Supplementary Material for EDeNN: Event Decay Neural Networks for low latency vision

Celyn Walters
https://www.surrey.ac.uk/people/celyn-walters
Simon Hadfield
https://www.surrey.ac.uk/people/simon-hadfield

Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, UK

1 Scalar regression with EDeNNs - Network details

This section relates to section 4.1 from the main submission. The objective was to predict X, Y and Z angular velocity from an event stream.

The network consists of 4 encoder layers followed by a fully connected layer. The encoder layers are made up of Event Decay Convolution (EDeC) kernels with our new formulation for partial convolutions to cater to sparse event data. The decoder layers consist of nearest-neighbour upsampling followed by 2D transpose convolutions. The activation function used in each layer was CELU [1]. Table 1 shows layer structure and output tensor shapes. The code was trained with supervision from the dataset of [2] with an initial learning rate of 1.0.

Layer type	Shape (C, D, H, W)
(Input)	2, 100, 180, 240
Encoder layer 1	16, 100, 89, 119
Encoder layer 2	32, 100, 44, 59
Encoder layer 3	64, 100, 21, 29
Encoder layer 4	128, 100, 10, 14
Bottleneck	256, 100, 8, 12
Fully connected	3, 100, 1, 1

Table 1: Output tensor shapes for each layer in the Event Decay Neural Network (EDeNN) for the optical flow task. Input layer consists of positive and negative events at the image resolution, and was padded from 346 × 260 to 352 × 272 for perfect division in the deeper layers.

2 Dense estimation with EDeNNs - Network details

This section relates to section 4.2 from the main submission. The objective was to predict dense optical flow from an event stream.

The network consists of 4 encoder layers, a bottleneck layer, 4 decoder layers, and a fully connected layer. The encoder layers are made up of EDeC kernels with our new formulation for partial convolutions to cater to sparse event data. The decoder layers consist of nearest-neighbour upsampling followed by 2D transpose convolutions. The activation function used in each layer was CELU [1]. The best model was trained with supervision from the MVSEC dataset [9] with an initial learning rate of 0.01. For the evaluation, pixel regions without input events or ground truth were masked, which is typical in other approaches. Table 2 shows layer structure and output tensor shapes.

Layer type	Shape (C, D, H, W)
(Input)	2, 100, 272, 352
Encoder layer 1	16, 100, 136, 176
Encoder layer 2	32, 100, 68, 88
Encoder layer 3	64, 100, 34, 44
Encoder layer 4	128, 100, 17, 22
Bottleneck	128, 100, 17, 22
Decoder layer 4 (prediction)	2, 100, 17, 22
Decoder layer 4	256, 100, 34, 44
Decoder layer 3 (prediction)	2, 100, 34, 44
Decoder layer 3	96, 100, 68, 88
Decoder layer 2 (prediction)	2, 100, 68, 88
Decoder layer 2	64, 100, 136, 176
Decoder layer 1 (prediction)	2, 100, 136, 176
Decoder layer 1	40, 100, 272, 352
Fully connected	2, 100, 272, 352

Table 2: Output tensor shapes for each layer in the EDeNN for the optical flow task. Input layer consists of positive and negative events at the image resolution, and was padded from 346 × 260 to 352 × 272 for perfect division in the deeper layers.

Table 3 shows the results seen in Figure 3 from the main submission. E-RAFT was evaluated on the original ground truth from the MVSEC dataset. The implementations of EV-FlowNet, FireNet and FireFlowNet are from [4]. The hardware used for obtaining the results and step times was an Intel i9-10900K with an NVIDIA GeForce RTX 3090.

Approach	AEE	$\%_{\text{outlier}}$	Resolution	t (ms)	t/res. $\times 10^9$ (ms)
E-RAFT [3]	0.46	0.49	$256 \times 256 \times 15$	0.0326	33.1459
EV-FlowNet [8]	0.47	0.25	$128 \times 128 \times 2$	0.0042	128.0029
FireNet [7]	0.55	0.35	$128 \times 128 \times 2$	0.0015	45.6160
FireFlowNet [6]	1.02	1.62	$128 \times 128 \times 2$	0.0010	30.9749
LIF-EV-FlowNet [4]	0.53	0.35	$128 \times 128 \times 2$	0.0063	191.8583
LIF-FireNet [4]	0.57	0.40	$128 \times 128 \times 2$	0.0023	69.5087
LIF-FireFlowNet [4]	0.84	1.15	$128 \times 128 \times 2$	0.0021	65.3463
2D CNN, partial (ours)	2.22	25.53	$352 \times 272 \times 24$	0.0070	3.0249
EDeNN, partial [5]	1.06	4.64	$352 \times 272 \times 24$	0.0133	5.7879
EDeNN, partial (ours)	0.82	2.20	$352 \times 272 \times 24$	0.0129	5.5939

Table 3: Comparison of event-based optical flow approaches on the test sequence 'outdoor_day1' from the MVSEC dataset [9]. t represents the average step time for the forward pass over the test sequence on identical hardware.

References

[1] Jonathan T. Barron. Continuously Differentiable Exponential Linear Units. (3):1–2, 2017. URL http://arxiv.org/abs/1704.07483.

[2] Mathias Gehrig, Sumit Bam Shrestha, Daniel Mouritzen, and Davide Scaramuzza. Event-based angular velocity regression with spiking networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[3] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. Dense Optical Flow from Event Cameras. In *IEEE Int. Conf. 3D Vis.(3DV)*, 2021.

[4] Jesse Hagenaars, Federico Paredes-Vallés, and Guido de Croon. Self-Supervised Learning of Event-Based Optical Flow with Spiking Neural Networks. *Advances in Neural Information Processing Systems*, 34(NeurIPS), 2021.

[5] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting Chun Wang, Andrew Tao, and Bryan Catanzaro. Image Inpainting for Irregular Holes Using Partial Convolutions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11215 LNCS:89–105, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01252-6_6.

[6] Federico Paredes-Valles and Guido C. H. E. de Croon. Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy. pages 3445–3454, 2021. doi: 10.1109/cvpr46437.2021.00345.

[7] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert E. Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 156–163, 2020. doi: 10.1109/WACV45572.2020.9093366.

[8] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, jun 2018. ISBN 978-0-9923747-4-7. doi: 10.15607/RSS.2018.XIV.062.

[9] Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The Multivehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, jul 2018. ISSN 2377-3766. doi: 10.1109/LRA.2018.2800793.