Multi-Modal Indoor Dataset for Event-based Monocular Depth Estimation by Mobile Robots

Ignacio Bugueno-Cordova[1,3,4], Luna Gava[2], Rodrigo Verschae[3], Javier Ruiz-del-Solar[4], Nicolás Navarro-Guerrero[1]

***Abstract*— This article introduces a multi-modal indoor dataset for event-based monocular depth estimation by mobile robots. The dataset was recorded on a humanoid platform and includes synchronized RGB, depth, event streams, and IMU data from Intel RealSense D435i, DAVIS346, and Prophesee EVK4 sensors. To provide a baseline, we implement a CycleGAN model that learns bidirectional mappings between the event-representation and the depth domain. We evaluate multiple state-of-the-art representations showing that event-based inputs could outperform frame-only inputs across accuracy, perceptual quality, and geometric reliability. The dataset and baseline together provide a reproducible testbed for event-based perception in indoor mobile robotics. The dataset will be available at the project website: `https://ibugueno.github.io/mmid-event-depth-dataset/`.**

I. Introduction

Social robots are becoming increasingly common in home assistance, healthcare, public venues, and education, where they must navigate dynamic indoor environments safely and interact effectively [1]. Social environments are particularly demanding, requiring robust human–robot interaction and advanced robotic vision methods to cope with frequent motion, clutter, and varying illumination [2]. Frame-based cameras, the dominant sensing modality, suffer from motion blur, latency, and reduced robustness in high-dynamic-range conditions, limiting mobile robot perception. Event-based cameras overcome these issues by asynchronously encoding brightness changes with microsecond resolution, high dynamic range, and minimal latency [3]. Although studied for depth estimation [4], optical flow [5], and odometry [6], their use in human–robot interaction remains largely unexplored, with only a prior simulation study [7] suggesting their potential but without real-world validation.

Most event-based depth datasets focus on autonomous driving or controlled indoor scenes rather than mobile robot navigation in indoor environments. MVSEC [8] provides stereo events with IMU and LiDAR depth for outdoor and handheld evaluation. DSEC [9] offers high-resolution stereo events under challenging illumination with synchronized LiDAR and GPS. EVIMO2 [10] includes dense depth and segmentation in indoor scenes with multiple moving objects. While these benchmarks advanced event-based perception, none address the challenges of indoor robot navigation, where narrow corridors, clutter, and variable lighting require low-latency and robust depth estimation.

Event-only models leverage the temporal consistency and high dynamic range of event streams, eliminating the need for frames. Prior work ranges from recurrent architectures [11] to transformer-based designs such as EReFormer [12], as well as distillation [13] and neighborhood-preserving methods [14]. Fusion approaches integrate events and frames for complementary cues, including asynchronous recurrent networks [15], unified transformers with temporal encoders [16], and reliability-oriented attention schemes [17]. Self-supervised and cross-modal strategies reduce annotation cost by enforcing consistency during training [18] or leveraging spiking neural networks for energy-efficient inference [19].

This article introduces a multi-modal indoor dataset for event-based monocular depth estimation by mobile social robots. The dataset includes synchronized RGB, depth, event streams, and IMU data from Intel RealSense D435i, DAVIS346, and Prophesee EVK4 sensors. We further evaluate this dataset with a CycleGAN architecture that learns bidirectional mappings between the event and depth domains, providing a baseline for future methods.

II. Preliminaries

A. Event data

Event cameras asynchronously encode brightness changes in the log-intensity signal $L(\mathbf{u}_k, t_k) = \log(I(\mathbf{u}_k, t_k))$ [3]. An event at pixel $\mathbf{u}_k = (x_k, y_k)^T$ and time t_k is triggered when the change since the last event exceeds a contrast threshold $C > 0$:

$$L(\mathbf{u}_k, t_k) - L(\mathbf{u}_k, t_k - \Delta t_k) \geq p_k C, \tag{1}$$

where $p_k \in \{-1, +1\}$ is the polarity and Δt_k the elapsed time. Over a time window, the sensor outputs an event stream $\mathcal{E}(t_N) = \{(\mathbf{u}_k, t_k, p_k)\}_{k=1}^{N}$, where each event consists of a pixel location, timestamp, and polarity, and the data is captured with microsecond resolution.

B. Event Representations

To leverage the spatiotemporal structure of event-based data, we describe state-of-the-art representations that convert asynchronous events into dense grid-like structures compatible with deep learning pipelines.

[1] L3S Research Center, Leibniz Universität Hannover, Hanover, Germany i.bugueno@ieee.org

[2] Event-Driven Perception for Robotics, Istituto Italiano di Tecnologia, Genova, Italy luna.gava@iit.it

[3] Robotics and Intelligent Systems Laboratory, Institute of Engineering Sciences, Universidad de O'Higgins, Rancagua, Chile

[4] Advanced Mining Technology Center and Department of Electrical Engineering, University of Chile, Santiago, Chile

1) Accumulative Polarity Frames: Given a sequence of events $\mathcal{E}(t_N)$ within a time window, this representation computes a dense frame $F \in \mathbb{R}^{H\times W}$ where each pixel accumulates polarities:

$$F_{\text{Acc. Events}}(x,y) = \sum_{k=1}^{N} \delta_{x_k,x}\delta_{y_k,y}(2p_k - 1), \tag{2}$$

with $\delta_{i,j}$ the Kronecker delta and $p_k \in \{0,1\}$. Values are quantized into $\{-1,0,+1\}$.

2) Surface of Active Events: SAE encodes the most recent timestamp of an event at each pixel location. Formally, for a sequence $\mathcal{E} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}$, the SAE is defined as

$$F_{\text{SAE}}(x,y) = \max\{\, t_k \mid (x_k,y_k) = (x,y)\,\}. \tag{3}$$

This representation highlights the temporal recency of activity, providing a continuous map of motion dynamics over the sensor plane.

3) Binary Event History Image: BEHI [20] highlights whether any event occurred at a pixel location during the whole interval:

$$F_{\text{BEHI}}(x,y) = \left(\sum_{k=1}^{N}[x_i = x, y_i = y, t_i < T]\right) > 0. \tag{4}$$

4) Tencode: Tencode [21] is an RGB encoding representation that maps event timestamps and polarity into image channels. Given $t_{\min}$ and $t_{\max}$ as the time bounds, and polarity $p \in \{0,1\}$, the RGB value at each pixel is defined as:

$$F_{\text{Tencode}}(x,y) = \begin{cases} (0, \frac{255(t_{\max}-t)}{\Delta t}, 255), & \text{if } p = 1, \\ (255, \frac{255(t_{\max}-t)}{\Delta t}, 0), & \text{if } p = 0, \end{cases} \tag{5}$$

where $\Delta t = t_{\max} - t_{\min}$.

5) Exponential Reduced Ordinal Surface: EROS [22] encodes event activity as a grey-level surface updated event-by-event. For an event $e_k = (v_x, v_y, t_k, p_k)$, the surface $E(x,y,t)$ is updated over a square neighborhood Ω_k of radius k_{EROS} centered at (v_x, v_y) by

$$E(x,y,t_k^+) = \begin{cases} E(x,y,t_k^-)\cdot d, & (x,y) \in \Omega_k,\ (x,y) \neq (v_x,v_y), \\ 255, & (x,y) = (v_x,v_y), \\ E(x,y,t_k^-), & (x,y) \notin \Omega_k, \end{cases} \tag{6}$$

where the decay factor is defined as

$$d = p^{\,1/k_{\text{EROS}}}, \qquad 0 < p < 1. \tag{7}$$

This update assigns the maximum intensity to the incoming event pixel while exponentially decaying the surrounding neighborhood, producing a dense representation of local spatiotemporal activity. Unlike other event representations (e.g.: SAE), the EROS representation does not decay after a fixed temporal window. Instead, it decays based on the occurrence of new events. This event-driven decay mechanism makes EROS inherently velocity-invariant. As a result, both the camera motion and independently moving objects produce consistent representations, enabling reliable feature extraction for navigation tasks.

III. The Dataset

A. Sensors & Setup

The dataset was recorded with three sensors: one depth sensor and two neuromorphic cameras. Table I lists the modalities, frequency, and resolutions.

TABLE I
SENSORS USED IN THE DATASET: MODALITIES, FREQUENCY, AND RESOLUTIONS.

Sensor	Data	Frequency	Resolution
Intel RealSense D435i	RGB frame	30 Hz	1920x1080
	Depth frame	30 Hz	1280x720
DAVIS346	Gray frame	Variable	346x260
	Events	1 MHz	346x260
	IMU	8 kHz	N/A
Prophesee EVK4	Events	1 MHz	1280x720

The sensors were mounted on Bender [23], a general-purpose social designed for human–robot interaction (Figure 2). The Intel RealSense D435i, DAVIS346, and Prophesee EVK4 were rigidly attached at head height to approximate the robot's visual perspective. Extrinsic calibration was obtained from physical offsets and refined using homography-based visual alignment.

B. Dataset collection

Data were collected teleoperating the Bender robot [23] in indoor environments relevant to mobile social robots, including corridors, classrooms, and offices with different lighting conditions and human activity. Each sequence contains synchronized RGB, depth, and event streams, along with derived event representations: Accumulative Polarity Frames, BEHI, SAE, Tencode, and EROS. Examples are shown in Figure 1.

TABLE II
DATASET SUMMARY FOR INTEL REALSENSE D345I, DAVIS346 AND PROPHESEE EVK4.

Data	Total
# Rosbag Sequences	50
Intel RealSense D345i RGB images	15,98K
Intel RealSense D345i Depth images	15,98K
DAVIS346 frames	15,98K
DAVIS346 events	282,88M
Prophesee EVK4 events	1,414B
DAVIS346 IMU samples	319,68K

IV. Baseline for event-based monocular depth estimation

A. Method

The Cycle Generative Adversarial Network (CycleGAN) is adapted to learn a bidirectional mapping between event-based representations $\mathcal{H}$ and monocular depth images $\mathcal{Z}$. This approach uses two generators, $G_{H\to Z}$ and $G_{Z\to H}$, along with discriminators D_Z and D_H. The adversarial

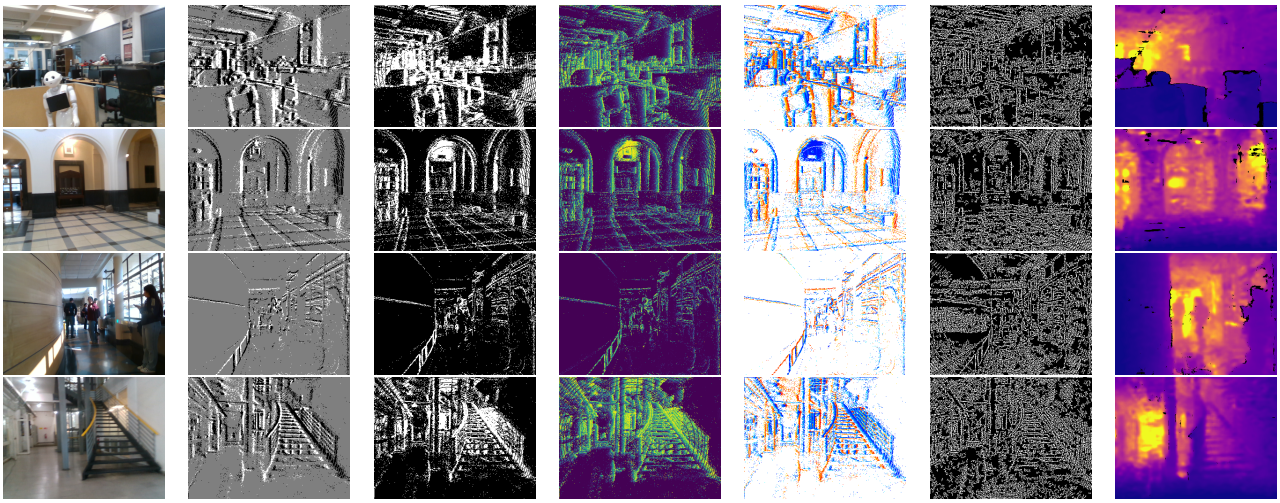
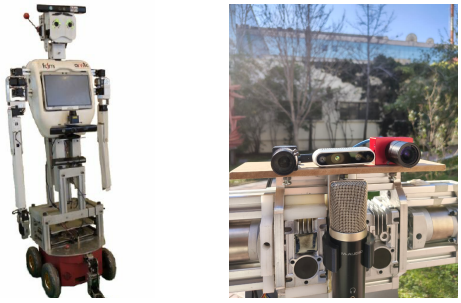


Fig. 1. Example of the dataset: RGB frames, Accumulative Polarity Events, BEHI, SAE, Tencode, EROS, and depth ground truth samples.



(a) (b)

Fig. 2. Robotic platform and sensor setup used. From left to right: Prophesee EVK4 event camera, Intel RealSense D435i, and DAVIS346.

objective enforces realism in each target domain. Meanwhile, cycle-consistency and identity losses preserve geometric and structural information for robust depth estimation in indoor navigation scenarios.

Let $\mathcal{H}$ denote the domain of event-based representations, obtained from asynchronous event streams $\{(t_j, x_j, y_j, p_j)\}$, and $\mathcal{Z}$ the domain of depth images. The generators are defined as $G_{H\to Z} : \mathcal{H} \to \mathcal{Z}, \quad G_{Z\to H} : \mathcal{Z} \to \mathcal{H}$. The adversarial loss for $G_{H\to Z}$ and its discriminator D_Z follows the least-squares GAN formulation:

$$\mathcal{L}_{\text{GAN}}(G_{H\to Z}, D_Z, \mathcal{H}, \mathcal{Z}) = \mathbb{E}_{z\sim p_{\mathcal{Z}}(z)}\big[(D_Z(z)-1)^2\big] + \mathbb{E}_{h\sim p_{\mathcal{H}}(h)}\big[(D_Z(G_{H\to Z}(h)))^2\big]. \quad (8)$$

Similarly, the adversarial loss for $G_{Z\to H}$ with discriminator D_H is given by $\mathcal{L}_{\text{GAN}}(G_{Z\to H}, D_H, \mathcal{Z}, \mathcal{H})$. To ensure bijective mapping between event-based representations and depth images, a cycle-consistency loss is introduced:

$$\mathcal{L}_{\text{cyc}}(G_{H\to Z}, G_{Z\to H}) = \mathbb{E}_{h\sim p_{\mathcal{H}}(h)}\big[\|G_{Z\to H}(G_{H\to Z}(h))-h\|_1\big] + \mathbb{E}_{z\sim p_{\mathcal{Z}}(z)}\big[\|G_{H\to Z}(G_{Z\to H}(z))-z\|_1\big]. \quad (9)$$

In addition, an identity loss encourages generators to preserve structure when the input already belongs to the target domain:

$$\mathcal{L}_{\text{id}}(G_{H\to Z}, G_{Z\to H}) = \mathbb{E}_{z\sim p_{\mathcal{Z}}(z)}\big[\|G_{H\to Z}(z)-z\|_1\big] + \mathbb{E}_{h\sim p_{\mathcal{H}}(h)}\big[\|G_{Z\to H}(h)-h\|_1\big]. \quad (10)$$

The overall objective combines the adversarial, cycle-consistency, and identity losses with weighting factors:

$$\mathcal{L}(G_{H\to Z}, G_{Z\to H}, D_Z, D_H) = \mathcal{L}_{\text{GAN}}(G_{H\to Z}, D_Z, \mathcal{H}, \mathcal{Z}) + \mathcal{L}_{\text{GAN}}(G_{Z\to H}, D_H, \mathcal{Z}, \mathcal{H}) + \lambda_{\text{cyc}}\,\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}}, \quad (11)$$

where λ_{cyc} and λ_{id} are hyperparameters that control the contribution of the cycle-consistency and identity terms, respectively.

B. Training details

The CycleGAN baseline was trained for 100 epochs with a batch size of 8 and the Adam optimizer. Table III summarizes the most relevant hyperparameters.

TABLE III
TRAINING HYPERPARAMETERS AND LOSS CONFIGURATION.

Parameter	Value
Epochs	100
Batch size	8
Learning rate (G)	2×10^{-4}
Learning rate (D)	2×10^{-4}
β_1	0.5
β_2	0.999

C. Metrics

To evaluate depth reconstruction, we use a set of evaluation metrics. Pixel-wise accuracy is measured using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Perceptual fidelity is

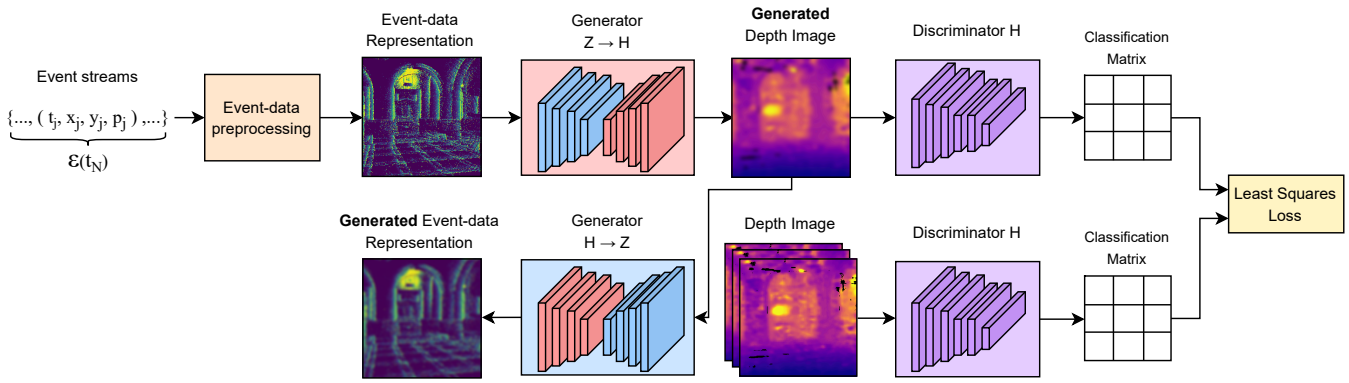


Fig. 3. Suggested CycleGAN architecture for event-based monocular depth estimation baseline. Event streams are preprocessed into dense representations, which are then translated into depth images and back through paired generators ($H \to Z$, $Z \to H$). Discriminators enforce realism in each domain using a least-squares loss, while cycle consistency enables reconstruction across event and depth domains.

quantified through the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM), which capture global contrast and structural consistency. Normalized Cross-Correlation (NCC) further evaluates alignment between reconstructed and ground-truth depth distributions. Finally, accuracy thresholds $\delta_{1.25}$, $\delta_{1.25^2}$, and $\delta_{1.25^3}$ indicate the proportion of pixels where the estimated depth is within increasing multiplicative bounds of the reference.

V. Results & Brief Discussion

Table IV and V summarize the validation performance of CycleGAN trained with different event representations.

TABLE IV
PIXEL-WISE AND PERCEPTUAL METRICS AT EPOCH 100. THE BEST PERFORMANCES CORRESPOND TO LOWER VALUES OF MSE, RMSE, MAE AND HIGHER VALUES OF PSNR AND SSIM.

Modality	MSE ↓	RMSE ↓	MAE ↓	PSNR ↑	SSIM ↑
RGB Frame	0.1023	0.2835	0.2269	11.89	0.500
Acc. Events	0.0930	0.2802	0.2279	11.73	0.470
BEHI	0.0964	0.2902	0.2378	11.27	0.464
SAE	**0.0856**	**0.2675**	**0.2155**	**12.14**	**0.503**
Tencode	0.0894	0.2779	0.2217	11.69	0.485
EROS	0.1052	0.2974	0.2457	11.23	0.501

TABLE V
GEOMETRIC RELIABILITY METRICS AT EPOCH 100. THE BEST PERFORMANCES CORRESPOND TO HIGHER VALUES OF NCC AND δ THRESHOLDS.

Modality	NNC	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
RGB Frame	0.333	0.260	0.457	0.608
Acc. Events	**0.364**	0.254	0.456	0.607
BEHI	0.249	0.228	0.430	0.594
SAE	0.342	0.255	0.460	0.606
Tencode	0.312	**0.262**	**0.463**	**0.610**
EROS	0.298	0.239	0.427	0.571

SAE achieved the best reconstruction error (lower MSE, RMSE, MAE) and the highest PSNR, indicating superior pixel-wise accuracy. Both SAE and EROS showed marginally higher SSIM scores, yielding slightly higher structural consistency. For geometric reliability, the δ-thresholds show that SAE and Tencode outperform other representations, suggesting better alignment with ground-truth depth. Accumulated events performed competitively, surpassing RGB frames in several metrics, which highlights the robustness of event-based inputs under challenging conditions. Overall, these results demonstrate that event representations, particularly SAE and Tencode, provide more reliable depth estimation than conventional frame-based input, establishing a strong baseline for future comparisons on the proposed dataset.

VI. Conclusions & Future Work

This article introduced a new dataset for event-based perception in indoor mobile robotics. It provides synchronized RGB, depth, event streams, and IMU data and includes a CycleGAN baseline that maps event representations to depth with cycle-consistent reconstruction.

Experimental results show that event-based inputs could outperform frame-only inputs, with SAE and Tencode yielding the most reliable accuracy and geometric consistency. These preliminary findings highlight the value of neuromorphic sensing for indoor robotics and provide a reproducible testbed for future methods.

Future work will expand this dataset with longer sequences, more diverse environments, and dynamic human–robot interaction scenarios. Beyond CycleGAN, we plan to design new architectures and to integrate event-based depth estimation into closed-loop navigation and social interaction tasks on real robots. A mid-term objective is to deploy these capabilities on humanoid platforms such as Pepper, enabling socially aware robots with low-latency, robust depth perception for navigation and human–robot interaction in classrooms.

VII. Acknowledgment

This work was partially funded by the FONDEQUIP Project EQM170041, the Basal project AFB230001, and the 2025 IEEE CIS Graduate Student Research Grants. Special thanks to G. Olguin, I. Romero, S. Pizarro, and S. Bugueno for their technical support in data capture.

REFERENCES

[1] K. Youssef, S. Said, S. Alkork, and T. Beyrouthy, "A Survey on Recent Advances in Social Robotics," *Robotics*, vol. 11, no. 4, p. 75, 2022.

[2] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke, "Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 1, pp. 12:1–12:66, Feb. 2023.

[3] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-Based Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.

[4] J. Furmonas, J. Liobe, and V. Barzdenas, "Analytical Review of Event-Based Camera Depth Estimation Methods and Systems," *Sensors*, vol. 22, no. 3, p. 1201, Jan. 2022.

[5] R. Guamán-Rivera, J. Delpiano, and R. Verschae, "Event-Based Optical Flow: Method Categorisation and Review of Techniques That Leverage Deep Learning," *Neurocomputing*, vol. 635, p. 129899, June 2025.

[6] J. Zhang, X. Yu, H. Sier, H. Zhang, and T. Westerlund, "Event-based Sensor Fusion and Application on Odometry: A Survey," in *IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, vol. CFP2540Z-ART, Lyon, France, 2025, pp. 1–6.

[7] I. Bugueno-Cordova, J. Ruiz-del-Solar, and R. Verschae, "Human-Robot Navigation using Event-based Cameras and Reinforcement Learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Nashville, TN, USA, 2025, pp. 5013–5021.

[8] 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 Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, July 2018.

[9] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, July 2021.

[10] L. Burner, A. Mitrokhin, C. Fermüller, and Y. Aloimonos, "EVIMO2: An Event Camera Dataset for Motion Segmentation, Optical Flow, Structure from Motion, and Visual Inertial Odometry in Indoor Scenes with Monocular or Stereo Algorithms, Tech. Rep. arXiv:2205.03467, May 2022.

[11] J. Hidalgo-Carrió, D. Gehrig, and D. Scaramuzza, "Learning Monocular Dense Depth from Events," in *International Conference on 3D Vision (3DV)*, Fukuoka, Japan, Nov. 2020, pp. 534–542.

[12] X. Liu, J. Li, J. Shi, X. Fan, Y. Tian, and D. Zhao, "Event-Based Monocular Depth Estimation With Recurrent Transformers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7417–7429, Aug. 2024.

[13] C. Hu, J. Jiang, Y. Li, M. Sun, and Z. Fang, "EDE-Distill: Boosting Event-Based Monocular Depth Estimation Performance via Knowledge Distillation," *IEEE Robotics and Automation Letters*, vol. 10, no. 8, pp. 8252–8259, Aug. 2025.

[14] L. Lin, K. Li, L. Yang, and T. Li, "EMDepth: Self-Supervised Event-Based Monocular Depth Estimation," in *International Conference on Neuromorphic Computing (ICNC)*, Chongqing, China, 2024, pp. 1–4.

[15] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining Events and Frames Using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2822–2829, Apr. 2021.

[16] A. Devulapally, M. F. F. Khan, S. Advani, and V. Narayanan, "Multi-Modal Fusion of Event and RGB for Monocular Depth Estimation Using a Unified Transformer-based Architecture," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, June 2024, pp. 2081–2089.

[17] T. Pan, Z. Cao, and L. Wang, "SRFNet: Monocular Depth Estimation with Fine-grained Structure via Spatial Reliability-oriented Fusion of Frames and Events," in *IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 2024, pp. 10 695–10 702.

[18] J. Zhu, L. Liu, B. Jiang, F. Wen, H. Zhang, W. Li, and Y. Liu, "Self-Supervised Event-Based Monocular Depth Estimation Using Cross-Modal Consistency," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023, pp. 7704–7710.

[19] S. Jain B and S. Tripathi, "Self-Supervised Event Based Monocular Depth Estimation using Spiking Neural Networks," in *International Conference on Computer Vision and Computational Intelligence (CVCI)*. Hong Kong, China: ACM, Aug. 2025, pp. 57–64.

[20] Z. Wang, F. Cladera, A. Bisulco, D. Lee, C. J. Taylor, K. Daniilidis, M. A. Hsieh, D. D. Lee, and V. Isler, "EV-Catcher: High-Speed Object Catching Using Low-Latency Event-Based Neural Networks," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8737–8744, 2022.

[21] Z. Huang, L. Sun, C. Zhao, S. Li, and S. Su, "EventPoint: Self-Supervised Interest Point Detection and Description for Event-Based Camera," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, Hawaii, 2023, pp. 5396–5405.

[22] L. Gava, M. Monforte, M. Iacono, C. Bartolozzi, and A. Glover, "PUCK: Parallel Surface and Convolution-kernel Tracking for Event-Based Cameras, Tech. Rep. arXiv:2205.07657, May 2022.

[23] J. Ruiz-del-Solar, M. Correa, R. Verschae, F. Bernuy, P. Loncomilla, M. Mascaró, R. Riquelme, and F. Smith, "Bender: A General-Purpose Social Robot with Human-Robot Interaction Capabilities," *ACM Transactions on Human-Robot Interaction*, vol. 1, no. 2, pp. 54–75, 2013.