SpiLiFormer: Enhancing Spiking Transformers with Lateral Inhibition —Supplementary Material—

Zeqi Zheng[1,2*], Yanchen Huang[2,3*], Yingchao Yu[4,2], Zizheng Zhu[2,5],
Junfeng Tang[1,2], Zhaofei Yu[6], Yaochu Jin[2†]
[1]Zhejiang University [2]Westlake University [3]Nanjing University [4]Donghua University
[5]University of Electronic Science and Technology of China [6]Peking University
{zhengzeqi, huangyanchen, jinyaochu}@westlake.edu.cn

A. Appendix

A.1. Adversarial Test

Our adversarial experiments use two well-established techniques: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD).

FGSM [3] is a single-step adversarial attack algorithm designed to generate adversarial examples efficiently. It computes the gradient of the loss function with respect to the input data and adds a small perturbation in the direction of the gradient's sign. The adversarial example x_{adv} is generated as:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y)) \tag{1}$$

where x is the original input, ϵ controls the perturbation magnitude, and $J(x, y)$ is the loss function.

PGD [9] is an iterative variant of FGSM that generates adversarial examples by repeatedly applying gradient updates. Starting from an initial perturbed input, PGD iteratively refines the perturbation while projecting the result back into a L_∞-norm ball of radius ϵ. The update rule at each iteration t is:

$$x_{\text{adv}}^t = \text{Clip}_{x,\epsilon}\left(x_{\text{adv}}^{t-1} + \gamma \cdot \text{sign}(\nabla_x J(x_{\text{adv}}^{t-1}, y))\right) \tag{2}$$

where γ is the step size, and $\text{Clip}_{x,\epsilon}(\cdot)$ ensures the perturbation remains within the allowed bounds.

A.2. Datasets

Our experimental evaluation includes five standard datasets, consisting of three static ones (ImageNet-1K, CIFAR-10, and CIFAR-100) and two event-based neuromorphic ones (CIFAR10-DVS and N-Caltech101).

ImageNet-1K: ImageNet-1K [1], formally known as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset, is one of the most influential benchmarks in computer vision research. It comprises over 1.28 million training images across 1,000 classes, along with 50,000 validation images and 100,000 test images.

CIFAR-10: CIFAR-10 [6] is a fundamental benchmark dataset in computer vision research, comprising 60,000 color images of size 32×32 pixels, distributed across 10 mutually exclusive classes.

CIFAR-100: CIFAR-100 [6] builds upon the design principles of CIFAR-10 while introducing a more challenging classification task. It consists of 60,000 color images of size 32×32 pixels, categorized into 100 finer-grained classes.

CIFAR10-DVS: CIFAR10-DVS [8] represents a neuromorphic adaptation of the original CIFAR-10 dataset, specifically designed for the evaluation of SNNs in event-based vision tasks. There are 10,000 samples, whose spatial size is 128×128.

N-Caltech101: N-Caltech101 [10] is a neuromorphic adaptation of Caltech101, containing 101 classes and 8,709 samples with a spatial resolution of 180×240 pixels.

A.3. Energy Consumption Calculation of SNNs and ANNs

The uniformity of convolution enables the subsequent batch normalization (BN) and linear scaling transformations to be seamlessly integrated into the convolutional layer as an added bias during deployment [2, 5]. Consequently, when estimating theoretical energy consumption, the impact of BN layers can be disregarded. Before computing the theoretical energy consumption for SpiLiFormer, we first determine the number of Synaptic Operations (SOPs) of spikes.

$$\text{SOP}^i = f_r \times T \times \text{FLOPs}^i, \tag{3}$$

where i denotes the i-th layer module in SpiLiFormer, f_r represents the firing rate of spike trains at the input of the layer module, and T refers to the simulation time step. FLOPs^i represents the number of floating-point operations

* Equal contribution. † Corresponding author.

in the i-th layer module, measured in terms of multiply-and-accumulate (MAC) operations. SOP^i refers to the count of spike-based accumulate (AC) operations. We assume that MAC and AC operations are executed on 45nm hardware [4], where $E_{\mathrm{MAC}} = 4.6pJ$ and $E_{\mathrm{AC}} = 0.9pJ$ according to previous studies [4, 7, 11, 12]. The theoretical energy consumption of SpiLiFormer is computed as follows:

$$E_{\mathrm{SpiLiFormer}} = E_{\mathrm{AC}} \times \left(\sum_{i=2}^{M} \mathrm{SOP}^{i}_{\mathrm{Conv}} + \sum_{j=1}^{N} \mathrm{SOP}^{j}_{\text{FF-LiDiff Attn}} + \sum_{p=1}^{R} \mathrm{SOP}^{p}_{\text{FB-LiDiff Attn}} \right) + E_{\mathrm{MAC}} \times \mathrm{FLOPs}^{1}_{\mathrm{Conv}}, \quad (4)$$

where $\mathrm{FLOPs}^1_{\mathrm{Conv}}$ represents the floating-point operations in the first convolutional layer, which processes the input image in RGB format. Subsequently, the SOPs from M convolutional layers, N layers of FF-LiDiff attention, and R layers of FB-LiDiff attention are summed and multiplied by E_{AC}. For ANNs, the theoretical energy consumption is determined as follows:

$$E_{\mathrm{ANN}} = E_{\mathrm{MAC}} \times \mathrm{FLOPs}. \quad (5)$$

A.4. Selection of the Optimal α Hyperparameter

We perform an ablation study on both static and dynamic datasets to select the optimal value of the hyperparameter α, as shown in Tab. 4. The results show that the model achieves peak accuracy when $\alpha = 0.5$, while other values lead to varying degrees of performance degradation. As a result, we set α to 0.5 by default in all subsequent experiments.

Datasets	α Value				
	0.3	0.4	0.5	0.6	0.7
CIFAR-10	96.35	96.40	96.63	96.36	96.16
CIFAR10-DVS	86.3	86.4	86.7	86.1	85.6

Table 4. Ablation Study of the α hyperparameter

A.5. Evaluation of Inference Latency

Datasets	Inference Time per Sample (ms)	
	w/o FB-LiDiff	w/ FB-LiDiff
CIFAR-10	0.7657	0.9322 (+21.7%)
CIFAR-100	0.6965	0.8581 (+23.2%)
CIFAR10-DVS	9.9433	10.0657 (+1.2%)
N-Caltech101	18.4273	19.8113 (+7.5%)
ImageNet-1K	58.9901	66.3027 (+12.4%)

Table 5. Inference time per sample (ms) across datasets.

We conduct a comprehensive evaluation of the inference latency introduced by FB-LiDiff due to its additional forward pass. As shown in Tab. 5, FB-LiDiff increases inference time by 1.2% to 23.2%, with over 20% overhead observed on static CIFAR datasets.

A.6. Supplementary Tables and Figures

Dataset	Methods	Time Step	Clean	FGSM Maximum Perturbation				PGD Iterations			
				0.05	0.1	0.2	0.3	5	10	30	50
CIFAR-10	QKFormer	4	96.18	68.33	65.67	59.51	53.33	33.94	25.56	17.96	16.8
	SpiLiFormer(Ours)	4	**96.63 (+0.45)**	**68.9 (+0.57)**	**66.05 (+0.38)**	**59.98 (+0.47)**	**54.46 (+1.13)**	**35.43 (+1.49)**	**25.7 (+0.14)**	**18.61 (+0.65)**	**17.13 (+0.33)**
CIFAR-100	QKFormer	4	81.15	35.45	31.18	26.27	22.07	18.48	13.43	10.09	9.12
	SpiLiFormer(Ours)	4	**81.63 (+0.48)**	**37.33 (+1.88)**	**34.19 (+3.01)**	**29.08 (+2.81)**	**24.3 (+2.23)**	**18.84 (+0.36)**	**14.04 (+0.61)**	**10.43 (+0.34)**	**9.67 (+0.55)**
CIFAR10-DVS	QKFormer	16	84.00	29.30	17.70	10.50	10.00	2.00	1.20	0.40	**0.50**
	SpiLiFormer(Ours)	16	**86.70 (+2.70)**	**34.50 (+5.20)**	**23.70 (+6.00)**	**15.60 (+5.10)**	**11.70 (+1.70)**	**3.50 (+1.50)**	**1.40 (+0.20)**	0.40	0.40
										0.00	-0.10
ImageNet-1K	QKFormer	1	80.10	37.38	32.39	27.13	23.83	9.49	4.44	2.04	1.81
	SpiLiFormer(Ours)	1	**81.54 +(1.44)**	**40.17 (+2.79)**	**36.01 (+3.62)**	**30.99 (+3.86)**	**27.45 (+3.62)**	**10.89 (+1.40)**	**5.14 (+0.70)**	**2.43 (+0.39)**	**1.95 (+0.14)**

Table 6. Adversarial robustness comparison between QKFormer and our model across four datasets, including CIFAR-10, CIFAR-100, CIFAR10-DVS, and ImagetNet-1K. For PGD, the attack strength is set to 8/255, with a step size of 2/255 per iteration.

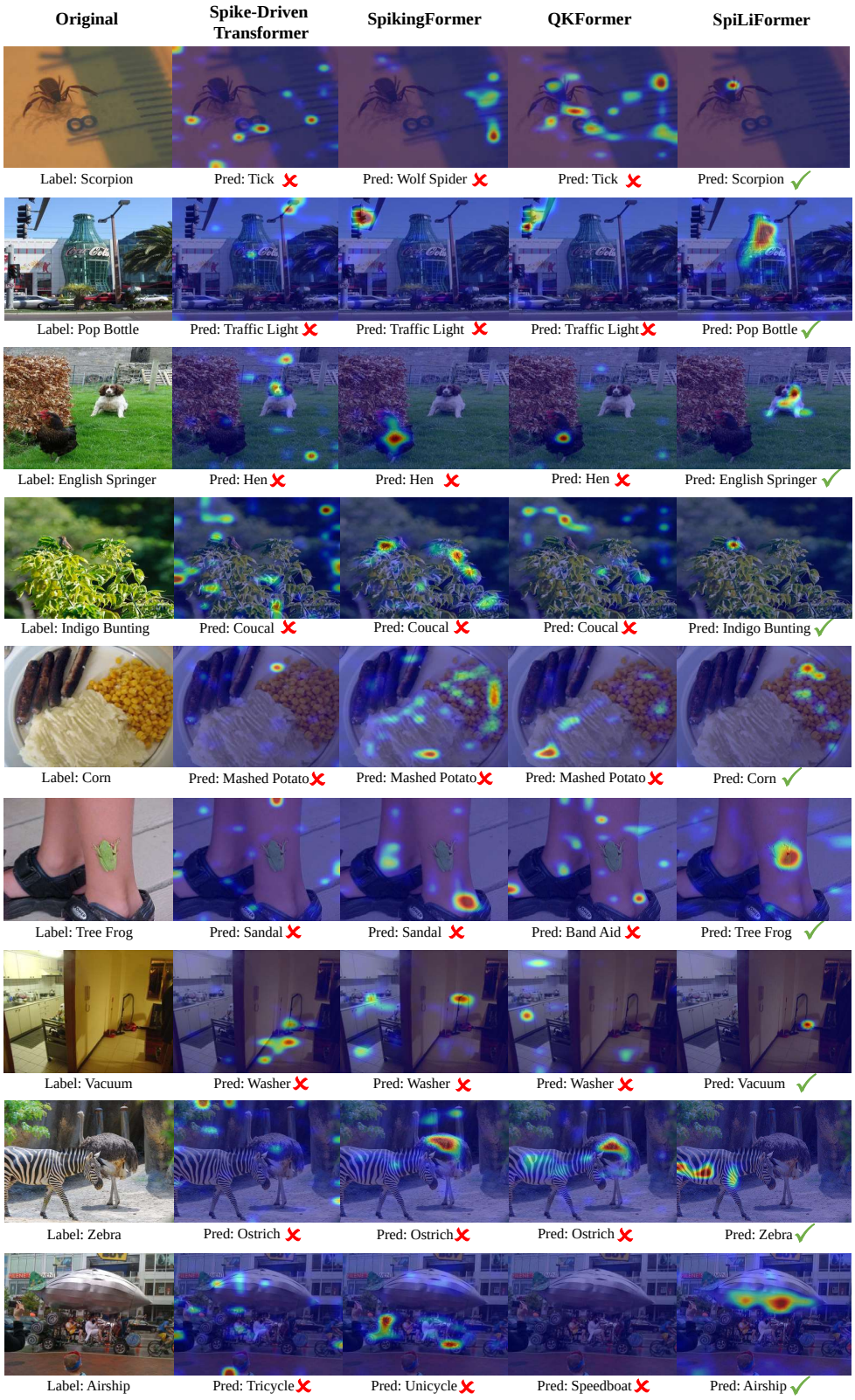

Figure 4. Comparative visualization of attention heatmaps from ImageNet-1K, with corresponding ground truth labels and model predictions annotated below each sample.

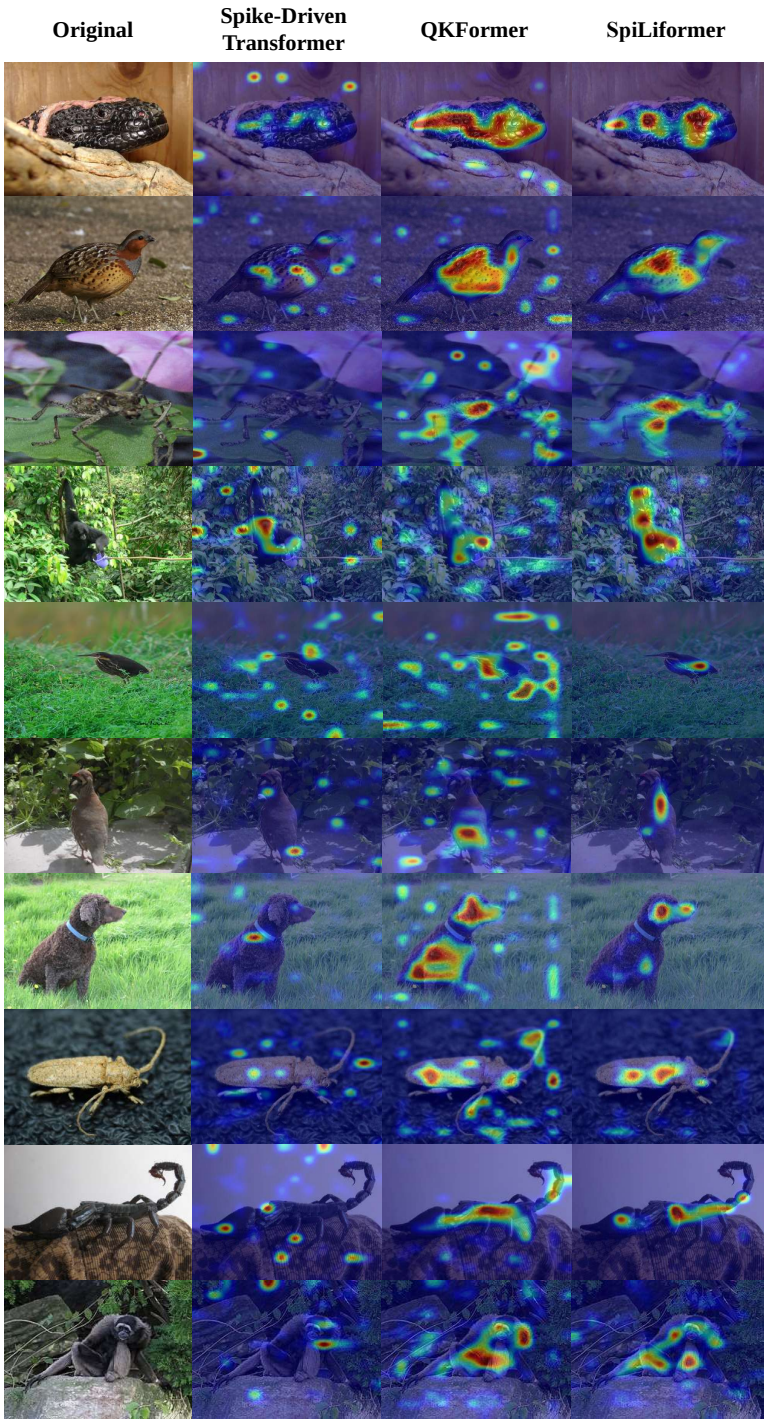

Figure 5. Representative samples from ImageNet-1K, demonstrating original images and their corresponding attention heatmaps across different models.

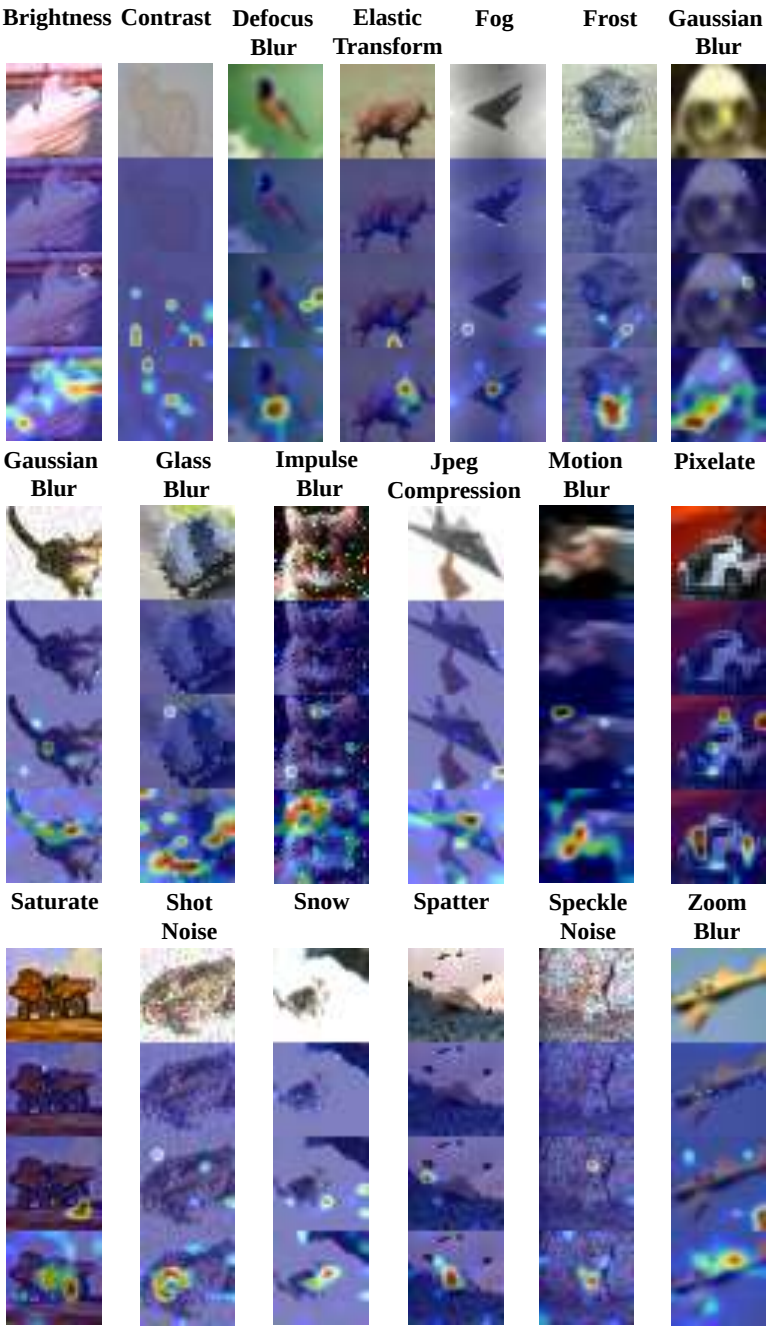

Figure 6. Visualization of CIFAR-10C. For all 19 types of corruptions, each column displays the following cases: the first image is the original corrupted image; the second and third images show the attention heatmaps of Spike-Driven Transformer and QKFormer, respectively; the last image visualizes the attention of SpiLiFormer.

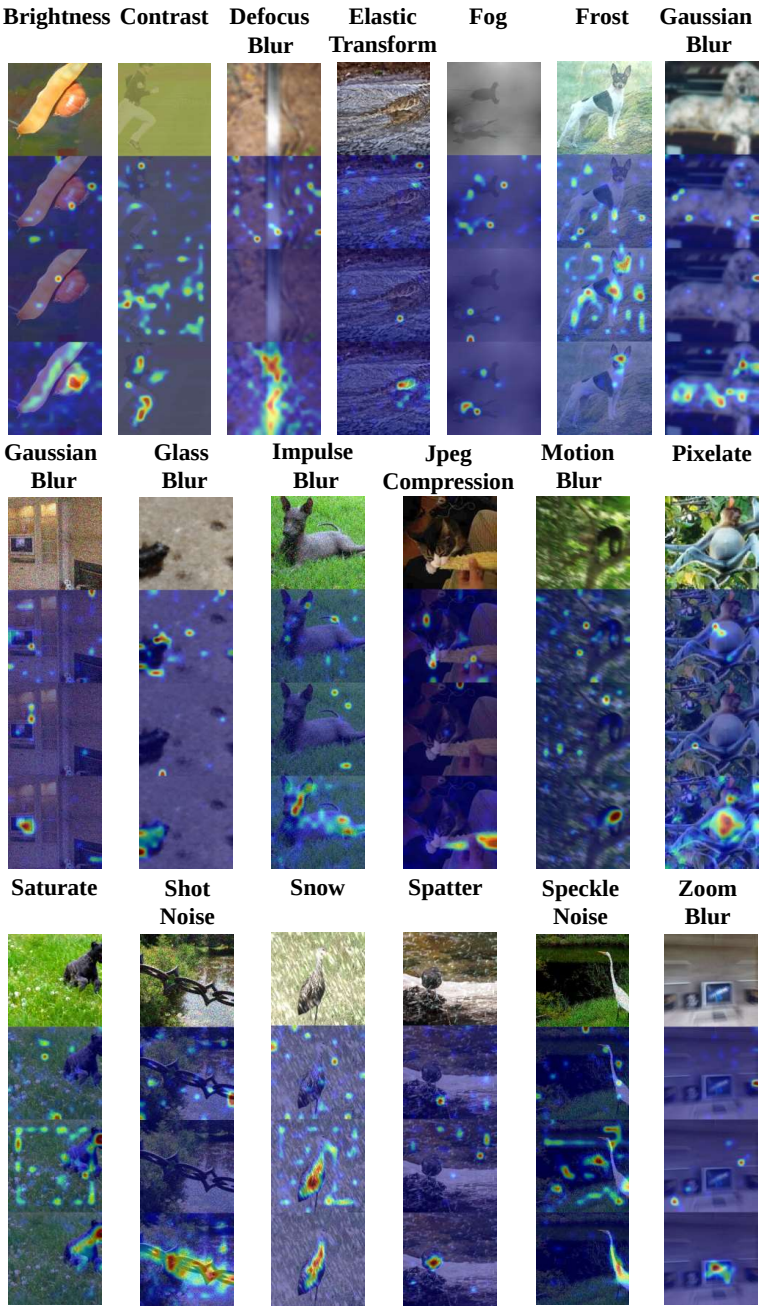

Figure 7. Visualization of ImageNet-1K-C for all types of corruptions. The layout and image order follow the same structure as illustrated in Fig. 6.

References

[1] 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. Ieee, 2009. 1

[2] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. *arXiv preprint arXiv:2202.11946*, 2022. 1

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[4] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*, pages 10–14. IEEE, 2014. 2

[5] Yangfan Hu, Huajin Tang, and Gang Pan. Spiking deep residual networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):5200–5205, 2021. 1

[6] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 1

[7] Souvik Kundu, Massoud Pedram, and Peter A Beerel. Hire-snn: Harnessing the inherent robustness of energy-efficient deep spiking neural networks by training with crafted input noise. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5209–5218, 2021. 2

[8] 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. 1

[9] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017. 1

[10] 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

[11] Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. *Advances in neural information processing systems*, 36, 2024. 2

[12] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, YAN Shuicheng, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023. 2