SPECTRA: Synchronized Stereo Event-Camera Driving Dataset for Diverse Perception Tasks

Jihed Dachraoui[1*], Anass El Moudni[1*], Elsa Planterose[2], Sebastien Kramm[1], Fabio Morbidi[3], Rémi Boutteau[1]

***Abstract*— Event-based vision is emerging as a transformative technology in the field of autonomous driving, offering high temporal resolution, low latency, and robust performance under challenging conditions such as low light and high dynamic range. To support the academic and practical success of this technology, the availability of diverse and high-quality datasets is critical for training and evaluating deep learning models effectively. Recognizing this need, we introduce a novel dataset specifically designed for event-based stereo vision in autonomous driving scenarios. Our dataset combines data from event-based cameras, RGB cameras, LiDARs, and IMUs, offering a multimodal foundation for addressing a wide range of perception tasks. It includes precise ground truths for object detection, depth estimation, and pose tracking, enabling researchers to develop and benchmark models across multiple tasks. The dataset is meticulously synchronized across all sensors to facilitate the exploration of sensor fusion strategies and the development of algorithms tailored for event-based perception.**

I. INTRODUCTION

The development of robust perception systems is a cornerstone of robotics and autonomous vehicle technologies. While traditional frame-based cameras are widely employed, their limitations in high dynamic range environments, fast motion, and challenging lighting conditions have spurred interest in alternative sensing modalities. Event cameras, which asynchronously detect pixel intensity changes, have emerged as a promising solution. Their high temporal resolution, low latency, and robustness to extreme lighting conditions make them particularly advantageous for capturing dynamic and complex scenes. Stereo event cameras expand these capabilities by enabling depth perception and 3D scene understanding, which are critical for tasks such as obstacle detection, navigation, and localization. Despite their potential, research into stereo event-based perception has been limited by the scarcity of high-quality datasets that capture real-world driving scenarios.

To bridge this gap, we introduce a novel dataset specifically tailored for stereo event cameras in autonomous driving applications. Our dataset captures diverse and challenging environmental conditions, including variations in illumination, weather, and motion dynamics. It features synchronized data from stereo event cameras, stereo high-resolution frame-based cameras, a lidar, and a GNSS/IMU system with RTK corrections. This multimodal dataset provides a comprehensive resource for advancing research in perception, localization, and mapping.

The main contributions of this work are:

- **Dataset Creation:** A high-quality dataset integrating stereo event cameras, stereo high-resolution frame-based cameras, lidar, and RTK-corrected GNSS/IMU data. The dataset is designed to address the unique challenges of autonomous driving in real-world environments.
- **High-Precision Ground Truth:** The inclusion of centimeter-level accurate localization data from RTK-corrected GNSS/IMU systems enables precise trajectory estimation and evaluation of localization algorithms.
- **Diverse Environmental Conditions:** Data collected under challenging scenarios such as night driving, adverse weather, and high-speed motion to ensure the dataset reflects the complexities of real-world operations.

This dataset is a significant step toward addressing the challenges of stereo event-based perception in robotics and autonomous systems. By offering synchronized, multimodal data under diverse conditions, it provides researchers with the tools necessary to advance the state of the art in perception and localization for real-world autonomous systems. .

II. STATE OF THE ART

Why driving datasets matter? Datasets underpin progress in autonomous driving: they provide the shared substrate on which algorithms for depth estimation, tracking/odometry, and SLAM are trained, validated, and fairly compared. The consolidation brought by KITTI [1]—synchronized cameras, LiDAR, GPS/IMU, and public leaderboards—demonstrated how carefully curated, real-world data can catalyse entire research threads. An analogous push is underway in event-based vision, where the sensing modality promises robustness to HDR and fast motion, but where the scarcity of standardized, driving-focused datasets has historically limited reproducibility and comparison.

Early driving resources (DDD17/DDD20 [2], [3]) targeted driver-assistance and steering-angle prediction. MVSEC (of-

*These authors contributed equally to this work

[1]Université Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie, Normandie Université, LITIS UR 4108, Rouen, France, {anass.el-moudni,jihed.dachraoui, sebastien.kramm,remi.boutteau}@univ-rouen.fr

[2] INSA Rouen Normandie, Université Rouen Normandie, Université Le Havre Normandie, Normandie Université, LITIS UR 4108, Rouen, France, elsa.planterose@insa-rouen.fr

[3] Université de Picardie Jules Verne, MIS laboratory, UR 4290, Amiens, France, fabio.morbidi@u-picardie.fr

ten written as MVSEC/LVSEC) [4] provided one of the first *multi-sensor*, driving-oriented benchmarks with synchronized events, frames, LiDAR, and IMU, plus derived ground truth for optical flow and SLAM. DSEC [5] substantially scaled up spatial resolution and scenario diversity (day/night), introduced improved optical-flow annotations, and inspired task-specific extensions (DSEC-Detection [6], DSEC-Semantics [7]). Complementary efforts broaden sensing and domains: VECtor [8] focuses on SLAM-oriented sequences, ECMD [9] augments stereo events with infrared imagery, and ViViD++ [10] tightly synchronizes events, RGB, thermal, LiDAR, and IMU. M3ED [11] spans road/drone/legged platforms with rich labels (LiDAR depth, semantics, optical flow), while CoSEC [12] introduces a beam-splitter rig that yields pixel-aligned event/RGB streams.

Observed limitations in MVSEC, DSEC, and M3ED : Despite their impact, commonly reported pain points remain for depth-centric evaluation:

- **Sparse depth supervision.** Depth is typically obtained by projecting LiDAR onto the image plane, yielding sparse or semi-dense supervision that limits supervised training of dense depth networks and complicates fine-grained error analysis.
- **Pose and timing nuances.** Even with rigorous calibration, long-term extrinsic drift, residual timestamp jitter, or per-sensor time offsets can introduce small alignment errors that accumulate in mapping/SLAM evaluations (notably on fast urban drives).
- **Condition coverage.** Night, adverse weather, lens artefacts, and high-speed manoeuvres are underrepresented or unevenly distributed across sequences, making robustness studies harder.

Our positioning: SPECTRA: To address these gaps for stereo event-based *driving* scenarios, SPECTRA is designed with three principles:

1) **Tight multi-sensor synchrony.** Stereo events and high-resolution frames are time-aligned to LiDAR and RTK-GNSS/IMU, with documented calibration and periodic revalidation to reduce long-horizon drift.
2) **Enhanced depth supervision.** In addition to native LiDAR projections, we provide *densified depth targets* produced by a dedicated fusion-and-regularisation pipeline, specifically crafted to supervise deep stereo event depth models while controlling bias and preserving metric scale.
3) **Driving-oriented protocols.** We release task-specific splits and metrics for stereo event depth and tracking/odometry, spanning day/night, adverse weather, and high-speed segments to stress-test robustness and generalization.

III. Dataset

The system consists of two Prophesee EVK4 event cameras, each featuring IMX636 sensors with a 1280x720 resolution, 1/2.5" sensor, and a 45 cm baseline. Additionally, it includes two FLIR Blackfly S cameras (BFS-PGE-13Y3C-C) with 1280x720 resolution (binned), 1/2" sensors, global shutter color capability, and a 47 cm baseline. The Blackfly cameras operate in auto-exposure mode, capturing data at 20 Hz. For 3D perception, the system integrates an Ouster OS1-32 LiDAR sensor with 32 vertical channels, 1024 horizontal points, and a 100-meter range, also operating at 20 Hz. Positioning and navigation are managed by the ixblue Atlans A7 system, which includes a Septentrio AsteRx4 GNSS receiver and a 6-DoF IMU, offering a 200 Hz update rate and NTRIP RTK corrections, ensuring positioning accuracy of ±0.6-1 cm. All sensors are oriented in the forward-facing direction. The system is synchronized using an Arduino Due card, and all raw data is collected and processed onboard a NUC 11TNKv7, which acts as the central storage and processing unit, the system is shown in Figure 1.

Fig. 1. Real prototype of our data-acquisition system. The sensor suite comprises dual RGB and dual event cameras, a LiDAR, and an IMU. An Arduino-based timing unit and a GPS-RTK receiver are mounted inside the enclosure for hardware-level synchronization.

A. Calibration and Synchronization

We perform extrinsic calibration among event/RGB cameras, IMU, and LiDAR with precise hardware time synchronization. The calibration pipeline is as follows: e2calib [13] and Kalibr [14] for camera–camera, livox_camera_calib [15] and MATLAB [16] for LiDAR–camera, and GRIL-Calib [17] for LiDAR–IMU; calibration files (`.yaml`, `.txt`) are included.

B. Sequences

We have collected over 4 TB of raw data from drives in urban/ suburban areas near Rouen and Saint-Étienne-du-Rouvray, France. To ensure high-quality standards, the data is carefully reviewed, and only sequences with superior recording and calibration quality are retained. The dataset currently consists of over 30 sequences captured during the day, night, dawn, and sunrise with several scenarios low and high speed, textured and textureless scenes, with high and low flows of people and vehicles with plans to expand it in the future by including additional sequences recorded under a wider range of conditions.

In building our dataset, we deliberately covered a broad range of scene dynamics high-speed runs, moderate urban cruising, straight-line traversals, and circular motion at roundabouts across industrial areas, vegetated roads, uptown and downtown streets. The event modality captures diverse

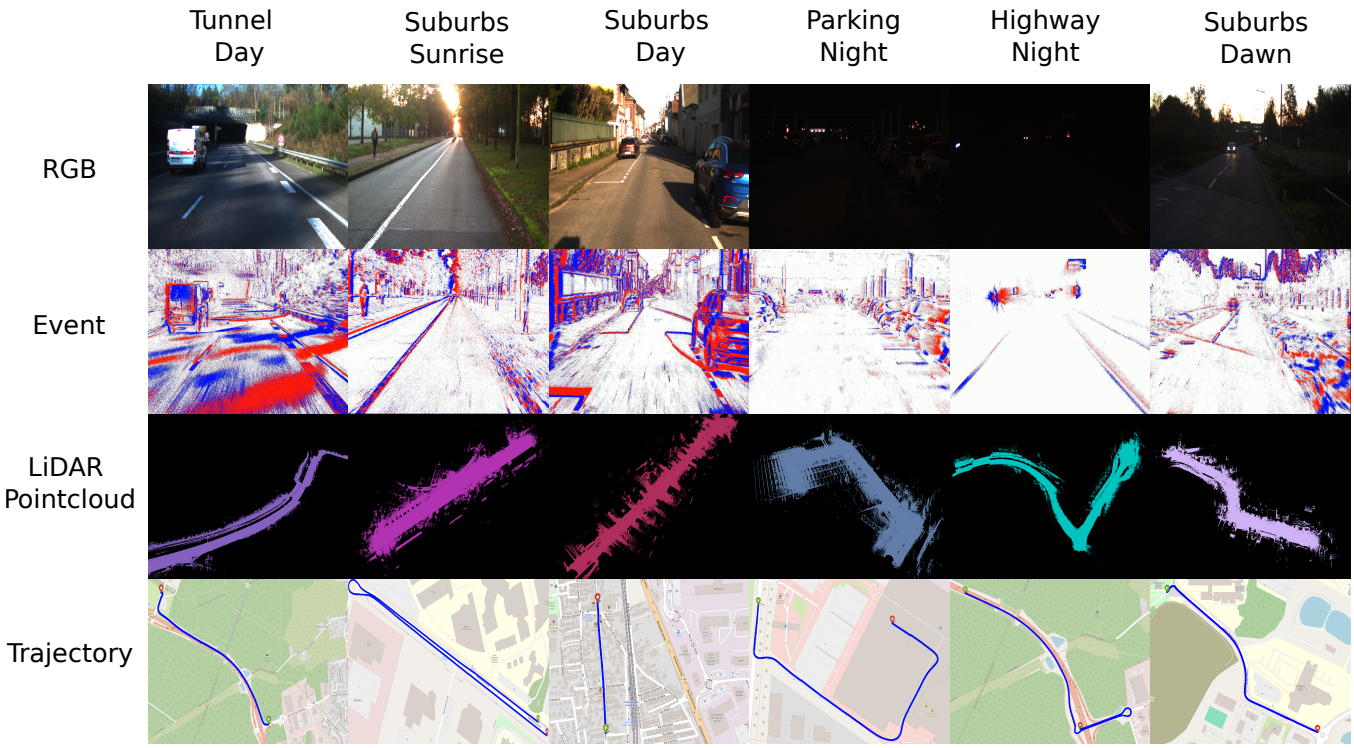

Fig. 2. Overview of the SPECTRA data acquisitions. Rows show sensing modalities (event camera, RGB, LiDAR, and IMU); columns group the different scenarios and environmental conditions (urban, suburban, highway, and varied weather/day–night). Each panel displays synchronized snippets for the corresponding modality and scenario.

objects, including pedestrians, traffic signs/lights, cars, buses, scooters, bicycles, strollers, and road-imprinted markings. We include both ego-motion sequences (vehicle moving, most pixels dynamic) and stationary intervals (e.g., red lights), where only independently moving objects trigger events making the data well suited to moving-object segmentation.

IV. Ground Truths Generation

We generate ground-truth labels by fusing LiDAR, IMU, and cameras in a tightly time-synchronized pipeline. The process yields (i) accurate ego-trajectory and a registered point cloud, (ii) per-frame sparse depth via LiDAR reprojection, (iii) dense depth maps via label-aware diffusion guided by semantics, and (iv) object-detection annotations.

A. Trajectory & Registered Point Cloud (LiDAR+IMU with Faster-LIO)

We estimate the platform pose with Faster-LIO, which tightly couples IMU preintegration with LiDAR scan matching and deskewing. Given hardware PPS-based timing, Faster-LIO outputs timestamped poses $\mathbf{T}_{WI}(t)$ and an accumulated, globally registered point cloud. This serves as both the odometry ground truth and the source of 3D points for depth supervision.

B. Sparse Depth via LiDAR-to-Camera Reprojection

Using calibrated extrinsics (IMU→camera and LiDAR→camera) and intrinsics, we transform each registered LiDAR point $\mathbf{p}^W$ into the camera frame at time t and project it onto the image plane. We keep only points with positive depth and apply a z-buffer to retain the closest surface per pixel, yielding a *sparse* but metrically accurate depth map aligned to the (left) camera/event frame. We limit points to a fixed range (e.g., $100m$) and discard out-of-FoV samples.

C. Dense Depth via Semantic, Label-Aware Diffusion

To address the sparsity typical of LiDAR-projected labels, we densify depth using image semantics:

- **Semantic masks.** We run a DeepLabV3+-style segmenter on the RGB frame aligned to the event geometry, producing class (or instance) regions.
- **Multi-modal seeding.** Within each semantic region, sparse depths are clustered with 1D K-means ($K \leq 3$) to capture multiple surfaces (e.g., hood vs. windshield).
- **Label-gated diffusion.** Cluster centers propagate to unlabeled pixels using a bilateral affinity (spatial and appearance), *restricted to the same semantic region* to preserve boundaries. A few Jacobi smoothing iterations optionally reduce artifacts while keeping original LiDAR seeds fixed.

The result is a *dense*, boundary-aware metric depth map that mitigates the sparsity and moving-object bias seen in prior driving datasets relying solely on raw LiDAR projections.

D. Object Detection Annotations

For object-detection ground truth, we employ a YOLO-family detector (YOLOv10) to pre-label bounding boxes on RGB frames (and, optionally, event reconstructions), followed by human verification/correction. Final annotations

are exported in standard YOLO format and aligned frame-by-frame with the rest of the modalities, enabling detection and tracking benchmarks.

Reproductibility: We release scripts for (i) Faster-LIO pose estimation, (ii) LiDAR reprojection with z-buffering, (iii) semantic inference, and (iv) label-aware diffusion, ensuring identical preprocessing across sequences and methods. An example of the modalities are shown in the Figure 2

V. Conclusion

This paper introduced **SPECTRA**, a synchronized, multimodal driving dataset centered on *stereo event cameras* and complemented by stereo RGB, a LiDAR, and RTK-aided GNSS/IMU. We detailed the acquisition platform, hardware timing via PPS fan-out, and comprehensive spatial–temporal calibration; described the dataset organization (raw rosbags and learning-ready exports); and outlined our pipeline for generating dense, learning-useful depth targets from LiDAR reprojected into the image plane with label-aware diffusion. The sequences span diverse operating conditions (day/night, tunnels, high-speed segments, urban/suburban texture and traffic), with aligned annotations (poses, masks, boxes) and consistent file formats to support both model-based SLAM/VO and deep learning.

What SPECTRA enables. By pairing tightly time-aligned stereo events with reliable trajectory and depth supervision, SPECTRA enables rigorous benchmarking for stereo event depth, VO/SLAM, moving-object segmentation, cross-modal fusion (event+RGB+LiDAR+IMU), and semi/self-supervised training in HDR and high-motion regimes. The dual distribution raw ROS topics and preprocessed PNG/HDF5/poses lowers the barrier for both robotics pipelines and vision models.

To address the limitations, we are finalizing a Renault ZOE capture platform with water-sealed mounts and improved timing fan-out for all-weather recording, expanding scenarios (snow/heavy rain/fog, highways, rural), and planning higher-density LiDAR captures. On the annotation side, we aim to (i) release per-pixel confidence for densified depth, (ii) extend labels for long-range movers and adverse weather, and (iii) host standardized splits and a public evaluation server for stereo event depth and odometry.

In summary, SPECTRA complements existing resources by combining strict time geometry alignment with learning-ready depth supervision at automotive ranges, across challenging real-world conditions. We will continuously refresh the corpus with new sequences and tooling to foster fair, reproducible comparisons.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd17: End-to-end davis driving dataset," *arXiv preprint arXiv:1711.01458*, 2017.

[3] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *2020 IEEE 23rd international conference on intelligent transportation systems (ITSC)*. IEEE, 2020, pp. 1–6.

[4] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The Multivehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception," *IEEE Rob. Autom. Lett.*, vol. 3, no. 3, pp. 2032–2039, 2018.

[5] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," *IEEE Rob. Autom. Lett.*, vol. 6, no. 3, pp. 4947–4954, 2021.

[6] D. Gehrig and D. Scaramuzza, "Low-latency automotive vision with event cameras," *Nature*, vol. 629, no. 8014, pp. 1034–1040, 2024.

[7] Z. Sun*, N. Messikommer*, D. Gehrig, and D. Scaramuzza, "Ess: Learning event-based semantic segmentation from still images," *European Conference on Computer Vision. (ECCV)*, 2022.

[8] L. Gao, Y. Liang, J. Yang, S. Wu, C. Wang, J. Chen, and L. Kneip, "VECtor: A Versatile Event-Centric Benchmarkfor Multi-Sensor SLAM," *IEEE Rob. Autom. Lett.*, vol. 7, no. 3, pp. 8217–8224, 2022.

[9] P. Chen, W. Guan, F. Huang, Y. Zhong, W. Wen, L.-T. Hsu, and P. Lu, "Ecmd: An event-centric multisensory driving dataset for slam," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 407–416, 2023.

[10] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: Vision for visibility dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6282–6289, 2022.

[11] K. Chaney, F. Cladera, Z. Wang, A. Bisulco, M. A. Hsieh, C. Korpela, V. Kumar, C. J. Taylor, and K. Daniilidis, "M3ed: Multi-robot, multi-sensor, multi-environment event dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 4015–4022.

[12] S. Peng, H. Zhou, H. Dong, Z. Shi, H. Liu, Y. Duan, Y. Chang, and L. Yan, "Cosec: A coaxial stereo event camera dataset for autonomous driving," *arXiv preprint arXiv:2408.08500*, 2024.

[13] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza, "How to calibrate your event camera," in *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, June 2021.

[14] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1280–1286.

[15] C. Yuan, X. Liu, X. Hong, and F. Zhang, "Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments," 2021. [Online]. Available: https://arxiv.org/abs/2103.01627

[16] MathWorks, "Lidar camera calibrator," 2021. [Online]. Available: https://fr.mathworks.com/help/lidar/ref/lidarcameracalibrator-app.html

[17] T. Kim, G. Pak, and E. Kim, "Gril-calib: Targetless ground robot imu-lidar extrinsic calibration method using ground plane motion constraints," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, p. 5409–5416, June 2024. [Online]. Available: http://dx.doi.org/10.1109/LRA.2024.3392081