SCARF: A Set of Centre Active Receptive Fields for Velocity Invariant Event Representation

Luna Gava[1], Mikihiro Ikura[1], Arren Glover[1], Chiara Bartolozzi[1]

***Abstract*— Event-cameras promise low-latency and high temporal resolution perception for various computer vision tasks, especially for resources constrained, highly dynamic scenarios such as robotics. The novel sensor circuitry (i.e. asynchronous, independent pixels) that enables these advantages also introduces new challenges for algorithm development. The temporal synchronisation of a global shutter can no-longer be relied upon, and instead temporal association of individually timestamped events becomes a non-trivial problem that must be addressed, especially when various parts of the scene move with different speeds. The proposed Set of Centre Active Receptive Fields (SCARF) attempts to solve temporal association by maintaining an active set of events that represents the contrast changes across the scene for every precise moment in time - integrating information both spatially and temporally - and doing so while also inherently accounting for variation in velocity across the image plane. There are no temporal parameters that need to be tuned to the motion present in a dataset. Experiments in this paper demonstrate SCARF produces a representation more similar to Sobel filters (i.e. a representation of intensity change similar to the data produced by event-cameras) than other velocity variant representations, while also achieving the lowest computational cost. The output can be sampled at sub-millisecond resolution as an edge image or as a set of sparse events, during fast motion, in real-time, and without motion-blur.**

I. INTRODUCTION

Computer vision is widely used across smart devices - including mobile phones, smart eyewear, robotics, and home security. Event-cameras have the potential to reduce power and computational requirements, while increasing accuracy in adverse conditions (fast motion and low-light) for many of these applications. A wide variety of algorithms have been explored [1] taking advantage of the low-latency, high-frequency, compressed signal with a high dynamic range for both dynamic (e.g. robotics) and static (e.g. surveillance) applications. Hardware and datasets (e.g. [2], [3], [4]) are becoming more readily accessible.

Event-cameras encode visual information in an asynchronous pixel-wise event-stream instead of an image frame. The event-stream contains all the information of a typical greyscale image but with a higher temporal resolution; i.e. the greyscale frame can be recreated by integrating data over time and interpolating intensity values [5]. However, the data is of a strictly different format and cannot be directly fed into current state-of-the-art vision algorithms. A key problem that must be addressed in novel event-driven algorithms is how to correctly and appropriately perform the temporal integration of data.

[1]All authors are with the Istituto Italiano di Tecnologia, Italy. {first.last@iit.it}

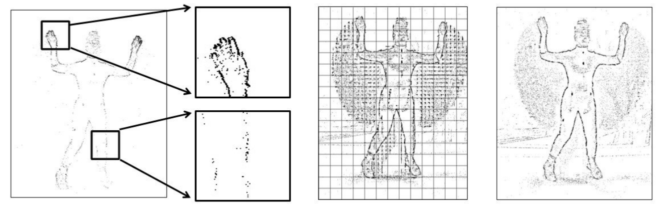


Fig. 1: To achieve a consistent representation of objects moving at different speeds, (a) a single temporal integration period cannot be used. One solution is to (b) keep local memory of a fixed amount of recent events, which results in artifacts and noise being “left behind” in regions which no longer contain texture. (c) SCARF solves this problem by implementing event removal mechanisms (not dependent on time), and therefore allow infinite memory of static objects, and consistent representation during speed variation.

Various methods exist for integrating event data, each with trade-offs. Fixed time windows work for velocity estimation [6] but fail for object recognition at varying speeds - slow motion underexposes, while fast motion reintroduces blur. Fixed event counts achieve velocity invariance for single objects, but struggle with multiple motions or motions with varying speeds. These strategies are used for both 2D image formation (e.g. [7], [8], [9]), as well as spatio-temporal volumes for deep learning pipelines (e.g. [10], [11], [12]).

Motion compensation [8] effectively removes motion-blur, however the initial spatio-temporal volume must contain enough motion for the algorithm to converge, but not so much as to invalidate an assumption of linear motion. The requirement to tune a temporal parameter detracts from it’s adaptability. Moreover, the heavy processing required for multiple objects [13] and non-linear motion [14] using motion compensation results in best use for off-line applications.

In contrast, velocity-invariant representations (e.g. [15], [16], [17], [18], [19], [20], [21], [22]) integrate events robustly under unknown and variable motions. They have been applied to feature detection [15], [17], [20], object detection [22] and human pose estimation [4].

The first velocity invariant representation for events camera was the Speed-Invariant Time Surface (SITS) [15] and augmentations of SITS have been proposed [16], [17], [22], as well as alternative algorithms [18], [20] and data transforms learned from data [19], [21]. Algorithms currently available have the downsides that they are either too slow for modern cameras, produce an output no longer grounded

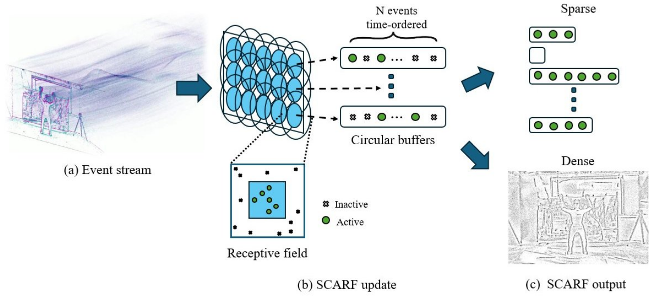


Fig. 2: The SCARF pipeline: (a) event stream - asynchronous pixels with microsecond temporal resolution, (b) SCARF update - the set of overlapping receptive fields with active regions (blue) covering the entire sensor plane; circular buffers store events in active (blue) and inactive (white) regions in temporal order, (c) SCARF output - the list of active events (sparse) or the visualisation of active events on the sensor plane (dense).

in scene luminance, do not maintain a sparse event-list, or the code is just not available. In addition, to the author's knowledge very few direct comparisons have been made of outputs of velocity invariance representations.

In this paper we introduce the **S**et of **C**entre **A**ctive **R**eceptive **F**ields (SCARF) as a velocity invariant representation. The data structure is composed of overlapping receptive fields (RF) across the sensor plane. Each RF stores a fixed number of events, split into active (covering the retina) and inactive regions to enable event removal without temporal decay. SCARF supports both sparse event queries and dense "image-like" outputs.

Unlike prior methods, it preserves contrast magnitude and requires only a few lightweight updates per event (copying it into a small amount, e.g. 4, of circular buffers), making it efficient and grounded in real scene dynamics.

The contributions of this work are:

- Introduction of the Set of Centre Active Receptive Fields (SCARF) representation that maintains a list of sparse events, represents scene gradient magnitude, and also improves accuracy and throughput.
- The use of event-cameras with SCARF allows an edge image to be generated both during fast motion and periods of no-motion, in real-time, at sub-millisecond resolution, and without motion-blur, which can be adapted to a variety of downstream vision tasks.
- Introduction of metric, curated data, pipeline, and code for evaluation of velocity invariant representations. To the author's knowledge, this is the first work that quantitatively evaluates directly velocity invariant event-based representations.

II. Related Work

"Event representation" covers various approaches, from encoding data in a 2D/3D space for neural networks [12] to compact feature-based scene descriptions [23], [24]. In this extended abstract, we focus on representations that closely preserve raw event data while integrating spatial-temporal information to achieve velocity invariance.

The first velocity-invariant representation, Speed Invariant Temporal Surface (SITS) [15], avoids timestamp dependency by assigning pixels pseudo-order values. Events decrement neighboring values, forming consistent patterns regardless of motion speed.

Several extensions followed: Chain SAE (CSAE) [16] uses linked lists for efficiency; Threshold Ordinal Surface (TOS) [17] imposes bounds for better corner stability; and Exponentially Reduced Ordinal Surface (EROS) [22] introduces exponential decay to reduce discretisation artifacts, applied in 6-DoF and human pose tracking. The Adaptive Exponential Decay Surface of Active Events (AED-SAE) [20], which adjusts decay using FFT-based bandpass filtering, improving over SITS and TOS in corner detection.

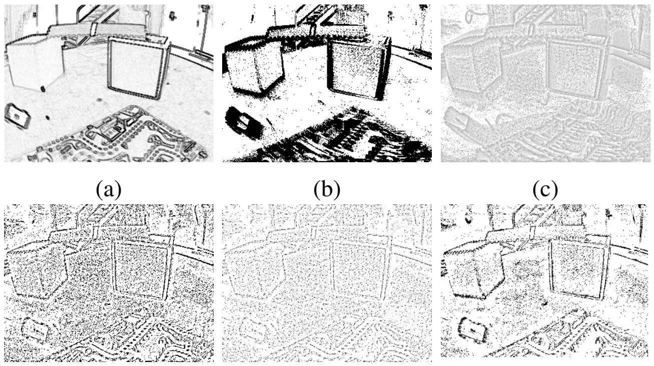
Learning-based methods include event-LSTM [19], which used temporal compression similar to contrast maximization, and Variable Kernel Speed Invariant Surface (VK-SITS) [21], which combines SITS' invariance with data-driven robustness.

Despite advances, the throughput of most methods is below the real-time threshold for modern cameras (often over 20M events/s). Trained models are useful if they improve accuracy, but otherwise require more time and resources. As far as can be ascertained, none of the current state-of-the-art maintain events for sparse computation with precise timing, but instead convert the event stream to an alternative data format.

III. A Set of Centre Active Receptive Fields

An event-camera outputs an asynchronous stream of *events* each defined by its pixel location, polarity (light increase/decrease), and timestamp: $< u, v, p, t >$. Events are triggered when the sensor's pixel detects a change in light beyond a threshold, typically occurring at object edges, or all edges within the scene during camera motion. Images are not extracted from the camera, rather individual pixel locations are asynchronously emitted from the sensor in the precise temporal order at which they occurred. An object moving quickly will produce more events than an identical object moving slowly - yet often both should be represented equally.

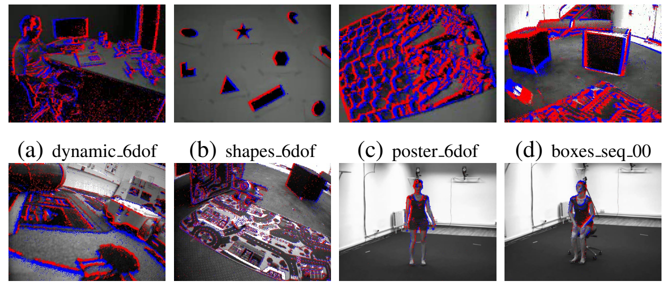
SCARF addresses the challenge of appropriate spatial temporal integration of events in the presence of multiple different velocities from objects, including camera motion itself. Naively accumulating a fixed number of events over the image plane re-introduces motion blur. Objects of various sizes and velocities require different numbers of events and therefore independent memory, see Fig. 1a. It is only possible to define a region of interest (ROI) around each object, within which a fixed number of events is stored, if its location and texture is known *a-priori*, see Fig. 1b. Making some assumptions about the scene, such *a-priori* information could be replaced by a uniform grid, as long as objects are of appropriate size such that no two independent objects of interest fit entirely within a single grid square. However, the



(a) (b) (c)

(d) (e) (f)

Fig. 3: (a) Sobel and different event-based representations: (b) Temporal Window (TW) (c) SITS (d) TOS (e) EROS (f) SCARF. Tested on the boxes_seq_00 from the evaluation sets of *EVIMO* [25].



(a) dynamic_6dof (b) shapes_6dof (c) poster_6dof (d) boxes_seq_00

(e) table-top_seq_00 (f) floor_seq_00 (g) cam2_s1_Dir (h) cam2_s1_Sit_1

Fig. 4: RGB images and events from selected sequences of the datasets experimented generated by (a),(b),(c) [26] DAVIS240 (d),(e),(f) [25] DAVIS 346C (g),(h) [27] input to RGB video to event-based simulator.

problem of events being "left behind" in grid squares when an object moves must still be solved, see Fig. 1c.

A. Centre-Active Receptive Fields

SCARF uses a grid of receptive fields (RF). each storing a fixed number N of events. To remove outdated events as objects move, each receptive field is divided into *active* and *inactive* regions. Only active-region events contribute to the output.

A single circular buffer stores events for both regions. As new events enter, older events are overwritten. As an object completely passes through a single RF the events that begin to fall in the inactive region erase the events in the active region. As such, infinitely persistent local memory is achieved while simultaneously able to represent RF with zero active events (i.e. no edges present).

The sensor is tiled with RFs whose active regions exclusively cover the image, while inactive regions overlap with neighbours by a ratio r (Fig. 2).

B. Implementation

SCARF updates with each incoming event, and supports asynchronous querying at any temporal resolution.

1) Initialisation: The sensor plane is divided into receptive fields (RFs), each with an active region, covering $k \times k$ pixels and then extended by the inactive region that overlaps neighbouring active RFs, defined by the ratio r. Therefore, each sensor pixel maps to one active RF and several inactive ones via precomputed connection maps A (active) and I (inactive). For each event at $< u, v >$, connection maps determine which circular buffers to update. Events added via A are tagged as active ($a = 1$); those via I as inactive ($a = 0$). This update is lightweight and involves only a copy of 32 bits of data, typically into 4 buffers per event, and independent of kernel size k, unlike other methods.

Edge images can be constructed incrementally: when an active event is added, the image is incremented by a constant c; when overwritten, it is decremented. This enables real-time, blur-free visual outputs with minimal overhead.

2) SCARF output: SCARF does not inherently have an output, instead it stores data which can be queried as needed by downstream and/or parallel running processes. The data SCARF stores is contained in the circular buffers of each receptive field. However, the circular buffers store both active and in-active labelled events, and it is only the active events that should be considered as relevant at the time of the query (hence they are "active").

Active events can be quickly extracted from each receptive field and used:

- **Sparse**: extract raw active events and iterate over them for algorithms, e.g. line detection, Hough transform.
- **Dense**: extract the optional Sobel-like image generated on-the-fly in parallel, for use in image-based pipelines, e.g. CNNs.

IV. EXPERIMENTS AND RESULTS

Almost no experiments have been performed to directly compare the output of event representations. We propose to compare velocity invariant event representations against greyscale images passed through a Sobel filter. While none (including SCARF) of the representation claim to reproduce a Sobel filter, it can be used as a ground truth as:

- Event cameras measure change in light levels, and inherently respond to edges in the scene. Sobel filter is a measure of change in light levels between pixels, also used in edge extraction techniques.
- We compare to velocity invariant representations such that any precise timing information is removed in favor of persistence of stationary edges. Greyscale images do not contain timing information, and have persistence of all stationary edges in the scene.

Such a metric imposes the experimental limitations of slow moving cameras, as to not introduce motion blur in the greyscale camera, with pixel-to-pixel matches between the greyscale and event sensors. We tested real [26], [25] and synthetic [27] datasets, shown in Fig. 4, which provide pixel-to-pixel matches of both events and greyscale frames,

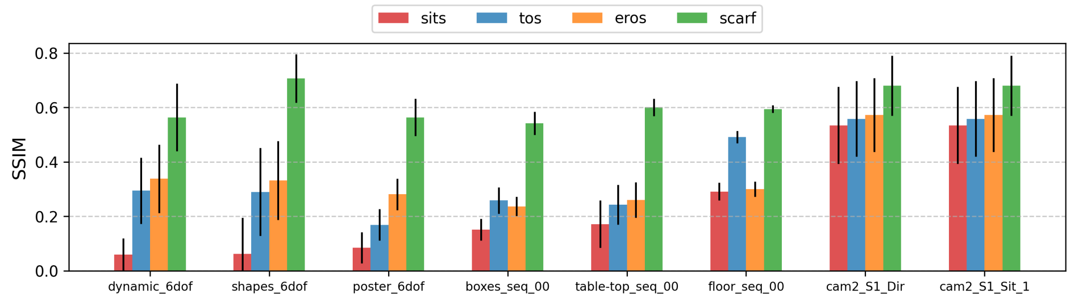


Fig. 5: Mean and standard deviation values over time for eight sequences taken from three different datasets.

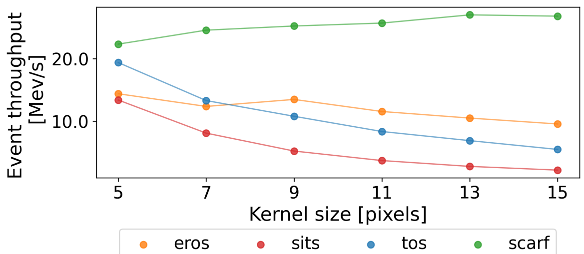


Fig. 6: Comparison in terms of event throughput for different event-based representations.

chosen for their varied motion and scene complexity (e.g., cluttered patterns and lighting variations).

We evaluate SCARF (Fig. 3f) against state-of-the-art event-based representations — SITS (Fig. 3c), TOS (Fig. 3d), and EROS (Fig. 3e) — focusing on accuracy and computational efficiency. A detailed analysis of SCARF parameters (e.g., block size and overlap ratio) is left for future work.

All tests ran on a Dell XPS 15 (i7 @ 5GHz, 32 GB RAM).

A. *Experiment 1: Comparison to Sobel Image*

We assess how well SCARF and other velocity-invariant representations capture edge information under varying scene dynamics, comparing their outputs to Sobel-filtered grayscale images.

Perceptual similarity is measured using SSIM (higher is better), computed only on pixels where event data is present. All outputs are normalized to a [0, 255] range for consistency. Parameters: kernel size = 9 for all methods; SCARF ($\alpha = 2.0$, $c = 0.3$, $r = 1.2$); EROS decay = 0.3.

1) Results: Figure 5 shows SCARF consistently achieves the highest SSIM, followed by EROS and TOS. SITS performs worst due to its design, which rarely clears pixels entirely. The low standard deviation across sequences indicates consistent, significant differences. Stationary-camera sequences (e.g., cam2_S1_Dir) show higher SSIM due to minimal background motion and localized hand movement.

B. *Experiment 2: Event Throughput and Cost*

We evaluate how many events per second each velocity-invariant representation can process, crucial for real-time performance. Sub-VGA cameras typically output 1–2M events/s, while HD sensors can exceed 50M. Throughput was

TABLE I: Computational cost per event and per update for different representations ($k = 9$)

Rep.	**Cost per Event** (μs)	**Cost per Surface Update** (μs)
SCARF	0.070	0.057
SITS	0.217	8.646
TOS	0.096	7.324
EROS	0.084	6.940

measured using the dynamic_6dof dataset, varying only the kernel size k, and averaging results over five trials.

1) Results: Figure 6 shows SCARF achieving the highest event throughput (25 M events / second) and remaining stable across kernel sizes, since its update cost per event is constant. A slight increase at larger k likely comes from reduced buffer overlap near sensor edges. Other methods slow significantly as k increases. SITS is the slowest due to per-event nested loops, making it unsuitable for HD cameras. TOS also degrades with k^2 complexity. EROS performs better due to vectorized updates but still falls below real-time rates for high-res input. In practice, $k = 9$ is the minimum for noise reduction, ruling out SITS and limiting EROS to controlled scenarios.

SCARF showed the highest efficiency, with the lowest per-event (0.07 µs) and surface read-out (0.06 µs) costs among all methods (see Table I).

V. CONCLUSION

The proposed Set of Centre-Active Receptive Fields (SCARF) unlocks the next generation of real-time, high-frequency artificial perception, where scene content is well represented, motion blur is eliminated, temporal resolution is maximized. SCARF maintains asynchronicity of event cameras and sparse access to raw events, while demonstrating the maintained events were more similar to Sobel representations. The state-of-the-art throughput can unlock modern HD event-cameras for real-time operation in a variety of applications including wearable devices and robotics. For the community already using event-cameras, SCARF can boost the performance of algorithms under variations in velocity, in real-time, and allows either an event-based or image-based downstream computation. For the vision community using traditional sensors, SCARF can enable the adoption of event-cameras for a sub-millisecond visual signal, on-demand, and achieved without motion blur.

References

[1] 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.

[2] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, 2021.

[3] 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.

[4] G. Goyal, F. Di Pietro, N. Carissimi, A. Glover, and C. Bartolozzi, "Moveenet: Online high-frequency human pose estimation with an event camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4024–4033.

[5] L. Wang, T.-K. Kim, and K.-J. Yoon, "Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8315–8325.

[6] J. L. Valerdi, C. Bartolozzi, and A. Glover, "Insights into batch selection for event-camera motion estimation," *Sensors*, vol. 23, no. 7, p. 3699, 2023.

[7] V. Vasco, A. Glover, and C. Bartolozzi, "Fast event-based harris corner detection exploiting the advantages of event-driven cameras," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 4144–4149.

[8] S. Shiba, Y. Aoki, and G. Gallego, "Event collapse in contrast maximization frameworks," *Sensors*, vol. 22, no. 14, p. 5190, 2022.

[9] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[10] G. Scarpellini, P. Morerio, and A. Del Bue, "Lifting monocular events to 3d human poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1358–1368.

[11] Z. Chen, J. Wu, J. Hou, L. Li, W. Dong, and G. Shi, "Ecsnet: Spatio-temporal feature learning for event camera," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 701–712, 2022.

[12] R. W. Baldwin, R. Liu, M. Almatrafi, V. Asari, and K. Hirakawa, "Time-ordered recent event (tore) volumes for event cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2519–2532, 2022.

[13] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, "Event-based motion segmentation by motion compensation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7244–7253.

[14] F. Hamann, Z. Wang, I. Asmanis, K. Chaney, G. Gallego, and K. Daniilidis, "Motion-prior contrast maximization for dense continuous-time motion estimation," in *European Conference on Computer Vision*. Springer, 2025, pp. 18–37.

[15] J. Manderscheid, A. Sironi, N. Bourdis, D. Migliore, and V. Lepetit, "Speed invariant time surface for learning to detect corner points with event-based cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 245–10 254.

[16] S. Lin, F. Xu, X. Wang, W. Yang, and L. Yu, "Efficient spatial-temporal normalization of sae representation for event camera," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4265–4272, 2020.

[17] A. Glover, A. Dinale, L. D. S. Rosa, S. Bamford, and C. Bartolozzi, "luvharris: A practical corner detector for event-cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10 087–10 098, 2021.

[18] R. Hu, Y. Xia, and Z. Sun, "A less noisy time surface for event-based visual odometry," in *2021 IEEE International Conference on Unmanned Systems (ICUS)*, 2021, pp. 299–304.

[19] L. Annamalai, V. Ramanathan, and C. S. Thakur, "Event-lstm: An unsupervised and asynchronous learning-based representation for event-based data," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4678–4685, 2022.

[20] J. Li, L. Su, C. Guo, X. Wang, and Q. Hu, "Asynchronous event-based corner detection using adaptive time threshold," *IEEE Sensors Journal*, vol. 23, no. 9, pp. 9512–9522, 2023.

[21] L. Acin, P. Jacob, C. Simon-Chane, and A. Histace, "Vk-sits: a robust time-surface for fast event-based recognition," in *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2023, pp. 1–6.

[22] 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.

[23] 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.

[24] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "Hats: Histograms of averaged time surfaces for robust event-based object classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1731–1740.

[25] A. Mitrokhin, C. Ye, C. Fermüller, Y. Aloimonos, and T. Delbruck, "Ev-imo: Motion segmentation dataset and learning pipeline for event cameras," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6105–6112.

[26] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[27] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.