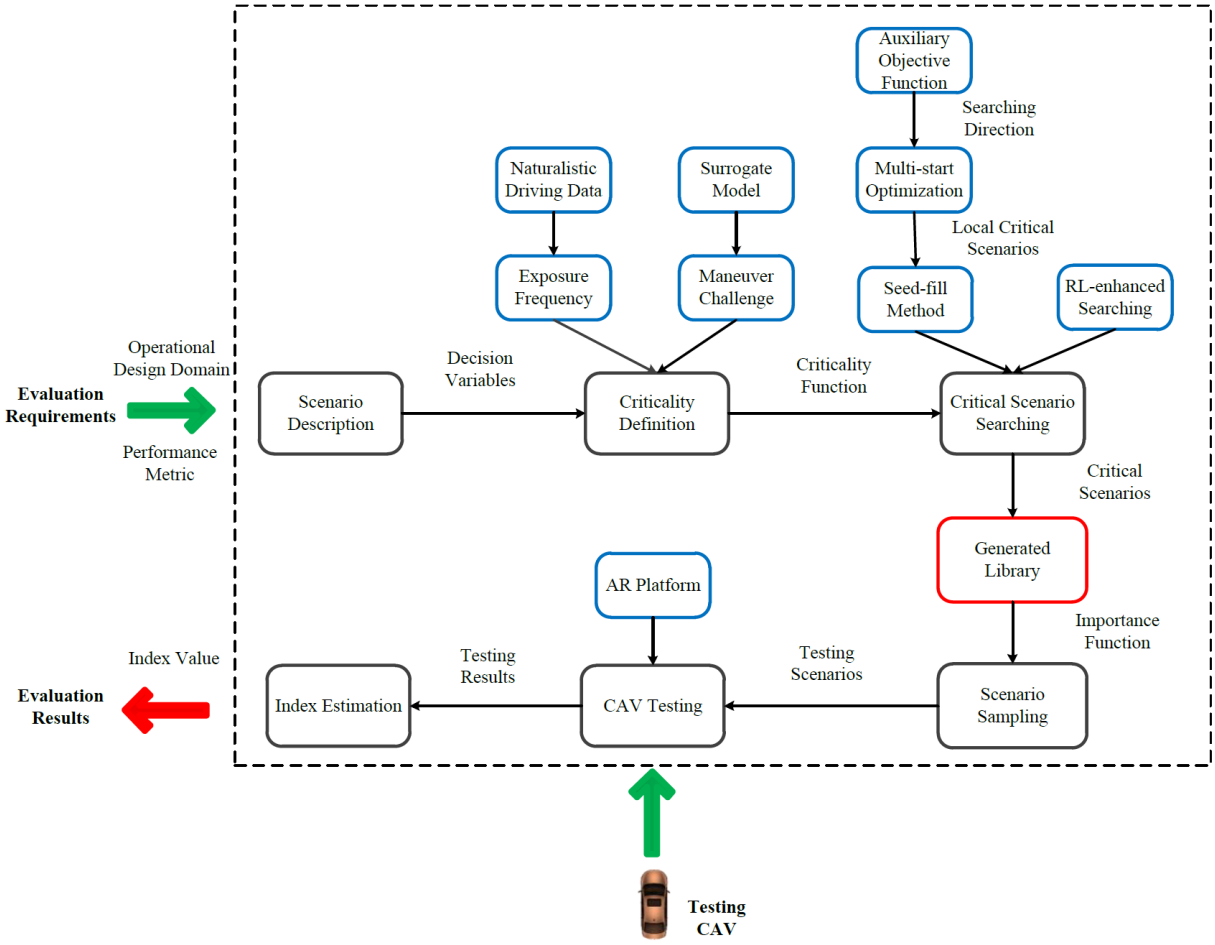


Figure 2 Proposed framework to the TSLG problem.

[Description: This figure shows a flow chart of the TSLG problem. The follow chart starts with evaluation requirements (operational design domain, and performance metric) as input to generate scenario description. Decision variables from the scenario description are input to the criticality definition. Other inputs for criticality definition are exposure frequency from naturalistic driving data and maneuver challenge from surrogate model. The output of criticality definition is the criticality function, which is the input for critical scenario searching. Other input to critical scenario searching include seed-fill method and RL-enhanced searching. The output of critical scenario searching is the critical scenarios, which are the input of the generated

library. Sampling from the generated library results in the testing scenarios. Combining with the AR platform, CAV can be tested, and the testing results can be used for index estimation. Finally, based on the index values, the evaluation results can be obtained.]

2.3.1 Definition of Criticality

The criticality of a scenario measures the importance in evaluating a performance metric. In ISO 26262 [21], the risk assessment of a scenario was defined as a combination of severity of injuries, exposure classification, and controllability classification. The exposure classification denotes the relative expected exposure frequency of the scenario where the injury can possibly happen. The controllability classification denotes the relative likelihood that the driver can act to prevent the injury. Inheriting the concepts of the risk assessment, we define the criticality of scenarios as

$$V(x|\theta) \overset{\text{def}}{=} P(S|x,\theta)P(x|\theta) \tag{1}$$

Where θ denotes the specified parameters in the ODD, x denotes the vector of decision variables, and S denotes the event of interest (e.g., accident) with a surrogate model (SM) of CAVs. The SM is designed to encode the common features of CAVs. A well-generated library should include more critical scenarios for most CAVs, and the introduction of the SM contributes to achieving this goal. An ideal SM should be calibrated from actual CAV driving data similar to human driving model calibration [22]. At the current stage, however, there is very little open CAV data available for public research. Therefore, we propose to calibrate the SM based on the human driving data, i.e., NDD. It is a reasonable starting point because of the following reasons. First, the common features of human drivers are the natural baselines for CAV evaluation. Critical scenarios for human drivers are the most straightforward testing scenarios for CAVs. Second, CAV is essentially an application of "artificial intelligence", the purpose of which is to mimic and outperform "human intelligence" [1]. Many CAV algorithms are obtained by imitating human driving behaviors, e.g., end-to-end learning method [23][24]. Third, a "human-like" CAV can improve safety in a mixed traffic condition, where CAVs and human-driven vehicles coexist on the roadway. A similar concept of "roadmanship" was recently proposed for CAV evaluation [25]. Therefore, it is reasonable to represent the common features of CAVs based on human naturalistic driving data.

The proposed definition is a conceptual generalization of the risk assessment in ISO 26262 [21]. The left term $P(S|x,\theta)$ measures the probability that CAVs encounter the event of interest in the scenario. The severity is encoded by determining the interested event, and the controllability classification is encoded by the probability. The right term $P(x|\theta)$ denotes the probability of the scenario occurring on public roads, which encodes the exposure classification. Different from the classification methods in ISO 262262, we generalize the concepts from safety to generic metrics, introduce the concept of SM, and define the criticality in a quantitative way. The justifications of this definition are theoretically proved regarding the evaluation accuracy and efficiency in [26].

To calculate the criticality, $P(x|\theta)$ can be obtained from NDD, and $P(S|x,\theta)$ is obtained by simulations of the SM.

The definition also indicates that both maneuver challenge $P(S|x,\theta)$ and exposure frequency $P(x|\theta)$ are critical for CAV evaluations. This is fundamentally different from most existing studies, which usually overvalue the infrequent scenarios. For instance, the worst-case scenario evaluation [14] focuses on the worst-case (i.e., most dangerous) scenarios for safety evaluation. The accelerated evaluation method for the car-following scenarios [19] maximizes the likelihood of the occurrence of accidents (e.g., crash or conflict), which generates the most infrequent scenarios. All these methods essentially focus on the most infrequent scenarios, which happen to be the most challenging scenarios for safety evaluation. However, for functionality evaluation, as an example, there is no explicit relation between the maneuver challenges (i.e., difficulty) and exposure frequency. All existing methods overvalue the challenging part but ignore the exposure frequency of scenarios. Taking an extreme example for conceptual explanation, the scenario that a meteor hitting a car is extremely dangerous but we cannot evaluate the performances of CAVs based on testing results from these extremely low frequent scenarios. The common and challenging scenarios are more critical for CAV evaluation.

2.3.2 Critical Scenario Searching

The next problem is how to search critical scenarios in the whole scenario space. The basic idea is to find local critical scenarios by optimization methods and then search their neighbor scenarios. However, directly using the criticality function as the objective function is problematic. Most scenarios are uncritical with zero criticality and zero gradient of criticality, i.e., local minimal. If a scenario is uncritical, its criticality function provides little information of searching direction for critical scenarios. Therefore, the optimization process degrades to a random sampling process, which is inefficient for complex scenarios.

To resolve this issue, an auxiliary objective function is designed to guide searching directions, and the seed-fill method is applied to search neighbor scenarios. The auxiliary objective function is designed as the combination of maneuver challenge and exposure frequency, similar to criticality definition. A commonly used multi-start optimization method is applied to obtain a number of local critical scenarios. Specifically, multiple initial points are generated by space filling methods (e.g., random sampling). After solving the optimization problem from each initial point, local critical scenarios are obtained. The parameters from the ODD are considered as constraints, e.g., speed limit, acceleration limit, perception range, etc. The number of initial points increases with the dimensions of the decision variables. The dimension of the decision variables can be greatly reduced by exploiting their specific structures, e.g., independence properties. Using the local critical solutions as starting points, other critical scenarios are expanded by the seed-fill method. Seed-fill, also called flood-fill, is a basic method in computer graphics [27] that determines the area connected to a given node in multi-dimensional arrays. The key idea is to exhaustively

explore the critical points of unexplored space rather than all of the space from the starting point outwards [28]. The criticality function instead of the auxiliary objective function is calculated in this step. The threshold of critical scenarios is theoretically analyzed in [26].

To illustrate the searching method, two typical non-convex objective functions, i.e., Peaks function and Ackley function [29], are studied, as shown in Figure 3 (a, c). The fifty and one-hundred initial searching points are sampled for the two functions respectively. In this illustration, the criticality function is calculated by the normalized objective function, and the threshold is manually selected. As shown in Figure 3 (b, d), critical scenarios of the both functions are effectively obtained by the proposed searching method as red areas.

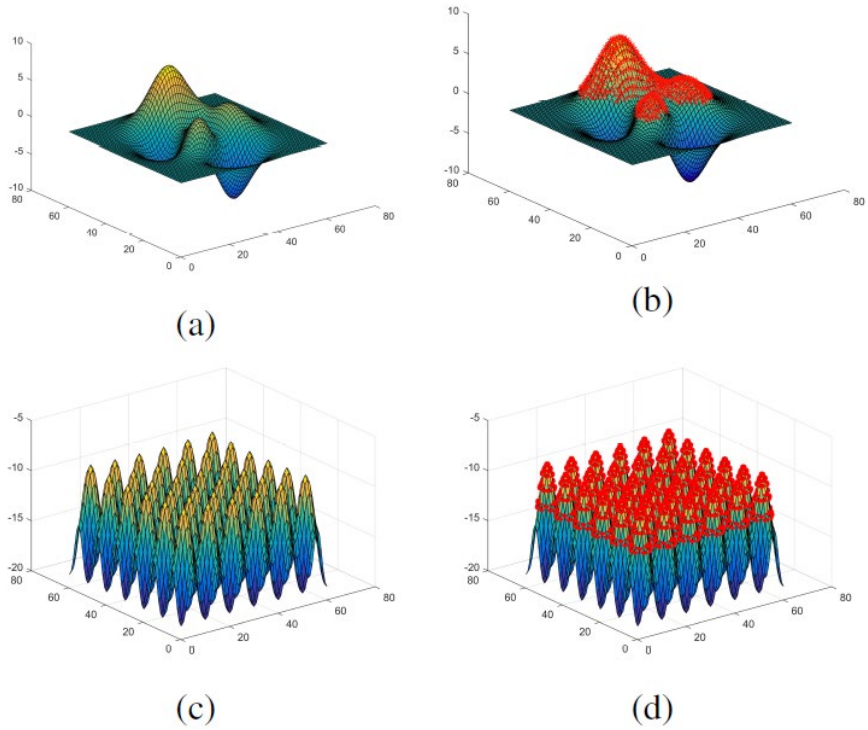


Figure 3 Critical scenario searching method for Peaks function (a) and Ackley function (c). The critical scenario are obtained as red points in (b) and (d) respectively.

[Description: This figure has four sub figures. The first subfigure shows the pdf of the Peaks function. The second subfigure shows the critical scenarios (in red dots) obtained from the Peaks function. The third subfigure shows the pdf of the Ackley function. The fourth subfigure shows the critical scenarios (in red dots) obtained from the Ackley function.

2.4 CAV Evaluation

After the library is generated, the next step is to evaluate CAVs with the generated library. As shown in Figure 2, three steps are designed, i.e., scenario sampling, CAV testing, and index estimation. The importance function is constructed based on the generated library.

The first step is to sample testing scenarios according to the generated library. The major challenge is how to balance exploitation and exploration. Critical scenarios are obtained based on the surrogate model (SM), which usually has dissimilarity compared with the testing CAV. Therefore, the generated library may miss some critical scenarios when testing a specific CAV. To solve this issue, besides sampling scenarios from the library according to their criticality values (i.e., exploitation), the scenarios outside the library is also sampled with a small probability (i.e., exploration). To better understand the trade-off between the exploitation and exploration, we compare the greedy sampling policy and ϵ greedy sampling policy. The greedy sampling policy greedily exploits the scenarios in the library. By this policy, all testing scenarios are sampled based on the normalized criticality values. The ϵ greedy sampling behaves greedily most of the time, but with small probability ϵ, it selects scenarios randomly outside the library with equal probability (i.e., exploration). This simple yet efficient method is commonly used for balancing exploitation and exploration [30].

The second step is to test the CAV with sampled scenarios. To provide a controllable, safe, and cost-effective testing environment, the augmented reality (AR) testing environment [31] is applied. Figure 4 is an illustration of the AR platform designed for Mcity, a newly established closed CAV testing facility at the University of Michigan. The platform combines the real-world testing facility and a simulation platform together. Movements of testing CAV in the real world are transmitted to the simulation platform by roadside units (RSUs), and the information of simulated BVs is fed back to testing CAV. The traffic control in the real world is synchronized with simulation. In this way, BVs in the simulation and testing CAV in the real-world can interact with each other. The initial conditions and maneuvers of BVs are determined by the sampled testing scenarios and imported in the AR platform as virtual vehicles. The testing CAV is running in the real testing facility, which responds to the maneuvers of virtual BVs. The testing can be repeated easily by sampling different scenarios from the library, which results in different BV movements. The total number of testing is determined by the required evaluation precision and confidence level [10][32][33].

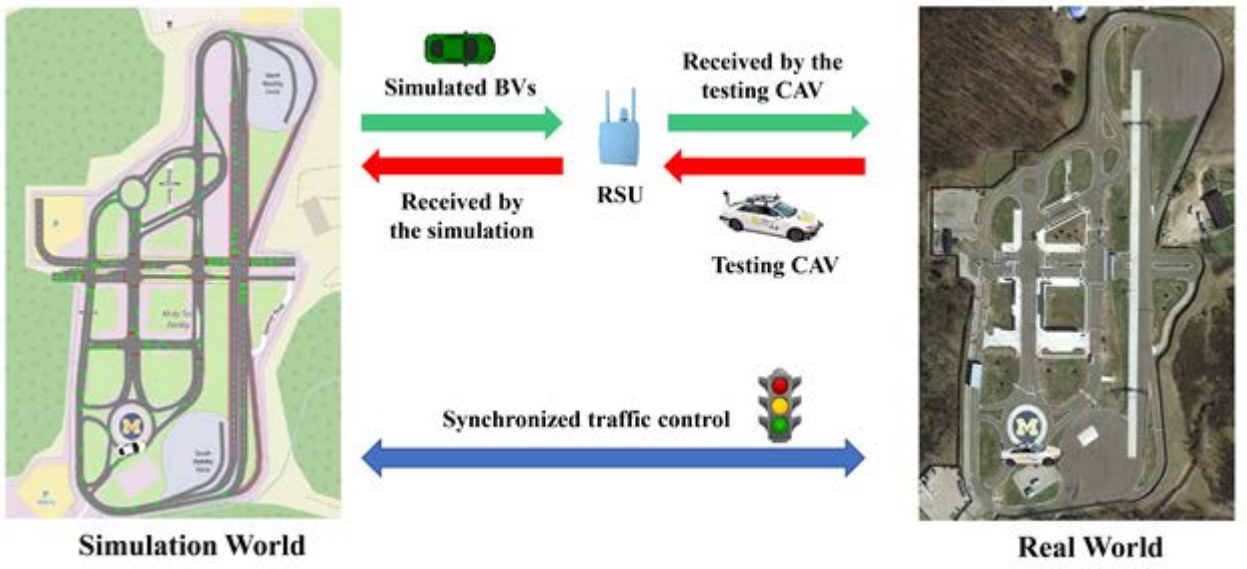


Figure 4 Augmented Reality Testing Platform at Mcity

[Description: This figure describes the augmented reality testing platform at Mcity. The left part shows the VISSIM simulation model of MCity while the right part shows a GoogleEarth map of the Mcity test facility. In the middle, two green arrows show how simulated BVs transmit their information to the testing CAV in the test facility. Two red arrows show ow real testing CAV transmit its information to the simulation environment. A blue arrow shows how traffic signal information are transmitted from the test facility to the simulation environment]

After the testing results are collected in the second step, the third step is to estimate the index value of the performance metric. The index value can be estimated as:

$$\hat{P}(A|\theta) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n}\frac{P(x_i\theta)}{P(x_i\theta)}P(A|x_{i,}\theta) \tag{2}$$

Where n denotes the total number of the sampled testing scenarios, $P(x_i|\theta)$ represents the scenario's occurrence probability in the naturalistic driving environment, $\bar{P}(x_i|\theta)$ denotes the importance function, i.e., the sampling probabilities depending on the policy (greedy or ϵ greedy), and $P(A|x_i,\theta)$ is estimated by the testing results.

More theoretical analysis regarding the accuracy and efficiency of the proposed can be found in [26].

3. Case Studies

Two case studies are designed to evaluate the proposed scenario library method as shown in Figure 5. (1) Cut-in case: a background vehicle (BV) makes a lane change in front of the testing CAV. (2) Highway exit case: the testing CAV needs to make a lane change to the right and exits the highway within a certain distance.

The cut-in case illustrates each step of the scenario library generation and evaluation framework regarding safety. A few specific questions are elaborated, i.e., auxiliary objective function design, NDD analysis, and SM construction. Moreover, because the cut-in case is low dimensional (i.e., two dimensions), it is convenient to visualize the results by figures and help readers better understand the proposed methods.

The highway exit case focuses on the functionality evaluation. Compared with safety evaluation, the major difference lies in the design of auxiliary objective function for the library generation, i.e. how to quantify the maneuver challenge regarding functionality. To this end, several new concepts are proposed, i.e., task, task solution, task solution difficulty, and task difficulty. The specific auxiliary objective function is designed for the highway exit case based on the concepts.

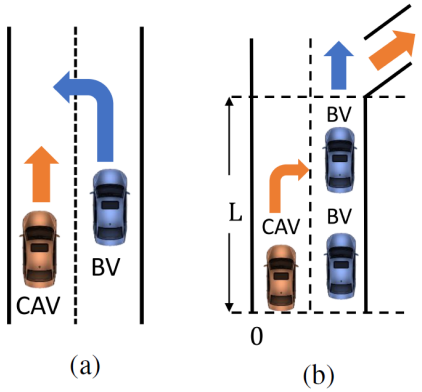


Figure 5 Case Studies (a) cut-in (b) highway exit

[Description: This figure has two subfigures. The left subfigure is an illustration of the cut-in case. A blue BV is cutting in a red CAV. The right subfigure is an illustration of the highway exit case. Two blue BVs are traveling on the right lane close to an exit ramp. A red CAV is traveling on the left lane and try to make a lane change to the right lane before the exit.]

3.1 Cut-in Case Study

Similar to most existing studies [7][10], the decision variable vector of the cut-in case is simplified as two dimensions, i.e.,

$$x = \left[R, \dot{R} \right]^T \tag{3}$$

Where R and $\dot{R}$ denote the range and range rate at the cut-in time respectively. For simplification, the BV is assumed to keep constant velocity after the cut-in behavior, and parameters of road environments are pre-determined. All these pre-determined parameters are denoted as θ. The accident rate is utilized to measure the safety performance of CAVs in the cut-in case. The road test method is simulated to estimate the accident rate as a baseline. Specifically, if a testing CAV drives on public roads, experiences *n* specified cut-in scenarios, and has *m* accident events, the accident rate of event *A* can be estimated by

$$P(A|\theta) \approx \frac{m}{n} \tag{4}$$

The public road test is simulated based on naturalistic driving data (NDD), so the method is denoted as NDD evaluation method in this report.

To provide searching directions for critical scenarios, an auxiliary objective function is designed as the combination of maneuver challenge and exposure frequency. First, the maneuver challenge is estimated by minimal normalized positive enhanced time-to-collision (mnpETTC). As discussed in [34][35], ETTC is one of most widely used indices of safety evaluation for varying velocity scenarios. Second, the exposure frequency of a scenario is estimated by the distance between the scenario and a common set (i.e., scenarios with high exposure frequency). The common set is determined by NDD analysis. More details about the definitions of mnpETTC and the distance to the common set can be found in [36].

Finally, the auxiliary objective function for safety evaluation in the cut-in case is formulated as:

$$\min_x J(x) = \min_x (mnpETTC(x) + w \times d(x, \Omega)) \tag{5}$$

Where $w \in (0,1]$ is a balance weight and $d(x, \Omega)$ is the distance between the scenario *x* and common set Ω.

NDD is analyzed to provide exposure frequency measurement, determine parameters of the auxiliary objective function, and calibrate the SM. The NDD from the Safety Pilot Model Deployment (SPMD) program at University of Michigan [37] is utilized for the cut-in case. The SPMD database is one of the largest databases in the world that recorded naturalistic driving behaviors over 34.9 million miles from 2,842 equipped vehicles in Ann Arbor, Michigan. In the database, there are 98 sedans equipped with the data acquisition system and MobilEye, which enables 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 from the database [10][38]: (a) the vehicles' speeds at the cut-in time belong to (2m/s; 40m/s); (b) the range at the cut-in time belongs to (0.1m; 90m). All these criteria are consistent with the pre-determined parameter θ. As a result, 414,770 qualified cut-in events are

successfully obtained. Figure 6 shows the location distribution of the events. The exposure frequency distribution (i.e., $P(x|\theta)$ is shown in Fig. 3, where brighter color denotes higher exposure frequency, i.e., the common set. The range and range rate are discretized by 2m and 0.4m/s respectively. The NDD evaluation method is equivalently sampling testing scenarios from this probability distribution.

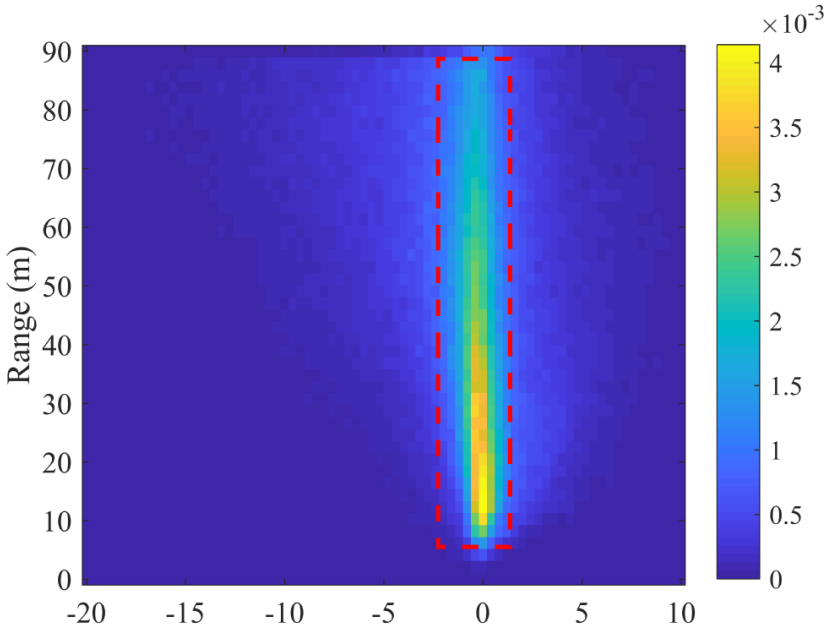


Figure 6 Distribution of the cut-in range and rage rate in NDD. The dashed red rectangle denotes the boundary of the common set.

[Description: This figure shows a heat diagram with range rate as the horizonal axis and range as the vertical axis. The color represents the probability of a cut-in event with a certain range and range rate. The dashed red rectangle denotes the boundary of the common set.]

SM construction is a very important step in the library generation process. In this case study, the commonly used intelligent driving model (IDM) is calibrated by the NDD [39] and selected as the SM for the car-following behaviors of CAVs after the cut-in event. The constraints of acceleration and velocity are added to make the model more practical (i.e., model accident-prone behaviors). Figure 7 shows the safety performance of the selected SM, where the SM has accidents in scenarios of the yellow region.

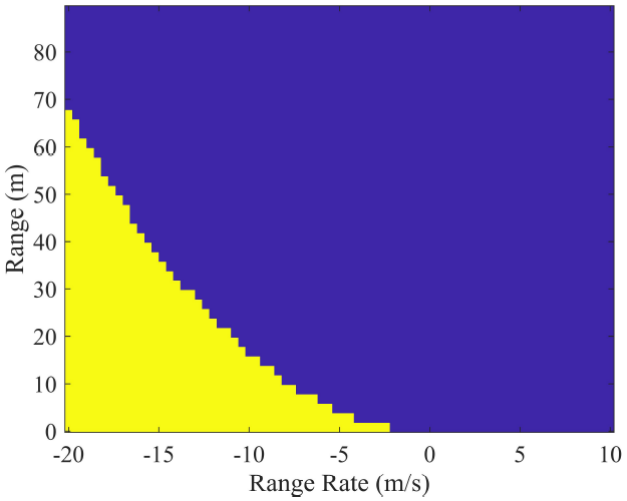


Figure 7 Safety performance of the SM, where the SM has accidents in scenarios of the yellow region

[Description: This figure shows a heat diagram with range rate as the horizonal axis and range as the vertical axis. The low left corner is represented as yellow color, which indicates the SM has accidents. Other areas are represented as blue color, which indicates the SM has no accidents.]

The optimization and seed-fill based method is applied to search for critical scenarios and construct the library. In this case, 50 points are uniformly sampled as the initial starting points. Figure 8 shows the obtained probability distribution after the library generation process. The color denotes the probability of a scenario, i.e., the normalized criticality. Compared with Figure 6, where only exposure frequency is considered, the new distribution encodes more domain knowledge, i.e., maneuver challenge and exposure frequency of scenarios. The library is constructed by the critical scenarios. In this case, the generated library contains a total number of 184 scenarios, which is about 5.38% of all scenarios.

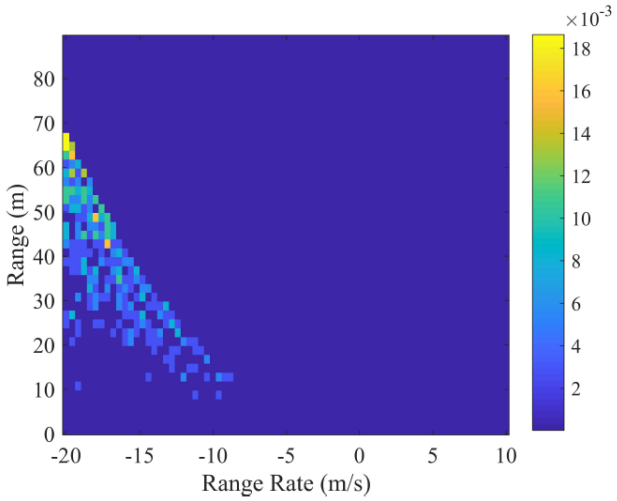


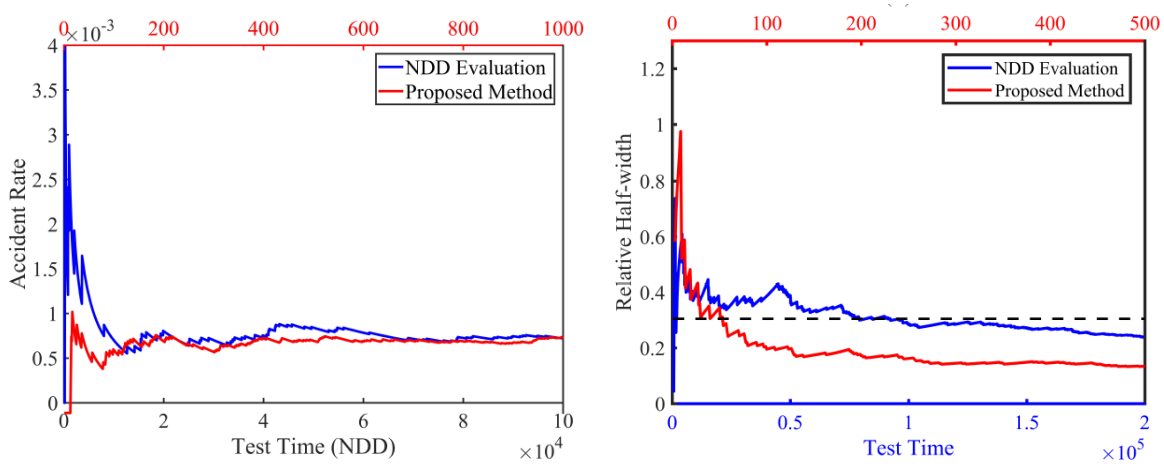
Figure 8 Generated library for the cut-in case. The color denotes the new scenario sampling probability.

[Description: This figure shows a heat diagram with range rate as the horizonal axis and range as the vertical axis. The color represents the sampling probability of a scenario with a certain range and range rate. Most of the high probability scenarios are sampled from the area with range from 30 to 70m and range rate from -10 to -20 m/s]

For field implementation, a real CAV should be tested. In this paper, simulation is used to validate the proposed method. Although a simulated CAV model cannot exactly reflect dynamics of a real CAV, it is used in this paper as a proof of concept to validate the proposed method.

A commonly used CAV model is selected, which combines adaptive cruise control and autonomous emergency braking functions (see [10] for details). The NDD evaluation method is applied as the baseline, where testing scenarios are sampled from the NDD distribution in Figure 6. For the proposed method, testing scenarios are sampled from the generated library in Figure 8. The ϵ greedy sampling policy is applied with ϵ = 0.05. The chosen CAV model is tested in the sampled scenarios, and an accident event is recorded if the vehicle range is smaller than a threshold.

Figure 9 shows the comparison of the two evaluation methods. The blue line denotes the results of NDD evaluation method, and the bottom x-axis denotes its number of tests. The red line denotes the results of the proposed method, and the top x-axis denotes its number of tests. As shown in Figure 9 (a), both methods can obtain accurate estimation of the accident rate for a predetermined relative half-width (e.g., β = 0.3). Figure 9 (b) shows that the proposed method achieves this confidence level after 51 tests, while the NDD evaluation method needs 9.63×10^4 tests. The proposed method is about 1,888 times faster than the NDD evaluation method (i.e., efficient). Because the most time-consuming and expensive step in the CAV evaluation process is expected to be the vehicle testing, the proposed method can significantly save both time and money compared to the NDD evaluation method.



(a) (b)

Figure 9 Safety evaluation results of the cut-in case: (a) estimation results of the accident rate; (b) relative half-width of the estimation results.

[Description: This figure has two subfigures to illustrate the result of the cut-in case study. The left subfigure shows the estimated accident rate between our proposed method in red curve and NDD evaluation in blue curve. Horizontal axis is the test time and the vertical axis is the accident rate. The figure shows both methods converge to the same accident rate. The right subfigure shows relative half-width convergence of the estimated results. Horizontal axis is the test time and the vertical axis is the relative half width. This figure shows our proposed method converges 1,888 times faster than the NDD evaluation given the required relative half width to be 0.3.]

3.2 Highway Exit Case Study

The highway exit case study is designed to evaluate the functionality of a CAV. As shown in Figure 5 (b), the decision variable vector of the highway exit scenario should include initial states of the CAV, number of BVs, and trajectories of each BV, which is high-dimensional. To simplify the problem and focus on the functionality evaluation, the initial position and velocity of the CAV are pre-determined as p_0 and v_0, the number of BVs is pre-determined as two, and all BVs keep their initial velocity unless the distance is less than a threshold d_{cf}, when the following BV will change its speed to be the same as the leading BV. As a result, the decision variable vector is formulated as:

$$x = [P_{0,1}, v_{0,1}, P_{0,2}, v_{0,2}]^T \tag{6}$$

where $p_{0,i}$, $v_{0,i}$ denote the initial position and velocity of the i^{th} BV. Although the simplified problem cannot exactly reflect the actual highway exit scenarios, it can be used as a demonstration of functionality evaluation.

The library generation methods are the same as the cut-in case, except for the auxiliary objective function design.

Similar to the cut-in case, the auxiliary objective function is composed of exposure frequency and maneuver challenge. To evaluate the maneuver challenge for generic functionality, four new concepts are proposed, i.e., task, task solution, task solution difficulty, and task difficulty. The "task" is defined based on the functionality, e.g., exit from the highway. The "task solution" denotes a feasible CAV trajectory to complete the task. The "task solution difficulty" denotes the difficulty in completing the task solution. Finally, the "task difficulty" denotes the difficulty of the task, which can be evaluated by the summation of all task solution difficulties as:

$$M_f(x) = \sum_{f \in F} W(f) \tag{7}$$

Where f is a feasible task solution and F is the set of all feasible task solutions; $W(f)$ is the difficulty in completing the task solution f. Note that $W(f)$ is negative and large $W(f)$ means higher difficulty.

For the specified highway exit case, the maneuver challenge is evaluated based on the proposed concepts. The task is to make a lane change to the right before reaching the off-ramp location. The task solution is defined as a feasible lane-change point $f = (t, p)$, where t is the lane-change time and p is the lane-change position. The feasible lane-change zone $f \in F$ is determined by maximal/minimal velocity (v_{max}; v_{min}), highway exit location (L), safe time-to-collision gaps (t_{min}), and reachability of the CAV. The reachability denotes whether the CAV can reach certain position at certain time considering the maximal/minimal acceleration (a_{max}, a_{min}) and maximal/minimal velocity. Figure 10 illustrates an example of the feasible lane change zone for a specific scenario, i.e., $= [-25, 34.5, -100, 40]^T$. The initial position of the CAV is set zero. The lane change boundary is determined by the maximal/minimal velocity and the off-ramp location, denoted as the red dashed line. The feasible lane change zone, i.e., F, consists of three isolated zones, separated by the trajectories of BVs. The gaps between F and the lane change boundary come from the reachability of CAVs. The gaps between F and the trajectories of BVs come from the safe time-to-collision gaps.

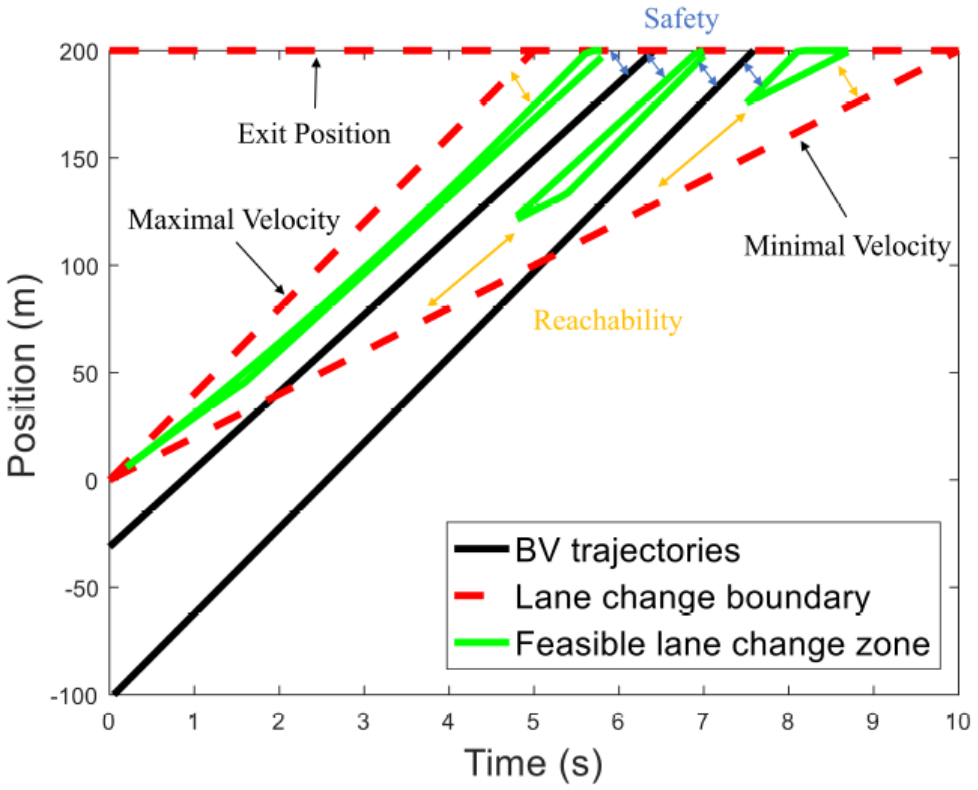


Figure 10 Task difficulty evaluation of the highway-exit case

[Description: This figure shows a time-space diagram in the high-exit case. Two black curves are BV trajectories. Red doted lines are boundaries of CAV trajectories with maximum speed, minimum speed, and exit position. Areas in the green curves are the feasible lane change zone for CAV.]

For simplicity, we assume all task solutions of this case have the same task solution difficulty. Finally, the auxiliary objective function of the highway exit case is designed as:

$$\min_x J(x) = \min_x (\frac{S(F)}{U_s} + w \times d(x, \Omega)) \quad (8)$$

Where $S(F)$ denotes the area of the feasible lane-changing zone; U_s is a normalization factor and w is the weight.

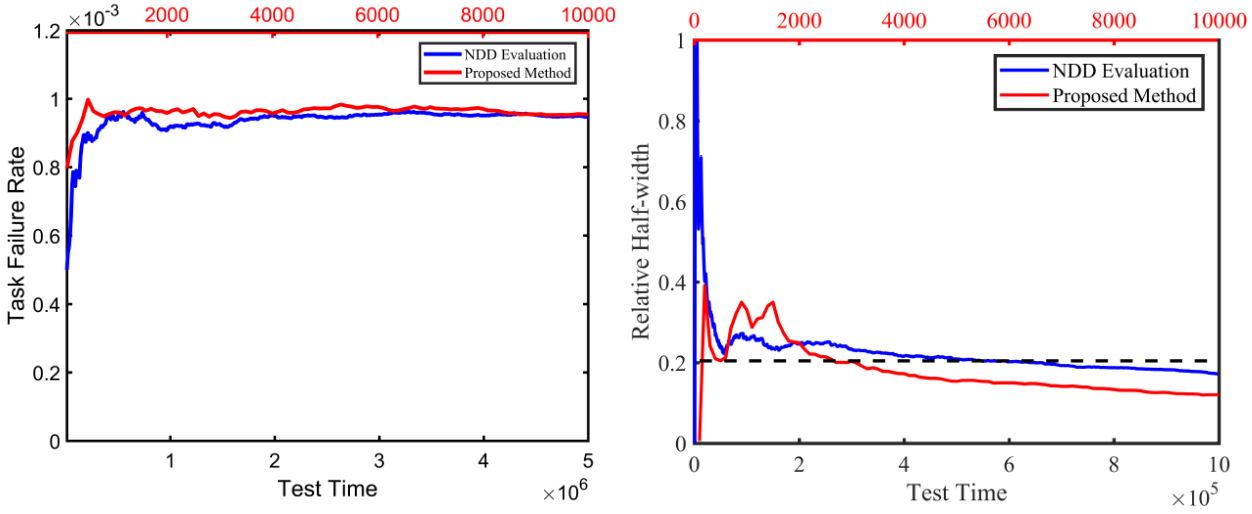
The NDD from the Integrated Vehicle- Based Safety System (IVBSS) project is used to provide exposure frequency information [40][41]. In the IVBSS project, 108 randomly sampled drivers from different ages used sixteen Honda Accords vehicles in an unsupervised manner for over a 40-day period. Query criteria are designed to extract car-following events from the database as: (1) vehicle was traveling on a highway; (2) vehicle was traveling at a speed of at least 20m/s (about 45mph); (3) cruise control function was not activated; (3) surface condition is dry; (4) light condition is day. The resulting dataset represents a total of 5×10^4 car-following events and 1.47×10^6 points of car-following trajectories. The exposure frequency of a scenario can be estimated as:

$$P(x|\theta) = P(p_{0,1}|\theta)P(v_{0,1}, R, v_{0,2}|\theta) \quad (9)$$

Where $R = p_{0,1} - p_{0,2}$, $P(p_{0,1}|\theta)$ denotes the initial position probability of the leading vehicle, which can be estimated by uniform distribution, and $P(v_{0,1}, R, v_{0,2}|\theta)$ is obtained from the distribution of car-following trajectories in the NDD. The MOBIL ('minimizing overall braking induces by lane changes') model is used as the SM in this case. The MOBIL model was proposed to derive human lane-changing rules for discretionary and mandatory lane changes [42]. It provides the utility measurement method for deciding which gap has a desirable lane change position. To predict the CAV's trajectories before the lane-change, the Model Predictive Control (MPC) [43] is applied, and the trajectory with higher predictive utility of lane change will be chosen as the solution to the task. After applying the critical scenario searching method, the testing scenario library of the highway exit case is generated. The total number of critical scenarios in the library is 1,895, which is about 0.12% of all scenarios.

A typical CAV lane-change model is evaluated in this case study, where the lane-change utility is evaluated by average travel time, average time gap density, and remaining travel time of different lanes (see details in [44]). Similarly, the NDD evaluation method is used as the benchmark. In the proposed method, testing scenarios are sampled from the generated highway exit library, and events of task failures (i.e., cannot exit from the highway) are recorded. Similar to the cut-in case, the ϵ-greedy sampling policy is applied with $\epsilon = 0.10$. The task failure rate is estimated to measure the functionality performance of the CAV model in the highway exit case.

After the estimated task failure rate converges to a certain estimation precision, the estimated task failure rate is obtained, and the evaluation process is completed. Fig. 8 shows the comparison of the two evaluation methods. The legends and axis are the same as the cut-in case. Similar with the previous case study, both methods can obtain unbiased estimation of the failure rate with the relative half-width ($\beta = 0.2$). Fig. 8 (b) shows that the proposed method achieves this estimation precision after 2.6×10^3 tests, while the NDD evaluation method takes 6.6×10^5 tests. The proposed method is about 255 times faster than the NDD evaluation method. The efficiency of the proposed method is influenced by the "dissimilarity" between the SM and the specific CAV model. It is the main reason why the efficiency of the proposed method in the highway case is lower than that in the cut-in case.



(a) (b)

Figure 11 The functionality evaluation results of the highway exit case: (a) estimation results of the task failure rate; (b) relative half-width of the estimation results.

[Description: This figure has two subfigures to illustrate the result of the highway exit case study. The left subfigure shows the estimated accident rate between our proposed method in red curve and NDD evaluation in blue curve. Horizontal axis is the test time and the vertical axis is the accident rate. The figure shows both methods converge to the same accident rate. The right subfigure shows relative half-width convergence of the estimated results. Horizontal axis is the test time and the vertical axis is the relative half width. This figure shows our proposed method converges 255 times faster than the NDD evaluation given the required relative half width to be 0.2.]

4. Findings and Recommendations

In this project, we proposed a unified framework to solve the testing scenario library generation (TSLG) problem for CAV evaluation. The framework can be used to generate testing scenario libraries for different scenario types, performance metrics, and CAV models.

A novel method was proposed to generate testing scenario libraries. The criticality of scenarios was defined as a combination of maneuver challenge and exposure frequency, which is more reasonable than that of most existing studies. A searching method is designed to efficiently obtain the critical scenarios.

To evaluate the maneuver challenge of scenarios, the surrogate model (SM) of CAVs was introduced, which contains the common features of CAVs. Although the dissimilarity between the SM and specific CAVs cannot be eliminated, it provides the theoretical foundation for progressively improving the efficiency by mitigating the dissimilarity. We believe that utilizing the domain knowledge (e.g., common features and NDD) has huge potentials for future study in this field. To validate the proposed method, the evaluation accuracy and efficiency were proved by theoretical analysis.

Two case studies are conducted to demonstrate the performance of the proposed method: Cut-in and highway exit. The cases were designed to reflect the general framework as well as unique features including auxiliary objective function design for different performance metrics (i.e., safety and functionality), Naturalistic Driving Data (NDD) analysis, and surrogate model (SM) construction. Results show that the proposed method can effectively and efficiently generate the testing scenario library, which can accelerate the evaluation process by a few magnitudes compared with the DNN evaluation method, with the same accuracy.

To the best of our knowledge, this is the first study that identifies the entire TSLG problem and solves it systematically for both different dimensions of scenarios, different performance metrics, and CAV models. It provides guidelines in generating testing scenario libraries for closed testing facilities to enable accurate and efficient CAV evaluation.

5. Outputs

The following outputs were generated during the performance of this project:

- Conference Presentations: 2019 TRB Annual Meeting and 2019 Automated Vehicle Symposium
- Conference Paper: Feng, S., Sun, H., Feng, Y. *, Yu, C., Bao, S., Misra, A., Zhang, Y., and Liu H.X., 2019. Testing Scenario Library Generation for Connected and Automated Vehicle Evaluation. *Transportation Research Board 98th Annual Meeting Compendium of Papers*, Washington DC, 2019.

- Journal Paper: Feng, S., Feng, Y., Yu, C., Zhang, Y., and Liu, H.X., Testing Scenario Library Generation for Connected and Automated Vehicles, Part I: Methodology. *Submitted to IEEE Transactions on Intelligent Transportation Systems.*
- Journal Paper: Feng, S., Feng, Y., Sun, H., Bao S., Misra, A., Zhang, Y., and Liu, H.X., Testing Scenario Library Generation for Connected and Automated Vehicles, Part II: Case Studies. *Submitted to IEEE Transactions on Intelligent Transportation Systems.*

6. Impacts

The impacts from the development of a testing scenario generation framework are significant. This has the potential to save automobile manufacturers and their suppliers millions of dollars in testing by improving the testing process with a few magnitudes. With the proposed framework, the automobile manufacturers don't need to deploy real vehicles on the road to perform NDD evaluation for billions of miles to collect statistic significant result. This cost savings can be cascaded to consumers, making the cost of a CAV more affordable. In turn, this may increase the penetration of CAVs faster.

References

[1] L. Li, Y.-L. Lin, N.-N. Zheng, F.-Y. Wang, Y. Liu, D. Cao, K. Wang, and W.-L. Huang, “Artificial intelligence test: a case study of intelligent vehicles,” Artificial Intelligence Review, vol. 50, no. 3, pp. 441–465, 2018.

[2] F. M. Favar`o, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, “Examining accident reports involving autonomous vehicles in california,” PLoS one, vol. 12, no. 9, 2017.

[3] N. Kalra and S. M. Paddock, “Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?”

Transportation Research Part A: Policy and Practice, vol. 94, pp. 182– 193, 2016.

[4] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, and M. Maurer, “Defining and substantiating the terms scene, situation, and scenario for automated driving,” in 2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE, 2015, pp. 982–988.

[5] L. Li, W.-L. Huang, Y. Liu, N.-N. Zheng, and F.-Y. Wang, “Intelligence testing for autonomous vehicles: a new approach,” IEEE Transactions on Intelligent Vehicles, vol. 1, no. 2, pp. 158–166, 2016.

[6] J. Zhou and L. del Re, “Reduced complexity safety testing for adas & adf,” IFAC, vol. 50, no. 1, pp. 5985–5990, 2017.

[7] H. Hunger, “Test specifications for highly automated driving functions: Highway pilot,” Tech. Rep., 2017. [Online]. Available: https://www.pegasusprojekt.de

[8] “Automated driving systems: A vision for safety,” Tech. Rep., 2017. [Online]. Available: https://www.nhtsa.gov/

[9] “Waymo safety report: On the road to fully self-driving,” Tech. Rep., 2017. [Online]. Available: https://waymo.com/safety/

[10] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, “Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques.” IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 3, pp.

595–607, 2017.

[11] W. G. Najm, S. Toma, J. Brewer et al., “Depiction of priority lightvehicle pre-crash scenarios for safety applications based on vehicleto- vehicle communications,” United States. National Highway Traffic Safety Administration, Tech. Rep., 2013.

[12] O. Carsten, N. Merat, V. Janssen, E. Johansson, M. Fowkes, and K. Brookhuis, “Human machine interaction and safety of traffic in europe,” HASTE Final Report, vol. 3, 2005.

[13] V. Karabatsou LMS, M. Pappas LMS, P. van Elslande INRETS, K. Fouquet INRETS, and M. Stanzel Volkswagen, “A-priori evaluation of safety functions effectiveness-methodologies,” 2007.

[14] D. Jung, D. Jung, C. Jeong, Y. Kou, and H. Peng, “Worst case scenarios generation and its application on driving,” SAE Technical Paper, Tech. Rep., 2007.

[15] G. E. Mullins, P. G. Stankiewicz, R. C. Hawthorne, and S. K. Gupta, “Adaptive generation of challenging scenarios for testing and evaluation of autonomous vehicles,” Journal of Systems and Software, vol. 137, pp. 197–215, 2018.

[16] F. Consortium et al., “Festa handbook version 2 deliverable t6. 4 of the field operational test support action,” Brussels: European Commission, 2008.

[17] H.-H. Yang and H. Peng, “Development and evaluation of collision warning/collision avoidance algorithms using an errable driver model,” Vehicle system dynamics, vol. 48, no. S1, pp. 525–535, 2010.

[18] K. Lee, Longitudinal driver model and collision warning and avoidance algorithms based on human driving databases, 2004.

[19] D. Zhao, X. Huang, H. Peng, H. Lam, and D. J. LeBlanc, “Accelerated evaluation of automated vehicles in car-following maneuvers,” IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 733–744, 2018.

[20] J. M. Hammersley and D. C. Handscomb, “General principles of the Monte Carlo method,” in Monte Carlo Methods. Springer, 1964, pp.

50–75.

[21] ISO, “Road vehicles – Functional safety,” 2011.

[22] T. A. Ranney, “Models of driving behavior: a review of their evolution,” Accident Analysis & Prevention, vol. 26, no. 6, pp. 733–750, 1994.

[23] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang et al., “End to end learning for self-driving cars,” arXiv preprint arXiv:1604.07316, 2016.

[24] J. Zhang and K. Cho, “Query-efficient imitation learning for end-to-end autonomous driving,” arXiv preprint arXiv:1605.06450, 2016.

[25] L. Fraade-Blanar, B. Marjory S., A. James M., and K. Nidhi, “Measuring automated vehicle safety: Forging a framework,” Santa Monica, CA: RAND Corporation, Tech. Rep., 2018.

[26] Feng, S., Feng, Y., Yu, C., Zhang, Y., and Liu, H.X., Testing Scenario Library Generation for Connected and Automated Vehicles, Part I: Methodology. https://arxiv.org/abs/1905.03419

[27] E.-M. Nosal, “Flood-fill algorithms used for passive acoustic detection and tracking,” in New Trends for Environmental Monitoring Using

Passive Systems, 2008. IEEE, 2008, pp. 1–5.

[28] M. Kalisiak and M. van de Panne, “Rrt-blossom: Rrt with a local floodfill behavior.” in ICRA, 2006, pp. 1237–1242.

[29] D. Ackley, A connectionist machine for genetic hillclimbing. Springer Science & Business Media, 2012, vol. 28.

[30] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. Cambridge, MA: MIT Press, 2011.

[31] Y. Feng, C. Yu, S. Xu, H. X. Liu, and H. Peng, “An augmented reality environment for connected and automated vehicle testing and evaluation,” in 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 1549–1554.

[32] L. Wasserman, All of statistics: a concise course in statistical inference. Springer Science & Business Media, 2013.

[33] S. M. Ross, Introductory statistics. Academic Press, 2017.

[34] K. Vogel, “A comparison of headway and time to collision as safety indicators,” Accident analysis & prevention, vol. 35, no. 3, pp. 427– 433, 2003.

[35] R. Chen, R. Sherony, and H. C. Gabler, “Comparison of time to collision and enhanced time to collision at brake application during normal driving,” SAE Technical Paper, Tech. Rep., 2016.

[36] Feng, S., Feng, Y., Sun, H., Bao S., Misra, A., Zhang, Y., and Liu, H.X., Testing Scenario Library Generation for Connected and Automated Vehicles, Part II: Case Studies. https://arxiv.org/abs/1905.03428

[37] D. Bezzina and J. Sayer, “Safety pilot model deployment: Test conductor team report,” Report No. DOT HS, vol. 812, p. 171, 2014.

[38] X. Gong, Y. Guo, Y. Feng, J. Sun, and D. Zhao, “Evaluation of the energy efficiency in a mixed traffic with automated vehicles and human controlled vehicles,” arXiv preprint arXiv:1806.00377, 2018.

[39] J. W. Ro, P. S. Roop, A. Malik, and P. Ranjitkar, “A formal approach for modeling and simulation of human car-following behavior,” IEEE

Transactions on Intelligent Transportation Systems, vol. 19, no. 2, pp. 639–648, 2018.

[40] J. R. Sayer, S. E. Bogard, M. L. Buonarosa, D. J. LeBlanc, D. S. Funkhouser, S. Bao, A. D. Blankespoor, and C. B. Winkler, “Integrated vehicle-based safety systems light-vehicle field operational test key findings report,” 2011.

[41] F. Feng, S. Bao, J. R. Sayer, C. Flannagan, M. Manser, and R. Wunderlich, “Can vehicle longitudinal jerk be used to identify aggressive drivers? an examination using naturalistic driving data,” Accident Analysis & Prevention, vol. 104, pp. 125–136, 2017.

[42] A. Kesting, M. Treiber, and D. Helbing, “General lane-changing model mobil for car-following models,” Transportation Research Record, vol. 1999, no. 1, pp. 86–94, 2007.

[43] J. B. Rawlings and D. Q. Mayne, Model predictive control: Theory and design. Nob Hill Pub. Madison, Wisconsin, 2009.

[44] J. Nilsson, J. Silvlin, M. Brannstrom, E. Coelingh, and J. Fredriksson, “If, when, and how to perform lane change maneuvers on highways,” IEEE Intelligent Transportation Systems Magazine, vol. 8, no. 4, pp. 68–78, 2016.