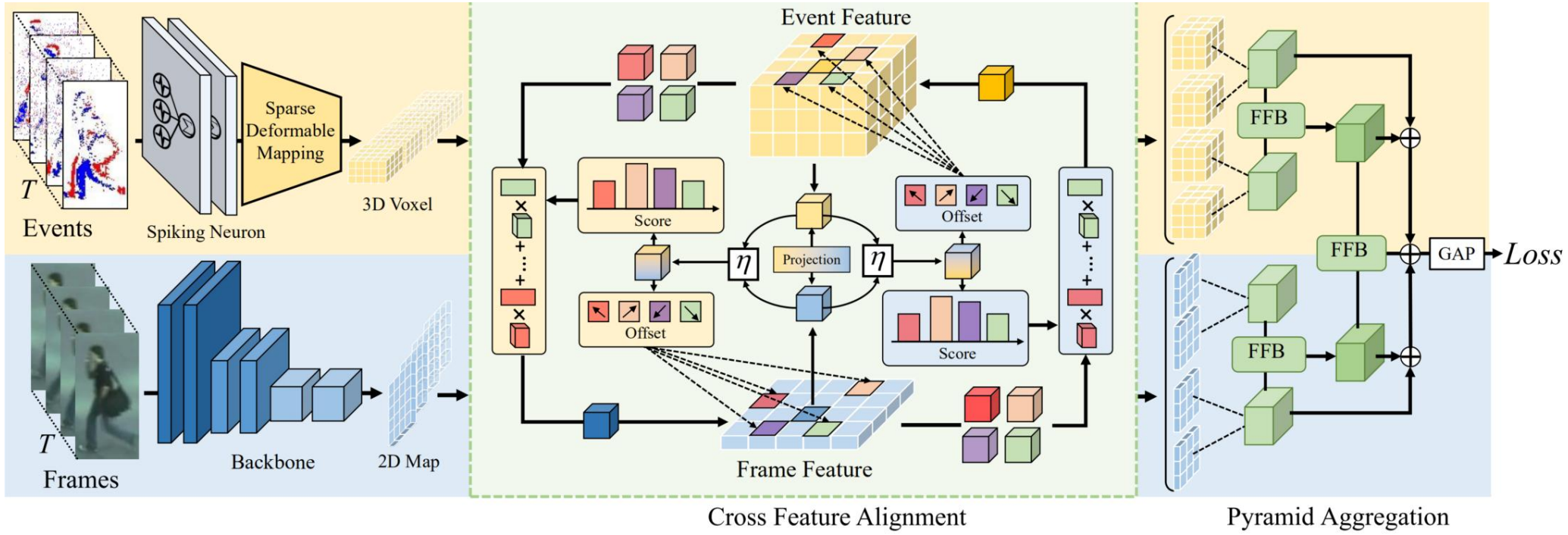
Event-guided Person Re-Identification via Sparse-Dense Complementary Learning

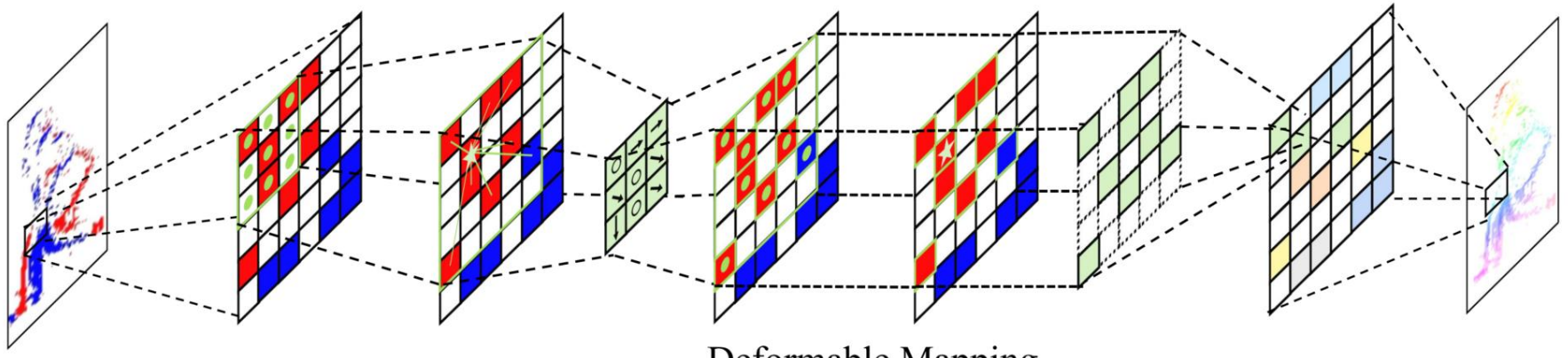
Chengzhi Cao, Xueyang Fu, Hongjian Liu, Yukun Huang, Kunyu Wang,
Jiebo Luo, Zheng-Jun Zha

University of Science and Technology of China, China
University of Rochester, USA

Poster ID: THU-AM-144

Framework Overview





Experiment

Methods		PRID-2011		iLIDS-VID		MARS	
Network	Input	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
GRL [35]	V	92.7	89.9	90.1	84.7	82.2	88.3
OSNet [59]	V	92.7	89.9	89.0	82.7	81.4	87.3
SRS-Net [45]	V	88.8	84.3	89.8	84.0	82.9	88.1
STMN [11]	V	92.8	88.8	84.1	77.3	81.8	88.3
CTL [33]	V	91.5	87.6	84.2	77.3	82.7	89.3
PSTA [49]	V	92.3	88.8	88.1	80.0	83.1	89.2
STGCN [53]	V	-	-	-	-	83.7	90.0
SINet [3]	V	-	96.5	-	92.5	86.2	91.0
RAFA [56]	V	-	95.9	-	88.6	85.9	88.8
MGH [52]	V	-	94.8	-	85.6	85.8	90.0
TCLNet [21]	V	-	-	-	86.6	85.1	89.8
STRF [2]	V	-	-	-	89.3	86.1	90.3
GRL [35]	E	21.4	11.2	30.2	18.0	27.7	16.7
OSNet [59]	E	22.2	10.1	27.9	16.7	30.9	19.3
SRS-Net [45]	E	17.2	9.0	32.7	19.3	20.9	10.0
STMN [11]	E	20.2	11.2	23.5	12.7	22.4	10.0
CTL [33]	E	20.4	13.5	28.4	18.0	25.6	12.7
PSTA [49]	E	22.2	12.4	22.4	10.0	22.7	12.0
GRL [35]	V+E	93.2	87.6	90.6	85.3	82.8	88.7
OSNet [59]	V+E	93.7	89.9	90.1	84.7	81.9	87.7
SRS-Net [45]	V+E	91.5	87.6	90.7	86.7	83.8	89.3
STMN [11]	V+E	94.0	91.0	87.2	81.3	83.4	89.0
CTL [33]	V+E	93.9	91.0	88.4	82.0	85.3	89.6
PSTA [49]	V+E	94.7	93.3	88.6	83.3	85.1	89.9
Ours	V+E	**96.9**	**96.5**	**93.2**	**92.7**	**86.5**	**91.1**

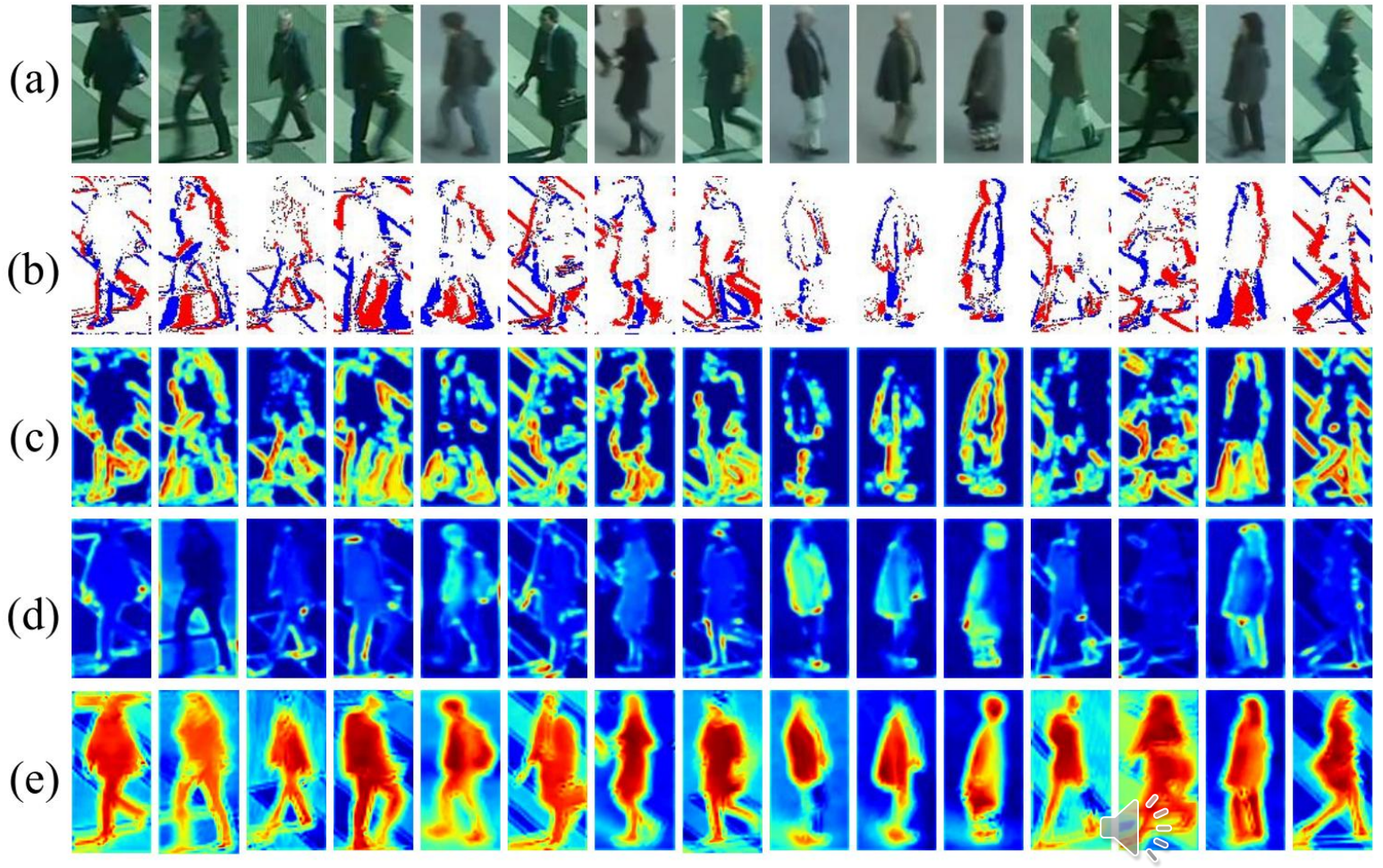
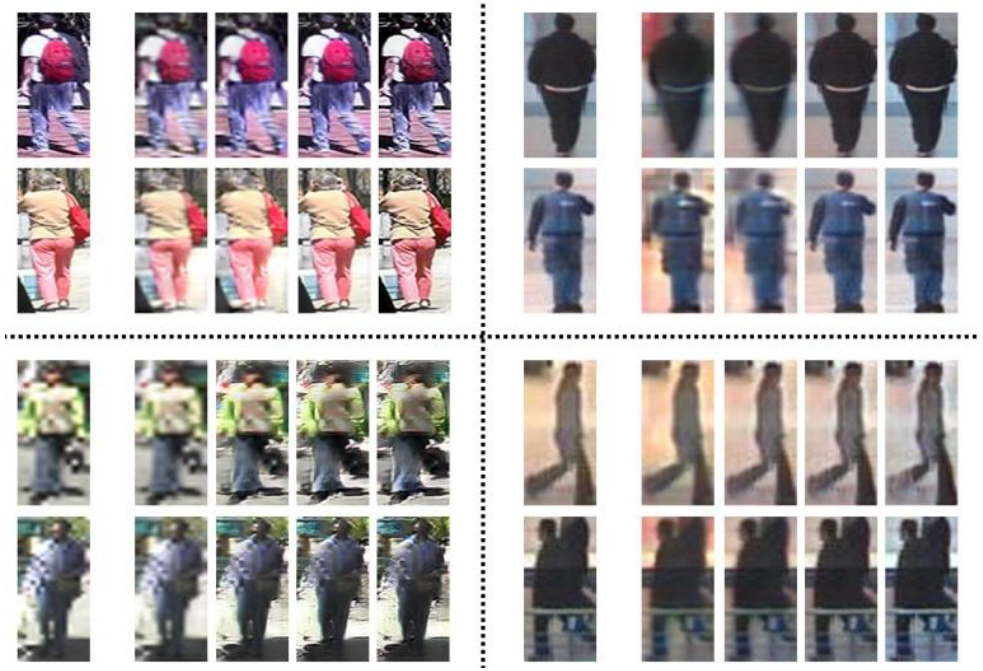
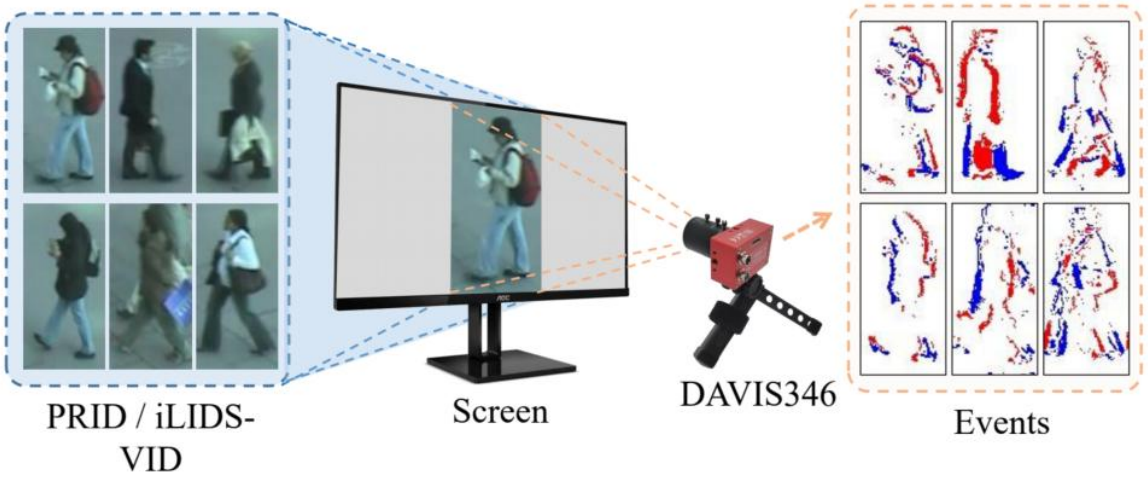


Figure 2. Visual examples of learned feature maps. From top to bottom: (a) original images, (b) corresponding events, (c) feature maps of events, (d) feature maps of PSTA [7] (w/o events), (e) feature maps of our network (w/ events).

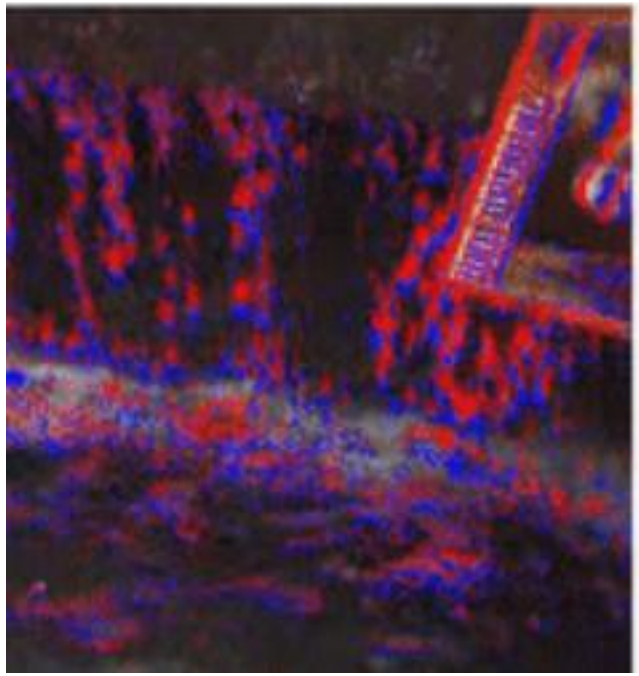
Introduction

Although video data can provide a wealth of appearance cues for identity representation learning, they also bring motion blur, illumination variations, and occlusions.





Introduction

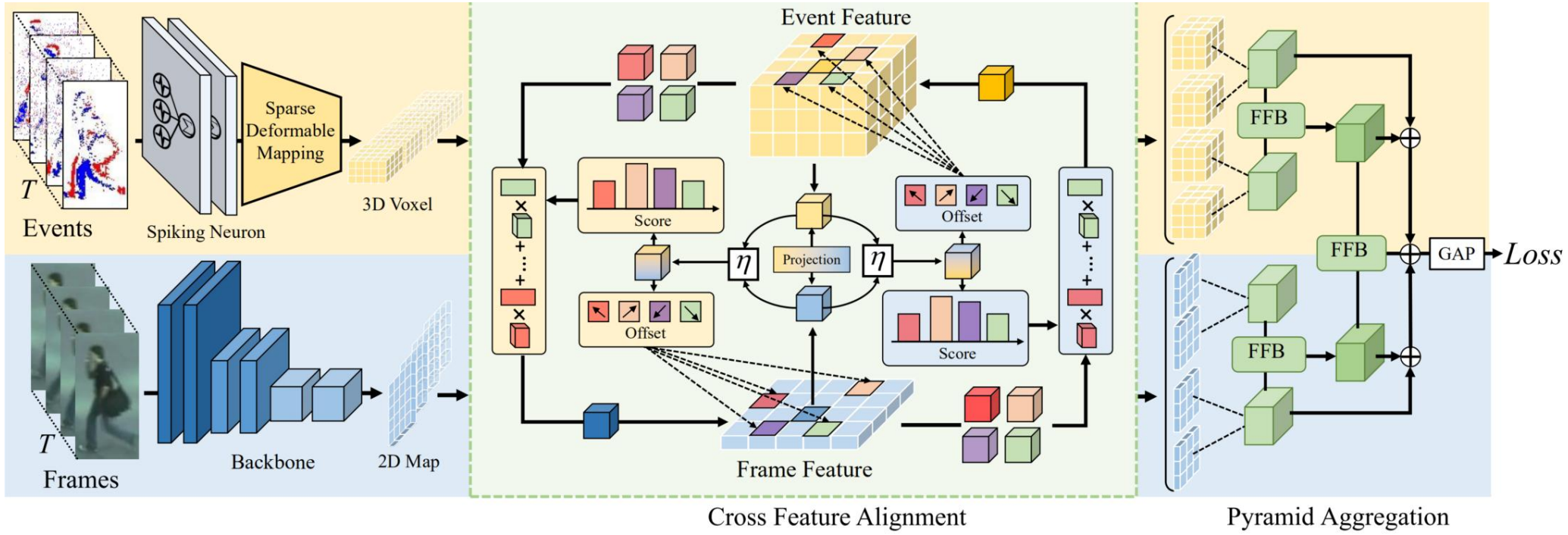


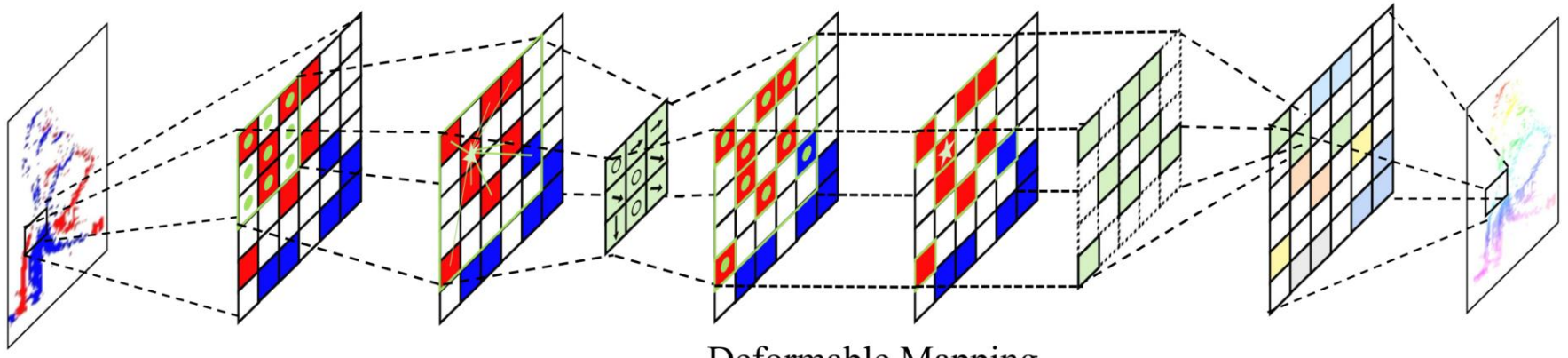
Event streams

- Low power
- Low latency
- High temporal resolution
- High dynamic range

Tulyakov, Stepan, et al. "Time lens: Event-based video frame interpolation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Framework Overview





Spike Neural Network

The polarity of event streams represents an increase or decrease of brightness at one pixel. Inspired by the dynamics and adaptability of biological neurons.

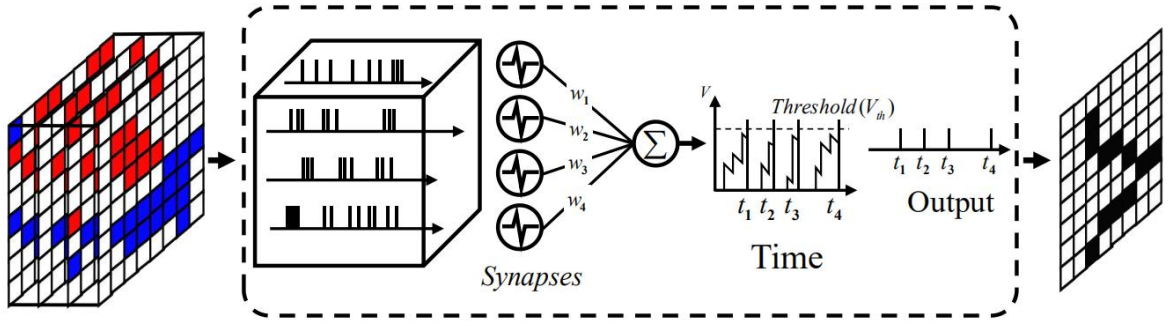


Figure 3. The structure of Leaky Integrate and Fire (LIF) Spike Neuron. The synaptic weight modulates the pre-spikes, which are then incorporated as a current influx in the membrane potential and decay exponentially. The post-neuron fires a post-spike and resets the membrane potential whenever the membrane potential reaches the firing threshold.

$$\tau_m \frac{dV(t)}{dt} = -V(t) + I(t),$$

$$I(t) = \sum_{i=1}^{n} (w_i \sum_{k} \psi_i(t - t_k)),$$

$$x_i^{t+1,n} = \sum_{j=1}^{n-1} w_{ij}^n o_j^{t+1,n-1},$$

$$u_i^{t+1,n} = u_i^{t,n} f(o_i^{t,n}) + x_i^{t+1,n} + b_i^n,$$

$$o_i^{t+1,n} = g(u_i^{t+1,n}),$$

Deformable Mapping

Most of recent works have shown that the number of spikes drastically vanishes at deeper layers, resulting in serious performance degradation. As shown in Figure 4. It clearly limits the application of SNN in computer vision. the deeper the SNN layer is, the more the number of spikes vanishes. But using deformable mapping can still preserve spatial information of events.

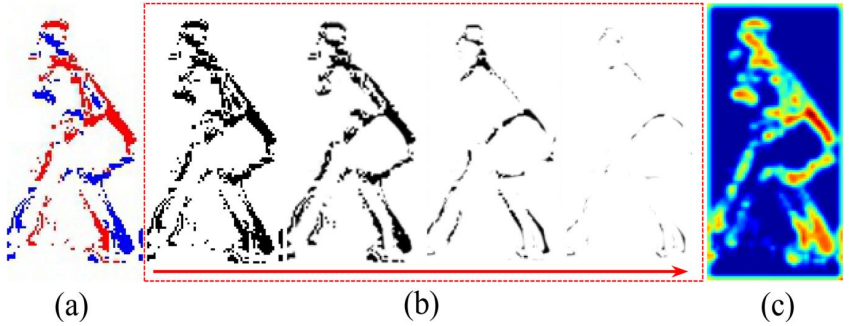
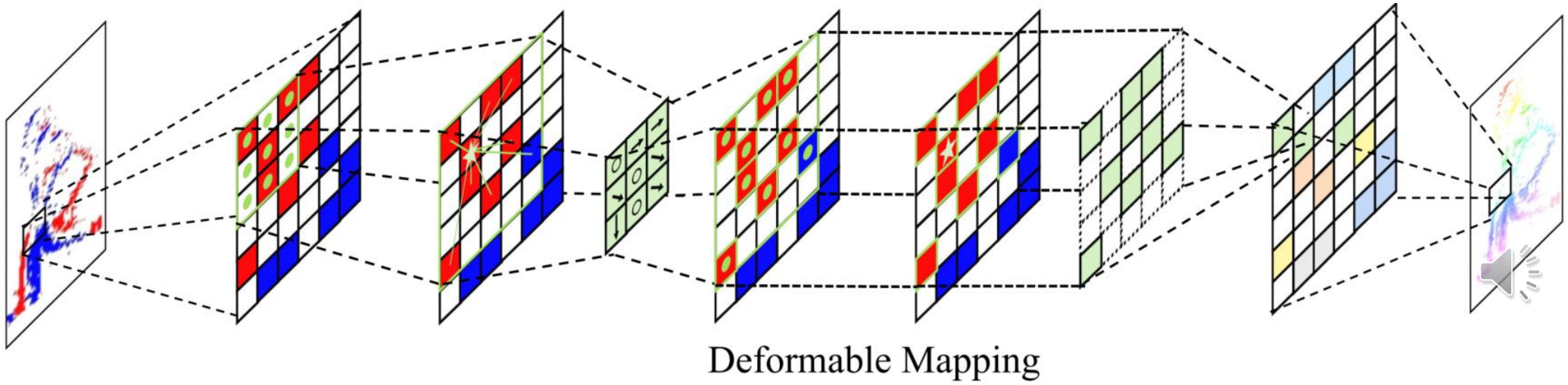


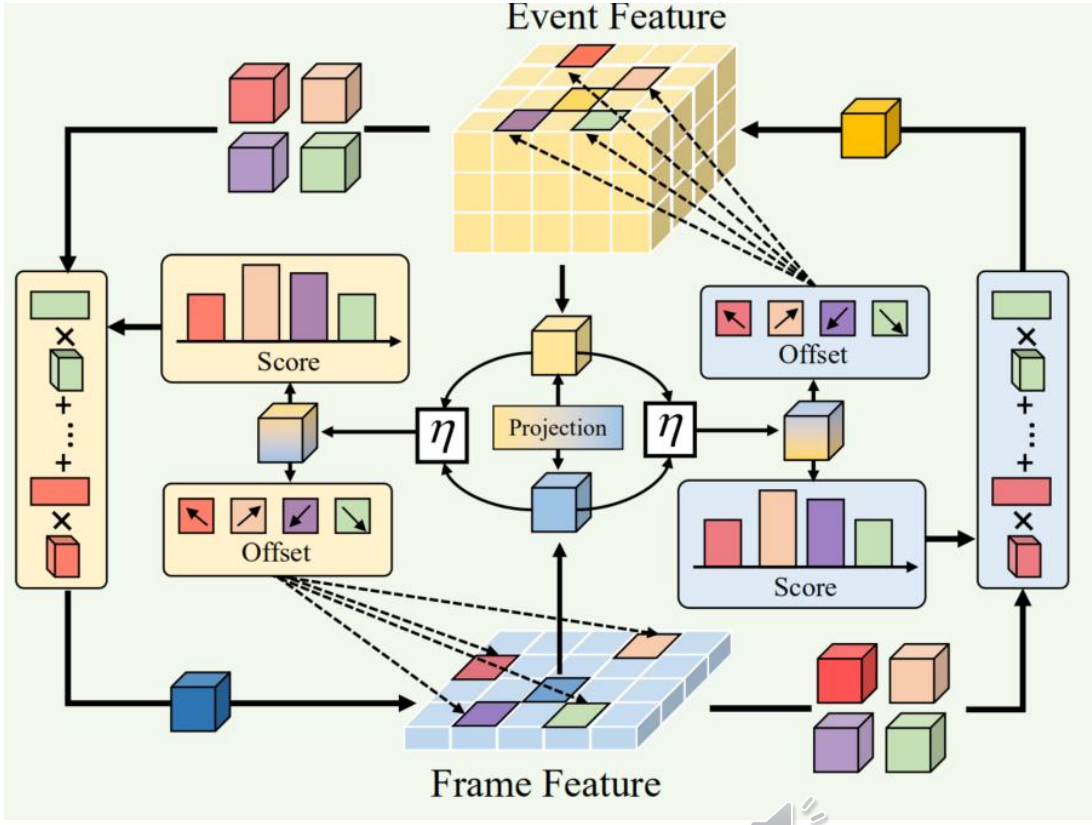
Figure 4. Visualization of features in SNN and deformable mapping. (a) presents events; (b) shows that the deeper the SNN layer is, the more the number of spikes vanishes. But using deformable mapping (c) can still preserve spatial information of events.



Cross Feature Alignment

This comple_x0002_mentary operation can not only conduct cross-domain re_x0002_lational modeling with the help of dynamically generated sampling offset but also maintain position consistency between events and frames to obtain the reference points.

$$\eta(Q_i, R_i) = \sum_{m=1}^{M} W_m [\sum_{i=1}^{I} MLP(Q_i) \cdot W'_m \mathbf{F}(R_i + \triangle R)],$$



Experiment

Methods		PRID-2011		iLIDS-VID		MARS	
Network	Input	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
GRL [35]	V	92.7	89.9	90.1	84.7	82.2	88.3
OSNet [59]	V	92.7	89.9	89.0	82.7	81.4	87.3
SRS-Net [45]	V	88.8	84.3	89.8	84.0	82.9	88.1
STMN [11]	V	92.8	88.8	84.1	77.3	81.8	88.3
CTL [33]	V	91.5	87.6	84.2	77.3	82.7	89.3
PSTA [49]	V	92.3	88.8	88.1	80.0	83.1	89.2
STGCN [53]	V	-	-	-	-	83.7	90.0
SINet [3]	V	-	96.5	-	92.5	86.2	91.0
RAFA [56]	V	-	95.9	-	88.6	85.9	88.8
MGH [52]	V	-	94.8	-	85.6	85.8	90.0
TCLNet [21]	V	-	-	-	86.6	85.1	89.8
STRF [2]	V	-	-	-	89.3	86.1	90.3
GRL [35]	E	21.4	11.2	30.2	18.0	27.7	16.7
OSNet [59]	E	22.2	10.1	27.9	16.7	30.9	19.3
SRS-Net [45]	E	17.2	9.0	32.7	19.3	20.9	10.0
STMN [11]	E	20.2	11.2	23.5	12.7	22.4	10.0
CTL [33]	E	20.4	13.5	28.4	18.0	25.6	12.7
PSTA [49]	E	22.2	12.4	22.4	10.0	22.7	12.0
GRL [35]	V+E	93.2	87.6	90.6	85.3	82.8	88.7
OSNet [59]	V+E	93.7	89.9	90.1	84.7	81.9	87.7
SRS-Net [45]	V+E	91.5	87.6	90.7	86.7	83.8	89.3
STMN [11]	V+E	94.0	91.0	87.2	81.3	83.4	89.0
CTL [33]	V+E	93.9	91.0	88.4	82.0	85.3	89.6
PSTA [49]	V+E	94.7	93.3	88.6	83.3	85.1	89.9
Ours	V+E	**96.9**	**96.5**	**93.2**	**92.7**	**86.5**	**91.1**

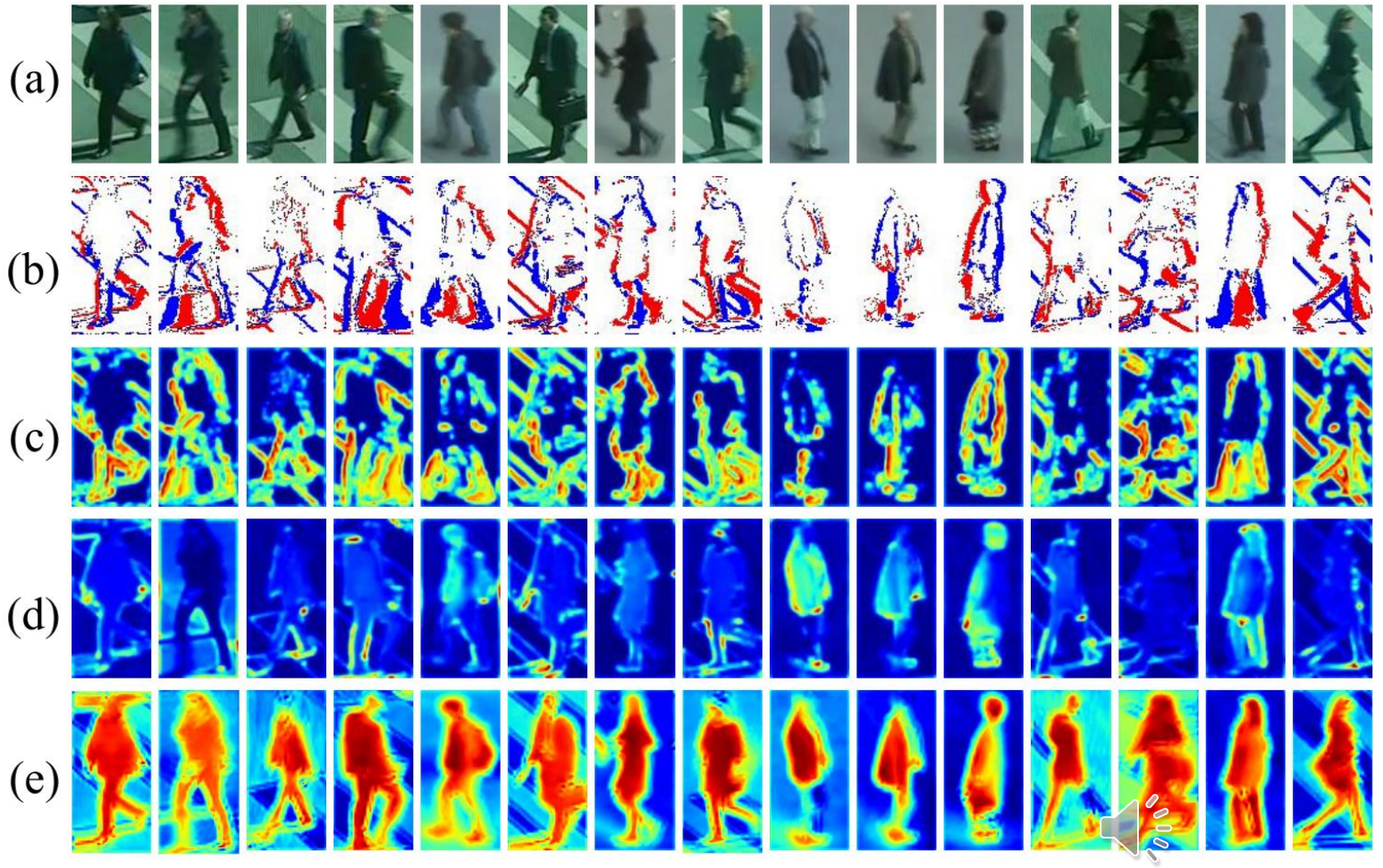


Figure 2. Visual examples of learned feature maps. From top to bottom: (a) original images, (b) corresponding events, (c) feature maps of events, (d) feature maps of PSTA [7] (w/o events), (e) feature maps of our network (w/ events).

Thank you !

Code: https://github.com/Chengzhi-Cao/SDCL