# Self-Supervised Pre-Training of Spiking Neural Networks by Contrasting Events and Frames

**Raghav Singhal**
Forschungszentrum Jülich
RWTH Aachen

**Jan Finkbeiner**
Forschungszentrum Jülich
RWTH Aachen

**Emre Neftci**
Forschungszentrum Jülich
RWTH Aachen

## Abstract

Artificial Neural Network (ANN) pre-training, followed by fine-tuning, is an established procedure to solve real-world problems where labeled data is scarce. Our work aims to adapt this established procedure to the domain of event-based vision and Spiking Neural Networks (SNNs). Event-based sensors, inspired by the retina, capture visual scenes with low latency and high dynamic range, making them suitable for many real-world vision problems. SNNs, inspired by biological neural networks, when implemented on neuromorphic hardware, enable energy-efficient and low-latency processing, making them well-suited for fully event-based pipelines. However, the lack of sufficiently large labeled datasets hinders the pre-training of SNNs. Here, we leverage joint frame and event data to forego labeling. We achieve this using self-supervised contrastive learning, where an ANN and SNN pair are jointly trained to assimilate (contrast) (un)related frame-event stream pairs. We show that the pre-trained SNN model reaches higher accuracy on several downstream visual classification benchmarks. These results signify that pre-training large-scale SNNs using raw data output from event cameras is possible and paves the way toward foundation SNN models.

## 1 Introduction

Inspired by the brain's remarkable capabilities, neuromorphic systems strive to capture its key mechanisms for low-power, versatile, and fast information processing [1–3]. Similar to the brain, neuromorphic hardware aims to directly leverage the physics of the hardware, leading to a many-fold increase in efficiency, parallelism, and response speed compared to conventional computers. While the algorithms for training such networks have progressed dramatically in the last decade [4–8], their applications still lag behind their conventional ANN counterparts. A key reason for this is the combined lack of suitable datasets, benchmarks, and training hardware, which collectively prohibit the highly practical workflow used in deep learning.

Pre-training is the standard practice in deep learning, thanks to improved deployment speed, generalizability, and robustness. This enables a common backbone network to extract meaningful low-level features for the task at hand. Pre-training relies critically on the availability of large datasets which capture various factors, such as background, noise, lighting, and inherent object variations [9]. This practice has become even more common in recent self-supervised and foundational models, whose training from scratch is often beyond the computational resources of most practitioners [10, 11].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

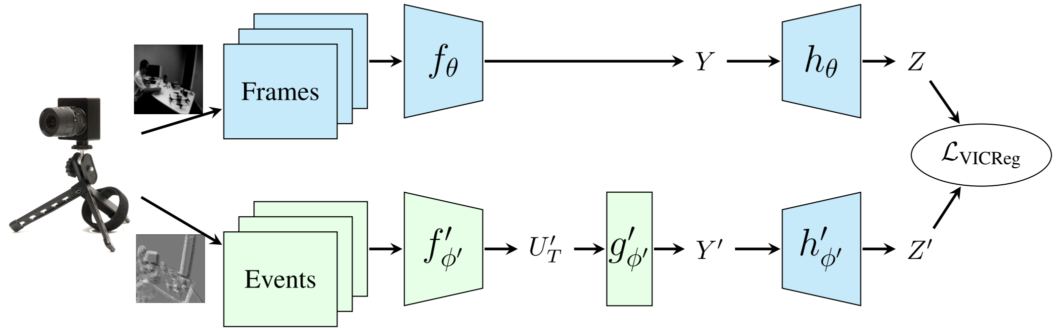

Figure 1: Given a batch of event streams and their corresponding frame-based images, we generate representations for images $Y$ by passing them through an encoder $f_\theta$. In parallel, event streams are encoded via $f_{\phi'}$ to produce a sequence of event representations $U'_T$, which are aggregated using an accumulator $g'_{\phi'}$, yielding representations $Y'$. These representations are fed into expanders $h_\theta$ and $h'_{\phi'}$ to produce embeddings $Z$ and $Z'$, which are jointly optimized using the VICReg loss. $f'_{\phi'}$ and $g'_{\phi'}$ (in green) are spiking in nature, while the other modules (in blue) are non-spiking.

Typical pre-training approaches are difficult to implement in neuromorphic computing due to the lack of labeled datasets on the scale of ImageNet and the higher training cost. While the latter is being addressed by more suitable hardware that emphasize locality and sparsity [12], the dataset labeling is a tedious process in videos and event streams. The lack of a neuromorphic SNN backbone is a critical obstacle to overcome before enabling their use in real-world applications. For example, in automotive traffic applications, the available event-based datasets are small compared to those of frame cameras [13]. This severely limits research in using novel vision sensors in real-world situations. A solid event-based vision backbone can mitigate this limitation by requiring a comparatively small fine-tuning dataset, which may be recorded in a well-controlled and consensual scenario.

Inspired by prior work in contrastive learning, we report here on a self-supervised algorithm to train a SNN on data recorded by neuromorphic vision sensors with simultaneous frame and event output. A majority of recent self-supervised techniques rely on data augmentations to achieve multiple views of the data [14, 15]. These different views are crucial to compute and minimize the similarity (contrast) losses [16–24]. Effective data augmentation techniques are critical in this context, however their application to event-based data has not been explored as extensively or rigorously as it has been for conventional image-based data. On the other hand, multimodal contrastive learning learns by associating pairs of data samples if they refer to the same object or concept. For example, CLIP [10] learns by associating a caption to an image if they are related while contrasting them if they are unrelated. Motivated by this approach, we show that it is possible to train an SNN using contrastive self-supervised learning by associating frame and event outputs to each other if they originate from the same visual sequence and disassociate them otherwise. This result is meaningful because very large datasets can be easily gathered using such cameras and pre-trained to provide a solid backbone for neuromorphic vision applications.

Our approach uses a joint architecture, as depicted in Figure 1, in which one branch is a conventional ANN and the other branch is an equivalent SNN. The ANN part embeds the frame output, while the SNN embeds the event output of the camera. The two branches are trained by minimizing the VICReg loss. The SNN follows Leaky Integrate and Fire (LIF) dynamics which are commonly used in software simulations and many digital and mixed-signal hardware emulations [25–29]. SNNs are compatible with modern ANN architectures such as Convnets [6, 30, 31] and transformers [32]. They can be trained using backpropagation through time using the well-established surrogate gradients approach [8]. We pre-train our model using a diverse corpus of frame-event pairs, synthesized from several open-source datasets, including DSEC [13], Brisbane-Event-VPR [33], DAVIS-DATA [34], MVSEC [35], and TUM-VIE [36]. Fine-tuning on downstream tasks, namely N-CARS [37] and CIFAR10-DVS [38], yields substantial performance improvements, demonstrating the effectiveness of our pre-training strategy. Our results demonstrate performance gains from self-supervised pre-training of SNNs, when fine-tuned on diverse downstream tasks.

## 2 Methods

**Core Methodology.** The overall architecture of our method is shown in Figure 1. We begin by processing a batch of event streams alongside their corresponding frame-based images. This data pairing, while originating from the same visual modality, can be viewed as multimodal since it captures different perspectives of the same sensory input. Image representations, $Y$, are generated by passing the images through an encoder, $f_\theta$. In parallel, event streams are processed by a SNN encoder, $f'_{\phi'}$, resulting in a sequence of spike-based representations, $U'_T$. To integrate these temporal spike representations, $U'_T$ is fed into an accumulator module, $g'_{\phi'}$, modeled as a Leaky Integrate layer with an infinite firing threshold. The final membrane potential, representing the cumulative spike activity across all time steps, forms the representation, $Y'$. Both sets of representations, $Y$ and $Y'$, are subsequently processed through their respective expander networks, $h_\theta$ and $h'_{\phi'}$, to produce the embeddings $Z$ and $Z'$.

These embeddings are optimized jointly using the VICReg loss function [39], ensuring robust alignment between the frame-based and event-based modalities (details about the loss can be found in [39]). VICReg is well-suited for this task as it independently regularizes each branch; it is thereby suitable for multimodal signal alignment tasks and helps to enhance the flexibility of joint-embedding self-supervised learning in such scenarios. During evaluation by fine-tuning on various downstream tasks, we only use the pre-trained SNN encoder, $f'_{\phi'}$, and discard the other modules.

**Pre-Training Dataset Processing.** The diversity and quality of data used during pre-training are critical for achieving robust model performance. We pre-train our model on a diverse corpus of frame-event pairs drawn from several event camera datasets, including DSEC [13], Brisbane-Event-VPR [33], DAVIS [34], MVSEC [35], and TUM-VIE [36]. Importantly, no ground truth labels are required, as our approach solely relies on the alignment between frames and event streams. Given the varying frame sampling rates across datasets, careful preprocessing is essential. For each dataset, we extract the last frame within each sampling period and aggregate the corresponding event stream into 10 time bins, ensuring consistency across the diverse sources. Through this process, we collected a total of 181,207 event-frame pairs across all datasets. A summary of the datasets is given in Table 1. DAVIS-DATA [34] includes 25 datasets featuring real-world event stream recordings alongside their corresponding greyscale frame-based images, encompassing scenarios such as office environments, urban settings, and simple objects and textures. The DSEC [13] and Brisbane-Event-VPR [33] datasets focus on event-based driving, capturing footage under a range of challenging lighting conditions. TUM-VIE [36] contains a diverse collection of handheld and head-mounted recordings in both indoor and outdoor contexts, including high dynamic range scenes. MVSEC [35] comprises recordings captured in various scenarios, including footage from a hexacopter, a car, and a motorcycle, all under different illumination levels and environmental conditions.

| Dataset | Image Frames | Sampling Period | Total Samples |
|---|---|---|---|
| DSEC [13] | RGB | 50 ms | 63,871 |
| Brisbane-Event-VPR [33] | RGB | 25 ms | 28,915 |
| DAVIS-DATA [34] | Grayscale | 46 ms | 25,825 |
| MVSEC [35] | Grayscale | 30 ms | 4,910 |
| TUM-VIE [36] | Grayscale | 50 ms | 57,686 |

Table 1: Overview of datasets used for pre-training. **Total Samples** denotes the total number of event-frame pairs we collected for each dataset.

We expect that the diverse scenes represented in this collection of datasets will promote improved knowledge transfer during the subsequent fine-tuning process on other datasets.

**SNN Details.** The spiking neurons in our SNN follow the LIF model dynamics [40]. The dynamics are described by the following discrete-time equations:

$$S[t] = \Theta(H[t] - V_{th}) \tag{1}$$

$$H[t] = V[t-1] + \beta(X[t] - (V[t-1] - V_{reset})) \tag{2}$$

$$V[t] = H[t]\ (1 - S[t]) + V_{reset}\ S[t] \tag{3}$$

Here, $X[t]$ is the input current at time step $t$, $H[t]$ represents the membrane potential before spiking, and $V[t]$ is the membrane potential after a spike occurs. The neuron fires when $H[t]$ crosses the threshold $V_{th}$, which is determined by the Heaviside step function $\Theta(x)$. The spike output at time step $t$ is denoted by $S[t]$, while $V_{reset}$ is the reset potential after a spike. The constant $\beta$ governs the membrane decay. Since the step function is non-differentiable, we apply the surrogate gradient method [8], to approximate the derivative of the function, using the arctan surrogate function [41].

**Implementation Details.** We use PyTorch and SnnTorch [42] to simulate spiking neurons for training and Tonic [43] for dataset processing. The ANN and SNN encoders are the ResNet-18 model [44] and its adapted spiking variant [45], respectively. In the spiking ResNet-18 model, we replace all ReLU activations with leaky integrate-and-fire (LIF) modules and substitute max-pooling operations with average pooling layers. The membrane decay potential $\beta$ is set to 0.5, and the reset potential $V_{reset}$ to 0. Both the ANN and SNN encoder networks consist of 512 output units. The expander networks comprise two-layer fully-connected models with batch normalization and ReLU activations, where each expander layer has 4096 units, as larger expander dimension sizes have been shown to enhance performance [46]. Notably, neither the encoder nor the expander modules share weights. We resize each event and frame data sample to a spatial resolution of $224 \times 224$ pixels. Following the training protocols established by BYOL [47], Barlow Twins [46], and VICReg [39], we jointly train the models using the LARS optimizer [48], with a learning rate defined as $lr = base\_lr \times {}^{batch\_size}/_{256}$, where $base\_lr$ is set to 0.2 and the batch size is 128. The learning rate follows a cosine decay schedule [49], starting from 0 with 10 warm-up epochs, ultimately reaching a final value of 0.002. We use the same VICReg loss regularization coefficients as used in the original paper [39], and pre-train the models for 200 epochs.

**Evaluation on Downstream Tasks.** We evaluate our pre-trained SNN model on two downstream benchmark event-based datasets for visual classification: CIFAR10-DVS [38] and N-CARS [37]. We resize the spatial resolution of all samples to 128 × 128 pixels across both datasets. CIFAR10-DVS, derived from the original CIFAR10 dataset, consists of 10,000 DVS recordings. We use a 9:1 train-validation split, allocating 9,000 samples for training and 1,000 for validation. N-CARS, a real-world event-based dataset for binary object classification, comprises 15,422 training samples (7,940 cars, 7,482 background) and 8,607 test samples (4,396 cars, 4,211 background). For both datasets, we bin events into 10 time steps and set the membrane potential decay factor $\beta$ to 0.5. We optimize our models using Adam [50] with an initial learning rate of 0.001, following a cosine decay schedule [49], and a weight decay of $10^{-4}$. The results are reported as the average over 3 different runs.

## 3 Results and Conclusions

| **Train Set %** | **w/o. PT** | **w. PT** | $\uparrow \boldsymbol{\Delta}$ |
|---|---|---|---|
| 1 | 12.12 | 21.27 | **9.15** |
| 10 | 35.03 | 39.25 | **4.22** |
| 50 | 46.91 | 56.47 | **9.56** |
| 100 | 60.02 | 64.13 | **4.11** |

(a) CIFAR10-DVS

| **Train Set %** | **w/o. PT** | **w. PT** | $\uparrow \boldsymbol{\Delta}$ |
|---|---|---|---|
| 1 | 62.09 | 78.41 | **16.32** |
| 5 | 79.63 | 83.58 | **3.95** |
| 50 | 90.83 | 91.89 | **1.06** |
| 100 | 92.38 | 92.91 | **0.53** |

(b) N-CARS

Table 2: Transfer learning on downstream tasks. Test accuracy achieved through full fine-tuning (supervised training) of the spiking ResNet-18 model after training on varying percentages of the CIFAR10-DVS and N-CARS training sets, with a linear classifier appended. **w/o. PT** and **w. PT** refer to networks initialized without and with the proposed pre-training scheme, respectively. $\uparrow \boldsymbol{\Delta}$ represents the gain from utilizing the pre-training method over no pre-training.

Table 2 presents the performance of the spiking ResNet-18 model when fine-tuned on varying percentages of the CIFAR10-DVS and N-CARS training sets, with a linear classifier appended. Our proposed pre-training scheme consistently leads to a significant improvement in performance compared to the baseline without pre-training (fully supervised training). This improvement is especially notable when fine-tuning on smaller subsets of the training data, demonstrating the intended effectiveness of our method in low-data regimes. Our results show successful demonstration of self-supervised transfer learning in SNNs, with substantial gains observed on downstream datasets that were not part of the original training set. This highlights the model's capacity to transfer knowledge and leverage representations learned during pre-training to enhance performance on unseen tasks. Additionally, the success of this approach opens new avenues for using SNN architectures in neuromorphic vision applications, laying a strong foundation for the development of versatile and generalizable SNN backbones. Our method also mitigates the challenges associated with limited exploration of data augmentation techniques for event-based data by simply enriching the event data with already existing frame-based images.

# References

[1] C. Mead. Neuromorphic electronic systems. *Proceedings of the IEEE*, 78(10):162936, 1990.

[2] G. Indiveri, B. LinaresBarranco, T.J. Hamilton, A. van Schaik, R. EtienneCummings, T. Delbruck, S.C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. SerranoGotarredona, J. Wijekoon, Y. Wang, and K. Boahen. Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience*, 5:123, 2011.

[3] Mike Davies. Benchmarks for progress in neuromorphic computing. *Nature Machine Intelligence*, 1(9):386388, 2019.

[4] Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural computation*, 30(6):15141541, 2018.

[5] Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass. Long shortterm memory and learningtolearn in networks of spiking neurons. *arXiv preprint arXiv:1803.09574*, 2018.

[6] Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424, 2020.

[7] Timo C Wunderlich and Christian Pehle. Eventbased backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1):117, 2021.

[8] E. O. Neftci, H. Mostafa, and F. Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradientbased optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):5163, Nov 2019.

[9] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, page 33203328, 2014.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[11] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive languageimage learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 28182829, 2023.

[12] Jan Finkbeiner, Thomas Gmeinder, Mark Pupilli, Alexander Titterton, and Emre Neftci. Harnessing manycore processors with distributed memory for accelerated training of sparse and recurrent models, 2023.

[13] 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.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[15] Sami Barchid, José Mennesson, and Chaabane Djéraba. Exploring joint embedding architectures and data augmentations for self-supervised representation learning in event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3902–3911, 2023.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[17] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[20] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

[21] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. pages 9630–9640, 10 2021.

[23] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[24] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022.

[25] E. Chicca, F. Stefanini, and G. Indiveri. Neuromorphic electronic circuits for building autonomous cognitive systems. *Proceedings of IEEE*, 2013.

[26] Paul A Merolla, John V Arthur, Rodrigo AlvarezIcaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spikingneuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668673, 2014.

[27] Simon Friedmann, Johannes Schemmel, Andreas Grübl, Andreas Hartel, Matthias Hock, and Karlheinz Meier. Demonstrating hybrid learning in a flexible neuromorphic hardware system. *IEEE transactions on biomedical circuits and systems*, 11(1):128142, 2017.

[28] M. Davies, N. Srinivasa, T. H. Lin, G. Chinya, P. Joshi, A. Lines, A. Wild, and H. Wang. Loihi: A neuromorphic manycore processor with onchip learning. *IEEE Micro*, PP(99):11, 2018.

[29] Charlotte Frenkel and Giacomo Indiveri. Reckon: A 28nm sub-mm2 task-agnostic spiking recurrent neural network processor enabling on-chip learning over second-long timescales. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 65, pages 1–3. IEEE, 2022.

[30] Saeed Reza Kheradpisheh and Timothée Masquelier. S4nn: temporal backpropagation for spiking neural networks with one spike per neuron. *arXiv preprint arXiv:1910.09495*, 2019.

[31] Loïc Cordone, Benoît Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[32] Rui-Jie Zhu, Qihang Zhao, and Jason K Eshraghian. Spikegpt: Generative pre-trained language model with spiking neural networks. *arXiv preprint arXiv:2302.13939*, 2023.

[33] Tobias Fischer and Michael Milford. Event-based visual place recognition with ensembles of temporal windows. *IEEE Robotics and Automation Letters*, 5(4):6924–6931, 2020.

[34] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017.

[35] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.

[36] Simon Klenk, Jason Chui, Nikolaus Demmel, and Daniel Cremers. Tum-vie: The tum stereo visual-inertial event dataset. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8601–8608. IEEE, 2021.

[37] 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.

[38] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017.

[39] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[40] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[41] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2661–2671, 2021.

[42] Jason K Eshraghian, Max Ward, Emre O Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 2023.

[43] Gregor Lenz, Kenneth Chaney, Sumit Bam Shrestha, Omar Oubari, Serge Picaud, and Guido Zarrella. Tonic: event-based datasets and transformations., July 2021. Documentation available under https://tonic.readthedocs.io.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11062–11070, 2021.

[46] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[47] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[48] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[49] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[50] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[51] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, jan 2022.

[52] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, apr 2022.

[53] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 12 2019.

[54] Mike Davies, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A Fonseca Guerra, Prasad Joshi, Philipp Plank, and Sumedh R Risbud. Advancing neuromorphic computing with loihi: A survey of results and outlook. *Proceedings of the IEEE*, 109(5):911934, 2021.

[55] Christian Mayr, Sebastian Hoeppner, and Steve Furber. Spinnaker 2: A 10 million core processor system for brain simulation and machine learning. *arXiv preprint arXiv:1911.02385*, 2019.

[56] Christian Pehle, Sebastian Billaudelle, Benjamin Cramer, Jakob Kaiser, Korbinian Schreiber, Yannik Stradmann, Johannes Weis, Aron Leibfried, Eric Müller, and Johannes Schemmel. The brainscales2 accelerated neuromorphic system with hybrid plasticity. *arXiv preprint arXiv:2201.11063*, 2022.

[57] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri. A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (dynaps). *IEEE Transactions on Biomedical Circuits and Systems*, PP(99):117, 2017.

[58] Bodo Rueckauer, IuliaAlexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and ShihChii Liu. Conversion of continuousvalued deep networks to efficient eventdriven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.

[59] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation, 2021.

[60] Xiang He, Dongcheng Zhao, Yang Li, Guobin Shen, Qingqun Kong, and Yi Zeng. Improving the performance of spiking neural networks on event-based datasets with knowledge transfer, 2023.

[61] Jason K Eshraghian, Max Ward, Emre Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D Lu. Training spiking neural networks using lessons from deep learning. *arXiv preprint arXiv:2109.12894*, 2021.

[62] Bernd Illing, Jean Ventura, Guillaume Bellec, and Wulfram Gerstner. Local plasticity rules can learn deep representations using self-supervised contrastive predictions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30365–30379. Curran Associates, Inc., 2021.

[63] Manu Srinath Halvagal and Friedemann Zenke. The combination of hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *bioRxiv*, 2022.

[64] Franz Scherr, Qinghai Guo, and Timoleon Moraitis. Self-supervised learning through efference copies. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4543–4557. Curran Associates, Inc., 2022.

[65] Yuhang Li, Youngeun Kim, Hyoungseob Park, Tamar Geller, and Priyadarshini Panda. Neuromorphic data augmentation for training spiking neural networks. In *European Conference on Computer Vision*, pages 631–649. Springer, 2022.

[66] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Eventmix: An efficient data augmentation strategy for event-based learning. *Inf. Sci.*, 644(C), oct 2023.

[67] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. Eventgan: Leveraging large scale image datasets for event cameras, 2019.

[68] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. pages 3583–3592, 06 2020.

[69] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 969–982. PMLR, 29–31 Oct 2018.

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

[71] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, jun 2018.

[72] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training, 2023.

[73] Asude Aydin, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. A hybrid ann-snn architecture for low-power and low-latency visual perception. *arXiv preprint arXiv:2303.14176*, 2023.

[74] Kenneth Stewart, Andreea Danielescu, Timothy Shea, and Emre Neftci. Encoding event-based data with a hybrid snn guided variational auto-encoder in neuromorphic hardware. In *Proceedings of the 2022 Annual Neuro-Inspired Computational Elements Conference*, NICE '22, page 88–97, New York, NY, USA, 2022. Association for Computing Machinery.

[75] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023.

[76] Alessandro Zanardi, Julian Zilly, Andreas Aumiller, Andrea Censi, and Emilio Frazzoli. Wormhole learning. pages 7899–7905, 05 2019.

[77] Alessandro Zanardi, Andreas Aumiller, Julian Zilly, Andrea Censi, and Emilio Frazzoli. Cross-modal learning filters for rgb-neuromorphic wormhole learning. 06 2019.

# A Appendix

## A.1 Related Work

Event-based cameras have garnered increasing interest due to their unique advantages over traditional frame-based cameras, including higher dynamic range and lower latency [51]. To process data from event-based cameras, one common approach is to slice event streams into frames, allowing standard architectures designed for image frames to be employed [52]. Alternatively, intensity images can be reconstructed from the event streams [53]. However, spiking neural networks (SNNs) offer a more natural fit for event-based data, particularly when considering the energy-efficient hardware available for SNNs [28, 54–57]. Despite this, efficiently training SNNs remains a significant challenge. A number of works have proposed methods to either distill or convert trained artificial neural networks (ANNs) into SNN models [58–60]. Advances in surrogate gradient techniques have made it possible to train SNNs in a manner that closely resembles standard deep learning methods [8, 61].

Large-scale model training, particularly for event-based data, increasingly relies on techniques that do not require manual annotation or labeled data. In the broader computer vision community, self-supervised learning (SSL) has gained prominence, with approaches like contrastive learning and joint-embedding architectures showing great promise [16–24]. These methods leverage data augmentations to create different views of the same image, but selecting the optimal combination of augmentations remains a challenge. Recent efforts have adapted SSL techniques to biologically-inspired settings, enabling some form of local learning in neural networks [62, 63], while others interpret the creation of different views in terms of causal and non-causal relationships in the world [64]. In addition, standard SSL techniques for image frames are now being adjusted for event-based data, with tailored augmentations for event streams [15, 65, 66]. There are also methods for converting video frames into event data for use in SSL [67–69].

In supervised learning, event-based data can be processed using standard ANNs as event frames [66, 70–72], hybrid SNN-ANN models [73, 74], or fully SNN-based architectures [15, 66]. To further enhance model performance, multimodal data can be incorporated instead of relying solely on vision-based data and random augmentations. Vision-language models (VLMs), such as CLIP [10], have shown strong performance when adapted to event-frames generated from event-based cameras [75]. Beyond text and vision, even different visual modalities, such as combining infrared and RGB data, have proven beneficial [76], with similar results observed when using RGB and event-based camera data together [77].