Does Road Diversity Really Matter in Testing Automated Driving Systems?

A Registered Report

Stefan Klikovits
klikovits@nii.ac.jp
National Institute of Informatics
Tokyo, Japan

Vincenzo Riccio
vincenzo.riccio@usi.ch
Università della Svizzera Italiana (USI)
Lugano, Switzerland

Ezequiel Castellano
ecastellano@nii.ac.jp
National Institute of Informatics
Tokyo, Japan

Ahmet Cetinkaya
ahmet@shibaura-it.ac.jp
Shibaura Institute of Technology
Tokyo, Japan

Alessio Gambi
alessio.gambi@fh-krems.ac.at
IMC University of Applied Science
Krems, Austria

Paolo Arcaini
arcaini@nii.ac.jp
National Institute of Informatics
Tokyo, Japan

Abstract

Background/Context. The use of automated driving systems (ADSs) in the real world requires rigorous testing to ensure safety. To increase trust, ADSs should be tested on a large set of diverse road scenarios. Literature suggests that if a vehicle is driven along a set of geometrically diverse roads—measured using various diversity measures (DMs)—it will react in a wide range of behaviours, thereby increasing the chances of observing failures (if any), or strengthening the confidence in its safety, if no failures are observed. To the best of our knowledge, however, this assumption has never been tested before, nor have road DMs been assessed for their properties.
Objective/Aim. Our goal is to perform an exploratory study on 47 currently used and new, potentially promising road DMs. Specifically, our research questions look into the road DMs themselves, to analyse their properties (e.g. *monotonicity*, *computation efficiency*), and to test correlation between DMs. Furthermore, we look at the use of road DMs to investigate whether the assumption that diverse test suites of roads expose diverse driving behaviour holds.
Method. Our empirical analysis relies on a state-of-the-art, open-source ADSs testing infrastructure and uses a data set containing over 97,000 individual road geometries and matching simulation data that were collected using two driving agents. By sampling random test suites of various sizes and measuring their roads' geometric diversity, we study road DMs properties, the correlation between road DMs, and the correlation between road DMs and the observed behaviour.

Keywords

road diversity measure, autonomous driving systems, behaviour diversity, testing

ACM Reference Format:
Stefan Klikovits, Vincenzo Riccio, Ezequiel Castellano, Ahmet Cetinkaya, Alessio Gambi, and Paolo Arcaini. 2022. Does Road Diversity Really Matter in Testing Automated Driving Systems?: A Registered Report. In *Proceedings of 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'22).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ESEM'22, September 18–23, 2022, Helsinki, Finland
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Automated driving systems (ADSs) are expected to drastically change the transportation industry by reducing the number of accidents, avoiding traffic congestion, and lowering fuel consumption. Nonetheless, reports of collisions involving ADSs [14, 42] and fatalities [17] emphasise the need for extensive validation and testing of the technology before they can be safely released onto public roads.

Thorough testing explores, i.e. *covers*, different aspects of the system under test (SUT)'s behaviour, and thus has the potential to find bugs and increase the confidence in the SUT's correctness [22]. Due to the complexity of the SUT, however, it is generally not possible to directly find test inputs that maximise behavioural diversity (BD). Therefore, to increase testing cost-effectiveness, existing research (e.g. [6, 10, 20]) proposed to generate test suites that maximise *test diversity*. The underlying assumption of generating diverse tests is that the more diverse the tests within a test suite (TS) are, the higher the SUT's BD will be, and thus, more of its functionality will be exercised [50]. Consequently, diversity-driven approaches such as Novelty search [27] have been applied to generate tests [19, 54].

Likewise, in the ADS domain many testing approaches aim to generate test suites of driving scenarios that feature various combinations of road structure (e.g. road geometry, lane markings), traffic participants (e.g. vehicles, pedestrians), and other environmental factors (e.g. weather, lighting). Arguably, roads are the most fundamental aspect of a driving scenario, as other aspects will typically be expressed in reference to their geometry (e.g. "overtaking on a straight/curvy road"). Thus, the goal must be to create a TS of diverse road geometries, to test as many vehicle behaviours as possible. This TS should ideally cover a range of different straights, bends, curves, and turns of various degrees of sharpness. The testing of ADSs is typically coverage-based, aiming to maximise the number of tested road shapes.[1] An adequate *diversity measure (DM)* should therefore reflect the prioritisation of variety of road geometries. Example 1 illustrates this concept over a set of simple roads. This interpretation of diversity, which is rooted in Software Testing, is different than the one from other disciplines. For instance, in Machine Learning, diversity in training data refers to "equality of distributions", which is required to avoid biased datasets. Such a

[1]Note that this research investigates road diversity. Test suite optimisation such as minimality or execution efficiency are orthogonal and considered future work.

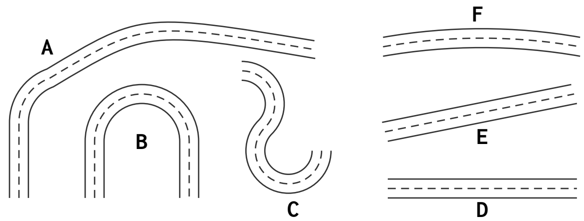

Figure 1: Different road geometries of varying diversity.

Example 1. The figure above displays a set of six road geometries, where roads **A**, **B**, **C** and **D** will very likely provoke very different driving behaviours, due to the differing curvatures and sharpnesses of turns. Road **E**, on the other hand, is a minimally rotated version of **D**, which therefore should not change a purely curvature-based DM when added to the TS. **F**, on the other hand, is similar but slightly bent. Even though its average curvature is comparable to **D** and **E**, it will most likely induce different ADS driving behaviour ("a smooth, long, continuous turn"). An adequately sensitive DM should therefore reflect the addition of **F** to the TS.

view is also common in some biological and natural settings [28], e.g. animal and plant population measurements.

Existing work proposed several DMs to express the diversity of roads, including measures based on *Jaccard similarity index* [24], *Iterative Levenshtein distance* [37], *discrete Fréchet distance* [33], and various other road features (e.g. *curvature*, *complexity*, and *direction coverage*) [38, 54].

To the best of our knowledge, the choice of DM has only been reported in published papers, but never justified. Furthermore, the underlying concept of road diversity measures seems to have never been studied before in the context of ADS testing. This raises the paramount questions of "*how much road geometries matter in testing ADSs*?" and "*how to measure the diversity of a set of roads?*" and motivates our research in finding a set of DMs that can be confidently used as objective measures of road diversity. Our research aims to answer the fundamental questions: *"Which DMs are well-suited for use in ADS testing?"* and *"is there any overlap (correlation) between the individual measures?"*. To answer these questions, we analyse the DMs' properties and check if they are indeed good indicators for vehicle behavioural diversity.

Desired properties of diversity measures. We believe that, in order to be useful for testing, a DM should guarantee certain properties which provide guidance towards the achievement of testing goals akin traditional coverage criteria, such as code coverage, which guide automatic test generation [21]. We expect adequate road DMs to have the following properties.

- *Monotonicity and growth* [47]: A test suite's diversity should never decrease when adding roads. Monotonicity is easily achieved by classical coverage criteria in which the test requirements to cover are fixed and known in advance. On the contrary, in domains like ADS, in which test requirements might not be known in advance, ensuring monotonicity is far from trivial. In addition to monotonicity, another aspect that is interesting to assess is *how* a DM grows as the number of test cases included in a TS increases. Indeed, a DM that grows for each individual test is able to discriminate better among roads than a DM that, although monotonic, increases only for some specific tests. Being able to discriminate better among roads is desirable, as it can possibly lead to test the driving agent in different conditions.
- *Insensitivity to duplicates* [45]: Additionally, the DM should be insensitive to duplicates (also known as "twin property" [45]) in the test suite, i.e. the DM value should neither increase nor decrease[2] if the same road is added twice. Note that this property contradicts *minimality of the test suite*, which is another desired property in testing; however, this is an orthogonal concern that should be treated independently and should not be accounted for by a DM.
- *Efficiency*: A DM should be also efficient to compute, i.e. computing a test suite's DM should take much less time than executing the test suite. While computing the coverage of classical coverage criteria is usually fast (as it consists of checking the coverage of each test requirement individually), computing a road DM could be expensive, as it may require pairwise comparison of all roads.
- *Additivity*: Where computational complexity is inevitable, a minimal property that should be guaranteed is that the DM computation should be additive, i.e. any new road added to the test suite should not require the re-evaluation of the diversity of all the roads, but only of the currently added road.

Correlation with behavioural diversity. In software testing, one of the most desired properties of structural coverage criteria is the ability to expose different behaviours of the SUT [22]. This also holds in our context, where higher DM values should effectively correlate with different observed behaviours. If a DM is *insensitive*, it would consider roads that actually trigger different types of ADS behaviours as similar: if this is the case, by relying on the coverage of the roads alone, some of those behaviours would not be triggered. On the other hand, a road DM that is *too sensitive* would flag similar roads as different, so wrongly expecting them to trigger different ADS behaviours. Additionally, such a DM would lead users to build unnecessary large test suites that do not increase BD.

Planned study. With the exploratory study proposed in this report, we aim to fill the lack of road DM research and investigate their effectiveness in ensuring an ADS's behaviour coverage. We therefore analyse the properties of a total of 47 DMs (see Section 3.2) that either have been used for measuring road diversity in previous research on testing, or have been used to measure diversity in related domains, such as measuring geometric diversity of lines and curves [1], or diversity in populations [45]. We perform our evaluation empirically on an extensive data set of roads, as mathematical analyses are not always possible (or are highly complex). This empirical study also provides quantitative estimates of effect sizes (e.g. value growth and efficiency) using real data, which are difficult to obtain through purely analytical methods.

2 Background and Related Work

The notion of diversity has been considered in various forms for the testing of ADSs. Here, we first introduce these different methods in Section 2.1. Then, in Section 2.2, we provide an overview of DMs that can be used to quantify the diversity of road geometries in a test suite.

[2] Absence of decrease is already guaranteed by monotonicity.

Table 1: Diversity Measures - Outline

Distance functions	Aggregation Methods	Direct DMs
Discrete Fréchet Distance	Weitzman Aggregation	Test Set Diameter
Partial Curve Mapping	Distance Entropy	Convex Hull of Curves
Dynamic Time Warping	Summing	
Normalised Relative Angle	Averaging	
Complexity Vectors	Averaging of Maxima	
Iterative Levenshtein Distance		
Jaccard Similarity Index		
Area Between Curves		
Manhattan Distance		

2.1 Diversity in ADS Testing

Several works on ADS testing aim to generate driving scenarios cost-effectively. Most of these works aim to achieve driving scenarios' diversity to avoid generating too similar scenarios that might expose the same issues multiple times.

Abdessalem et al. [3], Tuncali et al. [41], Zhu et al. [52], Majumdar et al. [30], and Zhong et al. [51] define diversity in terms of differences in the scenario configuration space. The scenario configuration space includes various parameters such as the initial position and speed of vehicles and pedestrians, the road layout, the placement of scenery elements and obstacles, as well as weather and lighting conditions. While Abdessalem et al. and Tuncali et al. did not measure how much scenarios differ, Zhu et al. used Euclidean distance to assess how different the scenarios are in the configuration space. Majumdar et al., instead, defined diversity based on the overall dispersion of the scenario parameters, whereas Zhong et al. proposed to discriminate diverse tests only if they are at a certain distance in the configuration space.

Riccio and Tonella [37] designed an Iterative Levenshtein distance to evaluate test diversity. However, differently from the work mentioned above, they considered only geometric properties of roads, i.e. the road shape. Other notable works that considered only road properties as test diversity discriminant are the ones by Zohdinasab et al. [54], Nguyen et al. [36], Gambi et al. [23][24], and Tang et al. [40]. In particular, Zohdinasab et al. computed high-level road features, like smoothness or complexity, and represented the tests into bi-dimensional feature maps such that roads mapped to the same map cells are considered similar. Notably, Zohdinasab et al. also used behavioural features to identify tests that expose similar behaviour of the ADS. This map representation has been later used for test selection [36], and test adequacy assessment [23]. Gambi et al. and Tang et al., instead, discriminated tests based on whether they take place on roads and intersections made of similar road segments.

Most of the existing approaches that take road geometry into account measure test diversity by aggregating similarity metrics computed over pairs of roads. The next section provides the necessary details on road diversity metrics and aggregation functions.

2.2 Diversity Measures for Road Geometry

In the literature, several methods have been described for computing DMs for roads. Furthermore, we also draw from other domains such as geometry and population diversity, to obtain potentially viable DM methods. We categorise these methods as (1) methods that can be computed by *aggregating pairwise distances* between roads, and (2) methods that can be *computed directly* by using the information of each individual road in a test suite.

Table 1 lists the approaches in each category. Diversity computation methods that belong to the first category are formed as a pair of a distance function and an aggregation method.

In the following, we overview existing distance functions and then explain various methods that can be used to aggregate the distances between road pairs to compute the test suite diversity.

2.2.1 Pairwise Distance Functions To the best of our knowledge, the following functions are commonly used for measuring dissimilarities between roads, (typically represented as curves):

Fréchet and discrete Fréchet distances. Fréchet [2] and discrete Fréchet [1, 33] distance functions are commonly used to measure distance between curves. Intuitively, when two vehicles move along two roads, the *maximum point-to-point distance* between them depends on the speed that they move. Fréchet distance corresponds to the shortest of all possible maximum point-to-point distances that can be obtained by varying the speeds. Discrete Fréchet distance provides an efficient way of approximating Fréchet distance for the case where the curves are described as sequences of straight segments. These distances have been used for characterising road dissimilarity [8] and the dissimilarity of the paths vehicles can take [18, 44].

Partial curve mapping distance. Partial curve mapping distance between curves representing two roads is defined as the sum of the discrepancy between the segments of two polygons representing normalised versions of the curves [48].

Dynamic time warping distance. Dynamic time warping was originally proposed for computing dissimilarity between temporal sequences where entries correspond to data obtained at consecutive time instants [4], but it has also been commonly used for checking dissimilarity between curves [16, 35].

Relative angle and normalised relative angle distances. Relative angles and normalised relative angles between curves as defined in [43] provide useful methods for computing pairwise distance measures that do not change when the curve representing one of the roads is rotated. This rotation invariance property is especially useful when testing trajectory planners [31, 46, 53] utilised in ADSs, as they depend on the sharpness of the turns on a road but not the entire orientation of the road (e.g. whether the road goes north or east).

Distance based on complexity vectors. In [38], roads were identified as sequences of smaller sections called frames. Then, curvatures and curvature-derivatives were used to compute the so-called complexity vectors for each frame. Given a pair of roads with multiple frames, the distance function of [38] finds the maximum of the distances between the complexity-wise closest frames of the roads.

Iterative Levenshtein distance. Curves representing roads can be described as lists of connected straight segments. The rotational differences between the orientation of consecutive segments form sequences of angles. Iterative Levenshtein distance for a pair of roads is defined in [37] as the Levenshtein *edit distance* [29] between the sequences of angles identified for each of the roads.

Jaccard similarity index. Jaccard similarity index (JS) for a pair of roads is defined in [24] by considering the sets of segments in those roads. JS takes a value between 0 and 1 corresponding to the ratio of the number of common segments of the roads to the total number of segments in the union of segment sets. Thus, the value 1 − JS can be used as a distance function quantifying the dissimilarity between a pair of roads.

Area between curves. The area between curves is a geometric measure that, casually speaking, measures "how much space fits" between two curves [26]. This measure allows fast computation of the pairwise distance between road curves.

Manhattan distance over feature vectors. Zohdinasab et al. [54] used Manhattan distance between feature vectors extracted from two roads as a measure of dissimilarity between them.

2.2.2 Aggregation Methods After a distance function is used on each pair of roads in a test suite, an aggregation method can be used for combining the distances to obtain a single value representing the DM. The properties of such a DM thus depend not only on the pairwise distance function but also on the method of aggregation. We report below aggregation methods commonly used to measure diversity:

Weitzman aggregation. Weitzman [45] proposed an aggregation method by considering ideal properties of the relationship between diversity measures and underlying pairwise distance functions. The computation of the proposed aggregation method involves set-based recursions, and is known to be slow in general.

Aggregation through summing. Distances between roads can be aggregated by taking their sum. This method is also used e.g. in biology to measure population diversity and corresponds to *functional attribute diversity* [11].

Aggregation using distance entropy. Distance entropy was proposed by Shi et al. [39] as a method for quantifying the diversity of a test suite for applications in software testing. When adapted to our problem setting, this method first constructs a weighted relationship graph of roads by using their pairwise distances. The aggregation is then achieved by computing the entropy of the weights in the minimum spanning tree of the relationship graph.

Aggregation through averaging all and averaging maximum distances. Averaging-based aggregation methods were previously used by [8] and [54]. In particular, [8] considered average of all pairwise distances between roads as an aggregation method in diversity computation. Moreover, [54] considered a more general setting where pairwise distances are defined for general test cases (not just roads). In the case of roads, the aggregation method of [54] corresponds to calculating the average of the distances from each road to the road that it is most distant to (i.e. averaging maximum distances).

2.2.3 Direct Computation DMs There are two diversity quantification methods that do not rely on distances between roads.

Test set diameter. Test set diameter was introduced by Feldt et al. [20] as a method for direct computation of DMs. It is linked to normalised compression distance for multisets [7], which uses Kolmogorov complexity of elements in a set.

Convex hull of curves. Area of the convex hull encompassing all road curves is a DM that can be computed directly without comparing roads with each other. Previously, convex hulls of curves have been used in [9] for checking if roads fit into a given map.

2.2.4 Analysis of the Relationship between Different DMs The relationship between different DMs can be characterised through correlation of the diversity values that they assign to test suites. In some special cases, correlation coefficients can be analytically derived. For instance, for the same underlying distance function, DMs obtained with aggregation through summation and averaging are perfectly correlated (with correlation coefficient 1), since one is a scaled version of the other. Analytical correlation analysis becomes more challenging when DMs use nonlinear aggregation methods or nonlinear operations in diversity calculations. This point is further discussed in the context of diversity of species in [13]. There, it is mentioned that while it is possible to show positive correlation between DMs, the exact value of the correlation coefficient is hard to derive analytically in many cases. In such cases, numerical methods are typically used (see, e.g. [34]).

Another aspect of correlation analysis of DMs is that correlation of aggregation-based DMs depends also on the relationship between the underlying distance functions. We note that for some DMs, the correlation between distance functions is preserved through aggregation. In particular, the correlation coefficient of two DMs defined by summing all pairwise distances obtained respectively with two different distance functions is equivalent to the correlation coefficient of the distance functions themselves. However, the correlation coefficient is hard to obtain analytically, since there is no straightforward transformation between distance functions for curves. In [26], some of these distance functions were compared through an empirical study on an optimisation problem.

3 Research Design

Given the importance of DMs in the testing of ADSs, an objective comparison of the methods is of vital interest. Specifically, we are interested in the properties of DMs, as well as whether and how strongly they are correlated. Additionally, we also would like to investigate the relationship between (road) DMs and behavioural diversity (BD) of the vehicle. To this extent, we select 47 DMs (see Section 3.2) and perform the—to the best of our knowledge—first exploratory study of DMs for road geometry.

3.1 Research Questions

We specifically investigate the following research questions (RQs):

RQ1 Which DMs guarantee beneficial properties such as *monotonicity*, *insensitivity to duplicates*, *efficiency*, and *additivity*? We expect that for a "good" DM, the diversity of a TS cannot decrease by adding more roads and remains constant when containing road duplicates. Furthermore, a DM should be efficient to calculate and extend (i.e. not require complete recalculation). In this RQ, we check the DMs individually for these four properties.

RQ2 Are DMs correlated among each other (i.e. pairwise)? Given that certain DMs measure similar (e.g. geometric) properties, we suspect that some of them might be correlated. This information is of interest, as the calculation of strongly correlated DMs might be redundant. In this RQ, we test our assumption by checking which DMs are correlated, and how strong their correlation is.

RQ3 What is the effect of road length on DMs? Conceptually, long roads could be seen as compositions of shorter road segments (e.g. turns and straights). Thus, certain DMs might "average out" or mask specific distinguishing features of the roads, such as e.g. when averaging a road with a short sharp turn after a long straight. On the other hand, geometric DMs might naturally favour longer roads. For instance, the *area between curves* of a slightly left bent and a slightly right bent road increases with the length of the roads. As this information is of practical interest to developers, who have to be aware of such properties, in this RQ, we study whether any DMs are correlated with the length of the road.

RQ4 Do the DM values correlate with the simulations' behavioural diversity? This RQ specifically investigates the main assumption that road diversity can be used as a proxy for BD.

- ***RQ4.a*** Is there a clear correlation between DMs and the BD that is calculated from vehicle observations (i.e. acceleration, brake, velocity, steering input, lateral position)?
- ***RQ4.b*** Does low (resp. high) road diversity imply low (resp. high) BD? While ***RQ4.a*** investigates general correlation, here we focus specifically on those TSs with low (resp. high) DM values and their correlation to BD. By intuition, one might suspect that while *generally high road diversity does not guarantee high BD, low road diversity certainly implies low BD.* This RQ will specifically evaluate such a correlation.
- ***RQ4.c*** Do TSs that exert low (resp. high) BD have similar DMs? This RQ can be thought of as "the inverse" of ***RQ4.b***. We focus specifically on TSs that yield low (resp. high) BD values and correlate them with the DMs of the TSs that exert them.
- ***RQ4.d*** What is the impact of road length on BD? Intuitively, we might suspect that BD is masked on longer roads (similar to the effect described in ***RQ3***), and hence, that shorter roads are preferable. In this RQ we analyse the impact of road length on correlation strength between DMs and BD.

3.2 Research Subjects

The subject of our research are the 47 diversity measures described in Section 2. Specifically, we will analyse all combinations of the nine pairwise distance measures and the five aggregation methods (see Table 1). Moreover, we will also consider the two direct DMs test set diameter and convex hull of curves.

We will perform our analyses on all 47 DMs, to get a complete picture of the DM landscape. Note that some properties for certain DMs can be deduced e.g. due to the nature of their aggregation function (monotonicity is for instance not assured when using *averaging*). Nonetheless, as we additionally aim to investigate the effect size, we will calculate these properties for all DMs.

Similarly, even though one might intuitively suspect certain DMs to be strongly correlated (***RQ2***), our goal is to test this hypothesis in a practical setting. The information on which and how strongly the DMs are correlated is of special interest, since it allows future users to avoid computing redundant DMs, i.e. those that produce very similar (if not equivalent) results.

3.3 Road Data Set

In the `SBST'2022 Tool Competition` [23], competitors provide search algorithms that generate non-intersecting two-lane roads which an autonomous driving agent should follow. Each generated road is simulated and provides timestamped observation records of the vehicle's *position*, *velocity*, *steering angle*, *brake* and *throttle* inputs. The roads—cubic interpolations of the Cartesian control points provided by the search algorithms—are handed to a simulator and allow calculation of *length*, *curvature*, *road heading*, *turn count* and aggregations thereof (e.g. *max curvature*), etc. Based on this data, we can extract further information such as *relative heading* w.r.t. the road, *total driven length*, *lateral vehicle position*, as well as *aggregate* values such as minimum, mean, maximum and standard deviation of steering input, as suggested by [25].

We use a large data set of 97,000 produced in the course of the competition as the basis for our research on road diversity. To generalise our data, we use roads generated randomly, as well as by three road generators, namely `AmbieGen`, `WOGAN` and `Frenetic` (see [23]), and simulation data provided by executing these roads using two autonomous driving agents (see Section 3.4).

Data Quality. A preliminary look into the data showed that both the road data and the simulation information is of high quality. The created roads have been analysed for self-intersections and maximum curvature by the `SBST` pipeline at creation time. Furthermore, we also checked that the simulation data is complete, i.e. the observation data for vehicle simulation (position, velocity, acceleration, steering/throttle/brake inputs, etc.) are available for every record, and the recordings have high enough frequency (e.g. between 5Hz and 20Hz). Nonetheless, we observed that a limited number of simulations have potentially invalid data. For instance, we discovered a small number of "faulty" simulations where the vehicle started driving in the wrong direction. We will thoroughly analyse the data and remove such simulations before running our experiments analysis on the data. To this extent, we will implement automatic ways to identify and remove these individuals from the data set, and we will also manually check the test samples. Furthermore, next to the road and simulation data itself, the `SBST` pipeline also produced and recorded some statistical information (driving direction coverage, road curvature measures) for each simulation. We will adapt our scripts to reproduce this data using our own (independent) implementation, thereby increasing confidence in our code and correctness of the existing data.

3.4 Driving Agents

To increase the generalisability of our results, we use driving simulations produced by two independent autonomous driving agents (DAs), i.e. `BeamNG.AI` and `Dave2`. Both considered agents are widely used in the literature [23, 24, 37, 54] and automatically perform the lane keeping task. They are, however, conceptually different (rule-based vs deep learning-based) and, thus, show different driving behaviour.

`BeamNG.AI` is the DA shipped with the `BeamNG.tech`[3] driving simulator. It defines the ego-car's trajectory before the simulation by leveraging a perfect knowledge of the road's geometry. In particular, `BeamNG.AI` plans a trajectory that maximises the car's speed (within

[3] https://beamng.tech/

pre-defined speed limits), while keeping the vehicle within the right lane as much as possible.

Dave2 exploits a deep learning architecture consisting of three convolutional and five fully-connected layers [5]. In particular, it learns a direct mapping from the on-board sensor camera input to the steering angle value to be passed to the ego-car's actuators. This means that it does not require previous knowledge of the road geometry.

In our study, we will perform all our analyses that take driving behaviour into account (i.e. ***RQ4***) w.r.t. each specific agent.

4 Execution Plan

In the following, we describe our specific research plan and provide the detailed protocol for the experiments for each RQ in Sections 4.1–4.4. Furthermore, shared experimental settings are presented before.

Test suites. The analysis of both road and behavioural diversity is based on the calculation of metrics for TSs. A TS is a set of randomly selected roads of fixed size. Our TSs are sampled with varying sizes of 10, 20, 50 and 100 roads. To generalise our results, we aim to perform each computation on 100 TS of each size[4].

Behavioural diversity. The metric for BD aims to express how much of a vehicle's behavioural range is covered within a TS. In this work, we say that the behaviour of a vehicle on a road is defined by the vehicle's *velocity*, *acceleration*, *braking*, *steering* input and *lateral position* on the road. Each simulation of a driving agent on a road produces records of these values in regular intervals (roughly every 0.1 seconds). Through aggregation, we calculate the *mean*, *minimum*, *maximum* and *standard deviation* values of each of the five observations as reported in [25], yielding a total of 20 values. For the comparison of BD of two road simulations, we calculate these 20 values for both roads and compute the (20-dimensional) normalised Euclidean distance. To expand our measurements from pairwise to set-of-roads metrics, we aggregate using the entropy approaches used for output diversity, as described in [32].

Driving agent. While the results of ***RQ1–3*** are independent of the specific DAs, for the analysis of ***RQ4*** we need to take the difference between DAs into account. Thus, we will separately analyse the correlation of DMs and BD for each DA.

Rotational adjustment. When thinking of road geometry, we typically only think of the shape of the road itself, but do not take its position and orientation into account. Thus, a perfectly straight road leading north will be seen as equivalent to a straight road of the same length leading east or west. To avoid being misled, we should therefore move all roads to the same starting location and also align them, before calculating the DMs. Nonetheless, some DAs such as Dave2 (which was trained on image data), take positioning of the sun and ego's own shadow into account for their behaviour. Thus, arguably, roads with the same geometry but different heading should be distinguished when analysing DMs for Dave2.

We will therefore duplicate our analyses for the TSs of Dave2 and report any correlation w.r.t. rotated as well as non-rotated roads. For the rotation, we will therefore do *data preprocessing* in the form of a *Procrustes analysis* [8, 15], leading to two transformations. First, the roads are relocated so that their starting points match. Second, the roads are rotated around the initial point with the angle of rotation obtained through an optimisation procedure [15]. These steps ensure that two roads with identical shapes yield zero distance.

[4]Even though we will use powerful computation infrastructure, certain DMs (e.g. those using Weitzman aggregation) are computationally (very) expensive for large TSs. We will therefore set a maximum computation time and adjust our analysis accordingly.

4.1 RQ1 – DM Properties

Monotonicity and growth. For some DMs and aggregation methods such as convex hull of curves, Weitzman aggregation and aggregation through summation, monotonicity can be mathematically proved. For these DMs and aggregation methods, we can simply report the theoretical results known from the literature. For other DMs and aggregation methods such as test set diameter and aggregation using distance entropy, instead, assessing monotonicity is more challenging as no theoretical results are known from the literature. For this latter category, we perform an empirical evaluation.

Moreover, in case a measure is monotonic, we are also interested in assessing "how and how much" it grows; specifically, we are interested in assessing to what extent each new test increases TS diversity. Indeed, measures for which each novel (non-duplicated) test increases the DM value are better at discriminating among different tests than DMs for which only a few, highly diverse tests lead to a noticeable change of the DM value; better discrimination among tests is desirable, as it can lead to exercise different ADS behaviours (think of adding road **F** in Example 1). Note that in the case of DMs via aggregation, diversity largely depends on the underlying complex distance functions, and, as a result, a precise mathematical assessment of the growth is hard to achieve. This is the case even for measures for which we can analytically check monotonicity (e.g. Weitzman aggregation [45]). Therefore, to quantify the growth, we perform an empirical evaluation. Namely, we evaluate the relative DM growth when adding new roads to the TS. Intuitively, smaller TSs should, on average, have a larger growth in diversity, due to the larger probability of adding a highly diverse road. Nonetheless, even large test suites' DM values should reflect the addition of individuals. We proceed as follows: (1) For each of the TSs (grouped by size), we compute the DMs. (2) We add n new roads to each TS, where n corresponds to $n \in [10\%, 20\%, 50\%, 100\%]$ of the TS's size and re-compute the DMs. (3) We then analyse the difference before and after extension, to report typical statistical metrics (mean, min, max, standard deviation) for each DM and TS size.

Insensitivity to duplicates. Similarly to monotonicity, for some DMs, this property can be determined directly from the DM definition whereas for other DMs an empirical approach is more suitable. In particular, mathematical properties of the distance functions can be used to show that DMs obtained through Weitzman aggregation and aggregation through summation are insensitive to duplication, but DMs obtained through averaging-based aggregation methods do not possess the insensitivity property, as duplication can decrease diversity for those DMs. Moreover, in case insensitivity to duplicates is not guaranteed, it is not possible to mathematically establish to what extent a DM is sensitive to duplication; therefore, we plan to assess this effect by means of an empirical study. We proceed as follows: (1) We calculate the TSs' DMs. (2) We then randomly duplicate n roads in each TSs, where n corresponds to

$n \in [10\%, 20\%]$ of the TS's size and re-compute the DMs. (3) We then check if any DM's value changed and analyse the difference in magnitude before and after the extension, using standard descriptive statistics (mean, min, max, standard deviation).

Efficiency. Although asymptotic efficiency of a DM can be obtained analytically, it cannot predict concrete results on real data. Therefore, we conduct an empirical assessment to precisely compare the different DMs and provide an initial evaluation of the DMs' computational efficiency. We proceed as follows. For each TS, we record the time it takes to compute each DM. Due to the total of 100 TSs of each size, we then confidently analyse and compare the DM computation times by size, listing their aggregated information (mean, median, standard deviation).

Additivity. Additivity measures the time it takes to re-calculate a TS's DM after being extended. As for efficiency, only an asymptotic analysis of additivity is possible starting from the DMs definition; in order to have a more realistic comparison of the different DMs, we conduct an empirical study as follows. (1) First we calculate each TSs' DMs. (2) We add n new roads to each TS, where n corresponds to $n \in [1 \text{ road}, 2 \text{ roads}, 10\%, 20\%]$ (Note the addition of 1 and 2 roads). (3) We re-calculate the DMs and record the time, before (4) we analyse the results.

4.2 RQ2 – Pairwise Correlation of DMs

Using the DMs calculated for each TS, we perform correlation analyses on the totality of the data and grouped by TS size. As discussed in Section 2.2.4, analytical derivation of correlation coefficients for certain pairs of DMs is prohibitively complex. For this reason, we perform an empirical evaluation of correlations. We select the correlation test to use depending on the analysed data. If the data is normally distributed, we use the Pearson's correlation test. Instead, if the data is not normally distributed, we use the Spearman's correlation test. Both tests produce as output a number in $[-1, 1]$, where 0 means that there is no correlation, while -1 and 1 indicate perfect negative and positive correlation. Intermediate values indicate different degrees of correlation, and can be interpreted using existing classifications, such as slight, low, moderate, high, and very high correlation [12]. When analysing correlation results between DMs, we report the corresponding classes and use these to draw conclusions. For instance, if two DMs are not or only weakly correlated, it could mean that they measure different characteristics of road geometries and that their use in combination might be more beneficial than using either DM individually. Instead, if two DMs are strongly correlated, it may hint at redundancy, and the properties of ***RQ1*** should be used to choose a better-suited one.

4.3 RQ3 – Correlation of DMs and Road Length

Here we analyse the effect of road length on the individual DMs. (1) We sample random TSs based on the length of the roads. As we are specifically interested in short (resp. long) roads, we have to adjust our TS sampling algorithm to assemble random TSs from the shortest (resp. longest) quantile of the roads. (2) We then calculate DMs for all TSs and perform an analysis using the same statistical tests described in Section 4.2, i.e. (3) we calculate DMs for test suites and (4) perform correlation analysis (as in ***RQ2***) between the DM values and the (TSs' average) road length.

The results of this analysis allow determining whether there is an effect of the road length on the diversity. Combined with the results of ***RQ4.d***, it helps users in selecting appropriate road lengths for ADS testing.

4.4 RQ4 – Correlation of DMs and BD

We analyse whether there is a correlation between the individual DMs and the BD. Unfortunately, due to the complexity of ADSs, it is typically and generally not possible to define a reliable mathematical model of an ADS's behaviour from which a correlation can be analytically deduced. Hence, we empirically study such a correlation and test the sensitivity of ADS behaviour to road diversity. This RQ aims to discover those DMs that are actually linked to BD, thus helping testers avoid using DMs that focus on irrelevant road features.

RQ4.a. First, we analyse the general correlation between road diversity (i.e. DMs) and BD. Remember that as BD is different for the individual DAs, we perform the analyses separately. Specifically, we proceed as follows: (1) For each TS, we calculate the DMs and (2) then, for each TS we calculate the BD (as described above). (3) Finally, we perform correlation analysis between these values.

RQ4.b. We analyse whether there is a specific relation between TSs with low (resp. high) diversity and the BD they expose. Therefore, for each DM we (1) first select the TSs that have low (resp. high) diversity according to the specific DM and (2) analyse the variance of the TSs' BD through correlation.

RQ4.c. Intuitively, we perform the "inverse analysis" of ***RQ4.b***. Specifically, we (1) select test suites that yielded low (resp. high) BD values in ***RQ4.a***, and (2) evaluate whether the TSs that exposed them have similarly low (resp. high) DM values.

RQ4.d. We study the effect of road length on the correlation between DMs and BD, to evaluate if for TSs with short roads, there is a stronger correlation between DMs and BD. Hence, we proceed as follows: (1) Using the similar-in-length TSs created in ***RQ3***, we calculate for each TS the values of the DMs and the BD (as in ***RQ4.a***) (2) We then perform a correlation analysis of each DM and BD pair. (3) Finally, we compare the yielded correlation coefficients and check if the correlation of DMs and BD of TSs with short (resp. long) roads is higher (resp. lower) than the one of "normal" TSs.

5 Discussion

5.1 Contributions and Impact

Our study will help in identifying a catalogue of DMs that are effective in ensuring behavioural diversity of the considered agents. Such DMs can be used for multiple tasks in the ADS testing process. Effective DMs can drive the generation of test suites that exercise diverse behaviours of the SUT, e.g. as a fitness function of a search-based test input generation algorithm. Another possible usage of the proposed DMs is the selection of a reduced number of diverse tests from an existing test suite, thus reducing the (high) cost of ADS testing.

Code availability. Evidently, the results of our analysis are specific to the implementation, the evaluation environment and data set used. In the spirit of open science, we make our analysis code publicly available, so other users and researchers can (a) reproduce the results of our study, (b) validate the algorithms and implementation of our DMs, and (c) extend this study by applying our correlation analyses to other diversity measures, behavioural metrics and data.

5.2 Threats to Validity

The validity of our study may be affected by different types of threats [49] that we discuss in the following.

Construct validity. A construct validity threat could be that the approaches we use to investigate our research questions are not appropriate. If this is the case, we may draw wrong conclusions regarding the investigated RQs; for example, we may observe correlation where there is none. To mitigate this threat, we carefully design the experiments to answer the RQs. To assess monotonicity, we will explicitly check that the diversity does not decrease when increasing the number of considered roads. For monotonic measures, we will check how they grow. We will proceed similarly for "insensitivity to duplicates". In order to check correlation, we will use either the Pearson's correlation test or the Spearman's correlation test depending on whether the data is normally distributed or not. Moreover, we will also consider the strength of the correlation when reporting the results. With regards to measures of BD, a threat could be that the ones we selected are not representative of the behavioural diversity in the ADSs that we use, or that they do not generalise. This could threaten the conclusions we draw in ***RQ4***. To address this threat, instead of coming up with our own BD measures, we looked at literature [25] to identify and select BD measures that have been used and studied before.

Internal validity. An internal validity threat could be that the observed correlation (if any) between road diversity and behaviour diversity is due to other confounding factors. For example, the behaviour of a driving agent depends not only on the road, but also on other elements of the driving scenarios such as road participants, regulatory elements, weather conditions, etc. In complex scenarios in which all these elements are present, it would very difficult (if possible at all) to isolate and measure the influence of the road structure on the driving behaviour. Therefore, there is the risk that we wrongly observe the correlation because the behaviour diversity is triggered by the different elements of the driving scenarios and not the road structure. To mitigate this threat, we use scenarios composed of only one road, without any other elements. The correlation of behaviour diversity with other traffic elements should be part of independent studies and we leave it as future work.

Other confounding factors could be the test suite size and the length of the road. Regarding test suite size, as for some metrics (i.e. those guaranteeing monotonicity) more tests in general lead to higher values, there is the risk of observing a positive correlation that, however, is only due to the test suite size. To mitigate this threat, we analyse the data by considering different test suite sizes separately, as described in Section 4. Regarding the length of the road, instead, we mitigate this threat by assessing the influence of road length on DMs (***RQ3***) and BD (***RQ4.d***).

External validity. A threat of this type is that the results of our study could be not generalisable to other types of roads, and other driving agents. To mitigate this threat, we sample from a very large set of roads created by different road generators, that can possibly generate roads of different shapes. Moreover, to assess the influence of road length on the diversity and correlation with behaviour diversity, we will artificially shorten roads to generate roads of different lengths. Different agents could react differently to different types of roads and so lead to different behaviour diversity values, and so different correlation with road diversity. To mitigate this threat, we consider two very different types of driving agent: `BeamNG.AI`, a rule-based driving agent exploiting the complete knowledge of the road geometry, and `Dave2`, a deep learning-based agent designed at NVIDIA [5].

6 Conclusions

As complete testing of a modern ADSs is impossible, current testing practices aim to expose a SUT to a suite of diverse scenarios, based on the assumption that high scenario diversity leads to a wide behavioural diversity of the SUT. In this report, we propose to conduct the first exploratory study to test this assumption. Due to the complexity of ADS scenarios, we focus our study on the arguably most fundamental part of a scenario description, namely its road geometry. In the past, a variety of road DMs has been applied in literature, but to the best of our knowledge, this use has only been reported, but never truly justified nor studied deeply.

The present report describes the research questions and execution plan for an exploratory study on 47 road geometry DMs that are either currently used in ADS testing or might be potentially beneficial to it. Specifically, our research questions analyse DMs' properties (***RQ1***), their pairwise correlation (***RQ2***), the relationship between DMs and road length (***RQ3***) and test the assumption that road diversity induces behavioural diversity (***RQ4***).

Our empirical analyses target a large data set of 97,000 individual road geometries and matching simulation data from two distinct driving agents. Based on this data, we will identify a catalogue of DMs that are effective in ensuring behavioural diversity for ADS agents and help developers effectively test their systems.

Acknowledgments

S. Klikovits, E. Castellano, and P. Arcaini are supported by ERATO HASUO Metamathematics for Systems Design Project (No. JPMJER1603), JST, Funding Reference number: 10.13039/501100009024 ERATO. S. Klikovits is also supported by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Research Activity Start-up No 20K23334. P. Arcaini is also supported by Engineerable AI Techniques for Practical Applications of High-Quality Machine Learning-based Systems Project (Grant Number JPMJMI20B8), JST-Mirai. A. Cetinkaya is supported by JSPS Grant-in-Aid for Early-Career Scientists No. 20K14771. A. Gambi was partially supported by the DFG project STUNT (DFG Grant Agreement n. FR 2955/4-1) and is also supported by the EU Project FLEXCRASH (Grant agreement ID: 101069674).

References

[1] Pankaj K Agarwal, Rinat Ben Avraham, Haim Kaplan, and Micha Sharir. 2014. Computing the discrete Fréchet distance in subquadratic time. *SIAM J. Comput.* 43, 2 (2014), 429–449.
[2] Boris Aronov, Sariel Har-Peled, Christian Knauer, Yusu Wang, and Carola Wenk. 2006. Fréchet distance for curves, revisited. In *European symposium on algorithms*.
[3] Raja Ben Abdessalem, Shiva Nejati, Lionel C. Briand, and Thomas Stifter. 2018. Testing Vision-based Control Systems Using Learnable Evolutionary Algorithms. In *Proc. of the 40th Int. Conference on Software Engineering (ICSE '18)*. ACM.
[4] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, Vol. 10. Seattle, WA, USA:, 359–370.

[5] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. 2016. End to End Learning for Self-Driving Cars. *CoRR* abs/1604.07316 (2016).

[6] Paulo MS Bueno, Mario Jino, and W Eric Wong. 2014. Diversity oriented test data generation using metaheuristic search techniques. *Information Sciences* 259 (2014), 490–509.

[7] Alessandro Calò, Paolo Arcaini, Shaukat Ali, Florian Hauer, and Fuyuki Ishikawa. 2020. Generating Avoidable Collision Scenarios for Testing Autonomous Driving Systems. In *IEEE 13th Int. Conf. on Software Testing, Validation and Verification (ICST)*. 375–386.

[8] Ezequiel Castellano, Ahmet Cetinkaya, and Paolo Arcaini. 2021. Analysis of Road Representations in Search-Based Testing of Autonomous Driving Systems. In *2021 IEEE 21st Int. Conf. on Software Quality, Reliability and Security (QRS)*.

[9] Ezequiel Castellano, Stefan Klikovits, Ahmet Cetinkaya, and Paolo Arcaini. 2022. FreneticV at the SBST 2022 Tool Competition. In *2022 IEEE/ACM 15th International Workshop on Search-Based Software Testing (SBST)*.

[10] Tsong Yueh Chen, Hing Leung, and IK Mak. 2004. Adaptive random testing. In *Annual Asian Computing Science Conference*. Springer, 320–329.

[11] Chun-Huo Chiu and Anne Chao. 2014. Distance-based functional diversity measures and their decomposition: A framework based on Hill numbers. *PloS one* 9, 7 (2014), e100014.

[12] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

[13] Paul A. DeBenedictis. 1973. On the correlations between certain diversity indices. *The American Naturalist* 107, 954 (1973), 295–302.

[14] DMV California. 2022. *Autonomous Vehicle Collision Reports*. State of California, Department of Motor Vehicles (DMV). Last accessed: July 23, 2022.

[15] Ian L. Dryden and Kanti V. Mardia. 2016. *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons.

[16] Alon Efrat, Quanfu Fan, and Suresh Venkatasubramanian. 2007. Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. *Journal of Mathematical Imaging and Vision* 27, 3 (2007), 203–216.

[17] @elonbachman. 2020. Tesla Deaths. Regularly updated at tesladeaths.com; version hosted on Zenodo will be updated periodically..

[18] Chenglin Fan, Jun Luo, and Binhai Zhu. 2010. Fréchet-distance on road networks. In *Int. Conf. on Computational Geometry, Graphs and Applications*. Springer.

[19] Robert Feldt and Simon Poulding. 2017. Searching for test data with feature diversity. *arXiv preprint arXiv:1709.06017* (2017).

[20] Robert Feldt, Simon Poulding, David Clark, and Shin Yoo. 2016. Test set diameter: Quantifying the diversity of sets of test cases. In *IEEE Int. Conf. on Software Testing, Verification and Validation*. IEEE, 223–233.

[21] Gordon Fraser and Andrea Arcuri. 2011. EvoSuite: Automatic Test Suite Generation for Object-Oriented Software. In *Proc. of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering (ESEC/FSE '11)*. ACM, 416–419.

[22] Gordon Fraser and José Miguel Rojas. 2019. *Software Testing*. Springer International Publishing, 123–192.

[23] Alessio Gambi, Gunel Jahangirova, Vincenzo Riccio, and Fiorella Zampetti. 2022. SBST Tool Competition 2022. In *15th IEEE/ACM International Workshop on Search-Based Software Testing, SBST 2022, Pittsburgh, PA, USA, May 9, 2022*.

[24] Alessio Gambi, Marc Mueller, and Gordon Fraser. 2019. Automatically Testing Self-Driving Cars with Search-Based Procedural Content Generation. In *Proc. of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2019)*. ACM, 318–328.

[25] Gunel Jahangirova, Andrea Stocco, and Paolo Tonella. 2021. Quality Metrics and Oracles for Autonomous Vehicles Testing. In *2021 14th IEEE Conference on Software Testing, Verification and Validation (ICST)*. 194–204.

[26] Charles F. Jekel, Gerhard Venter, Martin P. Venter, Nielen Stander, and Raphael T. Haftka. 2019. Similarity measures for identifying material parameters from hysteresis loops using inverse analysis. *International Journal of Material Forming* (may 2019). https://doi.org/10.1007/s12289-018-1421-8

[27] Joel Lehman and Kenneth O Stanley. 2011. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation* 19, 2 (2011).

[28] Tom Leinster and Christina A. Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology* 93, 3 (2012). https://doi.org/10.1890/10-2402.1

[29] Vladimir I. Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.

[30] Rupak Majumdar, Aman Mathur, Marcus Pirron, Laura Stegner, and Damien Zufferey. 2021. Paracosm: A Test Framework for Autonomous Driving Simulations. In *Fundamental Approaches to Software Engineering*, Esther Guerra and Mariëlle Stoelinga (Eds.). Springer International Publishing, 172–195.

[31] Matthew McNaughton, Chris Urmson, John M Dolan, and Jin-Woo Lee. 2011. Motion planning for autonomous driving with a conformal spatiotemporal lattice. In *2011 IEEE Int. Conf. on Robotics and Automation*. IEEE, 4889–4895.

[32] Hector D. Menendez, Michele Boreale, Daniele Gorla, and David Clark. 2022. Output Sampling for Output Diversity in Automatic Unit Test Generation. *IEEE Transactions on Software Engineering* 48, 1 (2022), 295–308. https://doi.org/10.1109/TSE.2020.2987377

[33] Axel Mosig and Michael Clausen. 2005. Approximately matching polygonal curves with respect to the Fréchet distance. *Computational Geometry* 30, 2 (2005).

[34] Maud A Mouchet, Sébastien Villéger, Norman WH Mason, and David Mouillot. 2010. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology* 24, 4 (2010).

[35] Mario E Munich and Pietro Perona. 1999. Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In *Proc. 7th Int. Conf. on Computer Vision*, Vol. 1. IEEE, 108–115.

[36] Vuong Nguyen, Stefan Huber, and Alessio Gambi. 2021. SALVO: Automated Generation of Diversified Tests for Self-driving Cars from Existing Maps. In *Proc. of the IEEE Int. Conf. on Artificial Intelligence Testing*. 128–135.

[37] Vincenzo Riccio and Paolo Tonella. 2020. Model-Based Exploration of the Frontier of Behaviours for Deep Learning System Testing. In *Proc. of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020)*. ACM, 876–888.

[38] Abbas Sadat, Sean Segal, Sergio Casas, James Tu, Bin Yang, Raquel Urtasun, and Ersin Yumer. 2021. Diverse complexity measures for dataset curation in self-driving. In *Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 8609–8616.

[39] Qingkai Shi, Zhenyu Chen, Chunrong Fang, Yang Feng, and Baowen Xu. 2016. Measuring the Diversity of a Test Set With Distance Entropy. *IEEE Transactions on Reliability* 65, 1 (2016), 19–27.

[40] Yun Tang, Yuan Zhou, Fenghua Wu, Yang Liu, Jun Sun, Wuling Huang, and Gang Wang. 2021. Route Coverage Testing for Autonomous Vehicles via Map Modeling. In *Proc. of the Int. Conf. on Robotics and Automation, ICRA*. IEEE, 11450–11456.

[41] Cumhur Erkan Tuncali, Georgios Fainekos, Hisahiro Ito, and James Kapinski. 2018. Sim-ATAV: Simulation-Based Adversarial Testing Framework for Autonomous Vehicles. In *Proc. of the 21st Int. Conf. on Hybrid Systems: Computation and Control (Part of CPS Week) (HSCC '18)*. ACM, 283–284.

[42] U.S. Department of Transportation, National Highway Traffic Safety Association. 2022. *Summary Report: Standing General Order on Crash Reporting for Automated Driving Systems (DOT HS 813 324)*. Technical Report.

[43] Michail Vlachos, Dimitrios Gunopulos, and Gautam Das. 2004. Rotation invariant distance measures for trajectories. In *Proc. of the 10th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*. 707–712.

[44] Jens Weise and Sanaz Mostaghim. 2021. Many-objective pathfinding based on fréchet similarity metric. In *Int. Conf. on Evolutionary Multi-Criterion Optimization*. Springer, 375–386.

[45] Martin L Weitzman. 1992. On diversity. *The Quarterly Journal of Economics* 107, 2 (1992), 363–405.

[46] Moritz Werling, Julius Zieglerand, Sören Kammel, and Sebastian Thrun. 2010. Optimal trajectory generation for dynamic street scenarios in a Frenet frame. In *Proc. IEEE Int. Conf. Robot. Autom.* 987–993.

[47] Elaine J. Weyuker. 1986. Axiomatizing Software Test Data Adequacy. *IEEE Trans. Softw. Eng.* 12, 12 (dec 1986), 1128–1138.

[48] Katharina Witowski and Nielen Stander. 2012. Parameter identification of hysteretic models using partial curve mapping. In *12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*. 5580.

[49] Claes Wohlin, Per Runeson, Martin Hst, Magnus C. Ohlsson, Bjrn Regnell, and Anders Wessln. 2012. *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated.

[50] Qing Xie and Atif M Memon. 2006. Studying the Characteristics of a "Good" GUI Test Suite. In *17th Int. Symp. on Software Reliability Engineering*.

[51] Ziyuan Zhong, Gail E. Kaiser, and Baishakhi Ray. 2021. Neural Network Guided Evolutionary Fuzzing for Finding Traffic Violations of Autonomous Vehicles. *CoRR* abs/2109.06126 (2021). arXiv:2109.06126

[52] Bing Zhu, Peixing Zhang, Jian Zhao, and Weiwen Deng. 2021. Hazardous Scenario Enhanced Generation for Automated Vehicle Testing Based on Optimization Searching Method. *IEEE Transactions on Intelligent Transportation Systems* (2021).

[53] Sheng Zhu and Bilin Aksun-Guvenc. 2020. Trajectory Planning of Autonomous Vehicles Based on Parameterized Control Optimization in Dynamic on-Road Environments. *J. Intel. Robot. Syst.* 100, 3 (2020), 1055–1067.

[54] Tahereh Zohdinasab, Vincenzo Riccio, Alessio Gambi, and Paolo Tonella. 2021. DeepHyperion: Exploring the Feature Space of Deep Learning-Based Systems through Illumination Search. In *Proc. of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2021)*. ACM, 79–90.