Testing DNN-based Autonomous Driving Systems under Critical Environmental Conditions

Zhong Li [1 2] **Minxue Pan** [1 3] **Tian Zhang** [1 2] **Xuandong Li** [1 2]

Abstract

Due to the increasing usage of Deep Neural Network (DNN) based autonomous driving systems (ADS) where erroneous or unexpected behaviours can lead to catastrophic accidents, testing such systems is of growing importance. Existing approaches often just focus on finding erroneous behaviours and have not thoroughly studied the impact of environmental conditions. In this paper, we propose to test DNN-based ADS under different environmental conditions to identify the critical ones, that is, the environmental conditions under which the ADS are more prone to errors. To tackle the problem of the space of environmental conditions being extremely large, we present a novel approach named TACTIC that employs the search-based method to identify critical environmental conditions generated by an image-to-image translation model. Large-scale experiments show that TACTIC can effectively identify critical environmental conditions and produce realistic testing images, and meanwhile, reveal more erroneous behaviours compared to existing approaches.

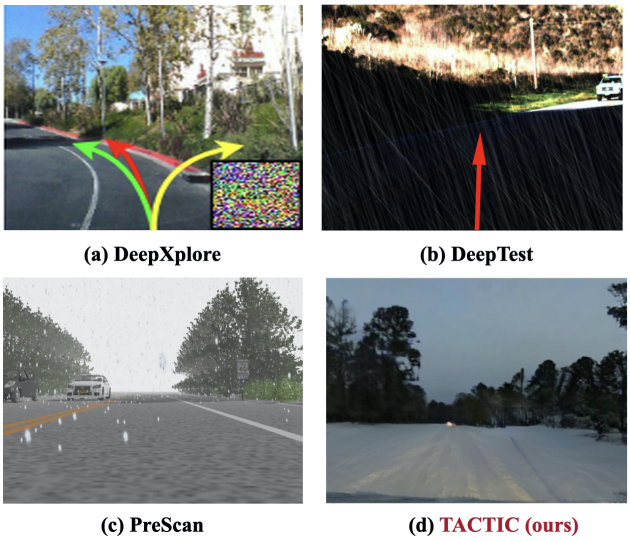


Figure 1. Testing driving scenes generated by DeepXplore (Pei et al., 2017), DeepTest (Tian et al., 2018), PreScan (Marketakis et al., 2009), and our approach TACTIC. Note that the coloured arrows in Figure 1 (a) and (b) are attached to present the predicted steering angles. TACTIC produces more realistic driving scenes compared to the other three approaches.

1. Introduction

Autonomous vehicles have achieved tremendous success over the past decade due to the significant advances in Deep Neural Networks (DNNs). We have witnessed a number of successful DNN-based autonomous driving systems (ADSs) (Grzywaczewski, 2017; Bojarski et al., 2016), where DNNs directly generate the control decisions (e.g., steering and braking) for the systems after processing inputs received from sensors (e.g., camera, Radar, Lidar). Unfortunately, such systems often demonstrate incorrect/unexpected behaviours that can lead to catastrophic accidents when deployed in real-world environments, for instance, the Tesla/Uber autonomous vehicle accidents (Boudette, 2017; Gibbs, 2018). There is an urgent need for a better approach to comprehensively and effectively testing these systems.

Recently, various approaches have been proposed for testing DNN-based ADSs which mostly rely on affine image transformation (Pei et al., 2017; Tian et al., 2018) or high-fidelity simulation (Marketakis et al., 2009; Abdessalem et al., 2016; 2018; Fremont et al., 2019) to generate independent driving scenes that can cause erroneous behaviours. However, there is little attention on studying the impact of environmental conditions (e.g., time-of-day, illumination, and weather, etc.) for DNN-based ADSs. Since DNN-based ADSs must be able to conduct proper operations under all possible environmental conditions in the real world, it is critical to understand which environmental conditions will

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China [2]Department of Computer Science and Technology, Nanjing University, Nanjing, China [3]Software Institute, Nanjing University, Nanjing, China. Correspondence to: Minxue Pan <mxp@nju.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

cause more erroneous behaviours.

In this paper, we first propose the problem of testing DNN-based ADS under different environmental conditions with the goal of identifying the critical environmental conditions under which the ADS are more prone to errors. This is a difficult problem, since it is infeasible to test ADS under the many environmental conditions in real-world. Even if only considering one environmental type, there can be a huge number of environmental conditions. For example, for the environmental type of rainy weather, there can be many different rainy conditions due to the different degrees of light and amounts of rain, etc. An alternative is to adopt synthesised test driving scenes. However, the driving scenes synthesised by most existing approaches can be unrealistic or even never happen in real-world environments. For instances, Figure 1(a)-(c) show three test scenes generated by the state-of-the-art testing techniques, i.e., DeepXplore (Pei et al., 2017), DeepTest (Tian et al., 2018), and PreScan (Marketakis et al., 2009). We can see that all these scenes lack the richness and authenticity of real-world images, and can rarely happen under real-world environments. Besides the problem of synthesised scenes lacking the authenticity of real-world scenes, it is difficult to fully specify an environmental condition by configuring the parameters for affine image transformation or high-fidelity simulation, for an environmental condition is very complex in nature. For example, one cannot exactly specify the accumulation of snow on the street or the amounts of rain for a driving scene. Even worse, the critical environmental conditions may be corner cases and only occupy a small fraction of the whole environmental condition space, making them difficult to be identified.

We address these challenges and propose TACTIC (**T**esting **A**DS under **C**ri**TI**cal **C**onditions), a novel testing framework that can identify critical environmental conditions for different environmental types and generate the corresponding driving scenes for effectively testing DNN-based ADSs. To specify the environmental conditions of an environmental type, instead of using configurations, TACTIC employs the Multimodal Unsupervised Image-to-Image Translation (MUNIT) (Huang et al., 2018) to deep-learn the features of environmental conditions from a set of sample scenes belonging to the same environmental type and encode them as a high-dimensional vector space called *style space*. A *style*, which is a vector in the style space corresponding to an environmental condition, contains extremely rich and complex environmental features. To identify the styles corresponding to the critical environmental conditions in the complex style space, TACTIC employs the search-based method with a carefully designed search objective accounting for both the number and diversity of erroneous behaviours. When styles are obtained, the style-learning process of MUNIT is reversed: styles are applied on existing test driving scenes to synthesise new ones. Because of the richness of styles, the synthesised scenes are realistic enough (as shown by Figure 1(d)) to test whether the DNN-based ADS can perform correctly under the environmental conditions corresponding to the obtained styles.

To evaluate TACTIC, we conducted a large-scale study on three well-known open-source DNN-based ADSs. The results demonstrate that TACTIC is effective in identifying critical environmental conditions for various environmental types. The driving scenes synthesised by TACTIC using the critical environmental conditions are not only more realistic, but also cause the DNN-based ADSs to produce more serious erroneous behaviours, compared to other approaches.

Overall, the main contributions of this paper are:

- To our knowledge, this is the first work that proposes to test DNN-based ADSs with the goal of identifying critical environmental conditions.
- We propose a novel approach named TACTIC, which employs the search-based method to effectively identify critical environmental conditions from the environmental condition space generated by MUNIT model.
- We perform a large-scale evaluation with TACTIC, and the results show that TACTIC is effective in revealing more serious erroneous behaviours for DNN-based ADS with the help of the obtained critical environmental conditions.

Our implementation of TACTIC and all the experimental data are publicly available[1] to facilitate future studies.

2. Related Work

Testing of DNN-based ADS Recently, there are various works emerging on testing DNN-based ADS (Pei et al., 2017; Tian et al., 2018; Zhang et al., 2018; Li et al., 2019; Huang et al., 2019; Odena et al., 2019; Zhou et al., 2020). DeepXplore (Pei et al., 2017) utilises a gradient-based differential testing techniques to detect behaviour inconsistencies among multiple DNN-based ADSs with neuron coverage guidance. As a following work of DeepXplore, DeepTest (Tian et al., 2018) further leverages neuron coverage to guide test generation for DNN-based ADS, which adopts domain-specific image transformations on driving scenes. DeepRoad (Zhang et al., 2018) proposes a GAN-based approach to generate realistic driving scenes for testing DNN-based ADS. DeepBillboard (Zhou et al., 2020) designs an algorithm to generate printable adversarial billboard that can trigger erroneous behaviours of DNN-based ADS. However, these works mainly focus on introducing erroneous behaviours to prove that the ADS under test contain

[1] https://github.com/SEG-DENSE/TACTIC

defects. They do not study the environmental conditions which may be the root causes for the erroneous behaviours. Differently, the goal of TACTIC is to identify critical environmental conditions under which the DNN-based ADS are prone to errors, and generate the corresponding driving scenes to such conditions for testing.

DNN Coverage Criteria. A variety of testing coverage criteria for DNNs have been proposed in order to guide test generation (Pei et al., 2017; Ma et al., 2018; Sun et al., 2018; Ma et al., 2019; Du et al., 2019; Odena et al., 2019). For example, DeepXplore (Pei et al., 2017) designs neuron coverage for DNNs to measure the percentage of neurons that are activated by a given set of test data. DeepGauge (Ma et al., 2018) generalises the concept of neuron coverage and proposes a set of multi-granularity coverage criteria which take the distribution of training data into consideration. TensorFuzz (Odena et al., 2019) debugs DNNs with coverage-guided fuzzing, which is to find adversarial inputs by mainly adding independent noise to individual inputs. It proposes to measure whether coverage has increased on a given test input by using an approximate nearest neighbour (ANN) algorithm to check how far away the nearest neighbour of the input is. In this work, we use two neuron-level coverage criteria selected from DeepGauge to measure the diversity of erroneous behaviours. Note that TACTIC can be easily generalised to other coverage criteria.

Metamorphic Relations for DNNs. In DNN testing, metamorphic relations (Chen et al., 2020) are widely studied to ameliorate the test oracle problem. A common metamorphic relation in testing DNN-based ADS is stated as the driving behaviours of a self-driving car should not change significantly or stay the same for the transformed images (Tian et al., 2018; Zhang et al., 2018; Han & Zhou, 2020). In this work, we leverage a metamorphic relation for steering angle control as the test oracle for detecting erroneous behaviours, and further explore its efficacy by preforming a large-scale study on three open-source DNN-based autonomous driving models and five real-world environmental types.

3. The TACTIC Approach

In this paper, we aim to test DNN-based ADS under different environmental conditions with the goal of identifying the critical environmental conditions under which the ADS are more prone to erroneous behaviours. We focus on DNN-based ADSs that perform end-to-end steering angle control, i.e., the DNN-based ADS can output the steering angles by taking the driving scenes captured by the car-mounted camera as inputs. We choose to study the steering angle control because it is a fundamental function of all kinds of ADS, and high-quality test datasets constructed and used by many existing approaches are available for experiments. TACTIC is not restricted to the steering angle control but can be extended to support the testing of other functions, e.g., car-braking, as discussed in Section 3.2.2.

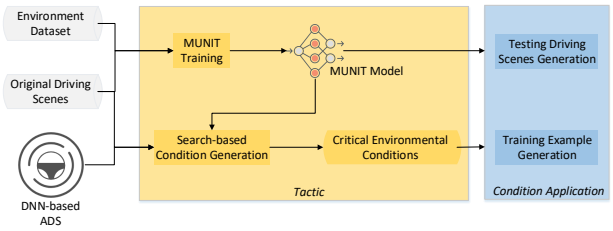


Figure 2. The overview of TACTIC. For a subject DNN-based ADS and a given environmental type, TACTIC first trains a MUNIT model to encode the environmental condition space of the environmental type. Then, TACTIC employs a search-based method, e.g., (1+1) ES in this work, to identify the critical environmental conditions based on the MUNIT model. Finally, testing driving scenes (or more training examples) are generated using the critical environmental conditions. For details see Section 3.

The workflow of TACTIC is shown in Figure 2. For a subject DNN-based ADS and a given environmental type, TACTIC first trains a MUNIT model using two sets of images: the original driving scenes which are used to synthesise testing scenes and the environmental dataset which contains driving scenes belong to the given environmental type. When the training process completes, the environmental condition space of the environmental type is encoded as a style space in the MUNIT model. Details are elaborated in Section 3.1. Then TACTIC sets a search objective which is to maximise the number and diversity of erroneous behaviours and employs a search-based method, e.g., (1+1) Evolution Strategy (ES) (Rechenberg, 1978) in this work, to search the style space for the environmental conditions that reach the search objective, i.e., the critical environmental conditions. Details are presented in Section 3.2 and Section 3.3. Finally, the original driving scenes are transformed by the MUNIT model into synthesised scenes under the critical environmental conditions, to test whether the subject DNN-based ADS can output consistent steering angles under the critical environmental conditions. It is worth mentioning that besides testing, the critical environmental conditions of TACTIC can also be used to synthesise more training examples to improve the performance of DNN-based ADSs by adversarial training (Pei et al., 2017; Tian et al., 2018), which we plan to study in future work.

3.1. Environmental Conditions

To encode the environmental condition space of an environmental type, we leverage MUNIT, a multimodal unsupervised image-to-image translation framework based on Generative Adversarial Networks (GAN). One insight of

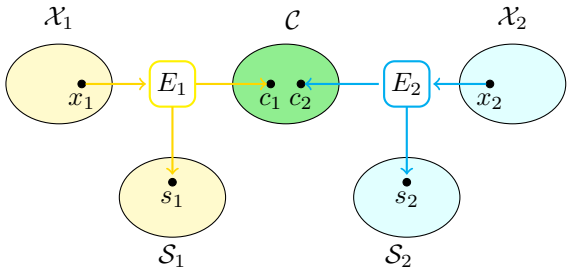


Figure 3. Structure of MUNIT (Huang et al., 2018). Images in each domain $\mathcal{X}_i$ are encoded to a shared content space $\mathcal{C}$ and a domain-specific style space $\mathcal{S}_i$ through an encoder E_i. Each encoder has an inverse decoder G_i omitted from this figure.

MUNIT is that the representation of images can be decomposed into a content space that is domain-invariant and a style space that captures domain-specific characteristics, as shown in Figure 3. Accordingly, when training a MUNIT model based on two sets of driving scenes that belong to two different environmental types, the domain-invariant content space will encode the information shared by scenes of the two different environmental types, such as the road information like the shapes of road and the roadside trees. The specific characteristics of each environmental type, e.g., the unique degrees of illumination, amounts of rain, and cloud patterns for the rainy weather type, would be encoded by the style space. Therefore, we can use the style space of an environmental type in a well-trained MUNIT model to represent the environmental condition space of the type.

After obtaining a MUNIT model, driving scenes under different environmental conditions of an environmental type can be realistically generated. Specifically, suppose that we have a MUNIT model M trained on two sets of driving scenes that belong to two environmental types: the environmental type $\mathcal{X}_1$ and the environmental type $\mathcal{X}_2$. To transform a driving scene x_1 in $\mathcal{X}_1$ to a driving scene x_2 in $\mathcal{X}_2$, the MUNIT model M first uses the encoder E_1 of $\mathcal{X}_1$ to decompose x_1 into a content vector c_1 and a style vector s_1, i.e., $(c_1, s_1) = E_1(x_1)$. Then, to produce x_2, M recombines the content c_1 with a style vector s_2 sampled from the style space $\mathcal{S}_2$ of $\mathcal{X}_2$ by the decoder G_2 of $\mathcal{X}_2$, i.e., $x_2 = G_2(c_1, s_2)$. In the above transformation process, by using another style vector s_2', we can generate a driving scene x_2', i.e., $x_2' = G_2(c_1, s_2')$, which has the same content as x_2 but a different environmental condition. Furthermore, we can also transform the entire set of driving scenes under different environmental conditions into those under the same environmental condition by using the same style vector. More specifically, let $\mathbf{X}_1 = \{x_1, x_2, ..., x_n\}$ be a set of diving scenes under different environmental conditions of $\mathcal{X}_1$. Through the encoder E_1, each scene $x_i \in \mathrm{X}_1$ is decomposed into a content vector c_i and a style vector s_i, i.e., $(c_i, s_i) = E_1(x_i)$. Then, we recombine each content vector c_i with the same style vector s sampled from the style space $\mathcal{S}_2$ of $\mathcal{X}_2$ by the decoder G_2 of $\mathcal{X}_2$ to generate a driving scene x_i', i.e., $x_i' = G_2(c_i, s)$. Finally, we obtain a set of driving scenes under the same environmental condition of $\mathcal{X}_2$, i.e., a set $\mathbf{X}_2 = \{x_1', x_2', ..., x_n'\}$. Figure 4 shows some driving scenes synthesised by respectively using the above two types of transformations, based on the MUNIT models used in TACTIC. More details about MUNIT can be found in Appendix A.

3.2. Search Objective for Critical Environmental Conditions

As introduced in Section 3.1, TACTIC uses the style space in a MUNIT model to represent the environmental condition space of a target environmental type. However, identifying the critical environmental conditions (i.e., critical style vectors) from the style space is still challenging. First, the style space in MUNIT is a high-dimensional vector space, and hence, a style vector in the style space is difficult for interpretation. For example, when transforming a sunny scene to a rainy scene, a style vector may encode information about the light effects, amounts of rain, cloud patterns, and even a mixture of all of them and many more, unlike the configurations in simulators where each configurable variable controls a certain simulation object. Second, the critical environmental conditions may be the corner cases which only occupy a small fraction of the entire space.

To alleviate these challenges, we use the search-based approach, which is proven to be effective for complex optimisation problems, to search the style space for the critical style vectors. The key to the success of the search is a proper fitness function that measures the fitness of the style vectors. The number of erroneous behaviours is certainly important metrics, since it indicates that the ADS are prone to errors. However, using such metrics alone may lead to the same type of errors repeatedly happening. Therefore, we design the fitness function for the style search to consider both the number and the types of errors, i.e., it is the combination of two effectiveness measures: (1) the ability to detect more diverse erroneous behaviours; and (2) the ability to detect more erroneous behaviours.

3.2.1. TESTING COVERAGE CRITERIA

In this work, we employ DNN coverage criteria to measure the ability of a style vector to detect diverse erroneous behaviours. Prior work has proved that increasing the DNN coverage can lead to more diverse behaviours (Pei et al., 2017; Tian et al., 2018; Xie et al., 2019; Huang et al., 2019), which, consequently, can increase the chance to detect more diverse erroneous behaviours. Therefore, a style vector is

Figure 4. Driving scenes synthesised using MUNIT. **Left:** We transform one original driving scene to driving scenes under different environmental conditions of an environmental type by using different styles. **Right:** We transform two driving scenes that have different environmental conditions to those under the same environmental condition of an environmental type by using the same style.

considered to have the potential to find defects of more diversity, if the driving scenes obtained by applying this style vector have higher DNN coverage measures. Many different DNN coverage metrics have been proposed. Since we need to distinguish a large number of style vectors by coverage, we carefully select two fine-grained coverage metrics on the sub-neuron level (Ma et al., 2018) for TACTIC to support, i.e., k-multisection Neuron Coverage (KMNC) and Neuron Boundary Coverage (NBC). It is worth mentioning, however, TACTIC is not bound to any specific metrics and new metrics can be easily supported.

Let O be a set of neurons of a DNN. Let $\phi(x, o)$ be the output value of neuron $o \in O$ for input x. When providing the training dataset D_{train} to the DNN model, the range of the observed neuron activation values of a neuron $o \in O$ is represented as $[low_o, high_o]$, where $low_o = \min_{x \in D_{train}} \phi(x, o)$ and $high_o = \max_{x \in D_{train}} \phi(x, o)$. Moreover, we refer the value range out of $[low_o, high_o]$, i.e., $(-\infty, low_o) \cup (high_o, +\infty)$, as the corner-case regions of the neuron o. Then, KMNC and NBC are defined as follows:

KMNC. Given a neuron $o \in O$, the KMNC measures how thoroughly the given input set T covers the range $[low_o, high_o]$. To quantify this, the range $[low_o, high_o]$ is divided into k equal sections (i.e., k-multisections), for $k > 0$. Let S_i^o be the i-th section with $1 \leq i \leq k$. Then if $\phi(x, o) \in S_i^o$, the i-th section is covered by at least one input $x \in T$. Therefore, the KMNC of a DNN obtained by the test input set T is defined as:

$$\mathsf{KMNC} = \frac{\sum_{o \in O} |\{S_i^o | \exists x \in T : \phi(x, o) \in S_i^o\}|}{k \times |O|}$$

NBC. Given a neuron $o \in O$, the NBC shows to what extent the corner-case regions are covered by the given input set T. If $\phi(x, o)$ belongs to $(-\infty, low_o)$ or $(high_o, +\infty)$, the corresponding corner-case region is covered by at least one input $x \in T$. Therefore, given the test input set T, let $\mathsf{UpperCornerNeuron} = \{o \in O | \exists x \in T : \phi(x, o) \in (high_o, +\infty)\}$ and $\mathsf{LowerCornerNeuron} = \{o \in O | \exists x \in T : \phi(x, o) \in (-\infty, low_o)\}$ be the sets of neurons that ever fall into the corner case regions, respectively. The NBC of a DNN achieved by the test input set T is defined as:

$$\mathsf{NBC} = \frac{|\mathsf{UpperCornerNeuron}| + |\mathsf{LowerCornerNeuron}|}{2 \times |O|}$$

Moreover, we respectively denote the regions that KMNC targets as KMNC regions and the regions that NBC targets as NBC regions in the rest of the paper.

3.2.2. METAMORPHIC ERROR ANALYSIS.

Following the existing work (Tian et al., 2018), we also adopt Metamorphic Relation (MR) (Chen et al., 2020) as the test oracle to determine whether a system behaviour is correct or not. In particular, the MR used in TACTIC is defined as that the steering angles should be consistent among the driving scenes transformed from the same ones by applying different style vectors, i.e., retain behaviour consistency. However, such an MR is too strong to be practical since the acceptable steering angle for a driving scene can be within a range (Tian et al., 2018). Therefore, the MR is further relaxed to a steering angle divergence within an error bound. Formally, let x_o be an original driving scene and x_t be a newly generated driving scene synthesised based on x_o in a target environmental type. The θ_o and θ_t represent the steering angles for x_o and x_t, respectively. Then the MR is defined as $|\theta_o - \theta_t| < \epsilon$, where the ϵ is a user-defined error bound.

Based on the MR, we can assess the tendency of behaviours to be erroneous ones. For a style vector, if the new driving scenes synthesised with this style vector have larger steering

angle divergences compared with the original driving scenes, then the style vector is considered to be able to detect more erroneous behaviours. TACTIC can be easily extended to test other functions of DNN-based ADS by providing a specification depicting the erroneous behaviours for those functions. For example, for the function of car-braking, a simple way to specify the erroneous behaviours can be that the two boolean variables indicating whether or not to brake respectively for the original and synthesised scenes have different values.

3.2.3. FITNESS FUNCTION DEFINITION.

Now we formally define the two effectiveness measures:

The ability to detect diverse erroneous behaviours of a style vector is measured by the increase in the testing coverage measured by KMNC or NBC. Let R_t be the set of uncovered KMNC/NBC regions so far during the search-based testing process and R_s be the set of regions that will be newly covered by the synthesised driving scenes using style vector s. The ability of s to detect diverse erroneous behaviours is calculated as: $F_c(s) = \frac{|R_s|}{|R_t|}$.

The ability to detect more erroneous behaviours of a style vector is measured by the mean steering angle divergences between the original dataset of driving scenes and the newly synthesised driving scenes with that style vector. Formally, let $g(x, s)$ be the function that transforms a driving scene x into another in the target environmental type using style vector s, and $f(x)$ be the function that returns the steering angle for the driving scene x. Let I_o be the original dataset of driving scenes, the ability of a style vector s to detect more erroneous behaviours is calculated as:

$$F_d(s) = \frac{1}{|I_o|} \sum_{x \in I_o} |f(x) - f(g(x, s))|.$$

Based on the above two effectiveness measures, we design the fitness function for TACTIC as follows:

$$F(s) = w_c * F_c(s) + w_d * norm(F_d(s)),$$

where s is a style vector, $norm(x) = \frac{x}{x+1}$ is a normalisation function (Greer & Ruhe, 2004; Zhang et al., 2008) for normalising the value of F_d in the same magnitude with F_c (i.e., within the same range of value $[0, 1]$), and w_c, w_d are the weights assigned to two measure, respectively, which allow for prioritisation of the measures. In this work, we simply treat the two measures as equal by setting $w_c = w_d = 1$, and leave the study of weight optimisation for future work.

3.3. Search of Critical Environmental Conditions

We are now ready to present how TACTIC identifies the critical environmental conditions of a given environmental type for a subject DNN-based ADS. We choose (1+1) Evolution Strategy (ES) (Rechenberg, 1978), which has shown to be effective in previous studies (Ali et al., 2013; Arcuri, 2013; Ji et al., 2018), as the search algorithm in TACTIC to identify the critical environmental conditions. TACTIC receives as inputs the subject DNN-based ADS N , the MUNIT model M trained for the given environmental type, and a set I_o of the original driving scenes. The output of TACTIC is a set S of the critical environmental conditions of the given environmental type.

Specifically, TACTIC iteratively calls the (1+1) ES to identify a critical environmental condition until a user-defined number of such conditions are obtained. Each iteration of TACTIC consists of the following main steps: *First*, the (1+1) ES is initialised with an initial individual, which is a style vector randomly sampled from the style space in the MUNIT model M. *Second*, the (1+1) ES evolves the individual s via selection and reproduction. In particular, a new individual s' is firstly generated based on the current individual s. Then, the fitness values of s and s' is evaluated as described in Section 3.2.3. After that, the individual with higher fitness value is kept as the current individual in the next evolution. *Lastly*, the (1+1) ES terminates the evolution process when the stopping condition is met and the current individual is added to the set S as a critical environmental condition. The stopping condition could be many, for example, the level of improvement during the evolution process or maximum time budget, etc. In this work, the stopping condition is set conservatively to that the fitness value has no improvement in 100 successive iterations.

TACTIC stops when a user-defined number of critical conditions are obtained, at which point TACTIC returns the critical environmental condition set S. In this work, we set the number to be four, as we observed in the experiments that increasing the number of style vectors when it is already above four would not result in a significant increase in the test coverage. Note that this number is related to the number of the driving scenes in the test dataset, and can be easily configured when necessary.

4. Experimental Evaluation

In this section, we present the experimental results to demonstrate the effectiveness of TACTIC. We mainly focus on evaluating whether TACTIC is able to find critical environmental conditions, which are used to synthesise driving scenes that can effectively reveal erroneous behaviours in different environmental types. Additionally, testing ADS under environmental conditions requires that the synthesised driving scenes realistically reflect reality, and therefore, we study the realism of the synthesised driving scenes.

4.1. Experimental Settings

Datasets and DNN-based ADSs. We conduct experiments on the Udacity dataset (Udacity, 2016). To demonstrate the effectiveness of TACTIC, we consider three popular pre-trained DNN-based ADSs, which have been widely used in previous work (Pei et al., 2017; Tian et al., 2018; Zhang et al., 2018; Zhou et al., 2020), i.e., Dave-orig (0bserver07, 2016), Dave-dropout (Navoshta, 2016), and Chauffeur (Emef, 2016). For each DNN-based ADSs, we study five environmental types (night, sunshine, rain, snow in daytime and snow in night) that are representatives of the runtime environments for DNN-based ADSs. The corresponding dataset of each environmental type is collected from YouTube. See more details about the datasets and targeted DNN-based ADSs in Appendix B.

Baselines. We consider two groups of baselines. (1) For the effectiveness comparison in Section 4.2 , we evaluate the effectiveness of TACTIC against two methods, i.e., an approach using randomly sampled environmental conditions (denoted as $\mathrm{R_c}$) and the state-of-the-art DeepRoad (Zhang et al., 2018). (2) For the image quality comparison in Section 4.3, we compare the realism of the driving scenes synthesised by TACTIC with three methods, i.e., DeepRoad (Zhang et al., 2018), DeepTest (Tian et al., 2018), and PreScan (Marketakis et al., 2009), which are the current state-of-the-art in testing DNN-based ADSs. Note that we do not include DeepTest and PreScan in the effectiveness comparison (c.f., Section 4.2) due to the synthesised driving scenes by these methods are unrealistic, as shown in Section 4.3.

Evaluation Metrics We use two metrics to evaluate the effectiveness of TACTIC: (1) the achieved test coverage (i.e., KMNC and NBC) and (2) the number of detected erroneous behaviours. For the achieved coverage, we focus on measuring the achieved coverage for CNNs used in the ADSs due to the coverage criteria are less effective on testing RNNs (Tian et al., 2018) and we set the number k of KMNC to 1000, consistent with DeepGauge (Ma et al., 2018). For the number of detected erroneous behaviours, we present the number of erroneous behaviours detected under four different error bounds (c.f. Section 3.2.2), which are consistent with the error bounds used in DeepRoad (Zhang et al., 2018). Regarding the quality of synthesised driving scenes, we conduct a user study to evaluate the realism of the synthesised driving scenes, since currently, user studies are the most effective standard to evaluate the realism of objects artificially synthesised.

4.2. Comparison with Baselines on Effectiveness

Setup. We execute TACTIC with two coverage-guiding strategies: KMNC (denoted as $\text{TACTIC}^{\text{KMNC}}$) and NBC (denoted as $\text{TACTIC}^{\text{NBC}}$), respectively, on each of the three subject DNN-based ADSs, under the five environmental types. In each execution, the entire set of driving scenes from the Udacity testing dataset is used as the original driving scenes (i.e., I_o in Section 3.3), and 4 critical environmental conditions are generated (c.f. Section 3.3), which are separately applied on the Udacity testing dataset to synthesise testing driving scenes (i.e., 4×5614 testing driving scenes are totally generated).

For $\mathrm{R_c}$, when testing the three subject systems, we randomly sample 4 style vectors (equal to the number of critical style vectors generated by TACTIC) for each of the environmental types and separately apply the 4 random style vectors on the Udacity testing dataset to synthesise testing driving scenes (i.e., 4×5614 testing driving scenes are totally generated). Moreover, to reduce the randomness of (1+1) ES used in TACTIC and $\mathrm{R_c}$, we conduct 10 runs for each experimental case and average the results.

For DeepRoad, we use the training datasets of the MUNIT models in TACTIC for training DeepRoad, and then generate driving scenes for each of the environmental types. Note that, given the Udacity testing dataset, DeepRoad can only generate the same number (5614) of testing driving scenes as the Udactiy dataset for each environmental type, since it transforms each driving scene into the other in a deterministic way. Therefore, for TACTIC, we apply just one critical style on the Udactiy testing dataset each time and then average the results for a fair comparison.

Due to the space limit, we report the results of NBC-guided TACTIC on Dave-orig in the main body of the paper. Results of KMNC-guided TACTIC on Dave-orig and results of TACTIC on Dave-dropout and Chauffeur can be found in Appendix C.1. The results on all the coverage-guiding strategies and the DNN-based ADSs are consistent.

Results. Table 1 and Table 2 summarise the results of comparing $\text{TACTIC}^{\text{NBC}}$ with $\mathrm{R_c}$ and DeepRoad, respectively. Better results are highlighted with a darker background. We discuss the results from two aspects: (1) the achieved test coverage, and (2) the number of detected erroneous behaviours. In terms of the achieved coverage, compared with $\mathrm{R_c}$, TACTIC achieves higher coverage in all environmental types, demonstrating that TACTIC can detect more diverse erroneous behaviours than $\mathrm{R_c}$. Compared with DeepRoad, TACTIC achieves slightly lower coverage in the environmental type of "Snow in Daytime" while obtaining higher coverage than DeepRoad in all the other four environmental types. We leave the analysis about why DeepRoad achieves higher coverage in Appendix C.1. In terms of the detected erroneous behaviours, compared with $\mathrm{R_c}$ and DeepRoad, TACTIC detects many more erroneous behaviours in all environmental types for both four error bounds. Furthermore, we also observe that, when the error bound increases, $\mathrm{R_c}$ and DeepRoad hardly detects erroneous behaviours while TAC-

Table 1. Results of comparing NBC-guided TACTIC with R_c on Dave-orig. Better results are highlighted with a darker background.

ENV. TYPE		NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
METHOD		TACTIC	R_c	TACTIC	R_c	TACTIC	R_c	TACTIC	R_c	TACTIC	R_c
COVERAGE	KMNC	73.68%	43.55%	56.24%	43.34%	50.71%	42.04%	54.41%	47.15%	72.09%	52.04%
	NBC	35.92%	3.18%	13.81%	1.97%	7.30%	2.00%	8.67%	3.88%	33.30%	10.10%
NUMBER OF ERRORS	10°	18675.1	2971.2	2629.8	1300.5	9795.2	1484.5	13450.8	3605.2	20879.0	7323.4
	20°	13790.0	269.7	196.5	103.8	3479.7	44.9	3396.8	197.7	17978.1	2583.0
	30°	8627.7	25.5	22.7	4.7	1933.8	0.1	845.1	10.6	14561.1	749.7
	40°	4303.7	1.9	0.7	0.0	687.7	0.0	209.5	0.8	12074.7	72.5

Table 2. Results of comparing NBC-guided TACTIC with DeepRoad on Dave-orig. Better results are highlighted with a darker background.

ENV. TYPE		NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
METHOD		TACTIC	DEEPROAD	TACTIC	DEEPROAD	TACTIC	DEEPROAD	TACTIC	DEEPROAD	TACTIC	DEEPROAD
COVERAGE	KMNC	54.59%	40.99%	45.37%	40.97%	40.71%	40.23%	41.80%	45.21%	60.03%	55.66%
	NBC	21.72%	5.99%	6.62%	2.05%	2.90%	2.05%	3.34%	3.81%	26.64%	21.50%
NUMBER OF ERRORS	10°	4885.0	613.0	708.8	355.0	3035.8	487.0	3708.0	1250.0	5041.0	3189.0
	20°	3898.3	114.0	41.5	40.0	1442.5	20.0	820.5	90.0	4418.0	1996.0
	30°	2662.5	33.0	4.8	3.0	816.5	0.0	154.3	2.0	3716.3	802.0
	40°	1620.5	9.0	0.0	0.0	304.5	0.0	47.3	0.0	2843.0	173.0

TIC still shows strong ability to detect erroneous behaviours. In summary, TACTIC can effectively identify critical environmental conditions, and reveal more diverse and more serious erroneous behaviours than R_c and DeepRoad.

Time Efficiency. We also analysed the efficiency of TACTIC. Compared to R_c and DeepRoad, TACTIC spent more time on testing DNN-based ADS, since TACTIC needs to search the style space to identify the critical environmental conditions. Specifically, about a total of 15 hours, including the time cost for driving scenes generation, model prediction and fitness function calculation, was required to identify a critical environmental condition in our experimental environment. Such a time cost is worthwhile, since TACTIC can obtain the critical environmental conditions under which the DNN-based ADS are more prone to errors by a relatively comprehensive exploration of the condition space of an environmental type. This cannot be done by existing approaches using synthesised driving scenes, and would require substantially more time for real-world tests (Kalra & Paddock, 2016).

4.3. Comparison with Baselines on Image Quality

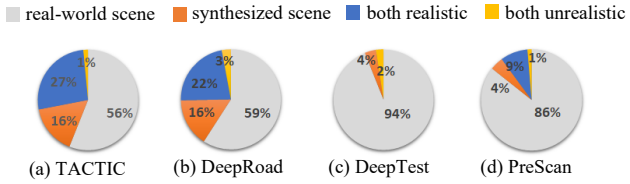
Setup. We conduct a user study to evaluate the realism of the testing driving scenes synthesised by TACTIC through comparing them with the test driving scenes synthesised by DeepRoad (Zhang et al., 2018), DeepTest (Tian et al., 2018), and PreScan (Marketakis et al., 2009). Specifically, we design an online questionnaire consisting of two questions: (1) "Which driving scene is more realistic?", and (2) "Which environmental type does the driving scene belong to?".

In the first questionnaire, 80 image pairs, each containing one synthesised image and one real image, were shown to the participants. Among the 80 synthesised images, each of the four approaches synthesised 20 images. The participants were asked to choose one of the four choices: the first image is more realistic, the second image is more realistic, both are realistic, and both are unrealistic. This question intends to compare the degrees of realism between the real-world driving scenes and the ones synthesised by different approaches.

In the second questionnaire, the same 80 synthesised images that were used in the first questionnaire were shown to the participants. The participants were asked to choose one environmental type for the driving scenes from the seven choices: rain, sunshine, night, fog, snow in daytime, snow in night, and image unrealistic. This question intends to compare the ability to synthesise testing driving scenes to realistically reflect the environmental conditions. All participants were given the same image pairs (or images).

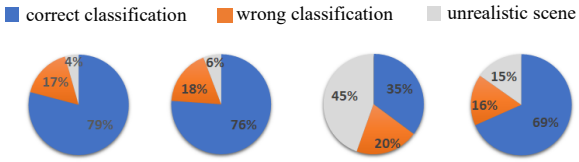
The questionnaire was distributed online, open to both students and industrial practitioners.

Results. We received answers from a total of 34 participants of which 12 are from industry and 22 are students. Figure 5 and Figure 6 present the results of answers to the first and the second questions, respectively. The results demonstrate that the testing driving scenes synthesised by TACTIC are more realistic than the ones synthesised by affine image transformation (e.g., DeepTest) or high-fidelity simulation (e.g., PreScan) and can better reflect real environments. Specifically, from Figure 5, it can be observed that there are nearly half of driving scene pairs where the synthesised driving scenes cannot be distinguished with the



(a) TACTIC (b) DeepRoad (c) DeepTest (d) PreScan

Figure 5. Comparison of realism among real-world driving scenes and synthesised scenes of (a) TACTIC, (b) DeepRoad, (c) DeepTest, and (d) PreScan. The grey sectors denote the ratios that the real-world scenes are selected. Therefore, the smaller the grey sectors are, the more realistic the synthesised scenes are.



(a) TACTIC (b) DeepRoad (c) DeepTest (d) PreScan

Figure 6. Ratios of correct classification of the driving scenes synthesised from (a) TACTIC, (b) DeepRoad, (c) DeepTest, and (d) PreScan. The blue sectors denote the ratios that the synthesised scenes are correctly classified as their corresponding environmental types. Therefore, the larger blue sectors suggest that the synthesised scenes can more realistically reflect the environmental types.

real-world ones in human perception for TACTIC (44%) and DeepRoad (41%). In contrast, for DeepTest and PreScan, there are on average 94% and 86% driving scene pairs where the real-world driving scenes are considered more realistic, respectively. As shown in Figure 6, in general, the testing driving scenes synthesised by TACTIC and DeepRoad can more realistically reflect the environmental types compared to the ones synthesised by DeepTest and PreScan.

Furthermore, we observe that the testing driving scenes synthesised by TACTIC and DeepRoad have comparable results for the degrees of realism. The reason is that both TACTIC and DeepRoad use a type of GAN to generate new testing driving scenes. However, as shown in Section 4.2, compared with DeepRoad, TACTIC manages to detect more diverse and more serious erroneous behaviours.

4.4. Ablation Study

Setup. We perform the ablation study to justify the selection of (1+1) ES in TACTIC. In particular, we implement a new version of TACTIC by replacing the search algorithm (1+1) ES with a random search (RS) and compare the effectiveness of the two versions. Except for the search algorithm, the other implementation and settings, e.g., the fitness function, are not changed. For each experimental case, we also conduct 10 runs and average the results to reduce the randomness in the search. Again, we mainly report the results of NBC-guided TACTIC on Dave-orig in the main body of this paper, and leave the complete results to Appendix C.2.

Table 3. Results of comparing the effectiveness of (1+1) ES and RS in NBC-guided TACTIC on Dave-orig. Better results are highlighted with a darker background.

ENV. TYPE	METHOD	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10^o	20^o	30^o	40^o
NIGHT	(1+1) ES	73.68%	35.92%	18675.1	13790	8627.7	4303.7
	RS	53.09%	10.73%	15833.3	4783.0	384.0	9.1
SUNSHINE	(1+1) ES	56.24%	13.81%	2629.8	196.5	22.7	0.7
	RS	45.48%	2.85%	1451.4	81.2	5.0	0.0
RAIN	(1+1) ES	50.71%	7.30%	9795.2	3479.7	1993.8	687.7
	RS	42.35%	2.00%	3633.9	188.1	3.8	0.0
SNOW IN DAYTIME	(1+1) ES	54.41%	8.67%	13450.8	3396.8	845.1	209.5
	RS	51.87%	6.80%	9730.0	748.8	20.0	14.3
SNOW IN NIGH	(1+1) ES	72.09%	33.30%	20879.0	17978.1	14561.1	12074.7
	RS	65.45%	25.69%	19341.5	14917.3	9955.4	4966.7

Results. Table 3 summarises the comparison results and we also discuss the results from the achieved test coverage and the number of detected erroneous behaviours. For the achieved test coverage, TACTIC using (1+1) ES achieves higher coverage than TACTIC using RS in both five environmental types. For the number of the detected erroneous behaviours, TACTIC using (1+1) ES obtains significantly more erroneous behaviours on all environmental types. Additionally, we also observe that the difference between the number of erroneous behaviours detected by the two approaches increases with the error bound. Such results demonstrate that (1+1) ES has better performance in identifying critical environmental conditions than random search.

5. Conclusion

In this paper, we propose to study the impact of different environmental conditions on the DNN-based ADS and raise the problem of identifying the critical environmental conditions that would make the system under test more prone to erroneous behaviours. We introduce a novel approach TACTIC, which employs a search-based method to search the environmental condition space of the given environmental type generated by MUNIT. Large-scale experiments demonstrate that TACTIC can effectively identify critical environmental conditions and synthesise realistic testing driving scenes. Compared to the state-of-the-art approaches, TACTIC can reveal more diverse and more erroneous behaviours for the popular DNN-based ADSs, and meanwhile, reach a satisfactory testing coverage.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Nos. 62032010 and 61972193) and the Fundamental Research Funds for the Central Universities of China (No. 14380027).

References

0bserver07. Nvidia-autopilot-keras: End to end learning for self-driving cars, 2016. URL https://github.com/0bserver07/Nvidia-Autopilot-Keras. [online, accessed 27-May-2021].

Abdessalem, R. B., Nejati, S., Briand, L. C., and Stifter, T. Testing advanced driver assistance systems using multi-objective search and neural networks. In Lo, D., Apel, S., and Khurshid, S. (eds.), *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, ASE 2016, Singapore, September 3-7, 2016*, pp. 63–74. ACM, 2016. doi: 10.1145/2970276.2970311. URL https://doi.org/10.1145/2970276.2970311.

Abdessalem, R. B., Nejati, S., Briand, L. C., and Stifter, T. Testing vision-based control systems using learnable evolutionary algorithms. In Chaudron, M., Crnkovic, I., Chechik, M., and Harman, M. (eds.), *Proceedings of the 40th International Conference on Software Engineering, ICSE 2018, Gothenburg, Sweden, May 27 - June 03, 2018*, pp. 1016–1026. ACM, 2018. doi: 10.1145/3180155.3180160. URL https://doi.org/10.1145/3180155.3180160.

Ali, S., Iqbal, M. Z. Z., Arcuri, A., and Briand, L. C. Generating test data from OCL constraints with search techniques. *IEEE Trans. Software Eng.*, 39(10):1376–1402, 2013. doi: 10.1109/TSE.2013.17. URL https://doi.org/10.1109/TSE.2013.17.

Arcuri, A. It really does matter how you normalize the branch distance in search-based software testing. *Softw. Test. Verification Reliab.*, 23(2):119–147, 2013. doi: 10.1002/stvr.457. URL https://doi.org/10.1002/stvr.457.

Arcuri, A. and Briand, L. C. A hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. *Softw. Test. Verification Reliab.*, 24(3):219–250, 2014. doi: 10.1002/stvr.1486. URL https://doi.org/10.1002/stvr.1486.

Badvboynofear. Driving in the snow, Mar 2018. URL https://www.youtube.com/watch?v=fm5nKWeVNGI. [online, accessed 27-May-2021].

Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016. URL http://arxiv.org/abs/1604.07316.

Boudette, N. E. Teslas self-driving system cleared in deadly crash. *The New York Times*, 19, 2017.

Capon, J. A. Elementary statistics for the social sciences: Study guide, 1991.

Chen, T. Y., Cheung, S. C., and Yiu, S. Metamorphic testing: A new approach for generating next test cases. *CoRR*, abs/2002.12543, 2020. URL https://arxiv.org/abs/2002.12543.

Du, X., Xie, X., Li, Y., Ma, L., Liu, Y., and Zhao, J. Deepstellar: model-based quantitative analysis of stateful deep learning systems. In Dumas, M., Pfahl, D., Apel, S., and Russo, A. (eds.), *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019*, pp. 477–487. ACM, 2019. doi: 10.1145/3338906.3338954. URL https://doi.org/10.1145/3338906.3338954.

Emef. Sdc, 2016. URL https://github.com/emef/sdc. [online, accessed 27-May-2021].

Fremont, D. J., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A. L., and Seshia, S. A. Scenic: a language for scenario specification and scene generation. In McKinley, K. S. and Fisher, K. (eds.), *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019, Phoenix, AZ, USA, June 22-26, 2019*, pp. 63–78. ACM, 2019. doi: 10.1145/3314221.3314633. URL https://doi.org/10.1145/3314221.3314633.

Gibbs, S. Ubers self-driving car saw the pedestrian but didnt swerve–report. *The Guardian*, 2018.

Greer, D. and Ruhe, G. Software release planning: an evolutionary and iterative approach. *Inf. Softw. Technol.*, 46(4):243–253, 2004. doi: 10.1016/j.infsof.2003.07.002. URL https://doi.org/10.1016/j.infsof.2003.07.002.

Grzywaczewski, A. Training ai for self-driving vehicles: the challenge of scale, 2017.

Han, J. C. and Zhou, Z. Q. Metamorphic fuzz testing of autonomous vehicles. In *ICSE '20: 42nd International Conference on Software Engineering, Workshops, Seoul, Republic of Korea, 27 June - 19 July, 2020*, pp. 380–385. ACM, 2020. doi: 10.1145/3387940.3392252. URL https://doi.org/10.1145/3387940.3392252.

Huang, W., Sun, Y., Huang, X., and Sharp, J. testrnn: Coverage-guided testing on recurrent neural networks. *CoRR*, abs/1906.08557, 2019. URL http://arxiv.org/abs/1906.08557.

Huang, X., Liu, M., Belongie, S. J., and Kautz, J. Multimodal unsupervised image-to-image translation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, pp. 179–196. Springer, 2018. doi: 10.1007/978-3-030-01219-9_11. URL https://doi.org/10.1007/978-3-030-01219-9_11.

Ji, R., Li, Z., Chen, S., Pan, M., Zhang, T., Ali, S., Yue, T., and Li, X. Uncovering unknown system behaviors in uncertain networks with model and search-based testing. In *11th IEEE International Conference on Software Testing, Verification and Validation, ICST 2018, Västerås, Sweden, April 9-13, 2018*, pp. 204–214. IEEE Computer Society, 2018. doi: 10.1109/ICST.2018.00029. URL https://doi.org/10.1109/ICST.2018.00029.

Kalra, N. and Paddock, S. M. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94:182–193, 2016.

Li, Z., Ma, X., Xu, C., Cao, C., Xu, J., and Lü, J. Boosting operational DNN testing efficiency through conditioning. In Dumas, M., Pfahl, D., Apel, S., and Russo, A. (eds.), *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019*, pp. 499–509. ACM, 2019. doi: 10.1145/3338906.3338930. URL https://doi.org/10.1145/3338906.3338930.

Ma, L., Juefei-Xu, F., Zhang, F., Sun, J., Xue, M., Li, B., Chen, C., Su, T., Li, L., Liu, Y., Zhao, J., and Wang, Y. Deepgauge: multi-granularity testing criteria for deep learning systems. In Huchard, M., Kästner, C., and Fraser, G. (eds.), *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018*, pp. 120–131. ACM, 2018. doi: 10.1145/3238147.3238202. URL https://doi.org/10.1145/3238147.3238202.

Ma, L., Juefei-Xu, F., Xue, M., Li, B., Li, L., Liu, Y., and Zhao, J. Deepct: Tomographic combinatorial testing for deep learning systems. In Wang, X., Lo, D., and Shihab, E. (eds.), *26th IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2019, Hangzhou, China, February 24-27, 2019*, pp. 614–618. IEEE, 2019. doi: 10.1109/SANER.2019.8668044. URL https://doi.org/10.1109/SANER.2019.8668044.

Marketakis, Y., Tzanakis, M., and Tzitzikas, Y. Prescan: towards automating the preservation of digital objects. In Chbeir, R., Badr, Y., Kapetanios, E., and Traina, A. J. M. (eds.), *MEDES '09: International ACM Conference on Management of Emergent Digital EcoSystems, Lyon, France, October 27-30, 2009*, pp. 404–411. ACM, 2009. doi: 10.1145/1643823.1643898. URL https://doi.org/10.1145/1643823.1643898.

McGowan, D. Driving on snow - greenville, nc 1-4-2018 at 7:00am, Jan 2018. URL https://www.youtube.com/watch?v=ps56tnQG8V0. [online, accessed 27-May-2021].

Navoshta. Behavioral cloning: End to end learning for self-driving cars, 2016. URL https://github.com/navoshta/behavioral-cloning. [online, accessed 27-May-2021].

Odena, A., Olsson, C., Andersen, D. G., and Goodfellow, I. J. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4901–4911. PMLR, 2019. URL http://proceedings.mlr.press/v97/odena19a.html.

Pei, K., Cao, Y., Yang, J., and Jana, S. Deepxplore: Automated whitebox testing of deep learning systems. In *Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017*, pp. 1–18. ACM, 2017. doi: 10.1145/3132747.3132785. URL https://doi.org/10.1145/3132747.3132785.

Rechenberg, I. Evolutionsstrategien. In *Simulationsmethoden in der Medizin und Biologie*, pp. 83–114. Springer, 1978.

Sun, Y., Huang, X., and Kroening, D. Testing deep neural networks. *CoRR*, abs/1803.04792, 2018. URL http://arxiv.org/abs/1803.04792.

Tian, Y., Pei, K., Jana, S., and Ray, B. Deeptest: automated testing of deep-neural-network-driven autonomous cars. In Chaudron, M., Crnkovic, I., Chechik, M., and Harman, M. (eds.), *Proceedings of the 40th International Conference on Software Engineering, ICSE 2018, Gothenburg, Sweden, May 27 - June 03, 2018*, pp. 303–314. ACM, 2018. doi: 10.1145/3180155.3180220. URL https://doi.org/10.1145/3180155.3180220.

Tours, D. C. Driving on california freeway from calabasas to universal studios hollywood.no music, Jun 2018. URL https://www.youtube.com/

watch?v=01_SVshk-MY. [online, accessed 27-May-2021].

Udacity. The udacity open source self-driving car project, 2016. URL https://github.com/udacity/self-driving-car. [online, accessed 27-May-2021].

Utah, J. Los angeles 4k - night drive, Feb 2019. URL https://www.youtube.com/watch?v=lTvYjERVAnY. [online, accessed 27-May-2021].

Vargha, A. and Delaney, H. D. A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132, 2000.

Vids, A. Rain on a car roof - 1 hour - asmr, Apr 2014. URL https://www.youtube.com/watch?v=O88fXBx-Qdg. [online, accessed 27-May-2021].

Xie, X., Ma, L., Juefei-Xu, F., Xue, M., Chen, H., Liu, Y., Zhao, J., Li, B., Yin, J., and See, S. Deephunter: a coverage-guided fuzz testing framework for deep neural networks. In Zhang, D. and Møller, A. (eds.), *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2019, Beijing, China, July 15-19, 2019*, pp. 146–157. ACM, 2019. doi: 10.1145/3293882.3330579. URL https://doi.org/10.1145/3293882.3330579.

Zhang, M., Zhang, Y., Zhang, L., Liu, C., and Khurshid, S. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In Huchard, M., Kästner, C., and Fraser, G. (eds.), *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018*, pp. 132–142. ACM, 2018. doi: 10.1145/3238147.3238187. URL https://doi.org/10.1145/3238147.3238187.

Zhang, Y., Finkelstein, A., and Harman, M. Search based requirements optimisation: Existing work and challenges. In Paech, B. and Rolland, C. (eds.), *Requirements Engineering: Foundation for Software Quality, 14th International Working Conference, REFSQ 2008, Montpellier, France, June 16-17, 2008, Proceedings*, volume 5025 of *Lecture Notes in Computer Science*, pp. 88–94. Springer, 2008. doi: 10.1007/978-3-540-69062-7_8. URL https://doi.org/10.1007/978-3-540-69062-7_8.

Zhou, H., Li, W., Kong, Z., Guo, J., Zhang, Y., Yu, B., Zhang, L., and Liu, C. Deepbillboard: systematic physical-world testing of autonomous driving systems. In Rothermel, G. and Bae, D. (eds.), *ICSE '20: 42nd International Conference on Software Engineering, Seoul, South Korea, 27 June - 19 July, 2020*, pp. 347–358. ACM, 2020. doi: 10.1145/3377811.3380422. URL https://doi.org/10.1145/3377811.3380422.

A. Multimodal Unsupervised Image-to-image Translation (MUNIT) Framework

In this section, we give a brief description of the Multimodal Unsupervised Image-to-Image Translation (MUNIT) framework proposed by Huang et al. (2018).

MUNIT is based on a partially shared latent space assumption. It first assumes that the latent space of images can be decomposed into a content space and a style space. Then, it further assumes that images in different domains share a common content space but not the style space. In our context, the domains are different environmental types (e.g., sunny and rainy weather types) and the images are driving scenes. Therefore, the content space captures information that is shared by scenes of different environmental types, such as the shapes of roads and the roadside trees. The style space captures the variants of visual representation of the same content from different environmental types, for instance, the unique degrees of illumination, amounts of rain, and cloud patterns for the rainy weather type. Intuitively, the corresponding style space of an environmental type can be used to represent the environmental condition space of the environmental type.

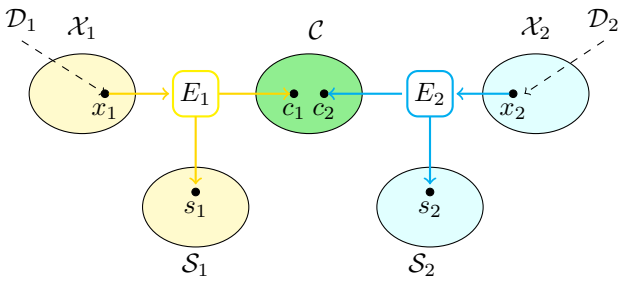


Figure 7. Structure of MUNIT. Images in each domain $\mathcal{X}_i$ are encoded to a shared content space $\mathcal{C}$ and a domain-specific style space $\mathcal{S}_i$ through an encoder E_i. Each encoder has an inverse generator G_i omitted from this figure. For each domain $\mathcal{X}_i$, there is also a discriminator D_i which detect whether the image belong to $\mathcal{X}_i$.

Figure 7 presents the structure of MUNIT, where $\mathcal{X}_1$ and $\mathcal{X}_2$ denote two different domains (i.e., two different environmental types in our context). For each domain $\mathcal{X}_i$, there are an encoder E_i which projects the images from $\mathcal{X}_i$ to a domain-invariant content space $\mathcal{C}$ and a domain-specific style space $\mathcal{S}_i$, and a decoder G_i which reproduces images of $\mathcal{X}_i$. Furthermore, D_1 and D_2 are two discriminators that detect whether the image belongs to $\mathcal{X}_1$ or $\mathcal{X}_2$, respectively. Specifically, these discriminators are excepted to differentiate whether the input images are real or synthetic ones (i.e., images produced by a well-trained generator). All E_i, G_i, and D_i in a MUNIT model are incarnated as deep neural networks, and learned by optimising the loss function that is consisting of the following costs:

- **Image Reconstruction Loss:** minimising the loss of image reconstruction for each $< E_i, G_i >$, which encourages reconstruction in the direction: image→latent→image.
- **Latent Reconstruction Loss:** minimising the loss of latent code reconstruction for each $< E_i, G_i >$, which encourages reconstruction in the direction: latent → image → latent.
- **GAN Loss:** achieving the equilibrium point in the minimax game for each $< G_i, D_i >$, where D_i aims at distinguishing between synthetic images produced by G_i and real images in $\mathcal{X}_i$.

In general, the total loss is defined as:

$$\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} \mathcal{L}_{\mathsf{GAN}_1} + \mathcal{L}_{\mathsf{GAN}_2} + \lambda_x(\mathcal{L}^{\mathsf{image}}_{\mathsf{recon}_1} + \mathcal{L}^{\mathsf{image}}_{\mathsf{recon}_2}) + \lambda_c(\mathcal{L}^{\mathsf{content}}_{\mathsf{recon}_1} + \mathcal{L}^{\mathsf{content}}_{\mathsf{recon}_2}) + \lambda_s(\mathcal{L}^{\mathsf{style}}_{\mathsf{recon}_1} + \mathcal{L}^{\mathsf{style}}_{\mathsf{recon}_2})$$

where $\mathcal{L}^{\mathsf{image}}_{\mathsf{recon}_i}$ represents the image reconstruction loss for $< E_i, G_i >$. $\mathcal{L}^{\mathsf{content}}_{\mathsf{recon}_i}$ and $\mathcal{L}^{\mathsf{style}}_{\mathsf{recon}_i}$ respectively represents the content code reconstruction loss and the style code reconstruction loss for $< E_i, G_i >$. $\mathcal{L}_{\mathsf{GAN}_i}$ denotes the GAN loss for $< G_i, D_i >$. λ_x, λ_c, λ_s are weights that control the importance of reconstruction terms.

Based on a well-trained MUNIT model, we can easily translate an image to another domain. For example, to translate an image $x_1 \in \mathcal{X}_1$ to $\mathcal{X}_2$, the MUNIT model first decomposes it into a content latent code c_1 and a style latent code s_1 by

the encoder E_1, i.e., $(c_1, s_1) = E_1(x_1)$. Then it uses the generator G_2 to produce an image x_2 by combining the content code c_1 and a style latent code s_2 sampled from the style space $\mathcal{S}_2$ of $\mathcal{X}_2$, i.e., $x_2 = G_2(c_1, s_2)$. By sampling different style latent code s_2 from the style space $\mathcal{S}_2$ of $\mathcal{X}_2$, multiple images with different appearances can be produced. For example, as shown in the left half of Figure 4, we can use one original driving scene to synthesise driving scenes under different environmental conditions of an environmental condition type. The original driving scene belongs to the sunny weather type. When transforming the original scene to the environmental type of 'Rain', we sampled two different style codes from the corresponding style space of the 'Rain' type. We can see that both synthesised driving scenes belong to the same environmental type but they have observable different degrees of light and amounts of rain. Besides this transformation, we can also use MUNIT to transform the entire set of driving scenes under different environmental conditions into those under the same environmental condition by using the same style code, as shown in the right half of Figure 4. We applied the same style code from the style space of the environmental type of 'Rain' on two different original driving scenes. It can be observed that the synthesised scenes are not only under the same environmental type, but also the same environmental condition by having similar degrees of light and amounts of rain.

B. Details for Datasets and Target DNN-based ADSs

In this section, we present details about the subject DNN-based ADSs and datasets used in our experiments.

B.1. Subject DNN-based ADSs

We focus on testing DNN-based ADSs that perform end-to-end steering angle control. Three popular pre-trained DNN-based ADSs that have been widely used in previous work (Pei et al., 2017; Tian et al., 2018; Zhang et al., 2018; Zhou et al., 2020), i.e., Dave-orig (0bserver07, 2016), Dave-dropout (Navoshta, 2016), and Chauffeur (Emef, 2016), are selected as the subject systems. Dave-orig and Dave-dropout both use the CNN models base on the DAVE-2 self-driving architecture from NVIDIA (Bojarski et al., 2016). Chauffeur is one of the top-ranked DNN models in the Udacity self-driving car challenge (Udacity, 2016). It consists of one CNN model and one LSTM model. For an input driving scene, the CNN first extracts features from the input and then the LSTM predicts the steering angle of the input based on the concatenation of 100 features extracted by the CNN from the previous 100 consecutive driving scenes.

Table 4. Details of datasets

DATASETS	NUMBER	DURATION	ENV. TYPE
UDACITY TRAINING	33808	-	SUNSHINE
UDACITY TESTING	5614	-	SUNSHINE
LOS ANGELES - NIGHT DRIVE (UTAH, 2019)	2000	1:17:03	NIGHT
DRIVING ON CALIFORNIA FREEWAY (TOURS, 2018)	1600	38:49	SUNSHINE
RAIN ON A CAR ROOF (VIDS, 2014)	1000	1:09:04	RAIN
DRIVING ON SNOW - GREENVILLE (MCGOWAN, 2018)	1200	28:56	SNOW IN DAYTIME
DRIVING IN THE SNOW (BADVBOYNOFEAR, 2018)	1200	1:04:53	SNOW IN NIGHT

B.2. Datasets

Table 4 presents the detailed information of all datasets used in our experiments. All three DNNs are trained on the Udacity dataset (Udacity, 2016), which is for the Udacity self-driving car challenge, containing 33808 training samples and 5614 testing samples. Each sample consists of a driving scene captured by a camera mounted behind the windshield of a driving car and the simultaneous steering angle issued by the human driver.

We study five environmental types (night, sunshine, rain, snow in daytime and snow in night) that are representatives of the runtime environments for DNN-based ADSs. The acquisition of the datasets used for training the MUNIT models is simple: we just searched videos longer than 20 minutes of driving in the five environmental types from YouTube, and conducted an automatic downsampling on the selected video to skip consecutive frames having similar contents and used the retained frames to construct the datasets. Table 4 shows the names of the videos and their links as in citations, the number of sampled frames, the duration of the videos, and the corresponding environmental types.

C. Further Experimental Details

In this section, we present further experimental details as well as some in-depth analysis.

C.1. Comparison with Baselines on Effectiveness

In order to evaluate the effectiveness of TACTIC, we compare TACTIC against two baseline methods, i.e., an approach using randomly sampled environmental conditions $\mathrm{R_c}$ and the state-of-the-art DeepRoad.

For TACTIC, we respectively execute TACTIC with two coverage-guiding strategies: KMNC (denoted as TACTIC$^{\mathrm{KMNC}}$) and NBC (denoted as TACTIC$^{\mathrm{NBC}}$) on each of the three subject DNN-based ADSs, under the five environmental types. In total, 30 experimental cases are executed (2 strategies $\times$ 5 environmental types $\times$ 3 systems). Furthermore, to reduce the randomness of (1+1) ES used in TACTIC, we conduct 10 runs for each case and average the results.

For $\mathrm{R_c}$, for each subject DNN-based ADS, we randomly sample 4 style vectors (equal to the number of critical style vectors generated by TACTIC) for each of the environmental types. In total, 15 experimental cases are executed (5 environmental types $\times$ 3 systems). We also repeat $\mathrm{R_c}$ 10 times for each experimental case and average the results.

For DeepRoad, we use the training datasets of the MUNIT models in TACTIC for training DeepRoad, and then generate driving scenes for each of the environmental types. Moreover, given the Udacity testing dataset, DeepRoad can only generate the same number (5614) of testing driving scenes as the Udactiy dataset for each environmental type, since it transforms each driving scene into the other in a deterministic way. Therefore, for TACTIC, we apply just one critical style on the Udactiy testing dataset each time and then average the results for a fair comparison.

Table 5. Average results of KMNC-guided TACTIC using (1+1) ES

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	69.12%	32.15%	21277.6	19499.2	15505.3	8952.7
	SUNSHINE	52.81%	9.90%	5668.1	540.5	22.8	0.4
	RAIN	46.21%	4.88%	13482.0	7084.2	3058.4	809.3
	SNOW IN DAYTIME	47.08%	3.65%	17938.7	6596.3	737.5	109.9
	SNOW IN NIGHT	72.95%	33.80%	21415.2	19480.5	17721.2	13677.6
DAVE-DROPOUT	NIGHT	51.55%	21.10%	16547.8	4846.6	917.5	113.4
	SUNSHINE	38.54%	10.32%	7031.0	793.5	30.0	0.2
	RAIN	40.07%	11.03%	13334.3	4707.7	454.6	67.8
	SNOW IN DAYTIME	39.82%	11.79%	18828.0	17709.0	11744.8	1966.2
	SNOW IN NIGHT	33.06%	8.48%	21539.9	18721.5	4701.6	7.3
CHAUFFEUR	NIGHT	78.23%	38.03%	1608.8	117.9	0.0	0.0
	SUNSHINE	69.41%	17.03%	439.1	16.0	0.0	0.0
	RAIN	61.56%	8.02%	15342.6	8377.9	1507.3	0.8
	SNOW IN DAYTIME	72.93%	20.23%	19188.2	5275.1	145.9	0.5
	SNOW IN NIGHT	71.87%	24.62%	9995.8	2040.4	25.8	0.0

Comparison results with $\mathrm{R_c}$. Table 5 and Table 6 summarise the average results achieved by TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$ over 10 repeated runs for each experimental case, respectively. Table 7 summarises the average results achieved by TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$ over 10 repeated runs for each experimental case. In Table 5, Table 6, and Table 7, Column *Coverage* is the achieved test coverage and Column *Number of Errors* is the number of detected erroneous behaviours. From these results. we have the following findings:

In terms of the achieved coverage, TACTIC achieves higher coverage than $\mathrm{R_c}$ in most cases. For Dave-orig, Dave-dropout, and Chauffeur, TACTIC$^{\mathrm{KMNC}}$ achieves on average 12.33%, 4.03%, and 5.87% higher coverage than $\mathrm{R_c}$, respectively, and TACTIC$^{\mathrm{NBC}}$ achieves on average 15.69%, 6.38%, and 7.42% higher coverage than $\mathrm{R_c}$, respectively. In terms of the detected erroneous behaviours, TACTIC detects more erroneous behaviours than $\mathrm{R_c}$. In total, for Dave-orig, Dave-dropout, and Chauffeur TACTIC$^{\mathrm{KMNC}}$ detects on average 470.77%, 1495.11%, and 738.80% more erroneous behaviours than $\mathrm{R_c}$, and TACTIC$^{\mathrm{NBC}}$ detects on average 330.47%, 1171.31%, and 501.72% more erroneous behaviours than $\mathrm{R_c}$. Furthermore, we also observe that, when the error bound increases, $\mathrm{R_c}$ hardly detects erroneous behaviours while TACTIC still shows a strong

Table 6. Average results of NBC-guided TACTIC using (1+1) ES

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	73.68%	35.92%	18675.1	13790	8627.7	4303.7
	SUNSHINE	56.24%	13.81%	2629.8	196.5	22.7	0.7
	RAIN	50.71%	7.30%	9795.2	3479.7	1993.8	687.7
	SNOW IN DAYTIME	54.41%	8.67%	13450.8	3396.8	845.1	209.5
	SNOW IN NIGHT	72.09%	33.30%	20879.0	17978.1	14561.1	12074.7
DAVE-DROPOUT	NIGHT	52.46%	22.09%	11114.4	2967.8	463.2	22.7
	SUNSHINE	43.26%	16.12%	1879.2	251.3	15.3	0.1
	RAIN	43.88%	13.75%	12872.3	5872.6	583.7	79.5
	SNOW IN DAYTIME	42.77%	13.92%	17648.9	14779.5	8012.4	930.0
	SNOW IN NIGHT	33.47%	8.55%	20973.3	8882.6	1831.0	21.1
CHAUFFEUR	NIGHT	78.93%	39.92%	2965.9	198.7	0.0	0.0
	SUNSHINE	69.85%	20.49%	306.3	22.8	0.0	0.0
	RAIN	62.19%	11.21%	12560.4	6627.6	915.5	3.4
	SNOW IN DAYTIME	71.36%	23.90%	12600.5	669.4	73.0	0.2
	SNOW IN NIGHT	72.61%	26.97%	2955.5	240.2	12.8	0.0

ability to detect erroneous behaviours. For example, for Dave-dropout, $\mathrm{R_c}$ detects no erroneous behaviours when the error bound is set to $40°$ in the environmental type of "Snow in Daytime", while TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$ still detect on average 1966.2 and 930 erroneous behaviours, respectively.

Following existing guidelines (Arcuri & Briand, 2014), to further investigate whether the differences between TACTIC and $\mathrm{R_c}$ are statistically significant, we use the nonparametric pairwise Mann-Whitney U test (Capon, 1991) and Vargha-Delaney Statistics $\hat{A}_{12}$ (Vargha & Delaney, 2000), where for two approaches A and B, A has significantly better performance than B if $\hat{A}_{12}$ is higher than 0.5 and the p-value is less than 0.05. The level of significance (α) is set to 0.05.

Table 8 reports the statistical test results comparing TACTIC$^{\mathrm{KMNC}}$ and $\mathrm{R_c}$, while those for comparing TACTIC$^{\mathrm{KMNC}}$ and $\mathrm{R_c}$ are reported in Table 9. The results show that: (1) for the achieved test coverage, TACTIC obtains significantly higher coverage than $\mathrm{R_c}$ in 86.67% (26 out of 30) experimental cases ($\hat{A}_{12}$ greater than 0.78 and p-value less than 0.01). For the other 4 cases, TACTIC also achieves comparable coverage to $\mathrm{R_c}$, with an average of only 0.66% lower coverage than $\mathrm{R_c}$. (2) for the number of detected erroneous behaviours, TACTIC detects significantly more erroneous behaviours than $\mathrm{R_c}$ for 91.67% (110 cases) of the 120 cases (4 error bounds and each bound has 30 experimental cases), by having $\hat{A}_{12}$ greater than 0.75 and p-value less than 0.04. For the 10 exceptional cases where the significance is not observed, the numbers of erroneous behaviours detected by both TACTIC and $\mathrm{R_c}$ are not large enough to pass the significance tests, but in no case, TACTIC detects fewer erroneous behaviours than $\mathrm{R_c}$.

In summary, we can conclude that TACTIC manages to detect more diverse and more serious erroneous behaviours than $\mathrm{R_c}$.

Comparison results with DeepRoad. Table 10 presents the results achieved by DeepRoad. Table 11 and Table 12 present the average results achieved by TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$, respectively. Recall that we separately apply each critical environmental condition produced by TACTIC and average the results to conduct a comparison with DeepRoad.

In terms of the detected erroneous behaviours, the comparison results are consistent with the results of comparing TACTIC with $\mathrm{R_c}$ that TACTIC always detects substantially more erroneous behaviours than DeepRoad on all three subject systems with the four error bounds. In total, for Dave-orig, Dave-dropout, and Chauffeur, TACTIC$^{\mathrm{KMNC}}$ detects on average 390.29%, 956.14%, and 514.88% more erroneous behaviours than DeepRoad, respectively, and TACTIC$^{\mathrm{NBC}}$ detects on average 341.92%, 931.90%, and 416.74% more erroneous behaviours than DeepRoad, respectively.

The comparison of the achieved coverage is more complex. For Dave-orig and Dave-dropout, TACTIC shows slightly higher coverage than DeepRoad, which, specifically, is that TACTIC$^{\mathrm{KMNC}}$ achieves on average 1.52% and 0.82% higher coverage, and TACTIC$^{\mathrm{NBC}}$ achieves on average 4.53% and 1.59% higher coverage than DeepRoad, respectively. For Chauffeur, the average coverage achieved by TACTIC is lower than DeepRoad, however, the difference is very small, which is an average of 0.71% between TACTIC$^{\mathrm{KMNC}}$ and DeepRoad, and 1.05% between TACTIC$^{\mathrm{NBC}}$ and DeepRoad.

Table 7. Average results of Using Random Styles ($\mathbf{R_c}$)

Model	Env. Type	Coverage		Number of Errors			
		KMNC	NBC	10°	20°	30°	40°
Dave-orig	Night	43.55%	3.18%	2971.2	269.7	25.5	1.9
	Sunshine	43.34%	1.97%	1300.5	103.8	4.7	0.0
	Rain	42.04%	2.00%	1484.5	44.9	0.1	0.0
	Snow in Daytime	47.15%	3.88%	3605.2	197.7	10.6	0.8
	Snow in Night	52.04%	10.10%	7323.4	2583.0	749.7	72.5
Dave-dropout	Night	36.70%	8.90%	965.2	42.1	3.1	0.0
	Sunshine	39.62%	12.51%	635.8	57.0	3.4	0.0
	Rain	33.07%	8.18%	878.6	66.4	4.0	0.0
	Snow in Daytime	37.83%	10.15%	701.3	57.3	5.5	0.0
	Snow in Night	31.68%	7.85%	2260.6	61.7	5.0	0.0
Chauffeur	Night	73.01%	20.21%	345.1	18.3	0.0	0.0
	Sunshine	68.28%	13.99%	294.5	11.2	0.0	0.0
	Rain	62.16%	10.18%	1048.2	163.2	0.0	0.0
	Snow in Daytime	63.05%	8.10%	3126.7	178.9	7.1	0.1
	Snow in Night	67.44%	16.77%	665.8	6.4	0.0	0.0

The reason is that the environmental condition to transform each testing driving scene generated by DeepRoad is independent of each other, whereas TACTIC uses one critical style vector corresponding to the same environmental condition to transform all the driving scenes. Consequently, the driving scenes generated by DeepRoad are under independent environmental conditions, and therefore, may result in higher coverage in some cases. This is especially the case for Chauffeur, since its CNN is only used as a feature extractor (cf. Appendix B), rendering the test coverage of Chauffeur more sensitive to diverse environmental conditions. Nevertheless, it is more important to study the critical environmental conditions instead of the various random ones, since most environmental conditions can be well handled by these popular DNN-based ADSs, as confirmed by the experimental results.

C.2. Ablation Study

To justify the selection of (1+1) ES in TACTIC, we implement a new version of TACTIC by replacing (1+1) ES with the random search (RS) and compare the effectiveness of the two versions. Except for the search algorithm, the other implementation and settings, e.g., the fitness function, are not changed. We denote TACTIC using RS with KMNC-guided as RS^{KMNC} and the one with NBC-guided as RS^{NBC} in the experiments. For each experimental case, we also conduct 10 runs and average the results to reduce the randomness of the random search.

Table 13 and Table 14 summarise the average results achieved by RS^{KMNC} and RS^{NBC} over 10 repeated runs for each experimental case, respectively. We discuss the comparison results from two aspects, i.e., the achieved test coverage and the number of detected erroneous behaviours. In terms of the achieved coverage, for Dave-orig, Dave-dropout, and Chauffeur, $\text{TACTIC}^{\text{KMNC}}$ achieves on average 7.34%, 3.88%, and 4.35% higher coverage than RS^{KMNC}, respectively, and $\text{TACTIC}^{\text{NBC}}$ achieves on average 9.98%, 4.21%, and 4.14% higher coverage than RS^{NBC}, respectively. In terms of the detected erroneous behaviours, in total, $\text{TACTIC}^{\text{KMNC}}$ detects on average 105.73%, 428.12%, and 120.24% more erroneous behaviours than RS^{KMNC}, and $\text{TACTIC}^{\text{NBC}}$ detects on average 62.98%, 619.51%, and 155.80% more erroneous behaviours than RS^{NBC} for Dave-orig, Dave-dropout, and Chauffeur, respectively. Additionally, we also find that the difference between the number of erroneous behaviours detected by the two approaches increases with the error bound. For example, when setting the error bound to 40°, for Dave-dropout, RS^{KMNC} and RS^{NBC} only respectively detect on average 0.1 and 0.2 erroneous behaviours in the environmental type of "Snow in Daytime", while $\text{TACTIC}^{\text{KMNC}}$ and $\text{TACTIC}^{\text{NBC}}$ respectively detect on average 1966.2 and 930 erroneous behaviours.

Furthermore, Table 15 and Table 16 respectively report the statistical test results comparing TACTIC using (1+1) ES and TACTIC using RS with two coverage-guiding strategies. From these tables, we can observe that: (1) In terms of the achieved coverage, $\text{TACTIC}^{\text{KMNC}}$ achieves significantly higher coverage than RS^{KMNC} for 73.33% (11 cases) of the 15 cases, and

TACTIC$^{\text{NBC}}$ achieves significantly higher coverage than RS$^{\text{NBC}}$ for 80% (12 cases) of the 15 cases, by having $\hat{A}_{12}$ greater than 0.7 and p-value less than 0.04. (2) In terms of the detected erroneous behaviours, TACTIC$^{\text{KMNC}}$ detects significantly more erroneous behaviours than RS$^{\text{KMNC}}$ for 91.67% (55 cases) of the 60 cases (4 error bounds and each bound has 15 experimental cases), and TACTIC$^{\text{NBC}}$ detects significantly more erroneous behaviours than RS$^{\text{NBC}}$ for 93.33% (56 cases) of the 60 cases, by having $\hat{A}_{12}$ greater than 0.76 and p-value less than 0.02. For the few exceptional cases, both TACTIC using (1+1) ES and using RS detect few erroneous behaviours.

Therefore, we can conclude that the critical environmental conditions produced by TACTIC using (1+1)ES can achieve higher test coverage and detect more serious erroneous behaviours than those produced by TACTIC using RS.

Table 8. Statistical test results for TACTIC^{KMNC} and R_c.

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00085	0.92	0.36667	0.45	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00008	1	0.22433	0.395	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00008	1	0.00004	1	0.00008	1	0.00009	1
	40°	0.00009	1	0.03879	0.65	0.00003	1	0.00005	1	0.00008	1
DAVE-DROPOUT	KMNC	0.00009	1	0.07023	0.7	0.00009	1	0.00021	0.97	0.00009	1
	NBC	0.00009	1	0.10606	0.33	0.00008	1	0.00008	1	0.01835	0.78
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.00007	1	0.00005	1	0.00008	1	0.00008	1
	40°	0.00003	1	0.18406	0.55	0.00003	1	0.00003	1	0.00746	0.75
CHAUFFEUR	KMNC	0.00009	1	0.00363	0.86	0.00008	0	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00009	1	0.00009	0	0.00009	1	0.00009	1
	10°	0.00009	1	0.00043	0.945	0.00009	1	0.00009	1	0.00009	1
	20°	0.00008	1	0.00432	0.85	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00003	1	0.00006	1	0.00003	1
	40°	FAILED		FAILED		0.03893	0.65	0.31317	0.545	FAILED	

* We highlight the cases where TACTIC^{KMNC} is worse than R_c with red.
** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.

Table 9. Statistical test results for TACTIC^{NBC} and R_c

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00008	1	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00008	1	0.00004	1	0.00008	1	0.00009	1
	40°	0.00006	1	0.01742	0.7	0.00003	1	0.00005	1	0.00009	1
DAVE-DROPOUT	KMNC	0.00009	1	0.00009	1	0.00009	1	0.00016	0.98	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00008	1	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00054	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.0088	0.815	0.00005	1	0.00008	1	0.00008	1
	40°	0.00003	1	0.18406	0.55	0.00003	1	0.00003	1	0.00037	0.9
CHAUFFEUR	KMNC	0.00009	1	0.00021	0.97	0.41023	0.535	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00008	1	0.10614	0.67	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00008	1	0.00054	1	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00003	1	0.00006	1	0.00298	0.8
	40°	FAILED		FAILED		0.00298	0.8	0.33505	0.54	FAILED	

* We highlight the cases where TACTIC^{NBC} is worse than R_c with red.
** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.

Table 10. Results of DeepRoad

Model	Env. Type	Coverage		Number of Errors			
		KMNC	NBC	10°	20°	30°	40°
Dave-orig	Night	40.99%	5.99%	613	114	33	9
	Sunshine	40.97%	2.05%	355	40	3	0
	Rain	40.23%	2.05%	487	20	0	0
	Snow in Daytime	45.21%	3.81%	1250	90	2	0
	Snow in Night	55.66%	21.50%	3189	1996	802	173
Dave-dropout	Night	33.06%	9.28%	402	29	1	0
	Sunshine	34.48%	10.31%	201	16	1	0
	Rain	31.97%	8.46%	241	21	1	0
	Snow in Daytime	34.99%	10.24%	182	13	1	0
	Snow in Night	31.29%	8.81%	2324	62	1	0
Chauffeur	Night	69.19%	18.64%	140	6	0	0
	Sunshine	65.08%	11.16%	86	5	0	0
	Rain	62.25%	11.30%	232	35	0	0
	Snow in Daytime	63.64%	9.55%	876	188	23	0
	Snow in Night	68.17%	20.43%	573	57	0	0

Table 11. Results of TACTIC^{KMNC} comparing with DeepRoad

Model	Env. Type	Coverage		Number of Errors			
		KMNC	NBC	10°	20°	30°	40°
Dave-orig	Night	48.78%	13.41%	5398.5	5013.3	4199.0	2704.5
	Sunshine	42.27%	4.23%	1812.5	182.0	7.3	0.0
	Rain	41.75%	3.17%	2631.5	1330.8	389.3	85.5
	Snow in Daytime	40.45%	2.29%	4404.0	1518.8	231.5	30.5
	Snow in Night	55.05%	22.22%	5342.5	4976.8	4492.5	3456.5
Dave-dropout	Night	36.58%	12.14%	4418.5	1268.3	211.3	22.3
	Sunshine	34.35%	9.00%	1813.8	216.8	6.3	0.0
	Rain	35.24%	9.51%	1751.0	393.5	53.0	6.5
	Snow in Daytime	34.46%	9.51%	4221.3	3506.5	1803.8	243.5
	Snow in Night	31.72%	8.63%	5421.8	4690.5	1106.0	1.8
Chauffeur	Night	69.05%	19.47%	380.3	22.8	0.0	0.0
	Sunshine	64.94%	10.94%	109.0	2.8	0.0	0.0
	Rain	61.28%	7.53%	3887.5	2199.0	353.3	0.0
	Snow in Daytime	65.37%	10.86%	4865.3	1340.5	36.5	0.0
	Snow in Night	66.65%	16.27%	2550.0	569.3	8.5	0.0

Table 12. Results of TACTIC^{NBC} comparing with DeepRoad

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	54.59%	21.72%	4885.0	3898.3	2662.5	1620.5
	SUNSHINE	45.37%	6.62%	708.8	41.5	4.8	0.0
	RAIN	40.71%	2.90%	3035.8	1422.5	816.5	304.5
	SNOW IN DAYTIME	41.80%	3.34%	3708.0	820.5	154.3	47.3
	SNOW IN NIGHT	60.03%	26.64%	5041.0	4418.0	3716.3	2843.0
DAVE-DROPOUT	NIGHT	39.97%	14.89%	2116.8	774.8	124.5	5.3
	SUNSHINE	33.93%	11.05%	281.0	14.0	0.8	0.0
	RAIN	34.71%	9.61%	3796.5	1921.8	186.9	28.0
	SNOW IN DAYTIME	35.57%	9.76%	4872.3	4122.5	2688.5	342.3
	SNOW IN NIGHT	30.54%	8.77%	5306.0	2433.5	580.0	11.3
CHAUFFEUR	NIGHT	68.25%	19.95%	370.3	41.3	0.0	0.0
	SUNSHINE	64.80%	11.17%	73.3	5.3	0.0	0.0
	RAIN	61.34%	7.65%	4146.8	2198.3	369.5	2.5
	SNOW IN DAYTIME	63.76%	10.69%	3042.3	160.5	16.8	0.0
	SNOW IN NIGHT	66.37%	14.97%	569.4	59.5	3.2	0.0

Table 13. Average eesults of KMNC-guided TACTIC using Random Search

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	52.48%	10.24%	15471.7	4710.5	411.7	0.0
	SUNSHINE	43.66%	2.26%	2364.9	142.6	9.4	0.1
	RAIN	42.55%	2.06%	3629.9	225.9	5.8	0.2
	SNOW IN DAYTIME	50.48%	5.43%	10496.0	1232.4	48.1	0.8
	SNOW IN NIGHT	64.56%	25.42%	19629.1	15729.0	10319.7	4325.2
DAVE-DROPOUT	NIGHT	38.10%	9.58%	4907.7	328.6	36.4	0.0
	SUNSHINE	38.03%	10.92%	1429.5	47.7	1.3	0.0
	RAIN	34.83%	8.58%	1599.1	135.6	11.0	0.2
	SNOW IN DAYTIME	37.36%	9.99%	2180.0	209.6	12.0	0.1
	SNOW IN NIGHT	32.04%	8.50%	18887.9	4175.9	13.9	0.0
CHAUFFEUR	NIGHT	73.75%	22.29%	1202.5	54.7	0.0	0.0
	SUNSHINE	68.83%	16.51%	407.7	9.8	0.0	0.0
	RAIN	61.76%	8.64%	3616.4	309.5	0.0	0.0
	SNOW IN DAYTIME	65.45%	9.99%	12976.1	600.0	1.6	0.0
	SNOW IN NIGHT	70.25%	20.99%	3475.6	253.1	0.0	0.0

Table 14. Average results of NBC-guided TACTIC using Random Search

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	53.09%	10.73%	15833.3	4783.0	384.0	9.1
	SUNSHINE	45.48%	2.85%	1451.4	81.2	5.0	0.0
	RAIN	42.35%	2.00%	3633.9	188.1	3.8	0.0
	SNOW IN DAYTIME	51.87%	6.80%	9730.0	748.8	20.0	14.3
	SNOW IN NIGHT	65.45%	25.69%	19341.5	14917.3	9955.4	4966.7
DAVE-DROPOUT	NIGHT	38.61%	10.51%	4497.2	340.6	32.6	0.0
	SUNSHINE	42.56%	15.56%	902.4	42.8	1.1	0.0
	RAIN	35.08%	8.94%	1056.7	86.2	1.7	0.0
	SNOW IN DAYTIME	42.17%	13.86%	975.5	111.0	9.9	0.2
	SNOW IN NIGHT	32.38%	8.51%	18267.9	4321.6	8.1	0.0
CHAUFFEUR	NIGHT	75.00%	27.53%	691.8	32.1	0.0	0.0
	SUNSHINE	69.39%	18.39%	296.3	15.2	0.0	0.0
	RAIN	62.29%	11.82%	2397.7	232.7	0.0	0.0
	SNOW IN DAYTIME	66.60%	10.99%	12336.4	337.8	0.0	0.0
	SNOW IN NIGHT	70.40%	23.59%	2461.3	137.2	0.0	0.0

Table 15. Statistical test results for TACTIC^{KMNC} and RS^{KMNC}.

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00455	0.85	0.00009	0	0.00009	1
	NBC	0.00009	1	0.00009	1	0.00007	1	0.00009	0	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	40°	0.00009	1	0.13903	0.605	0.00004	1	0.00005	1	0.00009	1
DAVE-DROPOUT	KMNC	0.00009	1	0.07923	0.7	0.00009	1	0.00454	0.85	0.00009	1
	NBC	0.00009	1	0.10606	0.33	0.00009	1	0.00009	1	0.42505	0.53
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00008	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.00007	1	0.00009	1	0.00009	1	0.00008	1
	40°	0.00009	1	0.18406	0.55	0.00007	1	0.00009	1	0.00746	0.75
CHAUFFEUR	KMNC	0.00009	1	0.04808	0.725	0.00921	0.185	0.00009	1	0.0001	0.99
	NBC	0.00009	1	0.00256	0.875	0.00009	0	0.00009	1	0.00009	1
	10°	0.00029	0.96	0.00628	0.835	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00008	0.92	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00009	1	0.00009	1	0.00009	1
	40°	FAILED		FAILED		0.03893	0.65	0.08403	0.6	FAILED	

* We highlight the cases where TACTIC^{KMNC} is worse than RS^{KMNC} with red.

** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.

Table 16. Statistical test results for TACTIC^{NBC} and RS^{NBC}.

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00009	1	0.00021	0.97	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00009	1	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00008	1	0.00008	1	0.00009	1	0.00009	1
	40°	0.00009	1	0.01742	0.7	0.00003	1	0.00009	1	0.00009	1
DAVE-DROPOUT	KMNC	0.00009	1	0.14486	0.645	0.00009	1	0.23633	0.6	0.00009	1
	NBC	0.00009	1	0.00009	1	0.00009	1	0.33873	0.44	0.43989	0.525
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00012	0.99	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.00014	0.98	0.00007	1	0.00009	1	0.00008	1
	40°	0.00003	1	0.18406	0.55	0.00004	1	0.00009	1	0.00037	0.9
CHAUFFEUR	KMNC	0.00009	1	0.00862	0.82	0.27252	0.415	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00009	1	0.42504	0.53	0.00009	1	0.00009	1
	10°	0.00009	1	0.02459	0.765	0.00009	1	0.00506	0.845	0.03201	0.75
	20°	0.00009	1	0.00018	0.975	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00003	1	0.00003	1	0.00298	0.8
	40°	FAILED		FAILED		0.00298	0.8	0.08374	0.6	FAILED	

* We highlight the cases where TACTIC^{NBC} is worse than RS^{NBC} with red.
** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.