

# Performance comparison of DVS data spatial downscaling methods using Spiking Neural Networks

## Supplementary Material

Amélie Gruel  
 CNRS, i3S, Université Côte d’Azur  
 Sophi-Antipolis, France  
 amelie.gruel@univ-cotedazur.fr

Jean Martinet  
 CNRS, i3S, Université Côte d’Azur  
 Sophi-Antipolis, France  
 jean.martinet@univ-cotedazur.fr

Bernabé Linares-Barranco  
 IMSE-CNM  
 Sevilla, Spain  
 bernabe@imse-cnm.csic.es

Teresa Serrano-Gotarredona  
 IMSE-CNM  
 Sevilla, Spain  
 terese@imse-cnm.csic.es

### 1. Study of spatial event downscaling influence on a classification task - Additional classifier

Fang et al. introduced at the ICCV 2021 conference a SNN classifier, consisting in a new spiking neuron model called Parametric Leaky Integrate-and-Fire (PLIF) [2]. This model aims to better represent the heterogeneity of biological neurons by learning the membrane time constant, whereas membrane parameters correspond usually to hyperparameters set before training. To reproduce the biological dynamic where all neighbouring neurons share similar properties, the time constant is the same for all neurons belonging in the same layer but differs between layers, thus implementing diverse phase-frequency responsiveness. The authors also opted for a max-pooling method instead of the average-pooling method traditionally used in SNN, thus preserving the asynchronous characteristic of neuron firing.

The authors implement a backpropagation learning algorithm, applied to a classification task. They present the results

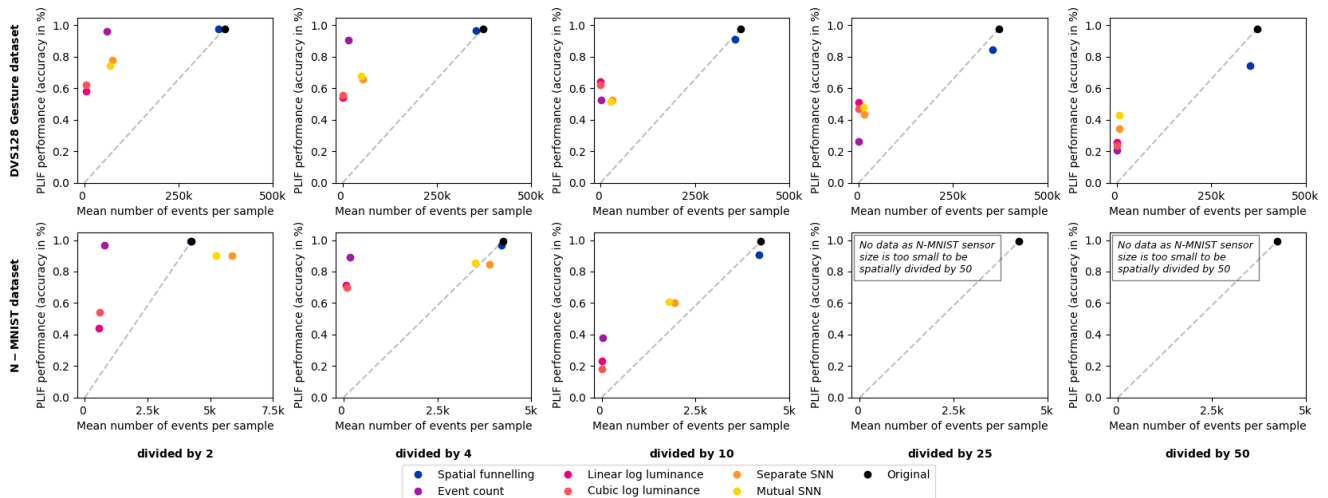


Figure 1: Evolution of the accuracy performance of the PLIF classifier applied to DVS128 Gesture (first line) and N-MMNIST (second line). The dashed line separates each plot between what we aim for – a decreased number of events for a significantly good accuracy (upper left) – and what we don’t aim for – a decreased accuracy for a great number of events (lower right). Globally most results are in the part that we aim for, with the notable exception of the ”SNN pooling” results.

obtained when classifying traditional RGB datasets, as well as neuromorphic datasets such as DVS128 Gesture [1], N-MNIST [5] and CIFAR10-DVS [4]. It is important to note that the authors chose to process the neuromorphic datasets as frames, and this precisely because of the high number of events. The events  $E$  are thus translated from their traditional representation  $E(x_i, y_i, p_i, t_i)$  (with  $x_i$  and  $y_i$  the pixel's coordinate of the  $i^{th}$  event,  $p_i$  the polarity and  $t_i$  the timestamp) into  $T$  slices of same time length. The frames  $F$  are represented as follows:

$$F(j, p, x, y) = \sum_{t=t_{min}}^{t_{max}} \delta(x_t, x) \cdot \delta(y_t, y) \cdot \delta(p_t, p) \quad (1)$$

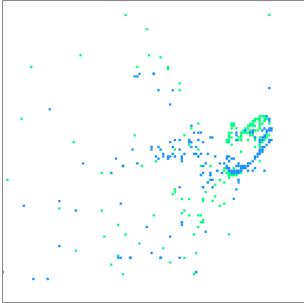
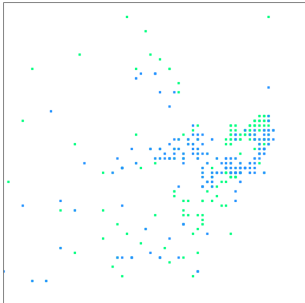
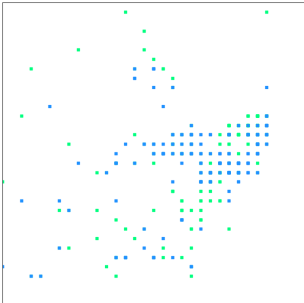
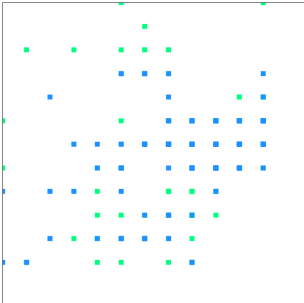
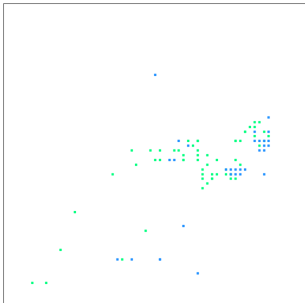
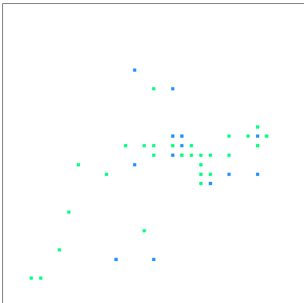
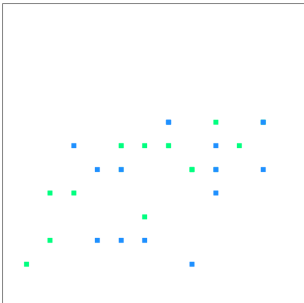

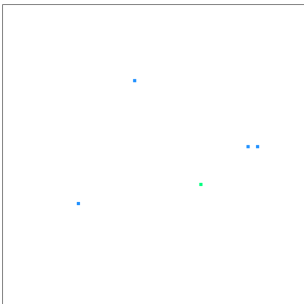
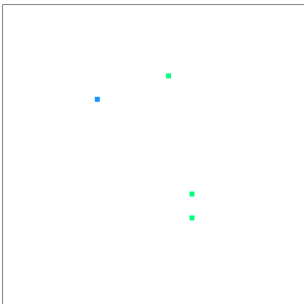
with  $t_{min}$  and  $t_{max}$  the minimal and maximal timestamp in the  $j^{th}$  slice and  $\delta$  the Kronecker delta function, which returns 1 if the parameters are equal, and 0 otherwise. When run on neuromorphic datasets, this classifier considers each frame as the input for one timestep;  $T$  is the number of timesteps on which the classifier is run for one epoch.

Comparing to the conclusions reached in the main paper, the performance results presented in Fig. 1 can be deemed surprising. Indeed, PLIF tends to the conclusion that the funnelling method is by far the most appropriate downscaling method, at least in the context of a classification task. This can be explained by the fact that the PLIF classifier, which takes in input frames, will have a better recognition performance when the number of events is higher, as it will have more information to choose from than in a sparser case. Indeed classifiers based on convolution such as PLIF tend to underperform on sparse data: they can handle fullscale event data, but quickly lose performance when the number of events drop, i.e. when the data is reduced using any methods except funnelling.

## 2. Visualisation of different spatial downscaling methods.

Tables 1 and 2 present one frame extracted respectively from the DVS 128 Gesture [1] and N-MNIST [5] datasets.

Three examples are presented for each method, one corresponding to a spatial downscaling by a factor of 2 (first column), another by a factor of 4 (second column) and the last by a factor of 10 (third column). The corresponding number of events produced by each method from this sample is indicated bellow each frame.

| Methods   | Downscaled by 2   | Downscaled by 4  | Downscaled by 10  |
|---|---|--|---|
| <p><b>Original</b><br/> <i>Number of events:</i><br/> <math>N_e = 670,118</math></p>  |   |  |   |
| <p>Simple event funnelling<br/> <i>Number of events:</i><br/> <math>N_e^{div=2} = 649,507</math><br/> <math>N_e^{div=4} = 649,414</math><br/> <math>N_e^{div=10} = 648,989</math></p>           |   |   |   |
| <p>Log-luminance<br/> with event count<br/> <i>Number of events:</i><br/> <math>N_e^{div=2} = 105,164</math><br/> <math>N_e^{div=4} = 25,400</math><br/> <math>N_e^{div=10} = 3,958</math></p>  |  |  |  |
| <p>Log-luminance<br/> with linear estimation<br/> <i>Number of events:</i><br/> <math>N_e^{div=2} = 7,016</math><br/> <math>N_e^{div=4} = 1,156</math><br/> <math>N_e^{div=10} = 143</math></p> |  |  |  |

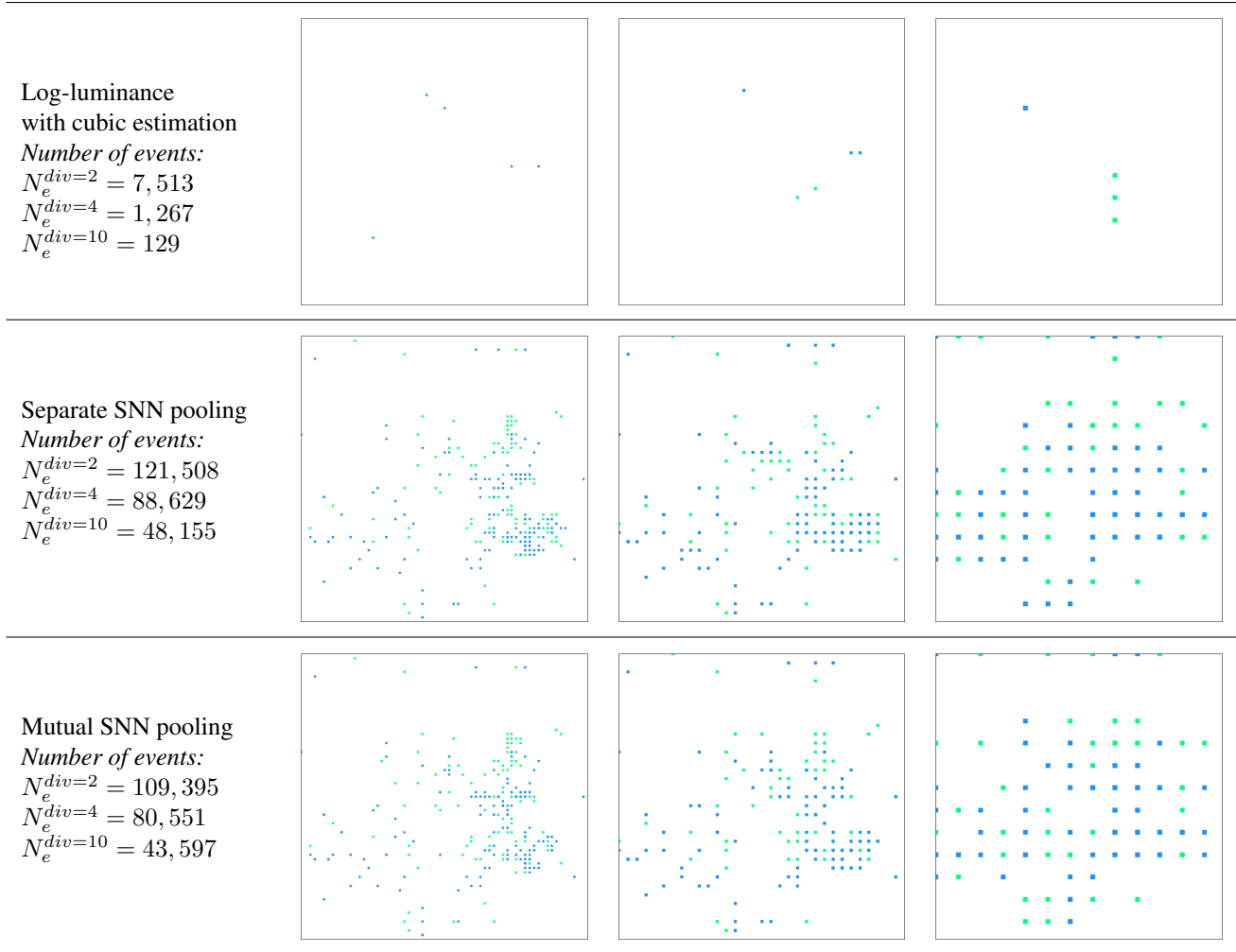
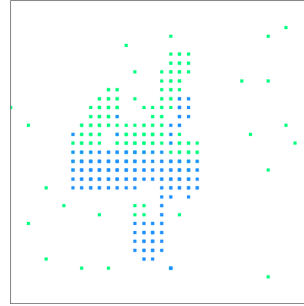


Table 1: Frames produced from the sample `user14_led_0.npz`, corresponding to the gesture "left hand clockwise" (class 6) from the DVS128 Gesture dataset, processed with the spatial downscaling methods described in [3] and the two novel methods introduced in this work. Each frame corresponds to the accumulation of the events occurring during the sample's first 5 ms. Green and blue pixels correspond respectively to positive and negative events.

**Original**

Number of events:

$$N_e = 3,848$$



| Methods  | Downscaled by 2 | Downscaled by 4 | Downscaled by 10 |
|--|-----------------|-----------------|------------------|
| Simple event funnelling<br>Number of events:<br>$N_e^{div=2} = 3,845$<br>$N_e^{div=4} = 3,844$<br>$N_e^{div=10} = 3,838$     |                 |                 |                  |
| Log-luminance with event count<br>Number of events:<br>$N_e^{div=2} = 716$<br>$N_e^{div=4} = 165$<br>$N_e^{div=10} = 19$     |                 |                 |                  |
| Log-luminance with linear estimation<br>Number of events:<br>$N_e^{div=2} = 566$<br>$N_e^{div=4} = 76$<br>$N_e^{div=10} = 3$ |                 |                 |                  |
| Log-luminance with cubic estimation<br>Number of events:<br>$N_e^{div=2} = 597$<br>$N_e^{div=4} = 84$<br>$N_e^{div=10} = 3$  |                 |                 |                  |

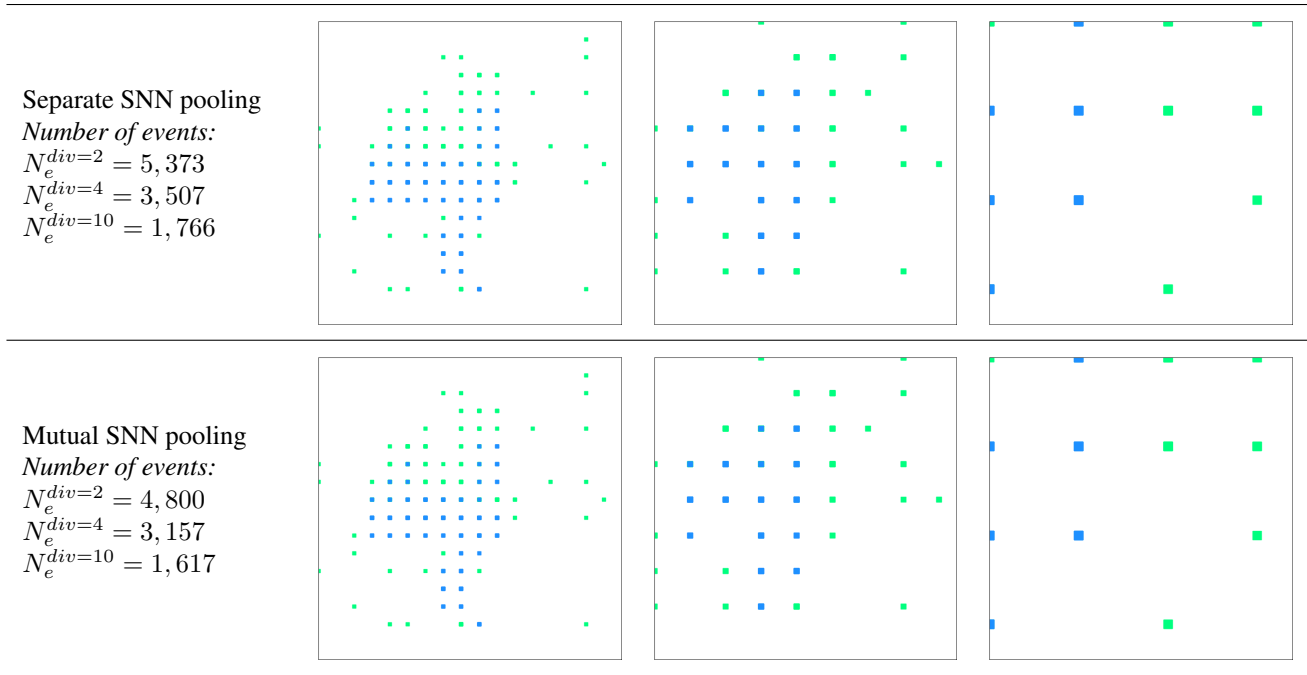


Table 2: Frames produced from the sample 19484.npz, corresponding to the number "4" from the N-MNIST dataset, processed with the spatial downscaling methods described in [3] and the two novel methods introduced in this work. Each frame corresponds to the accumulation of the events occurring during the sample's first 50 ms. Green and blue pixels correspond respectively to positive and negative events.

## References

- [1] A. Amir et al. A Low Power, Fully Event-Based Gesture Recognition System. In *CVPR*. IEEE, 2017.
- [2] Wei Fang et al. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *ICCV*, 2021.
- [3] Amélie Gruel, Jean Martinet, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. Event data downscaling for embedded computer vision. In *VISAPP*, 2022.
- [4] 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.
- [5] Garrick Orchard et al. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015.