N-ImageNet: Towards Robust, Fine-Grained Object Recognition with Event Cameras

Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim

Dept. of Electrical and Computer Engineering, Seoul National University, Korea

{82magnolia, wogur110, ssonpull519, 96lives, youngmin.kim}@snu.ac.kr

Abstract

We introduce N-ImageNet, a large-scale dataset targeted for robust, fine-grained object recognition with event cameras. The dataset is collected using programmable hardware in which an event camera consistently moves around a monitor displaying images from ImageNet. N-ImageNet serves as a challenging benchmark for event-based object recognition, due to its large number of classes and samples. We empirically show that pretraining on N-ImageNet improves the performance of event-based classifiers and helps them learn with few labeled data. In addition, we present several variants of N-ImageNet to test the robustness of event-based classifiers under diverse camera trajectories and severe lighting conditions, and propose a novel event representation to alleviate the performance degradation. To the best of our knowledge, we are the first to quantitatively investigate the consequences caused by various environmental conditions on event-based object recognition algorithms. N-ImageNet and its variants are expected to guide practical implementations for deploying event-based object recognition algorithms in the real world.

1. Introduction

Event cameras are neuromorphic vision sensors that encode visual information as a sequence of events, and have a myriad of benefits such as high dynamic range, low energy consumption, and microsecond-scale temporal resolution. However, algorithms for processing event data are still at their nascency. This is primarily due to the lack of a large, fine-grained dataset for training and evaluating different event-based vision algorithms. While the number of event camera datasets surged in the past few years, many fine-grained datasets lack size [39], whereas large-scale real-world datasets lack label diversity [50]. Large amounts of publicly available data are one of the key factors in the recent success of computer vision. For example, ImageNet [46] triggered the development of accurate, high
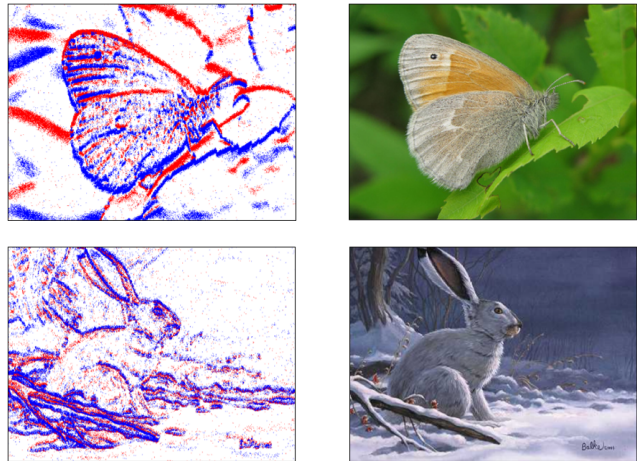

Figure 1: Sample events from N-ImageNet displayed along with their RGB counterparts from ImageNet [46]. Positive, negative events are shown in blue and red, respectively.

performance object recognition algorithms [15, 23] whereas MS-COCO [29] led to the advent of eloquent image captioning systems [60].

We provide N-ImageNet, an event camera dataset targeted for object recognition that surpasses all existing datasets in both size and label granularity as summarized in Table 1 and Figure 2. Since it is infeasible to manually obtain real-world instances of thousands of object categories, we opt to generate events by moving the sensor in front of an LCD monitor which displays images from ImageNet [46] as in [39, 27] using programmable hardware.

N-ImageNet is projected to function as a challenging benchmark for event-based object recognition algorithms. As shown in Table 1, evaluations of various classifiers on N-ImageNet demonstrate a large room for improvement, in contrast to popular benchmarks such as N-Cars [50] and N-Caltech101 [39]. We also experimentally justify the effectiveness of N-ImageNet pretraining. Models pretrained on N-ImageNet show a large amount of performance gain in various object recognition benchmarks, and are capable of

Dataset	# of Samples	# of Classes	Top Accuracy	Robustness Quantifiable?
N-Cars [50]	24029	2	95.8 [50]	✕
N-Caltech101 [39]	8709	101	90.6 [13]	✕
CIFAR10-DVS [27]	10000	10	69.2 [47]	✕
ASL-DVS [4]	100800	24	94.6 [20]	✕
N-MNIST [39]	70000	10	99.2 [20]	✕
MNIST-DVS [49]	30000	10	99.1 [20]	✕
N-SOD [44]	189	4	97.14 [44]	✕
DVS128-Gesture [2]	1342	11	99.62 [19]	✕
N-ImageNet	**1781167**	**1000**	**48.93**	◯

Table 1: Comparison of N-ImageNet with other existing benchmarks for event classification.

rapidly generalizing to new datasets even with a small number of training samples.

We further analyze the robustness of event-based object recognition algorithms amidst changes in camera trajectories and lighting conditions. Event cameras can operate in highly dynamic scenes and low light environments, but events produced in such conditions tend to have more noise and artifacts from motion blur [9]. We record variants of N-ImageNet under diverse camera trajectories and lighting, and quantify the significant performance degradation of event-based classifiers under environment changes. To the best of our knowledge, our dataset is the first event camera dataset capable of providing quantitative benchmarks for robust event-based object recognition, as shown in Table 1. In addition, we propose a simple event representation, called DiST (Discounted Sorted Timestamp Image), that shows improved robustness under the external variations. DiST penalizes events that are more likely to be noise, and uses sorted indices of event timestamps to ensure durability against speed change.

To summarize, our main contributions are (i) N-ImageNet, the largest fine-grained event camera dataset to date, thus serving as a challenging benchmark, (ii) N-ImageNet pretraining which leads to considerable performance improvement, (iii) N-ImageNet variants that enable quantitative robustness evaluation of event-based object recognition algorithms, and (iv) an event camera representation exhibiting enhanced robustness in diverse environment changes.

2. Related Work

Event Camera Datasets With rising interests in event-based vision, the field has seen a wide range of event camera datasets targeted towards various computer vision tasks, such as object detection [42, 8], optical flow estimation [61, 45], and image reconstruction [48, 33, 56, 55, 54, 58, 28].

For event-based object recognition in particular, diverse datasets [39, 49, 27, 4, 50, 34, 18, 2, 6, 30, 52, 53, 44] have been proposed, and can be categorized depending on whether the recordings consist of real-world objects or
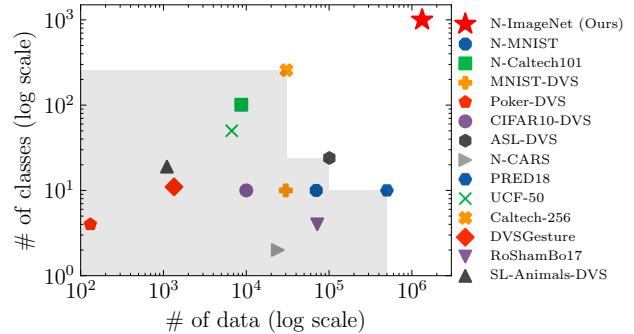

Figure 2: Comparison of N-ImageNet (the red star in the upper-right corner) against existing datasets (distributed in the area shaded in gray) in terms of dataset size and class count. Note that the axes are displayed in log scale.

monitor-displayed images. N-Cars [50] and ASL-DVS [4] contain event data obtained by directly capturing various real-world objects. Such datasets typically have a smaller number of labels compared to datasets acquired from monitor-displayed images since it is difficult to acquire fine-grained labels of real-world recordings.

Datasets such as N-MNIST, N-Caltech101 [39], MNIST-DVS [49] and CIFAR10-DVS [27], which belong to the latter category, are recorded by moving an event camera around monitors displaying images of well-known datasets like MNIST [25] and Caltech101 [11]. Monitor-generated event camera datasets can be considered synthetic in some aspects, but contain abundant labels from their original image datasets, which are beneficial for training and evaluating event-based object recognition algorithms. We follow the data acquisition procedure of these datasets, and generate a large-scale, fine-grained dataset by transforming ImageNet [46] data to event camera recordings. Furthermore, our experiments demonstrate that leveraging such a large-scale event dataset greatly improves the performance of event-based object recognition algorithms.

Event-Based Object Recognition Event camera data exhibits unique characteristics, namely asynchronicity and sparsity. Existing event-based object recognition algorithms could be classified by whether such characteristics are utilized. A large body of literature proposes models that perform asynchronous updates in a sparse manner [26, 38, 40, 35, 43, 2, 24, 50, 32, 5]. Recently proposed MatrixLSTMs [32] exhibit competitive results when evaluated on popular event-based object recognition benchmarks such as N-Cars [50] and N-Caltech101 [11]. MatrixLSTMs handle streams of event data using sparse updates of LSTMs [16], where an adaptive 'grouping' operation is used to update the network outputs asynchronously.

Algorithms that avoid the direct exploitation of the

sparse and asynchronous nature of event data also prevail [14, 57, 13, 37, 17, 22, 4]. These methods place more weight on performance, typically surpassing their sparse and asynchronous competitors in accuracy when tested on various object recognition datasets. Event Spike Tensors (EST) [14, 13] aggregate events using a learned kernel, resulting in a highly versatile representation of event data. Relatively simple encodings of event camera data have also been proposed, as in EV-FlowNet [61] and EV-Gait [57], where events are accumulated to form a four-channel image consisting of event counts and the newest timestamps of each pixel. We evaluate the performance of the aforementioned object recognition algorithms on N-ImageNet. Due to its scale and label diversity, N-ImageNet is capable of providing reliable assessments on different event-based object recognition algorithms.

Robustness in Event-Based Object Recognition Event cameras are known to successfully function in low-light conditions and dynamic scenes. However, the robustness of event-based classifiers under such conditions is a relatively unstudied problem. Most existing works [59, 42, 50] either only display qualitative results or experiment with synthetic adversaries. Sironi *et al*. [50] shows the robustness of their proposed event representation in various real-world objects, but the analysis is only made qualitatively. Wu *et al*. [59] quantitatively investigates the effect of noise on event-based classifiers, but the experiments are conducted on synthetic noise. Although there exist previous works such as Wang *et al*. [57] where real event camera noise is investigated, the experiments are carried out with a static camera under constant, ambient lighting. The N-ImageNet variants recorded under diverse lighting and camera trajectories enable realistic, quantitative assessment on the robustness of event-based object recognition algorithms.

3. Method Overview

3.1. Dataset Acquisition

N-ImageNet Dataset Following the footsteps of previous image-to-event conversion methods [39, 27, 49], we acquire N-ImageNet from an event camera that observes monitor-displayed images from ImageNet [46]. Since events are triggered by pixel intensity changes, external stimuli are solicited to generate event data. One viable solution for generating constant stimuli would be to keep the camera *still* and make the displayed images *move*, as in [27, 49]. However, as pointed out by Orchard *et al*. [39], such methods suffer from artifacts induced by the refresh mechanisms of monitors. Removing such artifacts requires an additional post-processing step in the frequency domain [27, 49], which is costly due to the immense number of images in ImageNet, and may alter the inherent subtleties in the raw measure-
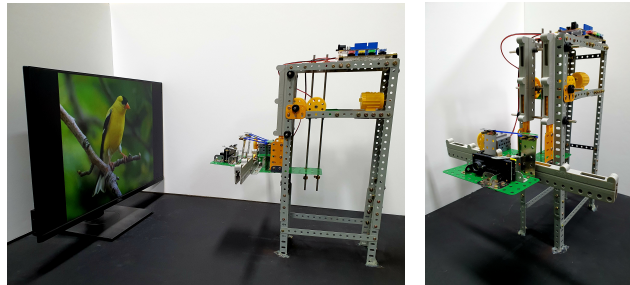

Figure 3: Custom hardware designed to convert RGB images to event camera data.

ments.

Instead, we opt to *move* the event camera around an LCD monitor displaying *still* images from ImageNet [46], as proposed in Orchard *et al*. [39]. We devise custom hardware to trigger perpetual camera motion as shown in Figure 3. The device consists of two geared motors connected to a pair of perpendicularly adjacent gear racks where the upper and lower motors are responsible for vertical and horizontal motion, respectively. Each motor is further linked to a programmable Arduino board [3], which can control the camera movement. In all our experiments, the event camera is vibrated vertically and horizontally on a plane parallel to the LCD monitor screen. The amplitude and frequency of vibration are controlled by the program embedded in the Arduino microcontroller [3].

Once the device is prepared, the event camera is mounted at the forefront of the device for recording event data. We use the 480×640 resolution Samsung DVS Gen3 [51] event camera for recording event sequences, and a 24-inch Dell P2419H LCD monitor for displaying RGB images from ImageNet [46]. The acquisition process is performed in a sealed chamber, to ensure that no external light will adulterate the recorded events. Under this setup, both the training and validation splits of ImageNet are converted to N-ImageNet. Thanks to the large scale and label granularity of ImageNet, N-ImageNet serves as a challenging benchmark for event-based object recognition and significantly boosts the performance of event-based classifiers via pre-training. These amenable properties of N-ImageNet will be further examined in Section 4.1.

Datasets for Robustness Evaluation We additionally present a benchmark to quantitatively assess the robustness of event-based object recognition algorithms. Event representations are vulnerable to alternations in camera motion or illumination, as even small changes can trigger a wide variety of event sequences. We simulate such changes using the programmable camera trajectory and monitor brightness from our hardware setup, and generate variants of the N-ImageNet validation split. We use these validation datasets

Dataset	Frequency (Hz)	Amplitude (mm)	Shape
Original	5	3	Square ↺
Validation 1	8.33	4.5	Vertical
Validation 2	5	3	Horizontal
Validation 3	5	6	Vertical
Validation 4	5	6	Horizontal
Validation 5	5	6	Square ↺

Table 2: Validation datasets made with various camera motion. ↺ indicates counterclockwise rotation. The amplitudes of square trajectories represent the lengths of the diagonals.

Dataset	Brightness Level	Gamma	Illuminance (lux)
Original	50	1	70.00
Validation 6	0	0.7	12.75
Validation 7	0	1	23.38
Validation 8	100	1	95.50
Validation 9	100	1.5	111.00

Table 3: Validation datasets generated under various brightness conditions.

to quantitatively evaluate the performance degradation of existing object recognition algorithms in Section 4.2.

Specifically, we present nine validation datasets to test the robustness of event-based object recognition algorithms amidst changes in motion or illumination. Table 2 lists five datasets with different camera trajectories. The frequency, amplitude, and trajectory shape of the camera movement are modified with the Arduino microcontroller [3]. Table 3 shows the monitor configurations of four additional validation datasets, designed to examine the effect of scene brightness changes on event-based classifiers. Note that Validation 6 and 9 datasets are intended to model scenes with exceedingly low/high illumination by using extreme monitor gamma values. We also report the illuminance measured at the position of the event camera since the same numerical configurations of different monitors may yield distinct displayed results. In all cases, other configurations are kept the same as the original N-ImageNet dataset.

3.2. Robust Event-Based Object Recognition

As many event-based classifiers are typically trained and tested in datasets captured in predefined conditions [39, 4], the performance degradation is inevitable in challenging scenarios that arise in real-life applications. In Section 4.2 we evaluate the robustness of existing object recognition algorithms with N-ImageNet variants and demonstrate that external changes indeed incur performance degradation. Even the best-performing event-based algorithms [14, 5] are fragile to diverse motion and illumination variations. Ironically, the main benefits of event cameras include the fast temporal response and high dynamic range.

We introduce Discounted Sorted Timestamp Image (DiST), which is designed specifically for robustness against changes in camera trajectory and lighting. The intuition behind DiST is twofold: (i) noisy events incurred from severe illumination can be suppressed with the evidence of the spatio-temporal neighborhood, and (ii) relative timestamps are robust against the speed of camera motion compared to raw timestamp values.

A typical output of an event camera is a sequence of events $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}$, where e_i indicates brightness change with polarity $p_i \in \{-1, 1\}$ at pixel location (x_i, y_i) at time t_i. Given an event camera of spatial resolution $H \times W$, let $\mathcal{N}_\rho(x, y, p)$ denote the set of events in $\mathcal{E}$ of polarity p, confined within a spatial neighborhood of size ρ around (x, y). For example, $\mathcal{N}_0(x, y, p)$ would indicate the set of all events of polarity p at the pixel coordinate (x, y).

DiST aggregates a sequence of events $\mathcal{E}$ into a 2 channel image $\mathbf{S} \in \mathbb{R}^{H \times W \times 2}$. DiST initiates its representation as the timestamp image [41]. The timestamp image is a 2 channel image, which stores the raw timestamp of the newest event at each pixel, namely $\mathbf{S}_o(x, y, p) = T_{\text{new}}(\mathcal{N}_0(x, y, p))$, where $T_{\text{new}}(\cdot)$ indicates the newest timestamp.

We first define the Discounted Timestamp Image (DiT) $\mathbf{S}_D$, obtained by subtracting the event occurrence period of the neighborhood from the newest timestamp in $\mathbf{S}_o$,

$$\mathbf{S}_D(x, y, p) = \mathbf{S}_o(x, y, p) - \alpha \mathbf{D}(x, y, p). \quad (1)$$

Here α is a constant discount factor and $\mathbf{D}(x, y, p)$ is the neighborhood event occurrence period,

$$\mathbf{D}(x, y, p) = \frac{T_{\text{new}}(\mathcal{N}_\rho(x, y, p)) - T_{\text{old}}(\mathcal{N}_\rho(x, y, p))}{C(\mathcal{N}_\rho(x, y, p))}, \quad (2)$$

where $T_{\text{old}}(\cdot)$ is defined similarly as $T_{\text{new}}(\cdot)$ and $\rho > 1$. For each pixel, the discount $\mathbf{D}(x, y, p)$ represents the time range $(T_{\text{new}}(\cdot) - T_{\text{old}}(\cdot))$ in which events are generated from the neighborhood $\mathcal{N}_\rho$, normalized by its event count $C(\cdot)$.

The discount mechanism is designed to be robust against event camera noise. Figure 4 illustrates the typical patterns for event sequences (left) and the resulting representation of DiST that resolves the discrepancies due to noise (right) in the 1-D case. Two dominant factors of event camera noise are background activities and hot pixels [9, 12]. Background activities are low frequency noise [12] more likely to occur in low-light conditions, caused by transistor leak currents or random photon fluctuations [9, 36, 21]. Figure 4 (a) indicates background activities, whose low frequency results in higher discounts from Equation 2. Hot pixels are triggered by the improper resets of events [9, 12], and are often spatially isolated [12] (Figure 4 (c)). Such pixels have a small neighborhood count $C(\cdot)$ and thus are highly discounted in Equation 2.
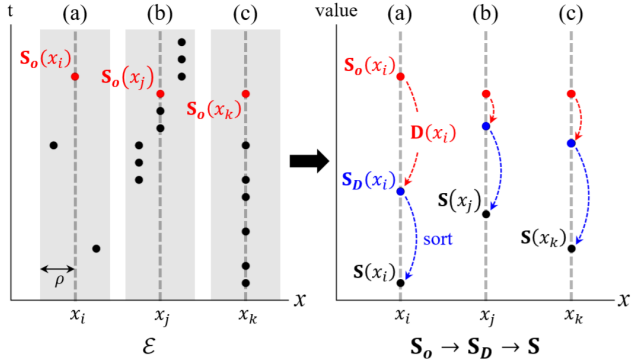
Figure 4: Suppressing event camera noise with the discount mechanism. One-dimensional events with a single polarity are shown. Events colored in red are the newest event of each pixel and the neighborhood regions $\mathcal{N}_\rho$ are shaded in gray. (a), (b), and (c) denote background activity, normal events, and hot pixels, respectively. Both the background activity and hot pixels in the raw event sequence (left) result in large discounts and are suppressed by DiST (right).

The final representation of DiST, $\mathbf{S}$ is the normalized sorted indices of $\mathbf{S}_D$,

$$\mathbf{S} = \texttt{argsort}(\mathbf{S}_D) / \max_{x,y,p} \texttt{argsort}(\mathbf{S}_D), 33 \quad (3)$$

where $\texttt{argsort}$ denotes the operation that returns the sorted indices of a tensor. The transformation from $\mathbf{S}_D$ to $\mathbf{S}$ is similar to Alzugaray *et al.* [1]. However, DiST performs a global sort where all pixels in $\mathbf{S}_D$ are sorted with a single ordering scheme, instead of the local, patch-wise sorting in Alzugary *et al.* [1]. The global sort allows for an efficient implementation of DiST, which is further explained in the supplementary material.

By using the normalized relative timestamps, i.e., sorted indices, DiST is resilient against the camera speed. To illustrate, consider the sequences of one-dimensional events (x, t, p) displayed in Figure 5. While the absolute values of the timestamps [41] are directly affected by the speed change, DiST remains constant.

We expect DiST to serve as a baseline representation for robust event-based object recognition. Its robustness against camera trajectory and scene illumination changes will be quantitatively investigated in Section 4.2.

4. Experimental Results

In this section, we empirically validate various properties of N-ImageNet. With its large scale and label diversity, N-ImageNet is not only a useful benchmark to assess various event-based representations, but can also boost the performance of existing algorithms via pre-training (Section 4.1). In Section 4.2, we investigate the robustness of event-based classifiers against diverse external conditions,

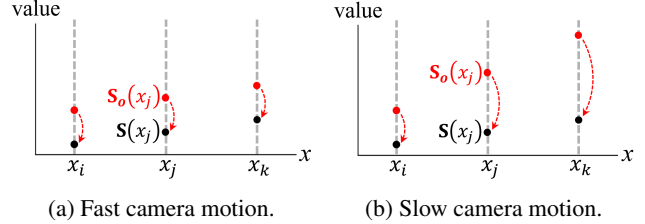(a) Fast camera motion. (b) Slow camera motion.

Figure 5: Robustness of DiST against event camera speed. Similar to Figure 4, one-dimensional events are displayed and red events denote the newest event. While the timestamps of red events vary with the camera speed, their relative timestamps obtained from sorting remain constant.

along with the efficacy of our proposed event representation, DiST.

Event Representations for Object Recognition We introduce the event representations used throughout our experiments. The representations are inputs to the object recognition algorithms, while the backbone classifier is fixed to ResNet34 [15]. This is because most event-based object recognition algorithms [14, 5] only differ in the input event representation and share a similar classification backbone. Eleven event representations are selected for evaluation on N-ImageNet and its variants, as shown in Table 4.

Two of the representations are learned from the data, namely MatrixLSTM [5] and Event Spike Tensor (EST) [14]. After the events are passed through LSTM [16] for MatrixLSTM and multilayer perceptrons for EST, the outputs are further voxelized to form an image-like representation.

The remaining representations can be classified based on how the timestamps are handled. Two of these representations discard temporal information, and only use the locations of events. Binary event image [7, 37] is generated by assigning 1 to pixels with events, and 0 to others. Event histogram [31] is an extension of the former, additionally keeping the event count of each pixel.

Three representations use the raw timestamps to leverage temporal information. Timestamp image [41] caches the newest timestamp for each pixel. Event image [57, 61] is a richer representation that concatenates the event histogram [31] and timestamp image [41]. Time surface [24] extends the timestamp image [41] in a slightly different manner, by passing each timestamp through an exponential filter. This allows the surface to place more weight on the newest events, which enhances the sharpness of the representation. The aforementioned representations with raw timestamps can be vulnerable to camera speed changes, as pointed out in Section 3.2.

We further include representations targeted to enhance robustness. HATS [50] improves the robustness against

Representation	Description	# of Channels	Accuracy(%)
MatrixLSTM [5]	Learned with LSTM	3	32.21
Event Spike Tensor [14]	Learned with MLP	18	**48.93**
Binary Event Image [7]	Binarized event occurence	2	46.36
Event Histogram [31]	Event counts	2	47.73
Event Image [57]	Event counts and newest timestamps	4	45.77
Time Surface [24]	Exponential of newest timestamps	2	44.32
HATS [50]	Aggregated newest timestamps	2	47.14
Timestamp Image [41]	Newest timestamps	2	45.86
Sorted Time Surface [1]	Sorted newest timestamps	2	47.90
DiT	Discounted newest timestamps	2	46.1
DiST	Sorted discounted timestamps	2	**48.43**

Table 4: N-ImageNet validation accuracy evaluated on various event representations.

event camera noise. Specifically, the outliers are smoothed by aggregating neighboring pixels of the time surface [24]. We use a slightly modified version of HATS [50] for more competitive results, and the details are provided in the supplementary material. Surface of active events with sort normalization [1], which we will refer to as sorted time surface, is robust against camera speed changes as the sorting generates relative timestamps. DiST, as explained in Section 3.2, is robust against both event camera noise and speed changes. We additionally report results on the variant of DiST without sorting, namely the Discounted Timestamp Image (DiT). Evaluation on DiT can shed light on the importance of the sorting operation in DiST.

Implementation Details All inputs are reshaped into a 224×224 grid to restrict GPU memory consumption and shorten inference time. All models are trained from scratch with a learning rate of 0.0003, except for the learned representations (MatrixLSTM [5] and EST [14]). The weights are initialized with ImageNet pre-training for these representations to fully replicate the training setup specified in the original works. We train these models with a learning rate of 0.0001. Further information regarding experimental details is provided in the supplementary material.

4.1. Evaluation Results with N-ImageNet

Event-based Object Recognition Table 4 displays the evaluation results of existing event-based object recognition algorithms on N-ImageNet. The accuracy of the best per-

forming model on N-ImageNet (48.9%), is far below that of the state-of-the-art model on ImageNet [10] (90.2%). The clear gap indicates that mastering N-ImageNet is still a long way to go.

Other examined models also exhibit a stark contrast in their reported accuracy on existing benchmarks and performance on N-ImageNet. For example, the test accuracy of the event histogram [32] on N-Cars is 94.5%, and the test accuracy of MatrixLSTM [5] on N-Caltech101 is 86.6%. These models show a validation accuracy of around 30 ~ 50% in N-ImageNet, further supporting the difficulty of N-ImageNet. N-ImageNet is a large-scale, fine-grained benchmark (Table 1) compared to any other existing benchmark and the inherent challenge will foster development in event classifiers that could readily function in the real world.

Assessment on Representations The evaluation of various representations on N-ImageNet allows us to make a systematic assessment of different design choices to handle event-based data. Interestingly, the performance of representations without temporal information (binary event image [37, 7] and event histogram [31]) are superior to representations directly using raw timestamps (timestamp image [41], time surface [24], and event image [57, 61]). The wide variations in raw timestamps deteriorate the generalization capacity of representations that directly utilize this information. This notion is further supported by the fact that the representations using relative timestamps (sorted time surface [1] and DiST) outperform those using raw timestamps.

It should also be noted that our proposed robust representation, DiST, successfully generalizes to large-scale datasets such as N-ImageNet, and shows performance on par with strong learned representations. EST [14] is the best performing model in Table 4, capable of learning highly expressive encodings of event data, thanks to its event aggregation using multilayer perceptrons. The performance of DiST is very close to that of EST, although it does not incorporate any learnable module in its event representation. The suppression of noise from discounting, and the resilience to variations in camera speed from using relative timestamps help DiST to generalize. If we either omit the discount (sorted time surface [1]) or the sorting mechanism (DiT), the performance is inferior to DiST, indicating the importance of the discounting and sorting operations. We further investigate the robustness of DiST in Section 4.2.

Efficacy of N-ImageNet Pre-Training Apart from being a challenging benchmark, the main motivation of N-ImageNet is to provide a large reservoir of event data to pretrain powerful representations for downstream tasks, echoing the role of ImageNet in conventional images. We validate the effectiveness of N-ImageNet pre-training by ob-

Dataset	N-Cars	CIFAR10-DVS	ASL-DVS	N-Caltech101
# of classes	2	10	24	101
Random	90.80	62.57	29.57	68.12
ImageNet	91.48	70.36	53.43	80.88
N-ImageNet	**94.73**	**73.72**	**58.28**	**86.81**

Table 5: Test accuracy of N-ImageNet pretrained models on existing event-based object recognition benchmarks, compared with ImageNet pretraining and random initialization.
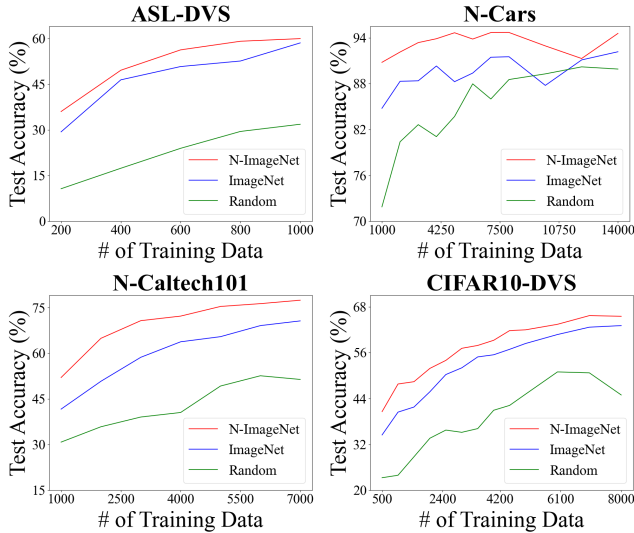

Figure 6: Test accuracy of N-ImageNet pretrained models in resource constrained settings. Each model is trained for 5 epochs with varying amounts of training data.

Factor	Trajectory		Brightness	
Change Amount	Small	Big	Small	Big
Validation Dataset Number	1, 2	3, 4, 5	7, 8	6, 9
MatrixLSTM [5]	33.00	25.62	28.91	23.60
Event Spike Tensor [14]	36.97	32.35	24.89	22.36
Binary Event Image [7]	36.68	31.82	30.94	25.54
Event Histogram [31]	38.72	32.49	33.01	27.72
Event Image [57]	36.52	30.96	32.26	27.04
Time Surface [24]	37.82	33.46	34.19	28.74
HATS [50]	38.95	33.28	33.26	28.22
Timestamp Image [41]	38.31	33.70	33.27	28.04
Sorted Time Surface [1]	38.92	33.69	33.47	28.38
DiT	38.21	33.61	32.66	28.42
DiST	**40.88**	**35.85**	**35.87**	**30.88**

Table 6: Mean accuracy measured on N-ImageNet variants with changes in camera trajectory and brightness.

fine-grained labels of N-ImageNet, which help models to generalize in challenging datasets where numerous labels are present. Furthermore, N-ImageNet pre-trained models outperform its competitors in N-Cars and ASL-DVS, which are recordings of real-world objects. This indicates that although N-ImageNet contains events from monitor displayed images, models pre-trained on it could seamlessly generalize to recognizing real-world objects.

As a practical extension to the previous experiment, we validate the generalization capability of N-ImageNet pre-trained models under resource-constrained settings. We train the same set of models for 5 epochs with initialization schemes from the previous experiment, under varying numbers of training samples. Figure 6 shows that N-ImageNet pre-training incurs a large performance improvement across all four evaluated datasets. The performance gain is further increased when the number of training samples is small. Such results imply that N-ImageNet pre-training provides strong semantic priors that enable object recognition algorithms to quickly adapt to new datasets, even with a few labeled samples.

4.2. Robust Event-Based Object Recognition

Validation Accuracy of N-ImageNet Variants Using the N-ImageNet variants created under various external conditions as described in Section 3.1, we examine the robustness of event-based object recognition algorithms. Table 6 shows the validation accuracy averaged over the trajectory-modified datasets and brightness-modified datasets. All models displayed in Table 4 are evaluated on the N-ImageNet variants. We group datasets according to the variation factor, i.e., brightness and trajectory, and the amount of discrepancy between the original setup and the modified

serving the capacity to generalize in new, unseen datasets. Four standard event camera datasets are used for evaluation: N-Caltech101 [39], N-Cars [50], CIFAR10-DVS [27], and ASL-DVS [4]. For seven event representations from Table 4, ResNet34 [15] is pre-trained on N-ImageNet and compared with ImageNet pre-training and random initialization. The seven representations selected are as follows: binary event image [7], event histogram [31], timestamp image [41], event image [57], time surface [24], sorted time surface [1], and DiST. In experiments explicated below, we report the averaged test accuracy of the seven representations on each dataset. Additional details about the experimental setup are specified in the supplementary material.

Table 5 displays the average test accuracy after training a fixed number of epochs for different initialization schemes. Note that we only use 800 samples from ASL-DVS [4] for training, as using the whole dataset made all model performances saturate near 99%. Networks initiated with N-ImageNet pre-trained weights outperform models from other initialization schemes by a large margin. Notably, the benefits of pre-training intensify as the number of classes in the datasets increases. This could be attributed to the

setup.

All tested models exhibit a consistent deterioration in performance when evaluated on the N-ImageNet variants. Furthermore, the amount of performance degradation intensifies as the amount of environment change increases, as shown in Table 6. These observations imply that many event-based object recognition algorithms are biased on their training setups, and thus fail to fully generalize in challenging, unseen environments. In spite of the consistent performance drop however, DiST outperforms its competitors under all external variations shown in Table 6. Notably, the ablated versions of DiST, i.e. sorted time surface [1], DiT, and timestamp image [41], all perform poorly compared to DiST. Along with the validation accuracy on the original N-ImageNet, this reinforces the necessity of both the discounting and sorting modules of DiST. DiST's capacity to generalize in unseen environmental conditions demonstrates its effectiveness as a robust representation for event-based object recognition.

Representation Consistency To further investigate the robustness of DiST, we quantify the content-wise consistency of various event representations. Seven representations from Table 4 are compared against DiST. Other three representations (MatrixLSTM [5], EST [14], event image [57]) are omitted as they have different number of channels, which may incur unfair comparison. For each representation, we assess the structural similarity index measure (SSIM) between the original representation from N-ImageNet and the representation obtained from N-ImageNet variants. To further elaborate, suppose $\mathcal{E}_{\text{orig}}$ and $\mathcal{E}_{\text{var}}$ are event sequences derived from the same image in the ImageNet validation dataset. Let $\mathbf{R}_{\text{orig}}$ and $\mathbf{R}_{\text{var}}$ be the event representations obtained from $\mathcal{E}_{\text{orig}}$ and $\mathcal{E}_{\text{var}}$ respectively. We report $\text{SSIM}(\mathbf{R}_{\text{orig}}, \mathbf{R}_{\text{var}})$, which measures how consistent each representations are amidst external condition changes.

As displayed in Figure 7, the contents of DiST are more consistent than other competing representations, which can be observed from its highest SSIM value. The contribution of discounting is greater than that of sorting in representation consistency, which can be seen from the SSIM difference of DiT and DiST. However, sorting is crucial for robust object recognition, as can be observed from Table 6, where a clear gap exists between DiT and DiST. Thus, the interplay between discounting and sorting as a whole enhances the robustness of DiST, further leading to improved performance in N-ImageNet variants.

Apart from the robustness of DiST, it must be noted that the N-ImageNet variants serve as the first benchmark for quantifying robustness in event-based classifiers. Although DiST shows a consistent improvement from previous models in robustness, it does not fully recover the original N-
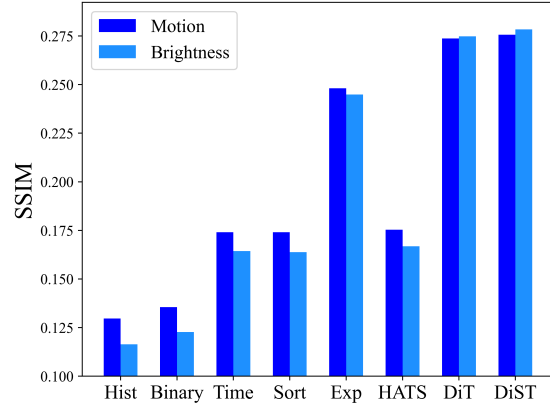

Figure 7: Structural similarity measure (SSIM) between the representations from N-ImageNet and its variants, grouped by changes in motion and brightness. High SSIM indicates that the structure of the representation is stable under external variations. Note that 'Time' and 'Exp' denote timestamp image [41] and time surface [24], respectively.

ImageNet validation accuracy reported in Table 4. We expect the N-ImageNet variants to spur future work in robust representations for event-based object recognition.

5. Conclusion

In this paper, we introduce N-ImageNet, a large-scale dataset for robust, fine-grained object recognition with event cameras. The performance of numerous event-based classifiers on N-ImageNet demonstrates the potential of our dataset as a challenging benchmark. N-ImageNet pretraining is empirically proven to be beneficial, boosting the performance of many existing object recognition algorithms, even under resource-constrained settings. Assessments carried out on the variants of N-ImageNet enable the quantitative evaluation of event classifiers' robustness, and demonstrate the bias present in many event-based object recognition algorithms. As a first step towards remedying such biases, a novel representation, DiST, is proposed that outperforms all the tested models in N-ImageNet variants. We expect N-ImageNet to foster the development of event-based object recognition algorithms that could readily function in real-life applications.

References

[1] I. Alzugaray and M. Chli. Ace: An efficient asynchronous corner tracker for event cameras. In *2018 International Conference on 3D Vision (3DV)*, pages 653–661, 2018. 5, 6, 7, 8

[2] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbrück, M. Flickner, and D. Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388 – 7397, Piscataway, NJ, 2017. IEEE. 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017); Conference Location: Honolulu, HI, USA; Conference Date: July 21-26, 2016. 2

[3] M. Banzi. *Getting Started with Arduino*. Make Books - Imprint of: O'Reilly Media, Sebastopol, CA, ill edition, 2008. 3, 4

[4] Y. Bi, A. Chadha, A. Abbas, , E. Bourtsoulatze, and Y. Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *2019 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2, 3, 4, 7

[5] M. Cannici, M. Ciccone, A. Romanoni, and M. Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision (ECCV)*, August 2020. 2, 4, 5, 6, 7, 8

[6] W. Cheng, H. Luo, W. Yang, L. Yu, S. Chen, and W. Li. Det: A high-resolution dvs dataset for lane extraction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1666–1675, 2019. 2

[7] G. Cohen, S. Afshar, G. Orchard, J. Tapson, R. Benosman, and A. van Schaik. Spatial and temporal downsampling in event-based visual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):5030–5044, 2018. 5, 6, 7

[8] P. de Tournemire, D. Nitti, E. Perot, D. Migliore, and A. Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020. 2

[9] T. Delbruck, Y. Hu, and Z. He. V2e: From video frames to realistic dvs event camera streams. *arXiv preprint arXiv:2006.07722*, 2020. 2, 4

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 6

[11] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. 2004. 2

[12] Y. Feng, H. Lv, H. Liu, Y. Zhang, Y. Xiao, and C. Han. Event density based denoising method for dynamic vision sensor. *Applied Sciences*, 10(6), 2020. 4

[13] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020. 2, 3

[14] D. Gehrig, A. Loquercio, K. Derpanis, and D. Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5632–5642, 2019. 3, 4, 5, 6, 7, 8

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. pages 770–778, 06 2016. 1, 5, 7

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 2, 5

[17] Y. Hu, T. Delbruck, and S.-C. Liu. Learning to exploit multiple vision modalities by using grafted networks, 2020. 3

[18] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in Neuroscience*, 10, 08 2016. 2

[19] S. U. Innocenti, F. Becattini, F. Pernici, and A. D. Bimbo. Temporal binary representation for event-based action recognition. *CoRR*, abs/2010.08946, 2020. 2

[20] D. Jack, F. Maire, S. Denman, and A. Eriksson. Sparse convolutions on continuous domains for point cloud and event stream networks. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 2

[21] A. Khodamoradi and R. Kastner. O(n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Transactions on Emerging Topics in Computing*, pages 1–1, 2017. 4

[22] D. Kostadinov and D. Scaramuzza. Unsupervised feature learning for event data: Direct vs inverse problem formulation, 2020. 3

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012. 1

[24] X. Lagorce, G. Orchard, F. Galluppi, B. Shi, and R. Benosman. Hots: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39, 07 2016. 2, 5, 6, 7, 8

[25] Y. LeCun, C. Cortes, and C. Burges. Mnist hand-written digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 2

[26] J. H. Lee, T. Delbruck, and M. Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience*, 10:508, 2016. 2

[27] H. Li, H. Liu, X. Ji, G. Li, and L. Shi. Cifar10-dvs: An event-stream dataset for object classification. *Frontiers in Neuroscience*, 11:309, 2017. 1, 2, 3, 7

[28] S. Lin, J. Zhang, J. Pan, Z. Jiang, D. Zou, Y. Wang, J. Chen, and J. Ren. *Learning Event-Driven Video Deblurring and Interpolation*, pages 695–710. 11 2020. 2

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1

[30] I. Lungu, F. Corradi, and T. Delbrück. Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In *2017 IEEE International Symposium on Circuits and Systems (IS-CAS)*, pages 1–1, 2017. 2

[31] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. 5, 6, 7

[32] N. Messikommer, D. Gehrig, A. Loquercio, and D. Scaramuzza. Event-based asynchronous sparse convolutional networks. 2020. 2, 6

[33] N. Messikommer, D. Gehrig, A. Loquercio, and D. Scaramuzza. Learning to see in the dark with events. 2020. 2

[34] D. P. Moeys, D. Neil, F. Corradi, E. Kerr, P. Vance, G. Das, S. A. Coleman, T. M. McGinnity, D. Kerr, and T. Delbruck. Pred18: Dataset and further experiments with davis event camera in predator-prey robot chasing. *arXiv preprint arXiv:1807.03128*, 2018. 2

[35] D. Neil, M. Pfeiffer, and S.-C. Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *Advances In Neural Information Processing Systems*, pages 3882–3890, 2016. 2

[36] Y. Nozaki and T. Delbruck. Temperature and parasitic photocurrent effects in dynamic vision sensors. *IEEE Transactions on Electron Devices*, 64(8):3239–3245, 2017. 4

[37] F. C. Ojeda, A. Bisulco, D. Kepple, V. Isler, and D. D. Lee. On-device event filtering with binary neural networks for pedestrian detection using neuromorphic vision sensors. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3084–3088, 2020. 3, 5, 6

[38] G. Orchard, R. Benosman, R. Etienne-Cummings, and N. V. Thakor. A spiking neural network architecture for visual motion estimation. In *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 298–301, 2013. 2

[39] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. 1, 2, 3, 4, 7

[40] G. Orchard, C. Meyer, R. Etienne-Cummings, C. Posch, N. Thakor, and R. Benosman. Hfirst: A temporal approach to object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2028–2040, 2015. 2

[41] P. K. J. Park, B. H. Cho, J. M. Park, K. Lee, H. Y. Kim, H. A. Kang, H. G. Lee, J. Woo, Y. Roh, W. J. Lee, C. Shin, Q. Wang, and H. Ryu. Performance improvement of deep learning based gesture recognition using spatiotemporal demosaicing technique. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1624–1628, 2016. 4, 5, 6, 7, 8

[42] E. Perot, P. de Tournemire, D. Nitti, J. Masci, and A. Sironi. Learning to detect objects with a 1 megapixel event camera. *arXiv preprint arXiv:2009.13436*, 2020. 2, 3

[43] J. A. Pérez-Carrasco, B. Zhao, C. Serrano, B. Acha, T. Serrano-Gotarredona, S. Chen, and B. Linares-Barranco. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward convnets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2706–2719, 2013. 2

[44] B. Ramesh, A. Ussa, L. Della Vedova, H. Yang, and G. Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in Neuroscience*, 14:135, 2020. 2

[45] B. Rueckauer and T. Delbruck. Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor. *Frontiers in Neuroscience*, 10:176, 2016. 2

[46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla,

M. Bernstein, A. C. Berg, and L. Fei-Fei. Im-ageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2, 3

[47] A. Samadzadeh, F. S. T. Far, A. Javadi, A. Nickabadi, and M. H. Chehreghani. Convolutional spiking neural networks for spatio-temporal feature extraction, 2020. 2

[48] C. Scheerlinck, N. Barnes, and R. Mahony. Continuous-time intensity estimation using event cameras. In C. Jawahar, H. Li, G. Mori, and K. Schindler, editors, *Computer Vision – ACCV 2018*, pages 308–324, Cham, 2019. Springer International Publishing. 2

[49] T. Serrano-Gotarredona and B. Linares-Barranco. Poker-dvs and mnist-dvs. their history, how they were made, and other details. *Frontiers in Neuroscience*, 9:481, 2015. 2, 3

[50] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman. HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification. *arXiv preprint arXiv:2018.00186*, June 2018. 1, 2, 3, 5, 6, 7

[51] B. Son, Y. Suh, S. Kim, H. Jung, J. Kim, C. Shin, K. Park, K. Lee, J. Park, J. Woo, Y. Roh, H. Lee, Y. Wang, I. Ovsiannikov, and H. Ryu. 4.1 a 640×480 dynamic vision sensor with a 9μm pixel and 300meps address-event representation. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 66–67, 2017. 3

[52] A. Vasudevan, P. Negri, B. Linares-Barranco, and T. Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 441–448, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. 2

[53] A. Vasudevan, P. Negri, B. Linares-Barranco, and T. Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 441–448, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. 2

[54] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang. Event enhanced high-quality image recovery, 2020. 2

[55] 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, 2020. 2

[56] L. Wang, I. S. M. Mostafavi, Y. Ho, and K. Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10073–10082, 2019. 2

[57] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6351–6360, 2019. 3, 5, 6, 7, 8

[58] Z. Wang, P. Duan, O. Cossairt, A. Katsaggelos, T. Huang, and B. Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[59] J. Wu, C. Ma, X. Yu, and G. Shi. Denoising of event-based sensors with spatial-temporal correlation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4437–4441, 2020. 3

[60] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. 1

[61] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. 2, 3, 5, 6