南京航空航天大學

Nanjing University of Aeronautics and Astronautics

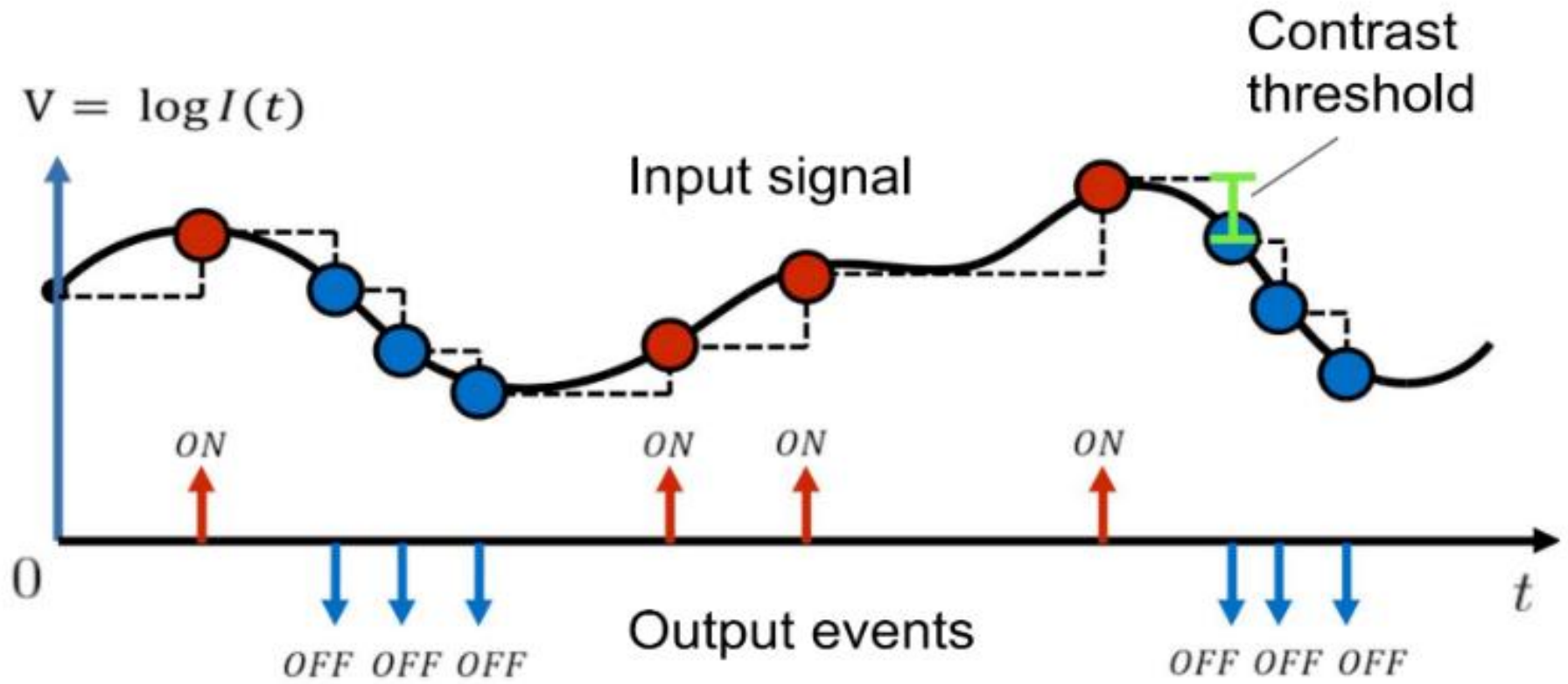
Event Transformer

Event Transformer

Zhihao Li[1], M. Salman Asif[2], and Zhan Ma[1]

[1] Vision Lab, Nanjing University, China
[2] University of California at Riverside, USA
lizhihao6@smail.nju.edu.cn, sasif@ece.ucr.edu,
mazhan@nju.edu.cn

Event camera

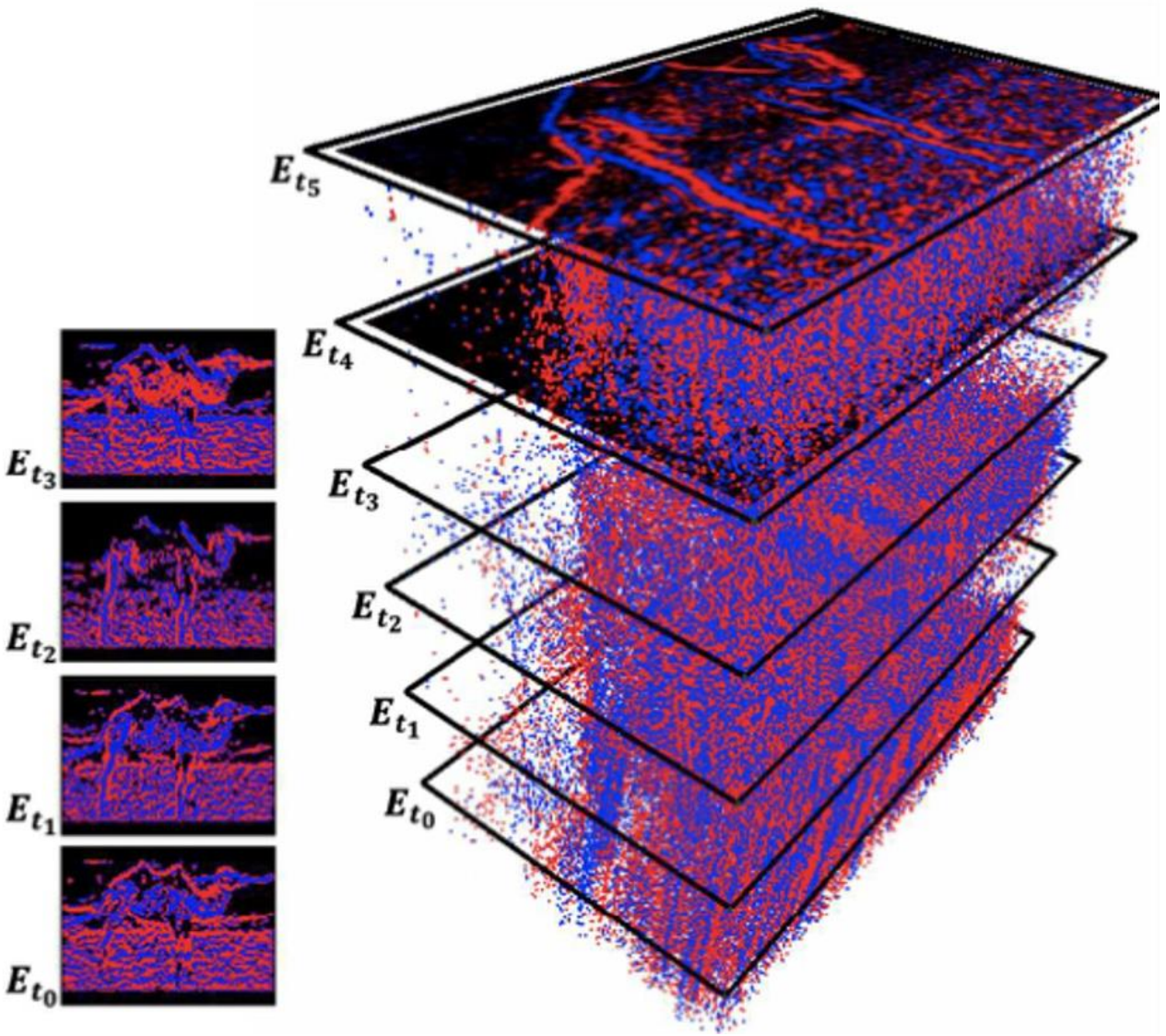


log I(x, y, t)- log I(x, y, t-Δt) = ±C

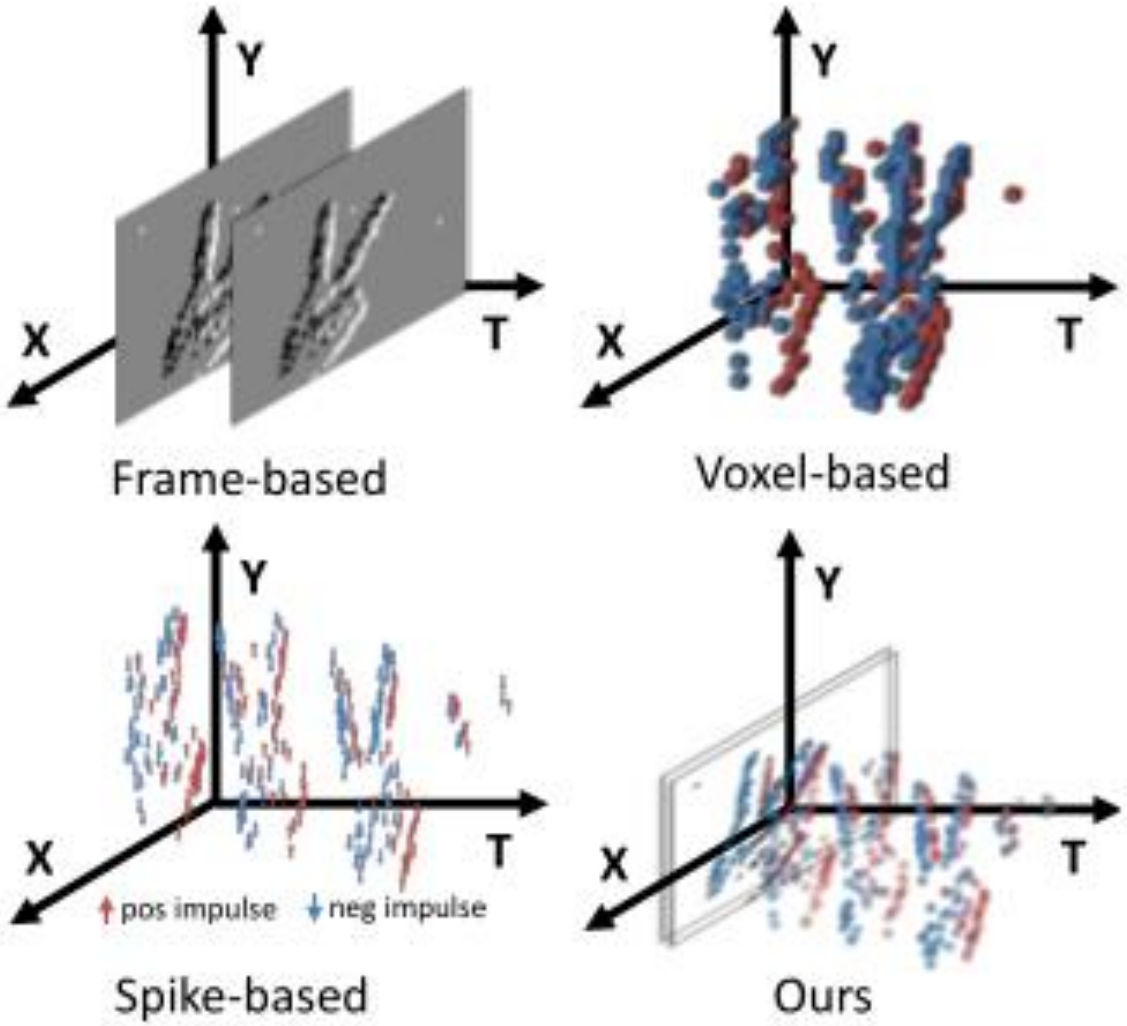
$$\mathbf{e}_k \triangleq (\mathbf{x}_k, t_k, p_k)$$

Event set: $\mathcal{E} = \{\mathbf{e}_k\}_{k=1}^{N}$

Event camera



Event camera



(a) Representation Forms

Event camera

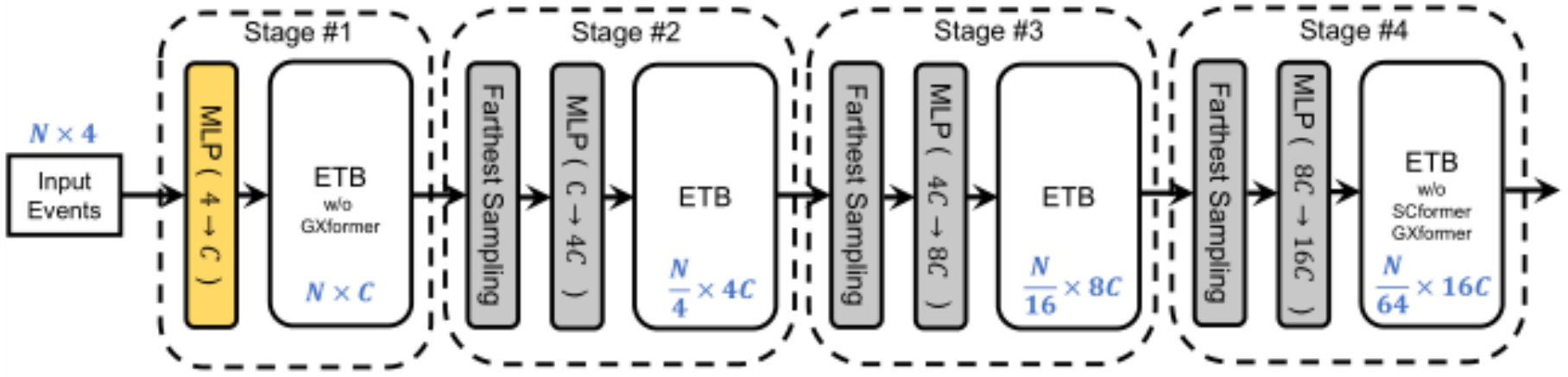
merit

1. Asynchronous
    a. Pixel-wise independent
    b. Brightness change
    c. Low bandwidth
2. Low latency
    a. No motion blur
3. High dynamic range (120dB)
4. Low power (mW)

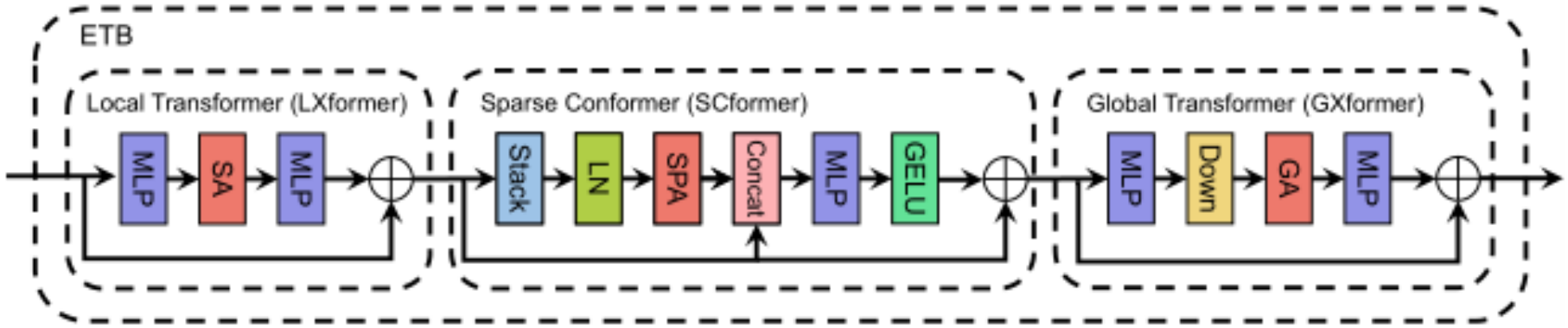
shortcoming

1. Data is asynchronous and not easy to process
2. Lack of texture information about the object

Method



(a) Event Transformer Backbone



(b) Event Transformer Block

farthest point down-sampling

1. The input set has N points, select a point p_0 from the set as the starting point, and obtain the sampling point set S={P0};

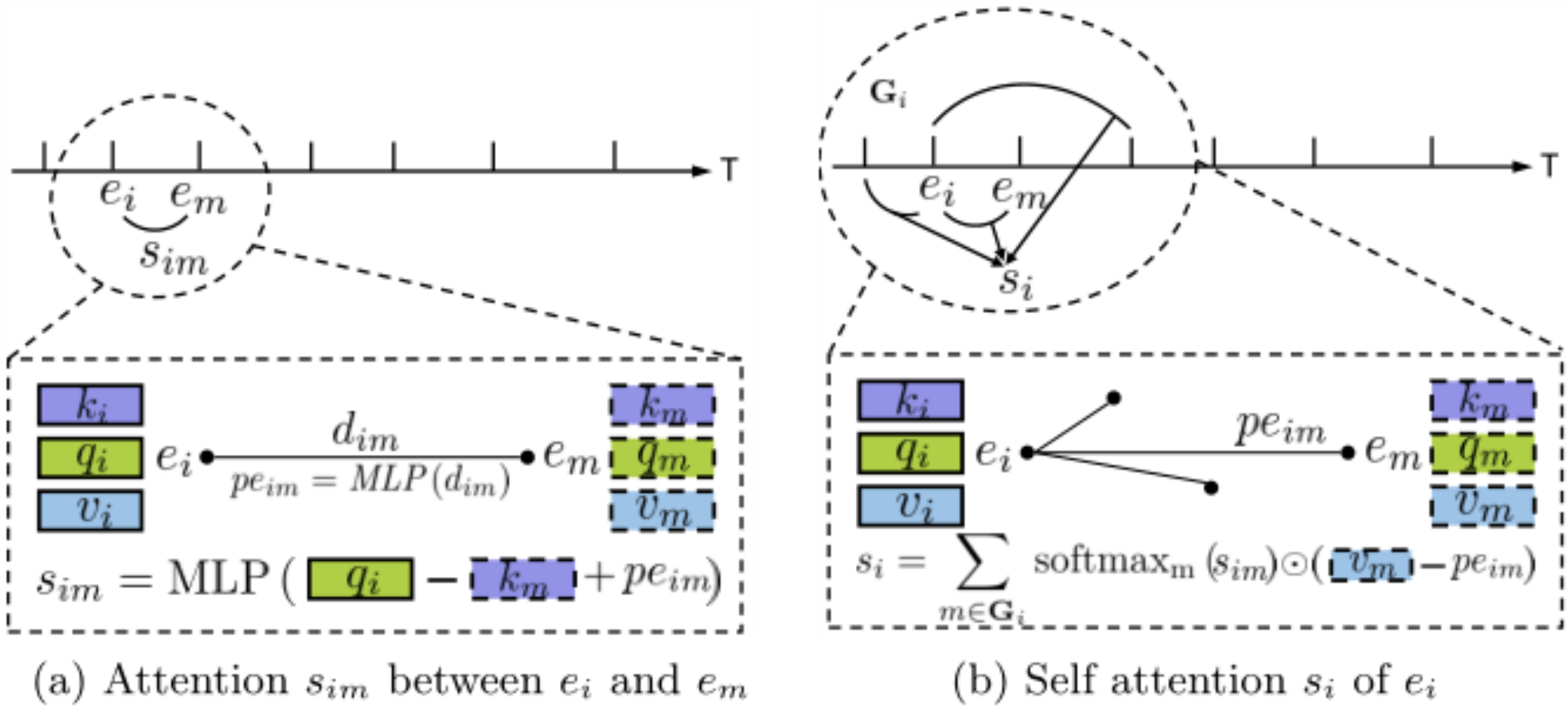
2. Calculate the distance from all points to p_0 , form an N-dimensional array L, select the point corresponding to the maximum value as p_1 , and update the sampling point set S={P0, P1};

3. calculate the distance from all points to p_1 , for each point p_i , if its distance from p_1 is less than L[i], then update L[i] = d(Pi, P1), therefore, the array L is stored in the nearest distance from each point to the sample point set S;

4. Select the point corresponding to the maximum value in L as P2, and update the sampling point set S={P0, P1, P2};

5. Repeat steps 2-4 until N target sampling points.

Local Transformer



(a) Attention s_{im} between e_i and e_m (b) Self attention s_i of e_i

Fig. 4: **Self attention in LXformer. (a)** The attention s_{im} between e_i and e_m is calculated using query q_i of e_i, key k_m of e_m and the distance d_{im} between e_i and e_j. **(b)** The self-attention s_i is calculated using the set G_i of events that are temporally close to event e_i.

Local Transformer

$$\mathbf{E} = \{e_1, \ldots, e_N\} \in R^{N \times 4} \qquad \mathbf{F} = \{f_1, \ldots, f_N\} \in R^{N \times C}$$

Q, K, and V

$$\mathbf{D} = \{d_{11}, \ldots, d_{ij}, d_{NN}\} \in R^{N \times N \times 4}$$

PE = MLP (D)

$$s_{im} = \mathrm{MLP}\left(q_i - k_m + pe_{im}\right), \tag{2}$$

$$s_i = \sum_{m \in \mathbf{G}_i} \mathrm{softmax}_m\left(s_{im}\right) \odot \left(v_m + pe_{im}\right). \tag{3}$$

$$y_i = f_i + \mathrm{MLP}\left(s_i\right). \tag{4}$$

Sparse Conformer

The sequence of events $\{(e_i, f_i)\}_{i=1}^{N}$ is converted to a 2D frame $I \in R^{H \times W \times (2+C)}$

L = LN (I) Q =SPConv$_q$(L), K = SPConv$_k$(L), V = SPConv$_v$(L)

event e_i and event $e_j \in$ W

$$s_{ij} = \mathrm{MLP}\left(q_{y_i,x_i} - k_{y_j,x_j} + pe_{ij}\right), \tag{5}$$

$$pe_{ij} = \mathrm{MLP}\left((y_i, x_i, p_i) - (y_j, x_j, p_j)\right)$$

$$s_i = \sum_{j \in W_i} \mathrm{softmax}_j\left(s_{ij}\right) \odot \left(v_j + pe_{ij}\right). \tag{6}$$

$$y_i = f_i + \phi\left(\mathrm{concat}\left(f_i, \mathrm{MLP}\left(s_i\right)\right)\right). \tag{7}$$

Global Transformer

The farthest point down-sampling with down-sample rate r → ˆE, for each $\boldsymbol{e}_j \in$ ˆE, its corresponding feature ˆ$\boldsymbol{f}_j \in$ ˆF is calculated as

$$\hat{f}_j = \max_k \left(\mathrm{MLP} \left(\mathrm{concat} \left(e_k, f_k \right) \right) \right), \tag{8}$$

$$s_i = \sum_{j \in \hat{\mathbf{E}}} \mathrm{softmax}_j \left(s_{ij} \right) \odot \left(\hat{v}_j + pe_{ij} \right), \tag{9}$$

$$y_i = f_i + \mathrm{MLP} \left(s_i \right). \tag{10}$$

Experiment

Table 2: Accuracy of Event-based Classification Using Various Datasets. N-M refers to N-MNIST, N-Cal to N-Caltech101, CIF10 to CIFAR10-DVS, N-C to N-Cars, and ASL to ASL-DVS.

Method	**Type**	**N-M** [21]	**N-Cal** [33]	**CIF10** [27]	**N-C** [39]	**ASL** [3]
H-First [34]	frame	0.712	0.054	0.077	0.561	-
HOTS [24]	frame	0.808	0.210	0.271	0.624	-
HATS [39]	frame	0.991	0.642	0.524	0.902	-
EST [18]	frame	0.990	0.753	0.634	0.919	0.979
AMAE [11]	frame	0.983	0.694	0.620	0.936	0.984
M-LSTM [5]	frame	0.986	0.738	0.631	0.927	0.980
MVF-Net [10]	frame	0.981	0.687	0.599	0.927	0.971
ResNet-50 [20]	frame	0.984	0.637	0.558	0.903	0.886
EventNet [37]	voxel	0.752	0.425	0.171	0.750	0.833
RG-CNNs [3]	voxel	0.990	0.657	0.540	0.914	0.901
EV-VGCNN [9]	voxel	0.994	0.748	0.651	0.953	0.983
VMV-GCN [43]	voxel	0.995	0.778	0.663	0.932	0.989
Gabor-SNN [26]	spike	0.837	0.196	0.245	0.789	-
PLIF [15]	spike	0.996	-	0.697	-	-
Ours	tensor	**0.999**	**0.789**	**0.712**	**0.954**	**0.999**

Experiment

Table 3: Model Complexity of Different Methods

Method	Type	Parameters	FLOPs	Top-1 Accuracy
EST [18]	frame	21.38M	4.28G	52.4
M-LSTM [5]	frame	21.43M	4.82G	63.1
MVF-Net [10]	frame	33.62M	5.62G	59.9
ResNet-50 [20]	frame	25.61M	3.87G	55.8
EventNet [37]	voxel	2.81M	0.91G	17.1
RG-CNNs [3]	voxel	19.46M	0.79G	54.0
EV-VGCNN [9]	voxel	0.84M	0.70G	65.1
VMV-GCN [43]	voxel	0.86M	1.30G	66.3
Ours	tensor	15.87M	0.51G	71.2

Ablation Study

Table 6: Ablation Study of Closest Events in LXformer

M **Closest Events in LXformer**	4	8	**16**	32	64
Top-1 Accuracy	65.8	69.0	**71.2**	70.9	69.6

Table 7: Ablation Study of Window Size in SCformer

Size of W **in SCformer**	1×1	3×3	5×5	7×7	9×9
Top-1 Accuracy	69.9	**71.2**	71.1	71.0	70.6

Table 8: Ablation Study of Down-sample Rate in GXformer

Down-sample rate r **in GXformer**	1	8	**32**	64	128
Top-1 Accuracy	69.8	70.3	**71.2**	70.6	70.1

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Thanks for Listening