Depth Estimation from Moving Stereo Event Cameras without Motion Cues

Jiahang Wu, Mikihiro Ikura, Luna Gava, Masayoshi Mizuno, Arren Glover, *Member, IEEE*, Chiara Bartolozzi, *Member, IEEE*

***Abstract*—Depth estimation is highly beneficial for robots performing either navigation or manipulation. Traditional cameras suffer from motion blur in dynamic and high-speed scenarios, to which event cameras are robust while also offering high temporal resolution, low latency, and high dynamic range. However, existing event-based methods require parameter tuning depending on the camera speed and require external measurements of camera motion. In this paper, we present a lightweight framework for real-time depth estimation using stereo event cameras (typically a front-end for SLAM). We propose the use of a velocity invariant event representation to remove parameter tuning due to camera speed, combined with Semi-Global Block Matching for fast depth estimation without requiring camera motion cues or external sensors. We achieve a consistent depth estimation under slow motion (extremely sparse data) and fast motion (motion blur). Our pipeline runs in real-time using only the CPU, with over 100 Hz output on the MVSEC dataset (i.e. $1.6\times$ faster than state-of-the-art), while also achieving a higher (or competitive) accuracy on publicly available datasets.**

***Index Terms*—Event Camera, Stereo Depth Estimation**

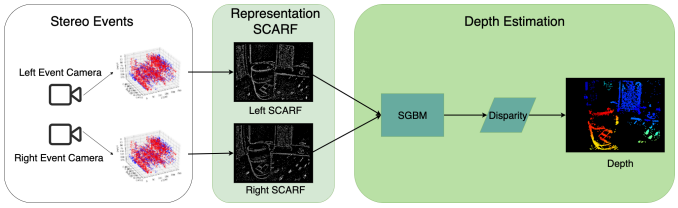

Fig. 1: Overview of the proposed stereo event-based depth estimation pipeline. Asynchronous events from calibrated left and right event cameras are first accumulated using the SCARF representation to generate image-like frames. A Semi-Global Block Matching (SGBM) algorithm is applied to compute disparity maps, which are converted into depth estimations. The proposed pipeline operates fully event-driven, without requiring camera pose or external sensor information, enabling real-time, high-frequency depth perception.

I. INTRODUCTION

Depth estimation is a core task in robotics, enabling applications such as object manipulation, scene understanding, and autonomous navigation. Stereo vision stands out as a power-efficient solution, especially in outdoor environments where active sensors like RGB-D cameras can struggle with sunlight interference and limited depth range. Conventional stereo systems are still limited in high-speed motion and dynamic lighting conditions due to the global gains and fixed shutter periods of frame-based camera technology. Event cameras have emerged as a potential alternative for high-speed and low-latency stereo vision [1].

Event cameras asynchronously detect per-pixel brightness changes, rather than capturing full image frames at a fixed frequency. They generate streams of events with microsecond latency, resulting in reduced power consumption and bandwidth requirements [2], [3]. These properties make event cameras ideal in robotics applications, such as high-speed feature tracking [4]–[6], depth estimation [7]–[9], SLAM [10], [11], and robot control [12].

This paper focuses on high-frequency and real-time depth estimation without motion cues on asynchronous event data as a front-end system, i.e., before a full SLAM pipeline, that includes mapping, camera motion estimation, and integration of those measurements. Our goal is to generate instantaneous and semi-dense depth maps in real-time using only stereo event streams. As illustrated in Fig. 1, we propose a novel real-time pipeline that (i) uses a velocity invariant event representation to enable instantaneous stereo depth estimation with an adaptive number of events, without temporal tuning parameters, and (ii) uses Semi-Global Block Matching (SGBM) to provide a CPU-only semi-dense output at a high output rate independently from the rate of input data.

The proposed pipeline achieves instantaneous semi-dense depth maps without requiring camera motion. Our method combines the strengths of event cameras and stereo matching and offers a lightweight and accurate solution for real-world robotic applications. The lightweight is defined in terms of input requirements, as our method relies solely on pure event data without motion cues, pose information, or additional sensors such as an IMU. Therefore, there is a possibility of a more robust front-end to integrate into SLAM. Our ***contributions*** are:

1) State-of-the-Art accuracy on most evaluated datasets, with higher event throughput and high-frequency depth outputs.
2) Instantaneous stereo matching with event cameras using a velocity-invariant representation, which eliminates the requirement of temporal tuning or external motion estimation – even in the presence of independently moving objects. This enables integration as a front-end to SLAM systems, improving depth accuracy.

Jiahang Wu, Mikihiro Ikura, Luna Gava, Arren Glover, and Chiara Bartolozzi are with Event-Driven Perception for Robotics, Istituto Italiano di Tecnologia, Genova, Italy. Masayoshi Mizuno is with Sony Interactive Entertainment Inc., Japan. Corresponding author: jiahang.wu@iit.it

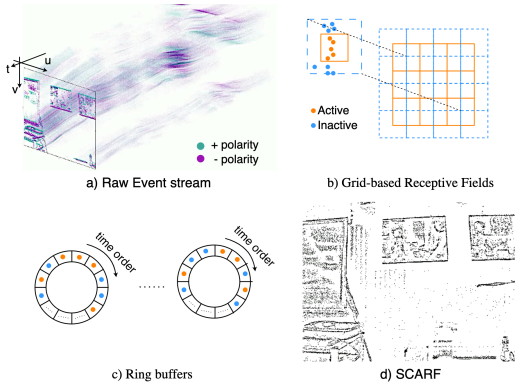

Fig. 2: Overview of SCARF. (a) Raw events are output from an event camera. (b) A grid structure shows active and inactive regions. (c) Each block maintains a ring buffer to store events. (d) SCARF displays all "active" events.

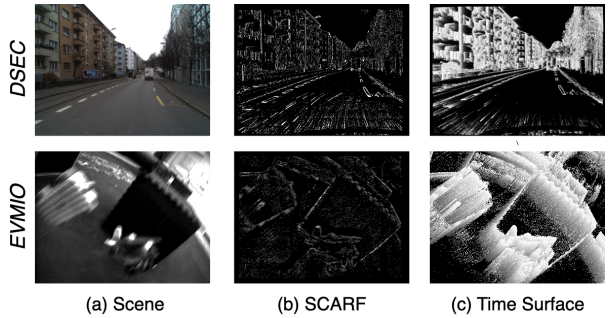

Fig. 3: SCARF vs Time Surface.

3) A CPU-based, semi-dense stereo pipeline using Semi-Global Block Matching (SGBM) over pixel space for real-time operation, independent of event rate. When combined with the SCARF representation, this maximizes event throughput in real-time.
4) Open-source code to enable algorithm benchmarking and comparison.

II. METHOD

Our proposed depth estimation method takes as input raw events from a calibrated stereo camera set and estimates the depth of the scene in a light-weight and high-frequency manner without camera motion information or external sensors. The overall pipeline is shown in Fig. 1. Inspired by the classical stereo matching paradigm [13], we extend it into the event-based domain. By converting asynchronous events into 2D velocity invariant matrix representations, we bound the amount of processing based on the number of pixels of the event sensor, rather than the unknown and highly variable number of events. Fast and high-frequency stereo matching can then be achieved using conventional algorithms. Choosing the appropriate representation allows real-time depth estimation directly from event data, without the need for a temporal time window or motion estimation.

In the pipeline, the event processing module first generates an event representation, called SCARF (Sec. II-A), that is then used for depth estimation. Secondly, the depth estimation module applies Semi-Global Block Matching (SGBM) [13] to estimate disparity over the most up-to-date left and right SCARF representations to estimate depth (see Sec. II-B).

A. Event Representation: SCARF

Event cameras generate asynchronous events. Each event $e = (x, y, t, p)$ is triggered asynchronously by per-pixel brightness changes, where (x, y) denotes the pixel location, t is the timestamp, and $p \in \{-1, +1\}$ is the polarity of the change.

To transform the sparse, asynchronous, and sequential event streams to represent spatial features, we integrate events into the SCARF representation. SCARF addresses the challenge of spatial-temporal event accumulation in scenes containing multiple objects moving at different velocities, including camera motion. Previous works, such as Time Surface [14], accumulate events over a temporal window, which, if not properly tuned to the scene dynamics, can lead to motion blur or missing information. Compared to Time Surface, as illustrated in Fig. 3, SCARF preserves relatively sharp edges with minimal redundant points and reduced motion blur, particularly for the objects with variant motion speed.

SCARF divides the sensor plane into several grid-based receptive fields (Fig. 2). The total number of receptive fields depends on the configured block size b and camera resolution. Each receptive field consists of an active region that spans the entire block in the grid, and an inactive margin that overlaps neighboring blocks. When an event arrives, it may fall within one or more receptive fields. For each receptive field, the event is added with a tag, active or inactive, depending on its location within that field. Each receptive field maintains a ring buffer to store its events. The buffers follow the First-In, First-Out (FIFO) principle, which means that the most recent event replaces the oldest event in the buffer. Only active events contribute to the output intensity of the corresponding image block. Events falling in the inactive border will be removed from active events, thereby "clearing" blocks of irrelevant data (i.e., sending the pixel intensity to 0).

B. Depth Estimation

To estimate depth from stereo event data, the Semi-Global Block Matching (SGBM) algorithm [13] is used on stereo SCARF representations. While SGBM, available in OpenCV, is designed for dense grayscale images, the algorithm still functions with the sparse representations generated by SCARF and offers the best trade-off between speed and robustness in traditional stereo literature.

The SGBM consists of three main components: (1) computation of a local matching cost for each disparity candidate, (2) a semi-global cost aggregation strategy along multiple directions with smoothness constraints, and (3) selection of the optimal disparity. Finally, the disparity map is converted into a depth map by the baseline B and focal length f of

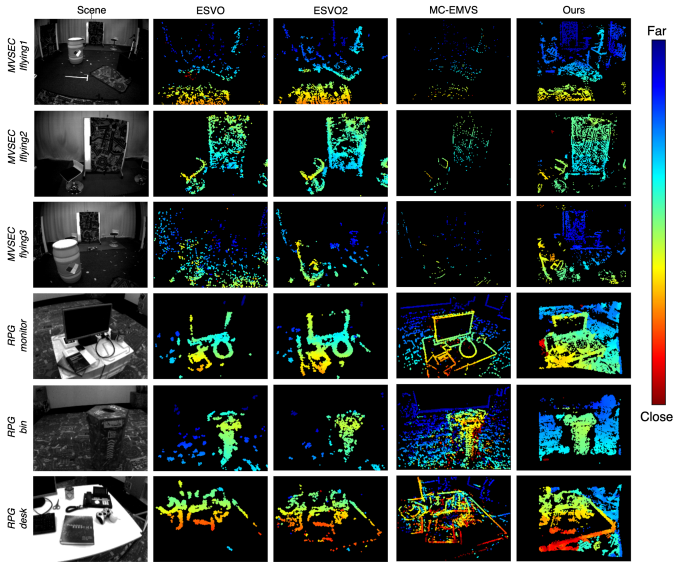

Fig. 4: Depth Estimation. Qualitative comparison of depth estimation results on several sequences using various stereo algorithms. The first column shows intensity images from Event camera (not used, just for visualization). Columns 2 to 5 show depth estimation results of ESVO [15], ESVO2 [11], MC-EMVS [16], and our method, respectively. Depth maps are color coded, from red (close) to blue (far) over a black background, in the range 1–6.25 m for 1-3 rows (*MVSEC* [17]) and the range 0.55–6.25 m for the 6-8 rows (*RPG* [7]).

the stereo camera. To adapt in our problem, we run it on the stereo SCARF and mask the produced depth map so that depth estimates are only given at pixels where events happened.

III. EXPERIMENTS

We report both qualitative and quantitative comparisons against state-of-the-art event-based stereo depth estimation methods, including MC-EMVS [16], ESVO [15], and ESVO2 [11] (Sec. III-A). We further analyze the run-time performance and computational complexity of each approach (Sec. III-B). In addition, Sec. III-C presents an evaluation on a self-collected indoor data from robots. Finally, we conduct an ablation study to investigate the comparison of SCARF v.s. Time Surface representations (Sec. III-D). The open-source code is run in the C++ environment for fair comparison. All methods are executed on an Apple M1 Max chip (10-core CPU) with 32 GB of memory. Our implementation is CPU-only and does not rely on any GPU acceleration.

A. Comparison of Stereo Depth Estimation Methods

Our method is evaluated on data sequences from three public event-based stereo datasets: *MVSEC* [17], *M3ED* [18], and *RPG* [7]. These datasets cover different event-camera types, spatial resolutions, and scene geometries, enabling comprehensive evaluation. Our algorithm works on undistorted and stereo-rectified coordinates, which are precomputed given the camera calibration.

TABLE I: Quantitative evaluation of our proposed method and baselines on MVSEC

Data Sequence	Algorithm	Mean Err [m] ↓	Median Err [m] ↓	Relative Err ↓
MVSEC (upenn_flying1)	ESVO	0.30	0.20	12%
	ESVO2	0.28	0.15	11%
	ESVO2 *Static*	0.35	0.93	13%
	MCEMVS	0.32	0.21	11%
	Ours	**0.26**	**0.14**	**9%**
MVSEC (upenn_flying2)	ESVO	0.49	0.28	30%
	ESVO2	0.35	0.43	16%
	ESVO2 *Static*	0.33	1.19	21%
	MCEMVS	0.33	**0.18**	14%
	Ours	**0.32**	0.22	**11%**

We compare our depth estimation results against three stereo methods and ground truth depth when available. The baseline methods are abbreviated by ESVO [15], ESVO2 [11], and MC-EMVS [16]. For a fair evaluation against our method, we additionally compare with ESVO2 [11], which is abbreviated as ESVO2 *Static*.

1) Qualitative Evaluation: Fig. 4 compares the quality of the depth estimation produced by the above stereo methods. The first column shows grayscale images from the datasets. None of the methods requires intensity information. Columns 2 to 5 show depth estimation results of ESVO, ESVO2, MC-EMVS, and our method, respectively.

Overall, our method produces sharper depth maps close to the scene contours compared to baseline approaches in indoor scenes. Our method reconstructs richer structural details and textures, resulting in denser and more coherent depth maps, whereas MC-EMVS produces significantly sparser reconstructions.

2) Quantitative Evaluation: We evaluate the quantitative performance of our method on multiple benchmark datasets and compare it against four baselines: ESVO, ESVO2, ESVO2 *Static*, and MC-EMVS. We summarize the results on *MVSEC*, using mean error, median error, and relative error. The mean error is defined as $\frac{1}{N}\sum_{k=1}^{N}|d_{\text{est},k} - d_{\text{gt},k}|$, and the relative error is defined as $\frac{1}{N}\sum_{k=1}^{N}\frac{|d_{\text{est},k} - d_{\text{gt},k}|}{d_{\text{gt},k}}$, where N is the number of valid pixels, and $d_{\text{est},k}$ and $d_{\text{gt},k}$ are the estimated depths and ground-truth at pixel index k.

In table I, the best results in each measurement are highlighted in bold. Since there is no ground truth in the RPG dataset, we will not report the quantitative results of the RPG dataset. Our method achieves the smallest errors across all metrics on the *MVSEC* sequences. Besides, our method outperforms ESVO2 *Static* in all evaluation metrics across multiple sequences. This suggests that the proposed depth estimation component can serve as a more accurate alternative in existing SLAM pipelines. Although we do not integrate our method into the full ESVO2 pipeline, the improved accuracy performance indicates its potential as a drop-in replacement.

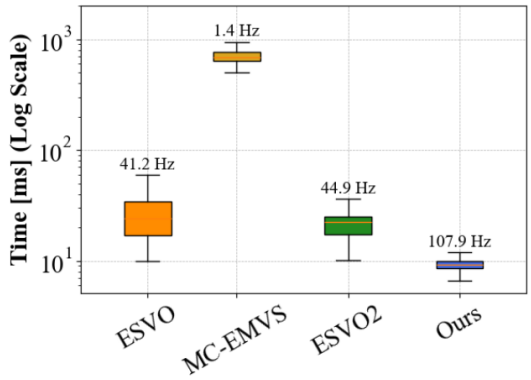

Fig. 5: Runtime analysis on *MVSEC Indoor_flying1*. The mean frequency depth estimation module across each method is shown above the corresponding boxplot.

B. Runtime Analysis

We evaluate the runtime performance on the *MVSEC_indoor_flying1* sequence using all tested methods and report their runtime statistics in Fig. 5. The boxplot shows that our method achieves over 100 Hz, which is the fastest among these four methods.

C. Experiments on Robots

To demonstrate our method's practicality in dynamic indoor environments, we collected our own dataset with stereo event cameras (ATIS Gen3) [19] on a robotic platform, R1 [20]. We manually control the robot to move at different speeds. To compare the output of Time Surface and SCARF, we plot the event rate over time and highlight selected timestamps (Fig. 6). The results show that SCARF consistently preserves clear structural contours regardless of the event rate (on the left and right), while Time Surface only produces clean contours when the event rate is low (on the left). The results demonstrate that our method is robust with respect to both ego-motion and highly dynamic object motion robust in real-world robotic perception tasks.

D. Ablation Study

We conducted an ablation study to evaluate the contribution of the SCARF representation. Specifically, we compared SCARF with Time Surfaces. Experiments are performed on MVSEC and M3ED datasets. Specifically, we replaced SCARF with the Time Surface as input, while keeping the rest of the pipeline fixed (i.e., using SGBM as the stereo matcher). As shown in Table II, our method outperforms the time surface variant across all three datasets. These results demonstrate that SCARF leads to better matching results under indoor conditions, as we witness in the Robot experiment.

IV. CONCLUSION

In this paper, we presented a lightweight and real-time pipeline for stereo depth estimation using event cameras. Unlike previous approaches that rely on camera motion estimation or computationally expensive neural networks, our method achieves high-frequency depth estimation without requiring pose or external sensors. Using a velocity-invariant event representation, SCARF, we transform asynchronous stereo events into an image-like frame with clear scene contours. Then we leverage Semi-Global Block Matching (SGBM) to produce disparity maps and calculate depth. Extensive experiments on public datasets demonstrate that our method achieves state-of-the-art performance on indoor sequences while maintaining high runtime efficiency. The proposed pipeline can run at over 100 Hz on a standard CPU.

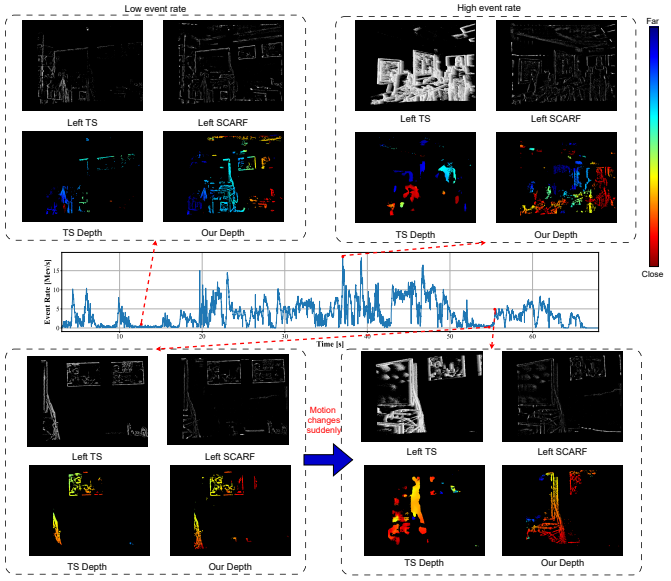

Fig. 6: Event rate and depth comparison between SCARF and TS on our datasets. Event rate over time (middle) on the R1 dataset with selected timestamps. The color of all depth maps follows the color bar.

TABLE II: Ablation study comparing the proposed SCARF representation with the Time Surface (TS) under the same SGBM depth estimation pipeline.

Data Sequence	Algorithm	Mean Err [m] ↓	Median Err [m] ↓	Relative Err ↓
MVSEC (upenn_flying1)	TS+SGBM	0.28	0.18	12%
	Our	**0.26**	**0.14**	**9%**
MVSEC (upenn_flying2)	TS+SGBM	**0.24**	**0.11**	12%
	Ours	0.32	0.22	**11%**
M3ED (spot_indoor_loop)	TS+SGBM	0.68	0.43	16%
	Our	**0.37**	**0.10**	**9%**
M3ED (falcon_indoor_flight1)	TS+SGBM	0.48	0.29	11%
	Our	**0.37**	**0.29**	**8%**

REFERENCES

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[3] B. Chakravarthi, A. A. Verma, K. Daniilidis, C. Fermuller, and Y. Yang, "Recent event camera innovations: A survey," *arXiv preprint arXiv:2408.13627*, 2024.

[4] A. Glover, L. Gava, Z. Li, and C. Bartolozzi, "Edopt: Event-camera 6-dof dynamic object pose tracking," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 18 200–18 206.

[5] M. Ikura, C. Le Gentil, M. G. Müller, F. Schuler, A. Yamashita, and W. Stürzl, "RATE: Real-time asynchronous feature tracking with event cameras," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 662–11 669.

[6] Y.-F. Zuo, W. Xu, X. Wang, Y. Wang, and L. Kneip, "Cross-modal semidense 6-dof tracking of an event camera in challenging conditions," *IEEE Transactions on Robotics*, vol. 40, pp. 1600–1616, 2024.

[7] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, "Semi-dense 3d reconstruction with a stereo event camera," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 235–251.

[8] M. Firouzi and J. Conradt, "Asynchronous event-based cooperative stereo matching using neuromorphic silicon retinas," *Neural Processing Letters*, vol. 43, pp. 311–326, 2016.

[9] S. Ghosh and G. Gallego, "Event-based stereo depth estimation: A survey," *arXiv preprint arXiv:2409.17680*, 2024.

[10] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[11] J. Niu, S. Zhong, X. Lu, S. Shen, G. Gallego, and Y. Zhou, "ESVO2: Direct visual-inertial odometry with stereo event cameras," *IEEE Transactions on Robotics*, 2025.

[12] M. Monforte, L. Gava, M. Iacono, A. Glover, and C. Bartolozzi, "Fast trajectory end-point prediction with event cameras for reactive robot control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4036–4044.

[13] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[14] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "Hots: a hierarchy of event-based time-surfaces for pattern recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1346–1359, 2016.

[15] Y. Zhou, G. Gallego, and S. Shen, "Event-based stereo visual odometry," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1433–1450, 2021.

[16] S. Ghosh and G. Gallego, "Multi-event-camera depth estimation and outlier rejection by refocused events fusion," *Advanced Intelligent Systems*, vol. 4, no. 12, p. 2200221, 2022.

[17] 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, 2018.

[18] 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.

[19] O. Holešovskỳ, R. Škoviera, V. Hlaváč, and R. Vitek, "Experimental comparison between event and global shutter cameras," *Sensors*, vol. 21, no. 4, p. 1137, 2021.

[20] A. Parmiggiani, L. Fiorio, A. Scalzo, A. V. Sureshbabu, M. Randazzo, M. Maggiali, U. Pattacini, H. Lehmann, V. Tikhanoff, D. Domenichelli *et al.*, "The design and validation of the r1 personal humanoid," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 674–680.