

jbDATA — A Collaborative Framework for Automatic Curation and Creation of a Dataset for AI Functions in Autonomous Driving

A. Nowzad¹, M. Schötz², H. Königshof⁵, R. Aufschläger⁸, F. Bally⁸, D. Bogdoll⁵, M. Heinrich⁵, W. Huang¹, D. K. Karunakaran², E. Kraft¹, R. Polley⁵, P. Reis⁵, K. Rösch⁵, C. Schlauch¹, M. K. Schöll⁸, P. Schörner⁵, C. Sigl³, T. Stiehm⁷, P. Szabados⁷, S. Weber³, Y. Yao¹, T. Zajusch⁴, N. Klein⁶

¹Aumovio SE · ²AVL Deutschland GmbH · ³b-plus technologies GmbH · ⁴EICT GmbH · ⁵FZI Research Center for Information Technology · ⁶Karlsruhe Institute of Technology (KIT) · ⁷Luxoft GmbH · ⁸Deggendorf Institute of Technology



Motivation & Contribution

- Passive "record everything, filter later" pipelines waste storage on redundant data while safety-critical anomalies stay underrepresented
- jbDATA's Smart Data Loop evaluates relevance, novelty, diversity, and coverage directly during acquisition
- Edge devices enrich the sensor stream with metadata and a lightweight environment model
- Trigger functions decide online what gets stored → closing the loop between acquisition and model development

25.9%

dataset balance improvement

82.2%

reduction in retained data volume

+5%

downstream model performance

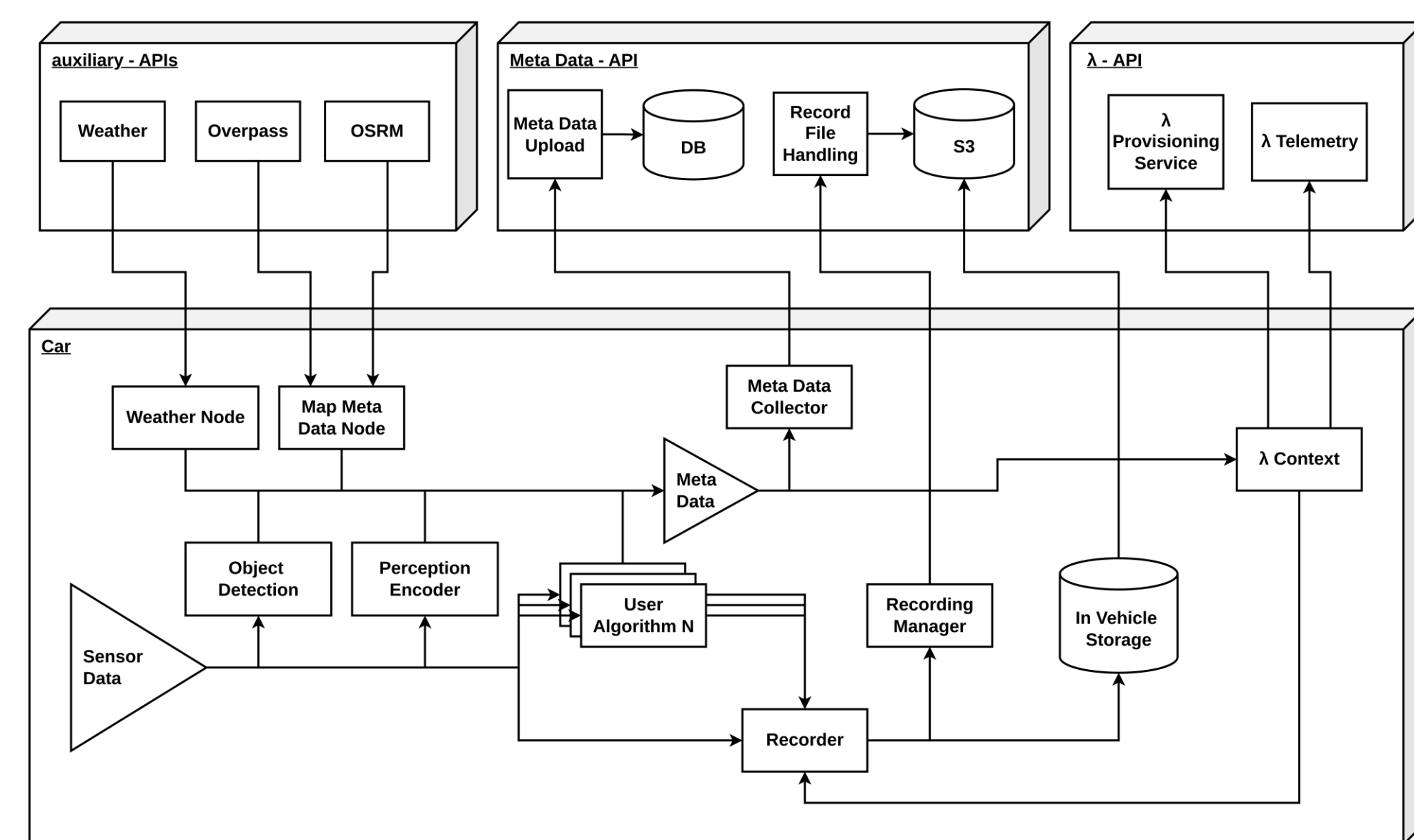


Fig. 3: Edge Platform Dataflow and Cloud Integration

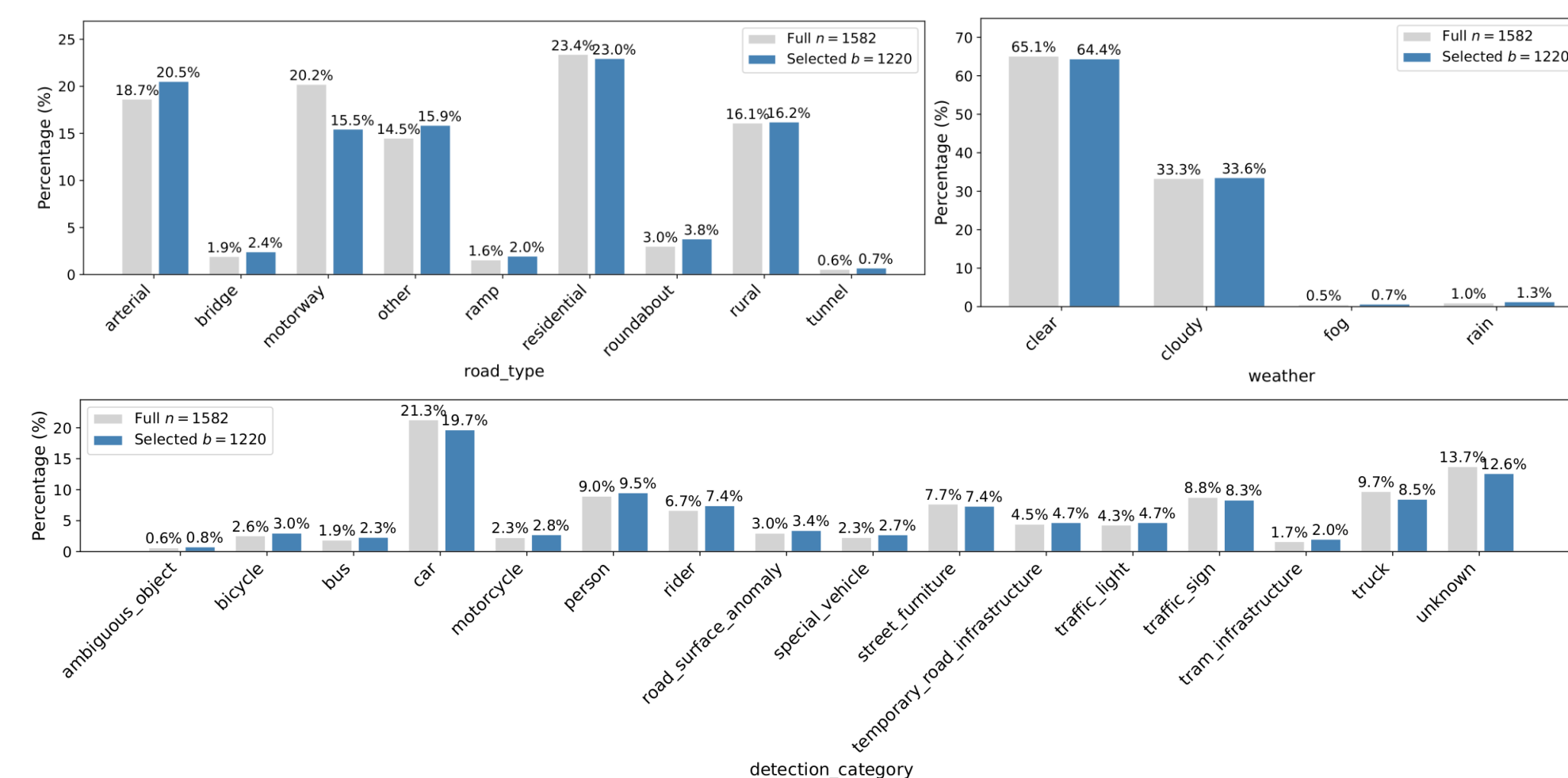


Fig. 4: Distribution comparisons between full (grey) and published (blue) datasets for the three considered pseudo-label categories.



Fig. 1: Research vehicles used in this work. Left to right: three recording vehicles (AVL, b-plus and FZI) contributing data to the published jbDATASET and two development vehicles used for Smart Data Loop development (THD and Valeo).

Closed-Loop Architecture

- Data collection is modeled as a discrete-time dynamical system: each sample is embedded, evaluated, and stored or discarded via a capture rule
- This makes acquisition path-dependent, continuously improving dataset diversity, balance, and coverage (Fig. 2)
- Realized on hardware via an edge-cloud architecture: on-vehicle triggers decide online what gets recorded and uploaded
- Cloud aggregates fleet-level statistics and redeploys refined trigger configurations back to the fleet (Fig. 3)

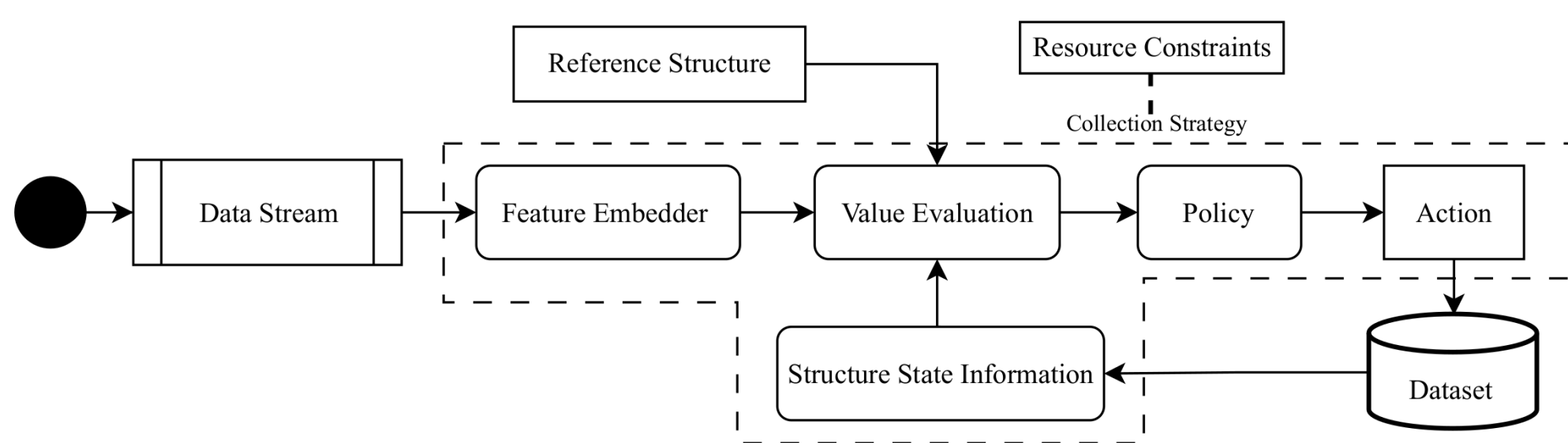


Fig. 2: Control-theoretic understanding of closed-loop dataset collection as a dynamical system



Fig. 5: Randomly selected camera frames illustrating the visual diversity of the publication subset.

Targeted Synthetic Enrichment

- Real-world data has limited coverage of rare objects, adverse weather, and long-tail scenarios
- ControlNet-based generation (trained on 1,000h of driving video) produces annotated images to specifically fill identified coverage gaps
- Wet/snowy/tunnel scenes: 62.8% of synthetic data; night/dawn-dusk: 53.5% - deliberately biased toward challenging conditions

77.65%

LA-mIoU layout agreement

94.4%

images rated error-free

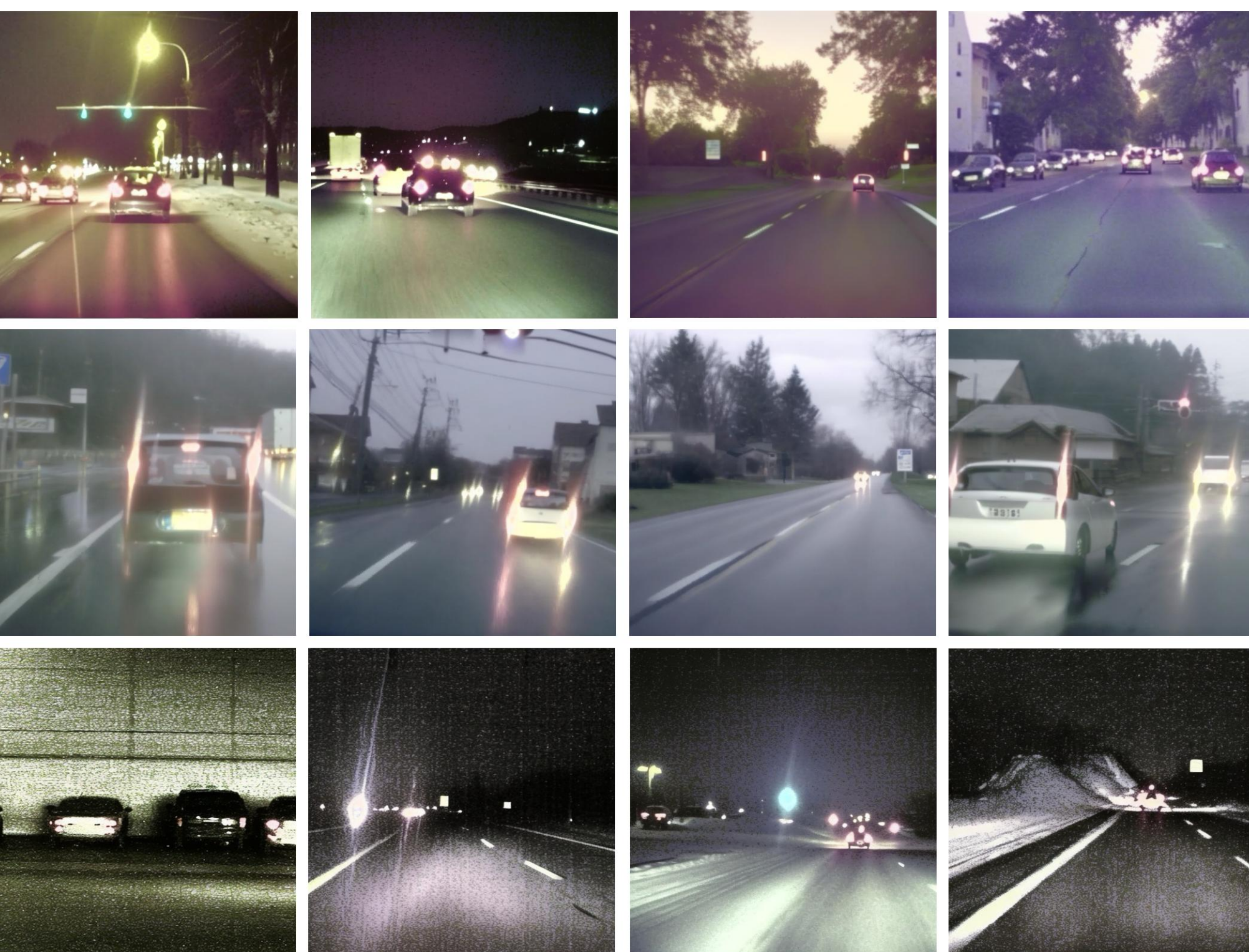


Fig. 6: Our synthetic dataset includes challenging visibility conditions. **Top row:** reduced camera visibility at night, dawn, and dusk, with artificial light sources. **Middle row:** impaired visibility in rainy scenes, including water droplets and lens distortions. **Bottom row:** images degraded by sensor noise under insufficient illumination.

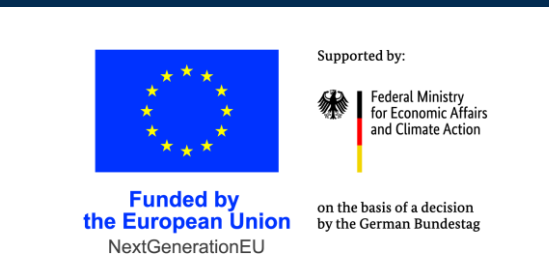


The final jbDATASET with 5 hours of real sensor data plus 10,000 synthetic frames is available at www.justbetterdata.de/dataset

For more information regarding the poster, contact: koenigshof@fzi.de

Consortium partners:
Aumovio · AVL · b-plus · FZI · Luxoft · Mercedes · TH Deggendorf · Valeo

just better DATA is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Energy.



[linkedin.com/company/ki-familie](https://www.linkedin.com/company/ki-familie)
justbetterdata.de