Recent advance in event-based vision : from deep learning perspective

Jianing Li Spiking Computing Group lijianing@pku.edu.cn

May 11, 2019

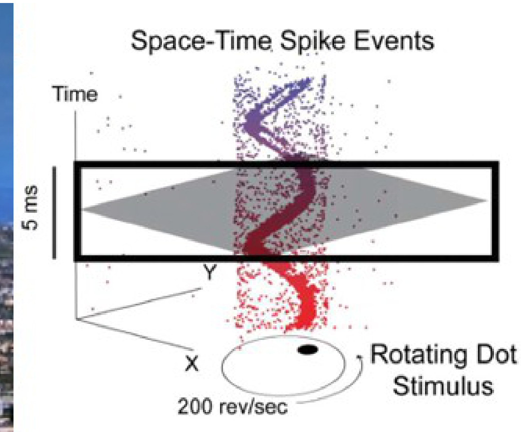
Overview

- **Introduction**
  - Event-based vision in CVPR 2019
  - Questions
- **Related works**
  - Time surface representations
  - Transformed images
- **End-to-end learning**
  - Events-to-video, CVPR 2019
  - Cv3dconv, IEEE Access 2019
  - EventNet, CVPR 2019
- **Discussion**
  - Better input representations for event data
  - Event-based vision in the future

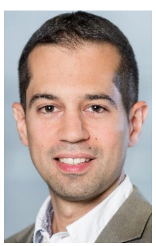
Overview

- **Introduction**
  - Event-based vision in CVPR 2019
  - Questions
- **Related works**
  - Time surface representations
  - Transformed images
- **End-to-end learning**
  - Events-to-video, CVPR 2019
  - Cv3dconv, IEEE Access 2019
  - EventNet, CVPR 2019
- **Discussion**
  - Better input representations for event data
  - Event-based vision in the future

Event-based vision workshops in CVPR 2019



- **Organizers:**



Davide Scaramuzza
UZH

Guillermo Gallego
UZH

Kostas Daniilidis
UPenn

Second International Workshop on Event-based vision and Smart Cameras – June 17, Long Beach

Event-based vision workshops in CVPR 2019

- **Call for papers and demos**
  - Event-based / neuromorphic vision.
  - Algorithm: Visual odometry, SLAM, 3D reconstruction, Optical flow estimation, image intensity reconstruction, recognition, stereo depth reconstruction, feature/ object detection and tracking, calibration, sensor fusion.
  - Model based, embedded or learning approaches.
  - Event-based signal processing, control, bandwidth control.
  - Event-based active vision.
  - Event-based camera datasets and/or simulators.
  - Applications in: robotics(navigation, manipulation, drones…), automotive, IoT, AR/VR, space, inspection, surveillance, crowd counting, physics.
  - Biologically-inspired vision and smart cameras
  - Novel hardware(cameras, neuromorphic processors, etc.) and/or software platforms.
  - New trends and challenges in event-based and/or biologically-inspired vision.

Event-based vision workshops in CVPR 2019

Invited speakers

Tobi Delbruck
ETH

Garrick Orchard
NUS

Jorg Conradt
KTH

Giacomo Indiveri
ETH

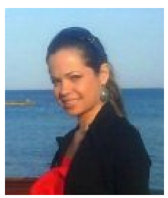
Piotr Dudek
Univ. Manchester

Andrew Davision
ICL

Cornelia Fermuller
Univ.Maryland

Yulia Sandamirskaya
ETH

Chiara Bartolozzi
Italiano di Tecnlogia



Margarita Chli
ETH

Robert Mahony
ANU

Invited companies

PROPHESEE
ATIS, France

SAMSUNG
DVS(640*480), SK

intel
Loihi, USA

Insightness,
DVS, Switzerland

inivation
DVS, Switzerland

cele pixel
DVS, China

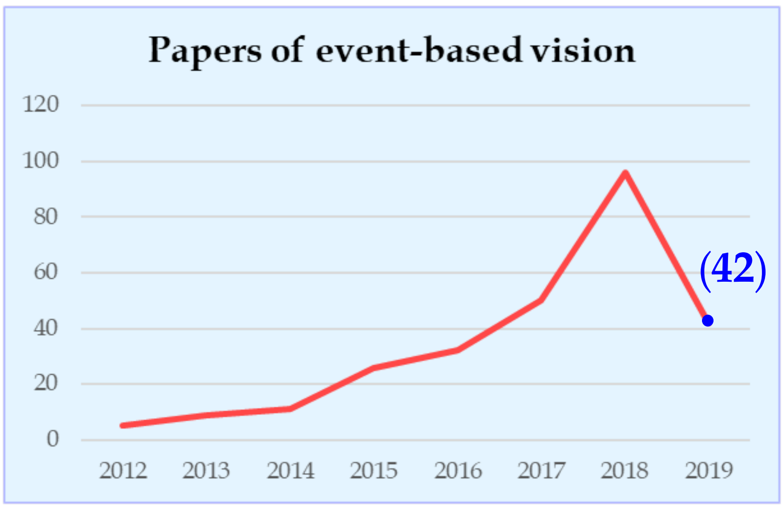
Event-based vision in CVPR 2019

- **Paper list (1+8)**
  - Bring a blurry frame alive at high frame-rate with an event camera, Liyuan Pan et. al, ***ANU*. (oral)**
  - Unsupervised event-based learning of optical flow, depth and ego-motion, Alex Z. Zhu et al, ***University of Penn***.
  - Events-to-video: bringing modern computer vision to event cameras, Henri Rebecq et al, ***UZH & ETH***.
  - EventNet: Asynchronous recursive event processing, Yusuke Sekikawa et al, ***Denso IT Laboratory***.
  - EV-Gait: Event-based robust gait recognition using dynamic vision sensors, Yanxiang Wang et al, ***HEU, China***.
  - Event-based high dynamic range image and very high frame rate video generation using conditional generation adversarial networks, S. M. Mostafavi et al, ***GIST***.
  - Speed invariant time surface for learning to detect corner points with event-based cameras, J. Manderscheid et al, ***PROPHESEE***.
  - Focus loss functions for event-based vision, Guilleromo Gallego et at, ***UZH & ETH***.
  - Event cameras, contrast maximization and reward functions: an analysis, T. N. Stoffregen et al, ***Monash University***.

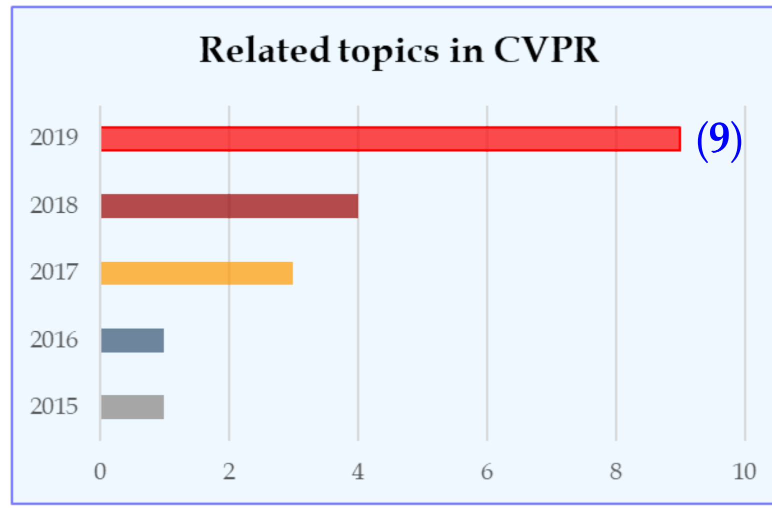
Papers in CVPR

Event-based vision

- Research papers in recent years
- Related topics in CVPR



Tab.1 Papers of event-based vision in recent years



Tab.2 Related topics in CVPR

Tips: all statistical papers mainly are about DVS, ATIS, DAVIS and CeleX.

We ask the questions

- **1** What are **new trends** and **challenges** in event-based vision?
- **2** How will **spatial-temporal data** meet **deep learning**?
- **3** Do you believe that **System theory** exists in **event-based vision**?

Overview

- Introduction
  - Event-based vision in CVPR 2019
  - Questions
- **Related works**
  - Time surface representations
  - Transformed images
- End-to-end learning
  - Events-to-video, CVPR 2019
  - Cv3dconv, IEEE Access 2019
  - EventNet, CVPR 2019
- Discussion
  - Better input representations for event data
  - Event-based vision in the future

Time surface representations

Time surface [1]

- Event streams

$$ev_i = [\mathbf{x_i}, t_i, p_i]^T, \quad i \in \mathbb{N}$$

- Time context

$$\mathcal{T}_i(\mathbf{u}, p) = \max_{j \leq i}\{t_j \,|\, \mathbf{x_j} = (\mathbf{x_i} + \mathbf{u}),\, p_j = p\}$$

- Computing time surface

$$\mathcal{S}_i(\mathbf{u}, p) = e^{-(t_i - \mathcal{T}_i(\mathbf{u}, p))/\tau}.$$

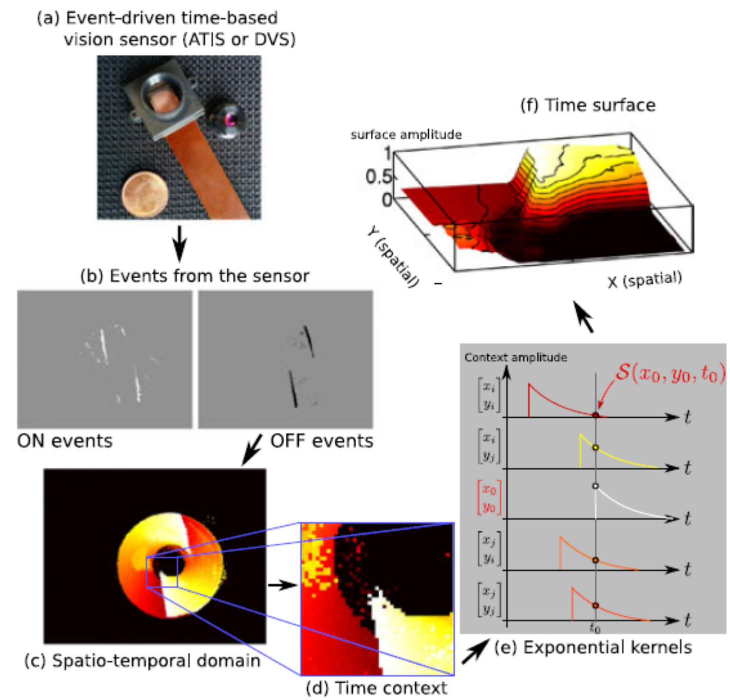


Fig.1 Time surface from the spatiotemporal events

[1] HOTS: A hierarchy of event-based time-surfaces for pattern recognition. Xavier Lagorce et.al . *PAMI* 2017.

Time surface representations

- **Local memory time surfaces** [2]
  - Time window

$$\mathcal{T}_{e_i}(\mathbf{z}, q) = \begin{cases} \sum_{e_j \in \mathcal{N}_{(\mathbf{z},q)}(e_i)} e^{-\frac{t_i - t_j}{\tau}} & \text{if } p_i = q \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathcal{N}_{(\mathbf{z},q)}(e_i) = \{e_j : \mathbf{x}_j = \mathbf{x}_i + \mathbf{z}, t_j \in [t_i - \Delta t, t_i), p_j = q\}$$

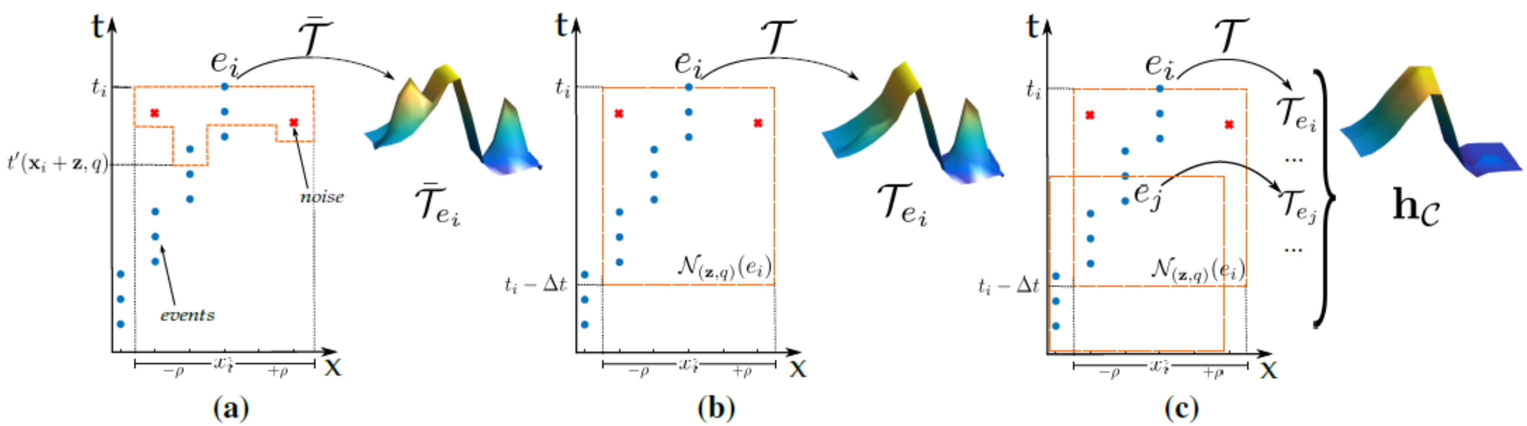


Fig.2 Time surface computation around an event, in presence of noise. (a)time surfaces; (b)local memory time surfaces; (c)HATS

[2] HATS: Histograms of averaged time surfaces for robust event-based object classification. Amos Sironi et.al . *CVPR* 2018.

Transformed images

Rate-based images [3]

- Integrating time window

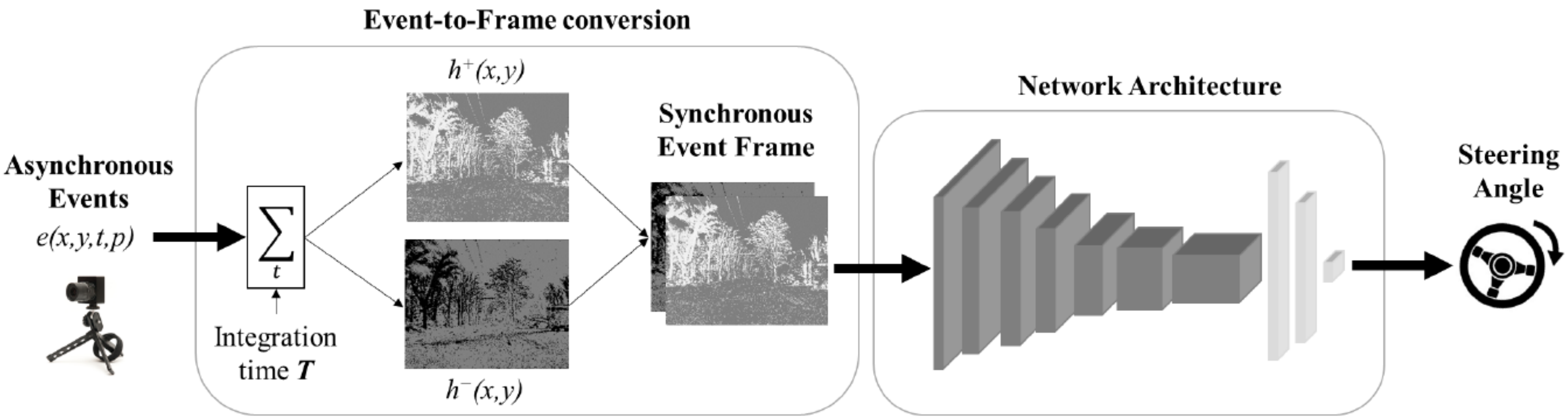
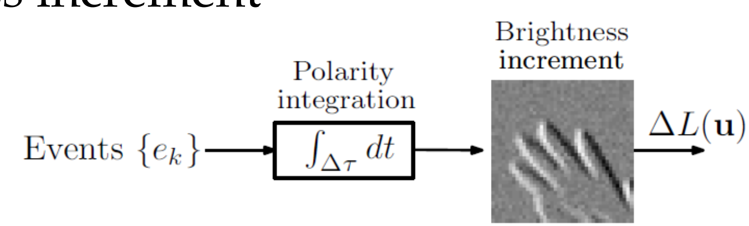


Fig.3 Event-to-frame conversion by rate-based strategy

Feature images [4]

- Brightness increment



[3] Event-based vision meets deep learning on steering prediction for self-driving cars. Ana I. Maqueda et.al . *CVPR* 2018.
[4] Asynchronous, photometric feature tracking using events and frames, Daniel Gehrig et al, ECCV 2018.

Overview

- Introduction
  - Event-based vision in CVPR 2019
  - Questions
- Related works
  - Time surface representations
  - Transformed images
- **End-to-end learning**
  - Events-to-video, CVPR 2019
  - Cv3dconv, IEEE Access 2019
  - EventNet, CVPR 2019
- Discussion
  - Better input representations for event data
  - Event-based vision in the future

Events-to-Video: bringing modern computer vision to event cameras

Henri, Rebecq, Rene Ranftl, Vladlen Koltun, **Davide Scaramuzza ***

CVPR, 2019

1 Introduction

- **Motivation**
  - Challenging illumination conditions
  - Fast motion

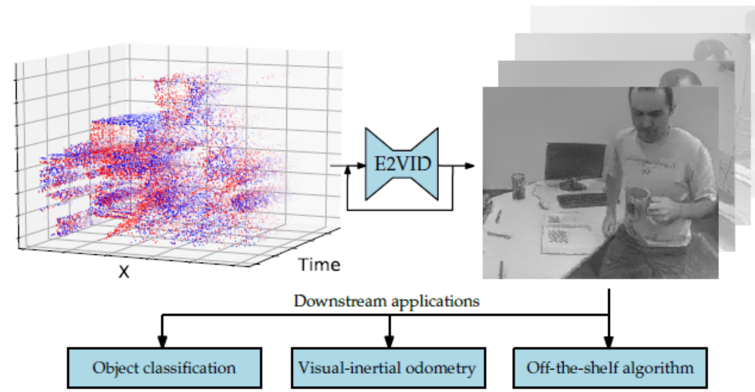


Fig.4 Converting spatial-temporal steam into high-quality video.

- **Contributions**
  - **Recurrent network architecture to reconstruct a video from spatial-temporal events**
  - Quantized assessment by transformed application
  - Providing a simulated and real events dataset

2 Problem statement

- **Events**
  - Asynchronous spatial-temporal point

$$e_i = < x_i, y_i, t_i, p_i >$$

- **Event representation**
  - Spatial-temporal voxel grid[5]

$$E(x_l, y_m, t_n) = \sum_{\substack{x_i = x_l \\ y_i = y_m}} p_i \max(0, 1 - |t_n - t_i^*|)$$

$$t_i^* \triangleq \frac{B-1}{\Delta T}(t_i - t_0)$$

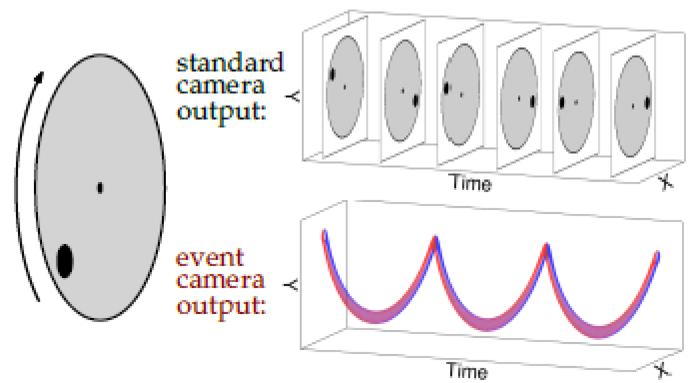


Fig.5 Comparison of conventional camera and event camera.

[5] Unsupervised event-based learning of optical flow, depth and ego-motion. Alex Zhihao Zhu et.al . *CVPR* 2019.

3 Method

Architecture

- Recurrent network

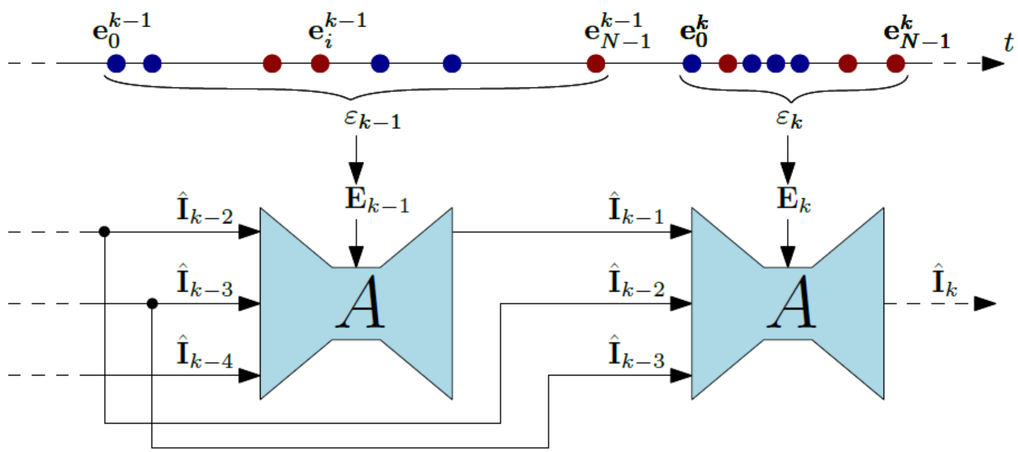


Fig.6 Each window is converted into E_K and passed through the network together with the last K reconstructed images to a new image $\hat{I}_k$.

Training strategy

- Dataset– event simulator ESIM [6]
- Loss functions

$$\mathcal{L}_K = \sum_{L=0}^{L} d_L(\hat{I}_{k-l}, I_{k-l})$$

[6] ESIM: an open event camera simulator. Henri Rebecq et.al . *CoRL* 2018.

3 Results

Representative results

- Event-camera dataset and simulator [7]

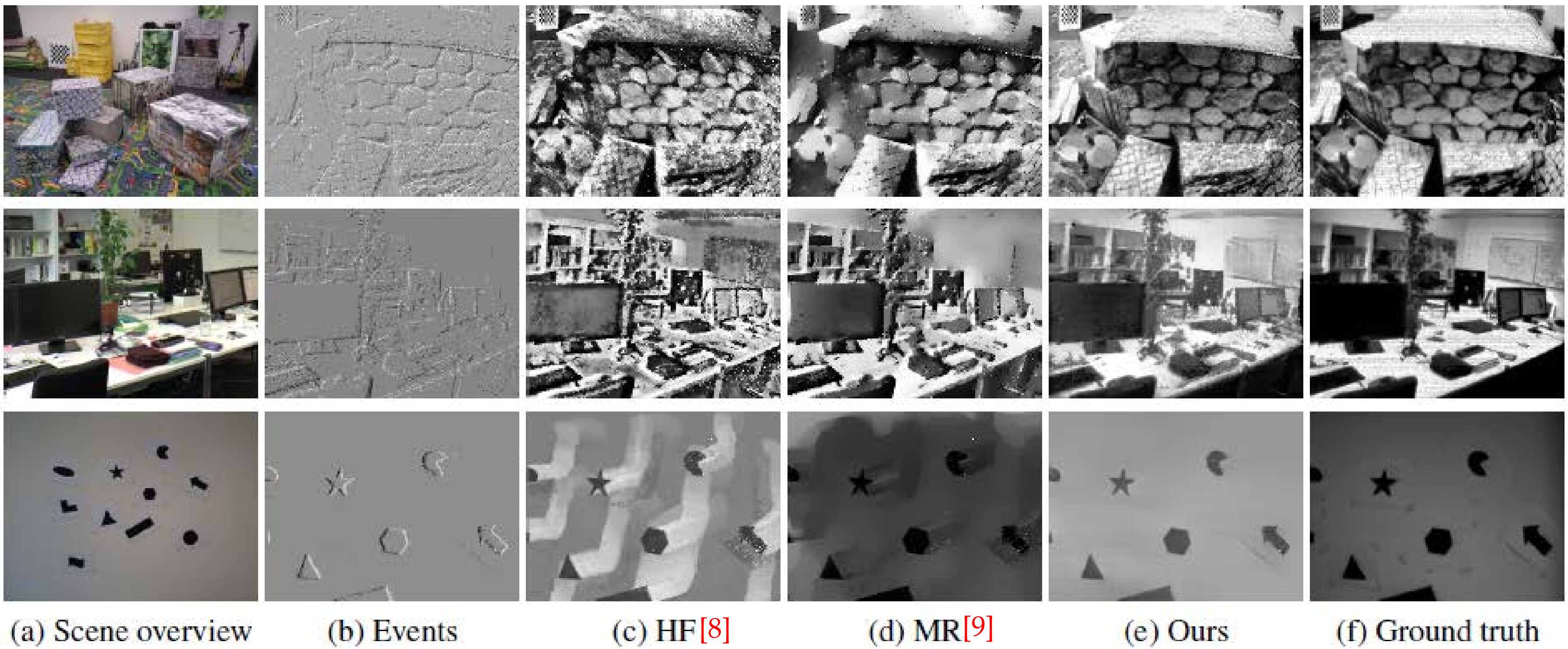


Fig.7 representative image reconstruction for event cameras.

[7] The event-camera dataset and simulator: event-based data for pose estimation, visual odometry, and SLAM. Elias Mueggler et al, IJRR, 2017.
[8] Continuous-time intensity estimation using event cameras, Cedric Scheerlinck et al, ACCV 2018.
[9] Real-time intensity-image reconstruction for event cameras using manifold regulation. Gottfried Munda et.al . ***IJCV*** 2018.

3 Results

Performance estimation

- MSE, SSIM and LPIPS

Dataset	MSE			SSIM			LPIPS		
	HF	MR	Ours	HF	MR	Ours	HF	MR	Ours
dynamic_6dof	0.10	0.11	**0.08**	0.39	0.44	**0.50**	0.53	0.53	**0.43**
boxes_6dof	0.09	0.07	**0.04**	0.45	0.47	**0.63**	0.51	0.54	**0.36**
poster_6dof	0.06	0.05	**0.04**	0.52	0.55	**0.68**	0.44	0.50	**0.32**
shapes_6dof	0.11	0.14	**0.10**	0.34	0.43	**0.44**	0.63	0.64	**0.53**
office_zigzag	0.09	0.06	**0.05**	0.36	0.43	**0.50**	0.54	0.55	**0.44**
slider_depth	0.08	0.08	**0.06**	0.48	0.51	**0.61**	0.50	0.55	**0.42**
calibration	0.07	0.06	**0.04**	0.41	0.41	**0.52**	0.55	0.57	**0.47**
Mean	0.09	0.08	**0.06**	0.42	0.46	**0.56**	0.53	0.55	**0.42**

Tab.3 Comparison to state-of-the-art image reconstruction methods on the Event Camera Dataset [7] .

[7] The event-camera dataset and simulator: event-based data for pose estimation, visual odometry, and SLAM. Elias Mueggler et al, IJRR, 2017.

4 Outlook

- **1** Will **CeleX overthrow image reconstruction for event cameras**?
- **2** How to further **exploit spatial-temporal information** for event data?
- **3 Is image reconstruction better for vision tasks**?

Constant velocity 3D convolution

Yusuke Sekikawa, Kohta Ishikawa, and **Hideo Saito** *

IEEE Access, 2019

1 Introduction

- **Motivation**
  - Exploiting spatial-temporal information
  - Taking advantages of event cameras
  - Fast computing

- **Contributions**
  - **cv3dconv, where a 3D kernel is represented as a 2D spatial kernel and constant velocity kernel**
  - Recursive convolution to fast computing
  - Significant improvement over the state-of-the-art architectures

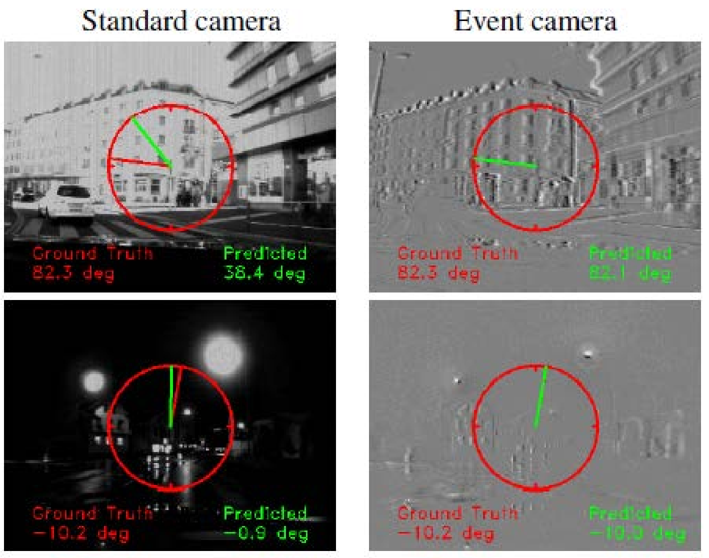


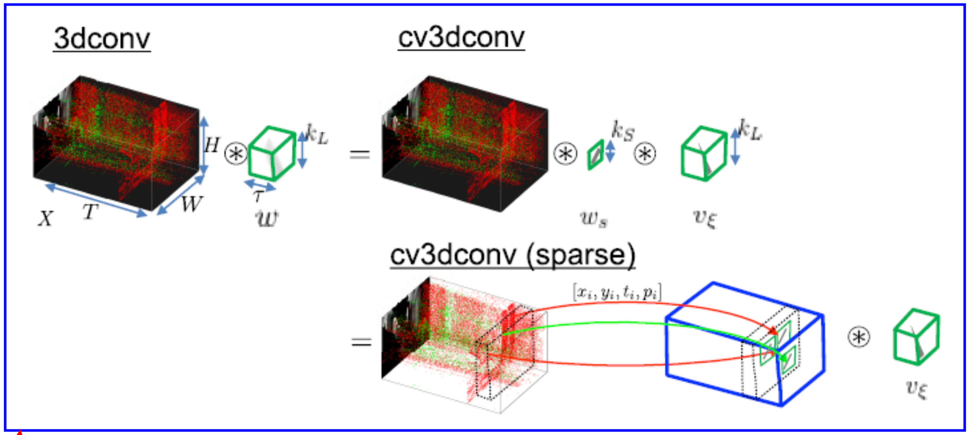
Fig.8 Steering angle performance on frames and event camera [10] .

[10] Event-based vision meets deep learning on steering prediction for self-driving cars. Ana I. Maqueda et.al . *CVPR* 2018.

2 Method

- **Architecture**
  - **cv3dconv**

$$z = X \circledast \omega_\xi = X \circledast \omega_s \circledast v_\xi$$



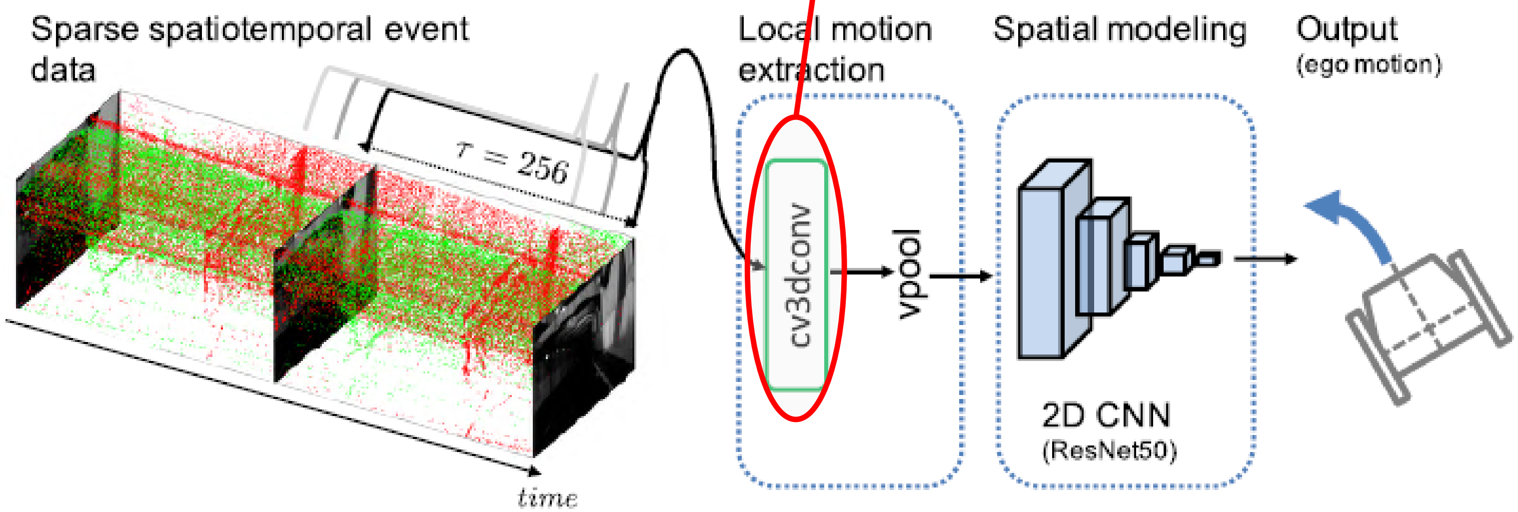


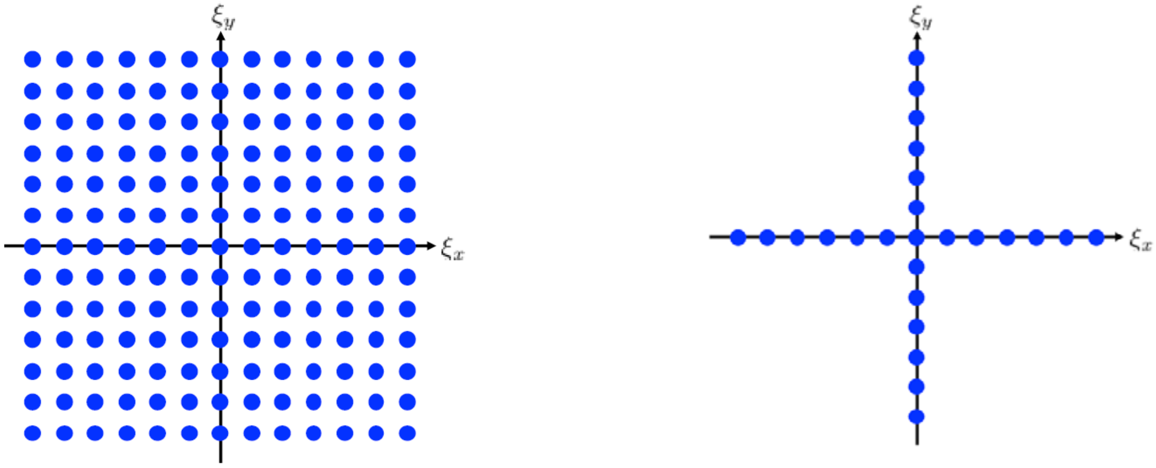
Fig.9 Overview of DNN architecture. The cv3dconv captures local spatial-temporal features from sparse spatiotemporal inputs, and then the subsequent 2D-CNN layers (ResNet 50) model the global spatial correlation of the extracted features.

2 Method

- **Constant velocity 3D convolution**
  - **Constant velocity approximation**

$$X \circledast v_{\zeta} = \sum_{i=0}^{\tau-1} X\left(x - i\xi_x, y - i\xi_y, t - i\right)$$

  - **Sampling strategy**



(a) (b)

Fig.10 Two kinds of velocity sampling strategy considered. (a) uniformly sampling; (b) sampling only along $\xi_x - \xi_y$ axes.

2 Method

- **Computing strategy**
  - **Recursive convolution with V_ξ**

$$z(x,y,t+1) = z\left(x-\xi_x, y-\xi_y, t\right) + \epsilon_{new} - \epsilon_{old}$$

$$\epsilon_{new} = X(x,y,t+1) \circledast w_s \circledast v_\xi(\cdot,\cdot,1)$$

$$\epsilon_{old} = X(x,y,t-\tau) \circledast w_s \circledast v_\xi(\cdot,\cdot,\xi)$$

  - **Fourier Convolution with V_ξ**

$$z(x,y,t) = [\mathcal{FT}^{-1}_{(1,2)}\hat{X}\left(\hat{x},\hat{y},\hat{x}\xi_x + \hat{y}\xi_y\right)] \circledast w_s$$

$$\hat{X} = \mathcal{FT}_{(3)}\mathcal{FT}_{(1,2)}X(\cdot,\cdot,t-\tau+1\colon t)$$

  - **Sparse event-wise convolution**

$$X(\cdot,\cdot,t) \circledast w_s = \sum_i S(x_i,w_s)$$

2 Method

Detailed framework

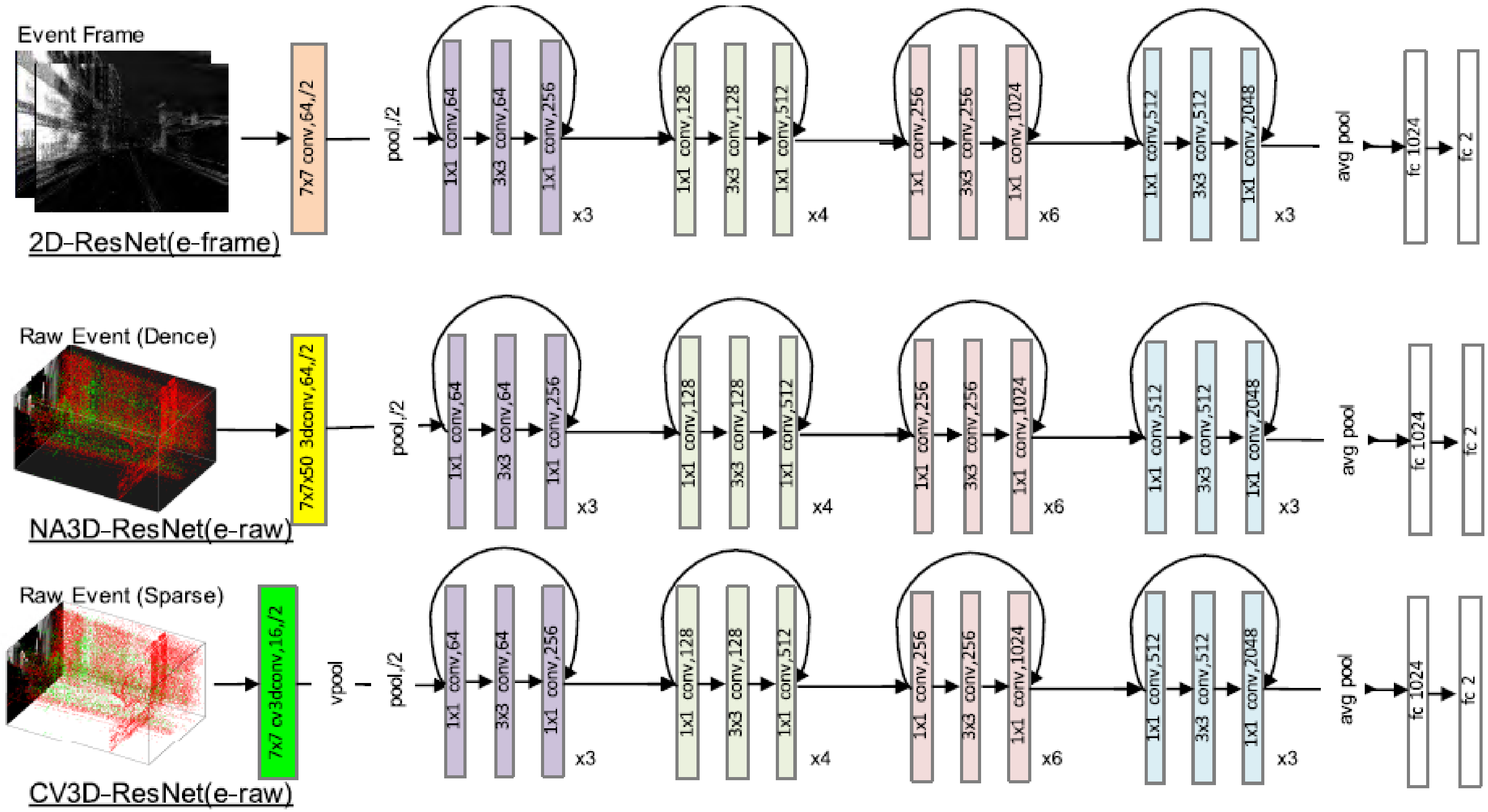


Fig.11 The detailed DNN frameworks.

3 Results

Experimental settings

Architecture Name	First two layers	Filters size of first layer	ν	Input
2D-ResNet(1)	2dconv,relu	$7 \times 7 \times 1 \times 2 \times 64$	-	Histogram
2D-ResNet(2)	2dconv,relu	$7 \times 7 \times 1 \times 4 \times 64$	-	Histogram+Timestamps
NA3DResNet	3dconv,relu	$7 \times 7 \times 256 \times 1 \times 64$	-	Dense Raw Event
CV3D-ResNet(1)	cv3dconv,vpool	$7 \times 7 \times 1 \times 1 \times 16$	13^2	Sparse Raw Event
CV3D-ResNet(2)	cv3dconv,vpool	$7 \times 7 \times 1 \times 1 \times 64$	13×2	Sparse Raw Event

Tab.4 Summary of each architectures.

Performance evaluation

Architecture Name	*Outdoor Night 3*				*Outdoor Night 3* + noise			
	EVA		RMSE		EVA		RMSE	
	$\Delta\theta$	ΔL	$\Delta\theta$	ΔL	$\Delta\theta$	ΔL	$\Delta\theta$	ΔL
2D-ResNet(1)	0.709	0.801	3.503	1.255	-0.032	0.024	8.542	5.595
2D-ResNet(2)	0.750	0.841	3.212	1.201	-0.445	0.010	14.473	4.415
NA3D-ResNet	0.952	0.944	2.113	1.173	0.434	0.423	5.983	2.203
CV3D-ResNet(1)	**0.955**	**0.948**	**1.553**	**0.860**	**0.661**	**0.542**	**3.934**	**2.013**
CV3D-ResNet(2)	0.950	0.939	2.198	1.216	0.489	0.477	4.992	2.202

Tab.5 Performance evaluation based on DDD17 dataset [11].

[11] DDD17: End-to-end DAVIS driving dataset, Jonathan Binas et.al. ***ICML workshops***, 2017.

3 Results

Computational complexity

Computing efficiency

		Number of sum-of-product operations	Ratio	Time [s] CPU	Time [s] GPU
3dconv		$TWH\nu(k_L^2\tau)$	1	1715.6	170.7
cv3dconv	Fourier-dense	$TWH(\ k_S^2 + \tau\log\tau + (\nu+1)\log(WH))$	17×10^3	185.7	11.7
cv3dconv	Fourier-sparse	$TWH(\ \alpha k_S^2 + \tau\log\tau + (\nu+1)\log(WH)\ )$	19×10^3	183.2	–
cv3dconv	sequential-dense	$TWH(k_S^2 + 4\nu)$	48×10^3	103.4	**5.54**
cv3dconv	sequential-sparse	$TWH(\alpha k_S^2 + 4\nu)$	$\mathbf{69 \times 10^3}$	**99.3**	–

Tab.6 Comparison with computational efficiency.

Parameters memory

	Number of parameters			Error	
	conv	fc	Total	Angle [deg]	Velocity [pix/τ]
3dconv	$101^2 \times 32 \times 32$	32×3	10.45M	6.91	1.93
cv3dconv	$31^2 \times 1 \times 32$	$(13^2 \times 32) \times 3$	0.048M	**3.78**	1.03
cv3dconv + vpool	$31^2 \times 1 \times 32$	$(13^2 + 32) \times 3$	**0.031M**	4.09	**0.76**

Tab.7 Comparison with parameters memory.

4 Outlook

- **1** How to further **exploit spatial-temporal information** for event data?

- **2** Could you design an **EventNet**, instead of 3D convolution strategy?

EventNet: asynchronous recursive event processing

Yusuke Sekikawa, Kohta Ishikawa, and **Hideo Saito ***

CVPR, 2019

1 Introduction

- **Motivation**
  - Further exploit spatial-temporal data
  - End-to-end learning for event streams

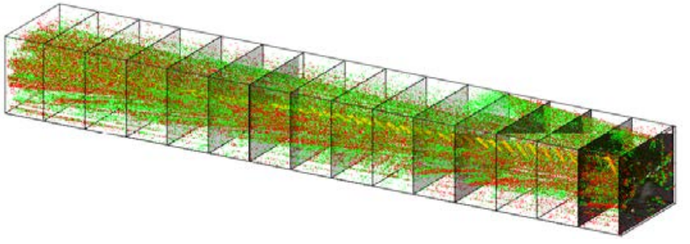


Fig.12 Snapshot from the MVSEC [12] .

- **Contributions**
  - **Recursive architecture using a novel temporal coding and aggregation scheme**
  - A lookup table (**LUP**) instead of multi-layer-perceptron (**MLP**) removing most of the sum-of-product operations
  - End-to-end learning by event-wise processing

[12] The multivehicle stereo event camera dataset: an event camera dataset for 3d perception. Alex Z. Zhu et.al. ***IEEE Robotics and Automation Letters,*** 2018.

2 Method

□ Architecture

■ PointNet [13]

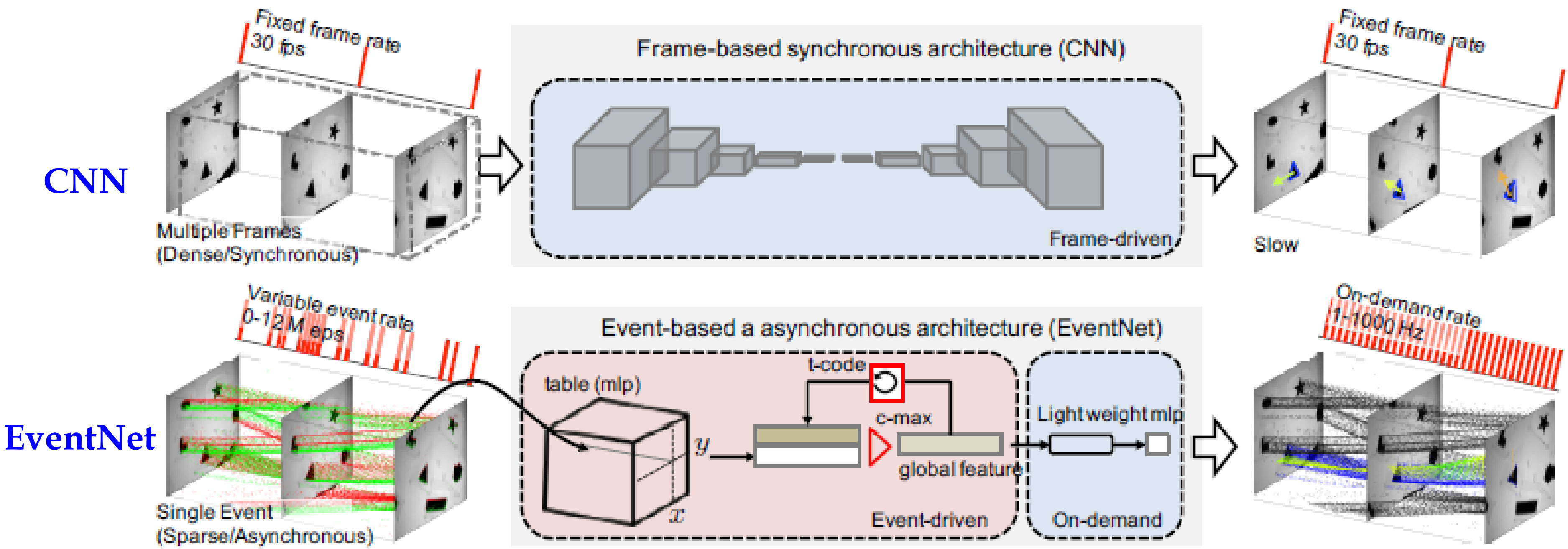


Fig.13 Overview of asynchronous event-based pipeline of EventNet in contrast to conventional frame-based CNN.

[13] PointNet: deep learning on point sets for 3D classification and segmentation Charles R. Qi et.al. ***CVPR,*** 2017.

2 Method

- **Framework**
  - **End-to-end learning**

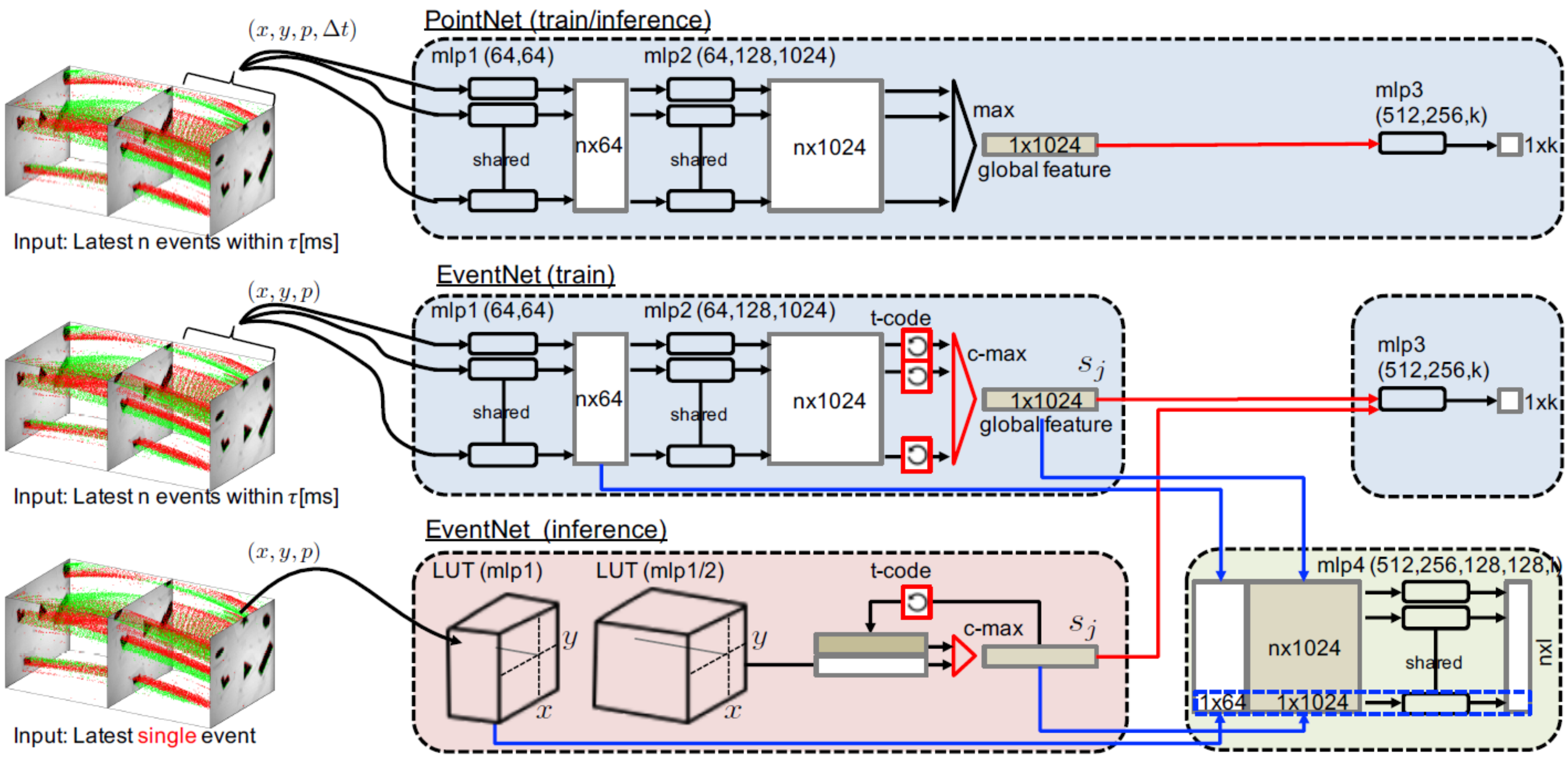


Fig.14 The framework of EventNet is shown in comparison with PointNet.

2 Method

- **Strategies**
  - **Symmetric function**

$$y_i = f(e_j) \approx g(\max(h(e_{j-n(j)+1}), \dots, h(e_j)))$$

Where $h: \mathbb{R}^4 \to \mathbb{R}^k$, max: $\underbrace{\mathbb{R}^K \times \cdots \times \mathbb{R}^k}_{n(j)} \to \mathbb{R}^K$, and g: $\mathbb{R}^K \to \mathbb{R}$. They approximate h and g using an **MLP**.

  - **Temporal coding**

$$\boldsymbol{h(e_i) \approx c(\mathrm{h}(e_i^-), \Delta t_{j,i}), \text{ where } e^- := (x, y, p)}$$

$$f(e_j) \approx g(\max(c(z_{j-n(j)+1}, \Delta t_{j,j-n(j)+1}), \dots, c(z_j, 0)))$$

Where $z_i = h(e_i^-) \in \mathbb{C}^K$. Using this formulation, we need to compute h only once for each observed event, however, c and max need to be computed for all events in time window every time a new event arrives.

2 Method

- **Strategies**
  - **Recursive processing**

$$a_{j,i} = c(z_i, \Delta t_{j,i}) = \left[|z_i| - \frac{\Delta t_{j,i}}{\tau}\right]^+ exp(-i\frac{2\pi\Delta t_{j,i}}{\tau})$$

$$\max\left(c(z_{j-n(j)+1}, \Delta t_{j,j-n(j)+1}), \ldots, c(z_j, 0)\right) = \max(c(s_j, \delta_{t_j}, h(e_{j+1}^-)))$$

$$f(e_{j+1}) \approx g(\max(c(s_j, \delta_{t_j}), \ldots, h(e_{j+1}^-))$$

$$s_j = max(c(s_{(j-1)}, \delta t_{j-1}), h(e_{j+1}^-))$$

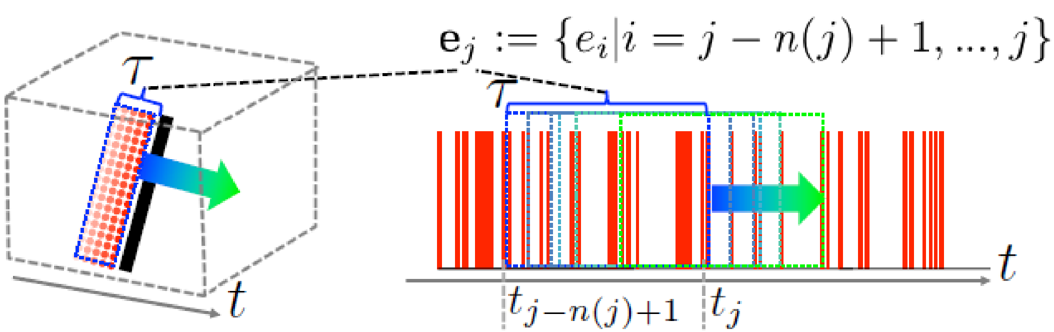


Fig.15 Recursive architecture for spatial-temporal events.

3 Results

Quantitative evaluation

- **ETHTED+** [7]
- **MVSEC** [12]

		ETHTED+			MVSEC	Real-time processing at 1 MEPS
		Semantic segmentation		Object-motion	Ego-motion	
		GA [%]	mIoU [%]	error [pix/τ]	error [deg/sec]	
PointNet		98.9	97.4(0.13)	3.14(0.08)	4.55	NO
EventNet		99.2	97.5(0.22)	3.11(0.28)	**4.29**	YES
Ablation	w/o TD	**99.4**	**98.8**(0.16)	**3.08**(0.32)	—	NO
	w/o TR	98.1	97.9(0.11)	3.74(0.06)		YES
	w/o ALL	98.3	97.1(0.25)	4.14(0.32)		NO

Tab.8 Quantitative evaluation using ETHTED+ and MVSEC.

Computational complexity

	#input mlp1	#input max	mlp1/2	max pool(+t-code)	total	mlp3	mlp4
PointNet	$n(j)$	$n(j)$	936.9×10^3	16.47×10^3	953.3×10^3	$\mathbf{0.58 \times 10^3}$	0.59×10^3
EventNet	1	2	**0.65**(29.27)	**0.36**	**1.01**	0.61×10^3	0.61×10^3

Tab.9 Computational times(us) for processing a single event with EventNet and PointNet.

[7] The event-camera dataset and simulator: event-based data for pose estimation, visual odometry, and SLAM. Elias Mueggler et al, IJRR, 2017.
[12] The multivehicle stereo event camera dataset: an event camera dataset for 3d perception. Alex Z. Zhu et.al. ***IEEE Robotics and Automation Letters,*** 2018.

4 Outlook

- **1** How to further design **local feature** representations?

- **2** How to transform other strategies for point cloud to event-based data?

- 3 Do you believe that **System theory** exists in **event-based vision**?

Summary

Representations	Disadvantages	Advantages
Image	**Lack of temporal information**	**Deep learning**
Time surface	**Complexity & Local feature**	**Spatial-temporal**
Feature	**Multi-steps**	**Complex vision tasks**
End-to-end CNNs	**Lack of datasets & loss function**	**Complex vision tasks**
End-to-end SNNs	**Complex vision tasks?**	**Temporal information**

Tab.10 Representations for spatial-temporal spikes from event cameras

Overview

- Introduction
  - Event-based vision in CVPR 2019
  - Questions
- Related works
  - Time surface representations
  - Transformed images
- End-to-end learning
  - Events-to-video, CVPR 2019
  - Cv3dconv, IEEE Access 2019
  - EventNet, CVPR 2019
- **Discussion**
  - Better input representations for event data
  - Event-based vision in the future

Discussion

- **Better input representations for event data**
  - Spiking neural network [14]
  - Point process theory + machine learning [15]

- **Event-based vision in the future**
  - Event-based cameras
  - Sparse and asynchronous **spatial-temporal point processes**

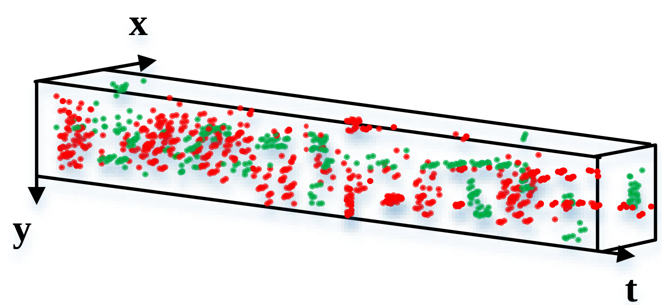


Fig.16 Asynchronous spatial-temporal spikes from event cameras.

[14] SLAYER: spike layer error reassignment in time. Sumit Bam Shrestha et al, NIPS, 2018.
[15] Learning time series associated event sequences with recurrent point process networks. Shuai Xiao et.al. ***TNNLS,*** 2019.

Q&A?

Thanks !