Supplementary Material for
Event-guided Person Re-ID via Sparse-Dense Complementary Learning

Chengzhi Cao[1], Xueyang Fu[1*], Hongjian Liu[1], Yukun Huang[1],
Kunyu Wang[1], Jiebo Luo[2], Zheng-Jun Zha[1]
[1]University of Science and Technology of China, China
[2]University of Rochester, USA
{chengzhicao@mail., xyfu@, jeffeey@mail., kevinh@mail., kunyuwang@mail.}ustc.edu.cn,
jluo@cs.rochester.edu, zhazj@ustc.edu.cn

1. Dataset

Because there is no corresponding event sequences in person Re-ID, we apply a display-camera system and simulator V2E [3] to generate events from three classical video-based datasets for Re-ID, including PRID-2011, iLIDS-VID and MARS. Some details are as follows.

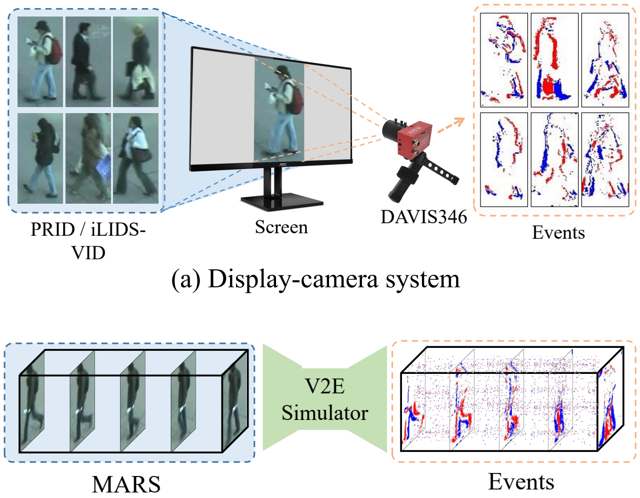
PRID-2011 [2]. There are 385 and 749 identities from two different cameras respectively. The video length ranges from 5 to 675 frames. To get corresponding event sequences from videos, we apply a display-camera system [1, 5] to record real-scenario person Re-ID data. Specifically, we utilize a screen (IPS, resolution 1920 × 1080, 144Hz) and DAVIS346 color camera. The screen is set at a 50-centimeter distance from the event camera. The display-camera system is shown in Figure 1(a).

iLIDS-VID [6]. 300 people were captured by two non-overlapping cameras, resulting in 600 image sequences. We apply the same display-camera system to record events.

MARS [8]. It contains 17,503 tracklets from 1261 identities and 3,248 low-quality tracklets captured by 6 cameras as distracters. There are 625 identities in the training set and 636 identities in the testing set. Because of the large scale of MARS, we use the event simulator V2E [3] to generate synthetic event sequences. V2E employs an accurate DVS model to synthesize event data from any real or synthetic traditional frame-based video. It can also use Super-SloMo [4] to upsample the temporal resolution of a normal camera video.

2. Analysis of Spiking Neural Network

We evaluate the contribution of each component and compare with standard Res-block to demonstrate the efficiency of our network. Our baseline has six Res-Block to extract event features. The results are reported in Table 1. As we can see, when we set the number of SNN from 1 to 2, the mAP and Rank-1 accuracy increases, but when the number of SNN is 3, the performance drops slightly. It clearly demonstrates the performance degradations in deep SNNs because the number of spikes drastically vanish at deeper layers.



(a) Display-camera system

(b) Video-to-Event Simulator

Figure 1. The generation of different dataset. (a) we deploy a display-camera system to observe real-scenario event data for PRID-2011 and iLIDS-VID dataset. (b) For MARS, we use the simulator V2E [3] to produce corresponding event streams.

Moreover, we add some Res-block in the following layers, but the mAP decreases slightly, which means that traditional convolution has destroyed the spatial distribution of event streams. It is not suitable for our complementary learning strategy. When we deploy deformable mapping in the deeper layers, the mAP and Rank-1 accuracy in PRID-2011 and iLIDS-VID dataset still increases continuously. This phenomenon illustrates that our deformable

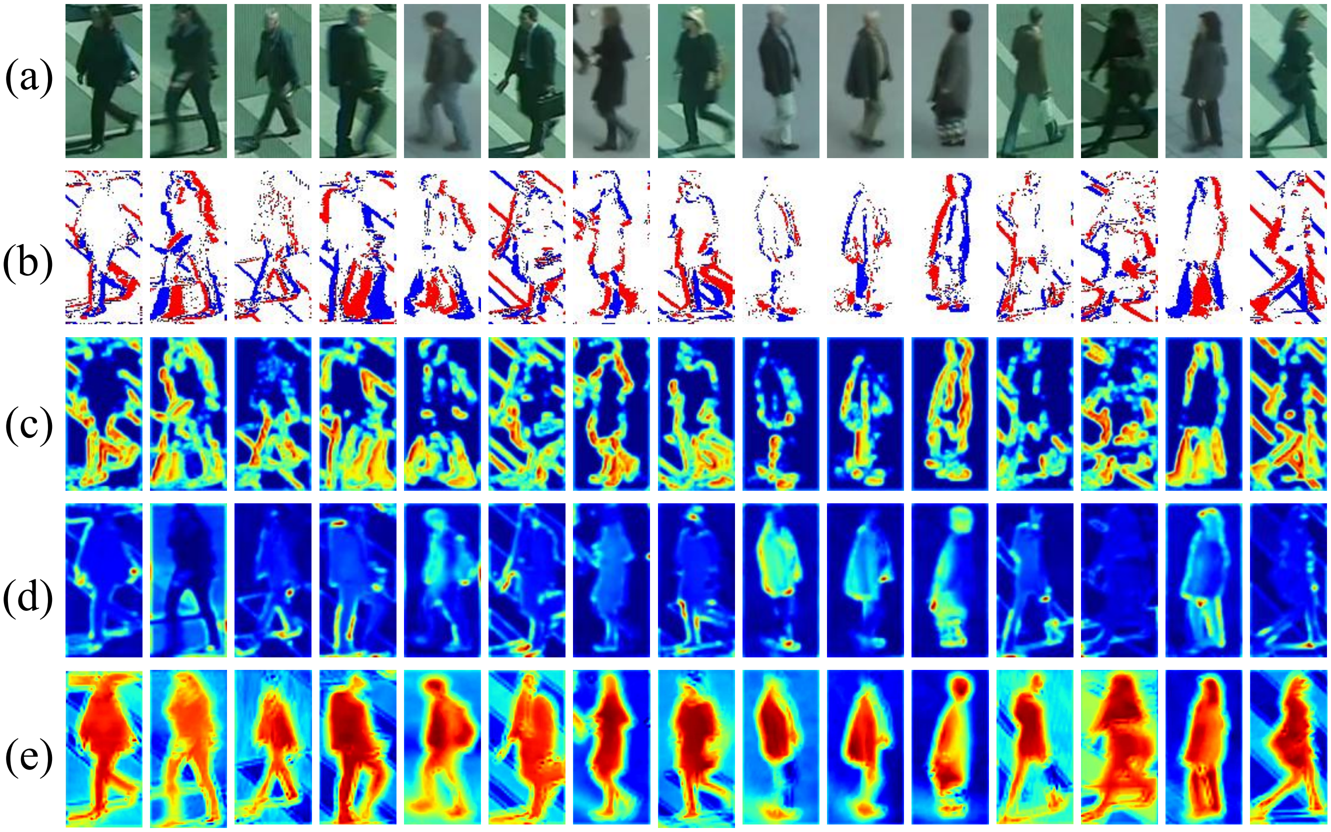


Figure 2. Visual examples of learned feature maps. From top to bottom: (a) original images, (b) corresponding events, (c) feature maps of events, (d) feature maps of PSTA [7] (w/o events), (e) feature maps of our network (w/ events).

Table 1. Quantitative analysis on different components in Spiking Neural Network. ”SN” and ”DM” represent spiking neuron and deformable mapping, respectively. ”Res” means Res block.

Modules			PRID-2011		iLIDS-VID	
Res	SN	DM	mAP	rank-1	mAP	rank-1
6	0	0	80.3	71.7	77.4	67.3
0	1	0	81.1	72.6	78.2	68.7
0	2	0	86.2	78.6	83.9	77.3
0	3	0	84.5	76.7	82.1	74.7
5	1	0	82.8	72.5	78.1	69.3
4	2	0	84.6	76.8	78.4	70.7
3	3	0	84.0	76.4	75.8	64.0
0	2	1	92.7	88.8	87.8	81.3
0	2	2	93.4	89.9	89.1	84.0
0	2	3	96.9	96.5	93.2	92.7
0	2	4	95.7	94.4	91.1	87.3

mapping operation can maintain the sparse distribution in event streams to provide extra brightness information.

3. Visualization

In this section, we show additional feature maps in our frameworks to demonstrate the effectiveness of events. As shown in the second row of Figure 2, for occluded frames, the feature of the baseline can not focus on the regions of pedestrians due to the influence of occlusions. But when adding events into the baseline, it still focuses on the better represented areas of each person.

References

[1] Peiqi Duan, Zihao W Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventzoom: Learning to denoise and super resolve neuromorphic events. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12824–12833, 2021. 1

[2] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person Re-identification by Descriptive and Discriminative Classification. In Anders Heyden and Fredrik Kahl, editors, Image Analysis. 2011. 1

[3] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2e: From Video Frames to Realistic DVS Events. 1

[4] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9000–9008, 2018. 1

[5] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2146–2156, 2021. 1

[6] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-identification by Video Ranking. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014. 2014. 1

[7] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 2

[8] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision – ECCV 2016, 2016. 1