Supplementary Materials for:
Exploring Joint Embedding Architectures and Data Augmentations for Self-Supervised Representation Learning in Event-Based Vision

Sami Barchid[1] José Mennesson[2] Chaabane Djéraba[1]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL
[2] IMT Nord Europe, Institut Mines-Télécom, Univ. Lille, Centre for Digital Systems
F-59000 Lille, France
[1]{sami.barchid, chaabane.djeraba}@univ-lille.fr
[2] jose.mennesson@imt-nord-europe.fr

A. Visualizations of EDAs

As a complement, we provide visualizations of all investigated EDAs. Figures 1, 2, and 3 illustrate examples of Common, Geometric, and Drop-based EDAs, respectively. The related videos are provided in the supplementary materials as `.gif` animations.

B. Performance Comparison against Fully-Supervised Models

As the proposed approach represents one of the first efforts in the field of event-based SSRL, a comprehensive comparison of the reported performance with prior studies [10, 23] is not feasible. However, we contextualize the results reported in Section 5.2 of the main text by juxtaposing them with those of fully-supervised works from the state-of-the-art. The remainder of this section comments on the reported comparisons.

Table 1 presents the performance comparison on the ASL-DVS dataset [3]. Our pre-trained 2D-CNN and fine-tuned 3D-CNN models achieve competitive performance on the Linear Evaluation Protocol, without the need for supervised finetuning on the ConvEnc. They are only outperformed by ESTF-Net [20], a heavier neural network architecture (a Vision Transformer) with ±46.7M parameters. These results demonstrate that our event-based SSRL framework can learn representations that surpass the capabilities of fully-supervised and more complex models. For instance, our pre-trained 2D-CNN (*i.e.*, a ResNet-18 architecture) achieves an increase of +1.48% in accuracy compared to EST [8], a ResNet-34 model pre-trained on ImageNet [4].

Table 2 provides the performance comparison on the N-Cars dataset [19]. Similarly to our observations for ASL-DVS, our reported ConvEncs on the Transfer Learning Protocol show competitive performance and are only outperformed by a heavier CNN architecture (ResNet-34) trained with EventMix. This suggests the good transferability of the learned representations for event-based object recognition tasks.

Method	Description	Accuracy (%)
RG-CNNs [3]	Graph Neural Network	90.1
EST [8]	Learned event representation + ResNet34 (pretrained on ImageNet [4])	97.9
MVF-Net [5]	Graph Neural Network	97.1
EV-VGCNN [6]	Voxel + Graph CNN	98.3
VMV-GCN [22]	Graph. Neural Network	98.9
AMAE [7]	Adaptive Motion Encoder + ResNet-34 (pretrained on ImageNet [4])	98.4
ESTF-Net [20]	Spatial and Temporal Vision Transformer	**99.9**
Ours	2D-CNN on Linear Protocol (*unsupervised features*)	99.38
Ours	2D-CNN - SemiSup-05%	97.06
Ours	3D-CNN - SemiSup-10%	99.70

Table 1. Performance comparison with fully-supervised methods on ASL-DVS [3].

Table 3 details the performance comparison on the N-Caltech101 dataset [16]. Our approach outperforms numerous fully-supervised models based on graph neural networks [3,5,17], but has lower scores than other works based on CNN architectures. It can be explained by the fact that the proposed event-based SSRL framework cannot learn discriminative features for a large number of categories (*e.g.* 101 classes for N-Caltech101). Still, the results remain encouraging since they are obtained with little to no supervision.

Table 4 provides the comparison of performance on the DVSGesture dataset [1]. The results indicate that the 3D-CNN model fine-tuned on 25% of the training set per-

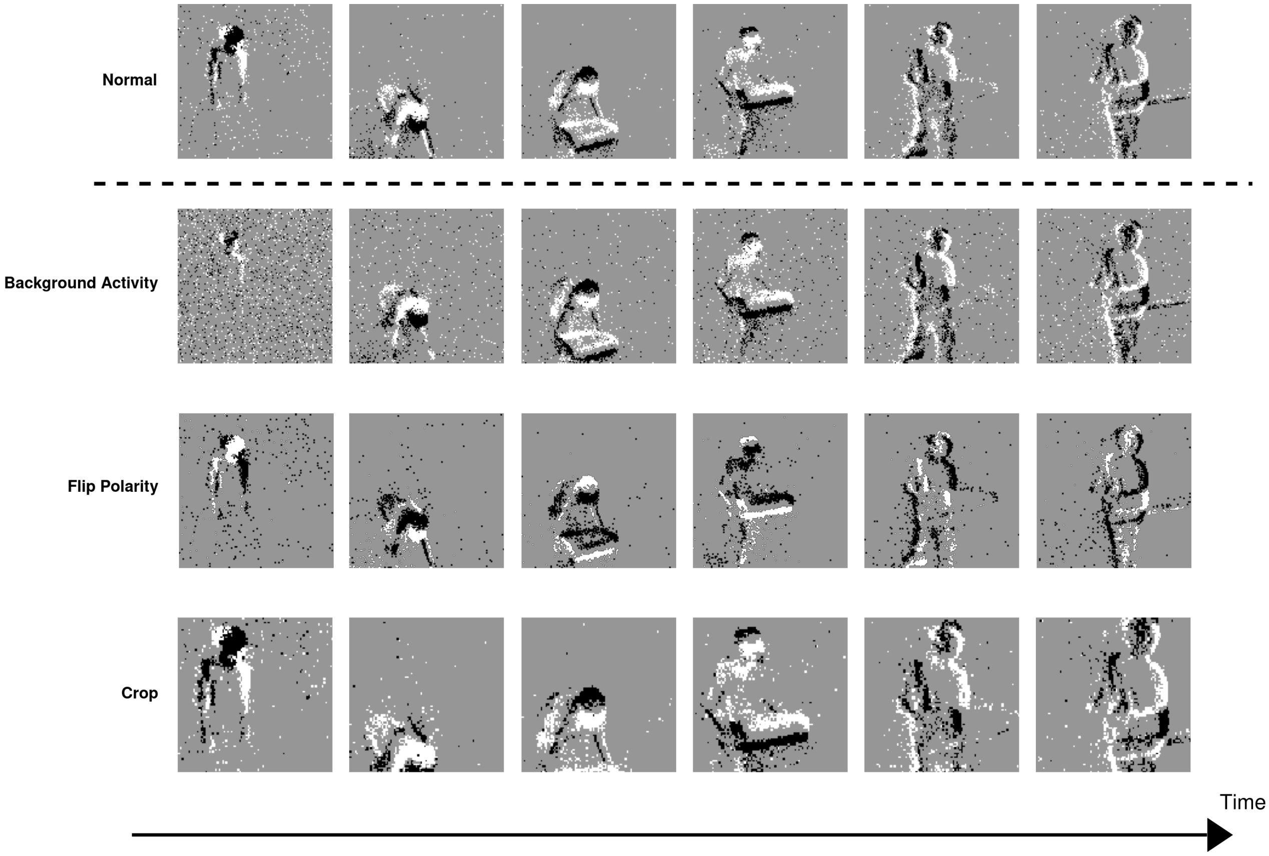

Figure 1. Examples of Common EDAs.

Method	Description	Accuracy (%)
RG-CNNs [3]	Graph Neural Network	91.4
EST [8]	Learned event representation + ResNet34 (pretrained on ImageNet [4])	91.9
Bina-Rep [2]	ResNet-18	92.04
MVF-Net [5]	Graph Neural Network	92.7
AsyNet [15]	Asyn. Sparse VGG13	94.4
EV-VGCNN [6]	Voxel + Graph CNN	95.3
EventMix [18]	ResNet-34 + EDA	**96.54**
Ours	3D-CNN on Transfer Learning Protocol (*features pretrained on ASL-DVS [3]*)	95.64
Ours	2D-CNN on Transfer Learning Protocol (*features pretrained on ASL-DVS [3]*)	94.61
Ours	CSNN$_{3D}$ on Transfer Learning Protocol (*features pretrained on ASL-DVS [3]*)	93.35

Table 2. Performance comparison with fully-supervised methods on N-Cars [19].

Method	Description	Accuracy (%)
RG-CNNs [3]	Graph Neural Network	65.70
AEGNN [17]	Graph Neural Network	66.80
MVF-Net [5]	Graph Neural Network	68.70
AsyNet [15]	Asyn. Sparse VGG13	74.50
VMV-GCN [22]	Graph. Neural Network	77.80
NDA [12]	Spiking ResNet-19 + EDAs	78.00
NDA [12]	Spiking VGG11 + EDAs	**81.70**
Ours	3D-CNN on Linear Protocol (*unsupervised features*)	69.46
Ours	2D-CNN - SemiSup-10%	64.64
Ours	2D-CNN - SemiSup-25%	72.79

Table 3. Performance comparison with fully-supervised methods on N-Caltech101 [16].

forms better than Bina-Rep [2], a ResNet-18 architecture trained with a specific spatiotemporal event representation technique. However, the other fully-supervised approaches achieve superior results compared to our event-based SSRL pretraining. These findings suggest that our approach has some limitations in learning optimal representations for spatiotemporal event-based vision tasks like activity recognition. This may be attributed to the design of our method, where the ConvEncs produce features computed over the entire sequence, which is more effective for extracting spatial information but less so for temporal information.

Table 5 presents the performance comparison on the DailyAction-DVS dataset [14]. Our CSNN$_{2D}$ with features pre-trained on DVSGesture [1] outperforms all previous works without the need for finetuning the ConvEnc.

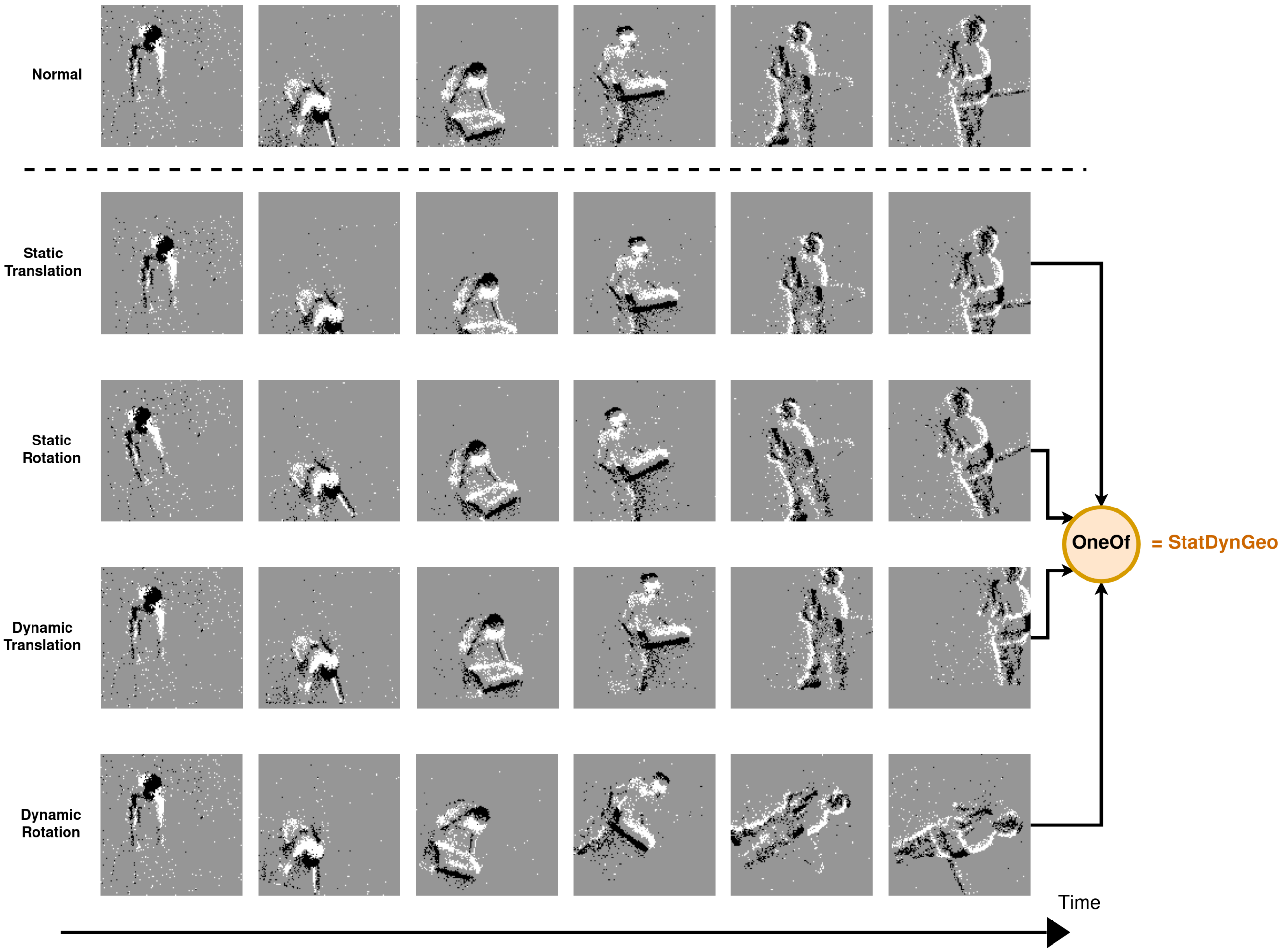

Figure 2. Examples of Geometric EDAs.

Method	Description	Accuracy (%)
Bina-Rep [2]	ResNet-18	87.88
TrueNorth [1]	CSNN (16 layers)	91.77
LIF-Net [9]	CSNN (8 layers)	93.40
Rollout [11]	Spiking VGG16	95.98
EventMix [18]	ResNet-18 + EDA	<u>96.75</u>
TA-Net [24]	CSNN + Temporal Attention	**98.61**
Ours	3D-CNN on Linear Protocol (*unsupervised features*)	89.77
Ours	3D-CNN - SemiSup-10%	81.44
Ours	3D-CNN - SemiSup-25%	90.15

Table 4. Performance comparison with fully-supervised methods on DVSGesture [1].

Method	Description	Accuracy (%)
[21]	SNN with SPA learning	68.30
[13]	SNN with SPA learning	76.90
Motion-based SNN [14]	Motion-sensitive neurons + SNN classifier	<u>90.30</u>
Ours	CSNN$_{2D}$ on Transfer Learning Protocol (*features pretrained on DVSGesture [1]*)	**91.03**

Table 5. Performance comparison with fully-supervised methods on DailyAction-DVS [14].

This highlights the high transferability of our event-based SSRL framework. However, it should be noted that previous works evaluated on this dataset are not deep neural networks like our ConvEnc. Therefore, the difference in complexity must be taken into account when comparing the results.

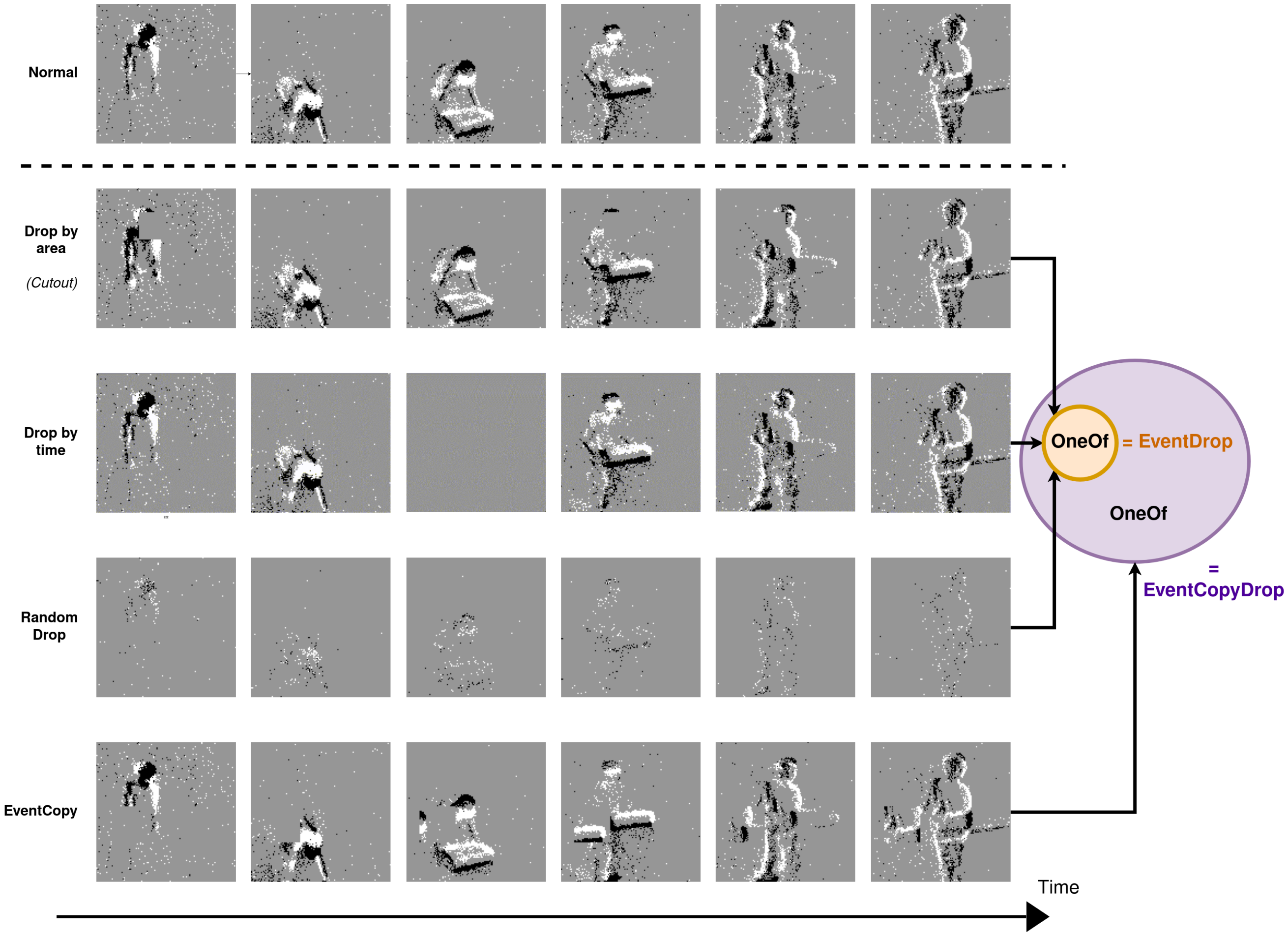

Figure 3. Examples of Drop-based EDAs.

References

[1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017. 1, 2, 3

[2] Sami Barchid, José Mennesson, and Chaabane Djéraba. Bina-rep event frames: A simple and effective representation for event-based cameras. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3998–4002. IEEE, 2022. 2, 3

[3] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020. 1, 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2

[5] Yongjian Deng, Hao Chen, and Youfu Li. Mvf-net: A multi-view fusion network for event-based object classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8275–8284, 2021. 1, 2

[6] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1181, 2022. 1, 2

[7] Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020. 1

[8] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 1, 2

[9] Weihua He, YuJie Wu, Lei Deng, Guoqi Li, Haoyu Wang, Yang Tian, Wei Ding, Wenhui Wang, and Yuan Xie. Comparing snns and rnns on neuromorphic vision datasets: Similarities and differences. *Neural Networks*, 132:108–120, 2020. 3

[10] Simon Klenk, David Bonello, Lukas Koestler, and Daniel Cremers. Masked event modeling: Self-supervised pretraining for event cameras. *arXiv preprint arXiv:2212.10368*, 2022. 1

[11] Alexander Kugele, Thomas Pfeil, Michael Pfeiffer, and Elisabetta Chicca. Efficient processing of spatio-temporal data streams with spiking neural networks. *Frontiers in Neuroscience*, 14:439, 2020. 3

[12] Yuhang Li, Youngeun Kim, Hyoungseob Park, Tamar Geller, and Priyadarshini Panda. Neuromorphic data augmentation for training spiking neural networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 631–649. Springer, 2022. 2

[13] Qianhui Liu, Haibo Ruan, Dong Xing, Huajin Tang, and Gang Pan. Effective aer object classification using segmented probability-maximization learning in spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1308–1315, 2020. 3

[14] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, pages 1743–1749, 2021. 2, 3

[15] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *European Conference on Computer Vision*, pages 415–431. Springer, 2020. 2

[16] 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. 1, 2

[17] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12371–12381, 2022. 1, 2

[18] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Eventmix: An efficient augmentation strategy for event-based data. *arXiv preprint arXiv:2205.12054*, 2022. 2, 3

[19] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018. 1, 2

[20] Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. *arXiv preprint arXiv:2211.09648*, 2022. 1

[21] Rong Xiao, Huajin Tang, Yuhao Ma, Rui Yan, and Garrick Orchard. An event-driven categorization model for aer image sensors using multispike encoding and learning. *IEEE transactions on neural networks and learning systems*, 31(9):3649–3657, 2019. 3

[22] Bochen Xie, Yongjian Deng, Zhanpeng Shao, Hai Liu, and Youfu Li. Vmv-gcn: Volumetric multi-view based graph cnn for event stream classification. *IEEE Robotics and Automation Letters*, 7(2):1976–1983, 2022. 1, 2

[23] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pretraining. *arXiv preprint arXiv:2301.01928*, 2023. 1

[24] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10221–10230, 2021. 3