Object Detection using Event Camera: A MoE Heat Conduction based Detector and A New Benchmark Dataset – Supplementary Material –

Xiao Wang[1], Yu Jin[1], Wentao Wu[2], Wei Zhang[3], Lin Zhu[4], Bo Jiang[1]*, Yonghong Tian[3,5,6]
[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]School of Artificial Intelligence, Anhui University, Hefei, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]Beijing Institute of Technology, Beijing, China
[5]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University, China
[6]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China
{*xiaowang, jiangbo*}*@ahu.edu.cn*, {*jy0x4f, wuwentao0708*}*@163.com*,
zhangwei1213052@126.com, {*linzhu, yhtian*}*@pku.edu.cn*

1. Related Work

• **Benchmark Datasets for Event-based Detection.** Event-based vision has gained significant attention due to its high temporal resolution and its ability to handle challenging conditions such as fast motion and varying lighting. To advance object detection in this domain, several notable datasets have been proposed. The SEVD [1] dataset provides a comprehensive synthetic event-based dataset for both ego-centric and fixed-camera traffic perception, allowing researchers to explore complex traffic monitoring scenarios. Similarly, the eTraM [8] dataset captures real-world traffic scenes using neuromorphic sensors, specifically designed for vehicle detection and tracking in urban environments. In the realm of automotive applications, the Gen1 [3] dataset introduces high-resolution event data recorded from vehicles, enabling precise object detection in high-speed situations and under challenging lighting conditions. Extending these capabilities further, the 1Mpx [7] dataset offers even finer detail for object detection, particularly useful in dynamic and low-light environments. These datasets mark significant advancements in event-based object detection, offering diverse benchmarks that span various domains. Different from these datasets, our proposed EvDET200K dataset provides high-definition event streams captured under different weathers and lightings and involves 10 categories.

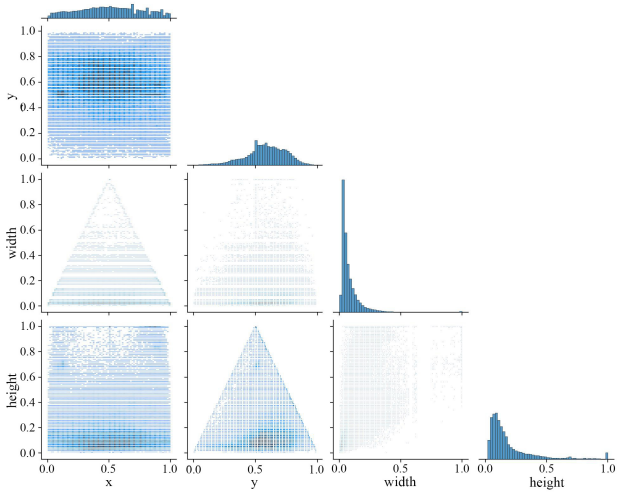

Figure 1. **Illustration of the labels correlogram of our proposed dataset**.

2. EvDET200K Benchmark Dataset

2.1. Detail of EvDET200K dataset

Tab. 1 compares several event datasets for object detection in terms of key characteristics such as sensor type, resolution, dataset scale, number of bounding boxes, dura-

* Corresponding Author: Bo Jiang

tion, number of classes, performance under various weather, lighting and so on. Our EvDET200K (2024) is a new event dataset with significant advantages, including the use of a high-definition Prophesee EVK4-HD sensor (1280×720px resolution), a large dataset scale (10,054 samples and 200K bounding boxes), and comprehensive coverage of various weather conditions (clear and rainy), lighting conditions (daytime and nighttime), and complex scenarios (multi-scene and multi-motion). Additionally, Our focus is on improving the detection of small objects, we plan to capture data from various perspectives, ensuring diversity across a wide range of scenarios. Notably, small objects make up 51% of the dataset, providing sufficient samples to support effective training. Compared to other datasets, EvDET200K stands out for its data diversity, detailed annotations, and adaptability to various tasks.

Fig. 1 illustration of the labels correlogram of our proposed dataset. Each row and column represent different labels, and each cell shows the correlation between those labels. Darker colors in the cells indicate stronger correlations, while lighter colors indicate weaker ones. The diagonal cells show the self-correlation of each label.

Fig. 3 showcases a selection of representative samples from our proposed EvDET200K dataset. These samples highlight the diverse scenarios, object categories, and environmental conditions captured in the dataset. They showcase the dataset's various object categories, lighting variations (daytime and nighttime), and dynamic scenes involving multi-object interactions. This diversity demonstrates the robustness and adaptability of EvDET200K, making it suitable for training models to achieve accurate and reliable event-based object detection.

2.2. Data Collection and Annotation

The EvDET200K dataset is captured using the PROPHESEE EVK4–HD event camera with a resolution of 1280 × 720. During the actual shooting process, we always adhere to the above principles to ensure that our proposed dataset contains rich event data and diverse challenges. We convert each video into five frame images and manually annotate them. For the annotation, we use the XYXY format to store bounding-box coordinates, with each annotation represented as a five-tuple (x_1, y_1, x_2, y_2, cls), where x_1, y_1 denote the top-left corner of bounding-box and x_2, y_2 denote the bottom-right corner, along with the class label cls. The annotations for each video are saved in a JSON file.

3. Experience

3.1. Dataset and Evaluation Metric

In addition to the newly proposed EvDET200K dataset, we also conducted a comparison with several state-of-the-art detectors on the N-Caltech101 [6] dataset to validate the generalization capability of our method. The N-Caltech dataset contains 101 object categories and approximately 9,000 event streams, which are split into training and test sets in an 8:2 ratio. This dataset features complex and variable backgrounds, which present significant challenges for detection algorithms. For evaluation metrics, we used the mean Average Precision (mAP) at different IoU thresholds, the most commonly used metric in object detection. We also report Precision and Recall to assess the accuracy of predictions and the ability to detect positive instances. Additionally, we measured the number of parameters, FLOPs, and FPS for each detector, providing a more comprehensive and accurate understanding of the models' performance.

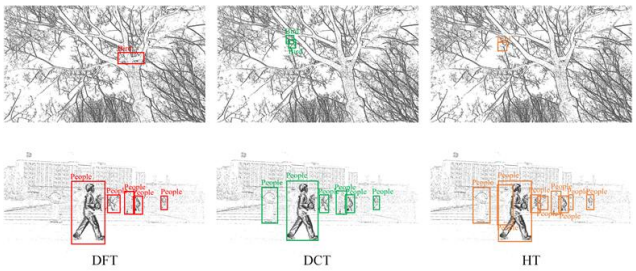

Figure 2. Visualization of the characteristics of different transformations.

3.2. Implementation Details

For training the detector, we set the number of epochs to 80. The model is optimized using the AdamW optimizer with an initial learning rate of 0.001 and weight decay of 0.0001. The batch size is set to 6, and the input image size is 640x640. Our code is implemented in Python using the PyTorch framework, and the experiments are conducted on a server equipped with an AMD EPYC 7542 32-Core Processor CPU and an NVIDIA RTX 4090 GPU. This configuration ensures efficient training and helps the model achieve stable convergence through the use of the AdamW optimizer's adaptive learning rate and regularization.

3.3. Ablation Study

Analysis on frequency of expert triggering across different stages. For the same scene, different transformations yield varying results, as shown in Fig. 2 In this event detection task, DCT is suitable for small targets, HT is suitable for complex scenarios, while DFT is more generalized. Therefore, effectively combining these transformations can achieve more flexible and superior performance. The activation frequency of different experts across different stages is shown in the Tab. 2. It reveals that the model prefers DFT in the shallow stages, and a noticeable divergence can be found in the latter two stages. We believe that DFT has an advantage in extracting generalized features, while the other

Table 1. Comparison of event datasets for object detection. (CL: clear, RA: rainy, DT: daytime, NT: nighttime, MS: multi-scene, MM: multi-motion.)

Dataset	Year	Sensor	Resolution	Scale	Bbox	Duration	Class	Real	Weather		Lighting		Object	
									CL	RA	DT	NT	MS	MM
N-Caltech [6]	2015	Simulator	-	9000	9K	1-10s	101		✓		✓		✓	
SEVD [1]	2024	Simulator	-	-	9M	2-30m	6		✓	✓	✓	✓	✓	
DDD17 [2]	2017	DAVIS 346B	346×260px	36	-	1-50m	7	✓	✓	✓	✓	✓	✓	
DDD20 [5]	2020	DAVIS 346B	346×260px	216	-	1-50m	7	✓	✓		✓	✓	✓	
Gen1 [3]	2020	Prophesee Gen1	304×240px	2357	255K	30-120s	2	✓	✓		✓		✓	
1Mpx [7]	2020	Prophesee Gen2	1280×720px	929	25M	30-120s	3	✓	✓		✓		✓	
DSEC [4]	2021	Prophesee Gen3.1	640×480px	60	390K	1-30m	8	✓	✓		✓	✓		
EvDET200K (Ours)	2024	Prophesee EVK4–HD	1280×720px	10054	200K	2-5s	10	✓	✓	✓	✓	✓	✓	✓

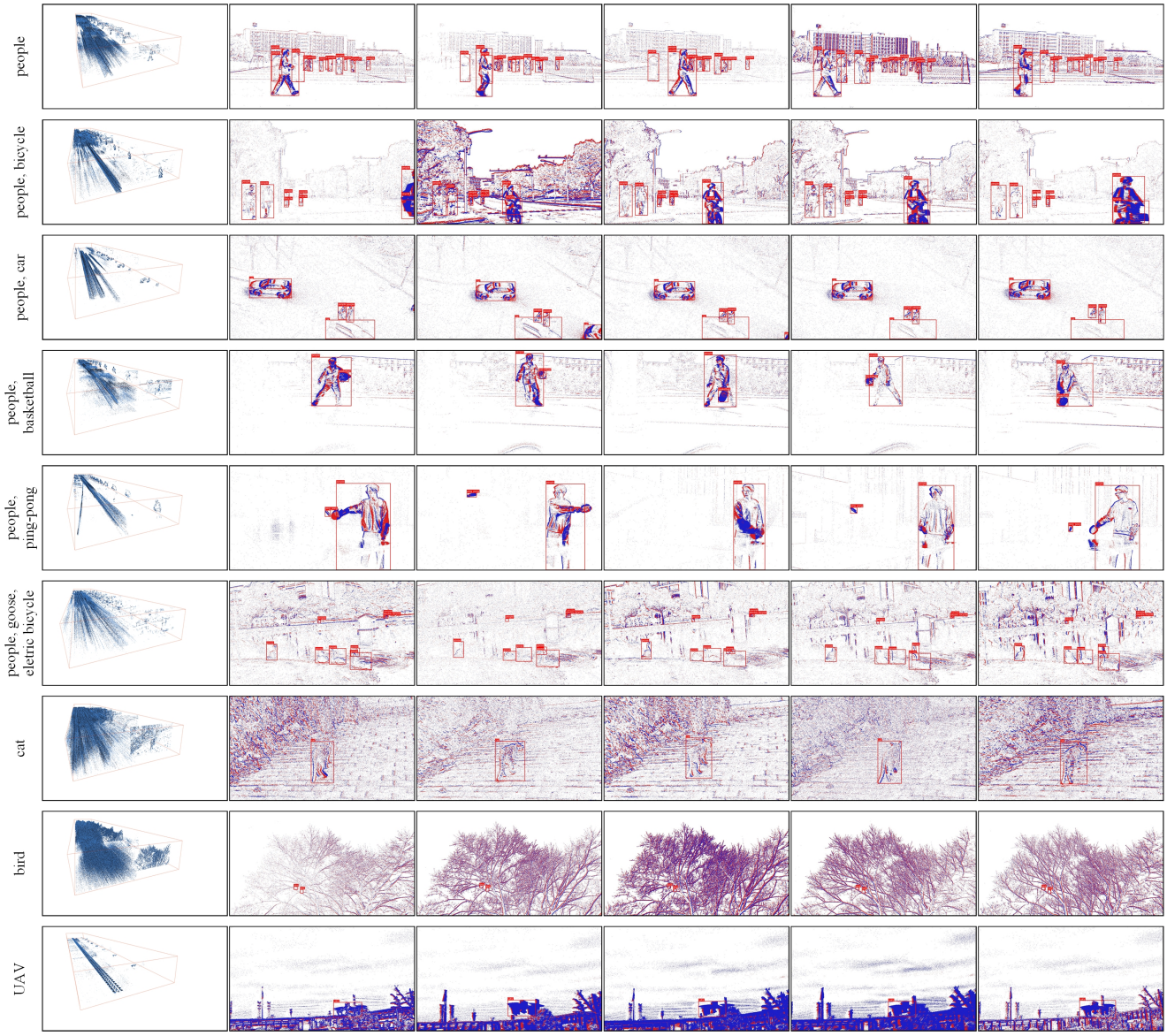

Figure 3. Illustration of some representative samples of our proposed EvDET200K dataset.

Table 2. The frequency of expert triggering across different stages.

	DFT	DCT	HT
Stage 1	43121	4314	2835
Stage 2	40565	6032	3673
Stage 3	24329	17328	8613
Stage 4	28653	16752	4865

Table 3. Ablation studies on different input resolution.

Resolution	**mAP**	**mAP@50**	**mAP@75**	**FLOPs**
256×256px	40.0	69.7	39.5	11.9G
448×448px	49.7	78.4	52.1	36.3G
512×512px	52.0	79.9	54.7	47.2G
640×640px	52.9	80.4	55.8	73.4G

two can better model data with different characteristics.

Adopting mixture of transformations (MoT) for better results mainly due to the following reasons: *1) Superior Signal Representation*: The integration of various transformations offers a richer signal representation. For instance, DCT excels in representing smooth areas in images, while DFT is adept at handling periodic features. *2) Computational Optimization*: The computational efficiency of HT may surpass that of DFT and DCT in certain scenarios. *3) Enhanced Robustness*: Different transformations exhibit varying robustness against different types of errors and noise. *4) Adaptability*: MoT allows for the adaptive selection of the most suitable transformation method based on the characteristics of the signal, thereby enhancing processing flexibility and efficiency.

Analysis on Different Resolutions of Event Stream. In this section, we investigate the impact of event stream resolution on detection performance. We conduct experiments with four different resolutions: 256×256px, 448×448px, 512×512px, and 640×640px. As shown in Tab. 3, we observe that the 256×256px resolution achieves 40.0/69.7/39.5, 448×448px achieves 49.7/78.4/52.1, 512×512px achieves 52.0/79.9/54.7, and 640×640px achieves 52.9/80.4/55.8. Intuitively, higher resolution event streams retain more spatial information, which can positively influence the model's performance. Our model's performance at lower resolutions is not particularly outstanding, but it still achieves strong results at intermediate resolutions, demonstrating its adaptability to scenarios with limited computational resources.

Analysis on Different Channels in Each Stage. Tab. 4 presents the changes in mAP, FLOPs, and parameters at different stages of the model with different channel configurations. We list two MHCO configurations: (2,2,6,2) and (2,2,18,2), each paired with two sets of channel configurations: (64, 128, 256, 512) and (96, 192, 384, 768). Specifically, with the (2,2,6,2) configuration, mAP increases from 50.4 to 52.6; with the (2,2,18,2) configuration, mAP rises from 51.9 to 52.9. This indicates that increasing the number of channels contributes to performance improvement. However, increasing the number of channels also leads to higher computational costs and an increase in the number of parameters. For the (2,2,6,2) configuration, FLOPs increase from 18.0G to 39.2G, and the number of parameters increases from 19.4M to 36.2M. It is important to balance the performance gain with the computational and storage overhead, depending on the specific application and hardware constraints.

Table 4. Ablation studies on different channels in each stage.

Layer	(2,2,6,2)		(2,2,18,2)	
Channel	(64,128,256,512)	(96,192,384,768)	(64,128,256,512)	(96,192,384,768)
mAP	50.4	52.6	51.9	52.9
FLOPs	18.0G	39.2G	33.1G	73.4G
Param	19.4M	36.2M	29.7M	58.7M

4. Visualization

• **Detection Results.** The comparison shown in Fig 4 illustrates the detection results of our proposed MvHeat-DET alongside DERT, SpikeYOLO, and RVT detectors. As seen in the figure, our detector demonstrates strong performance even in dense scenes, whereas the other detectors tend to suffer from missed detections or false positives in such environments.

• **Feature Maps.** Fig. 5 shows representative feature map visualizations of our proposed method on the EvDET200K dataset. It is evident that, even in challenging scenarios, our method is still able to focus on the key detection areas, demonstrating the effectiveness of our model. It also compares the feature maps of our method with the vHeat model. Our model is able to focus on the detection targets effectively from the second stage, while vHeat only begins to focus on the key areas at the fourth stage. This indicates that we can design smaller network architectures to achieve better balance between performance and calculate consumption.

5. Discussion

• **What is role of the $e^{-k(v_x^2+v_y^2)t}$?** We call $e^{-k(v_x^2+v_y^2)t}$ the thermal diffusivity coefficient, which serves as an adaptive filter in the frequency domain to facilitate visual heat conduction. Different frequency values manifest as distinct image patterns: high frequencies correspond to edges and textures, while low frequencies represent flat or smooth regions. By leveraging adaptive thermal diffusivity, MHCO selectively enhances or suppresses these patterns within individual feature channels. By aggregating the processed features across all channels, EvHeat-DET achieves a robust and comprehensive feature representation.

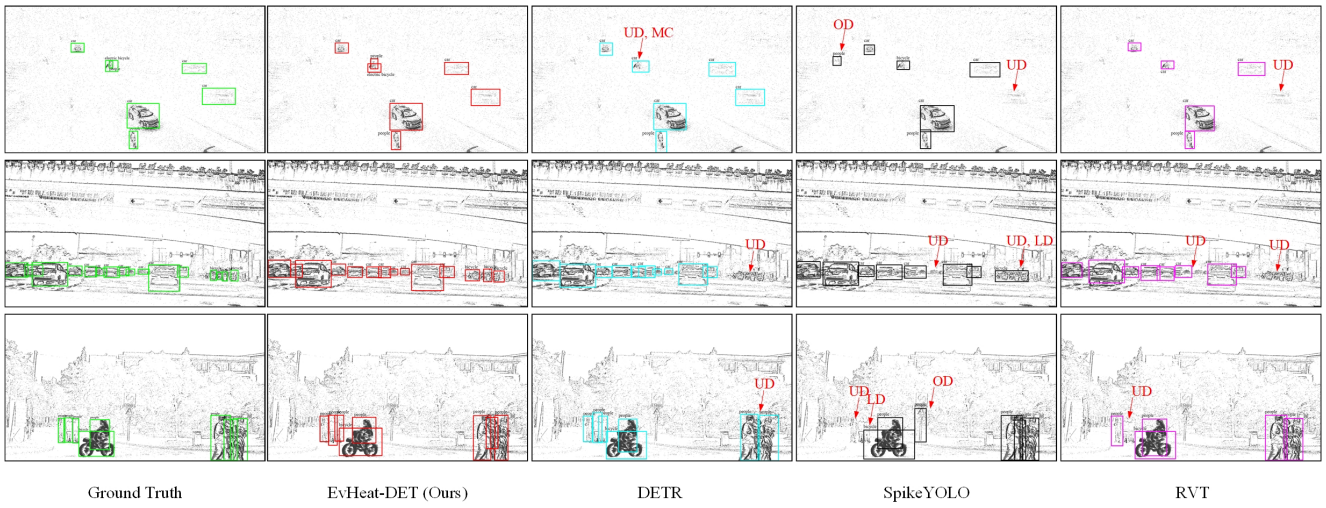

Figure 4. Visualization of the detection results of ours and other detectors. (MC: misclassification, UD: undetected, OD: over-detected, LD: large deviation.)

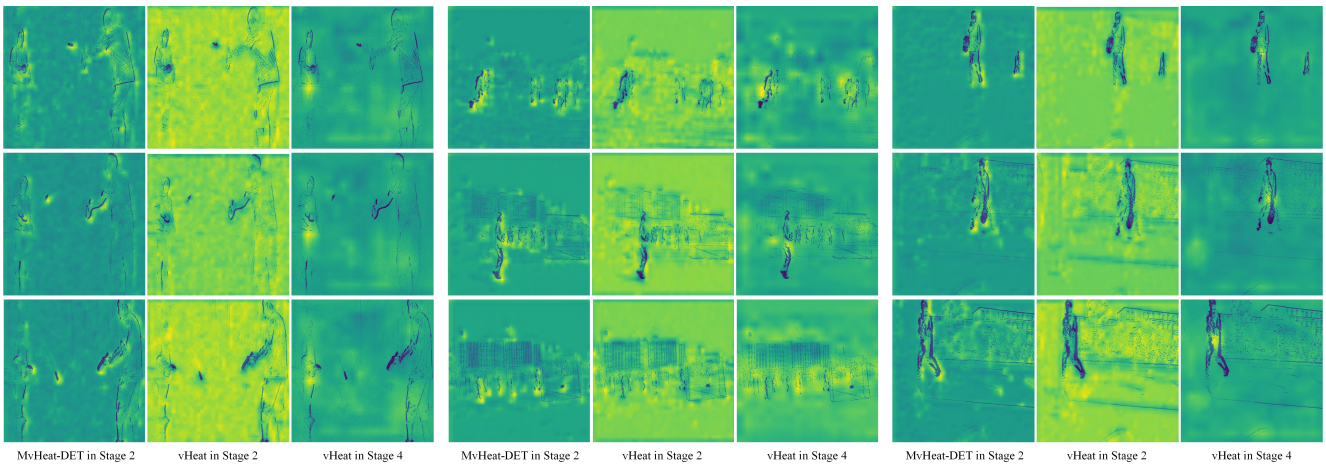

Figure 5. Visualization of the feature maps compared with vHeat.

• **What is the relationship/difference between MHCO and self-attention?** MHCO operates in the frequency domain, enabling it to influence all image patches through frequency filtering. This mechanism dynamically propagates energy via heat conduction, facilitating the perception of global information in the input image. Unlike self-attention, which relies on token similarity, MHCO is a distinctive attention mechanism rooted in the interpretable principles of physical heat conduction. Consequently, MHCO is more efficient than self-attention, as it avoids the computational overhead of evaluating pairwise relevance across all image patches.

• **Why choose MoE?** MoE significantly enhances model capacity by integrating a large number of expert networks. During inference, only a small subset of experts is activated, reducing computational cost. Each expert focuses on specific types of data or task features, demonstrating strong adaptability. In our design, DFT is effective for analyzing periodic signals and provides comprehensive frequency-domain information. DCT concentrates most of the energy in the low-frequency components, which typically contain richer global information in images, offering advantages in small object detection. HT is well-suited for representing sparse signals and excels at extracting edges and textures, features commonly found in small objects that appear near image boundaries. Moreover, HT has low computational complexity, making it suitable for resource-constrained scenarios. By combining these three types of experts, we aim to achieve a balance between detection performance and computational efficiency.

6. Limitation Analysis and Future Works

There is still room for improving our model. For instance, when the model is able to obtain good feature outputs in the shallow layers (as shown in Fig. 5), we can consider reducing the number of layers to decrease the model's complexity and achieve higher detection efficiency. Additionally, the current model does not fully leverage the temporal information. In the future, we will explore the use of 3D heat conduction inference, combined with the rich temporal information in the event stream, to enable more efficient detection.

References

[1] Manideep Reddy Aliminati, Bharatesh Chakravarthi, Aayush Atul Verma, Arpitsinh Vaghela, Hua Wei, Xuesong Zhou, and Yezhou Yang. Sevd: Synthetic event-based vision dataset for ego and fixed traffic perception, 2024.

[2] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset, 2017.

[3] Pierre De Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020.

[4] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

[5] Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2020.

[6] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

[7] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.

[8] Aayush Atul Verma, Bharatesh Chakravarthi, Arpitsinh Vaghela, Hua Wei, and Yezhou Yang. etram: Event-based traffic monitoring dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22637–22646, 2024.