SSBSE 2021

Search-Based Selection and Prioritization of Test Scenarios for Autonomous Driving Systems

Date: 2021-10-11

Presenter: Chengjie Lu

Chengjie Lu[1], Huihui Zhang[2], Tao Yue[1,3], Shaukat Ali[3]

[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China

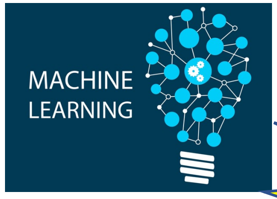
[2]Weifang University, Weifang, China

[3]Simula Research Laboratory, Oslo, Norway

CATALOGUE

Part 1
Motivation

Current Stage of Autonomous Driving Systems



Machine Learning

Catastrophic consequences

Planning & Routing

Environmental Perception

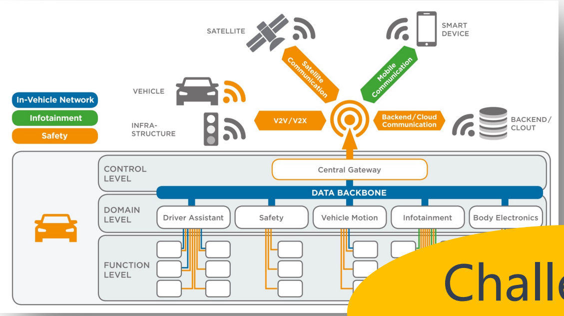
......

How to ensure the **safety** of ADSs?

Testing ADSs

Challenges of Testing ADSs

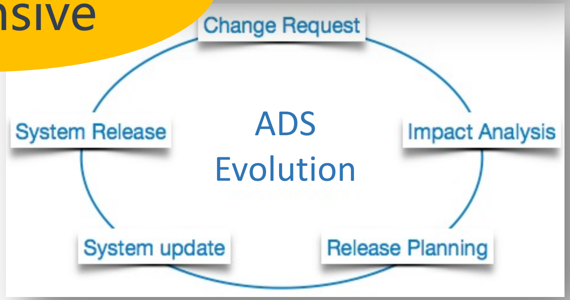
High complexity and heterogeneity of ADSs



High complexity and uncertainty of the operating environment

Challenging
Expensive

Testing an ADS is challenging and expensive.



ADSs evolve constantly with new functionalities introduced rapidly.

Testing multiple versions of ADSs is very expensive!

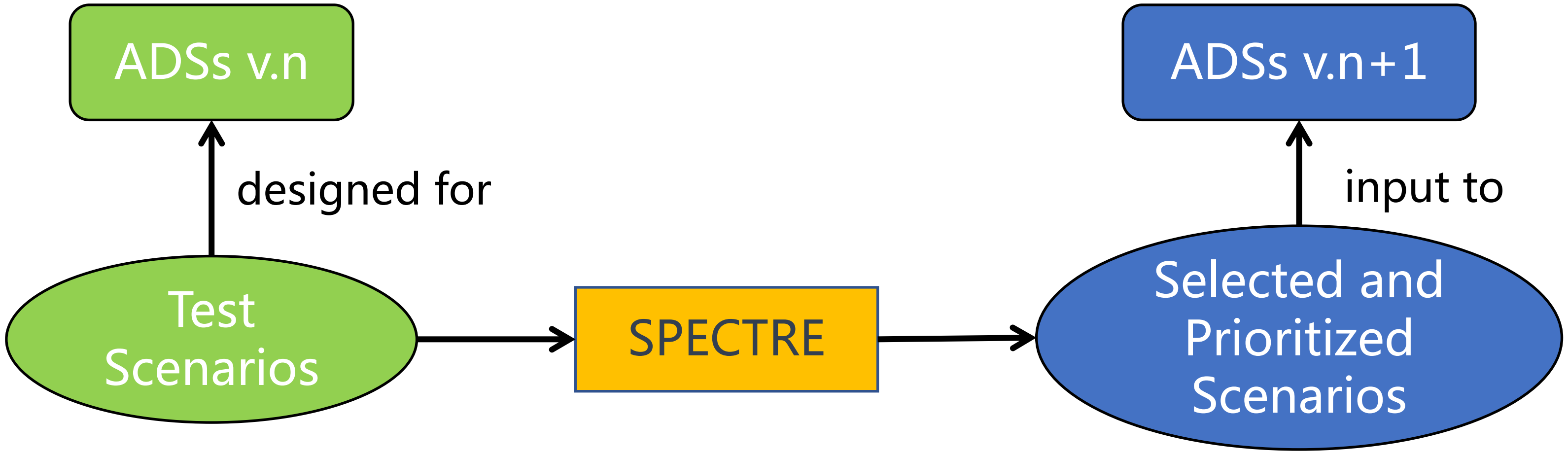
Motivation and Our Approach - SPECTRE

Motivation

Optimizing tests of ADSs, especially when testing a new version of an ADS.

SPECTRE

A search-based approach for Selecting and Prioritizing of test scenarios to test a new version of an ADS.



State-of-the-Art

Search-Based Testing of ADSs

NSGAII-SM: Identify critical behaviors of pedestrian detection vision based systems. {Ben Abdessalem, et al. 2016}

NSGAII-DT: Generate critical test scenarios for vision-based control systems. {Ben Abdessalem, et al. 2018}

AV-FUZZER: Generate AV safety violation scenarios. {Li et al. 2020}

Literature: **Generating** test scenarios.

SPECTRE: **Selecting and prioritizing** test scenarios.

Search-Based Test Case Prioritization

Test prioritization for regression testing {Li et al. 2007}, {Singh et al. 2010}, …

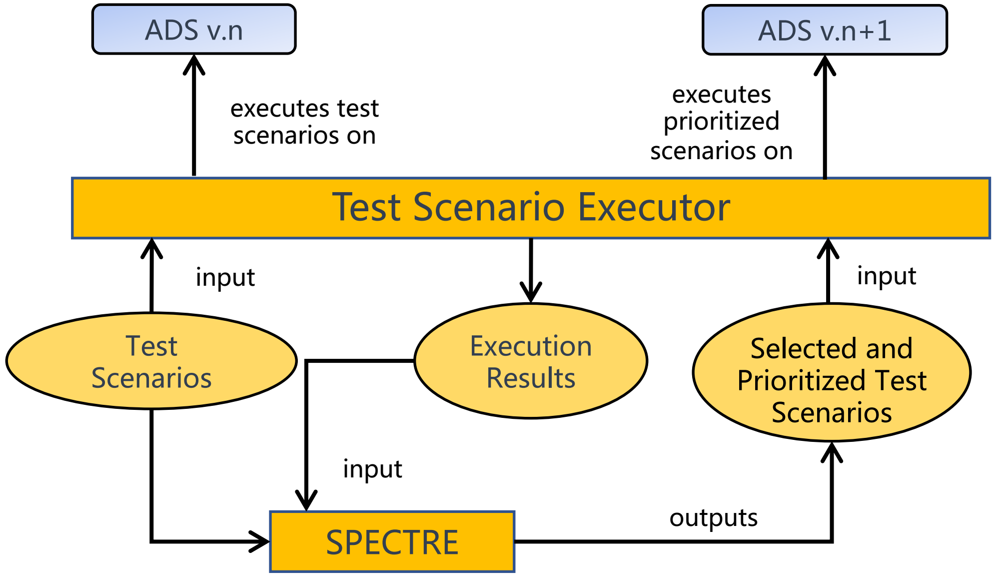
Literature: None of them studies on test prioritization for ADSs.

SPECTRE: Test scenario prioritization for ADSs

PART 2

Our approach: SPECTRE

SPECTRE's Overall Architecture



Execution Results (Attributes)

- Whether a collision occurred with the scenario
- Collision probability associated with the scenario
- The extend of demand on the ADS put by the scenario
- Diversity of the scenario as compared to others

Optimization Objectives

Problem Representation —Test Scenarios

Test Scenarios

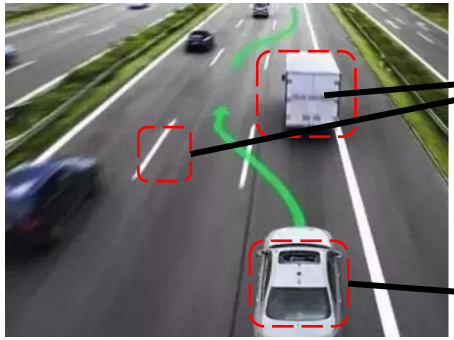
$SS = \{S_1, S_2, \dots, S_{ns}\} \xrightarrow{\text{Designed for}}$ ADS v.n

Selected and Prioritized Test Scenarios

$TS = \{S_1, S_2, \dots, S_{nts}\} \xrightarrow{\text{Testing}}$ ADS v.n+1

Properties specifying scenarios

$Properties = \{p_1, p_2, \dots, p_{np}\}$



Lane Changing

Environment
Weather
Road User
……

Ego Vehicle
Acceleration
Speed
……

Test Scenario Executor

Execution Results

Problem Representation — Attributes

Attribute-1 (Collision (COL))

$$COL \in \{True, False\}$$

Collision Scenario (S_{COL}): COL is True

Non-Collision Scenario (S_{NCOL}): COL is False

Attribute-2 (Collision Probability (CPT))

$$CPT = \begin{cases} \frac{SD - CD}{SD}, & CD < SD \\ 0.0, & else \end{cases}$$

$$SD = Fun_{SD}(\alpha_{ego}, \alpha_{obstacle}, v_{ego}, v_{obstacle})$$

$$CD = Fun_{CD}(Pos_{ego}, Pos_{obstacle})$$

Potential Collision Scenario (S_{PCOL}): $CPT \in (0, 1)$

Problem Representation — Attributes

Attribute-3 (Demand[1] (DEM))

UNIVERSITY OF WATERLOO WISE DRIVE

Measure how much difficulty the generated scenarios put the ego vehicle in.

Based on

Properties

Environment
Weather
Road User

Ego Vehicle
Acceleration
Speed

Concretely,

$\mathrm{p1} \rightarrow \mathrm{p1}_{Demand}, \mathrm{p2} \rightarrow \mathrm{p2}_{Demand}, \ldots\ldots, \mathrm{p}n \rightarrow \mathrm{p}n_{Demand}$

$Rain_{Demand}: 0(no), 1\ (light),\ \ 2\ (moderate), 3\ (heavy)$

High Demand property (P_{HighD}):
demand value ≥ the medium value

High Demand Scenario (S_{HighD}):
$\#P_{\mathrm{HighD}}$ > #property / 2

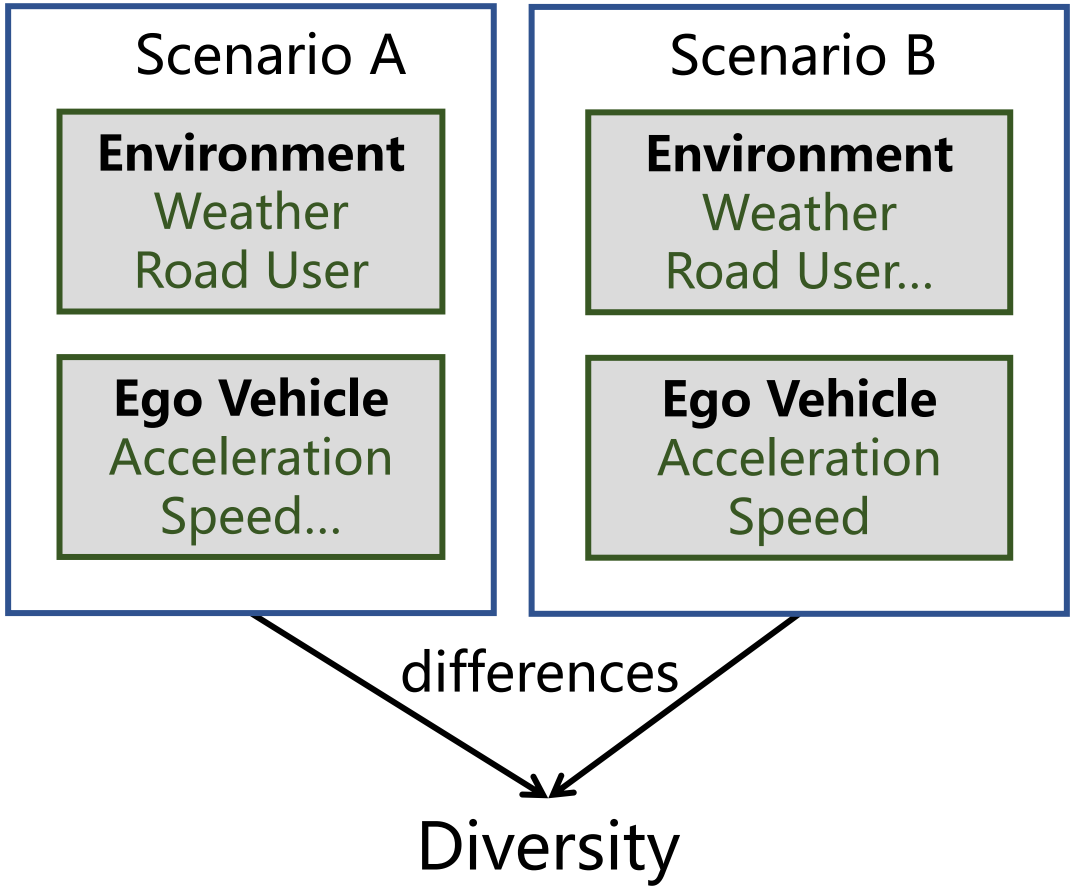
[1] Czarnecki, K.: Operational design domain for automated driving systems: Taxonomy of basic terms. Waterloo Intelligent Systems Engineering (WISE) Lab, University of Waterloo, Canada (2018)

Problem Representation — Attributes

Attribute-4 (Diversity (DIV))

Measure the **differences** among scenarios in a test suite.

Calculate based on the **differences of properties** of different scenarios.



Problem Representation — Optimization Problem

$SS = \{S_1, S_2, \dots, S_{ns}\}$

Desired Budget: Number of Test Scenarios (NTS)

Select a Test Suite (TS) with a particular permutation X

A selected and prioritized test suite TS has:

- the **maximum** number of collision scenarios,
- the **maximum** number of potential collision scenarios,
- the **maximum** number of high demand scenarios, and
- the **most diverse** scenarios,

Values of the four attributes of the test scenarios in TS will **descend**.

Optimization Objectives

PART 3
Evaluation

Evaluation — DataSet

Test setup

Simulator: LGSVL 5.0

ADS: Apollo 5.0

Test execution: nearly 1000 times

Data preprocessing
Remove duplication
Calculating diversity
…

90K Test Scenarios

Calculate Attributes

60K Test Scenarios

Evaluation — Experiment Design

Multi Objective Evolutionary Algorithms (MOEAs)

SPECTRE was integrated with 5 MOEAs:

NSGA-II, NSGA-III, IBEA, SPEA2 and *MOCell*

Parameters Settings

MOEAs used the default hyper-parameter settings from JMetal
NTS: 1000, 2000, …, 8000;
19properties: 5 vehicle properties, 14 environmental properties

Execution

SPECTRE was executed 30 times for each MOEA with each NTS.

Random Search (RS) was used for sanity check.

All the MOEAs performed significantly better than RS

Evaluation — Research Questions

RQ1 Comparisons of different MOEAs of merged search budget

How do the selected MOEAs compare to each other in terms of solving our optimization problem?

RQ2 Comparisons of different MOEAs of various search budgets

How the selected MOEAs compare to each other when solving optimization problems of various search budgets?

RQ3 Impact of search budgets on MOEAs

How does the search budget affect the effectiveness of the selected MOEAs?

RQ4 Time performance of MOEAs

How is the time performance of the selected MOEAs?

Evaluation — Evaluation Metric & Statistical Test

Quality Indicator

A smaller IGD value is better.

Inverted Generational Distance (IGD)

Statistical Test

Performed Mann-Whitney U test to test the significance of the differences and computed the Â12 statistics as the effect size. (**RQ1, RQ2**)

Performed the Spearman’**s** rank correlation (ρ) test to study the significance of correlation. (**RQ3**)

Evaluation — Result of RQ1

RQ1: Comparison of different MOEAs

- Statistical Results for comparing MOEAs when combining all NTS results

Metric	IBEA *vs.*				NSGA-II *vs.*			NSGA-III *vs.*		MOCell *vs.*
	NSGA-II	NSGA-III	MOCell	SPEA2	NSGA-III	MOCell	SPEA2	MOCell	SPEA2	SPEA2
$\hat{A}_{12}$	**0.121**	**0.287**	**0.008**	**0.207**	0.754	**0.044**	0.678	**0.009**	0.399	0.973
p-value	**<0.05**	**<0.05**	**<0.05**	**<0.05**	<0.05	**<0.05**	<0.05	**<0.05**	0.183	<0.05

IBEA is significantly better than the other MOEAs.

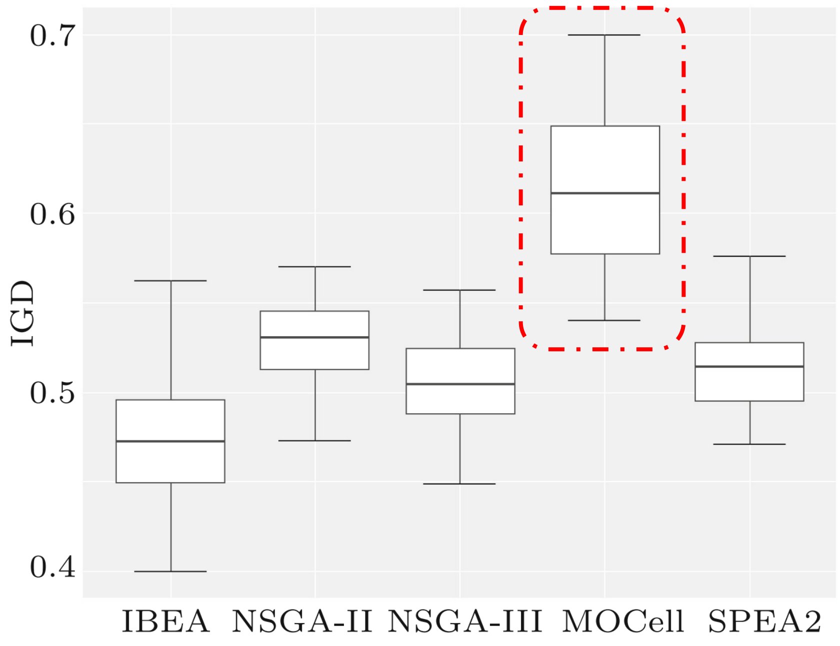
MOCell is significantly worst than the rest.

Ranking: IBEA, NSGA-III/SPEA2, NSGA-II, MOCell.

Evaluation — Result of RQ1

RQ1: Comparison of different MOEAs

- Descriptive statistics of IGD when combining all NTS results



A smaller IGD value is better.

MOCell performed the worst and also produced results with the largest variance.

The variances of IGD values of the other four MOEAs are smaller and comparable.

IBEA is recommended for solving our search problem!

Evaluation — Result of RQ2

RQ2: Pair-wise comparisons of MOEAs of various search budgets

- Results of comparing MOEAs for each NTS

NTS	IBEA *vs.*				NSGA-II *vs.*			NSGA-III *vs.*		MOCell *vs.*
	NSGA-II	NSGA-III	MOCell	SPEA2	NSGA-III	MOCell	SPEA2	MOCell	SPEA2	SPEA2
1000	**.027/<.05**	**.249/<.05**	**.009/<.05**	**.098/<.05**	.834/<.05	**.032/<.05**	.653/<.05	**.018/<.05**	**.289/<.05**	.972/<.05
2000	**.040/<.05**	**.239/<.05**	**.013/<.05**	**.052/<.05**	.826/<.05	**.121/<.05**	.629/.08	**.039/<.05**	**.227/<.05**	.922/<.05
3000	**.044/<.05**	**.258/<.05**	**.007/<.05**	**.062/<.05**	.904/<.05	**.128/<.05**	.707/<.05	**.016/<.05**	**.182/<.05**	.943/<.05
4000	**.061/<.05**	**.271/<.05**	**.029/<.05**	**.096/<.05**	.851/<.05	**.177/<.05**	.578/.304	**.071/<.05**	**.210/<.05**	.858/<.05
5000	**.068/<.05**	**.206/<.05**	**.002/<.05**	**.146/<.05**	.770/<.05	**.036/<.05**	.707/<.05	**.009/<.05**	.410/.234	.984/<.05
6000	**.038/<.05**	**.258/<.05**	**.014/<.05**	**.088/<.05**	.826/<.05	**.116/<.05**	.599/.191	**.046/<.05**	**.267/<.05**	.898/<.05
7000	**.100/<.05**	**.257/<.05**	**.004/<.05**	**.116/<.05**	.769/<.05	**.044/<.05**	.559/.438	**.023/<.05**	**.264/<.05**	.963/<.05
8000	**.144/<.05**	.359/.061	**.001/<.05**	**.204/<.05**	.808/<.05	**.060/<.05**	.636/.072	**.003/<.05**	**.289/<.05**	.976/<.05

Evaluation — Result of RQ2

RQ2: Pair-wise comparisons of MOEAs of various search budgets

- Ranking of MOEAs for each NTS value

NTS	Ranking	NTS	Ranking	NTS	Ranking	NTS	Ranking
1000	I, N-III, S, N-II, M	2000	I, N-III, S/N-II, M	3000	I, N-III, S, N-II, M	4000	I, N-III, S/N-II, M
5000	I, N-III/S, N-II, M	6000	I, N-III, S/N-II, M	7000	I, N-III, S/N-II, M	8000	I/N-III, S/N-II, M

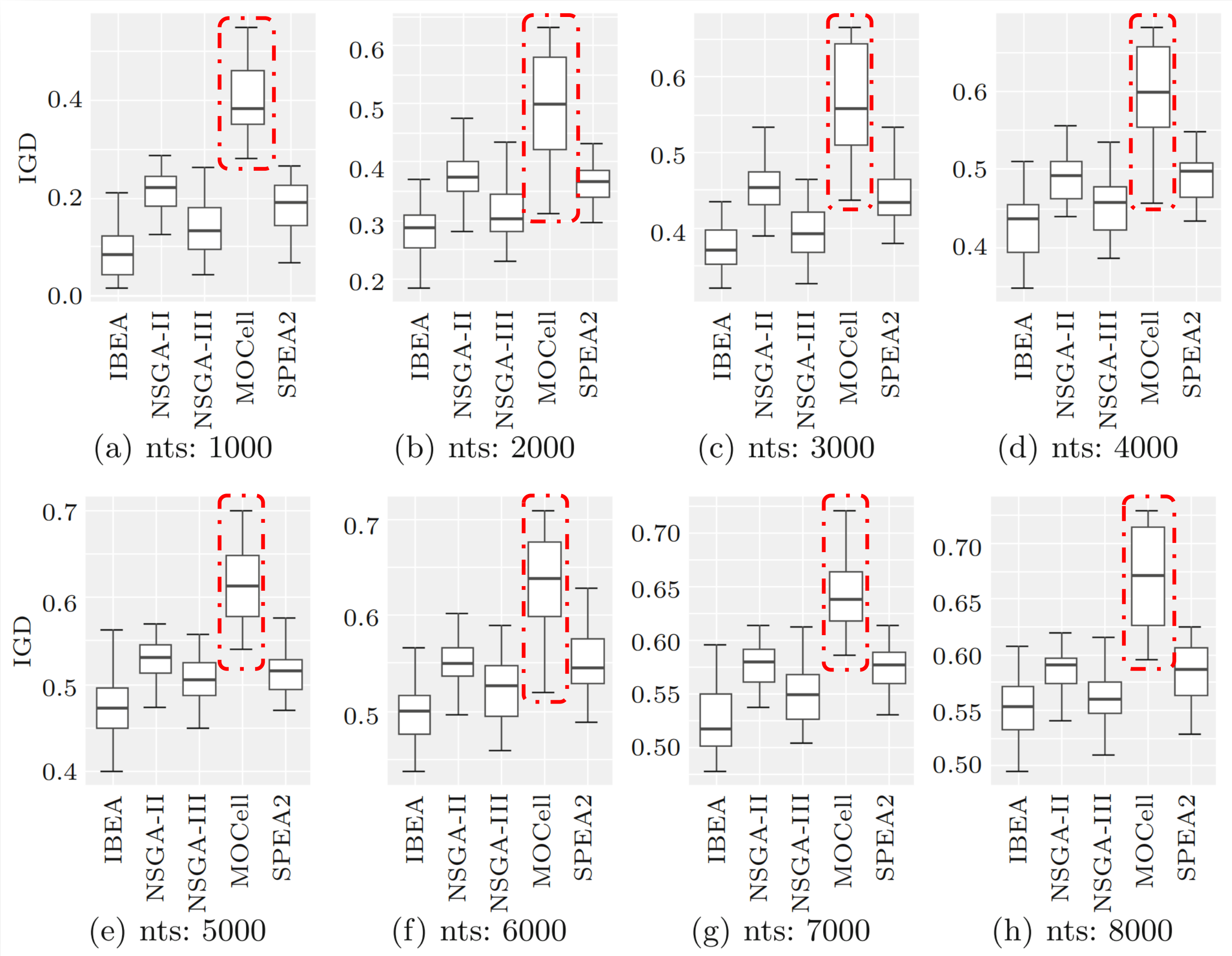
* **I**: IBEA, **N**: NSGA; **S**: SPEA2. **M**: MOCell; a/means two MOEAs have the same ranking.

IBEA (I) is significantly better than the other MOEAs.
MOCell (M) is significantly worst than the rest.

Evaluation — Result of RQ2

RQ2: Pair-wise comparisons of MOEAs of various search budgets

- Descriptive statistics of IGD in terms of various NTS



MOCell is the worst with the large variances.

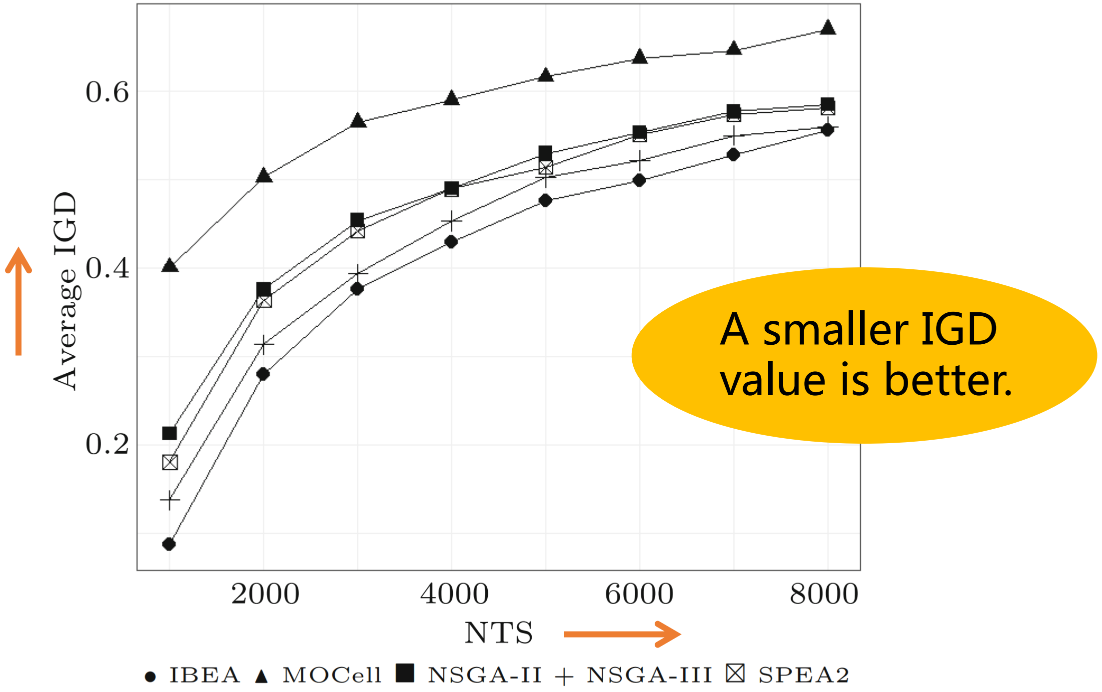
For the other MOEAs, we can observe smaller variances and they are comparable for most of NTS values.

Results are consistent with the results obtained for RQ1: IBEA is the best.

Evaluation — Result of RQ3

RQ3: Impact of search budgets on MOEAs

Results of IGD of various MOEAs when increasing NST



A larger NTS leads to a more challenging problem to solve.

Results of the Spearman's rank correlation test

Metric	MOSA				
	IBEA	NSGA-II	NSGA-III	MOCell	SPEA2
ρ	0.936	0.933	0.930	0.713	0.930
p-value	<0.05	<0.05	<0.05	<0.05	<0.05

There is a near perfect positive correlation between IGD and NTS, for MOCell, ρ=0.713 indicates a strong positive correlation.

The ability of the MOEAs producing high-quality solutions significantly decreases with the increase of NTS.

Evaluation — Result of RQ4

RQ4: Time performance of MOEAs

Average running time of each MOEA (Time Unit: Minute)

MOSA	NTS							
	1000	2000	3000	4000	5000	6000	7000	8000
IBEA	17.58	32.09	47.05	62.42	77.03	98.61	110.15	126.01
NSGA-II	16.19	30.89	46.14	61.87	76.82	97.35	108.92	122.09
NSGA-III	16.21	31.04	46.34	61.41	77.12	97.39	109.19	125.16
MOCell	17.36	33.01	49.77	66.52	82.22	106.89	118.56	136.68
SPEA2	16.81	31.10	46.23	61.81	77.73	97.93	111.35	126.14
Random	14.82	28.56	42.98	56.27	72.14	91.51	102.79	117.68

The time performance is practically **acceptable.**

A MOEA needs nearly 17 to 137 mins to solve the optimization problems of different complexity (i.e., NTS).

There are not much time differences among the studied MOEAs of each NTS that practically matter.

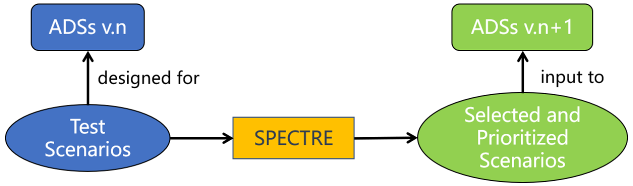
PART 4
Conclusion and Future Work

Conclusion

Challenging

Testing ADSs

It is important to optimize tests for ADSs, especially when **testing multiple versions of ADSs**.



SPECTRE, a multi-objective search approach for test scenario **selection and prioritization** for ADSs.

Evaluation

SPECTRE was integrated with five MOEAs, and evaluated on a large-scale dataset.

The evaluation results showed that **IBEA** performed the best in terms of producing high quality solutions, and therefore is recommended for addressing our optimization problem.

Future Work

Integrate with other ADSs

Continuously update attribute values

Parameter tuning for MOEAs

Search-Based Selection and Prioritization of Test Scenarios for Autonomous Driving Systems

Date: 2021-10-11

Presenter: Chengjie Lu

Chengjie Lu[1], Huihui Zhang[2], Tao Yue[1,3], Shaukat Ali[3]

[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China

[2]Weifang University, Weifang, China

[3]Simula Research Laboratory, Oslo, Norway