EventPillars: Pillar-based Efficient Representations for Event Data

Rui Fan[1,2], Weidong Hao[1,2], Juntao Guan[1,2,3], Lai Rui[1,2*], Lin Gu[4,5*], Tong Wu[1,2], Fanhong Zeng[1,2], Zhangming Zhu[1,2]

[1]Key Laboratory of Analog Integrated Circuits and Systems (Ministry of Education)
[2]School of Integrated Circuits, Xidian University, Xi'an 710071, China
[3]Hangzhou Institute of Technology, Xidian University, Hangzhou, China
[4]RIKEN AIP, Tokyo103-0027, Japan
[5]The University of Tokyo, Japan
2017301020002@whu.edu.cn, 24251111481@stu.xidian.edu.cn, guanjuntao@xidian.edu.cn, rlai@mail.xidian.edu.cn, lin.gu@riken.jp, 22111110489@stu.xidian.edu.cn, fhzeng2000@163.com, zhangmingzhu@xidian.edu.cn

Abstract

Event Cameras offer appealing advantages, including power efficiency and ultra-low latency, driving forward advancements in edge applications. In order to leverage mature frame-based algorithms, most approaches typically compute dense, image-like representations from sparse, asynchronous events. However, they are often unable to capture comprehensive information or are computationally intensive, which hinders the edge deployment of event-based vision. Meanwhile, pillar-based paradigms have been proven to be efficient and well-established for dense representations of sparse data. Hence, from a novel pillar-based perspective, we present *EventPillars*, an efficient, comprehensive framework for dense event representations. To summarize, it (i) incorporates the *Temporal Event Range* to describe an intact temporal distribution, (ii) *Activates* the *Event Polarities* to explicitly record the scene dynamics, (iii) enhances the target awareness by a spatial attention prior from *Normalized Event Density*, (iv) can be plug-and-played into different downstream tasks. Extensive experiments show that our EventPillars records a new state-of-the-art precision on object recognition and detection datasets with surprisingly $9.2\times$ and $4.5\times$ lower computation and storage consumption. This brings a new insight into dense event representations and is promising to boost the edge deployment of event-based vision.

Code — https://github.com/Fineshawray/EventPillars.git

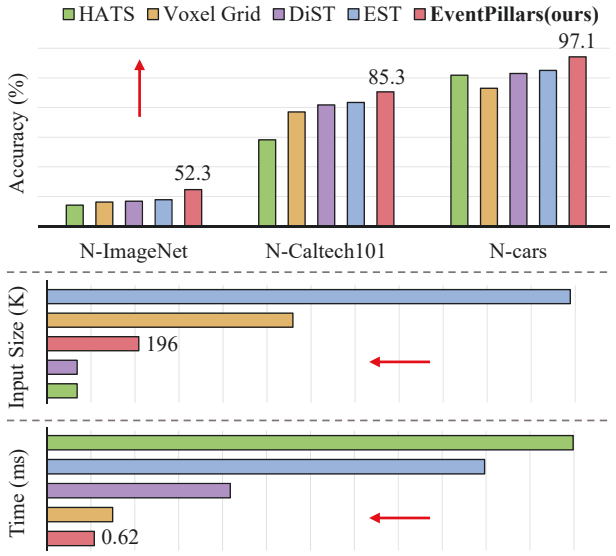

Figure 1: Comparisons of accuracy, input size (storage) and representation time on event-based object recognition given the fixed 224×224 input resolution. Red arrow indicates superior optimization direction.

1 Introduction

Event Cameras are novel, bio-inspired sensors that transmit logarithmic brightness changes termed "events". Unlike conventional frame-based cameras, they offer a unique trade-off between visual information and temporal resolution with minimal bandwidth and power consumption (Hamann et al. 2024), making them highly appealing for edge computer vision, e.g., object recognition and detection. However, by only encoding time, location, and polarity of these changes (Zubić et al. 2023), the stream of events is inherently sparse and asynchronous, thereby holding back their integration with mature frame-based algorithms.

To address this, prior approaches create dense representations from sparse events. Early research (Maqueda et al. 2018; Zhu et al. 2018, 2019) merely encoded one or two aspects of event information, resulting in suboptimal performance. Accordingly, subsequent studies (Gehrig et al. 2019; Kim et al. 2021; Zubić et al. 2023) demonstrated the criticality of comprehensive event information including count, polarity, and temporal information, despite often incurring substantial computational and storage overhead. However, few studies have adequately discussed the optimal trade-off between performance and efficiency of dense event representations. This unresolved tension may constrain the edge deployment of event-based vision.

On the other hand, pillar-based paradigms, a research line of dense representations tailored for sparse data such as point cloud, have shown the effectiveness of encoding explicit feature dimensions into individual channels. Consid-

*Corresponding author.

Representation	Dimensions	Description	Main Characteristics
Event Frame	$H \times W$	Sum polarities	w/o temporal & w/o polarity info.
Event Count	$2 \times H \times W$	Event counts	w/o temporal & implicit polarity info.
Voxel grid	$B \times H \times W$	Sum polarities by temporal bins	w/o polarity & implicit temporal info.
SAE	$2 \times H \times W$	Newest event timestamp	incomplete temporal & implicit polarity info.
HOTS	$2 \times H \times W$	Exponential of newest timestamps	incomplete temporal & implicit polarity info.
HATS	$2 \times H \times W$	Aggregated newest timestamps	w/o temporal & implicit polarity info.
DiST	$2 \times H \times W$	Sorted discounted timestamps	implicit polarity & implicit count info.
Event Image	$4 \times H \times W$	Event counts&newest timestamps	incomplete temporal & implicit polarity info.
EST	$2 \times B \times H \times W$	Sample event point-set into a grid	implicit polarity info. & compute costly
TORE	$2 \times K \times H \times W$	Time-ordered timestamp volumes	w/o count & implicit polarity info.
MDES	$M \times H \times W$	Mixed density event stack	w/o polarity info. & compute costly
ERGO-12	$12 \times H \times W$	Search across representations	implicit polarity info. & compute costly
EventPillars	$4 \times H \times W$	**Pillar-based framework**	**complete, explicit info.; compute efficiently**

Table 1: Comparison of prior image-like representations in event-based vision, ours is in **bold**. $H \times W$ denotes spatial resolution of input, $B = 9$ the number of temporal bins, $K = 3$ the depth and $M = 10$ the stacks number in the original related works.

ering the sparsity of event data, we argue that a similar pattern can be employed to design a *comprehensive* and *efficient* dense event representation framework under strict storage and speed constraints. However, when revisiting event data from a pillar-based perspective, there are still several limitations: (i) existing Newest Event Timestamp (NET) discards the earliest timestamp, resulting in inadequate descriptions for temporal distribution of events, (ii) binarized event polarity differs from the continuous reflectance of point cloud, necessitating a new approach for explicit polarity characterization and (iii) lack of spatial distributional information hinders the differentiation between target and noise.

To overcome these bottlenecks, we propose the Temporal Event Range (TER), which offers a simple but intact description for temporal information in each pillar. Then, we present the Activated Event Polarity (AEP) to efficiently summarize a continuous and explicit polarity information in each pillar. Finally, we devise the Normalized Event Density (NED) as an attention prior to guide target-noise differentiation, supplementing polarity and temporal components with spatial distributional information. By integrating these components, EventPillars achieves a *comprehensive* characterization for event data, while maintaining high computational and storage *efficiency* for edge applications.

We conduct extensive experiments across various tasks and datasets to evaluate our approach. By merely replacing the original representations, EventPillars yields significant improvements across all relevant baselines. For object recognition, accuracy gains of 3.6%, 4.6% and 3.4% are observed on N-Caltech101, N-Cars and N-ImageNet dataset respectively with 9.2× and 4.5× lower computation cost and input size, as shown in Fig. 1. In object detection, the mean average precision (mAP) also improves by 2.7% and 3.8% on Gen1 and 1 Mpx datasets. Notably, we achieve a new detection precision record on Gen1 dataset by **53.1%** mAP, which demonstrates the superiority and robustness of EventPillars. In general, this paper introduces a novel, pillar-based perspective into the study field of dense event representations and shows a promise in facilitating the edge deployment of event-based vision. Our main contributions are:

- We propose the Temporal Event Range, enabling a better description for temporal distribution of events.
- We put forward the continuous Activated Event Polarity to explicitly summary the binarized scene dynamics.
- We present the Normalized Event Density as a spatial attention prior to enhance target-noise differentiation.
- Based on the above representation components, we construct an efficient, comprehensive pillar-based framework termed EventPillars. With only 9.2× and 4.5× lower computation and storage consumption, it sets new precision records on both of recognition and detection benchmarks, which demonstrates its superiority at edge.

2 Related Works

Dense Event Representations Dense event representations involve transforming sparse events into grid-like tensors, making them compatible with frame-based architectures such as convolutional neural networks (CNNs). To start, the Binary Event Image (Cladera et al. 2020; Cohen et al. 2018) solely recorded occurence of events. Later Event Frame (Rebecq, Horstschaefer, and Scaramuzza 2017) aggregated events within a constant temporal window. Although simple and intuitive, it lacks crucial polarity and temporal information. In steering-angle prediction, Event Histogram (Maqueda et al. 2018) segregated events by opposite polarities to preserve polarity information, albeit implicitly encoded within two separate channels. Subsequently, Voxel Grid (Zhu et al. 2019) partitioned constant temporal window into several sub-windows, summing events within each, yet

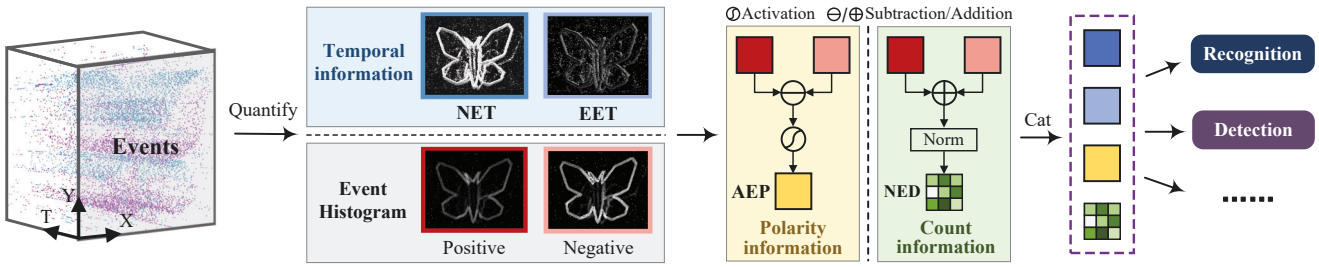

Figure 2: Overview of our EventPillars from raw, sparse events to comprehensive, plug-and-play dense representations.

neglecting polarity information. To summarize, these works share a common reliance on foundational event counts, but lack explicit description for event polarity or temporal information. This deficiency limits their capacity to fully capture the rich information inherent in event data.

Surface of Active Events (Zhu et al. 2018; Benosman et al. 2013) and Event Image (Wang et al. 2019) directly encode the newest event timestamps while disregarding earlier ones. Later, the Histogram of Time Surface (Lagorce et al. 2016) was introduced to capture local spatiotemporal dynamics, yet the computation is time-consuming and performance remains limited. The Time-Ordered Recent Event Volumes (Baldwin et al. 2022) encode timestamps in a form close to Voxel Grid, while at the expense of computational efficiency and without incorporating count information. Mixed-Density Event Stack (Nam et al. 2022) divides events into a sequence of overlapping event stacks to adapt varying scene dynamics, but discards polarity information.

Another line of research have sought to incorporate comprehensive information (count, polarity and time) to enhance performance. To this end, EST (Gehrig et al. 2019) devised an end-to-end learnable representation that samples event points into a grid. However, it retains implicit polarity and involves redundant computation. Subsequently DiST (Kim et al. 2021) employed a discount mechanism to mitigate the event noise in timestamp representations. More recently ERGO-12 (Zubić et al. 2023) conducted extensive hyperparameter searches for optimal components across various dense representations. Despite lacking intuitive interpretability, it achieves state-of-the-art performance on Gen1 dataset. Recent Hyper Histogram (Peng et al. 2023) uniquely stores time, polarity, and count information in 2B channels. This further demonstrates the advantages of using information across all dimensions of event data. Tab. 1 summarizes the main characteristics of these frameworks.

Pillar-based Perspective In the domain of point cloud, pillar-based encoders have been widely adopted for edge vision applications due to their computational efficiency. The early BirdNet (Beltrán et al. 2018) directly encoded point height, intensity and density information into a single image, enabling the transformation of sparse points into structured 2D representations, thus leveraging mature CNN designs on given hardware. Similarly, PIXOR (Yang, Luo, and Urtasun 2018) records height information and computes a 3D data cube. PointPillars (Lang et al. 2019) introduced an efficient encoder that learns features from stacked pillars, which can be scattered back to 2D pseudo-images. More recently, TinyPillarNet (Li, Zhang, and Lai 2023) explicitly records comprehensive information of sparse data. It categorizes the maximum and minimum coordinates on Z-axis and mean reflectance of points within each pillar as intrinsic information, while treating the number of points as distributional information for spatial attention priors. This approach significantly reduces computational overhead while comprehensively summarizing information from sparse points.

Inspired by these works, we propose a dense representation framework for event data from a novel pillar-based perspective. With significant performance improvements, our goal is to boost the edge deployment of event-based vision through the lens of dense event representations.

3 Preliminaries

Event cameras capture pixel-wise log brightness changes by a contrast threshold, each of which is triggered asynchronously and termed an "events". Following the common definition, an event can be described as a tuple $e_i = (x_i, y_i, t_i, p_i)$ triggered at pixel (x_i, y_i) at timestamp t_i with polarity $p_i \in \{-1, 1\}$, where p_i indicates the direction of change. Within a temporal window Δt, an event camera generates a set of events $\mathcal{E} = \{e_i\}_{i=1}^{N}$, where N is the number of triggered events. To build a dense representation $\mathcal{R}$ from the sparse events $\mathcal{E}$, a mapping $\mathcal{M} : \mathcal{E} \longrightarrow \mathcal{R}$ needs to be discovered (Zubić et al. 2023).

Let $\mathcal{N}(x, y) \in \mathcal{E}$ (Kim et al. 2021) denote the events on (x, y), which can also be regarded as events in a 1×1 "pillar" alone the T-axis. Taking a 1-channel image as an example, we then construct a dense representation $\mathcal{R} \in \mathbb{R}^{H \times W}$ through a selected pixel-wise mapping function

$$\mathcal{R}(x, y) = \mathcal{M}(\mathcal{N}(x, y)), y \leq H, x \leq W, \qquad (1)$$

where $H \times W$ is spatial resolution. Optimally, this mapping should strive to preserve the essential event characteristics, namely temporal, polarity and count information. In what follows, we elaborate on specific designs of EventPillars.

4 Innovative Framework: EventPillars

In this section, we present an innovative event representation framework, as illustrated in Fig. 2. Given a set of input events, our goal is to preserve various critical information while encoding them into a dense tensor. In Sec. 4.1, we propose the Temporal Event Range (TER) to capture intact

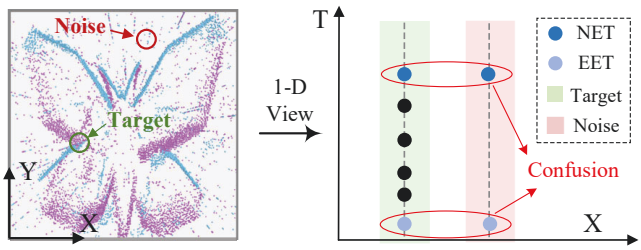

Figure 3: Illustration of intact timestamp and event noise disturbance. 1-dimensional events discarding polarity are shown. NET and EET can summarize the temporal range of events, yet they can not distinguish noise pillar (red) from the target (green) properly without the help of event count.

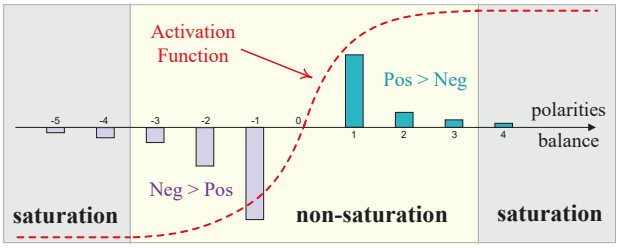

Figure 4: Mechanism of AEP. The balance of polarities from 100 samples of the N-Cars dataset were analyzed. By employing an appropriate continuous activation function, extreme balance of polarities can be filtered out while preserving the majority, thus avoiding the suppression to the principal polarity information.

spatiotemporal structure of asynchronous events. In Sec. 4.2, we devise the Activated Event Polarities (AEP) to emphasize continuous, explicit polarity information. In Sec. 4.3, we propose the Normalized Event Density (NED) to leverage count information as a spatial attention prior for target-noise differentiation. By concatenating the proposed TER, AEP and NED as a comprehensive event tensor encompassing count, polarity, and temporal information, our EventPillars enables plug-and-play and efficient inference with a host of algorithms, exhibiting superior performance.

4.1 Temporal Event Range

Timestamp Image (Zhu et al. 2018; Wang et al. 2019) have proven highly effective in event-based vision. Drawing parallels with pillar-based 2D projection (Lang et al. 2019) in point cloud processing, these pseudo-maps record the T-axis distribution of events in each pillar, akin to the Z-axis in point clouds. TinyPillarNet (Li, Zhang, and Lai 2023) has demonstrated that the height range in each pillar is significant for representing the 3D space occupation of objects. However, existing event representations typically record only the newest event timestamp (Zhu et al. 2018; Benosman et al. 2013), while neglecting earlier occurrences. This will fail to capture the spatiotemporal trajectories of event point cloud, which is crucial for downstream tasks (Peng et al. 2023).

Unlike EST, which samples event point-set, and DiST, which uses a discounted mechanism, we draw inspiration from pillar-based research to introduce the Temporal Event Range (TER) to overcome this limitation. As shown in Fig. 3, this component involves directly storing the newest and earliest event timestamps (NET&EET) disregarding polarity into dedicated feature channels. This enables an efficient, explicit representation of temporal event information without redundant computation.

Specifically, refer the prior definition (Kim et al. 2021), TER computes two 1-channel pseudo-maps $\mathbf{N} \in \mathbb{R}^{H\times W}$ and $\mathbf{E} \in \mathbb{R}^{H\times W}$ by

$$\mathbf{N}(x,y) = \frac{T_{new}(\mathcal{N}(x,y))}{\Delta t}, \tag{2}$$

$$\mathbf{E}(x,y) = \frac{T_{old}(\mathcal{N}(x,y))}{\Delta t}, \tag{3}$$

where T_{new} and T_{old} obtain the newest and earliest timestamp of $\mathcal{N}(x,y)$ in each pillar respectively, Δt denotes the temporal window normalizing timestamps into $[0,1]$. We set TER as the key component for temporal information.

4.2 Activated Event Polarity

Pillar-based point cloud explicitly encoders record the intrinsic reflectance r within each pillar into a single channel, which has proven instrumental in model comprehension of scene information (Beltrán et al. 2018; Lang et al. 2019; Li, Zhang, and Lai 2023). Analogously, event polarity $p \in \{-1, 1\}$ carries the brightness dynamics information. However, unlike continuous reflectance, the binary event polarity possesses entirely different physical properties and data forms. Hence, the effective summarization of polarity information from a pillar-based perspective emerges as a critical challenge for EventPillars.

Existing works either haven't explored polarity (Rebecq, Horstschaefer, and Scaramuzza 2017) or inefficiently handle it via separate channels (Maqueda et al. 2018), leading to suboptimal learning for dynamics. To address this, we propose using the balance of polarities to summarize this binary, intrinsic feature of events in each pillar, rather than merely regard it as a cancellation of opposite events (Maqueda et al. 2018). To this end, we propose the Activated Event Polarity (AEP), while mitigating the disturbance caused by extreme balance of polarities (saturation in Fig. 4).

Specifically, AEP computes a 1-channel pseudo-map $\mathbf{P} \in \mathbb{R}^{H\times W}$ by

$$\mathbf{P}(x,y) = \sigma\left(\frac{1}{C}\cdot\sum_{e_k} p_k\delta(x-x_k, y-y_k)\right), e_k \in \mathcal{N}, \tag{4}$$

where σ denotes a continuous activation function that smooths a meaningful range of polarity differences and filters out extreme, abnormal ones. δ is the Kronecker delta, and C is a factor that enhances the robustness of AEP across varying dynamic scenes and sensors, particularly in response to disparate balance of polarities. It is given by

$$C = \alpha \cdot \max_{\mathcal{N}} \left|\sum_{e_k} p_k\delta(x-x_k, y-y_k)\right|, \mathcal{N} \subset \mathcal{E}, \tag{5}$$

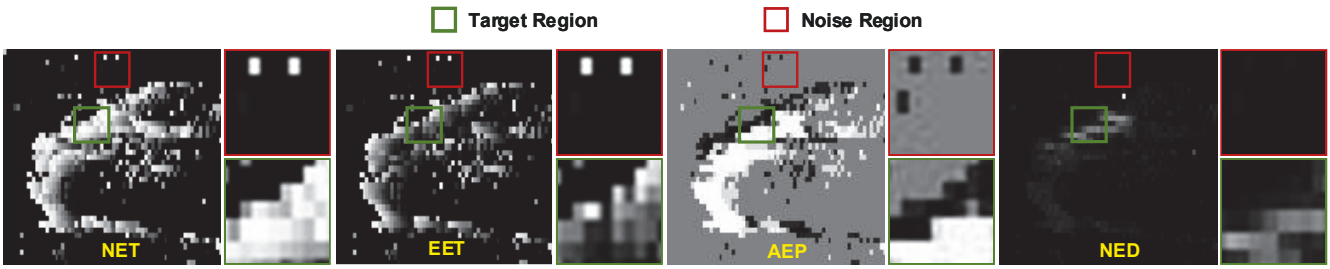

Figure 5: Real event noise in EventPillars. As a distributional prior, NED is less sensitive to typically small-quantity noise.

where $\alpha \in (0, 1]$ is a constant that controls the degree of scaling, and the maximum is taken over the balance of polarities among all event pillars.

In this way, distinct from prior approaches (Maqueda et al. 2018; Zhu et al. 2018), the proposed AEP preserves explicit and adaptive polarity information using only an efficient 1-channel representation. We establish AEP as the component that records intrinsic polarity information for EventPillars.

4.3 Normalized Event Density

In dynamic scenes, edge areas of moving objects typically trigger a higher number of events compared to event noise (Kim et al. 2021), serving as a natural spatial perception cue. Given that the temporal and polarity components do not directly distinguish between noise and targets, as illustrated in Fig. 3 and Fig. 5. Consequently, we propose utilizing the Event Frame (Rebecq, Horstschaefer, and Scaramuzza 2017), which records event quantities, as an attention map containing prior distributional information. This ideally complements other components, helping the model in more effectively differentiating targets and noise.

However, as noted in BirdNet (Beltrán et al. 2018), different sensors and imaging environments may record varying point cloud quantities. This variability can potentially mislead models trained on data collected under specific conditions when generalizing to other scenarios, thereby limiting downstream task performance. We argue that similar issues exist in Event Frame that solely record absolute quantities.

Hence, we introduce the concept of Normalized Event Density (NED). It normalizes the number of events in each pillar by the maximum one, thereby generating a relative distribution prior map on the XY plane. This approach can accommodate disparities in event quantities arising from different devices, sampling rates, or scene velocities.

Specifically, the process of NED can be expressed as

$$\mathbf{D}(x, y) = \frac{1}{M} \cdot \sum_{e_k} \delta(x - x_k, y - y_k), e_k \in \mathcal{N}, \quad (6)$$

where M denotes the maximum events number among all event pillars:

$$M = \max_{\mathcal{N}} |\sum_{e_k} \delta(x - x_k, y - y_k)|, \mathcal{N} \subset \mathcal{E}. \quad (7)$$

Here, $\mathbf{D} \in \mathbb{R}^{H \times W}$ represents the resulting 1-channel pseudo-map. We then incorporate NED as a distributional attention prior into EventPillars, establishing an efficient, comprehensive representation framework in conjunction with other components.

5 Experiments

We conduct extensive experiments to evaluate our proposed framework, and analyze the empirical results.

5.1 Setup

In this section, we introduce the primary baselines and corresponding datasets and show the implementation details.

Tasks & Datasets For object recognition, we employ three widely-used event-based datasets: N-Cars (Sironi et al. 2018), N-Caltech101 (Orchard et al. 2015), and N-ImageNet (Kim et al. 2021). Recorded using an ATIS event camera (Posch, Matolin, and Wohlgenannt 2010), N-Cars comprises 24,029 samples of background and cars, each consisting of events within 100ms temporal window. N-Caltech101, derived from recording the displayed Caltech101 dataset with a moving event camera, expands the number of classes to 100, encompassing 8,246 samples. N-ImageNet is a larger-scale dataset and recorded using Samsung DVS Gen3 (Son et al. 2017) with a similar methodology. We evaluate our EventPillars on these datasets for consistency with relevant baselines. Following prior works (Zubić et al. 2023; Kim et al. 2021), all inputs are resized to 224×224 resolution.

For object detection, we select Gen1 (de Tournemire et al. 2020) and 1 Mpx (Perot et al. 2020) automotive detection datasets to facilitate comparison with relevant baselines. Gen1 dataset, recorded by a 304×240 resolution QVGA event camera, includes 228k car and 28k pedestrian targets, with label frequencies of 1, 2, or 4 Hz. Adhering to existing evaluation metrics (Perot et al. 2020; Zubić et al. 2023; Gehrig and Scaramuzza 2023), we exclude targets with boundary below 10 pixels and diagonal below 30 pixels. The 1 Mpx dataset, recorded in similar scenarios, offers higher spatial resolution (720×1080) and labeling frequency (30 or 60 Hz). It comprises 25 million bounding boxes across three categories: cars, pedestrians, and two-wheelers. We remove targets with boundary below 20 pixels and diagonal below 60 pixels before halving resolution to 360×640 (Gehrig and Scaramuzza 2023).

Implementation Details For object recognition, we use ResNet-34 (He et al. 2016) pre-trained on ImageNet (Deng

Newest Event Timestamp	Earliest Event Timestamp	Acc (%)
✓		89.23
	✓	88.95
✓	✓	**92.86**

Table 2: Ablations of two temporal representations.

Representation	Activation	α	Acc (%)
Event Histogram	-	-	86.54
Activated Event Polarity (ours)	sigmoid	0.2	83.78
		0.4	84.12
		0.6	84.05
		0.8	83.63
	tanh	0.2	85.22
		0.4	**85.57**
		0.6	84.94
		0.8	84.70

Table 3: Test accuracy of AEP and its different setups, compared with Event Histogram that implicitly encode polarity.

et al. 2009) for a fair comparison, modifying the first and last layers to accommodate our framework and the number of classes. We train the network using an ADAM optimizer (Kingma and Ba 2017) with a learning rate of 2e-4 and a batch size of 32, maintaining other experimental setups consistent with prior works (Gehrig et al. 2019; Kim et al. 2021).

For object detection, we adopt optimal YOLOv6 head (Li et al. 2022) with Swin V2 transformer backbone (Liu et al. 2022) from ERGO-12 (Zubić et al. 2023), utilizing pre-trained weights from MS-COCO (Lin et al. 2014). Following ERGO-12, we expand channels from 4 to 12 to ensure aligned temporal windows thus to fairly compare with HOTS, TORE, ERGO-12 and etc in detection. We employ a batch size of 16 and an ADAM optimizer with a learning rate of 1e-4 for both Gen1 and 1Mpx datasets, maintaining other experimental details consistent with ERGO-12.

For a comparative analysis of computational efficiency across various representations, we follow the evaluation metrics proposed in prior work (Gehrig et al. 2019), which records the total time required to process a 100ms sample from the N-Cars dataset. This evaluation was conducted on a CPU (AMD EPYC, 64bits, 2.9GHz of RAM). All representations were implemented based on their publicly available source code, and we utilizes identical core operators, such as `torch.scatter`, to ensure a fair comparison.

5.2 Ablation Studies

In this section, we conduct ablation experiments to analyze our key innovations. All ablation studies are performed on the N-Cars dataset to optimize training cost.

Representation	Normalized	Acc (%)
Event Frame (count)	×	81.32
Normalized Event Density	✓	**86.06**

Table 4: Test accuracy of NED compared with Event Frame that uses count measurement.

NET	EET	AEP	NED	Acc (%)	Representation Time (ms)
✓				89.23	0.19
✓	✓			92.86	0.29
		✓	✓	91.33	0.33
✓	✓	✓		95.92	0.60
✓	✓		✓	95.37	0.31
✓	✓	✓	✓	**97.12**	**0.62**

Table 5: Complementarity of all representations in EventPillars, each of them continues improving the performance.

Temporal Representation We initially evaluate the effectiveness of our temporal component TER that consists of NET and EET. As in Tab. 2, while EET alone slightly underperforms NET, the combination of them significantly enhances accuracy from 89.23% (NET only) to 92.86%. This substantiates our assertion that comprehensive temporal information is essential in event-based deep learning.

Polarity Representation Next, we assess the effectiveness of the polarity information component, AEP. As indicated in Tab. 3, despite AEP's marginally lower performance compared to implicit polarity-recording histograms, this performance gap is understandable given the richer event count information inherent in Event Histogram. Notably, different activation functions and α setups, as mentioned in Sec. 4.2, influence the performance. Through experiments with various activation functions and α setups, we established that $tanh$ activation with $\alpha = 0.4$ provides optimal settings for AEP. Results in Tab. 5 further demonstrate the complementary advantages of explicit AEP with other components.

Density Representation Subsequently, we evaluate the proposed NED and its normalization design described in Sec. 4.3. Tab. 4 illustrates that NED significantly outperforms the unnormalized Event Frame with count measurement by 4.74%, validating the effectiveness of our design.

Comprehensive EventPillars Finally, we test the complementarity of main components constituting EventPillars through various combinations. Tab. 5 reveals that temporal information, as corroborated by numerous previous studies, is highly beneficial for event-based vision. Furthermore, the integration of polarity and temporal information elevates accuracy above 95%, already surpassing our primary comparison methods (Kim et al. 2021; Gehrig et al. 2019). Ultimately, with the introduction of density prior, our final classification performance reaches an impressive 97.12%. This

Representation	Acc (%)		
	N-Caltech101	N-Cars	N-ImageNet
HOTS	21.0	62.4	44.3
Event Histogram	71.3	86.1	47.7
Event Image	81.4	91.3	45.8
HATS	69.1	90.9	47.1
Voxel Grid	78.5	86.5	48.1
DiST	80.9	91.5	48.4
EST	<u>81.7</u>	<u>92.5</u>	<u>48.9</u>
EventPillars	**85.3**	**97.1**	**52.3**

Table 6: Comparisons of representations on N-Caltech101, N-Cars, N-ImageNet event-based recognition dataset, best is in **bold** and the second is <u>underlined</u>.

Representation	Representation Time (ms)	Speed (kEv/s)
HOTS	148.43	27.3
HATS	6.89	832.0
Voxel Grid	<u>0.86</u>	<u>4586.7</u>
DiST	2.40	1655.7
EST	5.73	691.9
EventPillars	**0.62**	**6397.0**

Table 7: Comparisons of representation time that processes a single sample from N-Cars dataset.

demonstrates the exceptional performance of the proposed EventPillars while reiterating the critical importance of preserving comprehensive event information in dense event representations for downstream tasks.

5.3 Benchmarks

Object Recognition Tab. 6 presents evaluation results of various dense representations across three primary event-based recognition datasets. Our EventPillars shows significant performance gains over all baselines (3.6% increase on N-Caltech101, 4.6% on N-Cars, and 3.4% on N-ImageNet), underscoring its performance superiority. Moreover, it utilizes only two additional channels compared to the representations with the fewest channels (e.g., HOTS, Event Histogram, Timestamp Image, HATS and DiST) as shown in Tab. 1, while substantially reducing input storage (4.5× reduction) relative to EST, the second-best performing method. This also demonstrates the high storage efficiency of our EventPillars, laying a solid foundation for the edge deployment of event-based high-level vision.

Object Detection Tab. 8 presents a comparative analysis of various representations in event-based detection. Notably, EventPillars significantly enhances detection performance, surpassing ERGO-12 by 2.7% and 3.8% respectively, which is, to our knowledge, the existing state-of-the-art approach

Representation	Augmentation	mAP (%)	
		Gen1	1 Mpx
HOTS	×	49.0	38.3
TORE	×	43.6	38.1
VoxelGrid	×	39.5	37.5
MDES	×	42.7	37.8
ERGO-12	×	49.3	40.0
ERGO-12	✓	50.4	40.6
EventPillars	×	<u>52.2</u>	<u>42.8</u>
EventPillars	✓	**53.1**	**44.4**

Table 8: Comparisons on Gen1 and 1 Mpx event-based detection dataset. Augmentation is identical with ERGO-12.

on Gen1. This demonstrates EventPillars' robustness and generalization across diverse high-level vision tasks, thereby corroborating the validity of our approach further.

Representation Time Previous works (Gehrig et al. 2019; Gehrig and Scaramuzza 2023) show that representation time (1ms-6ms) matches downstream network inference time (3-5ms), creating a significant bottleneck. As illustrated in Tab. 7, the representation time of our proposed EventPillars is 9.2× and 3.9× lower than the existing state-of-the-art methods EST and DiST. This high representation efficiency of EventPillars is particularly crucial in edge vision application scenarios with real-time requirements.

6 Discussion

One thing to note that our primary focus is on investigating effective event information encoding within a single temporal window. Consequently, we do not incorporate approaches that implicitly express temporal information through temporal bin partitioning, as employed in some works. Furthermore, to maintain consistency with other methods, we have not explored alternative temporal window sizes, which can potentially be a factor influencing the performance of EventPillars. In general, excessively short temporal window may result in insufficient events accumulation, while overly long window risks introducing too early scene dynamics unrelated to the target into EventPillars. This local spatiotemporal consideration of events presents a notable distinction from temporally synchronized point cloud.

7 Conclusion

In this paper, we propose EventPillars, a comprehensive and efficient dense event representation framework. Our work reexamines critical features of events in XYT space from a pillar-based perspective, transforming asynchronous, sparse event streams into synchronous, dense event tensors across polarity, count, and temporal information dimensions. It achieves state-of-the-art accuracy on both recognition and detection with extremely low storage and minimal computation cost, exhibiting plug-and-play applicability across diverse downstream tasks. We hope our efforts will accelerate the deployment of event-based vision in edge applications.

Acknowledgments

This work was supported in part by the National Science and Technology Innovation 2030-Major Projects under Grant 2021ZD0114400, and in part by the Young Scientists Fund of the National Natural Science Foundation of China No.62304162, and in part by the China Postdoctoral Science Foundation under Grant Number 2024M762532, and in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20241313, and in part by the Fundamental Research Funds for the Central Universities under Grant No.XJSJ24090. Dr. Lin Gu was supported by JST Moonshot R&D Grant Number JPMJMS2011 Japan.

References

Baldwin, R. W.; Liu, R.; Almatrafi, M.; Asari, V.; and Hirakawa, K. 2022. Time-ordered recent event (tore) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2519–2532.

Beltrán, J.; Guindel, C.; Moreno, F. M.; Cruzado, D.; Garcia, F.; and De La Escalera, A. 2018. Birdnet: a 3d object detection framework from lidar information. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3517–3523. IEEE.

Benosman, R.; Clercq, C.; Lagorce, X.; Ieng, S.-H.; and Bartolozzi, C. 2013. Event-based visual flow. *IEEE transactions on neural networks and learning systems*, 25(2): 407–417.

Cladera, F.; Bisulco, A.; Kepple, D.; Isler, V.; and Lee, D. D. 2020. On-device event filtering with binary neural networks for pedestrian detection using neuromorphic vision sensors. In *2020 IEEE International Conference on Image Processing (ICIP)*, 3084–3088. IEEE.

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

de Tournemire, P.; Nitti, D.; Perot, E.; Migliore, D.; and Sironi, A. 2020. A Large Scale Event-based Detection Dataset for Automotive. arXiv:2001.08499.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gehrig, D.; Loquercio, A.; Derpanis, K. G.; and Scaramuzza, D. 2019. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5633–5643.

Gehrig, M.; and Scaramuzza, D. 2023. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13884–13893.

Hamann, F.; Ghosh, S.; Martinez, I. J.; Hart, T.; Kacelnik, A.; and Gallego, G. 2024. Low-power Continuous Remote Behavioral Localization with Event Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18612–18621.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kim, J.; Bae, J.; Park, G.; Zhang, D.; and Kim, Y. M. 2021. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2146–2156.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lagorce, X.; Orchard, G.; Galluppi, F.; Shi, B. E.; and Benosman, R. B. 2016. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7): 1346–1359.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; Li, Y.; Zhang, B.; Liang, Y.; Zhou, L.; Xu, X.; Chu, X.; Wei, X.; and Wei, X. 2022. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv:2209.02976.

Li, Y.; Zhang, Y.; and Lai, R. 2023. Tinypillarnet: Tiny pillar-based network for 3d point cloud object detection at edge. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Maqueda, A. I.; Loquercio, A.; Gallego, G.; García, N.; and Scaramuzza, D. 2018. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5419–5427.

Nam, Y.; Mostafavi, M.; Yoon, K.-J.; and Choi, J. 2022. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6114–6123.

Orchard, G.; Jayawant, A.; Cohen, G. K.; and Thakor, N. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9: 437.

Peng, Y.; Zhang, Y.; Xiao, P.; Sun, X.; and Wu, F. 2023. Better and faster: Adaptive event conversion for event-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2056–2064.

Perot, E.; De Tournemire, P.; Nitti, D.; Masci, J.; and Sironi, A. 2020. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33: 16639–16652.

Posch, C.; Matolin, D.; and Wohlgenannt, R. 2010. 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.

Rebecq, H.; Horstschaefer, T.; and Scaramuzza, D. 2017. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization.

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

Son, B.; Suh, Y.; Kim, S.; Jung, H.; Kim, J.-S.; Shin, C.; Park, K.; Lee, K.; Park, J.; Woo, J.; et al. 2017. 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)*, 66–67. IEEE.

Wang, Y.; Du, B.; Shen, Y.; Wu, K.; Zhao, G.; Sun, J.; and Wen, H. 2019. EV-gait: Event-based robust gait recognition using dynamic vision sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6358–6367.

Yang, B.; Luo, W.; and Urtasun, R. 2018. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7652–7660.

Zhu, A.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2018. EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. In *Robotics: Science and Systems XIV*, RSS2018. Robotics: Science and Systems Foundation.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 989–997.

Zubić, N.; Gehrig, D.; Gehrig, M.; and Scaramuzza, D. 2023. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12846–12856.