DailyDVS-200: A Comprehensive Benchmark Dataset for Event-Based Action Recognition

Qi Wang[1*], Zhou Xu[1*], Yuming Lin[1], Jingtao Ye[1], Hongsheng Li[1], Guangming Zhu[1], Syed Afaq Ali Shah[2], Mohammed Bennamoun[3], and Liang Zhang[1✉]

[1] Xidian University, School of Computer Science and Technology, China
{qiwang0720, xuzhou0112}@stu.xidian.edu.cn, liangzhang@xidian.edu.cn
[2] Edith Cowan University afaq.shah@ecu.edu.au
[3] University of Western Australia mohammed.bennamoun@uwa.edu.au

Abstract. Neuromorphic sensors, specifically event cameras, revolutionize visual data acquisition by capturing pixel intensity changes with exceptional dynamic range, minimal latency, and energy efficiency, setting them apart from conventional frame-based cameras. The distinctive capabilities of event cameras have ignited significant interest in the domain of event-based action recognition, recognizing their vast potential for advancement. However, the development in this field is currently slowed by the lack of comprehensive, large-scale datasets, which are critical for developing robust recognition frameworks. To bridge this gap, we introduces ***DailyDVS-200***, a meticulously curated benchmark dataset tailored for the event-based action recognition community. DailyDVS-200 is extensive, covering 200 action categories across real-world scenarios, recorded by 47 participants, and comprises more than 22,000 event sequences. This dataset is designed to reflect a broad spectrum of action types, scene complexities, and data acquisition diversity. Each sequence in the dataset is annotated with 14 attributes, ensuring a detailed characterization of the recorded actions. Moreover, DailyDVS-200 is structured to facilitate a wide range of research paths, offering a solid foundation for both validating existing approaches and inspiring novel methodologies. By setting a new benchmark in the field, we challenge the current limitations of neuromorphic data processing and invite a surge of new approaches in event-based action recognition techniques, which paves the way for future explorations in neuromorphic computing and beyond. The dataset and source code are available at https://github.com/QiWang233/DailyDVS-200.

Keywords: Neuromorphic Sensors · Event-Based Action Recognition · Dynamic Range · Large-Scale Benchmark Dataset

1 Introduction

Action recognition is of significant importance across various domains, spanning from intelligent surveillance to video understanding. Current action recog-

✉ Corresponding author. * Equal contribution.

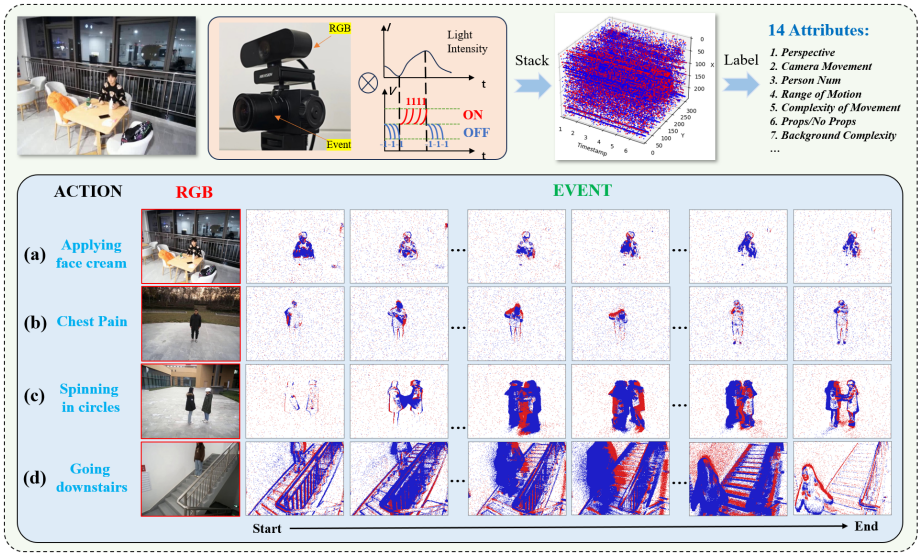
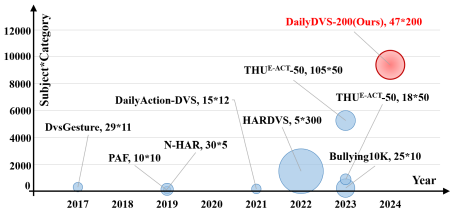


Fig. 1: The flow of our data acquisition process. We use both an RGB camera (above) and a DVS camera (below). Upon completion of the recording, the DVS camera generates event flow data, while the RGB camera captures the synchronized video stream. Subsequently, the data is processed to remove noise, and each sample is categorized based on its motion characteristics.

nition algorithms, leveraging large-scale frame-based benchmark datasets, have demonstrated remarkable performance. However, several challenges arise concerning the storage, transmission, and analysis of frame data. Redundant frames in video data contribute to unnecessary storage requirements and escalate power consumption burdens on devices. Moreover, frame-based cameras suffer from low frame rates and limited dynamic ranges, impeding their capability to effectively capture fast-moving objects and operate optimally under challenging lighting conditions like back-lighting and low light. So, the exploration and implementation of innovative solutions become imperative to address these limitations.

The development of dynamic vision sensors (also known as event cameras), such as DAVIS [6], CeleX [11], ATIS [44], and PROPHESEE[1], has introduced a new paradigm in visual perception. Unlike traditional cameras that capture images at fixed exposure rates, event cameras asynchronously record points in the scene where pixel brightness changes exceed a certain threshold, and output event in the form of tuples (t, x, y, p), where t represents the timestamp, (x, y) represents the two-dimensional coordinates, and p represents the polarity. Additionally, stacking events over a period of time can be viewed as a discrete 3D sequence along the time axis, as illustrated in the first row of Fig. 1. Since event cameras only capture parts of the scene where brightness changes occur, they

[1] `https://www.prophesee.ai`



DailyDVS-200					
Scale	22046	**14 attributes:**			
Class	200	**PR**	Props	**PE**	Perspective
Subject	47	**PO**	Posture	**DI**	Diurnality
Sensors	DVXplorer Lite	**DR**	Duration	**LO**	Location
Resolution	320×240	**AR**	Action Range	**DT**	Distance
Duration	1-20s	**PN**	Person Num	**HE**	Height
EventCount	18 billion	**CM**	Camera Motion	**SH**	Shadow
EventTime	24 hour	**IL**	Illumination Direction	**BC**	Background Complexity

Fig. 2: (Left) A comparison between existing datasets and our proposed DailyDVS-200 dataset for event-based action classification. (Right) Summary of Characteristics of DailyDVS-200. The nomenclature is **PR**: Props, **PO**: Posture, **DR**: Duration, **AR**: Action Range, **PN**: Person Num, **CM**: Camera Motion, **IL**: Illumination Direction, **PE**: Perspective, **DI**: Diurnality, **LO**: Location, **DT**: Distance, **HE**: Height, **SH**: Shadow, **BC**: Background Complexity.

significantly reduce redundant information and lower the storage and computational load on devices [28, 57, 63, 64]. Moreover, event streams emphasize the approximate outline of objects without recording specific color and texture features like traditional cameras, thus greatly protecting user privacy and eliminating concerns about privacy leakage during the use of related devices in daily life. Event cameras also exhibit high frame rates and a wide dynamic range, leading to densely packed event streams over time with minimal motion blur, enabling them to effectively capture fast-moving human actions. These attributes render them highly suitable for addressing the current challenges in action recognition.

In an ideal scenario, action recognition primarily involves systems analyzing motion characteristics of the human body trunk to predict and classify actions. However, in real-life situations, recognition systems not only capture human motion information but also record redundant information such as environmental backgrounds, as shown in the second row of Fig. 1, action (a)(b)(c) have no significant background information, while (d) prominently features staircase details. Thus, key human motion information is often disrupted by various factors such as camera motion, light intensity, scene complexity, *etc*. These factors affect the integrity of the data and increase the challenges of correct recognition.

Although some benchmark datasets have been proposed, many of them are synthetic datasets captured by pointing event cameras at screens displaying RGB images [42], or by converting commonly used RGB action recognition datasets into simulated event streams [4, 22, 24, 52]. However, simulated datasets often result in information loss due to their multi-stage nature, making it difficult to achieve the effectiveness of real event datasets in practice. While several large-scale real-world benchmark event datasets have been proposed recently [1, 4, 13, 15, 18, 29, 35, 38, 51, 58], such as DvsGesture [1], PAF [38], and Hardvs [58], they are limited by small scale, few categories, and limited diversity of individuals. Additionally, THU$^{\text{E-ACT}}$-50 [18] consists only of simple actions in fixed scenes, and although THU$^{\text{E-ACT}}$-50-CHL [18] introduces challenging ac-

Table 1: Comparison of event datasets for action recognition. Sub, MA, AA and DR denotes Subject, Multi-Attribute, Attribute Annotation and Duration of the action, respectively. Note that we only report these groups of real DVS datasets.

Dataset	Year	Sensors	Object	Scale	Class	Sub	Real	MA	AA	DR
ASLAN-DVS [4]	2011	DAVIS240c	Action	3,697	432	-	✗	-	-	-
MNISTDVS [49]	2013	DAVIS128	Image	30,000	10	-	✗	-	-	-
N-Caltech101 [42]	2015	ATIS	Image	8,709	101	-	✗	-	-	0.3s
N-MNIST [42]	2015	ATIS	Image	70,000	10	-	✗	-	-	0.3s
CIFAR10-DVS [27]	2017	DAVIS128	Image	10,000	10	-	✗	-	-	1.2s
HMDB-DVS [4, 24]	2019	DAVIS240c	Action	6,766	51	-	✗	-	-	19s
UCF-DVS [4, 52]	2019	DAVIS240c	Action	13,320	101	-	✗	-	-	25s
N-ImageNet [21]	2021	Samsung-Gen3	Image	1,781,167	1,000	-	✗	-	-	-
ES-lmageNet [32]	2021	-	Image	1,306,916	1,000	-	✗	-	-	-
DvsGesture [1]	2017	DAVIS128	Action	1,342	11	29	✔	✗	✗	6s
N-CARS [51]	2018	ATIS	Car	24,029	2	-	✔	✗	✗	0.1s
ASL-DVS [4]	2019	DAVIS240	Hand	100,800	24	5	✔	✗	✗	0.1s
PAF [38]	2019	DAVIS346	Action	450	10	10	✔	✗	✗	5s
DailyAction [35]	2021	DAVIS346	Action	1,440	12	15	✔	✗	✗	5s
HARDVS [58]	2022	DAVIS346	Action	107,646	300	5	✔	✔	✗	5s
THU$^{\text{E-ACT}}$-50 [18]	2023	CeleX-V	Action	10,500	50	105	✔	✗	✗	2-5s
THU$^{\text{E--ACT}}$-50-CHL [18]	2023	DAVIS346	Action	2,330	50	18	✔	✗	✗	2-5s
Bullying10K [13]	2023	DAVIS346	Action	10,000	10	25	✔	✗	✗	2-20s
DailyDVS-200 **(Ours)**	2024	DVXplorer Lite	Action	22,046	200	47	✔	✔	✔	1-20s

tions, it is still limited by the number and diversity of actions. Bullying10K [13] focuses exclusively on 10 types of bullying actions, making it difficult to demonstrate the superiority of models. Therefore, there is an urgent need for a comprehensive dataset that takes into account various factors to address these issues.

To bridge these aforementioned gaps, we present a novel and comprehensive dataset, termed DailyDVS-200. Our proposed dataset consists of 200 distinct action categories, collected from 47 subjects. As shown in Fig. 2 (left), the dataset we provide has the richest number of subject-category combinations. The detailed comparison with existing benchmark datasets can be found in Table 1. Our dataset systematically incorporates a wide range of real-life scenarios, including variations in *viewpoint*, *diurnal shifts*, *indoor and outdoor settings*, *lighting conditions*, *camera movements*, *actor counts*, *action duration*, *shadow effects*, *shooting elevations*, *distances*, *prop presence*, *action scopes*, *background complexities*, and *pose diversity*. Moreover, each data is annotated based on these attributes, more details can be found in Fig. 2 (right) and Fig. 4 (a).

Leveraging our newly introduced DailyDVS-200 dataset, we conduct a comprehensive evaluation of four distinct types of action recognition deep models across over ten diverse frameworks. Our findings reveal that the current state-of-the-art event-based action recognition networks continue to underperform when compared to traditional action recognition frameworks. *Additionally, we divided the dataset into 14 groups based on the annotation of attributes.* Through diverse group evaluations, we showed significant variations in the model's recognition performance under different action conditions. For instance, under the Swin-T [36] test, recognition performance significantly differs between dynamic and static camera conditions (27.84% vs. 58.05%), as well as across diverse action

ranges (Full-body: 59.24% vs. Limbs: 43.15% vs. Micro: 34.59%). Additionally, we undertake parallel tests under identical conditions using existing large-scale event datasets to authenticate the challenging nature of our dataset. In summary, our contributions are mainly reflected in the following aspects:

- We propose a large-scale neuromorphic dataset for action recognition, named DailyDVS-200. It consists of over 22k samples collected from 47 subjects, spanning 200 categories, and comprehensively reflects real-world challenges, with corresponding attribute annotations provided for each data point. *To the best of our knowledge, this is the first large-scale real-world neuromorphic dataset for action recognition with label provided.*
- To objectively evaluate the performance of different methods, we establish benchmarks for various types of recognition models on our dataset. This provides a wide baseline for future comparisons on the DailyDVS-200 dataset.
- We conducted group evaluation and analysis based on the annotation of attributes. This not only validates the impact of different attribute on event-based action recognition models but also introduces new research content for neuromorphic datasets, thereby driving advancements in the field.

2 Related Work

2.1 Event-based Dataset

Synthetic Datasets Currently, there are various publicly available event datasets, but early DVS datasets mainly originated from existing image classification datasets [27,42]. They primarily captured changes in brightness using DVS cameras and the relative motion of scenes, which posed significant challenges. Subsequently, Bi *et al*. [4] expanded the number of event datasets by converting publicly available frame-based action datasets into simulated event streams, including HMDB-DVS [24], ASLAN-DVS [22], and UCF-DVS [52]. However, these synthetic datasets somewhat suppressed the characteristics of event cameras, such as high dynamic range and high frame rate, by initially capturing images with regular cameras and then manually converting them into event streams.

Real-World Dataset for Event-Based Action Recognition In recent years, small-sized real-world event-based action datasets have emerged [1, 35, 38], but they have limitations. DvsGesture [1] has only 11 action categories with data from 29 individuals, PAF [38] includes only 10 actions and 450 records, and DailyAction [35] comprises 12 actions with 1,440 records at low resolution (128×128). Larger datasets like THU$^{\text{E-ACT}}$-50 [18] offer more samples but have limitations in term of consideration of challenging scenarios. Hardvs [58] is currently the largest dataset in terms of data volume, but it also has the smallest number of participants, with only 5 individuals, which presents significant limitations. Bullying10k [13] focuses on a single bullying-related scene and lacks diversity for

comprehensive action recognition research. These datasets, despite their merits, do not sufficiently advance event-based action recognition models or fully leverage the potential of event cameras.

2.2 Event Representation

Due to the asynchronous and discrete nature of event streams, they cannot be directly used for training. Therefore, it is necessary to convert them into suitable alternative representations for different action recognition methods. Frame-based representations of event streams are currently widely used because they allow simple generation of event frames by aggregating events within specific time windows [25, 39, 51, 62]. However, such designs primarily extract spatio-temporal information and do not fully exploit the time and polarity information contained in events. With the widespread adoption of Transformer networks in various event-based tasks, there has been a recent focus on converting events into suitable token forms [5, 43, 46]. For instance, Peng *et al*. [43] proposed a novel Group Token that reorganizes asynchronous events based on their timestamps and polarities, effectively leveraging the characteristics of events. Some researchers are also exploring new neural network models, such as Spiking Neural Networks (SNNs) [10, 41, 60], which simulate the pulse transmission of neurons to achieve information processing, making them more akin to the working mechanism of biological neural systems. Additionally, there are other representation methods, such as learning-based representations [8,19,37], graph-based representations [30,47] and TORE [2]. However, it remains unclear which representation method is most suitable for event streams at present.

2.3 Action Recognition

Historically, in frame-based action recognition, there have typically been two key steps: action representation [26,40,48,56] and action classification [23,33,50]. In recent years, deep learning techniques have integrated these two steps into an end-to-end learning framework, significantly improving action classification performance. To leverage information from all frames and model the inter-frame information correlation, Tran *et al*. [53] proposed 3DCNN to learn features in both spatial and temporal domains, but with high computational costs. Carreira and Zisserman [9] introduced I3D, which adapts well established image classification architectures, making training easier. Feichtenhofer *et al*. [17] proposed an efficient network, SlowFast, with both slow and fast pathways that can adapt to different scenarios by adjusting channel capacities, greatly enhancing overall efficiency. Additionally, various 3DCNN variants [16,54,65] have been proposed, further improving recognition efficiency. With the introduction of ViT [14], self-attention mechanisms [12,55] have been applied to action recognition [3,36], which has been shown to achieve good performance. Spiking neural networks have also been used for action recognition. However, due to the non-differentiability of discrete pulse signals, training SNNs poses challenges.

Several effective training methods have been proposed to address this challenge [10, 41, 59–61], but their effectiveness remains to be further investigated.

3 DailyDVS-200 Dataset

In this section, we provide a detailed overview of the data acquisition, annotation methodology, and analysis of action categories for our DailyDVS-200 dataset.

3.1 Data Structure

Data Modalities To meet the computation requirements of real life applications, we used a DVXplorer Lite sensor paired with an RGB camera to collect our dataset, details are shown in Fig. 1. We employed custom-built software to capture event data and save it as standard *aedat4* files, which include the event stream, IMU stream, and trigger stream. The spatial resolution of the event camera is 320×240, and the RGB camera is synchronized to ensure the quality of the capture and assist with subsequent annotation tasks.

Subjects To ensure the diversity and authenticity of the data, we selected 47 different subjects (26 males and 21 females) from hundreds of participants based on factors such as gender, physique, and height to participate in our dataset collection work. Each participant was assigned a unique ID number.

Action Classes To enhance the practical applicability of our DailyDVS-200 dataset, we meticulously curated and supplemented 200 daily action categories from commonly used video-based public datasets [7,20,34,52], a selection process that resulted in a dataset closely mirroring real-life scenarios. Finally, our dataset comprises 22,046 records, more detail about our DailyDVS-200 dataset can be seen in Fig. 2 (right) and Fig. 4.

The DailyDVS-200 dataset includes the following scenes: (1) **Household Activities**: *Sweeping*, *mopping*, *combing hair*, *washing towels*, *folding clothes*, *brushing teeth*, *washing face*, *cutting nails*, *etc.* (2) **Office Tasks**: *Nodding*, *clapping*, *sitting down*, *standing up*, *typing on a keyboard*, *moving a mouse*, *opening drawers*, *opening a laptop*, *etc.* (3) **Sports and Physical Activities**: *Jumping in place*, *running*, *dribbling basketball*, *long jump*, *skipping rope*, *push-ups*, *kicking a ball*, *swinging a badminton racket*, *etc.* (4) **Health-related Activities**: *Headache*, *chest pain*, *back pain*, *vomiting*, *leg massage*, *etc.* (5) **Interactions**: *Handshaking*, *toasting*, *hugging*, *high-fiving*, *arm wrestling*, *fist bumping*, *etc.* (6) **Bullying and Violence**: *Fighting*, *hitting*, *kicking*, *pushing*, *using objects to attack*, *etc.* (7) **Transportation-related Activities**: *Riding a bicycle*, *riding an electric scooter*, *walking with a backpack*, *etc.* This diverse range of action categories ensures the relevance of the proposed dataset to various real-world applications and provides a comprehensive basis for event-based action recognition research.

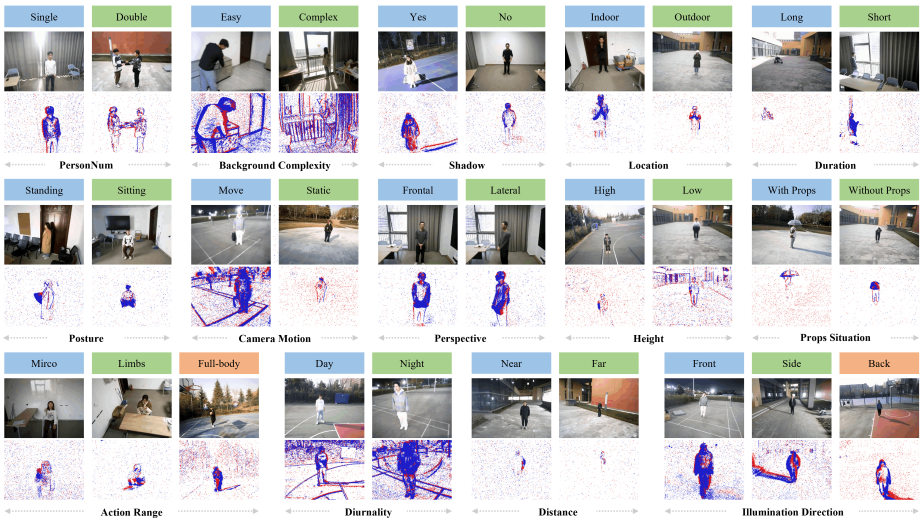


Fig. 3: A preview of our proposed DailyDVS-200 dataset and examples of our attribute annotations.

The dataset consists of a wide range of daily actions, considering various action characteristics: (1) **Fine-grained Micro, Limb, and Whole-body Movements**: Examples include finger movements such as *writing* and *trimming nails*, limb movements such as *rotating one arm* and *checking the time* as well as whole-body movements like *Tai Chi* and *walking with an umbrella*. (2) **Short-duration and Long-duration Actions**: It covers simple actions like *clapping hands* and *nodding* compared to longer actions such as *putting on shoes* and *dressing*. The time distribution of our dataset is illustrated in Fig. 4 (b). It can be observed that we have a diverse range of action duration. (3) **Actions with and without Props Interaction**: It includes actions involving interactions with objects such as *playing volleyball* and *playing table tennis*, as well as actions without object interaction like *going upstairs* and *going downstairs*.

In addition, as shown in Fig. 3, we pay close attention to the diverse performances of the same action under different settings: (1) **Different Perspectives**: We collect different views of the same action to simulate real-world perspectives, including front view, side view, top-down view, and bottom-up view. (2) **Different Distances**: Due to the limitations of the frame, the performance of the same action at different distances often varies significantly. Therefore, we capture actions from distances up to approximately 30 meters. (3) **Lighting Conditions**: i) Light Direction: Different light directions directly affect the performance of actions in the frame due to the event camera's sensitivity to light. Therefore, we collect data under front light, side light, and back light conditions. ii) Day and Night Conditions: Lighting conditions vary between day and night, which directly affect the performance of actions. Additionally, due to the high frame rate of event cameras, the flickering effect of low-frequency lights also has a sig-

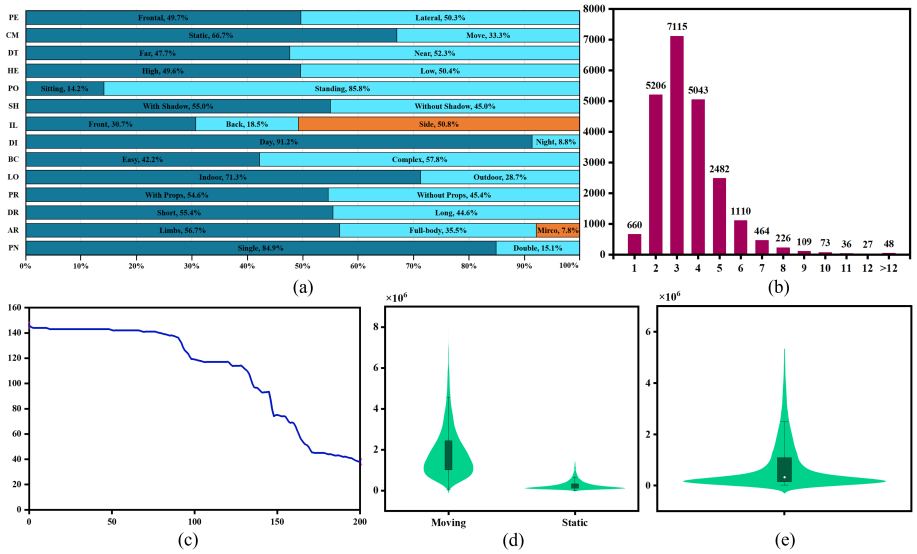


Fig. 4: Statistical data and analysis of DailyDVS-200. (a) Data proportions for the 14 attributes. (b) Distribution of data volumes for different time duration in seconds. (c) Number of images per class. (d) Distribution of Event Count compared between the moving and static. (e) Distribution of Event Count for all categories.

nificant impact. (4) **Different Camera Movements**: Event cameras focus only on the changes in light in the frame, making it easier to identify actions under fixed camera conditions. However, such data biases severely limit the application scenarios of event cameras. Therefore, we capture action data under certain camera movements, including varying degrees of background complexity.

In addition to these factors, which we believe have a significant impact on event data, we also consider other possible scenarios that may occur in real-world settings, including indoor and outdoor settings, sitting and standing postures for the same action, and the influence of shadows.

Collection Setups To replicate various scenarios encountered in daily life realistically, we set different data acquisition conditions for each participant and selected diverse locations such as different rooms, corridors, open squares, roads, and playgrounds. This ensures the richness and diversity of the scenes. Additionally, our data collection sessions are divided into three time periods: *morning*, *afternoon*, and *evening*, which closely reflect different conditions in real life. We have a variety of scenes to ensure the diversity of our dataset. During data collection, we select actions based on the current scene, and each action is captured three times: *once from a front view*, *once from a side view*, and *once from any view under camera movement*. Camera movement is achieved using a professional photography stand equipped with a mobile chassis. For each participant,

we also vary the camera's height (up to 3 meters) and data acquisition distance (up to 30 meters) to obtain multiple shooting perspectives.

Data Annotation After completing data acquisition for the action recognition task, we performed classification and labeling for each sample to ensure accurate identification of the represented actions. For each behavioral data, we further organized and divided them based on the scene, participant number, and camera movement status. Each category was named according to the participant's number, action name, scene number, and camera status to ensure the clarity and accuracy of the dataset's organizational structure. In Fig. 3, which provides examples of several groups labels, it can be observed that each attributes displays distinct differences, for example, *Illumination Direction* in the bottom-right corner. We can distinguish the approximate direction of light based on the shadows in the scene. Moreover, we conducted detailed attributes annotations for each data point based on its characteristics. We believe that the granular annotations provided for event-based action recognition can facilitate a deeper understanding of event data.

3.2 Benchmark Evaluations

To conduct standardized evaluations of the models tested on our benchmark dataset, we have defined precise criteria for two types of action classification assessments. The accuracy is reported as a percentage for each criterion. We utilize 4 NVIDIA GeForce RTX 4090 GPUs for all of the training and testing.

Cross-Subject Evaluation For the cross-subject evaluation, the 47 participants were divided into three groups: training set, validation set, and test set. The validation set consists of 8 individuals with the following IDs: 3, 4, 5, 24, 27, 31, 41, 43. The test set comprises 9 individuals with IDs: 4, 7, 10, 11, 16, 33, 37, 42, 45. The remaining 31 individuals form the training set.

Multi-Group Evaluation For multi-group evaluation, the model training settings are the same as those used in the Cross-Subject Evaluation. We first conduct training from scratch across participants, and then perform separate testing on the test set with different settings for various groups.

4 Experimental Results

In this section, we employ various methods to conduct cross-subject and multi-group evaluations on our dataset. Firstly, we test our dataset using multiple approaches, including frame-based, learnable-based, token-based, and spike-based methods. Secondly, within the frame-based approach, we explore the impact of frame gap and time step for generating event frames on the experiment results of our dataset. Additionally, we demonstrate the performance of multiple groups in

Table 2: Evaluation of different methods on our dataset. Best models results with different input types are **highlighted**.

Methods	Year	Input Type	Backbone	top-1 acc.(%)	top-5 acc.(%)
C3D [53]	2015	Frame	3D CNN	21.99	45.81
I3D [9]	2017	Frame	ResNet50	32.30	59.05
R2Plus1D [54]	2018	Frame	ResNet34	36.06	63.67
SlowFast [17]	2019	Frame	ResNet50	**41.49**	68.19
TSM [31]	2019	Frame	ResNet50	40.87	**71.46**
EST [19]	2019	Learnable	ResNet34	**32.23**	**59.66**
TimeSformer [3]	2021	Token	Transformer	44.25	74.03
Swin-T [36]	2022	Token	Transformer	**48.06**	**74.47**
ESTF [58]	2022	Token	ResNet18	24.68	50.18
GET [43]	2023	Token	Transformer	37.28	61.59
Spikformer [61]	2022	Spike	Transformer	**36.94**	**62.37**
SDT [59]	2024	Spike	Transformer	35.43	58.81

various event-based action recognition models. Finally, we validate our dataset's complexity using standard experimental setups from a large-scale action recognition dataset, with details and findings in the supplementary material.

4.1 Evaluations of Action Recognition models

Evaluation of different methods Our dataset was tested on 12 different action recognition algorithms, namely, C3D [53], I3D [9], R2Plus1D [54], Slowfast [17], TSM [31], EST [19], TimeSformer [3], Swin-T [36], ESTF [58], GET [43], Spikformer [61], and SDT [59]. This section employed cross-subject evaluation criteria, and our experimental results are reported in Table 2. As can be noted, among the frame-based methods (generated at intervals of 500ms), SlowFast [17] achieved the highest top-1 accuracy of 41.49%, while the TSM [31] achieved the highest top-5 accuracy of 71.46%. As for the token-based methods, Swin-T [36] achieved the highest accuracy in both top-1 and top-5, with 48.06% and 74.47% respectively, and also achieved the highest accuracy among all methods. In the spike-based methods, Spikformer achieved the highest top-1 and top-5 accuracy of 36.94% and 62.37%, respectively.

Evaluation of different groups We conducted fine-grained group testing on models trained on DailyDVS-200 dataset, including Slowfast [17], TSM [31], EST [19], Swin-T [36], Spikformer [61], and the results are presented in Table 3.

We observed that within the same framework, the performance varied significantly depending on factors such as different camera motion states, action scopes, number of actors, and camera heights. For instance, Swin-T [36] exhibited different performance under CM (27.84% vs 58.05%), AR (59.24% vs 43.15% vs 34.59%), PN (44.7% vs 66.88%), and HE (44.46% vs 52.14%). Notably, the impact of lighting conditions on model performance differed from traditional video classification. We found that the recognition performance was optimal under side

Table 3: Fine-grained group testing on different models. Best and second best results are **highlighted** and underlined.

Methods	1. BC		2. DR		3. IL			4. DT		5. HE		6. LO		7. CM	
	Complex	Easy	Long	Short	Back	Front	Side	Far	Near	High	Low	Indoor	Outdoor	Move	Static
SlowFast [17]	45.98	39.20	40.94	42.18	44.05	34.52	46.13	41.62	41.64	38.34	45.36	43.21	39.08	21.05	51.81
TSM [31]	45.37	40.84	39.68	44.69	42.91	36.60	47.16	40.71	44.09	38.90	46.51	43.41	40.93	26.88	50.16
EST [19]	35.22	30.59	29.42	34.52	32.89	24.71	38.26	28.93	35.28	27.45	37.64	33.76	29.82	13.52	41.51
Spikformer [61]	38.96	33.45	33.86	36.68	36.77	30.09	39.11	31.57	39.00	31.82	39.52	38.54	30.33	13.74	46.15
Swin-T [36]	**52.86**	**45.37**	**48.57**	**47.64**	**51.61**	**39.11**	**53.39**	**47.16**	**48.89**	**44.46**	**52.14**	**50.15**	**44.70**	**27.84**	**58.05**

Methods	8. PN		9. PE		10. PR		11. AR			12. SH		13. PO		14. DI	
	One	Two	Frontal	Lateral	No	Yes	Full-body	Limbs	Micro	No	Yes	Stand	Sit	Day	Night
SlowFast [17]	39.15	55.57	44.52	39.04	41.81	41.48	52.05	37.59	25.16	37.07	42.73	42.93	35.92	43.10	33.65
TSM [31]	40.07	55.90	46.13	39.17	46.66	38.98	53.31	37.21	33.02	42.75	42.39	43.11	39.61	44.20	33.02
EST [19]	29.88	45.56	33.42	31.15	34.81	30.08	41.18	28.55	19.81	35.18	31.55	32.70	30.13	34.74	18.55
Spikformer [61]	33.88	44.10	37.09	33.94	37.39	33.80	43.86	31.95	23.27	39.47	34.45	35.01	37.24	38.30	19.81
Swin-T [36]	**44.70**	**66.88**	**50.98**	**45.43**	**51.35**	**45.33**	**59.24**	**43.15**	**34.59**	**46.78**	**48.36**	**48.96**	**44.08**	**50.22**	**36.32**

Table 4: Evaluation of existing models with different frame settings on our dataset. Best and second best results are **highlighted** and underlined.

Methods	0.5s			0.25s			0.125s		
	gap0	gap2	gap4	gap0	gap2	gap4	gap0	gap2	gap4
C3D [53]	21.99	14.98	11.70	31.10	24.82	22.82	44.81	32.62	27.75
I3D [9]	32.30	22.94	20.82	45.39	29.54	29.42	59.10	36.70	37.50
R2Plus1D [54]	36.06	26.39	**24.97**	49.65	36.62	32.10	58.88	48.06	40.29
SlowFast [17]	**41.49**	**26.90**	24.55	**52.16**	33.28	25.43	**64.09**	44.81	44.64
TSM [31]	40.87	23.67	22.97	49.55	**37.94**	**32.37**	61.76	**51.48**	**48.77**

lighting, followed by backlit conditions, while it was poorest under front-facing lighting.

In the testing of the same attribute category, such as camera motion, we observed that token-based models (Swin-T [36]) outperformed Learnable-based model EST [19], frame-based traditional models (Slowfast [17] and TSM [31]) and spiking-based models (Spikformer [61] and SDT [59]) (27.84% vs 13.52% vs 26.88% vs 21.05% vs 13.74%). We analyzed this phenomenon, attributing it to the presence of excessive background events in mobile camera settings. Event-based recognition models, due to their high frame rate nature, tend to focus more on background noise, thereby impacting the model's recognition capability. Conversely, token-based models mitigate this effect. As shown in Fig. 4 (d), the number of events in mobile camera settings is significantly higher than in stationary camera settings.

To mitigate the bias in testing groups to camera motion attributes, we conducted training and testing on the TSM [31] model using different proportions of dynamic and static data. Our experimental results, as illustrated in Fig. 5 (a), show a noticeable decrease in the accuracy of action evaluated under both dynamic and static cameras as the proportion of dynamic data increases. Contrary to expectations, there was no improvement in accuracy with the increasing volume of event data captured by dynamic cameras. This suggests that current

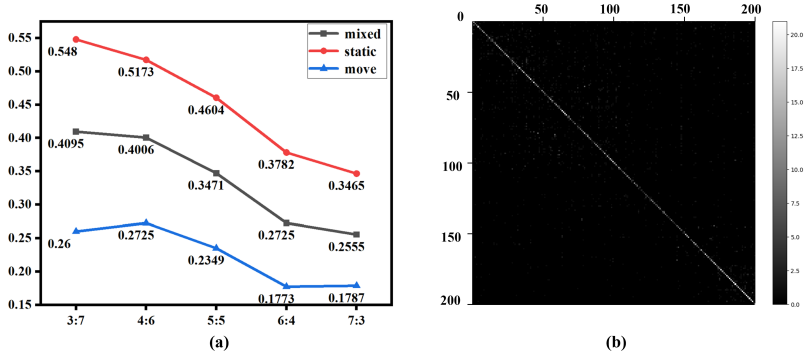


Fig. 5: (a) Evaluation of using different sizes of Moving camera set for action recognition. (b) Confusion matrix of Swin-T [36].

action recognition models struggle to effectively learn action data captured by dynamic cameras, highlighting an urgent issue that requires resolution.

Evaluation of different frame settings Table 4 provides a detailed comparison of several widely used frame-based action recognition models on DailyDVS-200 dataset with different frame settings. Since frame-based action recognition models reconstruct frame sequences from event stream data [45], frame sequences formed with different time step lengths often exhibit significant differences. We generated reconstructed frame sequences using three different fixed time step lengths of 0.5s, 0.25s, and 0.125s, respectively, and conducted experiments with frame intervals (gap) of 0, 2, and 4. Our experimental results revealed that increasing the event step length to capture additional temporal information contributes to improving the model's accuracy. However, increasing the frame interval leads to partial information loss, which is detrimental to model recognition.

Detailed analysis to actions and methods We also conducted a detailed analysis of the performance of different action recognition methods on our proposed DailyDVS-200 dataset. We took four best-performing methods (SlowFast [17], Swin-T [36], EST [19], Spikformer [61]) as examples for analysis. Firstly, we plot the confusion matrices of these methods. The confusion matrix of Swin-T [36] is shown in Fig. 5 (b) as an example, we can see there are still many actions cannot be correctly classified. Specifically, considering the large number of action categories, we also analyzed the top-10 accurate actions and top-10 incorrect actions for each method (see Table 5).

Among them, we found that the actions *hand in hand circling* and *push-up* ranked in the top-10 in accuracy across all four models. This could be attributed to uniqueness and large motion range of these two actions, resulting in strong spatiotemporal features. However, we also observed that some actions, such as *pitch* and *play table tennis* were prone to misclassification in most models. A

Table 5: Top-10 accurate and top-10 incorrect actions of different methods. The same action is marked with the same color.

Methods	Top-10 accurate actions	Top-10 incorrect actions	Methods	Top-10 accurate actions	Top-10 incorrect actions
Slowfast [17]	1. pull out the chair 2. open curtains 3. **hand in hand circling** 4. close the door 5. wash towels 6. fall down 7. turn the light on 8. **wipe the table** 9. cross your legs 10. **push-up**	1. hammer table 2. **play table tennis** 3. open window 4. clean windows 5. **pitch** 6. charge a phone 7. arrang cards 8. V sign 9. use a tablet 10. kick a ball	EST [19]	1. tie shoelaces 2. mutual bow 3. turn the light on 4. **push-up** 5. turn the light off 6. fall down 7. lie on the table 8. close window 9. cheers 10. **hand in hand circling**	1. **play table tennis** 2. open window 3. close curtains 4. slap the table 5. **pitch** 6. OK sign 7. make paper cuttings 8. charge a phone 9. hit people with things 10. headache
Swin-T [36]	1. **hand in hand circling** 2. arm wrestling 3. **push-up** 4. close curtains 5. open curtains 6. close the door 7. sit-up 8. **wipe the table** 9. moves heavy objects 10. cross legs	1. hammer table 2. **pitch** 3. take off headphones 4. write 5. blow nose 6. V sign 7. crush paper into a ball 8. take something from bag 9. chest pain 10. open the bottle	Spikformer [61]	1. **wipe the table** 2. close curtains 3. stand up 4. **hand in hand circling** 5. lie on the table 6. mutual bow 7. cross legs 8. clean the windows 9. **push-up** 10. go upstairs	1. **pitch** 2. **play table tennis** 3. plug in the power strip 4. take off shoes 5. trim nails 6. stomachache 7. play with hair 8. roll up sleeves 9. backache 10. headache

possible explanation is the fast speed and short duration of these actions, which pose challenges for feature extraction, thereby affecting the accuracy of classification. Furthermore, we found that two-person actions were rarely in the top-10 incorrect actions, while micro-actions such as the *V sign* and *OK sign* were seen difficult to accurately recognize. Another notable phenomenon is that *closing curtains* was well recognized in Swin-T [36], however, poorly recognized in EST [19], indicating significant variations among different models. Additionally, we conducted detailed testing in different scenarios and found that performance was consistently poor across all models in scenes labeled as *with shadow*, *front light*, *long distance*, *high-angle shot*, *night* and *outdoor*, which needed to pay more attention in the future. You can see the supplementary material for additional scenes evaluation details.

5 Conclusion

In this study, we introduce a comprehensive and challenging large-scale event-based action recognition dataset, named DailyDVS-200, which serves as a new benchmark for event-based action recognition. This dataset consists of 200 categories of everyday human actions performed by 47 individuals, resulting in over 22,000 event sequences. We meticulously consider the complexity of scenes, diversity of subjects, and variability of actions, providing 14 different attribute annotations for each data sample. Through evaluations and intragroup testing of over 10 different methods, we gain a nuanced understanding of the various scenarios in real-life settings where event cameras are utilized. Furthermore, comparison with other large-scale event datasets reveals that existing datasets exhibit high performance with traditional models, thus hindering innovation in event-based methods and the full utilization of event cameras' advantages. By introducing the proposed DailyDVS-200 dataset, we aim to provide new research directions for methodological innovation in this field.

Acknowledgements

This work was partly supported by Chinese Defense Advanced Research Program (50912020105), NSFC of China under Grants Nos. 62073258 and 62072352, and Natural Science Foundation of Shaanxi Province (2024JC-JCQN-66).

References

1. Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., et al.: A low power, fully event-based gesture recognition system. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7243–7252 (2017)
2. Baldwin, R.W., Liu, R., Almatrafi, M., Asari, V., Hirakawa, K.: Time-ordered recent event (tore) volumes for event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2519–2532 (2022)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
4. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y.: Graph-based spatio-temporal feature learning for neuromorphic vision sensing. IEEE Transactions on Image Processing **29**, 9084–9098 (2020)
5. de Blegiers, T., Dave, I.R., Yousaf, A., Shah, M.: Eventtransact: A video transformer-based framework for event-camera based action recognition. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–7. IEEE (2023)
6. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014). https://doi.org/10.1109/JSSC.2014.2342715
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
8. Cannici, M., Ciccone, M., Romanoni, A., Matteucci, M.: A differentiable recurrent surface for asynchronous event-based data. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 136–152. Springer (2020)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. Che, K., Leng, L., Zhang, K., Zhang, J., Meng, Q., Cheng, J., Guo, Q., Liao, J.: Differentiable hierarchical and surrogate gradient search for spiking neural networks. Advances in Neural Information Processing Systems **35**, 24975–24990 (2022)
11. Chen, S., Guo, M.: Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1682–1683. IEEE (2019)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Dong, Y., Li, Y., Zhao, D., Shen, G., Zeng, Y.: Bullying10k: A large-scale neuromorphic dataset towards privacy-preserving bullying recognition. Advances in Neural Information Processing Systems **36** (2024)

14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Duarte, L., Neto, P.: Event-based dataset for the detection and classification of manufacturing assembly tasks. Data in Brief **54**, 110340 (2024)
16. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 203–213 (2020)
17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
18. Gao, Y., Lu, J., Li, S., Ma, N., Du, S., Li, Y., Dai, Q.: Action recognition and benchmark using event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
19. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5633–5643 (2019)
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
21. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2146–2156 (2021)
22. Kliper-Gross, O., Hassner, T., Wolf, L.: The action similarity labeling challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3), 615–621 (2011)
23. Kong, Y., Fu, Y.: Human action recognition and prediction: A survey. International Journal of Computer Vision **130**(5), 1366–1401 (2022)
24. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
25. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: a hierarchy of event-based time-surfaces for pattern recognition. IEEE transactions on pattern analysis and machine intelligence **39**(7), 1346–1359 (2016)
26. Laptev, I.: On space-time interest points. International journal of computer vision **64**, 107–123 (2005)
27. Li, H., Liu, H., Ji, X., Li, G., Shi, L.: Cifar10-dvs: an event-stream dataset for object classification. Frontiers in neuroscience **11**, 309 (2017)
28. Li, J., Wang, X., Zhu, L., Li, J., Huang, T., Tian, Y.: Retinomorphic object detection in asynchronous visual streams. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1332–1340 (2022)
29. Li, Y., Dong, Y., Zhao, D., Zeng, Y.: N-omniglot, a large-scale neuromorphic dataset for spatio-temporal sparse few-shot learning. Scientific Data **9**(1), 746 (2022)
30. Li, Y., Zhou, H., Yang, B., Zhang, Y., Cui, Z., Bao, H., Zhang, G.: Graph-based asynchronous event processing for rapid object recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 934–943 (2021)
31. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7083–7093 (2019)

32. Lin, Y., Ding, W., Qiang, S., Deng, L., Li, G.: Es-imagenet: A million event-stream classification dataset for spiking neural networks. Frontiers in neuroscience **15**, 1546 (2021)
33. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR 2011. pp. 3337–3344. IEEE (2011)
34. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
35. Liu, Q., Xing, D., Tang, H., Ma, D., Pan, G.: Event-based action recognition using motion information and spiking neural networks. In: IJCAI. pp. 1743–1749 (2021)
36. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
37. Messikommer, N., Gehrig, D., Loquercio, A., Scaramuzza, D.: Event-based asynchronous sparse convolutional networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 415–431. Springer (2020)
38. Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., Knoll, A.: Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. Frontiers in neurorobotics **13**, 38 (2019)
39. Moeys, D.P., Corradi, F., Kerr, E., Vance, P., Das, G., Neil, D., Kerr, D., Delbrück, T.: Steering a predator robot using a mixed frame/event-driven convolutional neural network. In: 2016 Second international conference on event-based control, communication, and signal processing (EBCCSP). pp. 1–8. IEEE (2016)
40. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
41. Neftci, E.O., Mostafa, H., Zenke, F.: Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. IEEE Signal Processing Magazine **36**(6), 51–63 (2019)
42. Orchard, G., Jayawant, A., Cohen, G.K., Thakor, N.: Converting static image datasets to spiking neuromorphic datasets using saccades. Frontiers in neuroscience **9**, 437 (2015)
43. Peng, Y., Zhang, Y., Xiong, Z., Sun, X., Wu, F.: Get: group event transformer for event-based vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6038–6048 (2023)
44. Posch, C., Matolin, D., Wohlgenannt, R.: A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. IEEE Journal of Solid-State Circuits **46**(1), 259–275 (2010)
45. Rebecq, H., Horstschaefer, T., Scaramuzza, D.: Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization (2017)
46. Sabater, A., Montesano, L., Murillo, A.C.: Event transformer. a sparse-aware solution for efficient event data processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2677–2686 (2022)
47. Schaefer, S., Gehrig, D., Scaramuzza, D.: Aegnn: Asynchronous event-based graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12371–12381 (2022)
48. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on Multimedia. pp. 357–360 (2007)

49. Serrano-Gotarredona, T., Linares-Barranco, B.: Poker-dvs and mnist-dvs. their history, how they were made, and other details. Frontiers in neuroscience **9**, 481 (2015)
50. Shi, Q., Cheng, L., Wang, L., Smola, A.: Human action segmentation and recognition using discriminative semi-markov models. International journal of computer vision **93**, 22–32 (2011)
51. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: Hats: Histograms of averaged time surfaces for robust event-based object classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1731–1740 (2018)
52. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
53. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
54. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
56. Wang, H., Oneata, D., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. International journal of computer vision **119**, 219–238 (2016)
57. Wang, X., Li, J., Zhu, L., Zhang, Z., Chen, Z., Li, X., Wang, Y., Tian, Y., Wu, F.V.: Reliable object tracking via collaboration of frame and event flows. arxiv 2021. arXiv preprint arXiv:2108.05015
58. Wang, X., Wu, Z., Jiang, B., Bao, Z., Zhu, L., Li, G., Wang, Y., Tian, Y.: Hardvs: Revisiting human activity recognition with dynamic vision sensors. arXiv preprint arXiv:2211.09648 (2022)
59. Yao, M., Hu, J., Zhou, Z., Yuan, L., Tian, Y., Xu, B., Li, G.: Spike-driven transformer. Advances in Neural Information Processing Systems **36** (2024)
60. Zeng, Y., Zhao, D., Zhao, F., Shen, G., Dong, Y., Lu, E., Zhang, Q., Sun, Y., Liang, Q., Zhao, Y., et al.: Braincog: A spiking neural network based, brain-inspired cognitive intelligence engine for brain-inspired ai and brain simulation. Patterns **4**(8) (2023)
61. Zhou, Z., Zhu, Y., He, C., Wang, Y., Yan, S., Tian, Y., Yuan, L.: Spikformer: When spiking neural network meets transformer. arXiv preprint arXiv:2209.15425 (2022)
62. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
63. Zhu, L., Li, J., Wang, X., Huang, T., Tian, Y.: Neuspike-net: High speed video reconstruction via bio-inspired neuromorphic cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2400–2409 (2021)
64. Zhu, L., Wang, X., Chang, Y., Li, J., Huang, T., Tian, Y.: Event-based video reconstruction via potential-assisted spiking neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3594–3604 (2022)
65. Zhu, S., Yang, T., Mendieta, M., Chen, C.: A3d: Adaptive 3d networks for video action recognition. arXiv preprint arXiv:2011.12384 (2020)