SFOD: Spiking Fusion Object Detector

Supplementary Material

A. Derivation of Classification Loss Function Gradients

In this section, we present the derivation of gradients for the Mean Squared Error (MSE) and Cross-Entropy (CE) loss functions during backpropagation. To simplify, we assume the sample size of 1 and accordingly adjust the notation. The simplified formulas for MSE and CE are presented as Equations A and B, respectively.

$$\text{MSE} = \sum_{j=1}^{C}(y_j - a_j)^2 \tag{A}$$

$$\text{CE} = -\sum_{j=1}^{C} y_j \log(z_j) \tag{B}$$

A.1. Derivation of Mean Squared Error Loss Function Gradients

The gradient derivation of the MSE loss function is detailed in Equation C. From this derivation, it can be concluded that the gradients of the MSE loss function are proportional to the difference between the decoded values and the labels.

$$\frac{\partial \text{MSE}}{\partial a_j} = 2(a_j - y_j) \tag{C}$$

A.2. Derivation of Cross-Entropy Loss Function Gradients

The softmax function is shown in Equation D:

$$z_t = \frac{e^{a_t}}{\sum_{j=1}^{C} e^{a_j}} \tag{D}$$

Define k as the index where $y_k = 1$, the gradients of the CE loss function for $j = k$ is given by Equation E:

$$\begin{aligned}\frac{\partial \text{CE}}{\partial a_j} &= \frac{\partial \text{CE}}{\partial a_k} = \frac{\partial \text{CE}}{\partial z_k}\frac{\partial z_k}{\partial a_k} \\ &= -\frac{1}{z_k}\frac{(e^{a_k})\sum_{j=1}^{C} e^{a_j} - e^{a_k}e^{a_k}}{\left(\sum_{j=1}^{C} e^{a_j}\right)^2} \\ &= -\frac{1}{z_k} z_k(1 - z_k) \\ &= z_k - 1\end{aligned} \tag{E}$$

The gradients of the CE loss function for $j \neq k$ is given by Equation F:

$$\begin{aligned}\frac{\partial \text{CE}}{\partial a_j} &= \frac{\partial \text{CE}}{\partial z_k}\frac{\partial z_k}{\partial a_j} \\ &= -\frac{1}{z_k}\frac{0 \cdot \sum_{j=1}^{C} e^{a_j} - e^{a_k}e^{a_j}}{\left(\sum_{j=1}^{C} e^{a_j}\right)^2} \\ &= z_j\end{aligned} \tag{F}$$

Equations E and F can be combined as follows:

$$\frac{\partial \text{CE}}{\partial a_j} = z_j - y_j \tag{G}$$

From this derivation, it can be concluded that the gradients of the CE loss function are proportional to the difference between the post-softmax probability values and the labels.

B. Energy Consumption

The low energy consumption advantage of SNNs mainly stems from performing accumulation calculation (AC) only when neurons fire. According to Section 4.3, although our model includes multiplication and addition (MAC) operations, such calculations are minimal and rarely occur in our experiments. Thus, MAC is not considered when evaluating the energy consumption of SNNs. In non-SNNs, network computations primarily rely on MAC operations. Although there are some AC operations, their limited number and significantly lower energy consumption compared to MAC allow us to disregard them for simplicity. In Table 5, this is indicated by placing a greater than sign before the corresponding energy consumption values. Furthermore, in accordance with [9], we assume that data for different computations are realized as 32-bit floats in 45nm technology, with $E_{\text{MAC}} = 4.6pJ$ and $E_{\text{AC}} = 0.9pJ$. The formulas for calculating the energy consumption of SNNs and non-SNNs are presented as Equations H and I, respectively.

$$E_{\text{SNNs}} = T \times fr \times E_{\text{AC}} \times N_{\text{AC}} \tag{H}$$

$$E_{\text{non-SNNs}} = T \times E_{\text{MAC}} \times N_{\text{MAC}} \tag{I}$$

C. More Implementation Details

On the NCAR dataset, to ensure consistent spatial dimensions for all samples, we employ the nearest-neighbor interpolation method to resize them to a resolution of 64x64.

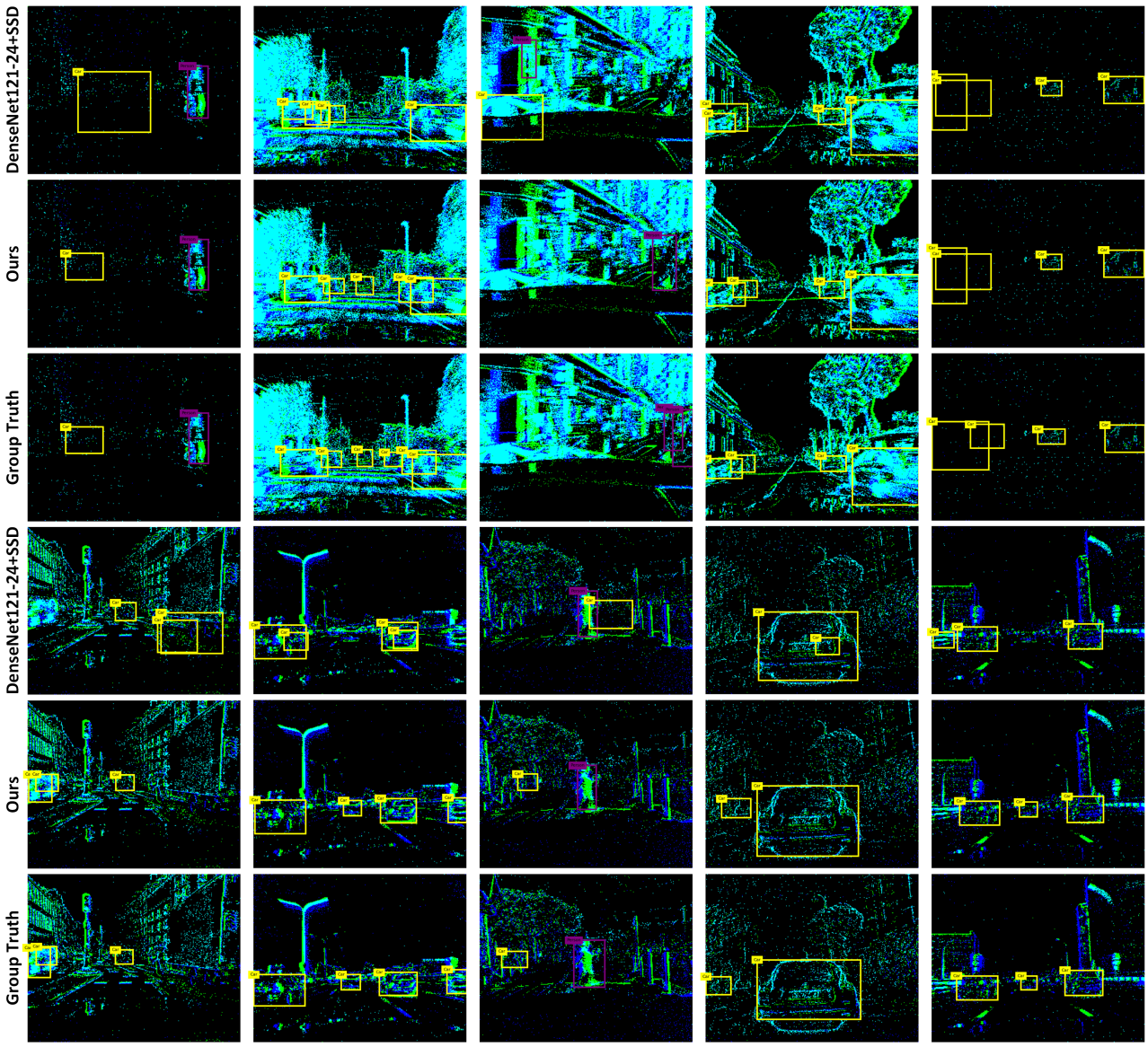

Figure A. **More visual comparison results on the GEN1 dataset.**

On the GEN1 dataset, given the 100 ms duration of NCAR dataset samples, accordingly, we extract the first 100 ms of each annotated box from this dataset as a sample. Furthermore, the samples in the GEN1 dataset have a consistent spatial resolution of 304x240. For the evaluation of the GEN1 dataset, following the criteria established in previous works [1, 3, 6, 8], we exclude bounding boxes with side lengths less than 10 pixels or diagonals shorter than 30 pixels.

Prior to each convolutional layer, we add a Batch Normalization (BN) layer [5], recommended by [1] for better performance and faster convergence. All convolutional layers are initialized using the He Initialization method [4], while biases in BN layers are set to 0, and weights are set to 1. The membrane time constant τ for PLIF neurons [2] is initialized to 2, with the threshold set to 1. To mitigate gradient explosion, gradient norms are clipped at a maximum value of 1. We opt for Atan as the gradient surrogate function. All model training is executed on a single Nvidia A40 GPU. Furthermore, considering the class imbalance between the foreground and background in detection model training, we employ the Focal Loss [7] as the loss function. The formula for the Focal Loss can be described as Equation J (for simplicity, consider the binary classification case). In this formula, p_t as shown in Equation K represents the predicted probability of a sample belonging to its true category.

α_t is a weighting factor for category balancing as shown in Equation L. Lastly, γ is a hyper-parameter that is adjustable to fine-tune the behavior of the loss function.

$$FL = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{J}$$

$$p_t = \begin{cases} p & \text{if } y = 1, \\ 1-p & \text{otherwise.} \end{cases} \tag{K}$$

$$\alpha_t = \begin{cases} \alpha & \text{if } y = 1, \\ 1-\alpha & \text{otherwise.} \end{cases} \tag{L}$$

D. More Visualization

In this section, we present more visual comparisons of our model with DenseNet121-24+SSD [1] and the Ground Truth across a broader range of scenarios in Figure A, further demonstrating the robust detection capabilities of SFOD in event camera object detection.

References

[1] 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. 2, 3

[2] 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. 2

[3] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13884–13893, 2023. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 2

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 2

[6] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. 2

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[8] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020. 2

[9] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6555–6565, 2023. 1