MUSES: The Multi-Sensor Semantic Perception Dataset for Driving under Uncertainty

Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool

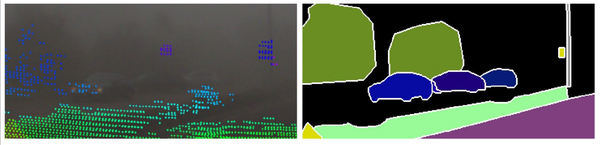
ETH zürich — Future Computing Laboratory

1 Abstract

We present the **multi-modal** MUSES **dataset** for driving under uncertainty and adverse conditions. MUSES includes 2500 images with diverse weather and illumination. Each image has high-quality 2D pixel-level **panoptic and uncertainty annotations**. It features calibrated and synchronized recordings from a frame camera, MEMS lidar, FMCW radar, HD event camera, and IMU/GNSS sensor, aiding sensor fusion for dense semantic understanding. MUSES enables the evaluation of multimodal perception systems in complex, real-world driving environments. Our dataset and bench-mark are publicly available at **https://muses.vision.ee.ethz.ch/**.

2 Motivation

- **Robust, all-weather visual perception** is essential for fully autonomous driving.
- Current datasets lack crucial **non-camera sensors** and **panoptic annotations** for challenging conditions.
- **Multi-modal** setup enables **better annotations**.
- **Uncertainty** estimates are crucial for **safety-critical downstream tasks**. MUSES provides ground truth annotations for aleatoric uncertainty in the RGB data.
- 2D panoptic annotations as highly adverse weather can yield **insufficient information for 3D annotation**:

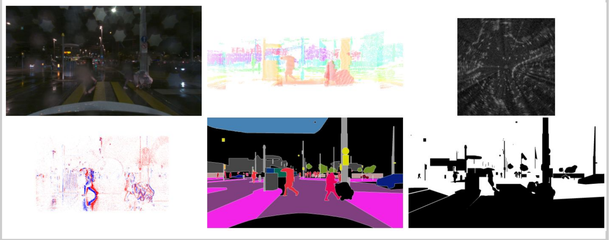


3 Supported Tasks & Public Benchmarks

- **Panoptic Segmentation**
- **Uncertainty-Aware Panoptic Segmentation**
- Semantic Segmentation
- Object Detection

4 Sensor Suite

Modality	Name	Specifications
Frame camera	TRI023S-CC	8-bit RGB, 30 Hz, 1920×1080, HFOV: 77°, VFOV: 43°
Event camera	Prophesee GEN4.1	1280×720, 15M events/s, HFOV: 64°, VFOV: 39°
Lidar	RS-LiDAR-M	10 Hz, avg. angular resolution: 0.2°, range: 200 m, HFOV: 120°, VFOV: 25°, 75K points/scan
Radar	Navtech CIR-DEV	4 Hz, range resolution: 43.8 mm, horizontal angular resolution: 0.9°, range: 330 m
IMU/GNSS	simpleRTK2B Fusion	RTK accuracy: <10cm, 30 Hz

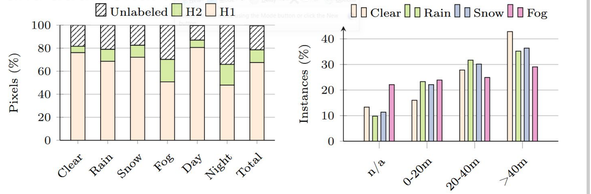


5 Dataset Statistics

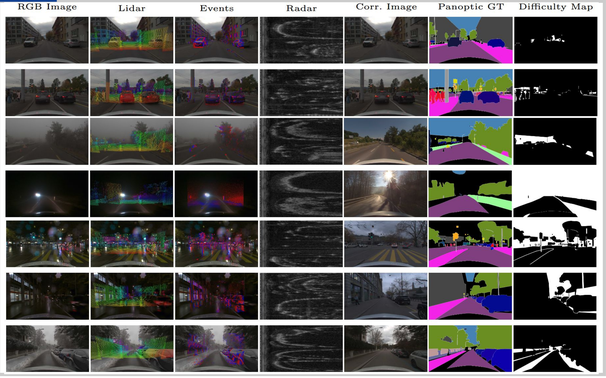
- 2500 Synchronized and calibrated multimodal recordings from **frame camera, lidar, radar** and **event camera** with panoptic annotations captured in various weather and illumination scenarios.
- For each adverse-condition scene a **corresponding reference frame** of the same scene taken under normal clear weather and daytime is provided.
- All the sensors are mounted over the windshield.
- **Recordings in Switzerland** during the period between November 2022 and July 2023.
- The dataset mostly consists of recordings which took place **in urban areas** but also on **highways** and **in rural regions**.
- Split along two axis: the weather **(clear/fog/rain/snow)** and the illumination conditions (**daytime/nighttime**).

6 Annotations

- **Panoptic segmentation annotations** created by a professional team of annotators.
- **19 semantic classes**, fully compatible with the evaluation classes of the Cityscapes dataset.
- Sophisticated **two-step annotation protocol:**
  - Step 1 (H1): Annotation of **frame camera** image alone
  - Step 2 (H2): Annotation of camera image with **auxiliary data**: all projected sensors, clear weather reference image and clear and adverse videos.
- Comparison of the two stages results in **class-level** and **instance-level aleatoric uncertainty** for the frame camera.
- **High-quality:** ~50% annotation time for **quality control** and **additional labels in Stage 2**:

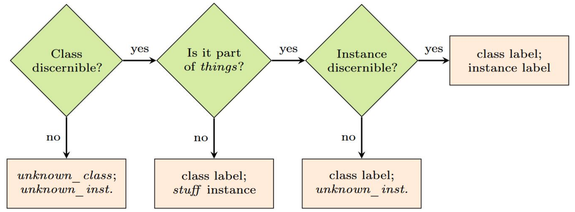


7 Visualization of Samples

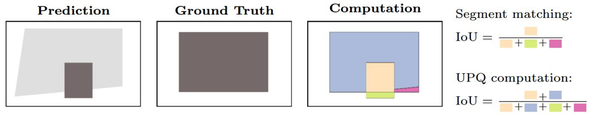


8 New Task: Uncertainty-Aware Panoptic Segmentation

- Ground truth uncertainty from two-step annotations:



- New metric **AUPQ** for the proposed task:
  - Reward models that are uncertain for pixels whose semantics are not unambiguously discernible from the available data (aleatoric uncertainty).
- Differentiate between **instance** and **class uncertainty**



9 Experiments

- We **benchmark** for smenatic segmetnation and multi-modal panoptic segmetnation
- Test the **generalization** of models trained on different datasets

Architecture	mIoU
DeepLabv3+ (ResNet101-D8) [10]	70.5
OCRNet (HRNetV2p-W48) [48]	71.9
SETR (ViT-L) [52]	71.1
SegFormer (MiT-B2) [45]	72.5
SegFormer (MiT-B5) [45]	74.7
Mask2Former (Swin-T) [11]	70.7
Mask2Former (Swin-L) [11]	**77.1**

Frame camera	Event camera	Radar	Lidar	Clear	Fog	Rain	Snow	Day	Night	All
✓				48.8	46.5	45.4	42.2	49.4	39.4	46.9
✓	✓			52.1	49.4	48.2	42.6	51.7	42.2	49.5
✓		✓		52.9	49.5	49.9	46.1	52.9	44.8	51.3
✓			✓	54.2	49.9	52.2	47.6	53.7	48.0	52.7
✓	✓	✓	✓	**55.3**	**50.3**	**53.8**	**47.9**	**54.1**	**49.7**	**53.6**

Training Dataset	Cityscapes-val [13]		ACDC-test [34]		MUSES-test		Mean
	mIoU	Δ	mIoU	Δ	mIoU	Δ	mIoU
Cityscapes [13]	83.7	—	65.7	-10.6	58.9	-18.2	69.4
ACDC [34]	70.9	-12.8	76.3	—	66.9	**-10.2**	71.4
MUSES	73.1	**-10.6**	72.0	**-4.3**	77.1	—	**74.1**