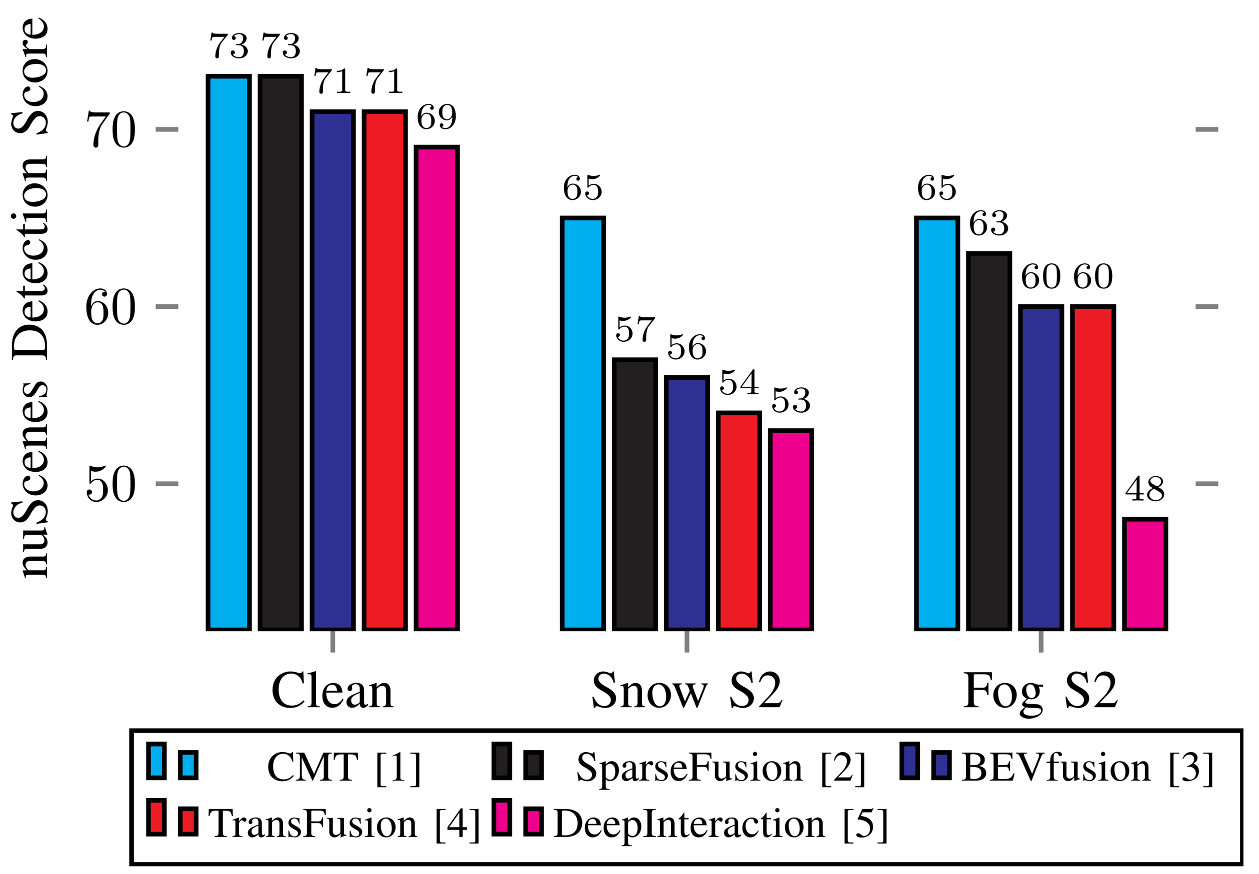
MultiCorrupt: A Multi-Modal Robustness Dataset and Benchmark of LiDAR-Camera Fusion for 3D Object Detection

Till Beemelmanns[1], Quan Zhang[2], Christian Geller[1], Lutz Eckstein[1]
[1]Institute for Automotive Engineering RWTH Aachen University, [2]Electrical Engineering and Computer Science TU Berlin

Introduction

Multi-modal 3D object detection models for autonomous driving have demonstrated exceptional performance on computer vision benchmarks like nuScenes. However, their reliance on densely sampled LiDAR point clouds and meticulously calibrated sensor arrays poses challenges for real-world applications. Issues such as sensor **misalignment**, **miscalibration**, and **disparate sampling frequencies** lead to spatial and temporal misalignment in data from LiDAR and cameras. Additionally, the integrity of LiDAR and camera data is often compromised by **adverse environmental** conditions such as inclement weather, leading to **occlusions** and **noise interference**.



Performance degradation of state-of-the-art multi-modal detectors for corruption *Snow* and *Fog* with a severity level of 2.

Method

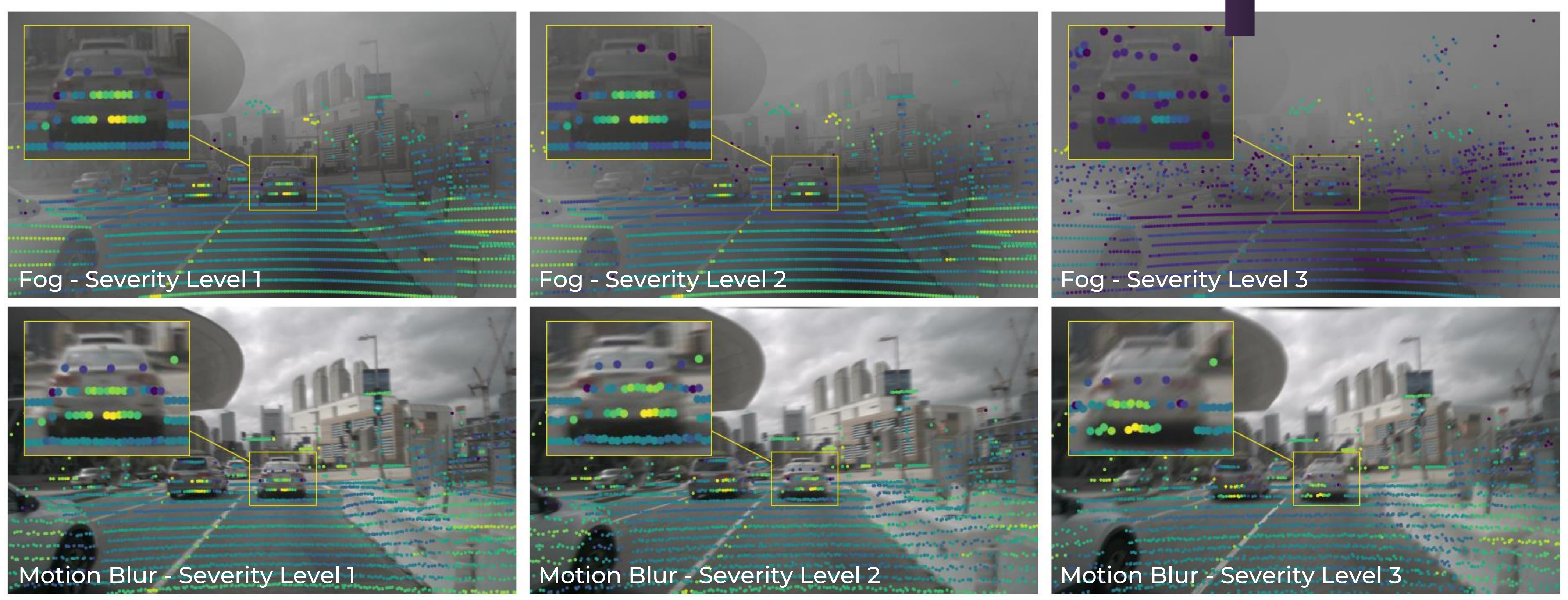
We introduce **MultiCorrupt**, a corrupted dataset based on nuScenes to evaluate LiDAR-camera fusion algorithms. We implement **ten synthetic corruptions** with **three severity levels**.

Corruption	Modality	Description
Darkness	C	Poisson Gaussian noise intensity s
Brightness	C	Addition of brightness in the HSV space
Points Reducing	L	Dropout points with probability p
Temporal Misalignment	LC	Frozen frame applied with probability p
Spatial Misalignment	LC	Extrinsic misalignment in degrees applied with probability p
Motion Blur	LC	Jitter noise from a Gaussian distribution with σ_t
Missing Camera	C	Dropping frames for multiple cameras with probability p
Beams Reducing	L	Number of beams remaining in the point cloud
Fog	LC	Approximated visibility in meters
Snow	LC	Approximated snowfall intensity in mm/h

We **benchmark** and **investigate the robustness** of five state-of-the-art 3D object detection models.

Method	mAP (%)	NDS (%)	Representation	Alignment	Fusion Mechanism
CMT [1]	70.28	72.90	BEV+images feature	learning & projection	self & cross attention
DeepInteraction [5]	68.72	69.09	BEV+images feature	learning & projection	cross attention
TransFusion [4]	66.72	70.84	BEV+images feature	projection	image as Q, LiDAR as K
Sparsefusion [2]	71.02	73.15	BEV+images feature	learning & projection	self-att. for LiDAR and images
BEVfusion [3]	68.72	71.44	BEV	depth and projection	concatenation

MultiCorrupt

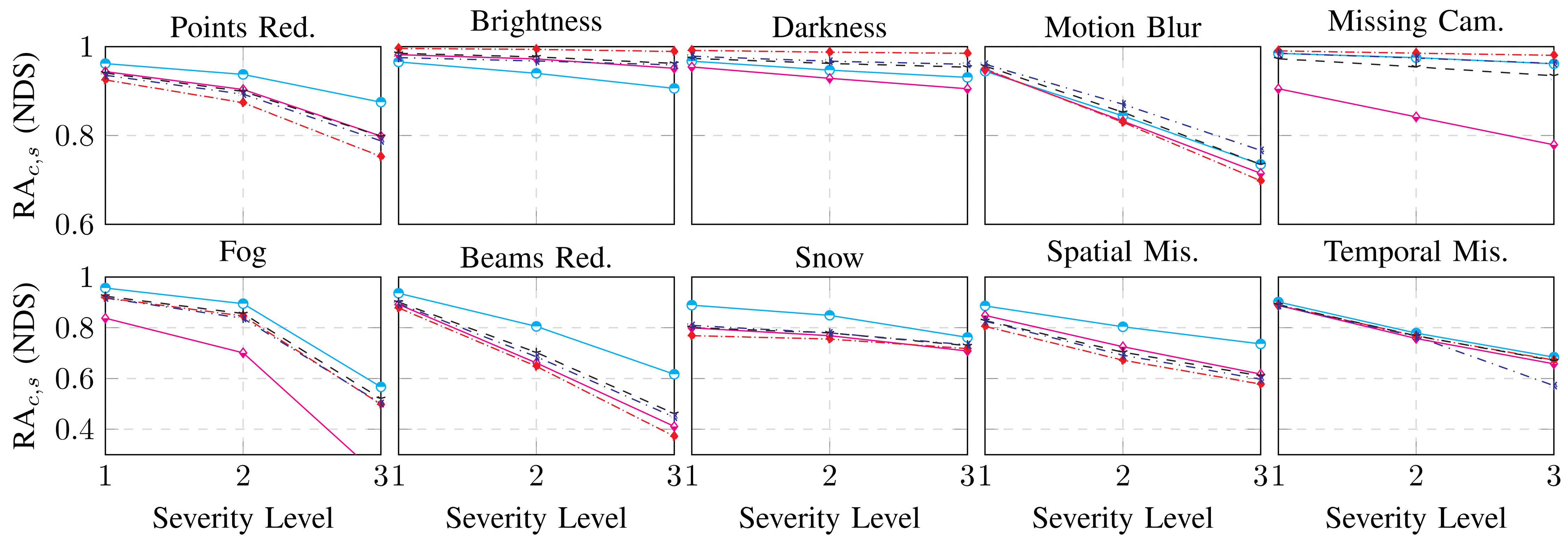


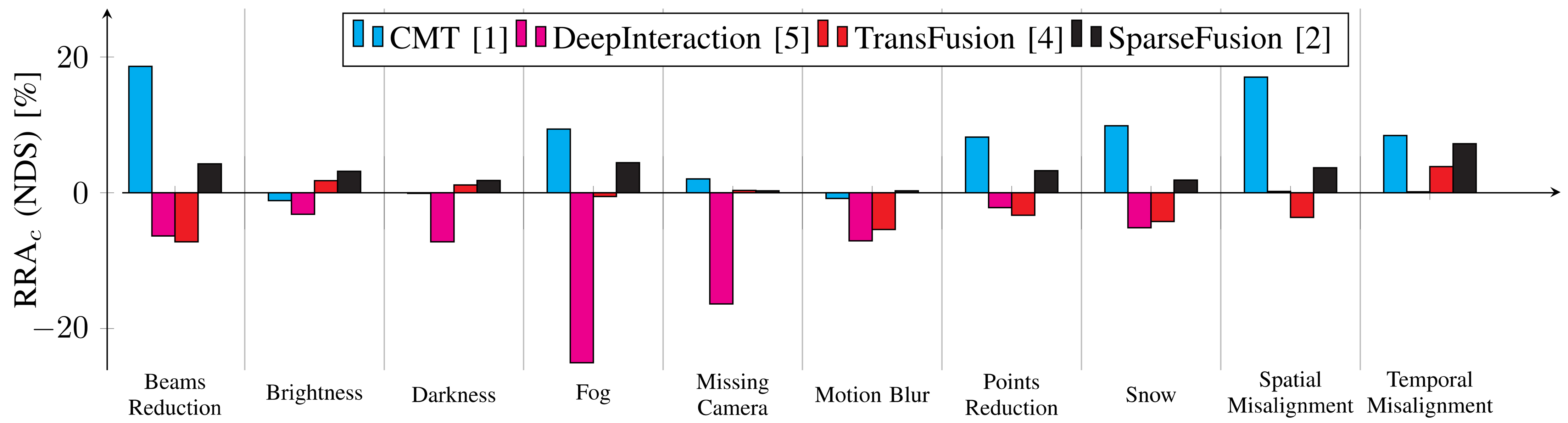
Example visualizations for *Fog* and *Motion Blur*. Visualizations for all corruptions available at github.com/ika-rwth-aachen/MultiCorrupt.

Results

Resistance ability RA_c computed with NDS score for all corruptions and severity levels.

Model	Beams Red.	Brightness	Darkness	Fog	Missing Cam.	Motion Blur	Points Red.	Snow	Spatial Mis.	Temporal Mis.	mRA
CMT [1]	**0.786**	0.937	0.948	**0.806**	0.974	0.841	**0.925**	**0.833**	**0.809**	**0.788**	**0.865**
DeepInteraction [5]	0.655	0.969	0.929	0.583	0.842	0.832	0.882	0.759	0.731	0.768	0.795
TransFusion [4]	0.633	**0.993**	**0.988**	0.754	**0.985**	0.826	0.851	0.748	0.685	0.777	0.824
SparseFusion [2]	0.689	0.975	0.963	0.767	0.954	0.848	0.879	0.770	0.714	0.777	0.834
BEVfusion [3]	0.676	0.967	0.969	0.752	0.974	**0.866**	0.872	0.774	0.705	0.742	0.830





Relative resistance ability RRA_c computed with NDS score using **BEVfusion** [3] as baseline model.

Metrics

The **Resistance Ability** RA_c is computed across the different severity levels with

$$RA_{c,s} = \frac{\mathcal{M}_{c,s}}{\mathcal{M}_{clean}} \qquad RA_c = \frac{1}{3}\sum_{s=1}^{3} RA_{c,s} \qquad mRA = \frac{1}{N}\sum_{c=1}^{N} RA_c$$

where $\mathcal{M}_{c,s}$ represents the measured metric (NDS or mAP) for corruption c at severity level s. The number of all corruptions is denoted by N, and $\mathcal{M}_{clean}$ is the performance on the original nuScenes dataset.

The **Relative Resistance Ability** RRA_c compares the relative robustness of each model for a specific corruption with a **baseline model**.

$$RRA_c = \frac{\sum_{s=1}^{3}(\mathcal{M}_{c,s})}{\sum_{s=1}^{3}(\mathcal{M}_{Baseline,c,s})} - 1 \qquad mRRA_c = \frac{1}{N}\sum_{i=1}^{N} RRA_c$$

The **Mean Relative Resistance Ability** $mRRA_c$ measures the relative robustness compared to a baseline model for all types of corruptions. We chose BEVfusion [3] as baseline.

Benchmark & Code

- **Open-Source Code** to create MultiCorrupt
- **Visualizations** of all corruptions
- More **detailed results** and **metrics**
- Updates with **new models**
- github.com/ika-rwth-aachen/MultiCorrupt

Conclusion

- Multi-modal 3D object detectors exhibit **different robustness** behavior depending on their specific **fusion**, **alignment** and **training strategies**
- **Robustness enhancing design choices** are **independent modality** handling, either through independent modality-spaces for Transformer tokens and queries or modality independent detection branches.
- **Masked-modal training** boosts robustness but requires further analysis if it is applicable across a variety of architectures.
- **Robustness diminishing factors** are **singular modality**-dependent query initialization or a deep coupling of multi-modal features early in the detection pipeline.

[1] Cross modal transformer via coordinates encoding for 3d object detection
[2] Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection
[3] Bevfusion multi-task multi-sensor fusion with unified bird's-eye view representation
[4] Transfusion: Robust lidar-camera fusion for 3d object detection with transformers
[5] Deepinteraction: 3d object detection via modality interaction

This work has received funding from the European Union's Horizon Europe Research and Innovation Programme under Gran Agreement No 101076754 - AIthena project.