PRISSMA Project
Plateforme de Recherche et d'Investissement pour la Sûreté
et la Sécurité de la Mobilité Autonome
04/2021 - 04/2024

[L3.6] CARRYING OUT THE SECOND EVALUATION TEST CAMPAIGN IN A CONTROLLED ENVIRONMENT AND PRODUCING TEST REPORTS.

[L3.6] RÉALISATION DE LA DEUXIÈME CAMPAGNE D'ESSAIS D'ÉVALUATION EN ENVIRONNEMENT CONTRÔLÉ ET PRODUCTION DES RAPPORTS D'ESSAIS

Main authors: A. Piperno (UTAC) , M. Dessaint (UTAC) , V. Srikanthan (UTAC), C. Serbouh (UTAC), Rémi Régnier (LNE), Mohamed Boudali (LNE), Pierre Duthon (Cerema), Frédéric Bernardin (Cerema), Jean-Baptiste Horel (INRIA), Khushdeep Singh Mann (INRIA), Alessandro Renzaglia (INRIA), Christian Laugier (INRIA), Elodie Chateauroux (Transpolis), Nicolas Merle (Transpolis), Yoan Nouet (Transpolis), Youssoufa Tahirou (Valeo), Richard Denis (Valeo), Oussama El Hamzaoui (Valeo), Ali Seba (IGN), Elias Ababou (IGN)

Keywords: Proof of concept, AI based vehicles homologation, repeatability tests, robustness tests, AI overfitting tests, anticipation tests, closed road tests, VIL, HIL, augmented reality,

Abstract. This deliverable is intended to provide the test reports from the second test campaign, which was carried out during the first quarter of 2023. This second campaign must validate the choices made following the first. During the second campaign, 5 POCs were carried out, one on the use of augmented reality as a means of testing on tracks between INRIA and Transpolis, another on tests in degraded conditions on benches between CEREMA and LNE, two carried out by UTAC and Transpolis on the generalisation of current tests to better include AI aspects, and finally one proposed by Valeo and IGN to evaluate the performance of a localization system for automated vehicles.

Résumé. Ce livrable est destiné à fournir les rapports d'essais de la deuxième campagne d'essais, réalisée au cours du premier trimestre 2023. Cette seconde campagne doit valider les choix effectués suite à la première. Lors de la seconde campagne, 5 POC ont été réalisés, l'un sur l'utilisation de la réalité augmentée comme moyen de test sur piste entre l'INRIA et Transpolis, un autre sur des tests en conditions dégradées sur bancs entre le CEREMA et le LNE, deux réalisés par l'UTAC et Transpolis sur la généralisation des tests actuels pour mieux inclure les aspects IA, et enfin un proposé par Valeo et l'IGN pour évaluer les performances d'un système de localisation pour les véhicules automatisés.

Table des matières

Chapter 1: UTAC homologation test reports

AI based vehicles could have some safety weak points regarding repeatability, robustness, anticipation and overfitting for official known tests. Therefore, UTAC PRISSMA WP3 team has built first answers and proposals to adapt or to create new homologation tests scenarios / protocols / testing tools / evaluation metrics during the WP3 POC number 1 tests in UTAC from February to July 2023. UTAC also proposed in January 2024 in L3.3 final proposals for tests, protocols and evaluation metrics. The 2024 PRISSMA WP3 POC number 2 aimed to confirm or fine-tune these proposals.

We only tested in 2024 the prototype ZOE NEXYAD « DREAMotorONE » research prototype with AI based anticipation, because it contains more IA and intelligence, and because its perception has been enhanced beginning 2024 allowing interesting tests more to fine-tune our conclusions and proposals for robustness and anticipation tests, in the four most interesting scenarios among new scenarios built in POC 1 in 2023.

Finally, these more complete and challenging tests, on a very intelligent vehicle, confirmed our January 2024 proposals (L3.3) for tests, protocols and evaluation/homologation metrics. We redo the 2023 scenarios with the 2024 vehicle settings, and 2024 vehicle reactions which are different and rather better in 2024 compared to 2023 ; but 2023 conclusions and proposals for requirements and evaluation are still the same and so are confirmed.
We also use this 2024 POC number 2 to build a method to identify and converge rapidly towards vehicle ODD limits (in order to make useful and difficult validation tests at the limits of ODD)
An A4 format poster at the end of this document summaries the whole UTAC WP3 PRISSMA results & proposals.

As previously explained in L3.3 in January 2024, today GRVA regulation group discussions are not very advanced and target to evaluate if existing or soon existing regulations could be sufficient to verify AI-based vehicles safety: Complex systems safety audit annexes , EU AI act, UN-ECE software update and cyber regulations (UN-ECE R155 & 156 regulations). These regulations mainly require audits (of AI and software development, validations, production, reparations, data management, and safety assessment for robustness and black box assessment) but not additional & standardized tests to evaluate vehicles on testing tracks. For example, overfitting tests and requirements are just an idea initially discussed in GRVA regulation group, but today neither Regulation nor Euro NCAP working groups work on such precise regulations with protocols and requirements for AI-based vehicles. So, we build proposals of scenarios, tests and requirements, with simple "KPI" basic metrics, coherent with today homologation metrics; And these proposals still remain to be presented and discussed to the regulation groups with States and OEMs and discussions could take a long time.

1. TEST PROGRAM

1.1. PROTOCOL VERSION

The following scenarios refer mainly to the ENCAP 2023 protocol for the geometry, the corridors, and the test speeds. Then, some variants of known scenario have been created for this study.

1.2. TESTS DESCRIPTION

Refer to deliverables **L3.5 and L3.3** shared in July 2023 and January 2024,

For these 2024 POC number 2 tests we decided to only test the prototype ZOE NEXYAD « DREAMotorONE » research prototype with AI based anticipation, because it contains more IA and intelligence, and because its perception has been enhanced beginning 2024 allowing interesting tests more to fine-tune our conclusions and proposals for robustness and anticipation tests, in the four most interesting scenarios among new scenarios built in POC 1 in 2023

The scenarios are in two categories:

- Robustness : stationary vehicle on emergency lane on highway
  This category allows to evaluate the robustness of the systems, it means to perform a specific scenario and change different parameters (Speeds, Angles, visual aspect…) and see the behavior.

- Pre-critical : Car to pedestrian (CPFA), object on lane (K16), stopped right lane on highway
  This category allows evaluating the anticipation of the systems on existing scenarios or new ones.

2. VEHICLE UNDER TEST

One vehicle has been tested during this campaign.

A first step of subjective testing has been done to have a first idea on the behavior of each vehicle.

After that, we can select which tests are relevant to perform more precisely with measurement system. The tests performed and the results are detailed in 5.3.

2.1. DESCRIPTION

ZOE prototype from French start-up NEXYAD with intelligent and anticipatory driving based on AI.

Two new innovative and intelligent proactive functions developed by NEXYAD:

Thus, NEXYAD has developed (and patented) two new functions of intelligent driving, and is in discussions with many French, German and Japanese manufacturers to market them: these driving functions use the estimation of the risk of NEXYAD and the consequently relevant safe speed to have to minimize the risk and stay in the green zone of driving risks (previous figure).

The risk is estimated according to the road map (arrival on a steep curve, a tight crossroads) and also what the vehicle sensors see (vehicle poorly parked, crowded crossroads, low field of vision

NEXYAD's two innovative and intelligent proactive driving functions are:

- **A safety assistant (named "safety coach") who alerts** the driver when his driving behavior is no longer prudent (risk to high) in relation to the driving context (accident reduction estimated by NEXYAD of at least 25%).

- **An intelligent and proactive ACC** that automatically regulates the vehicle speed according to the driving context (up to 75% accident reduction according to NEXYAD)

NB the difference in the result between 25% reduction of accident in alert mode compared to 75% of the intelligent ACC mode is explained by the fact that the driver may not immediately and always consider the warnings and not slow down.
These two new functions are being implemented on a prototype vehicle, the NEXYAD DREAMOTOR1, **see photos below**, which is therefore one of the most advanced prototype in the world (On PRISSMA there is no French actor among vehicle manufacturers and it’s difficult to know their skills and developments on these very upstream and very competitive subjects). NEXYAD is part of the French industrial research and development group of the Regions Normandy/Ile de France, called NEXTMOVE (previously MOVEO), which supported and facilitated these innovative projects.

The perception system of the vehicle has been enhanced beginning 2024 allowing interesting tests more to fine-tune our conclusions and proposals for robustness and anticipation tests, in the four most interesting scenarios among new scenarios built in POC 1 in 2023.

Anticipation functions interesting, towards more anticipation and less emergency maneuver. For instance, maximum deceleration braking of the car has been limited to 0,3 mS-2, against 1 ms-2 for usual vehicles , because the most used function is ACC with anticipation and not emergency automated braking .

The four more interesting scenarios we tested are:

Stationary car on emergency lane:

we redo our 2023 tests and protocols but with more and different influent factors enabling to reach vehicle limits (which were quite different than 2023 limits, so interesting to confirm of affine our requirements)

- stationary car overlap but also angle, see photo here below
- mud on the lidar glass, making perception more difficult (see photo here below)
- pedestrian standing near the stationary car on emergency lane
- longitudinal distance for obstacle / car perception
- accepted level of risk

These tests allow us to verify on a vehicle with different settings & different influents variables that our robustness tests and protocols and requirements are still relevant ; .



Angle & overlap for stationary Vehicle tests mud on LIDAR glass pedestrian near vehicle

Object on road (K16):

We redo this 2023 scenario with the 2024 vehicle settings, and 2024 vehicle reactions which are different and rather better in 2024 compared to 2023 ; but 2023 conclusions and proposals for requirements and evaluation are still the same and so are confirmed.

Car to pedestrian (CPFA):

We redo this 2023 scenario with the 2024 vehicle settings, and 2024 vehicle reactions which are different and rather better in 2024 compared to 2023 ; but 2023 conclusions and proposals for requirements and evaluation are still the same and so are confirmed.

Stationary right lane on highway:

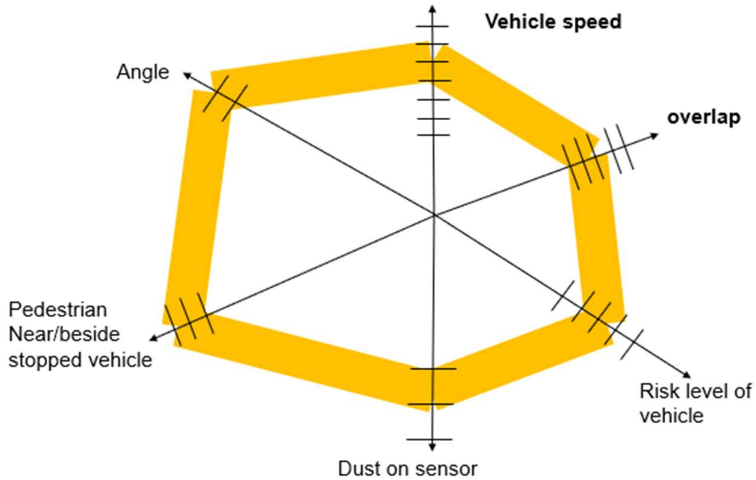
We redo this 2023 scenario with the 2024 vehicle settings, and 2024 vehicle reactions which are different and rather better in 2024 compared to 2023 ; but 2023 conclusions and proposals for requirements and evaluation are still the same and so are confirmed.

We also use this 2024 POC number 2 to build a method to identify and converge rapidly towards vehicle ODD limits (in order to make useful and difficult validation tests at the limits of ODD).

See figure below,

- we progressively Identified relevant factors of performance for each test scenario and limits for each factors ;
- Finally we focused our tests around these limit values.

- For example, it is useless to make many validation/verification tests at a speed of 30 km/h if the ODD limit is 70 km/h: tests at 60 km/h to 80 km/h are much more interesting.



ODD limits in yellow values for testing scenario factors

3. TESTING EQUIPMENT

3.1. MOTION MEASUREMENT

MOTION PACK 1	
Manufacturer Oxford Technical Solutions (OxTS)	
Unit model	
Sensors Accelerometers (Servo) / Gyros (MEMS)	
Data output rate 100 Hz	*Coupling method* GNSS / INS

3.2. DRIVING CONTROL SYSTEM

CONTROLLER	
Manufacturer Antony Best Dynamics (ABD)	
Unit model XR Omni	
Sampling rate 100 Hz	
Analog input voltage ± 10 V	*A / D conversion* 16 bits

3.3. HMI ANALYSIS

VIDEO VBOX	
	*Manufacturer* Racelogic

GOPRO	
	*Manufacturer* GoPro

3.4. ADDITIONAL EQUIPMENT

POWERMESH ANTENNA		
	*Manufacturer* Antony Best Dynamics (ABD)	
	Unit model TrackFi PowerMesh	*Communication* Wifi 5 GHZ / 2.4 GHz
	UTAC unit reference	

BASE STATION		
	*Manufacturer* Oxford Technical Solutions	*Unit model* GPS-Base-2G GLONASS
	Correction format RTCM V3	*Position accuracy* < 2 cm
	UTAC unit reference	

Targets propulsion systems

GLOBAL VEHICLE TARGET PLATFORM		
	*Manufacturer* Anthony Best Dynamics (ABD)	
	Platform unit model MKI / MKII P8500	*Communication* ABD Wifi 5 GHz
	MP Unit model RT3002 / RT3002G	
	UTAC unit reference	

	DDV0228 / DDV0233 / DDV0273

SOFT PEDESTRIAN TARGET RIG

Manufacturer Anthony Best Dynamics	*Unit model* SPT 20
Steering Robot SR60	*Sled height* 25 mm
Maximum speed 20 km/h with 15 kg payload	*Maximum acceleration* 0.8 g with 15 kg payload
UTAC unit reference	

EPT/EBT PLATFORMS

Manufacturer 4Active Systems (4A)	
Single belt unit model 4activeSB	*Dual belt unit model* 4activeSB
Single belt model dimensions Width : 492 mm Length: 990 mm Height: 26 mm Weight: 4kg	*Dual belt model dimensions* Width : 492 mm Length: 990 mm Height: 26 mm Weight: 4kg
Single UTAC unit reference -	*Dual UTAC unit reference* -

3.5. ROAD USERS TARGETS

GUIDED SOFT TARGET

Manufacturer DRI
Unit model Hatchback Soft Car 360™ (Ford Fiesta)
Dimensions
UTAC unit reference

3.6. VULNERABLE ROAD USERS TARGETS

Manufacturer 4Active Systems (4A)
Unit model 4activePA-adult
Model dimensions Body height: 1800 mm Shoulder width: 500 mm Weight: 4 kg
UTAC unit reference HUM00XX

EUROPEAN CHILD PEDESTRIAN TARGET	
 	*Manufacturer* 4Active Systems (4A)
	Unit model 4activePA-child
	Model dimensions Body height: 1154 mm Shoulder width: 298 mm Weight: 2 kg
	UTAC unit reference HUM00XX

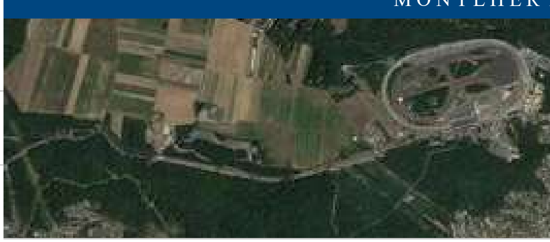
EUROPEAN BICYCLIST TARGET		
	*Manufacturer* 4Active Systems (4A)	
	Bicyclist unit model 4activeBS-adult	*Bike unit model* 4activeBS-adult
	Bicyclist model dimensions Body height: 1800 mm Shoulder width: 500 mm Weight: 4 kg	*Bike model dimensions* Handlebar height : 1200 mm Wheelbase: 1230 mm Weight: 6 kg
	Bicyclist UTAC unit reference HUM00XX	*Bike UTAC unit reference* -

3.7. ROAD SIGNS TARGETS

SPEED LIMIT 50 / 70 / 90		
	*Speed limit type* B14 - Explicit speed 50 - 70 - 90	
	Dimensions Diameter 1050 mm	*Specification* Class 2
	UTAC unit reference -	

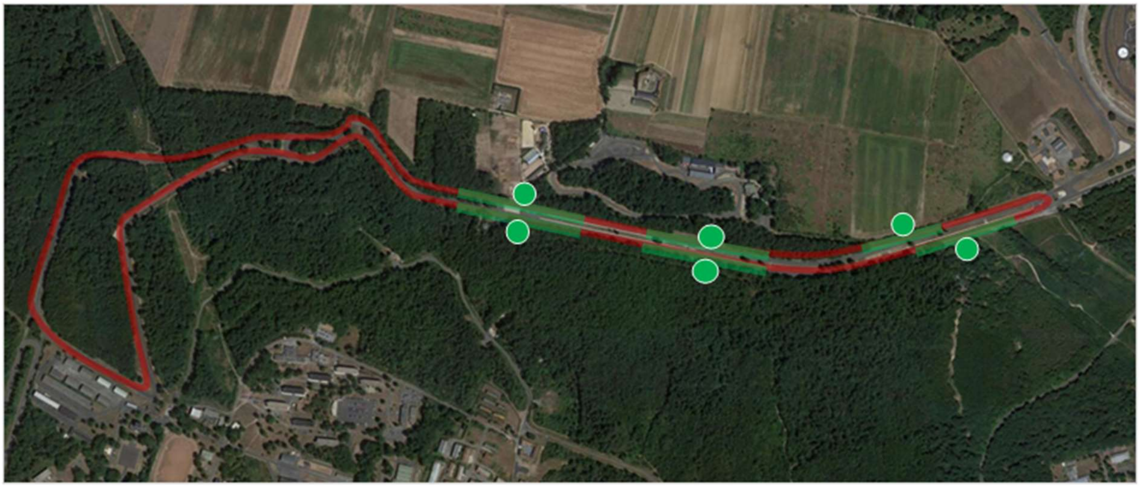
4. UTAC TEST TRACKS

4.1. LOCATION

MONTLHERY TEST UNIT	
	*Location* Autodrome de Linas-Montlhéry
	Address BP20212 91311 Montlhéry CEDEX France

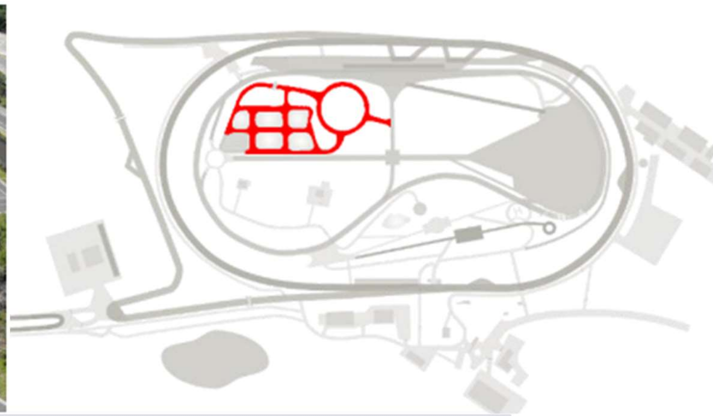
4.2. SPECIFICATIONS

4.2.1. MONTHÉRY UNIT - CR



4.2.2. MONTHÉRY UNIT – TEQMO HIGHWAY

4.2.3. MONTHÉRY UNIT – TEQMO CITY



« City » zone

Features
- Area : 38,500 m²
- 2 and 3 lanes 3.5 m wide
- Lighting & signaling
- 3 crossroads incl. 1 with roundabout and 1 with traffic lights, 7 « T » intersections
- 1 roundabout 50 m radius, 1 level crossing

Services
- Autonomous Shuttle
- Various projects
- Bus stations

5. TESTING RESULTS

5.1. POST-PROCESSING

We define the PASS/FAIL as:
- PASS: The system reacted and allowed to avoid the collision
- FAIL: The system didn't react OR reacted to late to avoid the collision

To go further in the analysis, we check the following values in the raw data (.txt file):
- **Maximum Speed (kph)** of the vehicle during the test

For that, we use the channel named "Speed (kph)" and we check the maximum during the test.

- **Minimum distance (m)** between the vehicle and the Target

This distance is 0 in case of Impact and in case of avoidance we use the channels named "Speed (kph)" and "Relative Longitudinal Distance (m)".

First, we find the index where the vehicle stops, it means when "Speed (kph)" reaches 0 kph.

Then, we check the "Relative Longitudinal Distance (m)" value at the same index.

- **Vehicle Impact Speed (kph)** in case of impact

This is the Vehicle Speed at the time of collision with the Target. We use the channels named "Speed (kph)" and "Relative Longitudinal Distance (m)".

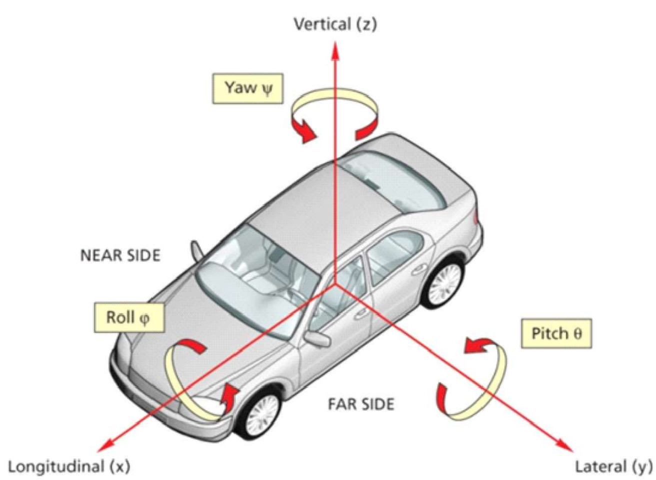
First, we find the index of the collision, it means where "Relative Longitudinal Distance (m)" reaches 0 m.

Then, we check the "Speed (kph)" value at the same index.

- **Vehicle Speed (kph)** at driver avoidance in case of it.

This is the Vehicle Speed at the time of driver avoidance (steering or braking). Depending on the action, we can find the index of the avoidance (huge variation) using "Yaw Velocity (°/s)" or "Forward Acceleration (m/s²)".
Then we check the "Speed (kph)" value at the same index.

5.2. REFERENCE DATA SYSTEM



5.3. DETAILS OF TESTS PERFORMED AND RESULT TABLE

Vehicle n°3 Zoe NEXYAD:

a) **Pre-critical:**

- **Part 1: July 2023**

First, we tested the vehicle without equipment some situation, which can generate anticipation of the system:

Category	Scenarios	Number of subjective tests	Successful	Keep for objective tests
Pre-critical	Stationary car	1	NO	NO
	Approach to roundabout	1	YES	NO
	Approaching strong curve	1	YES	YES

Then we performed some situation with measurement equipment.

All the tests are successful; here are the details of the post-processing:

Scenario	Date	Nbr of tests	VUT Speed (kph)	Success	Reaction	Anticipation	max speed (kph)	speed after curve 1 (kph)	speed after curve 2 (kph)
Approching Curve	03/07/2023	1	ACC	yes	yes	yes	83	29,9	53,9
Approching Curve	03/07/2023	1	ACC	yes	yes	yes	84	29,5	50,7

- **Part 2: Janv/April 2024**

We tested again the vehicle without equipment in some situation that can generate anticipation of the system:

Category	Scenarios	Number of subjective tests	Successful	Keep for objective tests
Pre-critical	Stationary car on highway (emergency lane)	2	YES	NO

	Highway Traffic Right Lane stopped (2 vehicles)	**1**	**YES**	**YES**
	Object on the road (sand filled barrier)	**2**	**YES**	**NO**
	CPFA	**9**	**MITIGATE**	**YES**
	CBLA	**2**	**YES**	**NO**

We decided to keep two scenarios for the pre-critical part:

- Highway traffic right lane stopped (2 vehicles)
- CPFA

Then we performed some situation with measurement equipment.

The scenario CPFA was well anticipated by the system, for the scenario with two vehicles stopped in the right lane, the system has no anticipation:

Scenario	Date	Nbr of Tests	VUT Speed (kph)	Target Speed (kph)	Success	Reaction	Anticipation	Comment	Max Speed (kph)	Remaining distance (m)	avoidance speed (kph)	impact speed (kph)
CPFA	10/04/2024	1	30	8	Yes	Yes	Yes	Detection OK, Vehicle not fully stopped	33,14	13,74	2,49	-
CPFA	10/04/2024	1	40	8	Yes	Yes	Yes	Detection OK, Vehicle not fully stopped	43,4	13,5	2,23	-
Highway traffic right lane stopped	10/04/2024	1	60	-	No	No	No	No detection, driver steering	64	-	63	-
Highway traffic right lane stopped	10/04/2024	1	70	-	No	No	No	No detection, driver steering	74,37	-	72,09	-
Highway traffic right lane stopped	10/04/2024	1	65	-	No	No	No	No detection, driver steering	66,19	-	63,9	-
Highway traffic right lane stopped	10/04/2024	1	70	-	No	No	No	No detection, driver steering	67,49	-	64,212	-

Conclusion for pre-critical scenarios:

We noticed better reactions in 2024 compared to 2023. NEXYAD improved their detection and response system, which allowed us to play more scenarios.
For example, in 2023, the system was not very reactive on a pedestrian crossing (CPFA), system reactions are better in 2024, which is why we were able to carry out this scenario.
The CPFA scenario proved to be relevant during this test phase, the pedestrian detection was well managed, the vehicle braked sufficiently.

Regarding the Highway traffic right lane stopped scenario, the system does not detect any danger, and therefore does not apply any braking prevention reaction.
Of course, there is no direct risk of collision because the targets are in the adjacent lane, but we could imagine a preventive "foot on the brake" as an average driver might do in this situation.

These tests allow us to verify on a vehicle with different settings & different influents variables that our robustness tests and protocols and requirements are still relevant.

Finally, these more complete and challenging tests, on a very intelligent vehicle, confirmed our January 2024 proposals (L3.3) for tests, protocols and evaluation/homologation metrics. We redo the 2023 scenarios with the 2024 vehicle settings, and 2024 vehicle reactions which are different and rather better in 2024 compared to 2023 ; but 2023 conclusions and proposals for requirements and evaluation are still the same and so are confirmed.

b) Robustness:

In the same way, we started to perform the scenarios without equipment to see the relevancy. We tested the scenario with a stationary car on the highway (emergency lane).

Category	**Scenarios**	**Number of subjective tests**	**Successful**	**Keep for objective tests**
Pre-critical	**Stationary car on highway (emergency lane)**	**2**	**YES**	**YES**

Then we performed some situation with measurement equipment.

Stationary car with an angle

Lidar masking with some mud

Here are the results of the tests performed with our measurement equipment:

Scenario	Date	Time	Nbr of tests	Risk Target**	VUT Speed (kph)	Overlap	Angle	Lidar masking	Stationary pedestrian	Lateral-distance-ped-VUT (m)	Success	max speed (kph)	remaining distance (m)	avoidance speed (kph)
Stationary car on Emergency Lane	09/04/2024	16:24	1	50	40	100	0	YES	NO	0	FAIL	40,86	0	40
Stationary car on Emergency Lane	10/04/2024	14:17	1	50	50	75	0	NO	NO	0	PASS	55,44	13,32	0
Stationary car on Emergency Lane	11/04/2024	10:45	1	50	60	75	25	NO	NO	0	FAIL	63,73	0	13
Stationary car on Emergency Lane	11/04/2024	15:24	1	30	40	100	0	NO	NO	0	PASS	42,91	20	0
Stationary car on Emergency Lane	11/04/2024	15:38	1	70	60	100	0	NO	NO	0	PASS	63,5	12	0
Stationary car on Emergency Lane	11/04/2024	17:03	1	30	60	75	0	NO	NO	0	FAIL	63,96	0	54
Stationary car on Emergency Lane	11/04/2024	17:25	1	70	60	75	0	NO	NO	0	FAIL	63,45	0	12
Stationary car on Emergency Lane	11/04/2024	17:48	1	70	60	75	0	NO	YES	0	PASS	62,63	12	0
Stationary car on Emergency Lane	12/04/2024	10:16	1	30	60	50	0	NO	NO	0	FAIL	61,88	0	8
Stationary car on Emergency Lane	12/04/2024	11:05	1	70	60	50	0	NO	YES	2	PASS	64,09	7,78	0

*****Risk Target or Lack of prudence target***: Internal parameter defined and set by NEXYAD. This parameter defines the level of lack of prudence allowed during the driving, from 0 to 100. A level of 100 for a not prudent driving and 0 for a very safe driving. In the tool, the lack of prudence target cannot be set at a level above 80.

<u>Conclusion for robustness scenario:</u>

The robustness tests proved to be interesting; we were able to vary different parameters to evaluate the performance of the system.
Some parameters are classic and come from the ENCAP protocols, such as vehicle speed and overlap.
We have also imagined new parameters to be varied, such as:

- **Target risk**, an internal parameter of NEXYAD, defined above.
  This parameter, set to 50, simulates the attention of an average driver. We noticed that this parameter set to 30 allowed better anticipation when the detection was effective, compared to a target risk of 70.
- **The angle** of the parked vehicle
  A test was carried out with a vehicle parked on the track, with an angle, this test is FAIL, while the same type of tests without an angle is PASS. NEXYAD confirmed that approaching a vehicle with an angle is a difficult situation now.
- **Lidar masking** with mud
  We noticed that when the lidar is partially masked with mud, the performance is greatly degraded. This led to ideas for an (automatic) Lidar cleaning system to avoid this kind of problem.
- **The presence of a pedestrian** in front of or next to the parked vehicle.
  The presence of a pedestrian in addition to the parked vehicle did not disturb the correct detection of the system, the reactions are correct.

Please note that as for any vehicle, robustness is not 100% perfect: for example, five impacts (or test driver avoidance) among 10 tests, as shown in the results table,
Note these 10 tests were a free choice among the many tests done and were done in very severe conditions. In more usual conditions the robustness of this vehicle is different, more than 90%, which is also the rate related to all tests done.
Despite vehicle is detected in perfect conditions (Euro NCAP protocol), we could observe that variations of conditions (overlap, angle, pedestrian standing near vehicle, vehicle speed, lidar obstruction (dust), could lead to impact (or test driver manual avoidance).
This can explain that the system is not fully robust to environment and that performance is not the same related to different conditions.

Finally, we redo our 2023 tests and protocols but with more and different influent factors enabling to reach vehicle limits (which were quite different than 2023 limits, so interesting to confirm of affine our requirements)

- stationary car overlap but also angle, see photo here below.
- mud on the lidar glass, making perception more difficult (see photo here below)
- pedestrian standing near the stationary car on emergency lane.
- longitudinal distance for obstacle / car perception
- accepted level of risk

These tests allow us to verify on a vehicle with different settings & different influents variables that our robustness tests and protocols and requirements are still relevant.

Finally, these more complete and challenging tests, on a very intelligent vehicle, confirmed our January 2024 proposals (L3.3) for tests, protocols and evaluation/homologation metrics. We redo the 2023 scenarios with the 2024 vehicle settings, and 2024 vehicle reactions, which are different and rather better in 2024 compared to 2023; but 2023 conclusions and proposals for requirements and evaluation are still the same and so are confirmed.

6. CONCLUSION

For 2024 POC 2 tests :

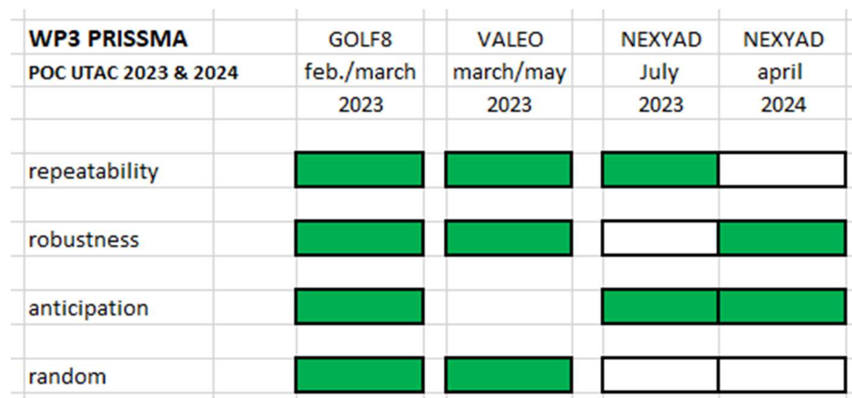
we only tested in 2024 the prototype ZOE NEXYAD « DREAMotorONE » research prototype with AI based anticipation, because it contains more IA and intelligence, and because its perception has been enhanced beginning 2024 allowing interesting tests more to fine-tune our conclusions and proposals for robustness and anticipation tests,
in the four most interesting scenarios among new scenarios built in POC 1 in 2023.
Finally, these more complete and challenging tests, on a very intelligent vehicle, confirm our January 2024 proposals (L3.3) for tests, protocols and evaluation/homologation.
metrics.

For the whole PRISSMA PROJECT (POC 1 in 2023 & POC 2 in 2024) :

We tested 3 different vehicles, which are technical references
(as largely explained in L3.2) and with different levels of autonomy:
level 1 for Golf8 and ZOE NEXYAD, level 4 for VALEO delivery robot.

The new 4 categories of tests and protocols built in 2022.
and detailed in L3.2 in January were tested in 2023 & 2024.
and their feasibility confirmed.

The beside figure synthesizes that the new tests proposed.
are feasible and OK for most of the 3 vehicles :



About our test results and our proposed requirements:

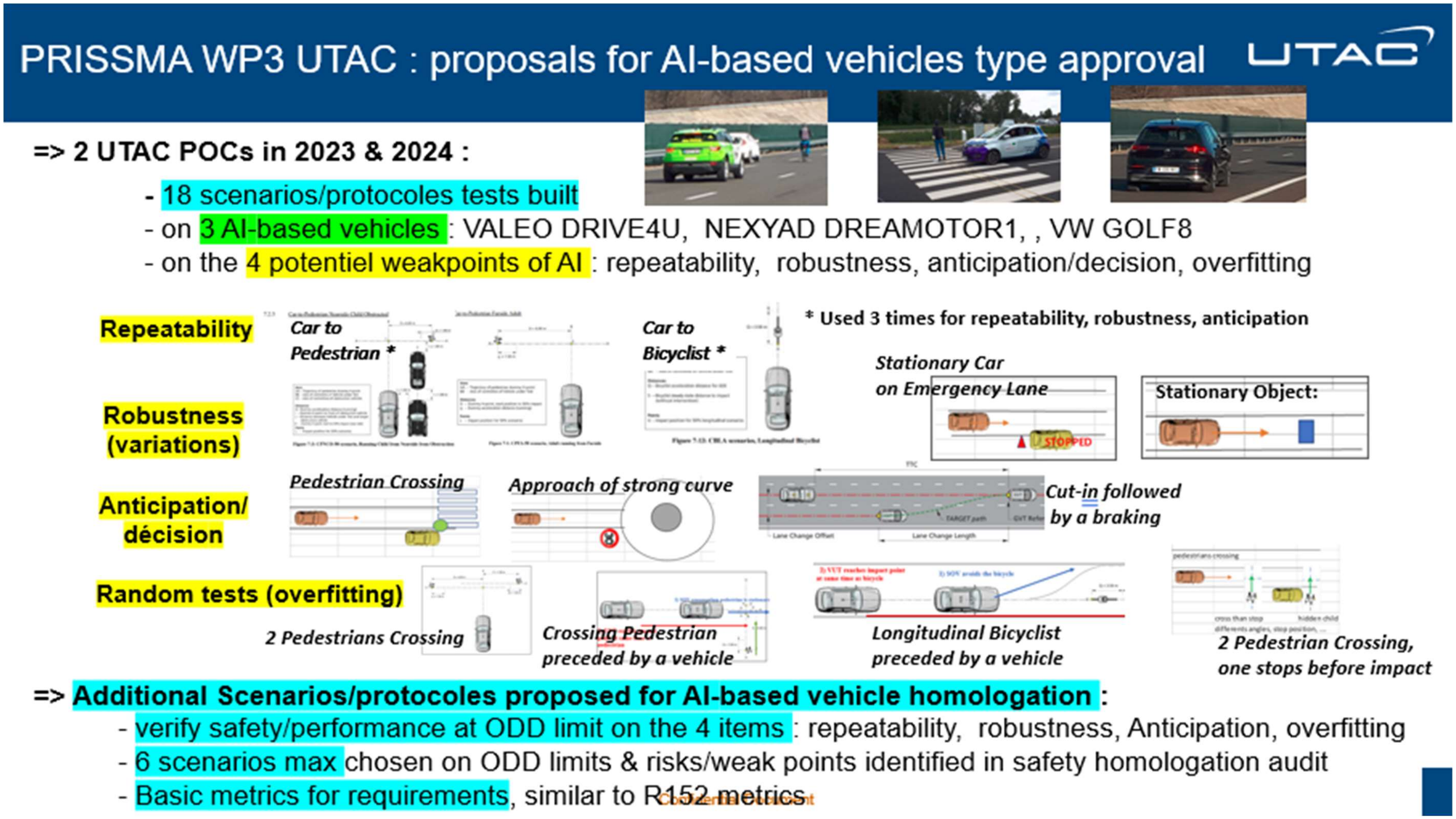
Repeatability tests & requirements: as mentioned in L3.2 in January 2024 with some results, no vehicle is perfectly repeatable ; The 3 vehicles tested are not perfect in repeatability and the main thing for safety is that they have no significantly lower performance than other vehicles without AI.

Robustness tests & requirements: we built and confirm feasibility of different tests and influent parameters to change during the tests (like Objects Speed, Angles, Overlaps…). The results of the tests show that the three vehicles tested are not perfect in robustness; The main thing for safety is that they have no significantly lower performance than other vehicles without AI.

Anticipation tests & requirements: we built and confirm feasibility of different new tests to evaluate vehicle anticipation. Two vehicles (Golf 8 and NEXYAD) showed real interesting anticipation skills so it would be good for safety to propose new tests and new evaluation of anticipation.

Random tests & requirements: we built and confirm feasibility of different new tests to evaluate vehicle anticipation. Two vehicles (Golf 8 and VALEO) showed real interesting skills to manage some of these new tests & scenarios, so it would be good to avoid type approval over-fitting to propose new tests for AI based vehicles type approval.

As an A4 format poster here is a summary of the whole UTAC WP3 PRISSMA results & proposals:


PRISSMA WP3 UTAC : proposals for AI-based vehicles type approval
UTAC
=> 2 UTAC POCs in 2023 & 2024 :
- 18 scenarios/protocoles tests built
- on 3 AI-based vehicles : VALEO DRIVE4U, NEXYAD DREAMOTOR1, , VW GOLF8
- on the 4 potentiel weakpoints of AI : repeatability, robustness, anticipation/decision, overfitting
Repeatability
Robustness (variations)
Anticipation/ décision
Random tests (overfitting)
Car to Pedestrian *
Car to Bicyclist *
* Used 3 times for repeatability, robustness, anticipation
Stationary Car on Emergency Lane
STOPPED
Stationary Object:
Pedestrian Crossing
Approach of strong curve
Cut-in followed by a braking
2 Pedestrians Crossing
Crossing Pedestrian preceded by a vehicle
Longitudinal Bicyclist preceded by a vehicle
2 Pedestrian Crossing, one stops before impact
=> Additional Scenarios/protocoles proposed for AI-based vehicle homologation :
- verify safety/performance at ODD limit on the 4 items : repeatability, robustness, Anticipation, overfitting
- 6 scenarios max chosen on ODD limits & risks/weak points identified in safety homologation audit
- Basic metrics for requirements, similar to R152 metrics

7. TESTING RESULTS

TIME INFORMATIONS		
Channel names	*Units*	*Comments*
Time	s	Time starts in the path
MP Time	s	GPS time of VUT
MP Time Tracker 1	s	GPS time of VRU or GST

VUT SPECIFIC INFORMATIONS		
Channel names	*Units*	*Comments*
Actual X (front axle)	m	X of the car (VUT) (at the bumper)
Actual Y (front axle)	m	Y of the car (VUT) (at the bumper)
Speed	kph	Absolute speed of the car (VUT)
Forward velocity	m/s	Forward speed of the car (VUT)
Lateral velocity	m/s	Lateral speed of the car (VUT)
Forward acceleration	m/s²	Forward acceleration of the car (VUT)
Lateral acceleration	m/s²	Lateral acceleration of the car (VUT)
Yaw angle	°	Yaw angle of the car (VUT)
Yaw velocity	°/s	Yaw velocity of the car (VUT)
Yaw acceleration	°/s²	Yaw acceleration of the car (VUT)

TARGET SPECIFIC INFORMATIONS		
Channel names	*Units*	*Comments*
Head tracker reference X position	m	Position of the VRU on X axis
Head tracker reference Y position	m	Position of the VRU on Y axis
Head tracker forward velocity	m/s	Speed of the VRU on its path
Head tracker forward acceleration	m/s²	Acceleration of the VRU on its path

RELATIVES VUT/TARGET SPECIFIC INFORMATIONS		
Channel names	*Units*	*Comments*
Time to Collision (longitudinal)	s	Remaining time before the VUT strikes the target, assuming that the VUT and the target would continue to travel with the speed it is travelling
Relative longitudinal distance	m	Difference between the longitudinal positions of the vehicle and the target
Relative lateral distance	m	Difference between the lateral positions of the vehicle and the target
Relative longitudinal velocity	m/s	Difference between the longitudinal speeds of the vehicle and the target
Relative lateral velocity	m/s	Difference between the lateral speeds of the vehicle and the target
Relative yaw	°	Difference between the yaw angles of the vehicle and the target

Chapter 2: INRIA/TRANSPOLIS POC

1. CONTEXT

1.1. INTRODUCTION

This section presents the details related to the second test campaign that was conducted in the framework of the PRISSMA POC proposed by Inria and Transpolis. The experiments were conducted over five days in early February 2024 at the Transpolis testing site.
The objective of this proof of concept is to showing the interest and potentiality of using an augmented reality framework developed by Inria as a tool to improve testing and validation of AI-based solution for autonomous vehicles in controlled environments. To do so, we tested for validation the AI-based perception software stack of INRIA's autonomous platform, represented by an automated Renault Zoe. This validation was achieved using a scenario-based approach where several dynamic virtual obstacles were introduced into the sensor data via INRIA's Augmented Reality (AR) system. The AR framework incorporates a data fusion methodology, enabling real-time augmentation of LiDAR sensor data. This allows for seamless integration of both real and virtual elements into testing scenarios, facilitating a smooth transition from simulation to real-world testing. The primary aim of this POC is to illustrate how augmented reality serves as a powerful tool for enriching testing scenarios in controlled environments. By doing so, it aims to demonstrate how augmented reality can make the evaluation and validation process more cost-effective and safer.
Before presenting the obtained results and their analysis, we first recall the main aspects of the Inria autonomous vehicle, its perception module and the augmented reality framework. However, a more detailed description of the objectives and components can be found in the documents describing the first experimental campaign and the protocol defined for the second campaign (PRISSMA deliverable 3.5 and 3.3 respectively).

1.1.1. INRIA’s autonomous driving platform

INRIA's autonomous platform is a Renault Zoe vehicle outfitted with multiple sensors for localization and perception. In particular, it is equipped with a Velodyne HDL-64 mounted on the roof, three Ibeo Lux LiDARs positioned at the front and 1 at the rear, Spectra SP90 RTK

Dual antenna GNSS for precise positioning, Xsens IMU for vehicle velocity and orientation data, a stereo camera, and 2 IDS cameras. The data from LiDARs is consolidated and synchronized using the IBEO fusion box. It also features automated steering, throttle, and brake commands, allowing for autonomous navigation through the embedded computer and software navigation stack. The focus of the experiments conducted for this POC was on the study of how the AI-based perception framework present on the vehicle reacts to the real-time fusion of real and virtual data obtained through the AR framework.

1.2. AUGMENTED REALITY

The augmented reality (AR) system consists of four key modules. Firstly, the Virtual Environment hosts a digital replica of the experimental vehicle and additional virtual elements for flexible testing scenarios. The Synchronization module ensures real-time alignment of the virtual vehicle with its real-world counterpart. Secondly, Sensor Emulation generates outputs from virtual sensors, focusing on LiDAR sensors for this POC. This module merges virtual and real sensor data to create a realistic AR perception. Lastly, the Visualization module presents merged sensor data from virtual and actual cameras in an intuitive format, aiding testers in understanding the AR scene. The AR framework shown in Figure 1operates in real time, ensuring seamless integration and accurate representation of the environment. Technical details on real-time merging of LiDAR data are available in [1].

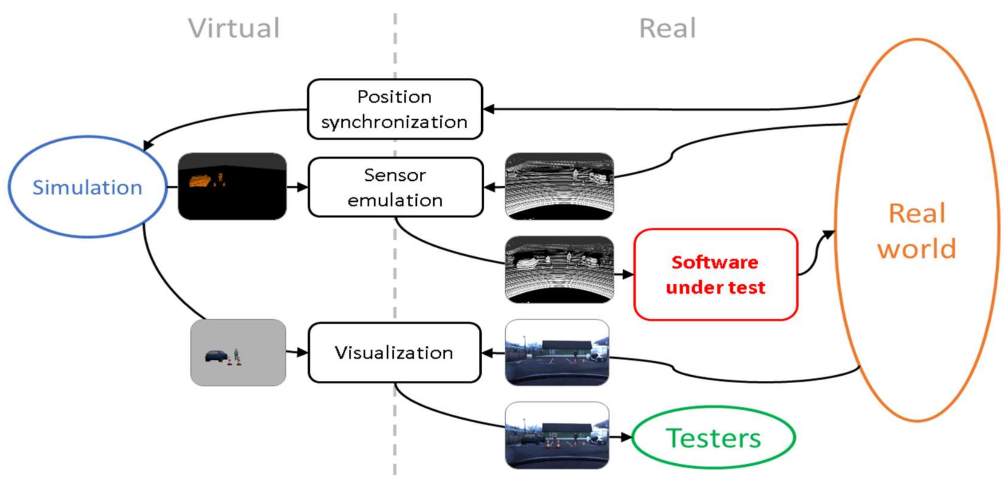


Figure 1: Augmented Reality (AR) framework

Augmented Reality system adeptly manages occlusions between real and virtual elements, ensuring accurate representation of test scenes. Experimental validation confirms the system's capability. This allows for seamless integration of the entire vehicle and its software into hybrid but realistic test environments. This innovative testing approach serves as a bridge between Vehicle-in-the-Loop and real world testing methodologies.

1.3. PERCEPTION MODULE

The perception module generates probabilistic occupancy grids via the CMCDOT framework, a Bayesian occupancy filter that infers occupancy probabilities, velocities, and collision risks with predicted obstacles. The prediction grid, as depicted in Figure 2, serves as a robust model for anticipating occupancy within an environment. By incorporating vital input data, such as occupancy probabilities and estimated velocity, this grid predicts the likelihood of cell occupancy. Each cell is projected based on its estimated velocity, facilitating the representation of movement. To mitigate noise, cells are subdivided into particles with specific accelerations and angular velocities.

Operating as a probabilistic distribution, the prediction grid offers insights into future occupancies within a three-second period. It amalgamates occupancy grids sourced from diverse sensor measurements, creating a unified representation that accumulates information over time. The velocity grid derived from LiDAR measurements remains the most precise estimation of motion. Furthermore, the prediction grid enhances understanding of occupancy by visualizing predictions over time. Static objects are depicted in white, while moving objects are portrayed using colors indicative of their estimated time of arrival. To ensure cautious behavior around moving objects, a significant uncertainty is introduced during the prediction process, resulting in the formation of clouds of predicted occupancy. This approach accommodates potential variations and uncertainties associated with object movement.

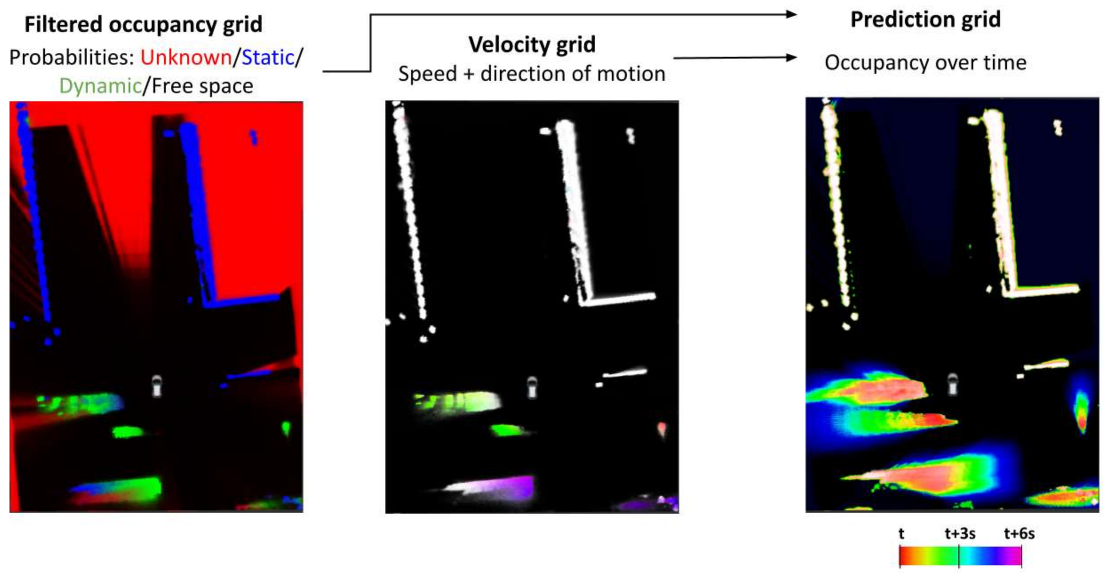


Figure 2: Prediction grid: a crucial component of the system. This grid predicts occupancy by projecting cells with velocity while accounting for noise and merging sensor data. It provides valuable insights into future occupancies, seamlessly integrating path planning and obstacle avoidance strategies. This comprehensive approach enhances understanding and navigation through dynamic environments.

1.4. LOCAL PLANNER

These ego-vehicle sensors serve as inputs for the navigation stack, which comprises three key modules: localization, perception, and navigation. Localization integrates data from multiple sensors using a Kalman filter, with a primary reliance on centimetric RTK GPS for accuracy. The perception module, as already described, is based on Bayesian inference and prediction. Finally, the navigation utilizes model predictive control (MPC) alongside a predictive collision detector (PCD) to guide the vehicle safely through the environment. The MPC module forecasts potential future trajectories based on different command samples (such as throttle, brake, and steering commands), while the PCD module calculates expected time to collision for each trajectory by predicting the behaviors of perceived obstacles. The command sample that minimizes collision risk is then selected and transmitted to the embedded car controllers for vehicle operation, as illustrated in **Figure 3**.

Overall, this navigation stack offers robust capabilities in localization, perception, path-finding and dynamic obstacle avoidance, making it well suited for navigating through dynamic and unstructured environments.

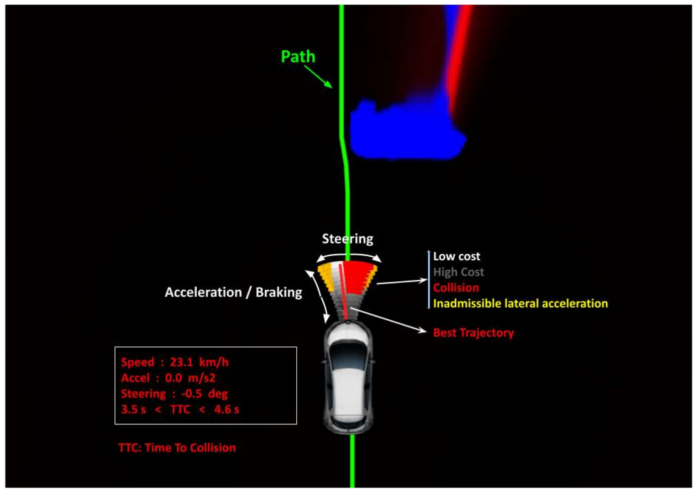


Figure 3: Local Planner, which is responsible for generating admissible commands, predicting trajectories, and assessing collision risks. It performs accurate and efficient computations to select the optimal trajectory within short time horizons.

2. EXPERIMENTS AND DATA COLLECTION

2.1. TRANSPOLIS TESTING FACILITY

The testing occurred at the Transpolis testing facility shown in **Figure 4**, where a variety of scenarios were replicated for this POC, particularly focusing on a long boulevard with an intersection within the City area. The City Area spans 30 hectares and encompasses a meticulously designed urban landscape with an extensive network of streets covering 12 kilometers, featuring two prominent boulevards with six lanes each. Divided into four sections, each section offers a distinct layout comprising intersections, crossroads, and parking areas. The infrastructure includes dedicated lanes for buses and bicycles, along with a ring road for convenient access. With 40 real buildings facilitating connectivity testing in various conditions, the area is equipped with adjustable facilities such as fiber optic cabinets, EV charging stations, and dynamic changing-message signs. Additionally, movable signs, traffic lights with GLOSA services and roundabouts cater to diverse testing needs, while luminescent road markings provide precise guidance.

The driving environments within the City Area offer varied surfaces, vegetation, and sloping terrain for comprehensive evaluations. Covering 7000 square meters, the City Area also functions as a parking facility and event space, demonstrating a commitment to advancing urban mobility through technological innovation.

The main boulevard intersection, where the experiment was conducted, features a 6 by 6 lanes configuration. The Zoé vehicle traverses the intersection from west to east lanes, while virtual obstacles move from north to south lanes. Notable static obstacles detected by LiDARs include four buildings positioned at each corner of the intersection, traffic lights, signs, and 2-meter-high concrete panels installed on the north side of the road traversed by the vehicle.

Figure 4: Satellite image of the Transpolis facility.

2.2. GENERATION OF GROUND TRUTH USING TRANSPOLIS SATELLITE IMAGE

In order to assess and analyze the results collected during these experiments, we first need to generate create ground-truth occupancy grids for our perception module (CMCDOT). This has been done by utilizing a satellite image of the Transpolis facility, as depicted in **Figure 3Figure 4**. Modifications have been also made to this satellite image in order to match the settings in the simulation environment as shown in **Figure 5**. The location of the Renault Zoé vehicle is pinpointed on the satellite image based on its geolocation during the test scenarios. Subsequently, an approximation of the ground truth occupancy grid, matching the dimensions of the CMCDOT grids, is extracted from the satellite image around the vehicle. This grid encompasses static objects and the surrounding environment.

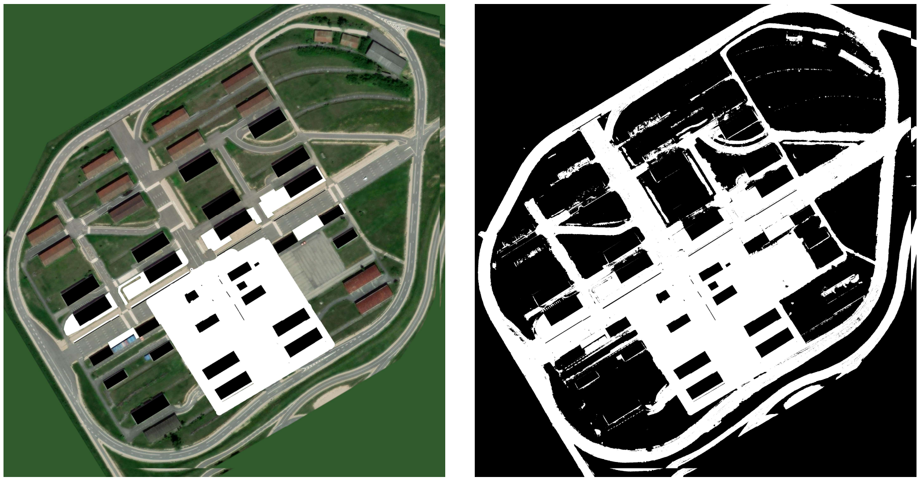


Figure 5: Left : Modifications to the satellite image of Transpolis in order to match the environment in simulation. Right: Binary image of Transpolis arena.

Utilizing the augmented reality (AR) framework, most of the dynamic objects populating the test scenarios are represented as virtual actors. The simulator governs these virtual actors, ensuring that their state (position, orientation, speed, and footprint) is accurately known at every moment of the test. They are geolocated onto the ground truth satellite image, and their footprints are superimposed onto it.

Similarly, employing the AR framework, we merge the real environment with the dynamic virtual actors during the test scenario, thereby generating a corresponding ground truth. This is achieved by combining the static ground truth of the Transpolis facility with the ground truth data of the actors from the simulator, as illustrated in Figure 6.

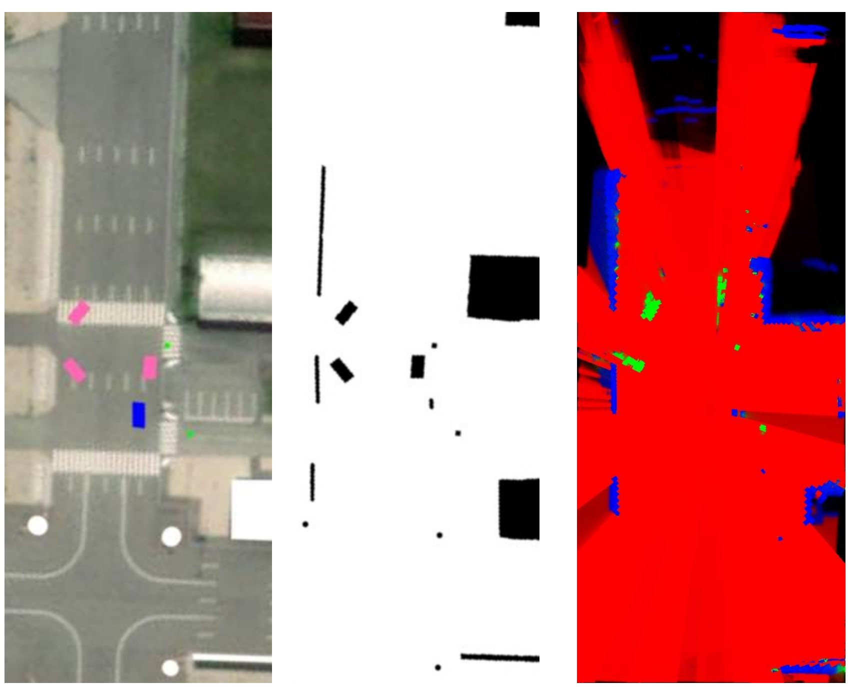


Figure 6: Left: Ground truth generation using the RGB satellite image of Transpolis arena, Middle: Binary ground truth image using the Figure 5 Transpolis image, Right: Occupancy grid from the perception module.

2.3. DATASET DESCRIPTION

During this second test campaign for the POC, numerous experiments were conducted involving the Zoé autonomous vehicle and dynamically augmented obstacles, following the five scenarios described in the following section. The test campaign lasted for 5 days, from the 12 to the 16 of February 2024 at the main intersection in the urban area of the Transpolis site. The first 3 days where dedicated to the calibration of the Zoé (perception and navigation parameters tuning for the test field and calibration of AR) and the tuning of the scenario (starting positions and start trigger points). The last 2 days were dedicated to the execution of the scenarios and the recording of the data. Tests with the pedestrian target occurred on the fourth day so; the first day of recording was focused on pedestrian scenarios and the second day on 20 bags for each scenario to complete the test campaign. For the 2 days of testing and recording, all parameters were kept constant to ensure the comparability of the results and to evaluate the reproducibility of the scenarios and the repeatability of the Zoé.

The Zoé's software architecture utilizes the ROS (Robot Operating System) framework, version Melodic, where software components are organized into nodes communicating via typed topics; for instance, the LiDAR driver sends point cloud messages on the LiDAR topic, which are then read by the CMCDOT node. To record experiment data, the rosbag tool captures messages from requested topics, storing them in timestamped binary files called bags. Subsequently, the rosbag player replays the recorded messages, maintaining their order and simulated time. Table 1 outlines the recorded topics, providing insights into the scenarios, ego-vehicle behavior, and interactions with augmented vehicles. 124 bags were recorded (scenario 1: 20, scenario 2: 20, scenario 3: 20, scenario 4: 21, scenario 5: 21, scenario pedestrian target: 22), representing around an hour of continuous driving and using a volume of 671 GB. The rosbags data description can be viewed in1.

3. SCENARIO DESCRIPTION

- **Scenario 1: Speeding vehicle from behind**
  - In this scenario, the autonomous vehicle (AV) encounters a situation where a vehicle approaches from behind at a significantly higher speed, prompting the AV to execute emergency braking. However, the obstacle vehicle does not have sufficient time to react, increasing the likelihood of a collision with the AV. Rosbags description for this scenario can be viewed in Table 2
- **Scenario 2: Opposite lane vehicle overtaking**
  - In this scenario, two vehicles travel in opposite lanes of the autonomous vehicle (AV). The rear vehicle, depicted in green, attempts to overtake the vehicle ahead, indicated in blue. This maneuver increases the risk of a collision with the AV as it navigates through its path. Table 3 presents the rosbags associated to this scenario.
- **Scenario 3: Vehicle from the right at intersection**

- In this scenario, the AV crosses an intersection simultaneously with another vehicle that refuses to yield the right-of-way. This refusal increases the likelihood of a collision between the two vehicles. Rosbags corresponding to this scenario are available in Table 4.

- **Scenario 4: Pedestrian crossing the road**
  - In this scenario, a pedestrian crosses the road at the same moment the AV passes. Despite clear visibility of the pedestrian, there is a risk of collision if the AV fails to react appropriately. Table 5 presents the associated rosbags for the scenario.

- **Scenario 5: Occluded pedestrian crossing the road**
  - In this scenario, a pedestrian crosses the road at the same moment the AV passes, but the pedestrian is partially obscured by parked vehicles. This occlusion increases the risk of a collision if the AV fails to detect the pedestrian in time. Table 6 presents the related rosbags.

3.1. SCENARIO 1

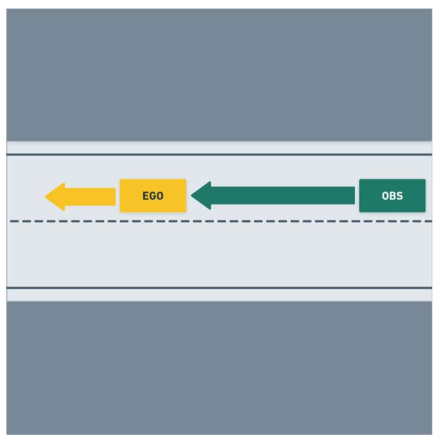


Figure 7 : Pictogram of Scenario 1

Figure 8: If a vehicle approaches the ego-vehicle (blue) from behind at a significantly higher speed, the ego-vehicle may need to execute an emergency braking maneuver. In such a scenario, the obstacle vehicle may not have adequate time to respond, increasing the risk of a collision with the ego-vehicle.

3.2. SCENARIO 2

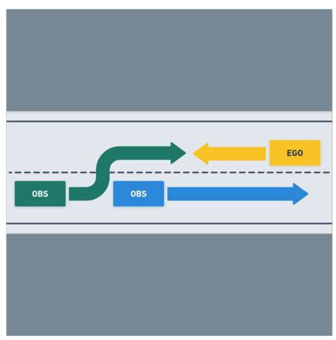


Figure 9: Pictogram of Scenario 2

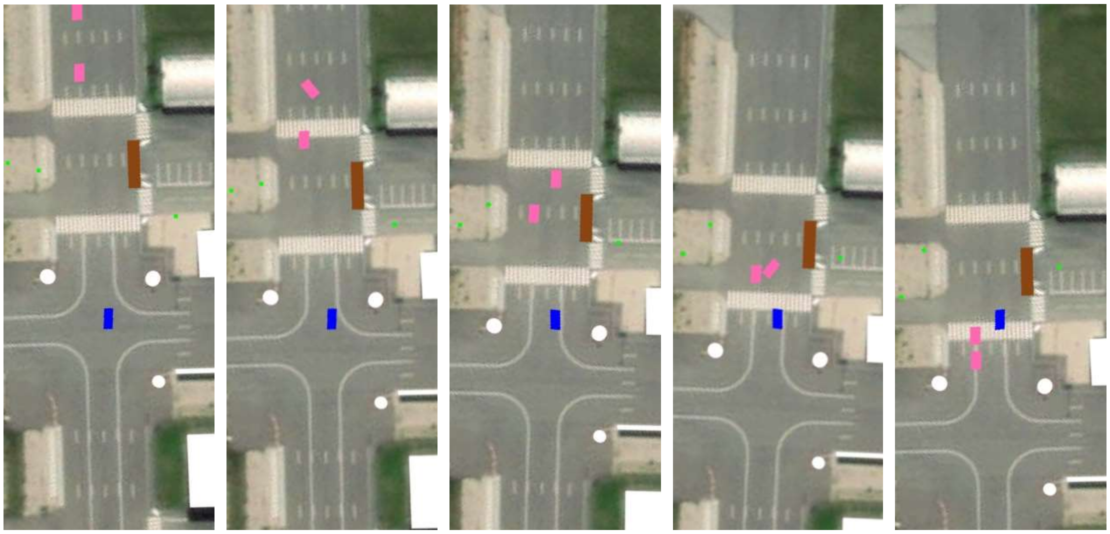


Figure 10: In this scenario involving two vehicles, both traveling in the opposite lane of the ego-vehicle (blue), the rear vehicle (green) attempts to overtake the vehicle ahead of it. This maneuver increases the likelihood of a collision with the ego-vehicle.

3.3. SCENARIO 3

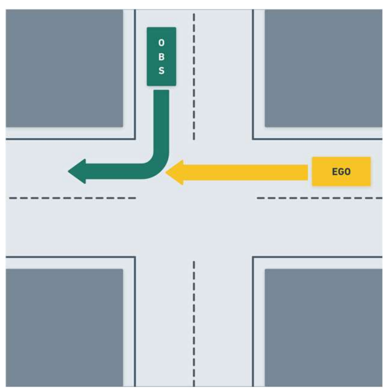


Figure 11: Pictogram of Scenario 3

Figure 12: In this particular scenario, referred to as a two-vehicle scenario, the ego-vehicle (blue) and another vehicle simultaneously approach an intersection. However, the other vehicle fails to yield the right-of-way, potentially resulting in a collision with the ego-vehicle.

3.4. SCENARIO 4

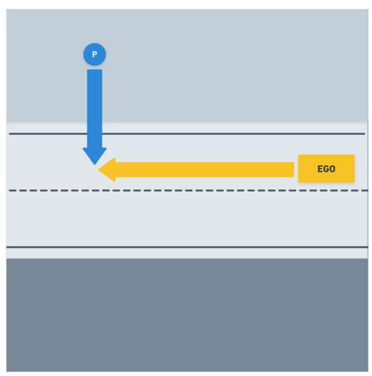


Figure 13: Pictogram of Scenario 4

Figure 14: In this scenario, a pedestrian crosses the road just as the ego-vehicle (blue) is passing, creating the potential for a collision. It is noteworthy that the ego-vehicle has a clear line of sight of the pedestrian during this moment.

3.5. SCENARIO 5

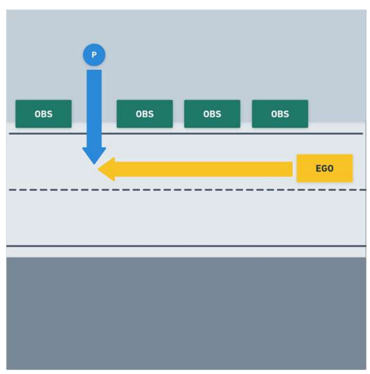


Figure 15: Pictogram of Scenario 5

Figure 16: In this scenario, a pedestrian crosses the road at the same time as the ego-vehicle (blue) passes, posing a potential collision risk. However, the pedestrian is obscured from the ego-vehicle's view by parked vehicles, heightening the likelihood of a collision.

4. SCENARIO EXECUTIONS DESCRIPTION

The selection and a detailed description of the scenarios used in this POC are outlined in PRISSMA Deliverable 3.3. This document provides an explanation of how and why each scenario has been chosen, based on the critical interactions identified in the Surca project [3, 4]. These scenarios are expected to be among the most relevant for testing the safety of an AV but also feasible to be executed at the Transpolis testing facility and repeatable for result analysis. Figure 7,Figure 9 Erreur ! Source du renvoi introuvable.**,** Figure 13, and Figure 15 show pictograms describing the scenarios, while Figure 8**,** Figure 10**,** Figure 12**,** Figure 14**,** Figure 16 show time-lapsed images of scenario execution.

4.1. SCENARIO 1 - SPEEDING VEHICLE FROM BEHIND

This scenario involves a vehicle that is approaching the Zoé from behind at a higher speed. Both cars are driving in the same lane with a 2 meters lateral offset. If the Zoé does not take any action it will be hit from behind by the faster vehicle. The scenario is designed to test the Zoé's ability to perceive and react from a danger coming from behind that can't be avoided by simply stopping. Two categories of outcomes were observed.

1. The Zoé perceived the vehicle and successfully avoids the collision by changing lane. In this case, the Zoé's tracking of the vehicle velocity is accurate enough to reduce the future occupancy to a realistic spread. Among the sampled trajectories, the planner is able to find a safe one that moves the Zoé out of the way of the incoming vehicle, usually by steering strongly to the left than braking.
2. The Zoé perceived the vehicle but is still brakes in front of it, leading to a rear collision. It is suspected that, in this case, the Zoé's tracking of the vehicle velocity was less accurate, resulting

in a future occupancy that is too spatially spread. This leads to an inability of the planner to find a safe trajectory and hence forcing the choice of a conservative emergency braking.

4.2. SCENARIO 2 - OPPOSITE LANE VEHICLE OVERTAKING

This scenario involves a vehicle that is overtaking another vehicle in the opposite lane. The overtaking vehicle is driving towards the Zoé and is expected to be unable to return to its lane in time to avoid a collision with the Zoé. The scenario is designed to test the Zoé's ability to perceive and react to a vehicle that is driving in the opposite lane and is invading its lane. The Zoé almost always performed an emergency braking to try to avoid the collision. The opposite lane vehicle starts its overtaking maneuver few seconds before the expected collision. Tracking of its occupancy is usually too late for the planner to find another trajectory than an emergency braking. According to Table 3: Scenario 2 , the Zoé is still able to avoid the collision in 70% of the bags. In two bags the tracking of the vehicle was accurate enough to allow the planner to find a safe trajectory avoiding the collision by steering to the right, out of the way of the incoming vehicle.

4.3. SCENARIO 3 - VEHICLE FROM THE RIGHT AT INTERSECTION

This scenario involves a vehicle that is approaching the Zoé from the right at an intersection. The vehicle is expected to give the right of way and to collide with the Zoé. None of the executions of this scenario led to a collision, the Zoé systematically stopped to avoid the collision. Similar to the scenario 1 shown in Figure 8, we observed two categories of outcomes: emergency stop when the future occupancy is too much spatially spread and slowdown with dynamic avoidance when the vehicle velocity is accurately tracked and the spread of future occupancy is reduced. Two cars driving on an opposite lane were added to increase the complexity of the scene. They were expected to not interfere with the Zoé but, when the Zoé avoided the first vehicle by steering to the left, it had to brake to avoid the second and third vehicles. It shows that the planner is able to navigate with multiple collision threats.

4.4. SCENARIO 4 - PEDESTRIAN CROSSING THE ROAD

This scenario involves a pedestrian that is crossing the road in front of the Zoé. The Zoé should detect the pedestrian and give the right of way, it is expected that the planner simply brakes before the pedestrian crosses the road. We observed that the Zoé avoid the collision in 75% of the bags. In the other 25%, the Zoé was not able to detect the pedestrian in time but usually stopped on the zebra crossing and then the pedestrian hit the side of the car. In several bags, the tracking of the pedestrian was accurate enough and the scenario timing allows the Zoé safely accelerate and cross before the pedestrian. However, almost all the bags show that the Zoé actions lack of smoothness and safety distance with a pedestrian, usually trying to pass just before it with few margins or stopping too close. It feels like the Zoé has an aggressive driving behavior.

4.5. SCENARIO 5 - OCCLUDED PEDESTRIAN CROSSING THE ROAD

This scenario is similar to the scenario 4 except that a bus occluded the pedestrian, reducing the time the pedestrian is visible and tracked by the Zoé's perception system. As we expected, the results are worse than the scenario 4: we observed the same collision rate but with two front collisions that were not observed in the scenario 4. We also observed more abrupt stops and fewer dynamic avoidance. This shows well the importance of the tracking duration of the obstacles for the planner to find better trajectories. The occlusion of the pedestrian by the bus reduces this tracking duration, leaving less time to react and less accurate prediction of future occupancy. The obstacle is unexpected and not predictable by the Zoé, this leads to an absence of anticipation and more abrupt emergency braking.

4.6. SCENARIO 5 WITH REAL PEDESTRIAN TARGET

This scenario is a variant of the scenario 5: while still using AR for most of the actors, the pedestrian crossing the road in front of the Zoé is a real pedestrian target designed for testing vehicle ADS and ADAS. The specifications of the pedestrian target and other targets available at Transpolis are in the test protocol of this POC in deliverable 3.3 of PRISSMA project. Figure 17 shows the setup of this scenario at Transpolis. Laser sensors positioned upstream of the pedestrian crossing detect the passage of the Zoé, triggering the system. The pedestrian accelerates quickly to 6.5 km/h, the scenario is configured so that the target will hit the Zoé in the middle of its front bumper if it does not stop. The displacement of the pedestrian is short, 3.5 m, it is directly in front of the Zoé in 2 sec, leaving little time for the Zoé to avoid it.

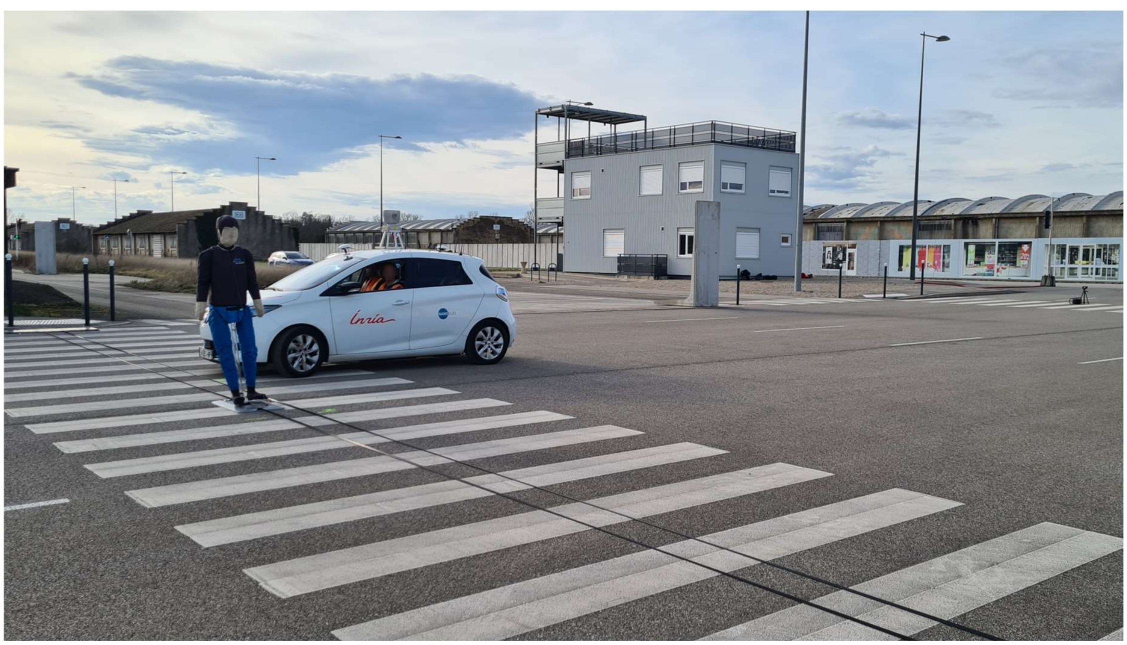


Figure 17 Photo of the setup of scenario 5 with pedestrian target. The target is crossing the road, pulled by the black belt, the Zoé stopped to avoid the collision. The tripod on the pedestrian crossing on the right is a laser system triggering the departure of the target. Surrounding AR actors cannot be seen in this photograph.

21 ROS bags of this scenario were recorded and we obtained a collision rate of 48%, which is a worse result than with the other two pedestrian scenarios, but by a small margin and the behavior of the Zoé in this scenario was similar to that of the other two scenarios. The main categories of behavior observed were:

- The Zoé is able to avoid the collision by abruptly braking in front of the pedestrian. Delays accumulate through the navigation pipeline from the first LiDAR measurements to the tracking of occupancy and velocity. When the delay is too important, the Zoé must strongly brake to avoid the collision. During three tests, the Zoé started braking too late and collided with the target.
- The Zoé perceives that the pedestrian is about to cross the street, and the planner accelerates to cross in front of the pedestrian, usually leaving little safety distance with the pedestrian.

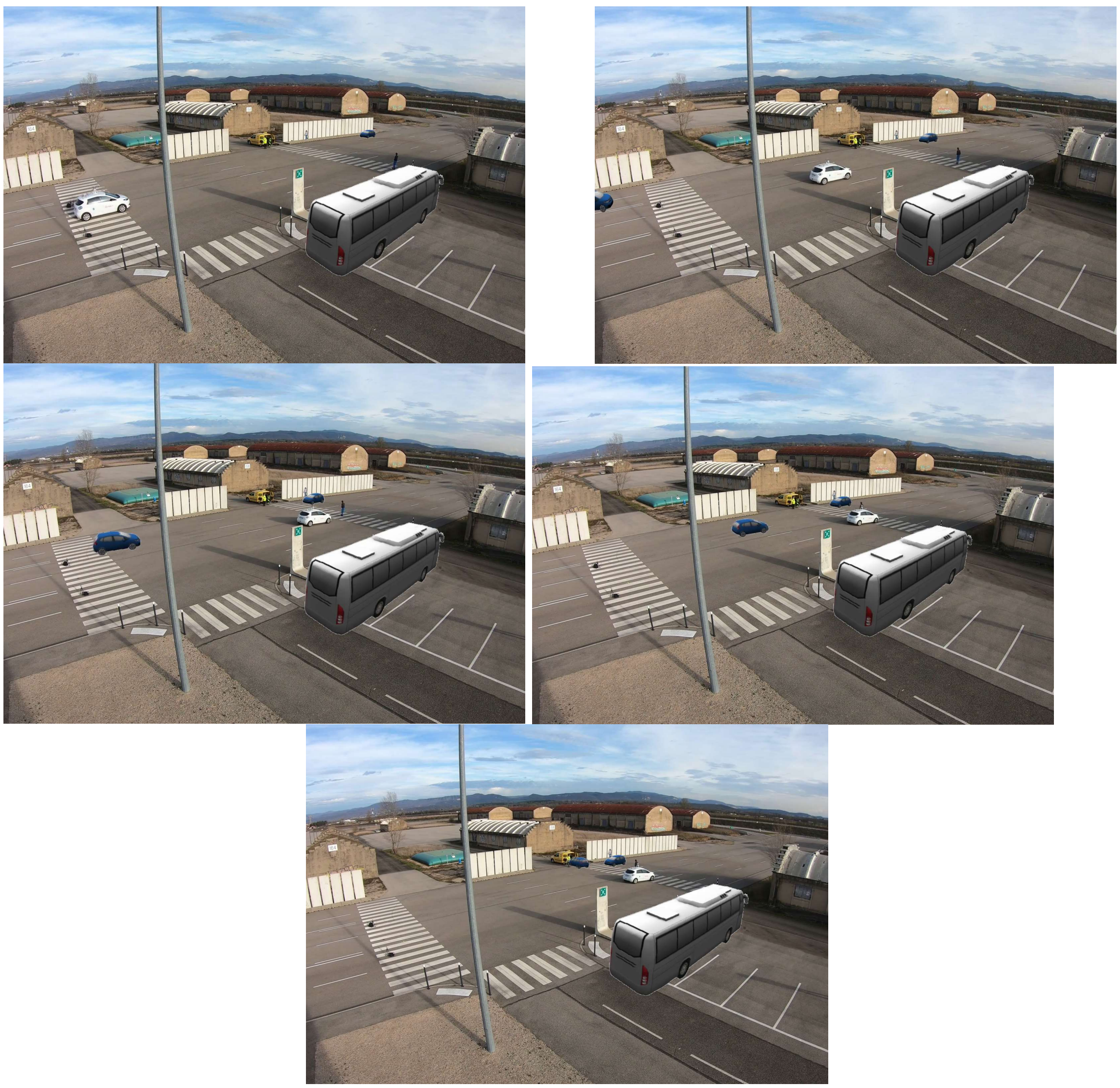


Figure 18 Timelapse of scenario 5 with pedestrian target. The video is augmented with AR actors, a digital twin of the camera was added in the virtual environment and the two videos were merged.

This scenario shows how the AR can be easily integrated with real actors: the design of this scenario is similar to the full-AR scenario 5, with the same virtual actors, except that the real target replaces a virtual pedestrian. We observed a similar behavior of the Zoé navigation system in both configurations, showing how AR can easily be used for the validation process of an AV in hybrid scenarios.

4.7. DISCUSSION

These experimental tests and the first qualitative analysis of their results and of the overall behavior of the autonomous Zoé already allow some conclusions to be drawn.
First, they show how the AR framework allowed us to consider critical scenarios based on accident reports and road user behavior, and to reproduce them safely in a controlled environment without the need to deploy complex and expensive resources. This is a crucial point to prove the feasibility of the proposed approach, which removes many practical and safety constraints in identifying the most relevant scenarios for testing and validating autonomous vehicles and their AI-based components.
Secondly, as expected, the behavior of the Zoé observed in the scenarios selected for this POC and executed at the Transpolis testing facility showed its ability to avoid collisions in most of the cases. However, these tests also revealed some limitation of the current version of the Zoé AI-based perception system to anticipate and react to other road users' actions in critical scenarios, the most important being:

1. Tracking of the other road users' velocity is not always accurate enough, potentially resulting in too much spatial dispersion of the future occupancy prediction.
2. When all possible future trajectories of the Zoé lead to a collision, the planner does not choose evasive maneuvers but emergency braking.

Finally, we identified a few needed improvements to the navigation module. For example, the planner does not always leave enough safety distance from other road users, and we should increase the margins when selecting safe trajectories. In addition, braking is often abrupt and uncomfortable for passengers, and we should have smoother braking commands.

These important results were only revealed by the use of critical scenarios and would not have been possible with more standard tests that do not involve potentially dangerous collisions, proving the importance of integrating the Augmented Reality framework into a validation process.

5. QUANTITATIVE ANALYSIS

5.1. EVALUATION OF OCCUPANCY GRID SIMILARITY

To provide a more thorough quantitative analysis of the data collected during the previously described scenarios, we evaluate the similarity of the occupancy grids generated by the Zoé perception system during the execution of the scenarios with their corresponding ground truth.

The main metrics adopted for this evaluation is the PFC-MSE metric [2], which simulates the behavior of a navigation algorithm on the occupancy grids and evaluates the similarity of the paths generated by the algorithm. The metric is based on the Mean Squared Error of the cost grids generated with the cost of simulated paths. Among the several metrics available in the literature, this is the most relevant for our use-case. Zoé's perception system provides several types of occupancy grids, the most relevant to evaluate is the filtered state grid (published under the ROS topic /state_grid). We evaluated every state grid generated by the Zoé and recorder during the POC, results are shown in the following figures, sorted by bags and scenarios.

5.2. RESULTS

5.2.1. Scenario 1

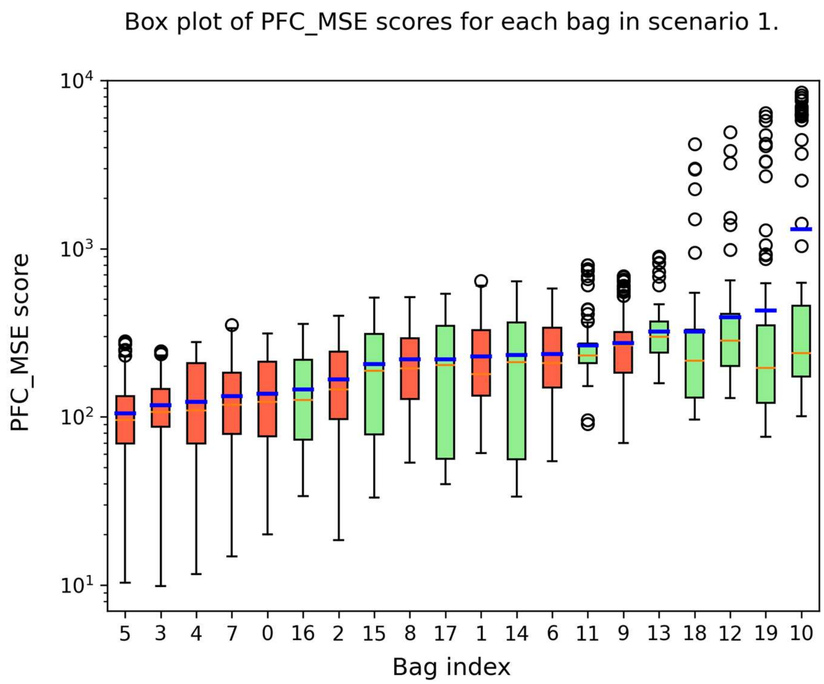


Figure 19 Box plot of PFC-MSE scores for every bag recorder for scenario 1. Y-axis is the PFC-MSE score in log scale, X-axis is the bag index of recording from 0 (first recorded bag) to 19 (last recorded bag). Mean PFC-MSE score sort bags during the bag. Median value is the yellow line; mean value is the blue line. Boxes colored in red correspond to bags leading to a collision and boxes in green correspond to bags leading to an avoidance of the collision.

In Figure 19, most score values from scenario 1 are in the range of 80 to 200, with median and quartiles showing consistent values across the bags, indicating consistent perception system performances. However, the mean scores are systematically higher than the median scores, suggesting a skewed distribution towards higher scores. This is because the PFC-MSE metric increases quadratically with a lower limit of 0 and an upper limit of the grid size squared. Bags with the highest mean scores correspond to bags with the most outliers. Half of the bags led to a collision, the other half to an avoidance. Firstly, as it is the scenario with the highest rate of collision, it is the most critical scenario for the Zoé. Secondly, bags leading to collision tend to have lower mean value (with PFC-MSE lower is better), so we cannot conclude that more accurate perception of the environment leads to a safer navigation. A possible explanation is that perception is just one component of the navigation pipeline, the planner, the controller and the localization produce errors that can lead to a collision. In other words, the perception system is not the only responsible for the safety of the navigation.

5.2.2. Scenario 2

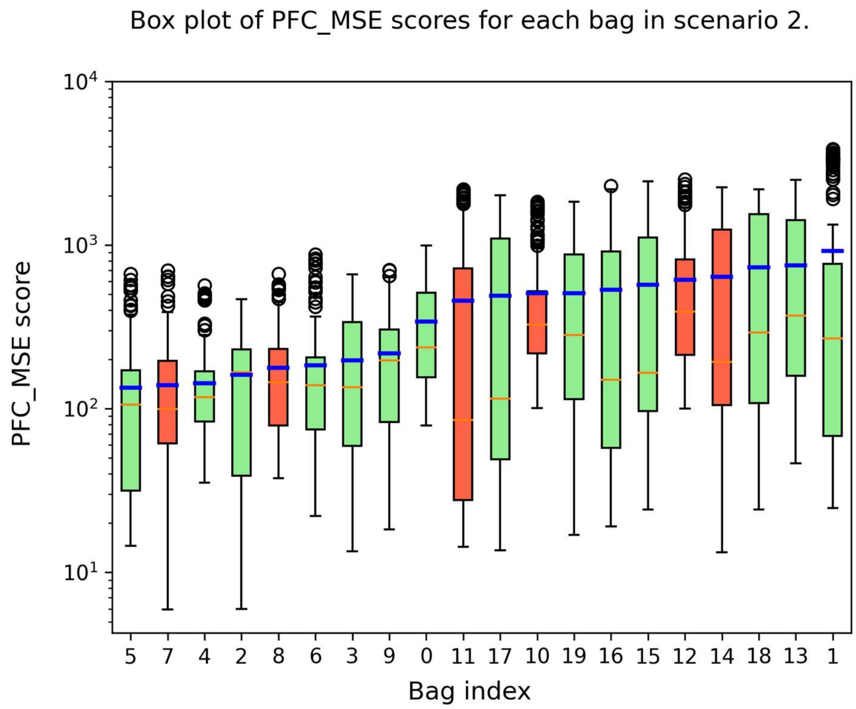


Figure 20 Box plot of PFC-MSE scores for every bag recorder for scenario 2. Y-axis is the PFC-MSE score in log scale, X-axis is the bag index of recording from 0 (first recorded bag) to 19 (last recorded bag). Mean PFC-MSE score sort bags during the bag. Median value is the yellow line; mean value is the blue line. Boxes colored in red correspond to bags leading to a collision and boxes in green correspond to bags leading to an avoidance of the collision.

In Figure 20, most score values from scenario 2 are in the range of 20 to 1000, with variations of median, quartiles and mean values across the bags. The distribution of the scores is more spread than in scenario 1, with a higher number of outliers. The critical situation and the scene of scenario 2 are more complex compared to scenario 1, the trajectory of the incoming vehicle is more difficult to predict and more actors are involved. Despite a less accurate perception, the collision is lower than in scenario 1, the Zoé is able to avoid the collision in 70% of the bags. This scenario is easier for the Zoé, its planner has a tendency to perform straight emergency stops when a collision is expected. In scenario 1, stopping in front of the incoming vehicle is not enough to avoid the collision, the Zoé needs to perform a dynamic avoidance maneuver while in scenario 2, if Zoé stops, the incoming vehicle might have moved back to its lane slightly before the collision.

5.2.3. Scenario 3

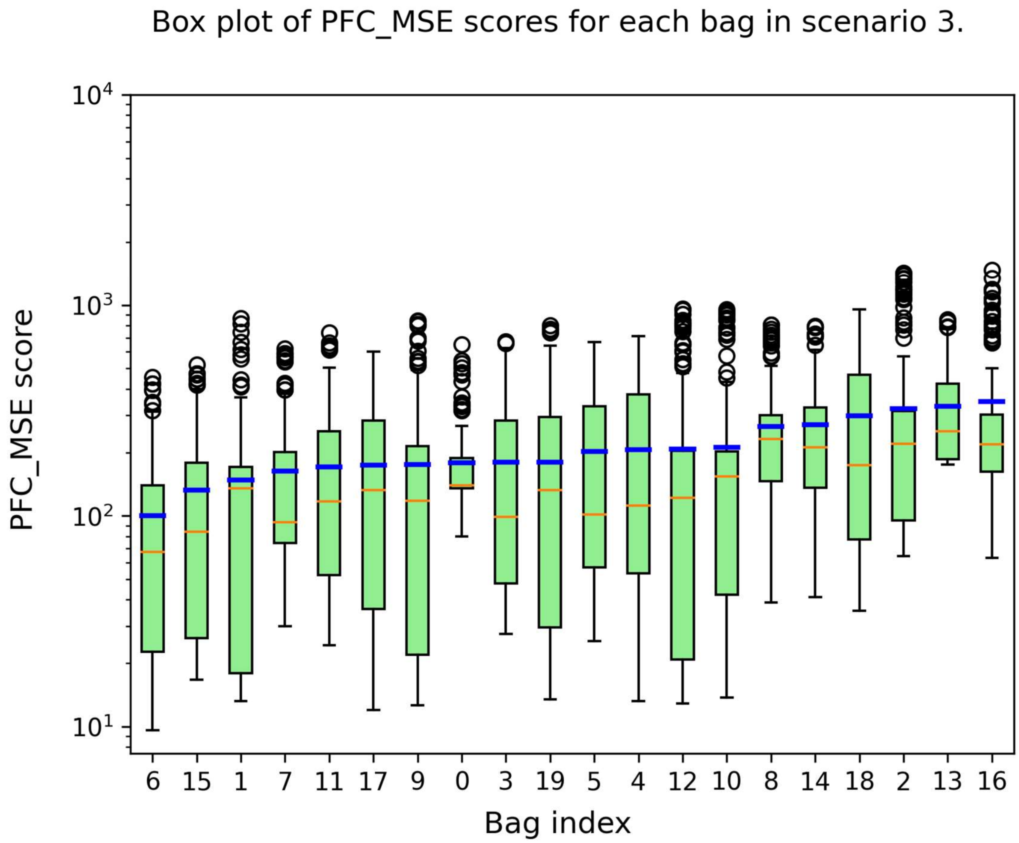


Figure 21 Box plot of PFC-MSE scores for every bag recorder for scenario 3. Y-axis is the PFC-MSE score in log scale, X-axis is the bag index of recording from 0 (first recorded bag) to 18 (last recorded bag). Mean PFC-MSE score sort bags during the bag. Median value is the yellow line, mean value is the blue line. Boxes are in green so, all bags led to an avoidance of the collision.

In Figure 21, most metric scores from scenario 3 are in the range of 20 to 200. While third quartiles and upper values are consistent across the bags, first quartile and lower values are not. Mean performances and worst performances are consistent, it allows finding an upper limit to perception accuracy can be found, while the lower limit cannot be defined which is less critical as it corresponds to a more accurate perception. Scenario 3 is the scenario with the lowest PFC-MSE scores and no bag leading to a collision. The most reasonable explanation is that the scenario is the easiest for the Zoé, the vehicle coming from the right is tracked for several seconds without occlusions, and the Zoé has enough time to anticipate its trajectory and to perform any avoidance maneuver. The increased tracking time compared to other scenarios might also be a reason for the lower PFC-MSE scores.

5.2.4. Scenario 4 and 5

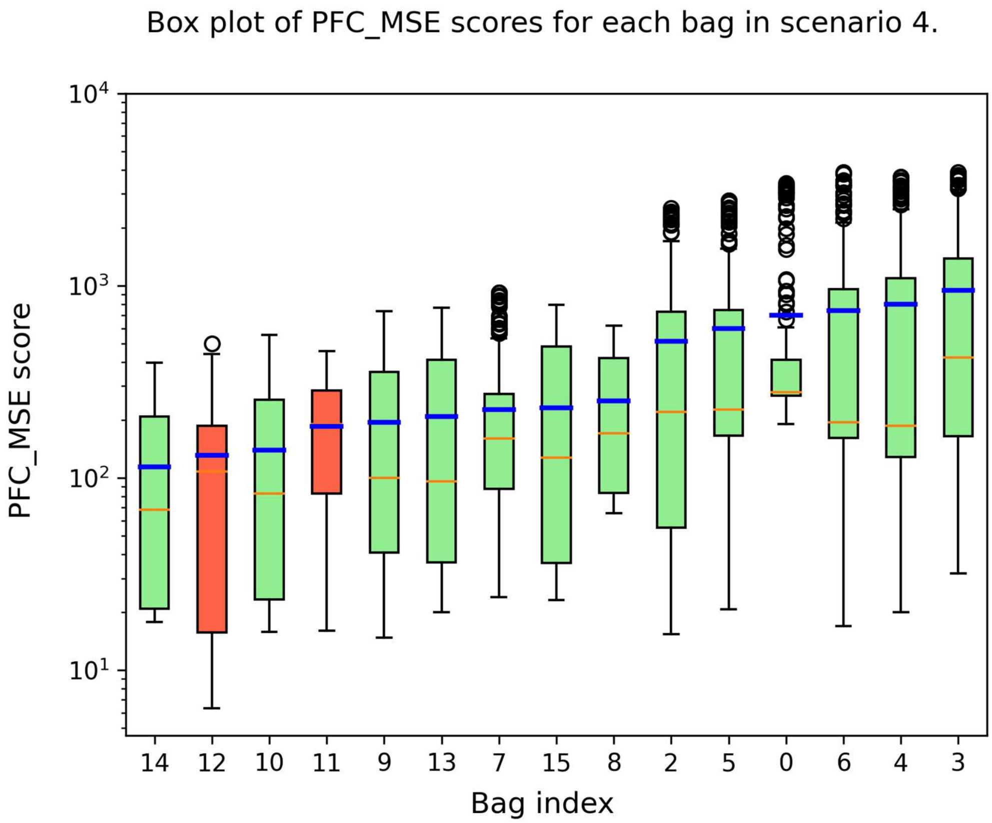


Figure 22 Box plot of PFC-MSE scores for every bag recorder for scenario 4. Y-axis is the PFC-MSE score in log scale, X-axis is the bag index of recording from 0 (first recorded bag) to 15 (last recorded bag). Mean PFC-MSE score sort bags during the bag. Median value is the yellow line, mean value is the blue line. Boxes colored in red correspond to bags leading to a collision and boxes in green correspond to bags leading to an avoidance of the collision.

Scenario 4 and 5 are almost identical, actor trajectories are the same, and the only difference is the occlusion of the pedestrian by the bus in scenario 5. PFC-MSE scores are slightly higher in scenario 5 but it is not significant enough. In fact, the occluded pedestrian in scenario 5 only occupies few cells in the occupancy grid; its impact on the global PFC-MSE score is low. The metric is not focused on evaluating the pedestrian perception (the critical element) but the perception of the environment as a whole. However, collision rate is higher in scenario 5, the particular perception of the pedestrian is less accurate in scenario 5 while the global perception of the scene is equally accurate in both scenarios.

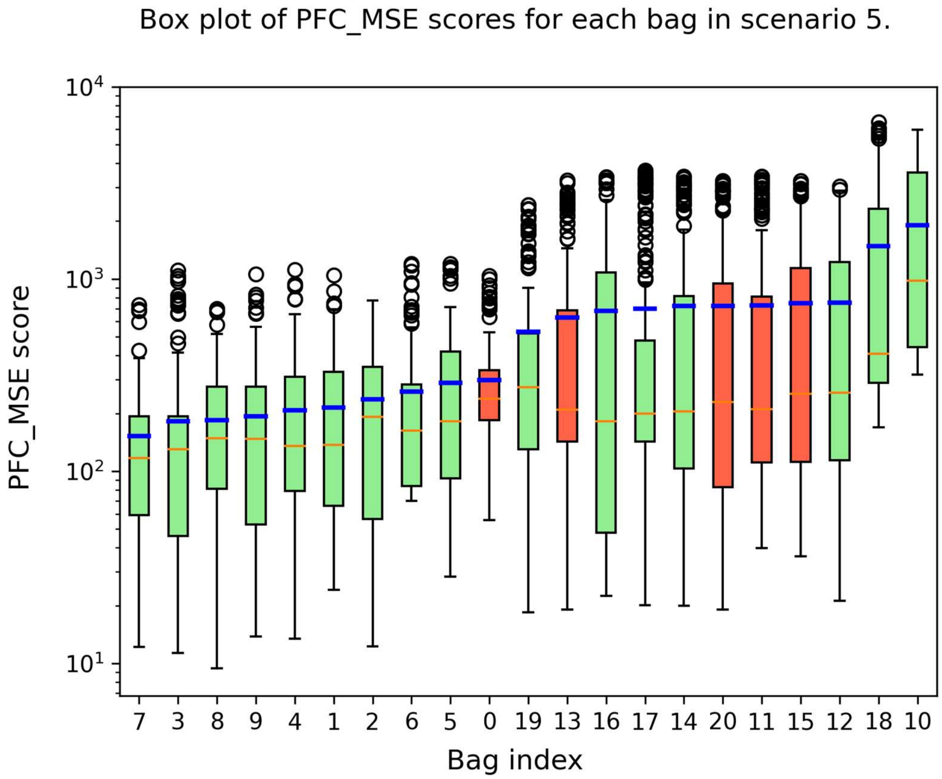


Figure 23 Box plot of PFC-MSE scores for every bag recorder for scenario 5. Y-axis is the PFC-MSE score in log scale, X-axis is the bag index of recording from 0 (first recorded bag) to 20 (last recorded bag). Mean PFC-MSE score sort bags during the bag. Median value is the yellow line, mean value is the blue line. Boxes colored in red correspond to bags leading to a collision and boxes in green correspond to bags leading to an avoidance of the collision.

5.2.5. All scenarios

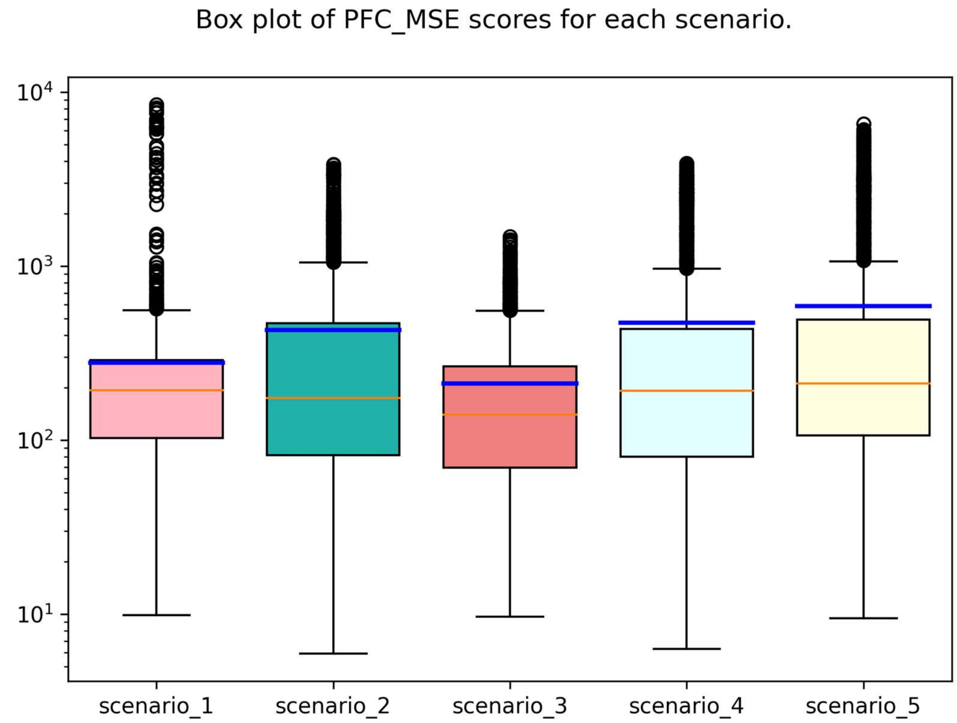


Figure 24 Box plot of PFC-MSE scores for the 5 scenarios. Y-axis is the PFC-MSE score in log scale, X-axis is the scenario name. For each scenario, the PFC-MSE scores are obtained by putting together scores from all the bags of the scenario. Median value is the yellow line, mean value is the blue line.

5.3. DISCUSSION

The study carried out on the occupancy grids generated during the tests of this POC, on the one hand, proved the feasibility and the interest of this analysis and validation protocol and, on the other hand, highlighted some limitations of the process and aspects that could be improved. Namely, it allowed identifying the following limitations:

1. The ground truth data used for comparison against the CMCDOT occupancy grids are an approximation. The ground truth was indeed generated starting from a low-resolution satellite image of Transpolis arena. The resolution of this image is lower than the resolution of the occupancy grids generated by CMCDOT. The grids were scaled to match each other resolution, losing information in the process. The ground truth data is also static and not up to date for the time the satellite image was taken, missing temporary or new modifications of the environment. Moreover, the ground truth was manually labeled and thus subject to human errors and interpretation of what is occupied or not. These incorrect details affected the precision of the data analysis, showing the importance of using a high-resolution digital twin for this phase.

2. The Augmented Reality framework relies heavily on the correct Zoé localization to merge virtual and real LiDAR point clouds. The accuracy of the AR actor merging in the point clouds is the accuracy of the localization. In addition, since the AR simulation data is used to label the

ground truth with actor occupancy, errors in the AR can lead to errors in the ground truth. These issues require the localization of the Zoé to be perfectly calibrated at the testing field before the tests (IMU, magnetometer, RTK GPS, etc.).

3. Bags were labeled for collision or avoidance based on the Zoé on-board visualization software during the test executions, which can have delays in displaying the occupancy grids and the actors. This is also subject to human errors and interpretation of the proximity of the obstacles to the Zoé.

REFERENCES

[1] T. Genevois, J.-B. Horel, A. Renzaglia, and C. Laugier, "Augmented reality on lidar data: Going beyond vehicle-in-the-loop for automotive software validation," in 2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022, pp. 971–976.

[2] J.B. Horel, R. Baruffa, L. Rummelhard, A. Renzaglia and C. Laugier, "A Navigation-Based Evaluation Metric for Probabilistic Occupancy Grids: Pathfinding Cost Mean Squared Error", IEEE International Conference on Intelligent Transportation Systems (ITSC), 2023.

[3] V. Judalet, L. Désiré, G. Saint Pierre, N. De Rus, G. Thiolon. Livrable 4.1: Choix des scénarios d'interactions VA / VL. *Projet Surca*, financé par la FSR et la DSR, 15 p. 2020.

[4] T. Serre, V. Battaglia, C. Chauvel. Livrable 6.1: Choix des scénarios d'interactions VL/piétons et cyclistes. *Projet Surca*, financé par la FSR et la DSR, 15 p. 2020.

1. ANNEX A - List of recorded ROS topics

Topic name	Topic type	Description
/zoe/velodyne_points	sensor_msgs/PointCloud2	Point clouds of the Velodyne HDL-64 LiDAR
/zoe/lux_right	sensor_msgs/PointCloud2	Point clouds of the front right, front center, front left and rear Ibeo Lux LiDARs
zoe/lux_center	sensor_msgs/PointCloud2	
/zoe/lux_left	sensor_msgs/PointCloud2	
/zoe/lux_rear	sensor_msgs/PointCloud2	
/temp/zoe/velodyne_packets	velodyne_msgs/VelodyneScan	Raw data measurements from the Velodyne HDL-64
/zoe/classified_cloud	sensor_msgs/PointCloud2	Merged point cloud from the 5 LiDARs with classification of ground
/zoe/us_right	sensor_msgs/Range	Front ultrasonic range sensors
/zoe/us_center	sensor_msgs/Range	
/zoe/us_left	sensor_msgs/Range	
/zoe/sp90_fix	sensor_msgs/NavSatFix	Satellite localization of the Zoé
/zoe/sp90_time_reference	sensor_msgs/TimeReference	
/zoe/fix	sensor_msgs/NavSatFix	
/zoe/fix_common	gps_common/GPSFix	
/zoe/raw_fix	sensor_msgs/NavSatFix	
/zoe/camera_front/image_rect_color	sensor_msgs/Image	Images stream of the front camera
/zoe/camera_front/camera_info	sensor_msgs/CameraInfo	Information about the camera and its calibration
/zoe/imu/mag	sensor_msgs/MagneticField	Magnetic compass of the Zoé IMU
/zoe/imu/data	sensor_msgs/Imu	IMU data (orientation, angular velocity and linear acceleration)
/navigation/dwa_result	dwa_dynamic_planner/Trajectory	Current trajectory of the Zoé generated by the local planner
/navigation/planner_result	dwa_dynamic_planner/PlannerResult	Status information on the local planner
/navigation/planner_status	dwa_dynamic_planner/PlannerStatus	
/zoe/velocity_grid	e_motion_perception_msgs/VelocityGrid	Grid of velocity vectors of the dynamic cells
/zoe/state_grid	e_motion_perception_msgs/FloatOccupancyGrid	Grid of filtered probability of occupied, dynamic, static and unknown
/zoe/occ_grid	e_motion_perception_msgs/FloatOccupancyGrid	Grid from one LiDAR point cloud of probabilities of occupied and unknown. Output of the LiDAR sensor model
/zoe/control/refs	ros_zoe_msgs/ControlRefs	Throttle, brake and steering commands sent to the hardware controller of the Zoé for automated driving
/tf	tf2_msgs/TFMessage	Dynamic and static transforms of the frames of the Zoé
/tf_static	tf2_msgs/TFMessage	
/zoe/velocity	geometry_msgs/TwistStamped	Velocity of the Zoé
/zoe/speed	geometry_msgs/TwistStamped	
/zoe/pose	geometry_msgs/PoseWithCovarianceStamped	Filtered Pose of the Zoé by a Kalman filter. Relative to a world fixed frame
/gazebo/set_model_state	gazebo_msgs/ModelState	States and status of the virtual Actors in Gazebo
/gazebo/link_states	gazebo_msgs/LinkStates	
/gazebo/model_states	gazebo_msgs/ModelStates	
/gazebo_scenario/rosparam	std_msgs/String	JSON serialization of all ROS parameters of the Zoé
/gazebo_scenario/scenario	std_msgs/String	JSON serialization of the scenario description and parameters

Table 1: The experiment captured diverse data streams using the rosbag tool.

1. ANNEX B - Short description of each scenario execution

ROS bag	Description
inria_zoe_2024_02_15_10_29_55/	Rear collision without trying lateral avoidance.
inria_zoe_2024_02_15_10_38_12/	Rear collision without trying lateral avoidance.
inria_zoe_2024_02_15_10_39_44/	Rear collision without trying lateral avoidance.
inria_zoe_2024_02_15_10_40_45/	Rear collision without trying lateral avoidance.
inria_zoe_2024_02_15_10_41_43/	Rear collision without trying lateral avoidance.
inria_zoe_2024_02_15_10_43_48/	Rear collision without trying lateral avoidance. Zoé abruptly stopped then restarted by going backward.
inria_zoe_2024_02_15_10_45_35/	Rear collision without trying lateral avoidance.
inria_zoe_2024_02_15_10_46_39/	Rear collision without trying lateral avoidance. Zoé abruptly braked twice.
inria_zoe_2024_02_15_10_48_23/	Rear collision without trying lateral avoidance. Zoé smoothly braked.
inria_zoe_2024_02_15_10_52_03/	Rear collision without trying lateral avoidance. Zoé abruptly braked.
inria_zoe_2024_02_16_11_22_37/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_24_33/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_26_30	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_29_08/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_31_05/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_32_31/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_33_25/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_34_58/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_36_36/	Collision avoided, Zoé dynamically avoided collision by steering to the left.
inria_zoe_2024_02_16_11_39_43/	Collision avoided, Zoé dynamically avoided collision by steering to the left.

Table 2: Scenario 1

ROS bag	Event description
inria_zoe_2024_02_15_09_53_04/	Collision avoided, Zoé abruptly braked twice.
inria_zoe_2024_02_15_09_54_37/	Collision avoided, Zoé smoothly braked.
inria_zoe_2024_02_15_09_56_11/	Collision avoided, Zoé smoothly braked due to false positive then smoothly braked for true obstacle.
inria_zoe_2024_02_15_09_57_37/	Collision avoided, Zoé smoothly braked.
inria_zoe_2024_02_15_09_59_25/	Collision avoided, Zoé smoothly braked but then performed an unexpected emergency stop.
inria_zoe_2024_02_15_10_01_06/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_15_10_02_51/	Collision avoided, Zoé abruptly braked.

inria_zoe_2024_02_15_10_04_04/	Front collision, Zoé did not tried to avoid obstacle.
inria_zoe_2024_02_15_10_05_40/	Front collision, Zoé braked to try to avoid obstacle.
inria_zoe_2024_02_15_10_07_31/	Collision avoided, Zoé did an erratic avoidance by brakin twice and steering right then left.
inria_zoe_2024_02_16_13_15_15/	Front collision
inria_zoe_2024_02_16_13_16_26/	Front collision
inria_zoe_2024_02_16_13_17_51/	Front collision
inria_zoe_2024_02_16_13_19_47/	Collision avoided
inria_zoe_2024_02_16_13_20_34/	Front collision, Zoé almost avoided collision.
inria_zoe_2024_02_16_13_21_36/	Collision avoided, Zoé almost collided.
inria_zoe_2024_02_16_13_22_38/	Collision avoided, Zoé almost collided.
inria_zoe_2024_02_16_13_23_31/	Collision avoided, Zoé almost collided.
inria_zoe_2024_02_16_13_25_01/	Collision avoided, Zoé almost collided.
inria_zoe_2024_02_16_13_26_17/	Collision avoided, Zoé steered to the right, almost collided.

Table 3: Scenario 2

ROS bag	Event description
inria_zoe_2024_02_16_10_43_56/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_45_03/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_47_19/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_49_56/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_51_01/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_52_17/	Collision avoided, Zoé abruptly avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_53_13/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_54_30/	Collision avoided, Zoé smoothly avoided incoming vehicles.
inria_zoe_2024_02_16_10_56_38/	Collision avoided, Zoé abruptly avoided incoming vehicles.
inria_zoe_2024_02_16_10_57_38/	Collision avoided.
inria_zoe_2024_02_16_10_58_42/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_10_59_59/	Collision avoided, Zoé abruptly avoided incoming vehicles.
inria_zoe_2024_02_16_11_00_54/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_11_11_18/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.
inria_zoe_2024_02_16_11_12_31/	Collision avoided, Zoé abruptly avoided incoming vehicles.
inria_zoe_2024_02_16_11_13_25/	Collision avoided, Zoé dynamically avoided vehicle from the right and then vehicles from opposite lane.

inria_zoe_2024_02_16_11_14_25/	Collision avoided, Zoé abruptly avoided incoming vehicles.
inria_zoe_2024_02_16_11_15_25/	Collision avoided, Zoé braked to avoid vehicle from the right.
inria_zoe_2024_02_16_11_16_41/	Collision avoided, Zoé abruptly avoided incoming vehicles.
inria_zoe_2024_02_16_11_17_43/	Collision avoided

Table 4: Scenario 3

ROS bag	Event description
inria_zoe_2024_02_16_13_59_48/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_16_14_02_07/	Collision avoided, Zoé abruptly braked twice.
inria_zoe_2024_02_16_14_03_03/	Collision avoided, Zoé abruptly braked toward pedestrian.
inria_zoe_2024_02_16_14_04_17/	No description from experiment.
inria_zoe_2024_02_16_14_05_10/	No description from experiment.
inria_zoe_2024_02_16_14_06_38/	No description from experiment.
inria_zoe_2024_02_16_14_07_48/	No description from experiment.
inria_zoe_2024_02_16_15_07_29/	Collision avoided.
inria_zoe_2024_02_16_15_08_25/	Collision avoided, Zoé almost collided with pedestrian.
inria_zoe_2024_02_16_15_09_16/	Collision avoided, Zoé almost collided with pedestrian.
inria_zoe_2024_02_16_15_10_08/	Collision avoided.
inria_zoe_2024_02_15_10_59_02/	Collision avoided.
inria_zoe_2024_02_15_11_02_10/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_11_03_33/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_11_04_44/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_11_06_09/	Side Collision, Zoé smoothly braked but could have avoided collision with stronger braking.
inria_zoe_2024_02_15_11_07_20/	Side Collision, Zoé abruptly braked twice
inria_zoe_2024_02_15_11_08_49/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_11_10_02/	Collision avoided, Zoé performed a noteworthy smooth braking.
inria_zoe_2024_02_15_11_11_17/	Outcome not reported, Zoé abruptly braked twice.
inria_zoe_2024_02_15_11_12_52/	Collision avoided, Zoé abruptly braked twice, but did not stop, then dynamically avoided the pedestrian.

Table 5: Scenario 4

ROS bag	Event description
inria_zoe_2024_02_15_12_07_28/	Front collision, Zoé abruptly braked thrice.
inria_zoe_2024_02_15_12_08_50/	Collision avoided, Zoé smoothly and dynamically avoided toward right.
inria_zoe_2024_02_15_12_10_01/	Collision avoided, Zoé smoothly and dynamically avoided toward right.
inria_zoe_2024_02_15_12_10_56/	Collision avoided, Zoé smoothly braked.
inria_zoe_2024_02_15_12_12_07/	Collision avoided, Zoé smoothly braked.
inria_zoe_2024_02_15_12_13_40/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_15_12_15_24/	Collision avoided

inria_zoe_2024_02_15_12_16_27/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_15_12_17_22/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_15_12_18_27/	Collision avoided, Zoé almost collided with pedestrian.
inria_zoe_2024_02_16_14_54_40/	Collision avoided
inria_zoe_2024_02_16_14_56_01/	Front collision
inria_zoe_2024_02_16_14_57_22/	Collision avoided, Zoé almost collided with pedestrian.
inria_zoe_2024_02_16_14_58_12/	Side Collision
inria_zoe_2024_02_16_14_59_05/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_16_15_00_11/	Collision avoided, Zoé abruptly braked.
inria_zoe_2024_02_16_15_01_07/	Collision avoided
inria_zoe_2024_02_16_15_01_59/	Collision avoided
inria_zoe_2024_02_16_15_03_58/	Collision avoided
inria_zoe_2024_02_16_15_04_55/	Collision avoided, Zoé slowdowned because of vehicle from opposite lane.
inria_zoe_2024_02_16_15_05_53/	Side Collision

Table 6: Scenario 5

ROS bag	Event description
inria_zoe_2024_02_15_14_48_46/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_14_52_56/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_14_56_41	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_14_59_04/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_15_05_43/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_15_10_18/	Driver intervention, collision avoided by manual braking, Zoé started braking too late.
inria_zoe_2024_02_15_15_12_56/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_15_17_16/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_15_18_31/	Driver intervention, collision avoided by manual braking, Zoé started braking too late.
inria_zoe_2024_02_15_15_44_43/	Collision avoided
inria_zoe_2024_02_15_15_47_00/	Collision avoided
inria_zoe_2024_02_15_15_49_22/	Collision avoided, Zoé dynamically avoided the pedestrian.
inria_zoe_2024_02_15_15_50_40/	Collision avoided, Zoé stopped to close to the target.
inria_zoe_2024_02_15_15_52_52/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_15_54_53/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_15_56_38/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_15_58_27/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_16_00_51/	Collision avoided, Zoé stopped too close to the target.
inria_zoe_2024_02_15_16_03_11/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_16_06_43/	Driver intervention, collision avoided by manual braking.
inria_zoe_2024_02_15_16_08_29/	Collision avoided, Zoé stopped too close to the target.

Table 7 Scenario pedestrian target

Chapter 3: CEREMA/LNE POC

1. INTRODUCTION

More and more intelligent systems on vehicles use AI (e.g., visual or mixed navigation, sign recognition, road tracking and obstacle detection). Certification up to SAE level 3 is now possible for vehicles featuring partially automated driving. The manufacturer must demonstrate that its vehicles ensure adequate safety conditions within their operational design domain (ODD), having conducted tests in diverse scenarios. In particular, this task concerns the first braking-related advanced driver assistance system (ADAS) that has been implemented as an ``Automatic Emergency Braking" (AEB). The qualification of these systems requires verification in all kinds of scenarios, including, for example, considering degraded weather conditions. For cost and safety reasons, these qualification tests cannot be carried out in real conditions, as some tests may present risks or have frequencies of occurrence too low to allow the collection of large series of data. For this reason, sensor simulation tools and degraded weather conditions (physical, numerical or hybrid) must be implemented. These simulation tools can be real (like in the PAVIN Fog and Rain platform), purely virtual (integrating sensor models, as in LEIA 1 and with the Cerema fog model) or can combine the physical sensor with simulated inputs (as is done in LEIA 2). The purely virtual simulators can be physically based or empirical (mainly based on Beer-Lamber theory). The latter family of simulators is realtime but the not the former. In the language of certification, which is now being established, we speak about X in the Loop (XiL) testing, with X representing the Software, the sensor (Hardware) or the entire Vehicle. In the HiL and ViL cases, we can imagine that the vehicle's real sensor is fooled by a screen system that makes the vehicle believe it's seeing things that don't exist. The advantage of not relying solely on software during simulation is that other disruptive elements can be considered during testing, such as sensor electronics, system response times, or vehicle dynamics in the case of ViL. These simulation tools need to be validated and qualified, as they may be used for certification. In particular, it is necessary to check these points:

- What scenarios should be considered to guarantee the results obtained on AI-based algorithms in the context of certification? In other words, what are the minimum scenario combinations to guarantee a given level of error and uncertainty during evaluation?
- The repeatability of an evaluation with the same tool: what is the uncertainty induced by the simulation tool on the evaluation?
- The reproducibility of a test from one tool to another: what are the differences in results between the different simulation tools (real or numerical)?

LNE and Cerema have different tools for AI systems evaluation at their disposal, that need to be qualified:

- Cerema's PAVIN Fog and Rain platform for producing artificial fog and rain.
- Cerema's K-HiL model that allows adding fog to real images in augmented reality mode.
- LNE's LEIA 1 simulator to create fully digitally simulated images.
- LNE's LEIA 2 to replay videos recorded and or fully simulated in front of a real camera, in order to address the HiL purpose.

The aim of this document is to present the results obtained to verify the various objectives mentioned above. These results should enable simulation tools to be compared with each other, and to characterize scenarios that enable repeatable evaluation with a known level of error.

1.1. CONTEXT

As it is not possible to deal with every type of function at once, we've chosen to concentrate on the pedestrian detection function. This function is a priority, since it guarantees the safety of Vulnerable Road Users (VRUs). It is also already approved for the AEB function. Our three objectives are as follows:

First, how can we guarantee that we have tested a wide enough range of conditions? AI-based algorithms are black boxes and it is, therefore, very difficult to find their boundary conditions. Indeed, the typology, position, and orientation of the pedestrian can influence the results of the algorithm. Similarly, the environment, disturbing objects, and occlusions can influence the detection. Beyond these geometric issues, weather conditions also have strong impacts, e.g., illumination, camera glare, fog, rain, and snow. Interest in this issue is recent in the field of autonomous vehicles and is the subject of numerous studies, but at present, the works listed in the literature only present particular cases and not a global solution.

Even if all the conditions required for successful validation have been identified, it is impossible to reproduce them all in real-world conditions. For this, one solution is to use numerical simulation. Many numerical simulators dedicated to autonomous vehicles exist. Most offer variants regarding pedestrians, environments, or weather, but only a few are calibrated against real-world conditions, to our knowledge. The second question is: how can we validate the realism and representativeness of a digital simulator? Will the behavior of artificial intelligence be the same in front of different simulators? To address more exhaustive scenarios, the data can be partially or totally simulated, so X-in-the-loop simulators appear to allow to use augmented reality mechanisms. These are simulation tools of this type that we propose to test in this protocol (K-HiL model and LEIA 1).

Beyond numerical simulation, real simulation methods are used to simulate adverse weather conditions. This is the case with the PAVIN fog and rain platform, which can reproduce adverse weather conditions on demand. This platform is calibrated from a meteorological point of view (calibration of intensities, drop size, and velocity). A real physical test must be qualified from a repeatability point of view. In the same way, the repeatability of virtual simulators is closely linked to the determinism of the simulator algorithms. Several sources can affect the determinism of the simulation, they can be classified into two categories, first are due to the algorithms of simulator itself such as randomness, and stochastic processes and the others are due to the hardware and operating system, which hosts the simulator such as floating-point arithmetic or parallelism and concurrency between processes. This is essential in the context of certification tests, where test laboratories are often qualified and audited, making repeatability tests and uncertainty measurements mandatory. Can this type of platform guarantee the repeatability of tests, as well as a standard deviation on the results obtained with AI?

In an attempt to answer these questions, this protocol introduces a new pedestrian database, focusing on weather (clear weather and fog) and an associated evaluation method of detection tools. That database comprises real data, gathered in clear weather and artificial fog conditions within the PAVIN fog and rain platform, and numerically simulated data (using the digital twin), executed in HiL mode, from a simplistic model prevalently used in most numerical simulators outlined in existing literature. Both real and simulated data are annotated with 2D pedestrian detection bounding boxes.

1.2. METHODOLOGY OVERVIEW

The aim of the present study is to characterize and evaluate the protocols and simulation tools enabling AI algorithm certification, including degraded weather conditions (fog). The evaluation of the used proving AI-based algorithm is outside the scope of this study. The proposed method is therefore as follows. First, an AI-based algorithm, which is applied to the intelligent vehicle and representative of the state of the art, is chosen. This algorithm will be used as a proving algorithm for the qualification of the simulation tools. A metric applied to this algorithm will be chosen. Then identical datasets will be prepared using the different simulation tools available. These datasets will have to include adapted scenarios to evaluate the identified proving algorithm. Indeed, they will include data in clear weather and foggy conditions, but also repeated scenarios to verify repeatability. Finally, the proving algorithm and the associated metric will be applied to all the datasets. A comparison of the scores obtained for each dataset will allow verifying reproducibility from one simulation tool to another. At the same time, this method will make it possible to discuss the repeatability of tests with a single simulation tool, and the minimum protocols to be put in place to guarantee error-free evaluation of pedestrian detectors. The following sections present in detail the protocol, the tests carried out and the databases obtained, the metrics implemented and the simulation tools involved.

2. EVALUATION PROTOCOL

2.1. FUNCTION TO BE ASSESSED

As already mentioned, there is a very wide variety of algorithms using AI for autonomous vehicles. Then we have chosen to limit ourselves to the example of pedestrian detection. As a reminder, the aim of the present study is to characterize and evaluate the protocols and simulation tools that are used for the certification, and not the pedestrian detection algorithm itself, which only serves as proving algorithm.

Concerning the pedestrian detection algorithm, the third version of YOLO detection algorithm, which stands for "You Only Look Ones", was chosen in this analysis. It is indeed a very common algorithm in the literature on object detection. Moreover, it is very easy to handle. The library of objects available in this version contains 80 items. The algorithm requires two main parameters: the confidence threshold (a value between 0 and 1) of the labeling and the object to label in the images. Only the class "person" is labeled in this study and the confidence threshold chosen is explained in the following section. A frame can get multiple detections with different level of confidence even though only one pedestrian is walking in the scene into our database.

2.2. DATABASE

In this study, we want to compare the following simulation tools: the PAVIN Fog and Rain platform, the K-HiL fog model, the LEIA 1 digital simulator (digital twin), the LEIA 2 simulator to better address HiL purpose. To this end, we propose to acquire the same data for these different simulation tools. We will then compare the results obtained in the different cases. We will also try to measure the uncertainty for some of them.

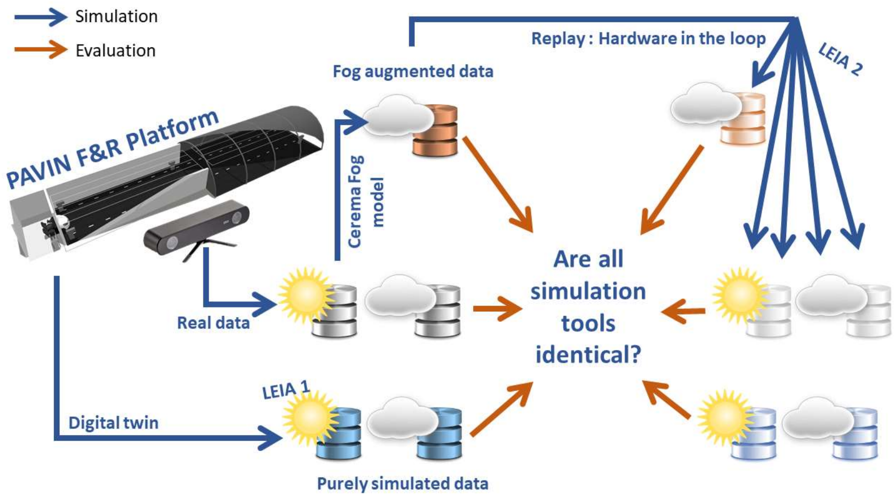


Figure 25: Presentation of the protocol used to obtain the various databases to be compared as part of the POC.

Figure 25 shows the different databases that will be compared. As it will be described in next sections, each is a complete database containing pedestrians walking in a road scene. Our proving algorithm (AI pedestrian detector) will be applied on each of these databases. The aim is to compare whether the scores obtained on each of them are similar.

The process used to obtain these different databases is complex and needs to be described in detail. Indeed, some of the simulation tools used enable real data to be augmented (Cerema's K-HiL model), while others enable data to be replayed to add the Hardware in the Loop aspect (LNE's LEIA 2). Figure 25 shows a schematic diagram of how each database is created.

First of all, a first database is created in clear weather and fog conditions on the PAVIN platform. To achieve this, a camera records real pedestrian making their way through a road scene (dark gray base). The PAVIN platform can reproduce artificial fog conditions on demand. After the actual tests, a database with real images acquired in clear weather and fog conditions is available (dark gray base).

Next, fog is added to the real data acquired in clear weather using the K-HiL model (dark orange base). This model enables fog to be simulated digitally over an image acquired by a real camera. Once the model has been applied, a second database with digitally simulated fog is available.

Thanks to a digital twin of the platform (3D model), the same scenarios are reproduced independently in the LEIA 1 simulator (dark blue base). It enables the same data to be created in a virtual world (full 3D simulation). This makes it possible to obtain a third database with clear weather and fog conditions.

From these three simulation tools, one real database (dark gray), one SiL database (dark blue) and one HiL database (dark orange) are obtained. To better address HiL simulations, we use the LEIA 2 simulator. This simulator enables us to replay a database in front of the real camera, in order to obtain images from the real camera, as if it had filmed the scene itself. This is important in the context of vehicle evaluation, as it enables the entire processing chain to be included in the evaluations (sensor, electronics, cables, central processing unit, etc.). The LEIA 2 simulator is therefore used to replay the PAVIN, LEIA 1 databases (dark colors) in front of the camera, resulting in 3 new databases taking HiL into account (bright colors).

In the end, there are 5 databases from various simulation tools to compare. Each of these five variants contains identical clear weather and fog conditions. These six variants are named and summarized in Table 8. In particular, the variants replayed in LEIA 2 have the same name with an *.

Table 8: Nomenclature and description of the databases used in the POC.

Variant name	Location / acquisition method	Type	Peds num.	Total number of videos
PAVIN	A real camera records pedestrian on the PAVIN platform. The platform can reproduce clear weather or fog.	Real	100	3 weather conditions * 100 pedestrians * 2 sequences = 600
K-HiL	Camera data from the PAVIN database (clear weather) is reused. Using the K-HiL simulator, digital fog is added to the images.	HiL	100	2 weather conditions * 100 pedestrians * 2 sequences = 400
LEIA	The platform's digital twin is used to recreate scenarios in an entirely virtual world, thanks to the LEIA 1 simulator.	SiL	36	3 weather conditions * 36 pedestrians * 2 sequences = 216
PAVIN*	PAVIN database replay into LEIA 2.	HiL	100	3 weather conditions * 100 pedestrians * 2 sequences = 600
LEIA*	LEIA database replay into LEIA 2.	HiL	36	3 weather conditions * 36 pedestrians * 2 sequences = 216

Table 8 shows the nomenclature of each database. It also shows the volume of data in each database. As described in the next part, the weather conditions chosen include clear weather, and two fog conditions (medium fog and dense fog). In addition, for each pedestrian/weather combination, the route was replayed and recorded twice, so that repeatability measurements could be made. As a result, a total of 2,432 video sequences are available. Each sequence lasts around one minute, so there are around 40 hours of real, partially or fully simulated videos in the final database.

The first objective of the study is therefore to compare the similarity of the 6 variants. For this purpose, a metric is defined in the next section. The second objective of the study is to measure the repeatability and uncertainty of a pedestrian detector evaluation. To this end, each of the databases will be randomly split into sub-sections. The metric will then be applied to each part as shown in Figure 26. This will also be analyzed. The general structure of the tests has been described; the following section presents the metrics used for the evaluation.

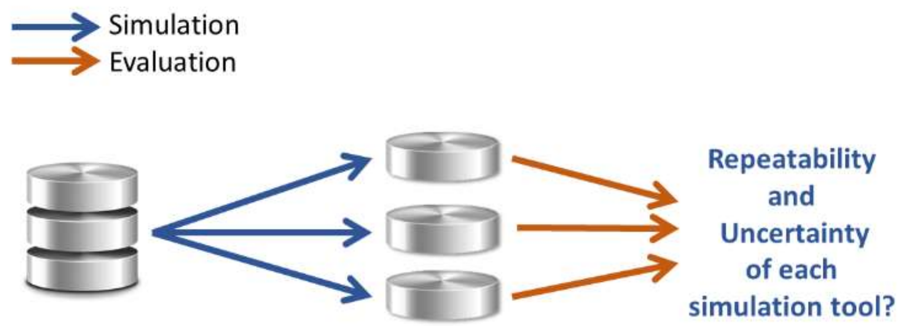


Figure 26: Method used to check repeatability and uncertainty. One of the variants is divided into subgroups, then the score is measured on each subgroup.

2.3. DESCRIPTION OF TESTS

2.3.1. SENSORS

As described above, the aim is to use an AI-based pedestrian detector for cameras to validate simulation tools. The stereo camera ZED2i (See next figure) from StereoLab has been chosen and purchased by Cerema for the data acquisition. Indeed, the latter will allow the testing of monocular detection and recognition algorithms (by taking only one channel) but also stereoscopic. This will allow proposing a database in agreement with the literature. Cerema will also make acquisitions in parallel with a thermal camera (Xenics). This will allow labeling the images of the ZED2i camera in dense fog conditions, thanks to a preliminary geometrical calibration. In fact, the pedestrian is almost invisible on the ZED camera in dense fog, which makes labelling very complicated. The different instruments were positioned at the beginning of the greenhouse (See Figure 30). Finally, the PAVIN platform's usual sensors record meteorological conditions.

Figure 27: StereoLab's ZED2i camera.

2.3.2. METEOROLOGICAL CONDITIONS

The objective of the scenarios defined for this study is to collect videos containing pedestrians moving in a scene subjected to various weather conditions (clear weather and two types of fog) and seasons using clothing representative of summer or winter.

Fog is characterized in meteorology by the Meteorological Optical Range (MOR), also called visibility, and noted as V. MOR, expressed in meters, corresponds to the distance at which the human eye no longer perceives contrast on a calibrated white-and-black target. The smaller the MOR, the denser the fog. It is considered that there is the presence of fog for a MOR below 1000 m in meteorology and below 400 m in road context.

The three types of weather conditions chosen are:

- Clear weather (CW): it allows having a reference scene without disturbances due to the presence of fog.

- Medium fog (MF): the visibility is of 23 m allowing modifying the general aspect of the objects of the scene by leaving detectable all the elements of the visible scene.
- Heavy fog (HF): the visibility is of 10 m allowing elements of the background to disappear for stereo camera but not for thermal camera.

These MOR values were chosen to obtain critical fog conditions. Thus, it is certain that these conditions will challenge the proving detection algorithm. Subsequently, the scores obtained by the latter will drop down, which will allow us to check whether the scores are similar for physical fog and numerically simulated fog. The next figure shows an example of the images obtained for the three weather conditions of the real data.

Figure 28: Three weather conditions for a daytime configuration of the scene with (from left to right): clear weather (CW), medium fog (MF), and dense fog (DF).

2.3.3. SCENE

To recreate a realistic environment, an urban scene with different elements was created in the PAVIN Fog and Rain platform: a Renault Megane vehicle, trees, a wooden picnic table, different traffic signs, ground marking strips, and orange traffic cones, as well as four calibrated targets (a large black and a large grey (50 x 50 cm), and a small white and a small black (30 x 30 cm). A 3D model (digital twin) with all the elements of this scene is also available with the dataset.

Figure 29: Daytime scene of the PAVIN platform for the PRISSMA tests.

For each trial, the pedestrians follow the same path through the platform and repeat it twice, consecutively, to ensure repeatability. Following the different colored lines in the next Figure, the path allows the pedestrian to be presented from the front (paths 4 and 7), the back (path 1) and the side (path 2, path 3, path 5, and path 6), in relation to the camera position (the red star in the next Figure). In addition to walking at a moderate pace, the pedestrians also find themselves sitting on the bench at the picnic table.

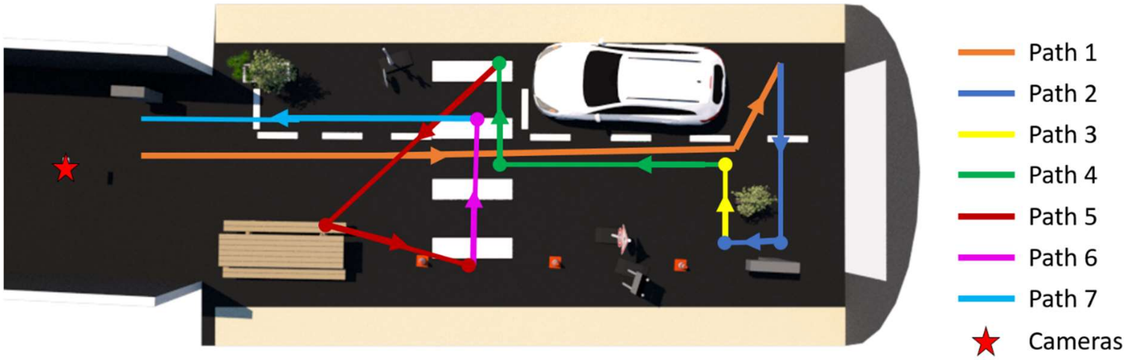


Figure 30: Path of the pedestrians during the tests following the colored lines and arrow directions.

2.3.4. PEDESTRIAN

The databases set up contain both real data and numerically simulated data. In this way, the pedestrians present in both types of database are described successively.
First, real pedestrians are described. To be representative of a wide variety of pedestrians, different characteristics have been made variable to form the batch of 100 different pedestrians (Figure 31) such as:

- Clothing: 50% of the clothing is representative of summer weather and 50% of winter weather.
- Accessories: a selection of pedestrians carry accessories with different sizes.
- Gender: 60% of the pedestrians are male and 40% are female.

To add a seasonality in the scene (summer/winter), the pedestrians have been dressed with clothes characteristic of high or low temperatures such as: hats, caps, shorts, pants, coats, etc and as much as possible, a variability of the color of the clothes has been respected (bright colors, dark or light colors). Wigs have also been used to increase the number of female pedestrians. Different sizes of accessories have been used in the tests. The objective is to have an impact on the overall silhouette of the pedestrian in an attempt to fool the detection algorithm. Considering the accessories worn by the pedestrian is crucial to guarantee his safety. An object worn by the pedestrian that would not be detected by the detection algorithm of an autonomous vehicle could endanger the pedestrian.
The data can be classified into four sub-lists:

- Small: for small accessories, such as a small backpack, a helmet, a plant, etc.
- Large: for large accessories, such as a large cardboard box, a snowboard, an open umbrella, etc.
- No accessories: when the pedestrian is not wearing any accessory or the accessory does not alter the pedestrian's overall silhouette (e.g., a headlamp, a yellow fluorescent vest, a cell phone).
- All: all pedestrians, regardless of the accessory sizes.

Table 9 shows the distribution of the number of pedestrians by the accessory size category and a thumbnail of the 100 pedestrians in the PAVIN database is shown in Figure 31.

Table 9: Number of pedestrians per sub-list of accessory sizes.

Accessory Size	Number of Pedestrians
Small	25
Large	33
No Accessories	42
All	100

Figure 31: Thumbnail of the 100 pedestrians of the PAVIN database.

On the simulation side (LEIA database), numerous pedestrians are also available. The 4DV library offer a wide choice of human in terms of gender, ethnicity, age and type of clothing. Figure 32 summarizes the humans used in the simulation. The appearance of the human can also be modified to change the hair or clothing color. In 4DV each human can be set with 3 appearances which means we have 36 different humans.

Figure 32: Thumbnail of the 36 pedestrians in LEIA database.

3. METHOD

3.1. A METRIC BASED ON A PEDESTRIAN DETECTION ALGORITHM

As explained above, our approach to comparing and qualifying physical and digital test equipment is based on analysis of the results obtained by a detection algorithm, rather than on analysis of the raw images themselves.

To do so, it is therefore necessary to have a pedestrian detection algorithm, a database labeled with a ground truth, and a detection algorithm evaluation metric. In this study, we have chosen to use the AUC score.

Concerning the detection algorithm, the third version of YOLO detection algorithm, which stands for "You Only Look Ones", was chosen in this analysis. It is indeed a very common algorithm in the literature on object detection. Moreover, it is very easy to handle. The library of objects available in this version contains 80 items.

The algorithm requires two main parameters: the confidence threshold (a value between 0 and 1) of the labeling and the object to label in the images. Only the class "person" is labeled in this study and the confidence threshold chosen is explained in the following section.

A frame can get multiple detections with different level of confidence even though only one pedestrian is walking in the scene into our database.

As reminder, the objective is not the evaluation of YOLO algorithm but to use a popular object detection algorithm to evaluate main characteristics of the database, and to compare digital and physical artificial fogs.

In object detection, a metric widely used to evaluate the validity of a detection is the intersection over union (i.e. IOU) between bounding boxes as shown in the next Figure.

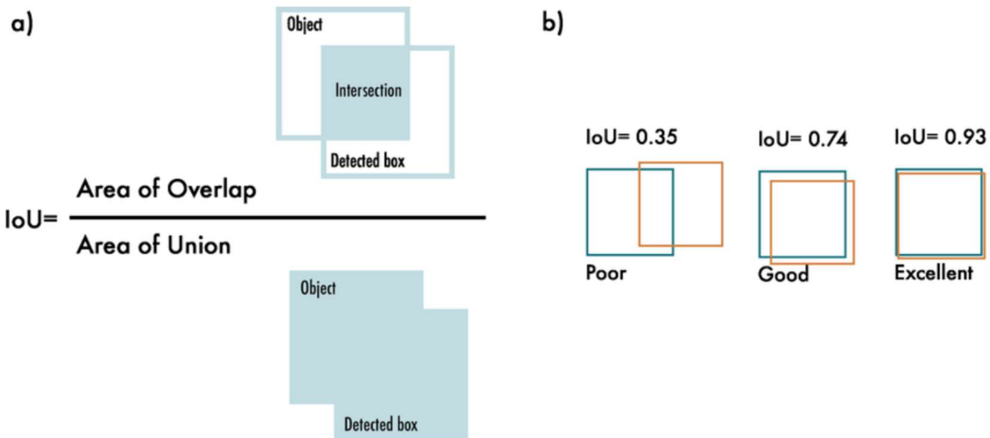


Figure 33: Intersection over Union (IoU). a) The IoU is calculated by dividing the intersection of the two boxes by the union of the boxes; b) examples of three different IoU values for different box locations \cite{terven2023}.

The intersection is calculated following the equation:

$$IOU(frame) = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

The precision-recall curve is then calculated based on the results of intersection over union values. The curve shows the trade-off between precision and recall for different confidence threshold values from the YOLO algorithm.

As an example, the different detections obtained by the YOLO algorithm, for different levels of confidence, from 0.3 to 1, on two images from the database are presented in the next Figure. The left image shows the 9 YOLO labels with two labels far from the pedestrian present in the scene, yet for one of them a confidence value greater than 0.5.

Then, the area under the curves (AUC) score is calculated. A large AUC value represents both high recall and high precision. A high precision value indicates a low false positive rate (good confidence value but no ground truth label), and a high recall value indicates a low false negative rate (low confidence value but ground truth has a label).

Figure 34: Example of YOLO detections on two clear weather images with different pedestrians. Colors: Green is for confidence > 0.9, yellow is for 0.9 > confidence > 0.7, orange is for 0.7 > confidence > 0.5, Red is for 0.5 > confidence > 0.3.

3.2. A METRIC BASED ON CONTRAST EVALUATION

A first "high-level" metric was proposed in the previous section. This verifies whether images from digital twins, with or without hardware in the loop, enable an AI-based algorithm to be evaluated in a way that is similar to reality. In addition to this, it is interesting to propose a "low-level" metric, at the opposite end of the processing chain. To do this, we are proposing a metric based on a contrast measurement directly on the raw image pixel data. This metric is directly inspired by the definition of the physical quantity describing fog density: the Meteorological Optical Range (MOR). In fact, fog density is characterized by a measure of contrast. To calculate the contrast ratio, we first need to define two zones in the image, one dark and one bright, which must be at an equal distance from the camera. It is important to take two zones at equal distances because the impact of fog is proportional to the distance between the target and the camera. Once the two zones have been defined, the pixel values in each zone are averaged to obtain the luminances L_{dark} and L_{bright}. The contrast is then calculated using the following expression:

$$Contrast = C = \frac{L_{bright} - L_{dark}}{L_{bright}}$$

Contrast therefore varies from 0 to 1. The denser the fog, the lower the contrast. This approach to calculating contrast was applied to 9 pairs of zones (dark and light), from all the images in the various databases. The following figure shows the zones used. As shown in the figure, the choice was made to take pairs of zones at different distances to see if the variation in contrast is indeed similar for different depths. This will be of particular interest in validating the fog addition approach of the LEIA database, which is based on the addition of fog layers at different depths.

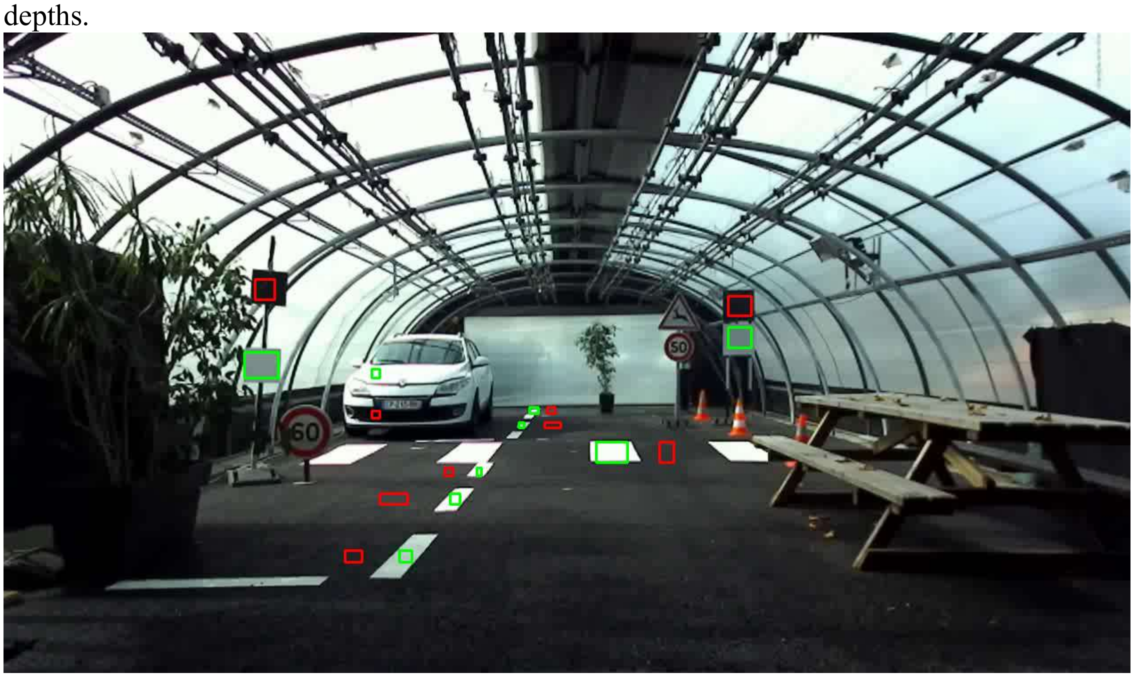


Figure 35: Presentation of the 9 measurement zones for the contrast method

The two metrics used for this study, and the database variants, were presented. The following section shows the results obtained.

4. RESULTS AND DISCUSSION

4.1. QUALITATIVE ANALYSIS OF DIFFERENT SIMULATION VARIANTS

Before starting to analyze the results, it is important to take a visual look at the images obtained for the different databases. The following figure therefore shows the images obtained for a similar pedestrian, in the same location, under different weather conditions and for different simulation modalities.

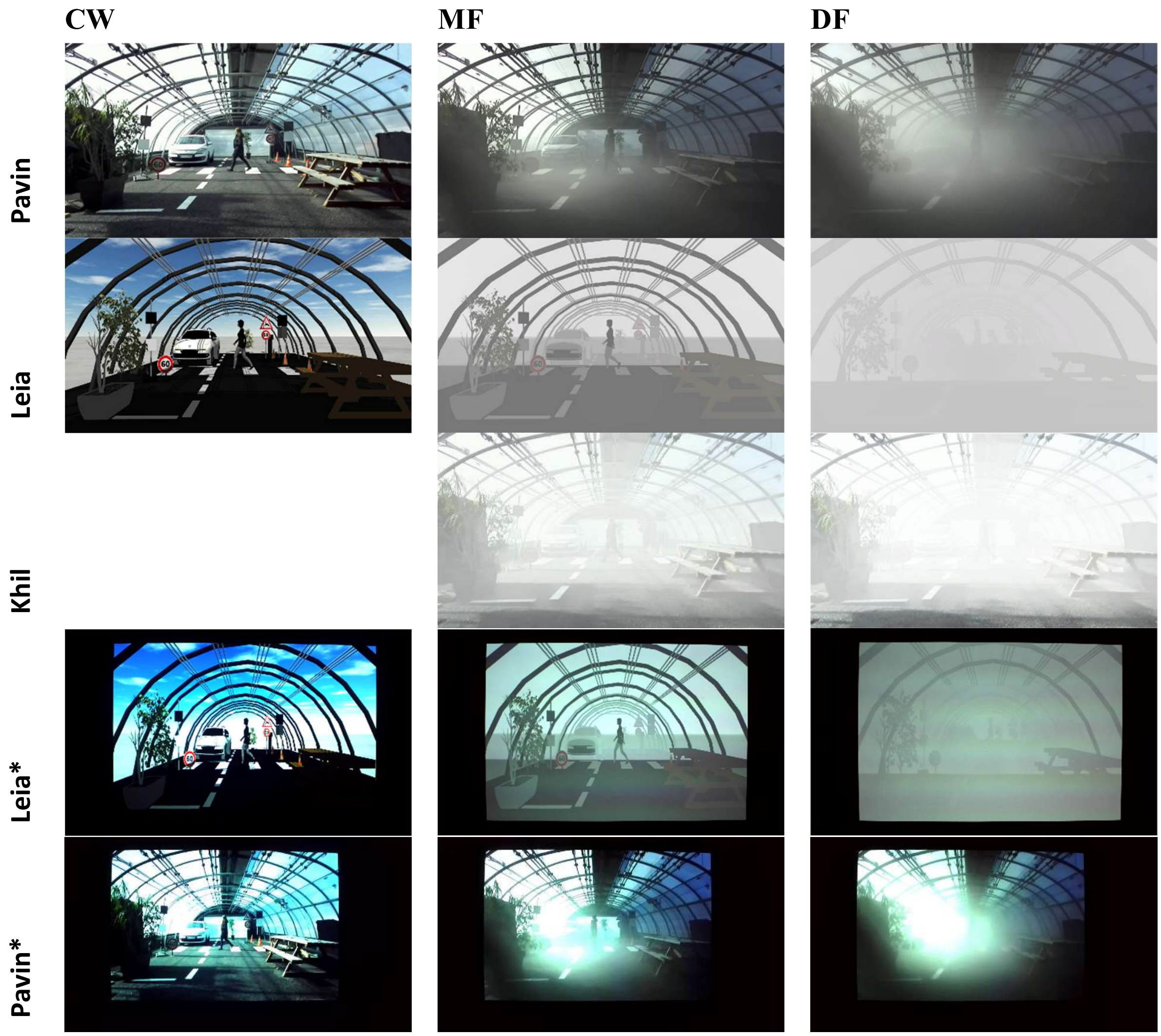


Figure 36: Example of an image for different weather conditions and simulation modes

In this figure, we can see that there are variations in rendering between the different simulation modalities. For example, clear-weather images from Leia are more contrasted than those filmed in real life on Pavin. The replay version * of the databases also significantly increases contrast. As far as fog is concerned, the models tend to make the fog visually too clear, compared with the real data from Pavin. For heavy fog, the Leia version also lacks contrast. In the case of Pavin*, we can also see that the camera is overexposed in places, while other areas are under-exposed. This means that the dynamic range of the video projection equipment used is not sufficient. Similarly, in the case of the two versions *, the framing is not ideal, resulting in black borders (and therefore under-resolution in the useful area). Further results will show whether these differences really matter: do artificial intelligence algorithms behave in the same way with the different modalities? What's more, thanks to the proposed contrast method, the qualitative findings we've just made can be confirmed quantitatively with a metric.

4.2. OVERALL RESULT PER DATABASE: HIGH LEVEL AND LOW LEVEL METHODS

Table 10: AUC scores obtained for the Yolo algorithm on the different variants of the database, for an IOU of 0.5.

	CW	MF	DF
Pavin	0.93	0.89	0.29
Leia	0.92	0.86	0.35
Khil		0.80	0.25
Pavin*	0.09	0.12	0.02
Leia*	0.84	0.84	0.42

First, we propose to analyze the overall results obtained by the Yolo algorithm on all the databases. The table above shows the AUC score obtained for each database variant, in the three weather conditions of clear weather, medium fog and dense fog. The AUC score varies from 0 to 1; the higher the AUC score, the better the algorithm. The first observation is that data from the Pavin* database completely prevent the algorithm from detecting pedestrians. The AUC scores for this variant are extremely low. A check of the detections shows that the algorithm has great difficulty in detecting pedestrians on Pavin* images, missing them on around three-quarters of the images. This is because the Pavin* images have very degraded dynamics. This is due to a change in equipment (different video protection for Pavin* and Leia*). Replaying an existing database is therefore not particularly appropriate. With regard to pure simulation tests, we can see that the Leia database achieves fairly similar results to Pavin, with a difference for dense fog, which appears less complicated for the algorithm in the Leia database (AUC = 0.35) than in the Pavin database (AUC = 0.29). The addition of the HiL aspect in the Leia* base, with the replay in front of the real camera, lowers the score for clear weather, while increasing it for heavy fog. In the case of clear weather, this can no doubt be explained by the fact that the image dynamics are very specific. The camera is filming a bright screen in a very dark room. It therefore has difficulty in making a good exposure setting. This could be compensated for by taking care to fix the exposure settings of the cameras when testing in Leia 2. The final variation is the Khil variant, in which fog is added to the initially fog-free real images. For this variant, the effect of the fog is stronger than for the Pavin base with real fog. This suggests that the background luminance and fog density settings should be reviewed for this model.
Beyond the results between databases, the table also shows that the choice of fog densities is not very well chosen for the evaluation of an algorithm. Medium fog seems too light (detection too easy), while dense fog is too dense (detection almost impossible). For evaluation purposes, therefore, much more variable fog ranges are required.
Following this analysis of the output of an artificial intelligence-based algorithm, it is interesting to see what the contrast levels of the different databases are. As explained in the method, the contrast ratio was measured at 9 points in the image. We propose here an average result for the 9 zones, to check whether some variants are indeed more contrasted than others.

Table 11: Contrast score obtained on different variants of the database.

	CW	MF	DF
Pavin	0.76	0.30	0.12
Leia	0.89	0.19	0.02
Khil		0.11	0.02
Pavin*	0.66	0.25	0.13
Leia*	0.77	0.29	0.04

From the point of view of contrast, the databases * are close to reality. Although they appear to be far apart visually, they are actually quite close numerically. On the other hand, we can see that the Leia and Khil simulations do not represent reality well in terms of contrast. It is interesting to note that the results obtained by the high-level and low-level metrics are not correlated at all. This shows that the contrast ratio of the images (although a very good representative of fog) is not sufficient to assess the quality of computer-generated images for the evaluation of artificial intelligence algorithms. Other factors such as saturation, colorimetric and resolution should perhaps be taken into account. In addition to the overall results for each simulation variant, it is interesting to go further to verify the impact of the pedestrian's distance, the impact of the pedestrian himself or even how uncertainty evolves with temporal sub-sampling. This is the subject of the following sections.

4.3. IMPACT OF DISTANCE ON RESULTS

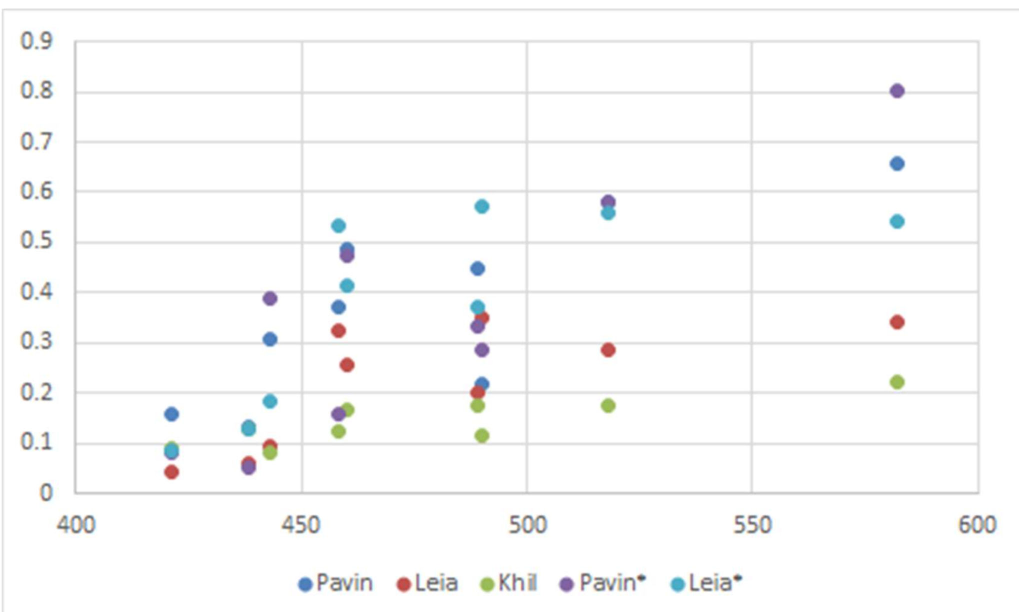


Figure 37: Contrast ratio as a function of measurement distance for Medium Fog (MF). The distance is given here as the vertical position of the measurement pixel.

The figure above shows the contrast ratio as a function of distance from the measurement target. For all the databases, the contrast ratio correlates well with the distance from the target. This result is normal according to the theory of the physical laws of fog. However, the slope of the curve is not the same for all the simulation variants. In particular, the Leia and Khil models are not sufficiently contrasted. This shows that the fog generated is too strong compared with real fog. It is also interesting to see that the replay on Leia 2 (* versions of the databases) allows us to obtain a rather realistic evolution of the contrast ratio as a function of distance. This contrasts with the results obtained with the high-level analysis for the PAVIN* database in particular. For the Pavin* database, the results of the high-level analysis (considering AI) are mediocre, whereas for the low-level analysis (on contrast) the results are quite good. Therefore, there is something in the replayed images that poses a problem for the AI. This shows that the simulator validation process cannot be limited to low-level metrics. They will also have to be qualified and verified using high-level metrics, involving AI systems. This shows the complementary nature of simulators, which can be used to run a large number of scenarios, and test chambers, which can be used to calibrate simulators on a sub-sample of scenarios.

4.4. IMPACT OF PEDESTRIAN VARIETY

The Pavin database was also created to check the impact of pedestrians on the analysis of AI algorithms. The Pavin database contains a variety of pedestrians, with and without objects in their hands. In addition, the number of pedestrians is large enough to allow statistical verification of the impact on pedestrian variability. This section therefore focuses on the analysis of the impact of pedestrians on the Pavin variant only.

Table 12: AUC score obtained for the Pavin database, for an IOU of 0.5

Accessories	CW	MF	DF
No accessories	0.93	0.89	0.29
Small	0.93	0.89	0.29
Large	0.83	0.75	0.24

Concerning pedestrian accessories, in both cases (CW or MF), the sequences with pedestrians carrying large accessories (orange dots) obtain lower scores, demonstrating a weakness of the YOLO algorithm for this type of configuration. If we consider the scores for an IOU equal to 0.5, the best scores of AUC are encountered for the pedestrians carrying small accessories or not carrying any, giving AUC of 0.89 and 0.93, respectively, in CW conditions and 0.89 and 0.89, respectively, for MF conditions. Pedestrians carrying large accessories are challenging to detect for YOLO with scores below 0.83. Therefore, accessories have a strong impact on results. Indeed, if we take the 'No Accessories' condition (i.e., pedestrian without accessory) as the reference, we degrade the score obtained by 11% in CW and 16% in MF for large accessories. The impact of accessories is, therefore, significant, but appears to be independent of weather conditions. These initial results on the type of pedestrian used for the tests can be supplemented by a second analysis of the uncertainty associated with pedestrians.

In this section, we aim to estimate the amount of data required for a database of pedestrians in the same urban environment to guarantee accurate and reproducible results during the evaluation of a tool, such as the YOLO pedestrian detection algorithm. This can help estimate the amount of data sufficient for a pedestrian database suitable for autonomous vehicle sensor evaluation. In this context, two factors influence the quantity of data: the number of pedestrians and the number of images (acquisition frequency) present in the database. These two elements will be evaluated in turn in this section. Since pedestrians have different clothing, genders, and shapes, a random selection of Np pedestrians of the 42 pedestrians from "No accessories" sub-list has been repeated 100 times with Np = [2, 5, 10, 15, 20, 25]. Next table shows the mean AUC and standard deviation values for the 100 random selections of Np pedestrians for CW and MF conditions. The first finding is that the average AUC score is not worse, which is reassuring. We find relative deviation values lower than 10% if at least 5 pedestrians are selected for CW and 10 for MF. If we consider obtaining a relative deviation below 5% for either CW or MF, 15 pedestrians are necessary. The relative deviation is lower for CW than for MF. Conversely, if we take the current case of a single pedestrian (the case of EuroNcap), the relative error in the result is greater than 16%.

Table 13: Effects over 100 iterations of a random selection of Np pedestrians on AUC values for clear weather and medium fog conditions. For no accessories sub list of pedestrians only.

	CW			MF		
Np	Mean AUC	Std Deviation	Relative Std Deviation (%)	Mean AUC	Std Deviation	Relative Std Deviation (%)
2	0.74	0.09	**11.6**	0.63	0.11	**16.9**
5	0.74	0.05	**6.7**	0.63	0.08	**12.3**
10	0.75	0.03	**4.6**	0.63	0.04	**6.5**
15	0.75	0.03	**3.5**	0.64	0.03	**5.4**
20	0.75	0.02	**3.0**	0.64	0.03	**4.3**
25	0.75	0.01	**2.1**	0.64	0.02	**3.6**

5. CONCLUSION

The aim of the Cerema-LNE POC was to verify the key points to be considered for the approval of vehicles using AI, in test conditions requiring the use of digital or physical simulation. This POC was based on tools such as the PAVIN Fog and rain platform and the Leia 1 and 2 simulators. The work focused on an example of a critical application for road safety: the detection of VRUs (pedestrians) in foggy conditions. This POC demonstrated several key elements:

- the boundary conditions for the AI algorithms are not easy to find. In the example of fog, the pedestrian detection algorithms quickly go from very good detection to no detection. Finding the boundary test cases (scenario, pedestrian route, pedestrian environment, and weather conditions) is therefore not easy. This is due to the structure of the algorithms based on AI, in which numerous threshold effects come into play. This raises even more questions about the test scenarios to be put in place, especially as these scenarios will have to evolve over time as the detection capabilities of the algorithms evolve.
- The different simulation methods (HiL or not, pure digital simulation, etc.) do not produce the same results on AI based algorithms. There is therefore a lot of work to be done on improving and validating the systems put in place (simulators, HiL injection...). While the geometry of the scene appears to be good, lighting conditions, surface properties and sensor characteristics need to be considered to obtain images that are closer to reality.
- Another major issue is the variability of pedestrians. At present, trials are carried out with a single standardised pedestrian. However, our results show that the error in evaluation exceeds 16% in this case. It is therefore important to review the test procedures to take account of this pedestrian variability. For example, by taking 10 pedestrians, this

uncertainty is reduced to less than 10% (less than 5% for 15 pedestrians). In addition, starting to consider more complex pedestrians (for example those carrying objects) could be an interesting perspective. Our results show that the score falls in the latter case.

- As mentioned above, our results show that the different simulation variants (pure numerical, HiL, etc.) do not lead to the same scores when evaluating AI algorithms. We have then compared the images using a low-level metric. It is surprising to see that the two types of comparison (high level = close to the AI algorithms and low level = close to the raw images) do not give the same results. Some simulation modalities obtain very different scores on the AI algorithms, whereas the low-level metric shows that the images are similar. As AI algorithms are black boxes, it is therefore difficult to define metrics that can tell whether images from two simulation modalities are similar or not. A great effort of research needs to be done in this area, in search of reliable combined metrics.

Chapter 4: Transpolis POC: Crossing a traffic light intersection

1 INTRODUCTION

Transpolis POC is focused on the ability to validate the functioning of a system able to determine the status of a traffic light when reaching an intersection. It follows PRISSMA deliverable 3.3 presenting the protocols.

One of PRISSMA difficulty is to dispose of a mature AI brick to run some type-approval tests. A COMMA 3X based on open pilot was used as an AI to validate. Its experimental mode has a traffic light reading option. V2X is also a way for the vehicle to be able to know the traffic light phase, as results, the protocol presented in PRISSMA L3.3 gave a list of several test scenarios for a complete system mixing camera vision and V2X to determine the traffic light status. Since no such system was available, only few scenarios were tested.

This chapter presents:

- The system under tests,
- The tests equipment,
- The tests run and their results.
- Further works using the data with another AI,
- A discussion.

2 THE SYSTEM UNDER TESTS

Due to a lack of mature and open system to be able to run validation or type-approval tests, Transpolis choose to present a POC with the following specifications:

- An autonomous shuttle has to cross an intersection equipped with traffic lights and a Road Side Unit (RSU),
- Its only possible maneuver is to go straight through a cross intersection,
- No other road users will be considered during the tests,
- The POC focusses on the ability of the shuttle system to be able to determine the status of the traffic lights,
- For functional safety reasons, a redundant system using vison based by camera (AI) and V2X was studied in this POC.

A mature system doing the fusion of RSU information and vision by camera was not available. Consequently, all the test scenarios defined in L3.3 procedure were not possible to be carried out but some safety and practical considerations are given in the discussion.

2.1 Embedded systems

The vision-based system chosen is an Openpilot AI installed on a hardware named Comma 3X. This system was installed on a Ford Focus (2018 model)

When plugged on the CAN of the vehicle, this system activates as an ACC and a lane centering, giving a level 2 of automation ADAS.

Comma 3x technical specifications are the following:

- CAMERAS
  - Three 1080p cameras w/ 140 dB of dynamic range: dual-cam 360° vision and a narrow cam to see far-away objects. The third camera is turn inward of the vehicle.
- PROCESSOR
  - Qualcomm Snapdragon 845
- CAN FD ENABLED
  - Supports CAN FD vehicles without extra hardware
- STORAGE
  - 128GB built in storage
- CONNECTIVITY
  - LTE
  - Wi-Fi
  - High-Precision GPS
- NIGHT-VISION
  - IR LEDs for interior night-vision monitoring
- DISPLAY
  - 2160x1080 OLED display
- PORTS
  - OBD-C port (USB-C w/ CAN)
  - USB 3.1 Gen 2 port

(a)

(b)

Figure 38 : (a) COMMA 3X display – Installation inside the Ford Focus

OpenPilot experimental mode was set on the COMMA 3X and the traffic lights detection option was activated.

A Lacroix city On-Board Unit (Figure 39) is also installed in the vehicle. Its antennas are magnetically fixed on the vehicle roof and it is powered on the 12V socket of the vehicle.

It communicates in Wi-Fi with a laptop.

Figure 39 : Lacroix OBU

2.2 Tracks and road Equipment

Tests were carried out in TRANSPOLIS urban area in les FROMENTAUX site. The ground network and track configuration are shown on Figure 40. The path of the POC is presented in red.

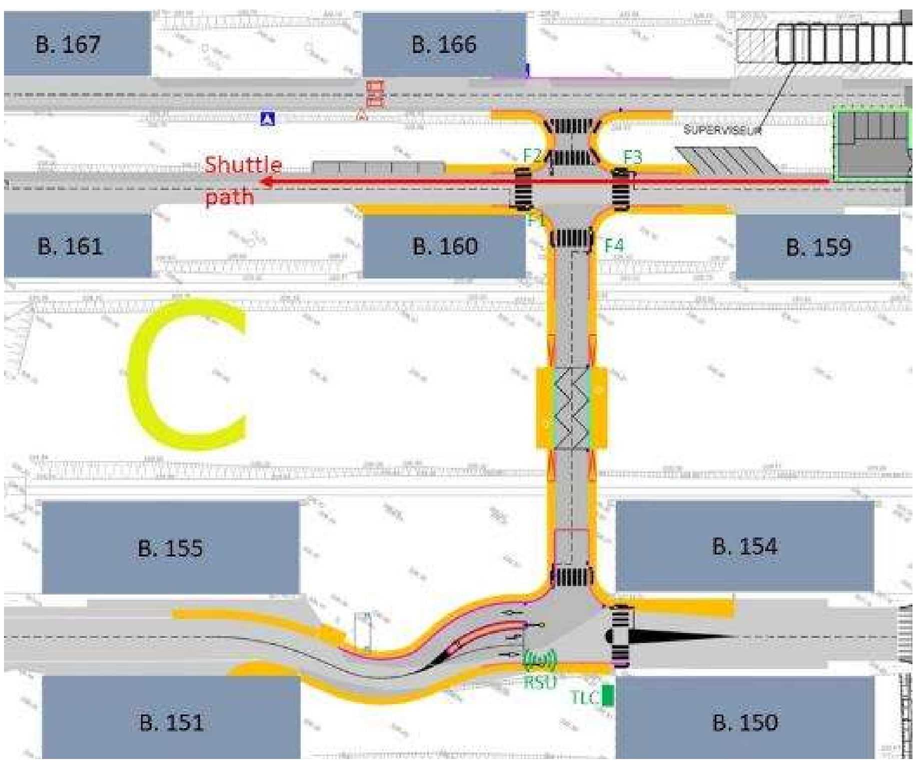


Figure 40 : Intersection configuration in TRANSPOLIS city area C. Scale: Building 150 is 50m long. RSU: roadside unit. TLC: Traffic light controller

Figure 41 : Picture of the POC intersection. F3 in the front plan

The intersection is equipped with a full and classical French traffic light system:

- A traffic light Controller (TLC) Lacroix traffic TRAFFY
    - Aluminium street cabinet 800, 1250 / 800 / 420, RAL 1015
    - General electrical protection 32A 300 mA
    - CPU GPS
    - Traffic lights cards
    - Command agent for 2 positions, Lacroix city
- Four Alumix traffic lights equipped with Equinoxe LEDs
- Four Aluminium posts
- Four R12 pedestrian signals.

All data sheets of the equipment were given in PRISSMA l3.3 annex.

The controller is programmed using a Wintraffy software. This software is installed on a Windows computer connected to the controller and allows to define all the intersection configuration and to set traffic lights cycles and programs shown in

Figure 42 : Traffic light program used during the trials

An ITS-G5 Road Side Unit of Lacroix city manufacturer is connected with ethernet to the TLC.

The RSU version is the 4.6.1 – It is not connected to a PKI server.
The RSU emits SPATEM and MAPEM messages at a frequency of 1Hz.

2.3 Tests equipment

An IMU (inertial measurement unit) with dual antenna and RTK correction is installed in the vehicle to measure its position with a cm accuracy (see Table 14).

To be able to analyse the synchronisation of the information from several systems, a device filming with up to 4 cameras was used (see Figure 43). This system also records the position of the vehicle and the universal time using the GPS.

Table 14 : OxTS RT3003 XG Inertial measurement Unit and GPS with dual antenna	
	<u>Performance (accuracy):</u> Position: Glonass L1,L2 0.01m Speed : 0.05km/h RMS Roll/Pitch: 0.03° Cap : 0.1° Angular speed : 0.01°/s Acceleration: 0.01%/ Taux 100m/s² Frequency : 100Hz

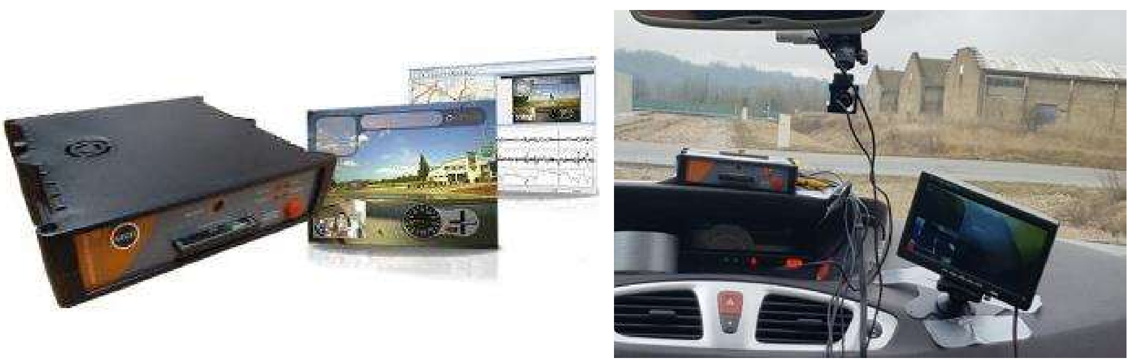


Figure 43 : PDRIVE system (tests&training no longer produced)

3 TEST PROCEDURES

The COMMA 3X was used as an adaptive cruise control, controlling the speed of the vehicle on the path. The IA included in the COMMA 3X was running as a black box. No metrics could be extracted in logs to be able to understand the detection of the traffic lights. However, it was possible to record the videos of the tests directly on the COMMA 3X in logs.
The ability of the COMMA 3X to detect the traffic light status and to control the speed of the vehicle was tested randomly with respect to the traffic light timing.

The OBU allows evaluating the coverage of a RSU making possible to analyse the coverage on the specific path of the POC.
This coverage was analysed by two means:

1- Running a RSSI coverage measure on the path. The RSSI is the Received Signal Strength Indicator. That path was run twice at a low speed (10km/h) and the RSSI of the received messages were recorded by OBU as well at the coordinates of the vehicle when receiving the messages.
2- The messages were logged in two static positions on the path heading towards F3. Position 1 is in non-line of sight of the RSU and position 2 is in line of sight of the RSU.

A full system making the fusion of RSU information and vision by camera was not available. Therefore, all test scenarios defined in L3.3 procedure were not possible to be carried out.
To evaluate the validity and the synchronisation of the two information sources, a videorecorder was used to film OBU timing HMI on a laptop, the status of the traffic light through de windscreen and the screen of the COMMA 3X during the random tests.

4 RESULTS

COMMA 3X traffic light detection

Over the twenty trials recorded, the COMMA 3X showed a consistent and good management to the intersection crossing only once. During this trial, the vehicle arrived at the red lights, stop, and stated again when the light turned green.
This behavior could not be reproduced later making reproducibility or robustness tests etc. useless. Consequently, all the test scenarios imagined in the L3.3 such as Scenario V-camera-S.x, Scenario V-camera-D-Rob.x, etc. could not be evaluated.
For all other trials, the system considered the intersection but not the traffic light. It stopped at the transversal road marking and crossed the intersection, whenever the light was green or red.

V2X coverage

Figure 44 presents the results of the coverage trial carried out. The path was run twice at low speed. It shows the masking of the buildings going from a poor reception (low RSSI) up to no message reception at all (behind the building 161 on the upper left of the Figure).

Figure 44 : Coverage on the path – Red for poor RSSI (-91 dBm) up to green for strong RSSI (-75dBm)

Two recordings of the messages were carried out in 2 static positions showed on Figure 45. Position 1, in non-line of sight of the RSU, over 1 minute, only 31 SPATEM were received. Position 2, in line of sight of the RSU, over 1 minute, 60 SPATEM were received.

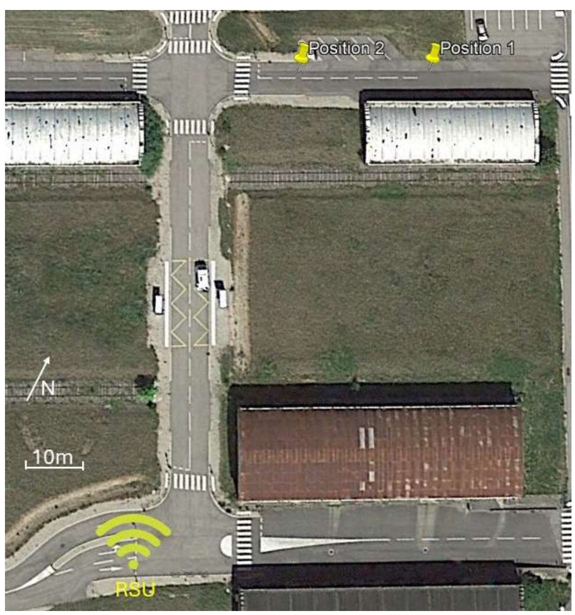


Figure 45 : Positions on the path where static recordings were done

Position	Duration	SPATEM received	Rate
1	1 min	31	51.7%
2	1 min	60	100%

Table 1: message reception evaluation on the path

Information synchronization validation

In the protocol, Scenario V-F-D was proposed to assess that OBU and camera information are consistent while running on the pathway. As the camera vision is not working, the tests equipment was set in order to evaluate its performance. A system recording a video from 3 cameras was used to analyze the synchronization of the information gather from different means has shown on Figure 46.

Such device is useful to record the synchronization of the HMI information with some event in a scenario. However, in this case, the display on the HMI may not be instantaneous introducing a bias. Moreover, the OBU interface displayed some irrelevant information such as negative timing or unknown timings, making difficult to interpret the results.

This means that to analyze the correct synchronization of various sources, logs with a common time base shall be recorded. Usually, the GPS time is used. In the case of the OBU information, it is possible to record the messages using Wireshark. Then, it is possible to extract the time from the messages using field such as the minEndTime or likelyTime expressed in UTC time stamps.

Figure 46: Example of video recording the Comma 3X, the OBU and the windscreen using the PDRIVE

The use of the videos is helpful when logs are not available / possible. For example, the visual status of the traffic light can be recorded using a system such as the PDRIVE which records videos and GPS time.

These tests then require some post processing treatments. They can be time consuming.

5 POC KNOWLEDGE ENHANCEMENT

The COMMA 3X logs gave videos of the various trials. The open pilot AI did not work, the videos extracted from the COMMA system were used to evaluate another AI.

Traffic Lights Detection and Color Recognition using YOLOv8 by F. Alam [1] was used.

The learning phase of the model was processed with a database of 2097 labeled pictures of traffic lights. The confusion matrix is shown on Figure 47.
After the training, the model was able to recognize a traffic light, to detect its color and to give an estimation of its confidence. The output are the videos with bounding boxes around the detected traffic light and the label of its color and the level of the confidence.

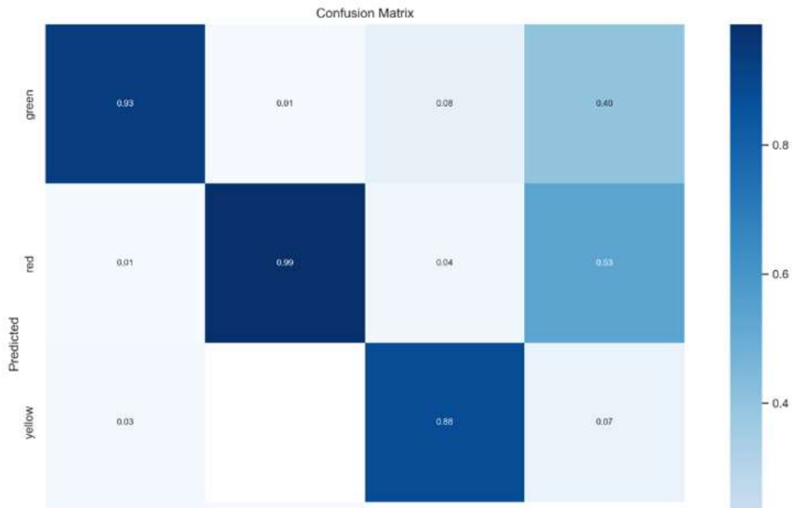


Figure 47: Confusion matrix given by the training.

Performances on a fixed video

A still video of 1 min/1200 frames on position 2 (Figure 45) was analyzed with a custom Matlab program. Since the image is fixed, it is possible to detect the status of the light by isolating the appropriate pixels and without programming a complex AI.

This video was captured in April at 05:06 pm in a street of the city track of Transpolis heading west.

The results of these analyses are presented in Table 15.

Table 15: Analysis of the performance of the yolo v8 model on a 1200 frame fix video o a traffic light cycle

Color of the traffic light	NF	NF of correct color detected	% of correct detection	NF of no detection	NF of Errors of detection
Red	760	347	45.66%	413	0
Yellow	120	0	0%	120	0
Green	320	320	100%	0	0

NF: number of frames/ images

Figure 48: Illustration of one frame of the video during the green phase.

The model was not able to detect the yellow phase on the video, it detected the red phase only on 45,66% of the cases but worked well on the green phase.

These short analyses are representative of what shall be done during a development / validation process of the model. The bad performances of the detection can be due to the poor sample of picture used for the training of the model.

Repeatability of the results

A 144 images video was treated 5 times by the yolo v8 model. The video was captured on a random trail with the car running toward the traffic light. It was red at the beginning of the video and it turned green.

The results showed a perfect repeatability of the results with the same detections and confidence indicator level (see Figure 49)

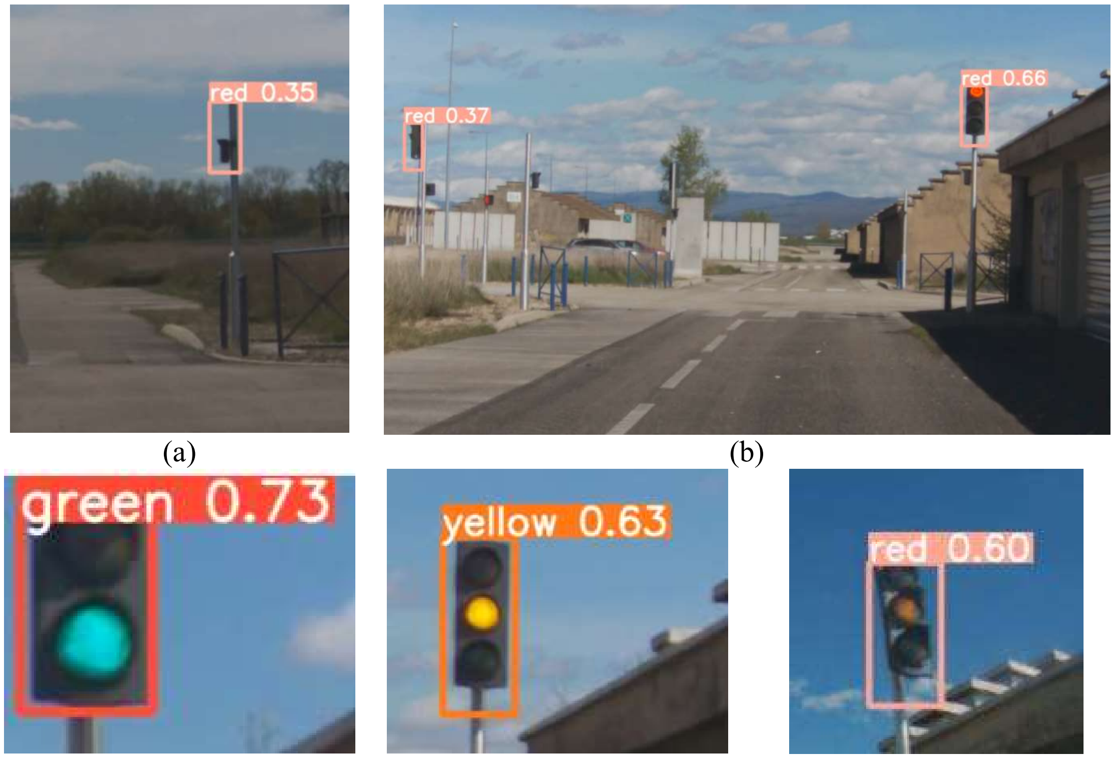
Figure 49 : illustration of a repeatability analysis.

Performances on a set of videos

A sample of 10 videos of the trials done on two traffic lights on Transpolis tracks was visual analyzed, giving the following results:

- The model is not able to detect the traffic light when the vehicle is at a distance of more than 25m. This shall be due to the poor sample of pictures used for the training.
- False detections were observed on pedestrian lights, this false positive is reproducible and was systematically observed (see Figure 49 a)
- An antagonist traffic light was falsely detected once (see Figure 49 b)
- The model was able to detect a green phase on a partial traffic light image (see Figure 49 c)

- The model was able to detect the yellow phase unlike the results obtained from the still video analyses (see Figure 49 d).
- An error of detections of the color was found for the yellow on the wide-angle camera (see Figure 49 e).



(a) (b)

(c) (d) (e)

Figure 50 : illustrations of the visual analyses of the performances

This kind of process is very time consuming and only 10 videos were treated since the development of the model is not the objective of our POC. Performance indicators such as the number of errors of color detection, the number of false detection or the number of non-detection can be estimated for a significant number of intersections and traffic light configurations and phase. The results of the validation tests and the validation method shall be presented in the manufacturer documentation for a type-approval.

6 DISCUSSION

Transpolis POC was designed to define a protocol to validate a system able to determine the traffic light status for an automated shuttle to cross an intersection.

One of PRISSMA difficulty is the lack of mature system available and “open” to run some instructive and significant tests on an AI brick.

The Openpilot AI installed on the COMMA 3X seemed promising to be able to run some test on Traffic lights. This system is designed to give a Level 2 of automation in a common car. It means that it controls the speed and the direction of the vehicle. The experimental mode used during this campaign was supposed to be able to handle the traffic light status. It worked only once with the car stopping at the red light and starting again when it turned green.

For all other trials, the system considered the intersection and not the traffic light. Openpilot is an American system, designed in California. The traffic light intersections in the USA are different from those in Europe, the traffic lights are located on the side of the intersection. In order to better understand the behavior of the system, we added a road marking 10m before the traffic light to try to reproduce a US intersection but the system behavior did not change. We also tried the system on another intersection, at the crossing of the two main avenues of Transpolis city track, the traffic light posts are mounted on concrete blocks to be moved. One of the traffic lights was moved but the road marking and the light position were not consistent with a US intersection and the system did not work better.

Type-approval tests are run considering the system as a black box that should work in all conditions. Consequently, the simple tests run on the tracks were sufficient to show that the system is not working.

However, the type-approval work for such system will be mainly based on an audit of the manufacturer documentation.

This documentation shall present the system architecture and the validations done. The short work presented here of the yolo v8 model is only a small illustration of the king of work that shall be part of the development and validation activities the manufacturer shall run and present in its documentation. The work done here was not deepened since it is very time consuming to analyze videos by visual inspections. Another AI could be trained to detect the false positive or misdetections but validation an Ai using another Ai is not an option.

Robustness tests and weather condition effects were not investigated here. However, these kinds of tests can be run on open roads. Moreover, vision-based detections can be developed and tested using large sets of videos or even synthetic images including these aspects.

In the validation process, tracks are used for the first trials of the system before testing it extensively on open roads. Tracks are also used for type approval tests. However, the question of using the technical service tracks for the development/validation phase and the type-approval tests may create a bias since the tracks data could be used for the AI training making a problem of over learning of the type-approval conditions.

The potential system at first imagined by Transpolis was based on vision by camera and a V2X module for functional safety.

V2X systems such as OBU, RSU and their connection with the traffic light controller are not yet certified according to international standards. The certification tests are not in the scope of PRISSMA since there is no AI in these systems.

For type approval of the full AI + V2X system, the manufacturer shall present clearly in the documentation the strategies adopted to handle the intersections. For example, the SPATEM allow to know the duration of traffic light phase, it means that a strategy can be implemented to save energy or to maximize passenger comfort. The vision-based brick can be used as a control of the SPATEM information, or also as a back up in case of message reception problem.

Traffic light phases do not always have a fix duration. A pedestrian can ask for the priority pushing a button, or the phase duration can be defined by a traffic management system according to the city congestion. These variations shall be taken into account in the SPATEM and the vehicle strategy. This kind of configuration was not tested but a protocol could be set for tests on Transpolis city tracks.

All information about the choices and developed strategies shall be presented in the documentation. The technical service would then define a set of tests to be run on tracks or on open roads for the type-approval. Running these tests on tracks have controllability and reproducibility advantages. For instance, it is possible on tracks to test the behavior of the vehicle when the intersection controller is out of order (blinking yellow).

For automated shuttle deployment in France, several stages are defined in the decree 2021 – 873 [2] before its operation approval. The shuttle is a part of the transport system that also includes the pathway and the infrastructure equipment. The RSU coverage of the pathway shall be evaluated and validated. Results shall be produced in the safety documentation to be reviewed by a qualified body.

The V2X system also imply cybersecurity issues. This topic is important for type-approval and safety of the deployments. In PRISSMA, WP5 is dedicated to cybersecurity.

7 REFERENCES

[1] F. Alam, «https://github.com/farukalamai/traffic-lights-detection-and-color-recognition-using-yolov8?tab=readme-ov-file,» 2023. [En ligne].

[2] Ministère de la transition écologique, *Décret n° 2021-873 du 29 juin 2021 portant application de l'ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation,* https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043729532, 2021.

Chapter 5: VALEO/IGN POC: protocol for evaluating/characterizing the performance of a localization system for automated vehicles (based on artificial intelligence), and associated ground truth system (in a controlled environment).

1 INTRODUCTION

In deliverable L3.2, a protocol for evaluating the performance of a localization system for automated vehicles in a controlled environment (test track) was described.

It was then recommended that the ground truth system (used for this task and equipping the test track) should return the absolute 3D position (in a global reference frame, e.g. RGF 93) and heading of the vehicle under test, equipped with the localization system under test.

Time stamps (based on GNSS time synchronization) associated with these absolute positions and heading were to be provided.

In deliverable L3.3:

- A description of the ground truth system in a controlled environment has been provided (sensors and other equipment used by the ground truth system, concept of operation based on photogrammetry, calibration phase, targets to be installed on the test vehicle, etc.).

- A possible post-processing method was then proposed to evaluate the performance of a trial localization system by comparing the data returned by the ground truth system (i.e. 3D absolute position and heading with time stamps) and similar data calculated by the trial localization system.

A proof of concept was then carried out in 2024 to validate/assess the suitability of the proposed protocol, ground truth system and post-processing method, with the aim of characterizing/validating a localization system for automated vehicles.

2 PROOF OF CONCEPT: DESCRIPTION AND ANALYSIS

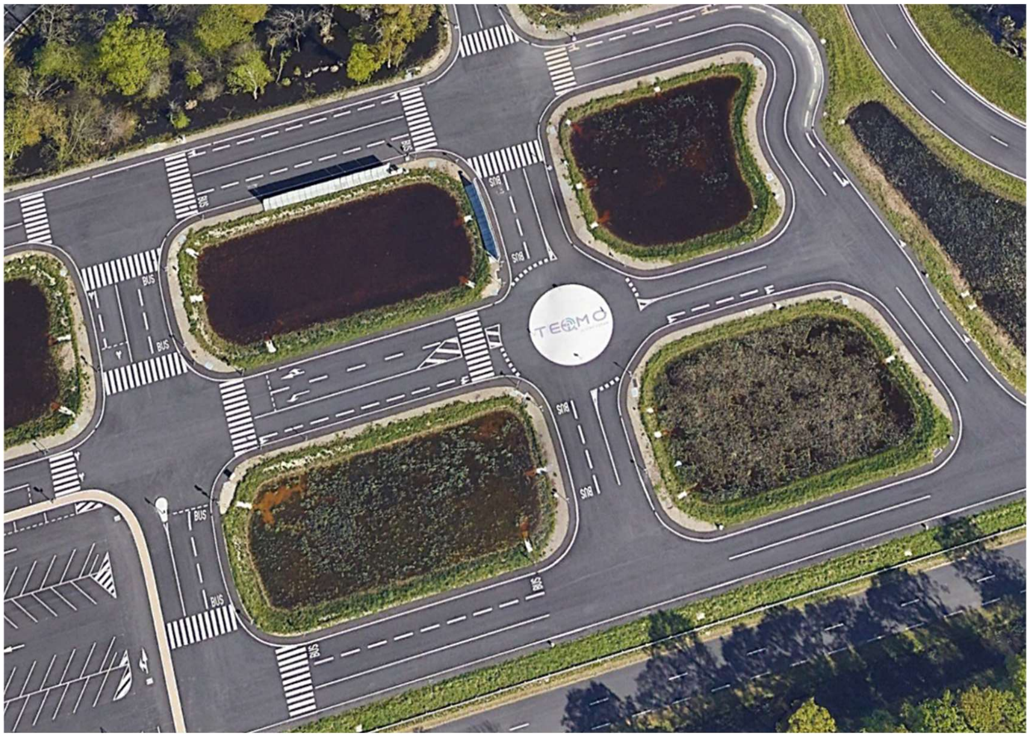
2.1 Test location, date and schedule

The POC was carried out at UTAC Linas Monthléry, in the TEQMO zone, on January 30, 2024, according to the following schedule:

Etape	Début	Fin	IGN	Valéo
Installation	8H30	10H00	Installation matériel	
Équipement du véhicule	10H00	10H30	Pose des cibles et des GNSS sur le véhicule	Mise-à-disposition du véhicule Véhicule stationné immobile au centre
Topométrie + pause déjeuner	10H30	12H30	Mesures topométriques des cibles terrain et des cibles véhicule	Mise-à-disposition du véhicule Véhicule stationné immobile au centre
Carto HD véhicule	12H30	13H00	Fin stations scanner laser + fin paramétrage caméras Chaussée accessible	Carte HD Valéo (avec matériel IGN installé)
Opérations	13H00	15H00	Prises de vues Chaussée accessible	Phase de roulage véhicule

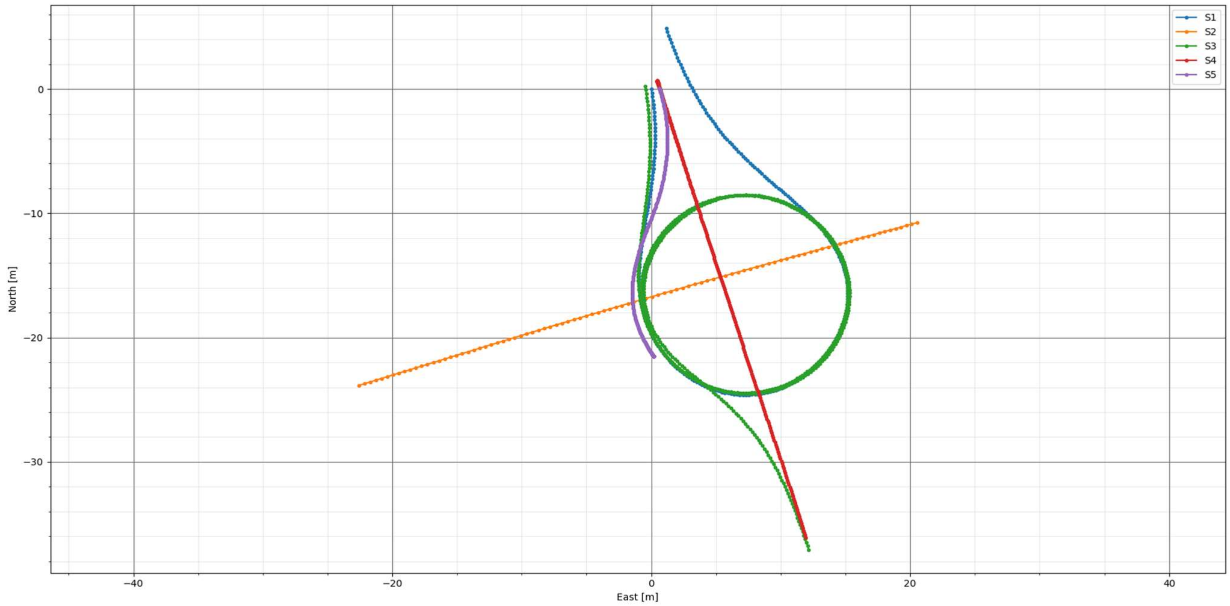
2.2 Scenarios

The following validation scenarios took place at the central traffic circle of the TEQMO UTAC test track.



The test area is the UTAC traffic circle

Scenario	Time (s)	Speed (km/h)	Description
S1	13.8	21	Enter the traffic circle from the north, drive around it and exit via the north exit
S2	5.0	35	Arriving from the east, the vehicle accelerates on the traffic circle in a straight line towards the west
S3	29.0	17	Entering the traffic circle from the north, the vehicle makes two complete turns and exits via the south exit
S4	14.2	14	In reverse, the vehicle enters the southern entrance to the traffic circle, crosses it in a straight line and exits via the northern exit
S5	12.6	10 -> 0	The vehicle enters the traffic circle via the north entrance and stops



Test scenarios

3 VEHICLE UNDER TEST

3.1 Ground truth system (internal to vehicle)

The vehicle used for the tests contains a ground truth system for localization based on a high-performance inertial unit and an RTK GNSS receiver.

During all the tests carried out, the ground truth system estimated an average uncertainty of 0.026 m in the vehicle's position and 0.01 degrees in its estimated heading.

The ground truth system calculates the position of the center of the vehicle's rear axle in the WGS84 coordinate system.

This type of ground truth system is commonly used for automated vehicles to characterize their localization system in open road conditions.

3.2 Localization system under test

On the test vehicle, the embedded localization system is based on Lidar point cloud localization.

During the installation phase, a point cloud map is created and optimized to better represent the vehicle's environment.

In the test phase, the vehicle's Lidar sensors are used as input to a map matching algorithm to provide a global position of the vehicle in real time.

We will evaluate the possibility of using the photogrammetric position estimation system as ground truth in a controlled environment (on a runway) when validating the performance of the Lidar point cloud positioning algorithm.

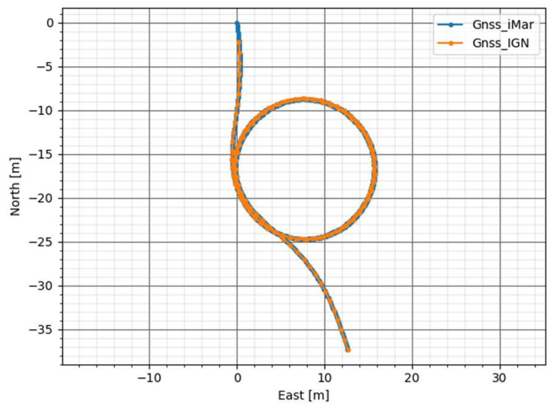
3.3 Results

a) <u>Validation of the reference system using photogrammetry</u>

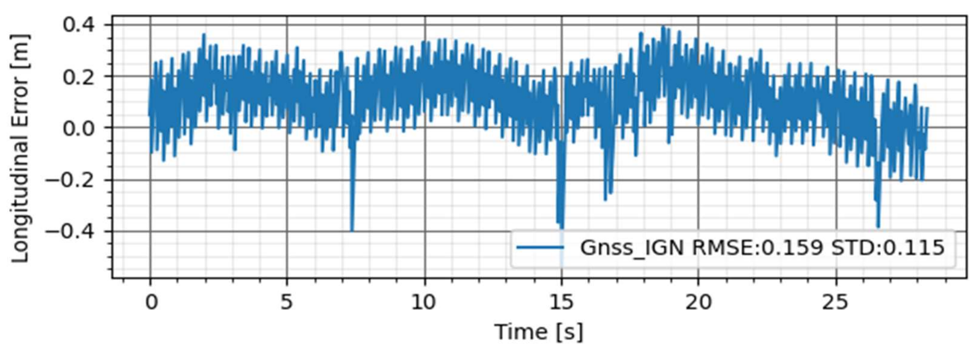
Before checking the results of the Lidar point cloud localization algorithm, we wanted to evaluate/comparison the two ground truth systems available to us during this POC:

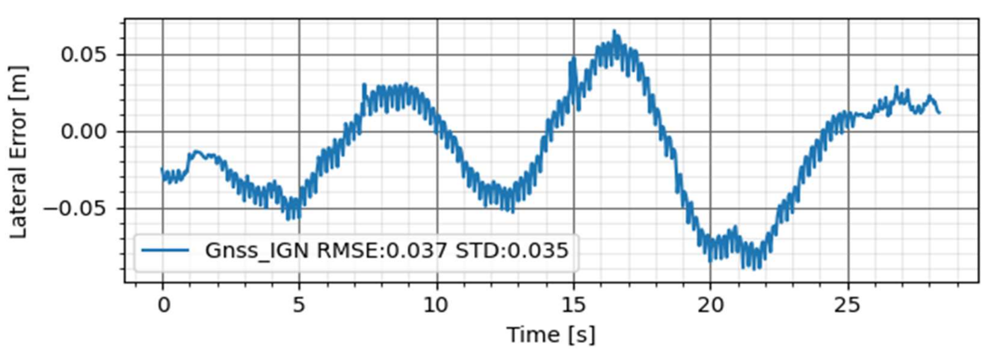
- The photogrammetric positioning system developed by IGN
- The positioning system based on a high-performance inertial reference frame coupled with RTK GNSS.

On the example of scenario 3 - two complete turns of the traffic circle (which presents a complicated case in terms of evaluation), we found differences/errors in lateral, longitudinal and heading of 3.7 cm, 16 cm and 1.74 degrees on average respectively.

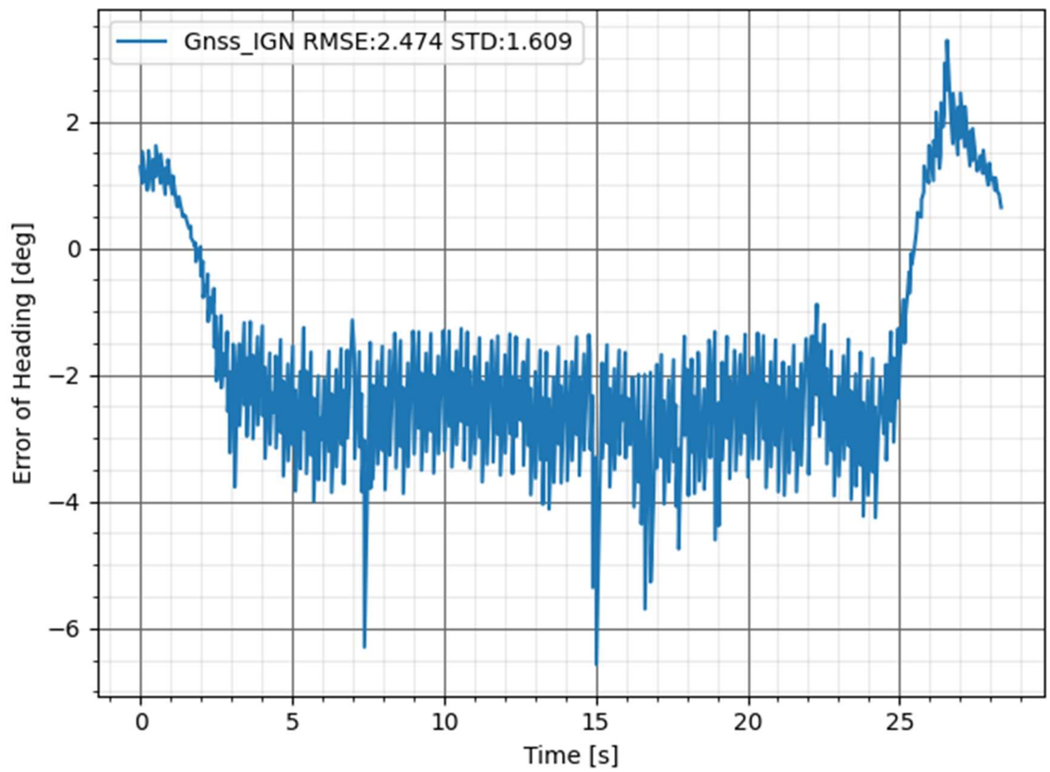


Scenario 3 test trajectory





Lateral and longitudinal error for scenario 3 test



Course error for scenario 3 test

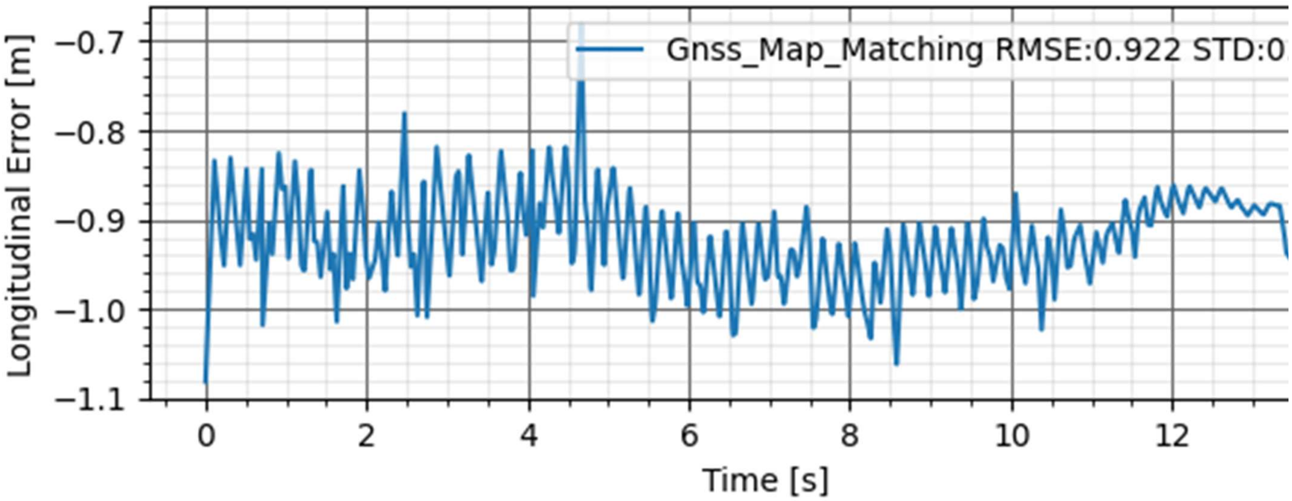
Rotational movements make estimating the difference (between the position value estimated by the system under test, and the reference value returned by the ground truth) complicated by the low frequency of the photogrammetry positioning system compared with the inertial unit data.

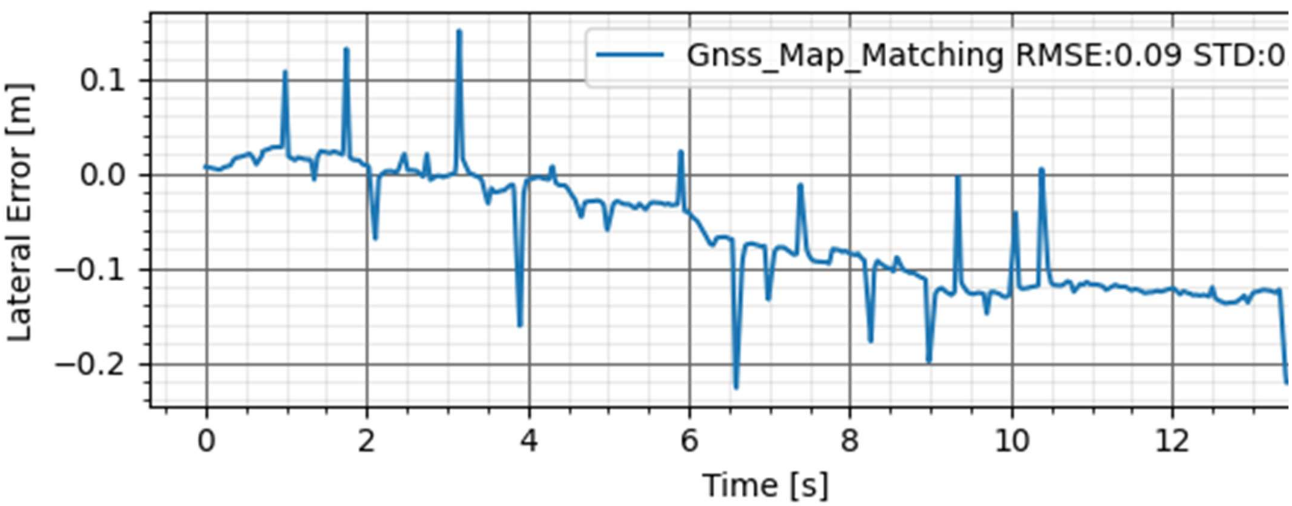
In scenario S2, where the vehicle travels in a straight line, we found differences of 2.7 cm, 30 cm and 0.31 degrees in lateral, longitudinal and heading, respectively.

b) Lidar Validation of a Lidar point cloud location system

To validate the accuracy of the Lidar point cloud-positioning algorithm, we calculated the lateral and longitudinal errors as well as the heading error for the different test scenarios.

Overall, we observed the at-tended performance of our positioning algorithm under the POC test conditions. Example with scenario S4:





Positioning accuracy of Lidar point cloud positioning algorithm using photogrammetric positioning system as ground truth

3.4 Analysis

a) <u>Performance</u>

The photogrammetric ground truth system (aimed at controlled environments) provides reference data (position and heading) consistent with the ground truth system commonly used in open road environments (RTK+IMU).

This photogrammetry-based system aims for greater accuracy than the RTK+IMU-based solution. Further tests are required to confirm this hypothesis.

b) <u>Synchronization</u>

The synchronization of positioning information is an important element in performance evaluation. Both systems under test had time synchronization based on GNSS signals. When evaluating localization performance, a script creates linear series based on a common time-tamp repository between the system under test and the ground truth. Linear interpolation is applied between two timestamps (if data is missing).

Although the synchronization of each of the GNSS systems can guarantee the same time base, it is preferable to have a common synchronization method (a synchronization signal sent by the ground truth system at the start of the tests, for example).

c) Data frequency

When data frequency on the ground-truth system is low compared with the sub-test system, we can observe an increase in lateral error when tests include bends (traffic circle turns) and longitudinal error when we have straight lines.

The lower the data frequency, the more difficult it is to assess performance accurately.

An increase in the frequency of measurements or estimates by the ground truth system is essential to provide a better granularity of positions and thus enable a better assessment of the localization algorithm's performance.

Increasing frequency can be achieved by a simple EKF on positioning results, for example.

d) Reliability of ground truth data

The photogrammetric positioning system provides no information on its status. It may be useful to have feedback on the status of the system, independently of the accuracy measurement provided.

For example, if the accuracy of the location measurement depends on the orientation of the cameras relative to the vehicle, it would be interesting to have another input indicating the status of the cameras themselves (which cameras are used, frame rate, internal synchronization status, illumination level, rain detection, etc.). By using these elements during the analysis phase, it is possible to automatically target the tests that offer the best performance for the ground-truth system. Another approach would be for the photogrammetric ground-truth system to filter out or a posteriori correct unreliable data (e.g. caused by a disturbance on the system's cameras), so that the reference data for characterizing/validating the localization system under test can be used with complete confidence.

e) Estimation stability

Unfortunately, we did not have the opportunity to test the stability of the ground truth system's position estimation. A stability test consists in continuously providing estimates of the vehicle's position at a standstill. The point cloud of positions obtained characterizes the stability of the position measurement by the ground truth system.

f) Operating limits

The aim of our tests was to evaluate the capabilities of the photogrammetric positioning system and the possibility of using it as a ground-truth system to validate a localization algorithm for autonomous vehicles.

We were unable to test the system's limits under unfavorable conditions. It is important to know whether the system detects a loss of accuracy and indicates this to the tester, so that it can be taken into account when evaluating a localization algorithm.

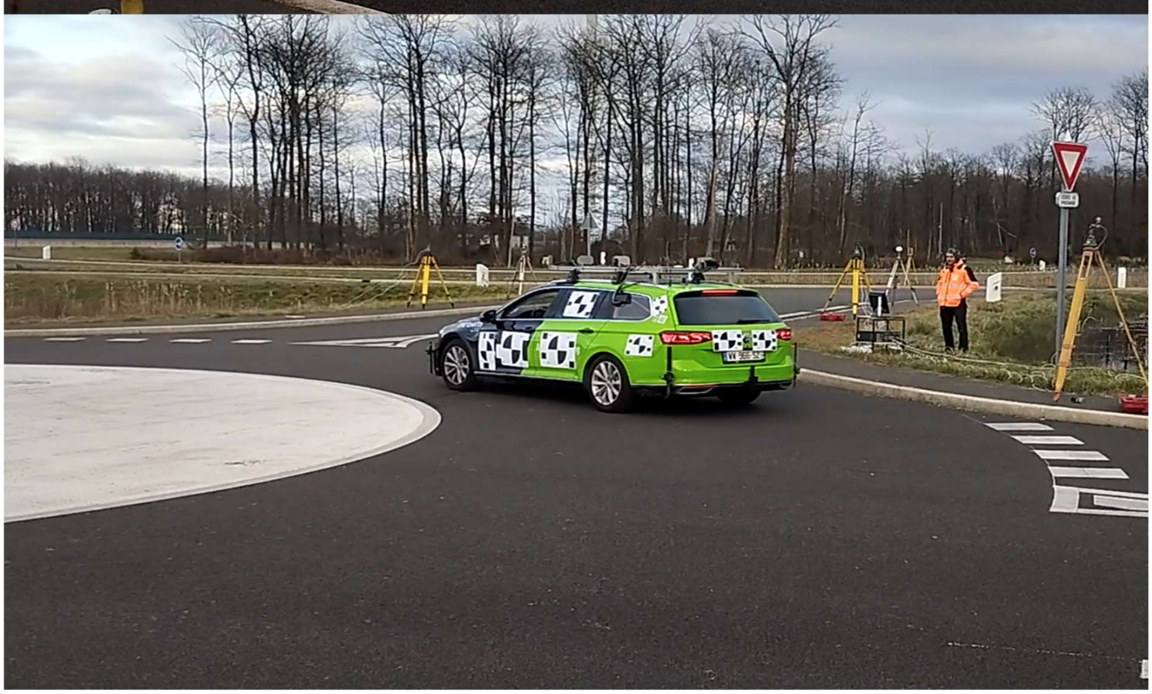
g) Scope of operation

The current configuration of the photogrammetry-based localization system limits the test area. The maximum speed we were able to reach was 36 km/h, as the area was limited. It would be desirable to extend the system (by adding cameras) in order to validate its performance at higher linear and angular speeds. For example, over the entire test area.

Another line of investigation would be for the photogrammetric ground-truth system to be able to operate in wider conditions (lower luminosity, e.g. twilight, night or adverse weather conditions e.g. rain, fog, etc.), as the localization systems under test are bound to operate in these degraded conditions.

4 ILLUSTRATIONS

The following photos illustrate the test scenarios carried out during the POC (TEQMO traffic circle).




CRE1DART05