3ET: Efficient Event-based Eye Tracking using a Change-Based ConvLSTM Network

Chen, Qinyu ; Wang, Zuowen ; Liu, Shih Chii; Gao, Chang

DOI
10.1109/BioCAS58349.2023.10389062

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)

2023 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-0026-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/BioCAS58349.2023.10389062

3ET: Efficient Event-based Eye Tracking using a Change-Based ConvLSTM Network

Qinyu Chen[*], *Member IEEE*, Zuowen Wang[*], Shih-Chii Liu[*], *Fellow IEEE*, Chang Gao[†], *Member IEEE*
[*]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland
[†]Department of Microelectronics, Delft University of Technology, Netherlands

***Abstract*—This paper presents a sparse Change-Based Convolutional Long Short-Term Memory (CB-ConvLSTM) model for event-based eye tracking, key for next-generation wearable healthcare technology such as AR/VR headsets. We leverage the benefits of retina-inspired event cameras, namely their low-latency response and sparse output event stream, over traditional frame-based cameras. Our CB-ConvLSTM architecture efficiently extracts spatio-temporal features for pupil tracking from the event stream, outperforming conventional CNN structures. Utilizing a delta-encoded recurrent path enhancing activation sparsity, CB-ConvLSTM reduces arithmetic operations by approximately 4.7× without losing accuracy when tested on a `v2e`-generated event dataset of labeled pupils. This increase in efficiency makes it ideal for real-time eye tracking in resource-constrained devices. The project code and dataset are openly available at https://github.com/qinche106/cb-convlstm-eyetracking.**

***Index Terms*—Pupil tracking, event cameras, sparsity, ConvLSTM, healthcare, AR/VR.**

I. Introduction

The process of eye movements often reveals our mental processes and comprehension of the visual realm. Implementing eye tracking technology offers many possibilities in augmented reality/virtual reality (AR/VR) domains, enabling techniques like foveated rendering to offer a more compelling and efficient visual experience [1], [2]. Eye tracking has potential benefits in wearable healthcare applications. For instance, it can aid in identifying eye movement disorders associated with diseases like Parkinson's or Alzheimer's, thereby enabling early diagnosis and regular assessments [3], [4].

Considering the energy and computational constraints of mobile headsets and the high sampling rate needed for applications like predictive foveated rendering for enhanced VR experiences, the eye tracking system should be lightweight to facilitate low-power and low-latency operation [5]. Eye tracking systems relying on high-speed and high-resolution cameras are costly and power-intensive. Notably, most of the image is redundant in near-eye eye tracking as the only significant movement typically originates from the eye region, with the rest remaining static.

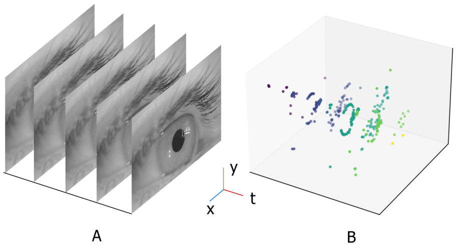

Fig. 1. Comparison between frames and events for the same 53 ms eye movement motion. A) Example video from the LPW dataset [6]. Frames are sampled at a fixed frame rate (95 kHz); B) Using the `v2e` simulator [7], the video frames in A are converted to realistic synthetic DVS event streams. In this example, 5 frames of size 240×180 produce only 310 events.

Dynamic Vision Sensors (DVS) or event-based cameras, which capture brightness changes as sparse events, emerge as a viable solution for near-eye tracking (Fig. 1). The sparsity in the DVS events originating from eye movements or pupil size changes can significantly reduce computational demands compared to conventional camera-based systems. Furthermore, as event-based cameras only record changes in brightness levels, they protect the user's privacy by avoiding the collection of detailed iris data.

Eye tracking is a significant field in computer vision [8]–[10], yet it's relatively unexplored with event cameras due to the scarcity of relevant event-based datasets [11], [12]. Two common approaches guide recent advances in event-based eye tracking algorithms, mirroring those of traditional computer vision: (1) The 3D model-based method locates key points corresponding to the image's geometrical features and fits them to a 3D eye model using optimization techniques. Despite their accuracy, these methods have limitations in resource-limited platforms like headsets, which often require frequent user-specific calibrations. (2) The appearance-based method employs Convolutional Neural Networks (CNNs) to track the eye within the raw image. However, CNNs tend to isolate spatial features, treating input data as independent and often ignoring the potential temporal context in the data sequence. Our study is aligned with the second approach, emphasizing pupil center detection, an integral component of eye tracking techniques.

This paper introduces an Efficient Event-based Eye Tracking (3ET) solution that integrates a ConvLSTM architecture merging convolution operators with recurrent units, thereby enhancing its efficiency in extracting sparse spatio-temporal features from event streams. We also propose a Change-Based ConvLSTM (CB-ConvLSTM) network to alleviate the computational burden. This innovative network introduces high sparsity into

Corresponding Author: Chang Gao

This work was funded by the Swiss National Science Foundation BRIDGE - Proof of Concept Project (40B1-0_213731).

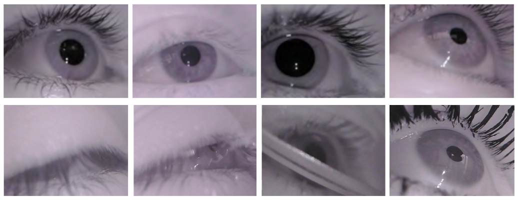

Fig. 2. Diverse set of images from LPW dataset [6]. The first row shows different eye appearances. The second row shows some difficult cases, e.g. eyelid occlusion, glasses occlusion, and heavy makeup.

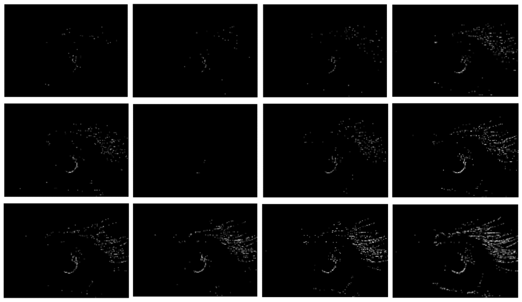

Fig. 3. A set of continuous event-based frames using voxel grid representation from event-based LPW dataset using DVS simulator `v2e` tool.

the process without compromising performance. Experiments demonstrate that our approach yields more than 30% higher accuracy than the CNN-based model while achieving a 4.7× reduction in arithmetic operations without any accuracy loss compared to the standard ConvLSTM-based model when used on an event-based pupil dataset.

II. Methods

A DVS pixel triggers an event in response to a localized change in brightness, yielding a sparse event stream in certain scenes. This inherent sparsity of input data presents opportunities for improved processing speed and heightened computational efficiency within postprocessing algorithms. The primary aim of our work is to devise a cost-efficient algorithm for pupil detection and tracking that is friendly for real-time inference by effectively exploiting the intrinsic sparsity in event data streams. We first outline the ConvLSTM architecture that is implemented to process this temporal data stream. Subsequently, we describe our proposed CB-ConvLSTM architecture, designed specifically to reduce the computational overhead of the network.

A. Dataset: Event-based Dataset for Eye Movement Tracking

Owing to the non-availability of a DVS dataset tailored specifically for eye tracking applications, we resort to `v2e` [7], an open-source DVS simulator, to transform an existing RGB dataset, namely Labeled Pupils in the Wild (LPW) [6], into a synthetic dataset composed of DVS event streams. The LPW dataset includes 66 high-quality videos capturing the eye region, each spanning approximately 20 seconds. Several representative examples are displayed in Fig. 2. To circumvent instances where no events are produced over a prolonged period, we selectively incorporated one-third of the video clips that show above-average speed in eye movements.

The produced event streams, depicted as a sequence of events, are stored as `.h5` files within the `v2e` simulator. Each *event*, marked by index i in an event stream, is expressed as $e_i = (x_i, y_i, t_i, p_i)$, where (x_i, y_i) signifies the pixel location, t_i is the timestamp, and $(p_i = \pm 1)$ represents the polarity or the direction of the brightness change. In this study, we employ the constant time-bin count representation [13], indicated as $V(x, y)$, where (x, y) is the pixel location.

The events occurring within a time window ΔT are translated into a $H \times W$ frame, with H and W representing the height and width of the event-based frame, respectively. This transformation is illustrated in the equation below,

$$V(x, y) = \sum p_i * I(x, y, t_i, x_i, y_i, T1, T1 + \Delta T) \quad (1)$$

where the summation includes all events e_i in the event stream, and $I(x, y, t_i, x_i, y_i, T1, T1 + \Delta T)$ is the indicator function that equals to 1 when $x = x_i$, $y = y_i$ and $T1 < t_i \leq T1 + \Delta T$. The function $V(x, y)$ yields the accumulated value of all events that occur within the time window from $T1$ to $T1+\Delta T$ at each location (x, y).

For the purpose of this research, we establish ΔT at 4.4 ms to ensure synchronicity with the frame rate of the source RGB dataset. This synchronization guarantees the precise alignment of the event frame with the corresponding labels. Within the simulator, the initial 640×480 frames are converted to the 240×180 size output of the DAVIS240 [14], and the resultant DVS output is further resized to 80×60 to reduce network training time. The final synthetic event-based dataset comprises 11k event-based frames derived from the 22 videos. Fig. 3 exhibits a selection of continuous event-based frames from a video sequence.

B. Pupil tracking using ConvLSTM Network on event-based LPW dataset

In previous work, deep CNN models [8], [9] were used to detect the pupil center. The networks performed reasonably well on the RGB dataset. As can be observed from Fig. 2, the pupil in the image has a distinct outline in most cases, which aids pupil detection. However, the model faces challenges when processing the sparse event-based frames depicted in Fig. 3. Many event-based frames lack sufficient events to enable reliable prediction on a per-frame basis. We suggest integrating recurrent structures like LSTM units into the neural network to tackle this issue. Such structures are better suited for interpreting temporal information across the sequence of eye movements in a video, thus improving the prediction accuracy on sparse, event-based frames.

However, the conventional LSTM is computationally expensive because of the all-to-all connections in the architecture. It does not inherently preserve or encode any spatial information. To solve this problem, the ConvLSTM [15] was proposed to additionally capture the spatial dependencies in the data. In this work, the ConvLSTM is used to capture spatial-temporal dependency in the sparse event-based frame stream.

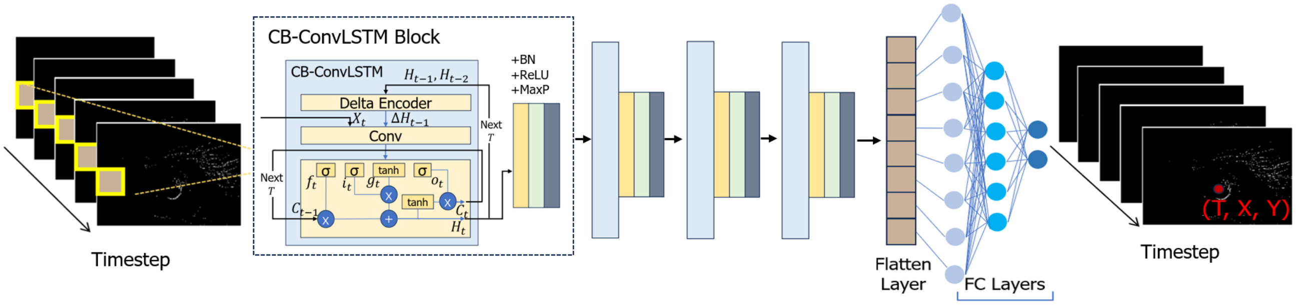

Fig. 4. The pupil tracking network using Change-Based ConvLSTM (CB-ConvLSTM) units on event-based LPW dataset

A ConvLSTM unit is composed of an input gate i, a forget gate f, a cell gate g, an output gate o, and a memory cell state c. Gates i, f, and g control the update of the cell c state. Gate o determines the proportion of cell memory that is transferred to the hidden state output h. The update equations of a ConvLSTM layer are given as:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_{xi} * \mathbf{X}_t + \mathbf{W}_{hi} * \mathbf{H}_{t-1} + \mathbf{b}_i\right) \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_{xf} * \mathbf{X}_t + \mathbf{W}_{hf} * \mathbf{H}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{g}_t &= \tanh\left(\mathbf{W}_{xg} * \mathbf{X}_t + \mathbf{W}_{hg} * \mathbf{H}_{t-1} + \mathbf{b}_g\right) \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_{xo} * \mathbf{X}_t + \mathbf{W}_{ho} * \mathbf{H}_{t-1} + \mathbf{b}_o\right) \\
\mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \\
\mathbf{H}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{C}_t\right)
\end{aligned} \tag{2}$$

where $*$ and $\odot$ signify the convolution and Hadamard functions, respectively. $\mathbf{X}_t$ is the input video frame tensor, $\mathbf{H}_t$ is the hidden state tensor, $\mathbf{C}_t$ is the memory cell tensor, $\mathbf{W}$ denotes weight matrices, and $\mathbf{b}$ denotes bias terms.

The model used for the pupil tracking dataset consists of four ConvLSTM layers and two FC layers, with a total of ~0.42 million parameters. The ConvLSTM layers comprise 8, 16, 32, and 64 hidden nodes, respectively, using a kernel size of 3×3. The outputs of each ConvLSTM layer go through a batch normalization function and a ReLU activation, followed by a max pooling layer for downsampling. The first FC layer has 128 hidden neurons, and the second FC layer generates two outputs corresponding to the pupil center's x and y coordinates. Note that the operations in the FC layers will be executed T times, where T signifies the length of the input sequence in the time dimension.

C. Change-Based ConvLSTM (CB-ConvLSTM) Network for Inducing Activation Sparsity

The inherent sparsity of the input event streams offers a significant opportunity to curtail computational costs related to network processing, as was studied in previous works [16], [17]. An in-depth evaluation of each network stage underscores the fact that within a ConvLSTM block, convolutions are the most computation-demanding modules.

Within every ConvLSTM block, the input consists of two elements: the input tensor $\mathbf{X}_t$ and the hidden state $\mathbf{H}_{t-1}$. The first component, $\mathbf{X}_t$, is naturally sparse, attributed to its origin from event-based frames in the context of the first ConvLSTM layer and due to the application of the ReLU activation in subsequent ConvLSTM layers. In contrast, the second input part, $\mathbf{H}_{t-1}$, is predominantly dense.

In light of this, we introduce Change-Based ConvLSTM (CB-ConvLSTM), aimed at inducing sparsity into hidden states. Instead of employing $\mathbf{H}_{t-1}$, we utilize the change between $\mathbf{H}_{t-1}$ and $\mathbf{H}_{t-2}$ as the recurrent input feature. Additionally, we set a threshold θ for this change to foster increased sparsity. The formulation of the thresholded change $\Delta\mathbf{H}_{t-1}$ is presented below:

$$\Delta\mathbf{H}_{t-1} = \begin{cases} (\mathbf{H}_{t-1} - \mathbf{H}_{t-2}), & (\mathbf{H}_{t-1} - \mathbf{H}_{t-2}) \geq \theta, \\ 0, & (\mathbf{H}_{t-1} - \mathbf{H}_{t-2}) < \theta. \end{cases} \tag{3}$$

The equations of a CB-ConvLSTM unit are shown below:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_{xi} * \mathbf{X_t} + \mathbf{W}_{hi} * \mathbf{\Delta H}_{t-1} + \mathbf{b}_i\right) \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_{xf} * \mathbf{X}_t + \mathbf{W}_{hf} * \mathbf{\Delta H}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{g}_t &= \tanh\left(\mathbf{W}_{xg} * \mathbf{X}_t + \mathbf{W}_{hg} * \mathbf{\Delta H}_{t-1} + \mathbf{b}_g\right) \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_{xo} * \mathbf{X}_t + \mathbf{W}_{ho} * \mathbf{\Delta H}_{t-1} + \mathbf{b}_o\right) \\
\mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \\
\mathbf{H}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{C}_t\right)
\end{aligned} \tag{4}$$

This modification induces a high level of temporal sparsity in convolutions considering that $\Delta\mathbf{H}_{t-1}$ comprises zero values. This approach sets itself apart from preceding change-based networks [18], [19] by exclusively employing a delta encoder for the hidden path, thereby circumventing the necessity to accumulate previous matrix-vector multiplication results. As a result, both computational and memory overhead associated with inducing and leveraging temporal sparsity is significantly reduced. Leveraging the CB-ConvLSTM unit as a base, we construct a network for pupil tracking, maintaining the same topology with the network described in Section II-B as illustrated in Fig. 4.

III. Experimental Setup & Results

The accuracy of pupil detection is assessed by the Euclidean distance between the predicted pupil center and the ground truth. Distances shorter than p pixels are regarded as successful in the estimation of the detection rate [9]. In this context, p_3, p_5, and p_{10} represent p = 3, 5, and 10, respectively.

In the training phase, the Mean Squared Error (MSE) is employed as the loss function, and model parameters are updated using the Stochastic Gradient Descent (SGD) optimizer.

TABLE I
SPARSITY LEVEL IN CB-CONVLSTM AND VANILLA CONVLSTM

θ		CB-ConvLSTM 0.5	 0.2	 0.1	 0	ConvLSTM -
ConvLSTM 1	Inp. sp[1]	0.955	0.955	0.955	0.955	0.955
	Hid. sp	0.999	0.999	0.998	0.733	0.244
ConvLSTM 2	Inp. sp	0.606	0.630	0.636	0.767	0.672
	Hid. sp	0.998	0.994	0.969	0.804	0.179
ConvLSTM 3	Inp. sp	0.596	0.589	0.541	0.735	0.572
	Hid. sp	0.997	0.969	0.928	0.677	0.195
ConvLSTM 4	Inp. sp	0.543	0.424	0.465	0.604	0.500
	Hid. sp	0.989	0.934	0.878	0.623	0.246
	Inp. sp	0.763	0.760	0.754	0.850	0.778
	Hid. sp	0.998	0.988	0.969	0.738	0.216
	Tot. sp	0.924	0.916	0.903	0.750	0.330
Network	**Tot. sp**	0.853	0.845	0.833	0.692	0.304

[1] Inp. sp, Hid. sp, Tot. sp denote the input tensor sparsity, the hidden state tensor sparsity, and the total sparsity.

TABLE II
DETECTION RATE, MODEL PARAMETERS, AND NUMBER OF OPERATIONS OF CB-CONVLSTM, VANILLA CONVLSTM, AND CNN MODELS

θ		CB-ConvLSTM 0.5	 0.2	 0.1	 0	ConvLSTM -	CNN -
Detection rate (%)	p_3	88.50	88.50	88.70	88.88	88.70	57.80
	p_5	96.70	96.76	96.90	97.07	97.10	77.40
	p_{10}	99.20	99.20	99.13	99.50	99.40	91.40
#Parameters (M)		0.42				0.42	0.40
#Flops (M)		9.00	9.49	10.22	18.86	42.61	18.40

A learning rate of 0.001 is maintained over 30 epochs with a batch size of 16. The data allocation for training and validation follows an $80/20$ proportion. We augment the dataset by using a stride of 1 through event-based streams, thus generating multiple overlapping clips from a single event video stream, each displaced by one frame.

Owing to the spatial and temporal sparsity of event streams, many frames often contain minimal information. To address this issue, our eye tracking model integrates a recurrent ConvLSTM structure. As illustrated in Fig. 5, the sequence length in the temporal dimension significantly impacts the detection rate in the ConvLSTM-based model. The detection rates, indicated by p_3, p_5, and p_{10}, raise to 88.8%, 97.0%, and 99.5%, respectively, representing an increase of 17.4%, 15.8%, and 6.9% when the sequence length is increased from 2 to 40. The extension in sequence length aids in acquiring more temporal information, thereby increasing the detection rate, especially when detecting those frames with less information (see Fig. 6).

The proposed CB-ConvLSTM structure was further evaluated for its contribution towards network sparsity and its impact on the detection rate, and a comparison was made with the vanilla ConvLSTM and CNN models. Table I presents the sparsity levels of various layers and components within the proposed CB-ConvLSTM and the standard ConvLSTM. Table II highlights the detection rate, model parameters, and the number of operations performed by the CB-ConvLSTM at different θ values, as well as the standard ConvLSTM and a CNN model.

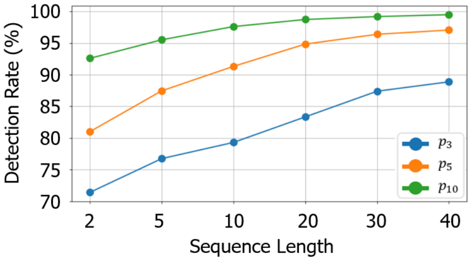

Fig. 5. Detection rate under different sequence length

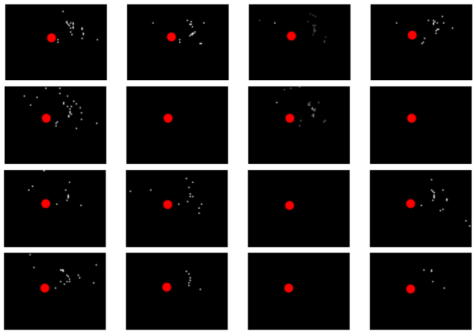

Fig. 6. Examples of 16 continuous event-based frames with predicted results. The predicted result is shown as a red dot. Even if few events are generated, our approach can still detect the location of the pupil.

By increasing θ from 0 to 0.5 to explore a broader design space, the temporal sparsity increases from 69.2% to 85.3% without adversely affecting the model's precision. Compared with the standard ConvLSTM, the CB-ConvLSTM demonstrates about 3× greater sparsity. The comparison CNN model, equipped with similar parameters as the ConvLSTM network and composed primarily of six convolution layers with 32, 32, 32, 64, 64, and 128 filters respectively, yields a sub-optimal detection rate (p_3 = 57.80%).

Regarding computational operations, the CB-ConvLSTM performs 8.8X fewer calculations during convolutions, which leads to a 4.7× reduction in computations for the entire network compared with the standard ConvLSTM and merely half the computational effort demanded by the CNN model.

IV. CONCLUSION

This work proposes a computationally efficient event-based eye tracking solution, specifically targeting pupil center detection. By leveraging the ConvLSTM network, we capture the spatio-temporal sparse features inherent in event streams more effectively. Moreover, to reduce computational demands, we introduce the novel CB-ConvLSTM structure to induce high temporal sparsity in activations to reduce arithmetic operations in convolutions. Results indicate that our approach achieves a network sparsity of 85.3%, facilitating a 4.7× reduction in arithmetic operations, all while maintaining accuracy, relative to the conventional ConvLSTM-based model. CB-ConvLSTM has the potential to be implemented on specialized hardware exploiting spatio-temporal sparsity [19], [20] to further reduce the inference latency and energy cost.

References

[1] A. S. Fernandes, T. S. Murdison, and M. J. Proulx, "Leveling the playing field: A comparative reevaluation of unmodified eye tracking as an input and interaction modality for VR," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2269–2279, 2023.

[2] W. Fuhl, G. Kasneci, and E. Kasneci, "Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2D and 3D segmentations, 2D and 3D landmarks, 3D eyeball, gaze vector, and eye movement types," in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2021, pp. 367–375.

[3] E. Pretegiani and L. M. Optican, "Eye movements in parkinson's disease and inherited parkinsonian syndromes," *Frontiers in Neurology*, vol. 8, p. 592, 2017.

[4] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. L. Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *Proceedings of the 10th ACM Multimedia Systems Conference*, 2019, pp. 255–260.

[5] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein, "Event-based near-eye gaze tracking beyond 10,000 Hz," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2577–2586, 2021.

[6] M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling, "Labeled pupils in the wild: A dataset for studying pupil detection in unconstrained environments," in *Proc. ACM International Symposium on Eye Tracking Research and Applications (ETRA)*, 2016, pp. 139–142.

[7] Y. Hu, S. C. Liu, and T. Delbruck, "v2e: From video frames to realistic DVS events," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021.

[8] K. I. Lee, J. H. Jeon, and B. C. Song, "Deep learning-based pupil center detection for fast and accurate eye tracking system," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 2020, pp. 36–52.

[9] S. Eivazi, T. Santini, A. Keshavarzi, T. Kübler, and A. Mazzei, "Improving real-time CNN-based pupil detection through domain-specific data augmentation," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–6.

[10] J. Liu, J. Chi, W. Hu, and Z. Wang, "3d model-based gaze tracking via iris features with a single camera and a single light source," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 2, pp. 75–86, 2020.

[11] C. Ryan, B. O'Sullivan, A. Elrasad, A. Cahill, J. Lemley, P. Kielty, C. Posch, and E. Perot, "Real-time face & eye tracking and blink detection using event cameras," *Neural Networks*, vol. 141, pp. 87–97, 2021.

[12] Y. Feng, N. Goulding-Hotta, A. Khan, H. Reyserhove, and Y. Zhu, "Real-time gaze tracking with event-driven eye segmentation," in *2022 IEEE on Conference Virtual Reality and 3D User Interfaces (VR)*. Los Alamitos, CA, USA: IEEE Computer Society, mar 2022, pp. 399–408.

[13] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, p. 154–180, jan 2022.

[14] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128× 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[15] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[16] Z. Wang, Y. Hu, and S.-C. Liu, "Exploiting spatial sparsity for event cameras with visual transformers," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 411–415.

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

[18] D. Neil, J. H. Lee, T. Delbruck, and S.-C. Liu, "Delta networks for optimized recurrent network computation," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 2584–2593.

[19] C. Gao, T. Delbruck, and S.-C. Liu, "Spartus: A 9.4 TOp/s FPGA-based LSTM accelerator exploiting spatio-temporal sparsity," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[20] Q. Chen, Y. Huang, R. Sun, W. Song, Z. Lu, Y. Fu, and L. Li, "An efficient accelerator for multiple convolutions from the sparsity perspective," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 6, pp. 1540–1544, 2020.