Deep Directly-Trained Spiking Neural Networks for Object Detection (Supplementary Materials)

Qiaoyi Su[1], Yuhong Chou[2,7], Yifan Hu[3], Jianing Li[4], Shijie Mei[5,7], Ziyang Zhang[6], Guoqi Li[1,7*]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]College of Artificial Intelligence, Xi'an Jiaotong University
[3]Department of Precision Instrument, Tsinghua University
[4]School of Computer Science, Peking University
[5]School of Vehicle and Mobility, Tsinghua University
[6]Advanced Computing and Storage Lab, Huawei Technologies Co. Ltd.
[7]Institute of Automation, Chinese Academy of Sciences

suqiaoyi2020@ia.ac.cn, huyf19@mails.tsinghua.edu.cn, lijianing@pku.edu.cn, guoqi.li@ia.ac.cn

A. Proof of Gradient Norm Equality

Definition 1 ***(General Linear Transform)*** *Let $f(x)$ be a transform whose Jacobian matrix is $\boldsymbol{J}$. f is called general linear transform when it satisfies:*

$$E\left[\frac{\|\boldsymbol{f}(\boldsymbol{x})\|_2^2}{\text{len}(\boldsymbol{f}(\boldsymbol{x}))}\right] = \phi\left(\boldsymbol{J}\boldsymbol{J}^T\right) E\left[\frac{\|\boldsymbol{x}\|_2^2}{\text{len}(\boldsymbol{x})}\right]. \tag{1}$$

Lemma 1 (Multiplication) *(Theorem 4.1 in [1]) Given $\boldsymbol{J} := \prod_{j=L}^{1} \boldsymbol{J}_j$, where $\{\boldsymbol{J}_j \in \mathbb{R}^{m_j \times m_{j-1}}\}$ is a series of independent random matrices. If $(\prod_{j=L}^{1} \boldsymbol{J}_j)(\prod_{j=L}^{1} \boldsymbol{J}_j)^T$ is at least the 1^{st} moment unitarily invariant, we have*

$$\phi\left((\prod_{j=L}^{1} \boldsymbol{J}_j)(\prod_{j=L}^{1} \boldsymbol{J}_j)^T\right) = \prod_{j=L}^{1} \phi(\boldsymbol{J}_j {\boldsymbol{J}_j}^T). \tag{2}$$

Lemma 2 (Addition) *(Theorem 4.2 in [1]) Given $\boldsymbol{J} := \prod_{j=L}^{1} \boldsymbol{J}_j$, where $\{\boldsymbol{J}_j \in \mathbb{R}^{m_j \times m_{j-1}}\}$ is a series of independent random matrices. If at most one matrix in $\boldsymbol{J}_j$ is not a central matrix, we have*

$$\phi(\boldsymbol{J}\boldsymbol{J}^T) = \sum_j \phi(\boldsymbol{J}_j {\boldsymbol{J}_j}^T). \tag{3}$$

Proposition 1 *For EMS-Block1 and EMS-Block2, the Jacobian matrix of the block can be represented as $\phi(\boldsymbol{J_j J_j^T}) = \frac{2}{\alpha_2^{j-1}}$.*

Proof A.1 ***proof of EMS-Block1.*** *Since the EMS-Blocks have 2 paths, the residual path and the shortcut path while separately name the Jacobian matrix of two paths (of block) as $\boldsymbol{J}_{res}$ and $\boldsymbol{J}_{sc}$. l is the layer number of the block, and it will be omitted where there is no ambiguity.*

For the residual path with 2 LCB blocks, according to General Linear Transform, we have

$$\alpha_2^{l,res} = \phi(\boldsymbol{J}_{res}\boldsymbol{J}_{res}^T)\alpha_2^{l-1},$$

The shortcut path is similar

$$\alpha_2^{l,sc} = \phi(\boldsymbol{J}_{sc}\boldsymbol{J}_{sc}^T)\alpha_2^{l-1},$$

Here, α_2^{l-1} is the 2^{th} moment of the input data from $(l-1)^{th}$ block. Because the initialized BN layer have the output with variance 1 and mean 0, $\alpha_2^{l,res} = \alpha_2^{l,sc} = 1$.Thus

$$\phi(\boldsymbol{J}_{res}\boldsymbol{J}_{res}^T) = \frac{1}{\alpha_2^{l-1}},$$

$$\phi(\boldsymbol{J}_{sc}\boldsymbol{J}_{sc}^T) = \frac{1}{\alpha_2^{l-1}}.$$

By addition principle

$$\begin{aligned}&\phi(\boldsymbol{J}_{EMS-Block1}\boldsymbol{J}_{EMS-Block1}^T)\\ &= \phi(\boldsymbol{J}_{res}\boldsymbol{J}_{res}^T) + \phi(\boldsymbol{J}_{sc}\boldsymbol{J}_{sc}^T)\\ &= \frac{2}{\alpha_2^{l-1}}.\end{aligned}$$

Proof A.2 ***proof of EMS-Block2.*** *Comparing with EMS-Block1, the EMS-Block2 extra have a concatenation at the shortcut path.*

[*] Corresponding author

Stage	ResNet-10	ResNet-18	ResNet-34
Conv1	3×3, 32, stride 2		
Conv2_x	$\begin{bmatrix} \text{3x3, 32} \\ \text{3x3, 64} \end{bmatrix} * 1$	$\begin{bmatrix} \text{3x3, 32} \\ \text{3x3, 64} \end{bmatrix} * 2$	$\begin{bmatrix} \text{3x3,32} \\ \text{3x3, 64} \end{bmatrix} * 3$
Conv3_x	$\begin{bmatrix} \text{3x3, 64} \\ \text{3x3, 128} \end{bmatrix} * 1$	$\begin{bmatrix} \text{3x3, 64} \\ \text{3x3, 128} \end{bmatrix} * 2$	$\begin{bmatrix} \text{3x3, 64} \\ \text{3x3, 128} \end{bmatrix} * 4$
Conv4_x	$\begin{bmatrix} \text{3x3, 128} \\ \text{3x3, 256} \end{bmatrix} * 1$	$\begin{bmatrix} \text{3x3, 128} \\ \text{3x3, 256} \end{bmatrix} * 2$	$\begin{bmatrix} \text{3x3, 128} \\ \text{3x3, 256} \end{bmatrix} * 6$
Conv5_x	$\begin{bmatrix} \text{3x3, 256} \\ \text{3x3, 512} \end{bmatrix} * 1$	$\begin{bmatrix} \text{3x3, 256} \\ \text{3x3, 512} \end{bmatrix} * 2$	$\begin{bmatrix} \text{3x3, 256} \\ \text{3x3, 512} \end{bmatrix} * 3$

Table 1. **Model structures for ablation experiments.** x represents the current module repeated x times and the first module transformed in a reduced dimension. Compared with the original ResNet structure, the number of channels are resized here, while the final FC layer is removed.

According to the discussion about concatenation in [1], we have

$$\phi\left(\boldsymbol{J}_j\boldsymbol{J}_j^T\right) = \frac{c_{j-1}}{c_j} + \frac{\delta_j}{c_j}\phi\left(\boldsymbol{H}_j\boldsymbol{H}_j^T\right),$$

Here $\boldsymbol{J}_j$ denote Jacobian matrix of the block of shortcut path without maxpooling layer. H_j denote the Jacobian matrix of the LCB block. c_{j-1} and c_j denoted as the channel numbers for input and output of concatenation. And $\delta_j = c_j - c_{j-1}$. It is trivial that by adding the maxpooling layer and using general linear transform, shortcut path can be expressed as

$$\phi(\boldsymbol{J}_{sc}\boldsymbol{J}_{sc}^T) = \frac{\alpha_2^{maxpool}}{\alpha_2^{l-1}}(\frac{c_{j-1}}{c_j} + \frac{\delta_j}{c_j}\phi\left(\boldsymbol{H}_j\boldsymbol{H}_j^T\right))$$

$$= \frac{1}{\alpha_2^{l-1}}(\frac{\alpha_2^{maxpool}c_{j-1}}{c_j} + \frac{\alpha_2^{maxpool}\delta_j}{c_j}\left(\frac{\alpha_2^{bn}}{\alpha_2^{maxpool}}\right)),$$

Since the 2^{th} moment α_2^{l-1} is strictly controlled by the BN layers of former block, the $\alpha_2^{maxpool}$ is fixed too. Thus, let $\alpha_2^{bn} = \frac{2c_j - \alpha_2^{maxpool}c_{j-1}}{\delta_j}$ by proper initializing of BN layers, $\phi(\boldsymbol{J}_{sc}\boldsymbol{J}_{sc}^T) = \frac{1}{\alpha_2^{l-1}}$ holds.

The other part of EMS-Block2 is similar with EMS-Block1, thus

$$\phi(\boldsymbol{J}_{EMSblock2}\boldsymbol{J}_{EMSblock2}^T) = \frac{2}{\alpha_2^{l-1}}.$$

Proposition 2 *For the EMS-ResNet, $\phi(\boldsymbol{J}\boldsymbol{J}^{\boldsymbol{T}}) \approx 1$ can be satisfied by control the 2^{th} moment of the input.*

Proof A.3 *MS-Block is a typical resblock that have already been discussed in [1]. Using general linear transform and addition principle, we have*

$$\alpha_2^{l-1}\phi(\boldsymbol{J}\boldsymbol{J}^T) = \alpha_2^l = \alpha_2^{l-1} + 1.$$

And α_2^{l-1} is comes from the EMS-Block1 or EMS-Block2, where α_2^{l-1} is fixed at 2. Thus, $\phi(\boldsymbol{J}_{MS-Block}\boldsymbol{J}_{MS-Block}^T) = \frac{3}{2}$.

By using multiplication principle, the whole blocks have the property

$$\phi(\boldsymbol{J}\boldsymbol{J}^T) = \frac{3}{\alpha_2^0},$$

where α_2^0 is the 2^{th} moment of the output of BN in encoding layer.

After initialized the BN in encoding layer, α_2^0 can be controlled to 3 and then $\phi(\boldsymbol{J}\boldsymbol{J}^T) \approx 1$ holds.

B. Datasets Introduction

COCO2017 Dataset COCO2017 Dataset [3] is a large-scale object detection, segmentation, key-point detection, and captioning dataset. For object detection, its training set and test set contain 118K and 5K images, respectively. The instances of 80 categories are labeled with their classes and bounding boxes respectively.

GEN1 Automotive Detection Dataset Event cameras possess outstanding properties compared with the traditional frame cameras. They have high dynamic range to overcome motion blur. Furthermore, objects are captured well even in low-light or overexposed scenes. The event e_n (defined in **Sec 4.1**) in the event camera represents the change in light intensity I of the pixel (x_n, y_n), which can be formulated as:

$$\ln I(x_n, y_n, t_n) - \ln I(x_n, y_n, t_n - \Delta t_n) = p_n\theta_{th} \quad (4)$$

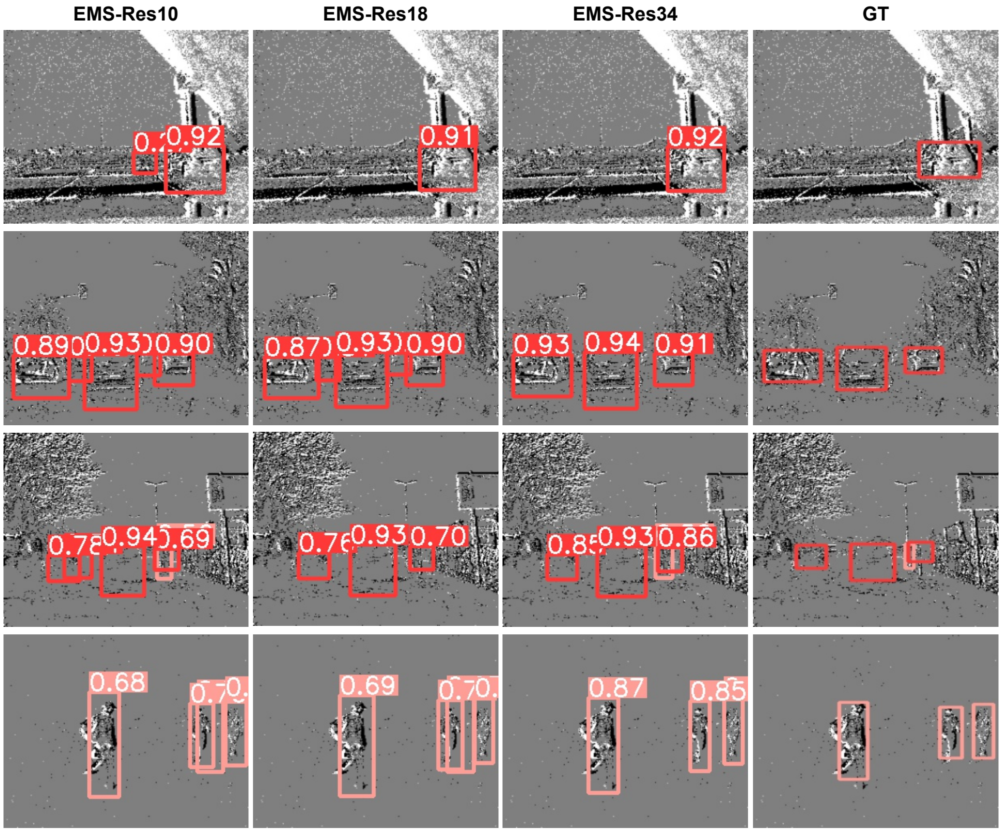

Figure 1. **More detailed comparison figures on the Gen1 dataset**.

where Δt_n represents the temporal sampling interval.

As the largest event camera-based dataset currently available, Gen1 dataset [2] contains two categories (pedestrians and cars), and 39 hours of automotive recordings in diverse scenarios. Gen1 is labeled manually by the gray level estimation feature of the ATIS sensor [4] with a resolution of 304×204 pixels and more than 255,000 bounding box annotations are yielded in total.

Model	mAP @0.5	mAP @0.5:0.95	Params	Firing Rate	Energy Efficiency
Sew-Res18	0.345	0.183	9.743M	24.20%	3.31×
MS-Res18	0.345	0.184	9.678M	32.32%	3.55×
EMS-Res18	**0.362**	**0.201**	**9.523M**	38.75%	**5.98**×

Table 2. Ablation studies of different residual blocks on COCO2017 dataset.

C. More Detailed Experiments

Here we provide more details on the experiments in **Sec 5.3** using the COCO2017 dataset. They are all based on the channel number reduction models (see Table 1), because we aim to validate the reliability of our experimental conclusions, not to achieve the optimal performance. In addition, we explore the impact of the last layer of LIF on the model performance. We use 4 Nvidia A100 GPUs and the SGD optimizer with a learning rate of 1E-2 for training.

Different Residual Blocks We conduct additional experiments on the COCO2017 dataset to fully illustrate the effectiveness of EMS-ResNet. We train all the models for only 120 epochs with a batchsize of 64, and the time steps are set to 3. As shown in Table 2 , our model presents optimal performance with a high spiking rate, which may imply an increase in spiking rate, enabling better model feature extraction. At the same time, our model is fully spiked, and even with a high spike rate, it is still more energy-efficient than other models. We set the energy consumption of the ANN with the same structure as the baseline, denoted as 1×, and our full spike EMS-ResNet reduces the energy consumption up to 5.98 times.

Model	mAP @0.5	mAP @0.5:0.95	Params	Firing Rate
EMS-Res10	0.203	0.091	6.387M	30.01%
EMS-Res18	0.268	0.132	9.523M	28.56%
EMS-Res34	0.335	0.178	14.58M	29.55%

Table 3. Impact of different number of residual blocks on COCO2017 dataset.

Numbers of Residual Blocks We explore the effect of network depth on performance on the COCO2017 dataset. Here we set the time step to 1 and train for only 50 epochs. As shown in Table 3, the network converges faster and recognizes objects more accurately as the depth increases.

Dataset	Model	T	Params	mAP @0.5	mAP @0.5:0.95
COCO	Non-spiking	3	9.523M	0.318	0.165
	Spiking	3	9.523M	0.305	0.157
Gen1	Non-spiking	5	9.343M	0.566	0.286
	Spiking	5	9.343M	0.565	0.286

Table 4. Impact of spiking/non-spiking detection layer on the model performance.

Spiking Detection layer For the object detection task, it is necessary to consider how to convert the features of the spike trains into continuous value representations of the bounding box coordinates. This can be achieved by using either a non-spiking detection layer that directly feeds the last neuronal membrane potential or a spiking detection layer that uses rate-coding before different detection layers. From the experimental results (see Table 4), these two conversion methods have little effect on the performance of the model.

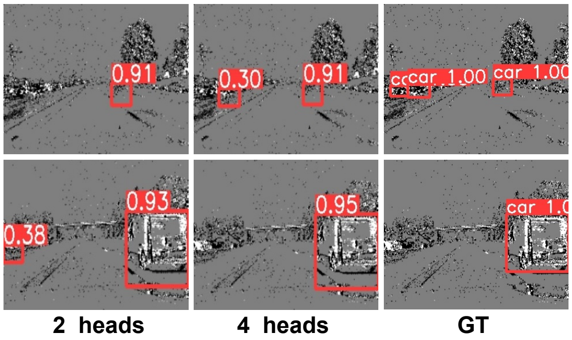

Figure 2. **Detection results of different scale detection heads on the Gen1 dataset.**

Numbers of Detection Heads We explore the impact of the number of detection heads on performance on the Gen1 dataset (Table 5). We compare 2-scale and 4-scale detection heads with the same backbone. From Figure 2, it can be seen that when the detection head scale is larger, the detailed information of the detection is more rich.

Heads Scale	mAP @0.5	mAP @0.5:0.95	Params	Firing Rate
2	0.565	0.286	9.34M	20.09%
4	**0.617**	**0.321**	10.04M	22.23%

Table 5. Impact of different number of detection heads on Gen1 dataset.

D. Detection Results Presentation

In the main text, we abbreviate the comparison results on the Gen1 dataset into a relatively small plot due to space constraints, and here we enlarge the results to be able to observe the details better (see Figure 1). In addition, we present capture videos on the Gen1 dataset in the video folder of the Supplementary Materials.

References

[1] Zhaodong Chen, Lei Deng, Bangyan Wang, Guoqi Li, and Yuan Xie. A comprehensive and modularized statistical framework for gradient norm equality in deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):13–31, 2022.

[2] 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.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2010.