ClearSight: Human Vision-Inspired Solutions for Event-Based Motion Deblurring -Supplementary Material-

Xiaopeng Lin*, Yulong Huang*, Hongwei Ren, Zunchang Liu, Hongxiang Huang,
Yue Zhou, Haotian Fu, Bojun Cheng†
The Hong Kong University of Science and Technology (Guangzhou)
{xlin746, yhuang496, hren066, zliu361, hhuang516, yzhou883, hfu373}@connect.hkust-gz.edu.cn,
bocheng@hkust-gz.edu.cn

1. Background

1.1. Human Visual System

The Human Visual System (HVS) facilitates visual perception by transmitting information from specialized retinal cell types through the optic nerve to the brain. This transmission employs two distinct pathways. The parvocellular pathway, originating from midget cells, delivers essential color and pattern information to parvocellular layers, essential for discerning fine visual details. The magnocellular pathway, initiated by parasol cells, provides depth and motion cues to magnocellular layers [1], as shown in Figure 1 in the main manuscript.

Visual signals are first relayed to the lateral geniculate body of the thalamus, serving as a critical sensory relay station, before advancing to the primary visual cortex for preliminary processing, including the visual attention [4]. Specifically, the visual attention mechanism within HVS consists of two primary components: the first is a baseline increase in neural activity that elevates neuron activity across specific visual areas, as shown in Figure 1 in the main manuscript; the second component is gain modulation, which aims to enhance the magnitude of neural responses. The visual association cortex processes information relayed from the visual cortex, engaging in sophisticated analysis of high-level semantic content from visual images. Consequently, this advanced processing capacity allows the human visual system to robustly identify and track objects within complex and dynamic environments.

1.2. Spiking Neural Networks

Spiking Neural Networks, as bioinspired computational frameworks, are inherently suited to handle the asynchronous and sparse characteristics of event data [2, 3]. The most common neuron model is the Leaky Integrate-and-Fire (LIF) model with iterative expression [6]. At each timestep t, the neurons in the l-th layer integrate the postsynaptic current $\boldsymbol{c}^l[t]$ with previous membrane potential $\boldsymbol{u}^l[t-1]$, the mathematic expression is illustrated in Equation (4):

$$\boldsymbol{u}^l[t] = (1-\frac{1}{\tau})\boldsymbol{u}^l[t-1] + \boldsymbol{c}^l[t], \tag{1}$$

where τ is the membrane time constant. $\tau > 1$ as the discrete step size is 1. The postsynaptic current $\boldsymbol{c}^l[t] = \mathcal{W}^l * \boldsymbol{s}^{l-1}[t]$ is calculated as the product of weights $\mathcal{W}^l$ and spikes from the preceding layer $\boldsymbol{s}^{l-1}[t]$, simulating synaptic functionality, with $*$ indicating either a fully connected or convolutional synaptic operation.

Neurons produce spikes $\boldsymbol{s}^l[t]$ via the Heaviside function Θ when the membrane potential $\boldsymbol{u}^l[t]$ surpasses the threshold V_{th}, as depicted in Equation (5):

$$\boldsymbol{s}^l[t] = \Theta(\boldsymbol{u}^l[t] - V_{\text{th}}) = \begin{cases} 1, & \text{if } \boldsymbol{u}^l[t] \geq V_{\text{th}} \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

After the spike, the neuron updates the membrane potential $\boldsymbol{u}^l[t]$ according to the reset mechanism as shown in Equation (3):

$$\boldsymbol{u}^l[t] = \boldsymbol{u}^l[t] - V_{\text{th}}\boldsymbol{s}^l[t], \tag{3}$$

where the $V_{\text{th}} \in \mathbb{R}$ is generally a global scalar that controls the firing and reset process for the neurons in each layers.

2. Theory Analysis of NCM

We first recall the standard LIF and LIF with Initialization from ANN (NCM), then we compare the graidents difference between LIF and the NCM.

Standard LIF Model The most common neuron model is the Leaky Integrate-and-Fire (LIF) model with iterative expression. At each timestep t, the neurons in the l-th layer

*Equal contribution.
†Corresponding author.

integrate the postsynaptic current $\boldsymbol{c}^l[t]$ with the previous membrane potential $\boldsymbol{u}^l[t-1]$, the mathematical expression is illustrated in Equation (4):

$$\boldsymbol{u}^l[t] = \beta\left(\boldsymbol{u}^l[t-1] - V_{\text{th}}\boldsymbol{s}^l[t-1]\right) + \boldsymbol{c}^l[t], \tag{4}$$

$$\boldsymbol{s}^l[t] = \Theta(\boldsymbol{u}^l[t] - V_{\text{th}}) = \begin{cases} 1, & \text{if } \boldsymbol{u}^l[t] \geq V_{\text{th}} \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $\beta \cdot \triangleq 1 - \frac{1}{\tau} \in (0,1)$. The postsynaptic current $\boldsymbol{c}^l[t] = \mathcal{W}^l * \boldsymbol{s}^{l-1}[t]$ is calculated as the product of weights $\mathcal{W}^l$ and spikes from the preceding layer $\boldsymbol{s}^{l-1}[t]$, simulating synaptic functionality, with $*$ indicating either a fully connected or convolutional synaptic operation. Neurons produce spikes $\boldsymbol{s}^l[t]$ via the Heaviside function Θ when the membrane potential $\boldsymbol{u}^l[t]$ surpasses the threshold V_{th}, as depicted in Equation (5), where the $V_{\text{th}} \in \mathbb{R}$ is generally a global scalar that controls the firing and reset process for the neurons in each layers.

LIF with Initialization from ANN. To fully elevates neuron activity across blurry areas, the threshold V'_{th} is redesigned according to the initial membrane potential as:

$$\boldsymbol{u}^l[t] = \begin{cases} V_{\text{init}} & , \text{ if } t = 0 \\ \beta\left(\boldsymbol{u}^l[t-1] - V'_{\text{th}} \odot \boldsymbol{s}^l[t-1]\right) + \boldsymbol{c}^l[t] & , \text{ otherwise} \end{cases}, \tag{6}$$

$$\boldsymbol{s}^l[t] = \Theta(\boldsymbol{u}^l[t] - V'_{\text{th}}) = \begin{cases} 1, & \text{if } \boldsymbol{u}^l[t] \geq V'_{\text{th}} \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

$$V'_{\text{th}} = 1 - \sigma(V_{\text{init}}), \tag{8}$$

where σ is the Sigmoid function, which rescales the initial feature to the range of 0 to 1. Unlike the global scalar threshold $V_{\text{th}} \in \mathbb{R}$ in vanilla LIF in Equation (4) and (5), the threshold $V'_{\text{th}} \in \mathbb{R}^{H \times W \times C}$ has the same dimension with the feature map V_{init}, providing more fine-grained control over the reset and firing processes of the neurons in the same layer.

Error backpropagation for LIF: To better present the derivation, we simplify the notation by using the subscript t and omitting the layer index l. For example, we use U_t to represent $\boldsymbol{u}^l[t]$. If there are errors in the expression $\frac{\partial \mathcal{L}}{\partial U_i}$ for $i = 1, \cdots, T$ from backpropagation in the previous layer, the error gradients with respect to U_t for the standard LIF model are calculated as follows:

$$\frac{\partial \mathcal{L}}{\partial U_t} = \frac{\partial \mathcal{L}}{\partial S_t} \cdot \frac{\partial S_t}{\partial U_t} + \sum_{i=t+1}^{T} \frac{\partial \mathcal{L}}{\partial S_i} \cdot \frac{\partial S_i}{\partial U_i} \prod_{d=1}^{T-i} \frac{\partial U_{t+d}}{\partial U_{t+d-1}} \tag{9}$$

$$= \frac{\partial \mathcal{L}}{\partial S_t} \cdot \Theta'(U_t - V_{\text{th}}) \tag{10}$$

$$+ \frac{\partial \mathcal{L}}{\partial S_{t+1}} \cdot \Theta'(U_{t+1} - V_{\text{th}}) \cdot \beta \tag{11}$$

$$+ \cdots \tag{12}$$

$$+ \frac{\partial \mathcal{L}}{\partial S_T} \cdot \Theta'(U_T - V_{\text{th}}) \cdot \beta^{(T-t)} \tag{13}$$

Error backpropagation for Init-LIF: The gradient propagation for initialization from the ANN is then expressed as:

$$\frac{\partial \mathcal{L}}{\partial U_t} = \frac{\partial \mathcal{L}}{\partial S_t} \cdot \Theta'(U_t - V'_{\text{th}}) \tag{14}$$

$$+ \frac{\partial \mathcal{L}}{\partial S_{t+1}} \cdot \Theta'(U_{t+1} - V'_{\text{th}}) \cdot \beta \tag{15}$$

$$+ \cdots \tag{16}$$

$$+ \frac{\partial \mathcal{L}}{\partial S_T} \cdot \Theta'(U_T - V'_{\text{th}}) \cdot \beta^{(T-t)} \tag{17}$$

Specifically, when the time step $t = 1$:

$$\frac{\partial \mathcal{L}}{\partial U_1} = \frac{\partial \mathcal{L}}{\partial S_1} \cdot \Theta'(U_1 - V'_{\text{th}}) + \cdots + \frac{\partial \mathcal{L}}{\partial S_T} \cdot \Theta'(U_T - V'_{\text{th}}) \cdot \beta^{(T-1)}. \tag{18}$$

The gradients are then backpropagated to the ANN feature map ($m = V_{\text{init}}$) as:

$$\frac{\partial \mathcal{L}}{\partial m} = \frac{\partial \mathcal{L}}{\partial U_1} \cdot \frac{\partial U_1}{\partial U_0} \cdot \frac{\partial U_0}{\partial m} + \sum_{i=1}^{T} \frac{\partial \mathcal{L}}{\partial S_i} \cdot \frac{\partial S_i}{\partial V_{\text{th}'}} \cdot \frac{\partial V_{\text{th}'}}{\partial m} \tag{19}$$

$$= \sum_{i=1}^{T} \frac{\partial \mathcal{L}}{\partial S_i} \cdot \Theta'(U_i - V'_{\text{th}}) \cdot \left(\beta^{i-1} - \sigma'(m)\right), \tag{20}$$

We can see that the error gradients backpropagated to the ANN module contain all event information, as they include all timestep membrane potential information, which is generated from the event inputs.

3. Datasets and Experiments

3.1. Datasets

We evaluate the BDHNet with GoPro, REBlur and MSRBD datasets with both synthetic and real-world scenarios.

GoPro: We evaluate the deblurring performance on GoPro dataset [5], which is the benchmark dataset for the image motion deblurring. It consists of 3214 pairs of blurry and sharp images, with 2103 pairs for training and 1111 pairs for testing. The resolution of all images is 1280×720

Figure 1. Qualitative comparisons under GoPro dataset. Best viewed on a screen and zoomed in.

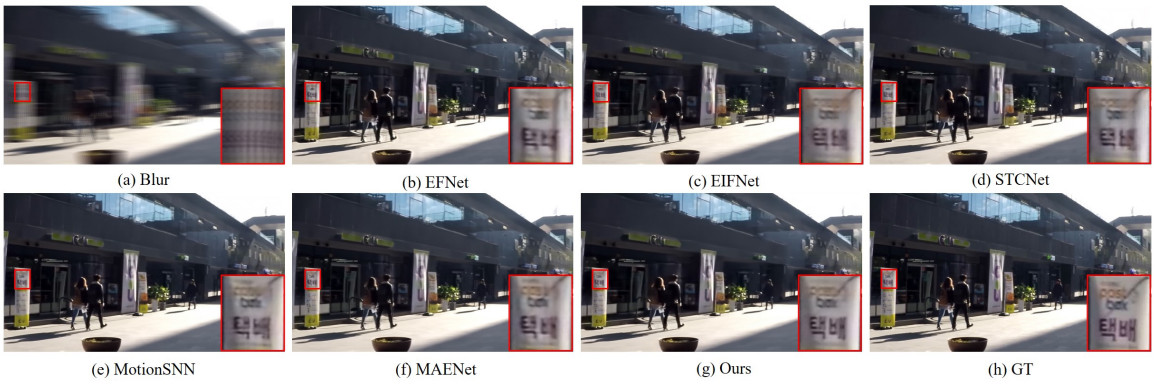

Figure 2. Qualitative comparisons under GoPro dataset. Best viewed on a screen and zoomed in.

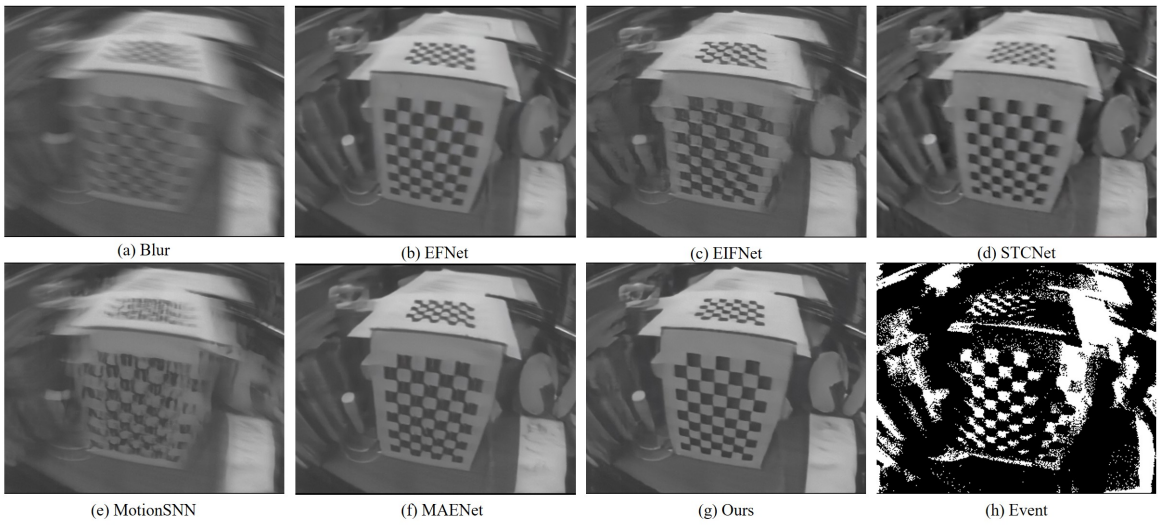

Figure 3. Qualitative comparisons under REBlur dataset. Best viewed on a screen and zoomed in.

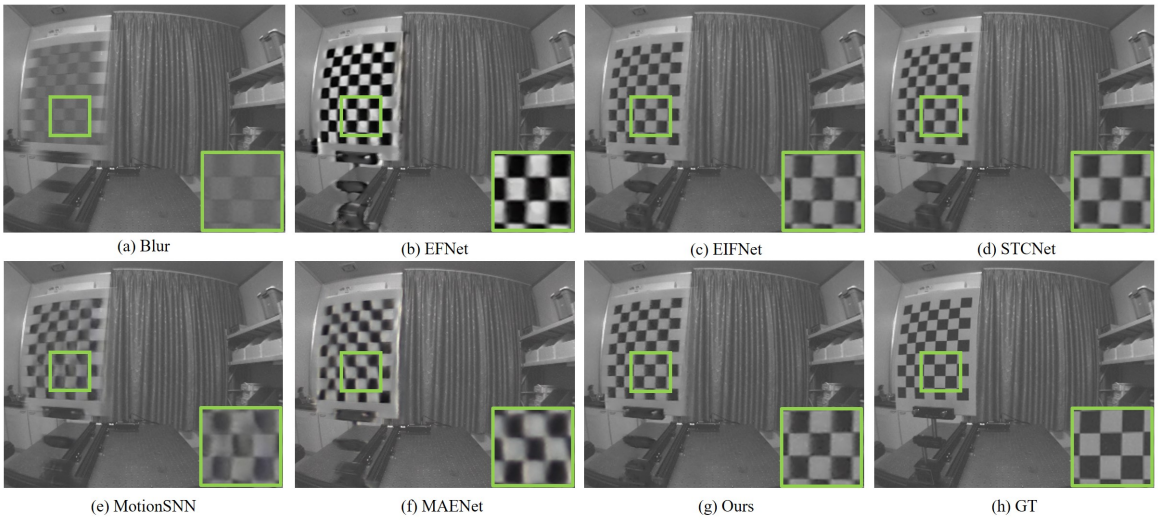

Figure 4. Qualitative comparisons trained with GoPro without fine-tuning under REBlur dataset. Best viewed on a screen and zoomed in.

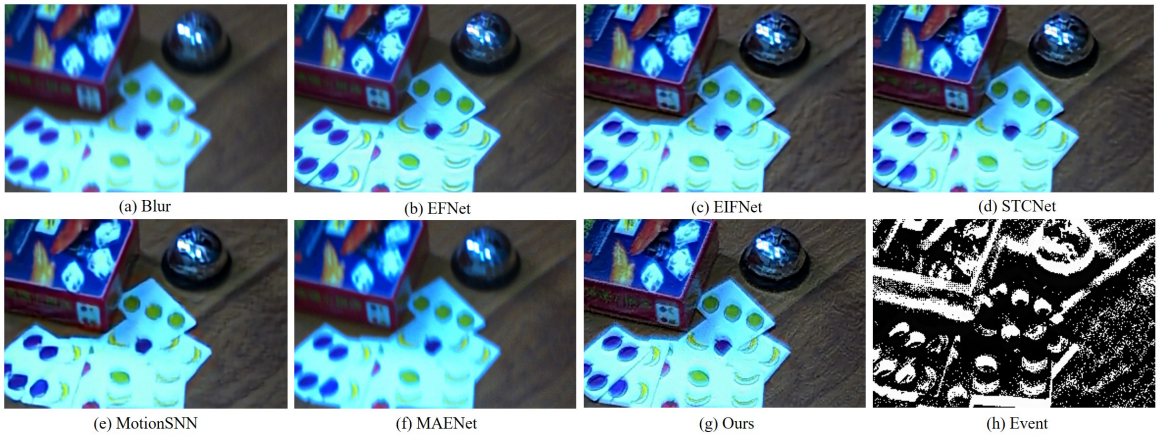

Figure 5. Qualitative comparisons trained with GoPro without fine-tuning in MS-RBD dataset. Best viewed on a screen and zoomed in.

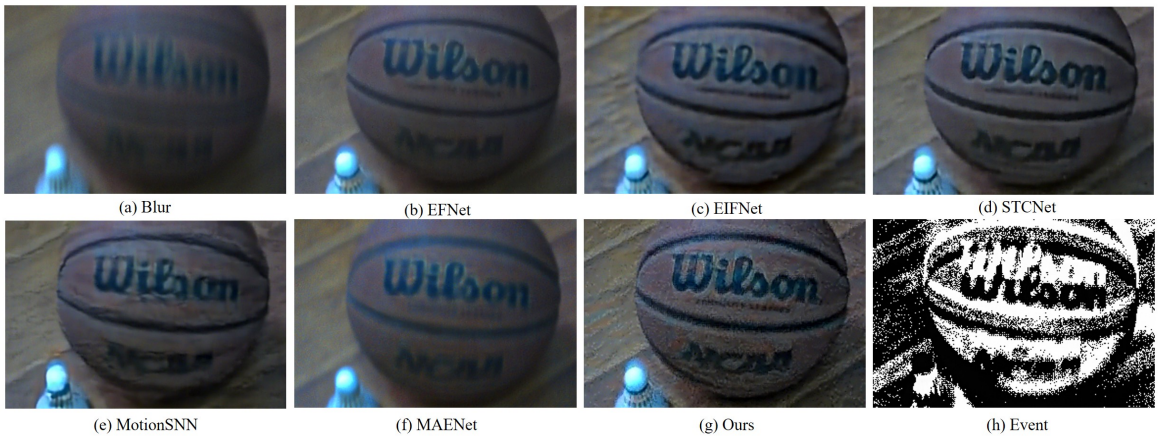

Figure 6. Qualitative comparisons trained with GoPro without fine-tuning in MS-RBD dataset. Best viewed on a screen and zoomed in.

and the blurry images are produced by averaging several adjacent high-speed sharp images. The event data is generated through the ESIM simulator. In this work, the raw event data is shaped into voxel-based representation for each image following EIFNet and the timestep in $\mathcal{V}$ is set to $b = 12$.

REBlur: REBlur dataset [5], captured by DAVIS for real event-based motion deblurring, comprises 1,389 image pairs with 486 designated for training and 903 for testing. It contains diverse linear and nonlinear indoor motions. Each image has a resolution of 260×360, consisting of real-world event data with the corresponding blurry and sharp images.

MS-RBD: MS-RBD dataset [7] is the multi-scale blurry dataset captured in the real-world scenario. The dataset contains 32 sequences of data with 22 indoor and 10 outdoor scenes. The resolution of all images is 288×192 with the corresponding events. We evaluate the deblurring performance on MS-RBD with a focus on the generalization ability in the real-world scenes, where the blur caused by camera ego-motion and dynamic scenes.

3.2. Supplementary Comparisons

Figure 1 and Figure 2 are the comparison results in the GoPro dataset. Figure 3 is the deblurring results in the REBlur dataset. As certain studies release only the weights trained on the GoPro dataset, we seek to ensure an impartial evaluation of the algorithmic performance and assess the generalization capabilities. To this end, we present the visualization results of inference on the REBlur and MS-RBD datasets, utilizing the GoPro-trained weights without fine-tuning, as shown in Figure 4, Figure 5, and Figure 6.

References

[1] Per Brodal. *The central nervous system*. oxford university Press, 2010. 1

[2] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34:21056–21069, 2021. 1

[3] Yulong Huang, LIN Xiaopeng, Hongwei Ren, FU Haotian, Yue Zhou, LIU Zunchang, Bojun Cheng, et al. Clif: Complementary leaky integrate-and-fire neuron for spiking neural networks. In *Forty-first International Conference on Machine Learning*. 1

[4] Nancy Kanwisher and Ewa Wojciulik. Visual attention: insights from brain imaging. *Nature reviews neuroscience*, 1 (2):91–100, 2000. 1

[5] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*, pages 412–428. Springer, 2022. 2, 5

[6] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018. 1

[7] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Generalizing event-based motion deblurring in real-world scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10734–10744, 2023. 5