Multimodal Estimation of Movement and Depth Based on Events for Scene Analysis

Vincent Brebion — PhD Defense

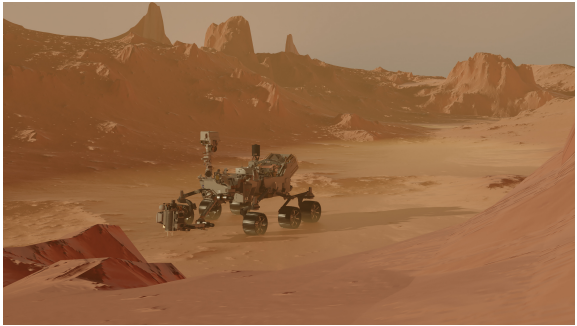
Université de technologie de Compiègne, CNRS, Heudiasyc, SIVALab

January 11, 2024

SIVALab/
Renault
Group

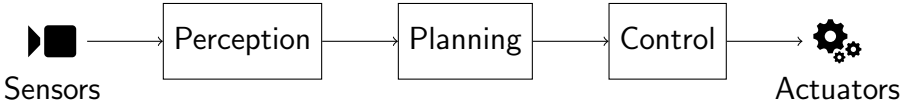
Robots are everywhere

(a) [Wikimedia 2022]

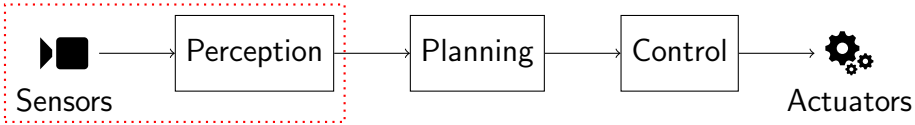


(b) [Wikimedia 2021]

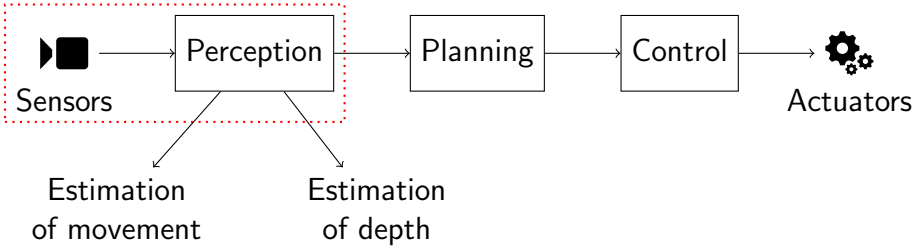
Robotic pipeline



Robotic pipeline



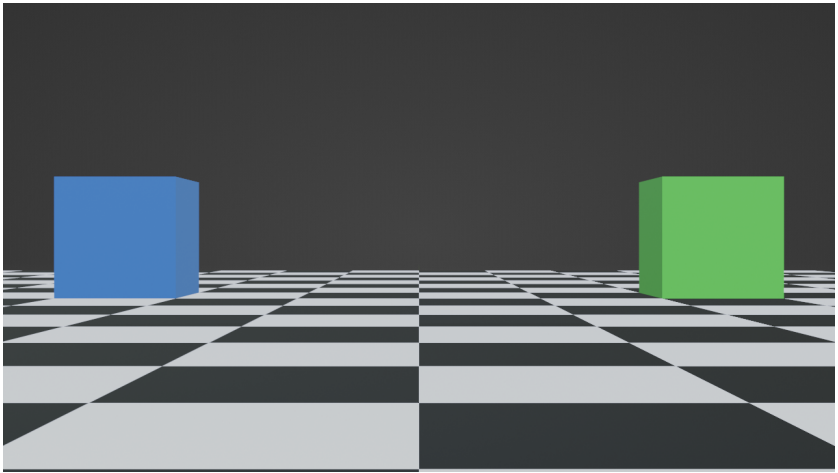
Robotic pipeline



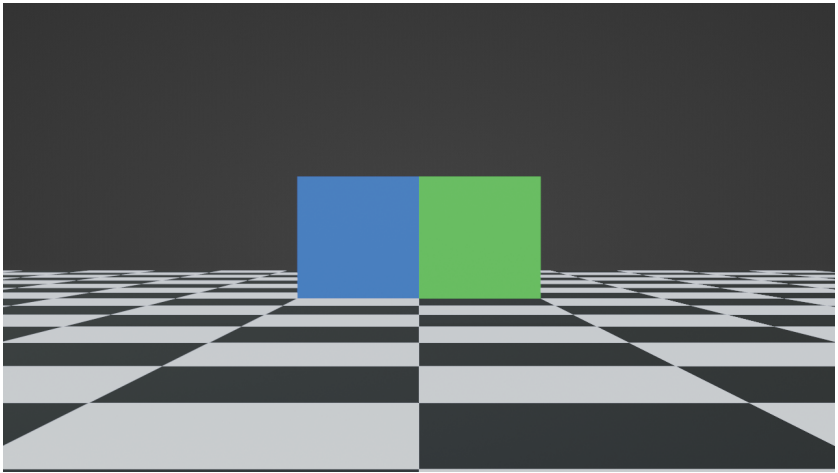
Estimation of movement

- Analysis of the motion of every element in the scene
- 2 subcomponents:
  - Ego-motion
  - Motion of every other mobile object

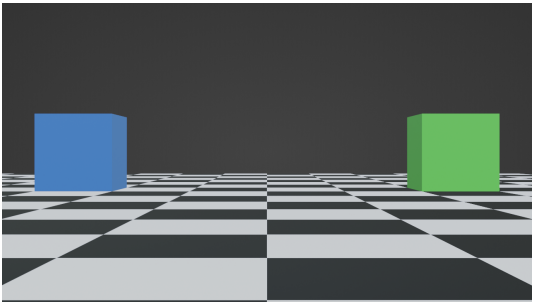
2D motion alone is not enough

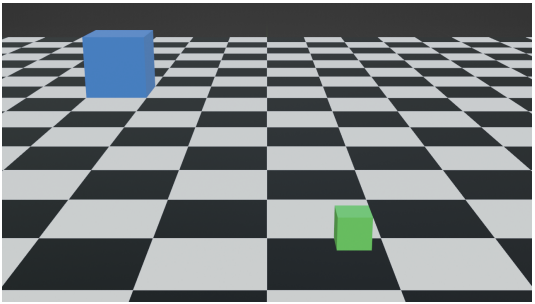


2D motion alone is not enough

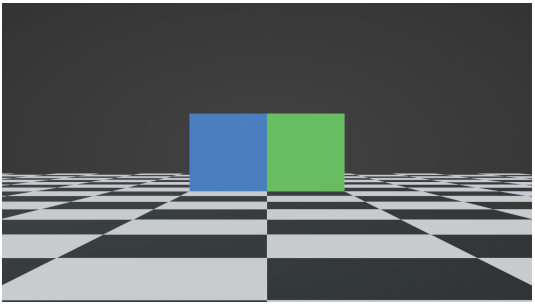


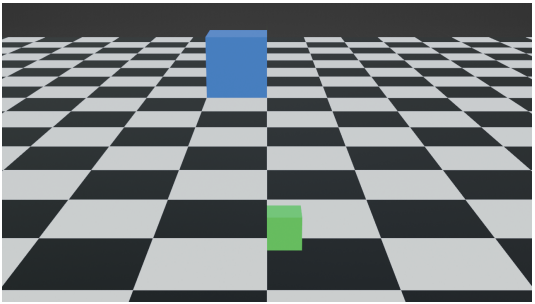
2D motion alone is not enough



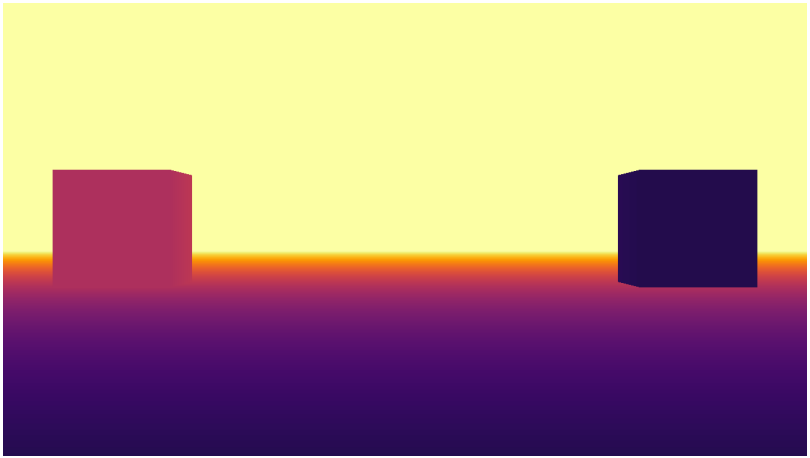


2D motion alone is not enough





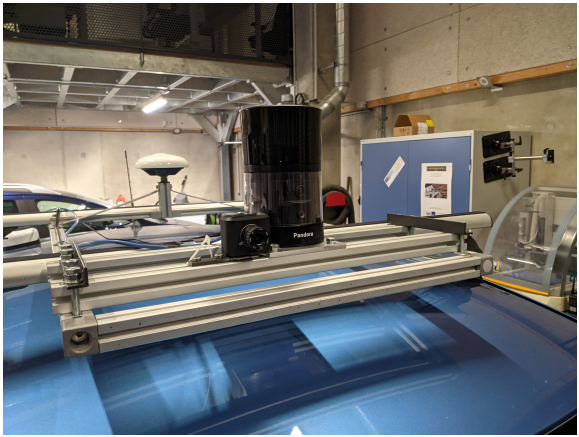
Estimation of depth



Estimation of depth

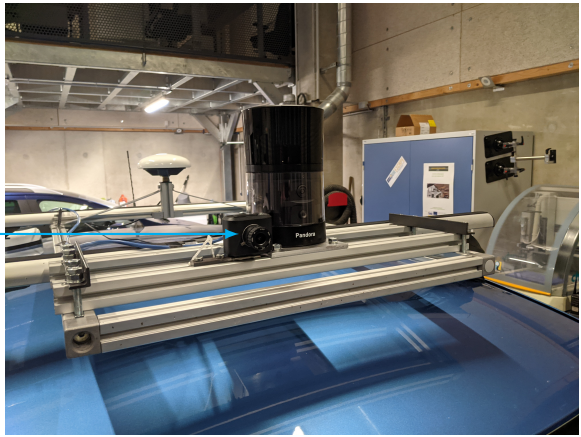
- Critical component when evolving in the real world
- Estimation of movement and of depth serve complementary purposes

Two sensors



Two sensors

Event camera
(usage: motion & depth)



Two sensors

Event camera
(usage: motion & depth)

LiDAR
(usage: depth)

Application to the automotive domain

Figure: Crossroads (scene generated with CARLA [Dosovitskiy et al. 2017])

Application to the automotive domain

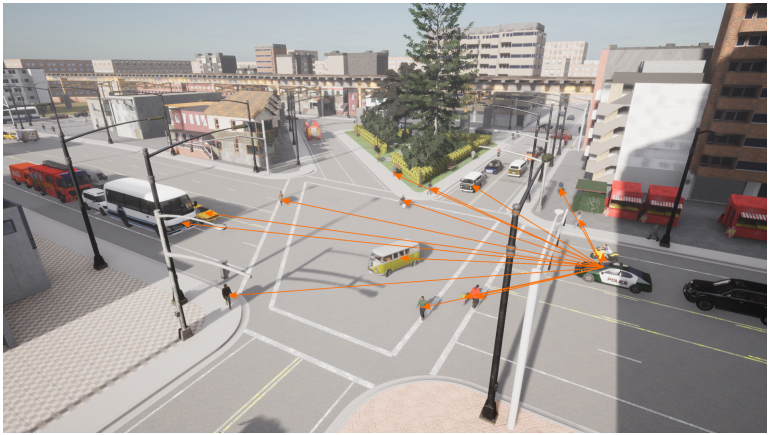
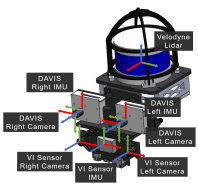
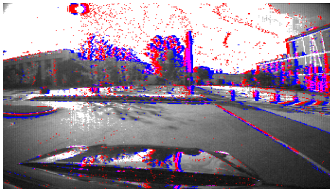


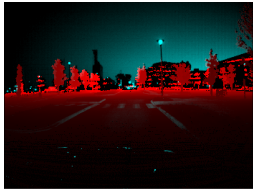
Figure: Crossroads (scene generated with CARLA [Dosovitskiy et al. 2017])

The MVSEC dataset [Zhu et al. 2018b]

- Reference in the literature
- Automotive scenes
- Event and LiDAR data
- Ground truth for motion & depth







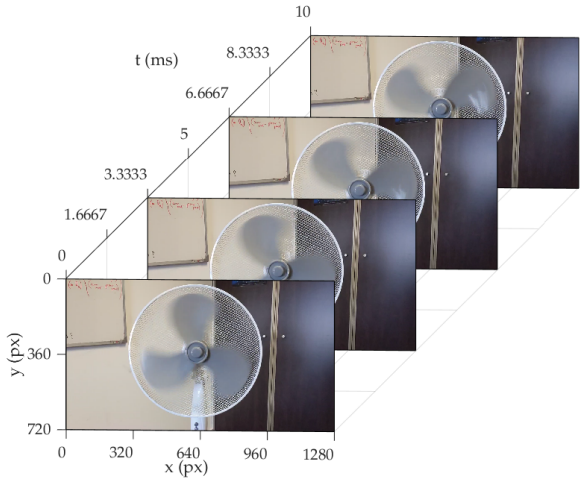
Outline

Outline

Frame-based cameras

- All pixels respond synchronously
- Light accumulation during a short time window
- Dense representation

Frame-based cameras



Core principle of an event camera

- Every pixel is independent and asynchronous
- An event $e \doteq (\mathbf{x}, t, p)$ is generated for pixel $\mathbf{x} = (x, y)$ at time t if:

$$|\Delta L(\mathbf{x}, t)| \geq \delta$$

with

$$\Delta L(\mathbf{x}, t) \doteq L(\mathbf{x}, t) - L(\mathbf{x}, t - \Delta t)$$

and

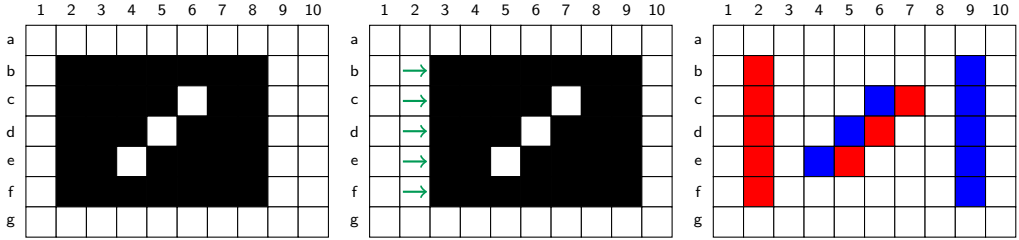
$$p \doteq \text{sgn}(\Delta L(\mathbf{x}, t))$$

Triggering events

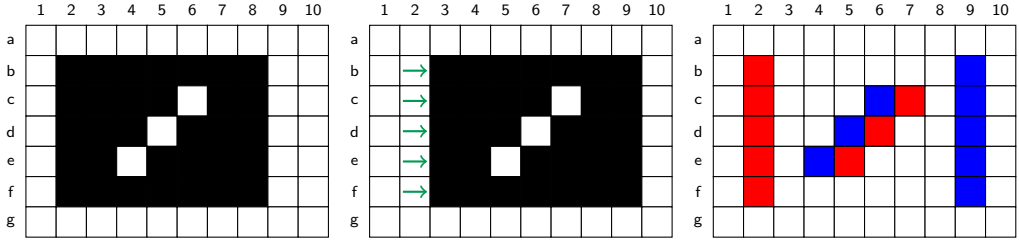
Events can be triggered by two highly different reasons:

1. **lighting changes** in the observed scene;
2. **relative motion** between the camera and the other objects in the scene.

Triggering events with motion



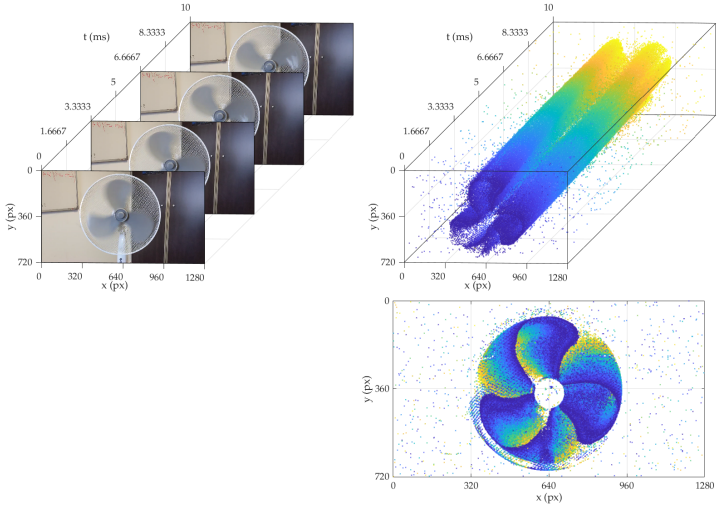
Triggering events with motion



Key property 1

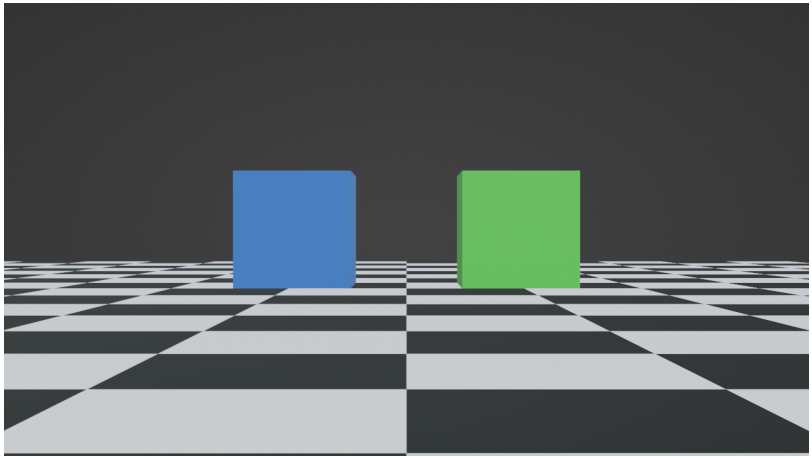
Under motion, events are only produced for **edges** and **textures** of objects

Real example: a rotating fan

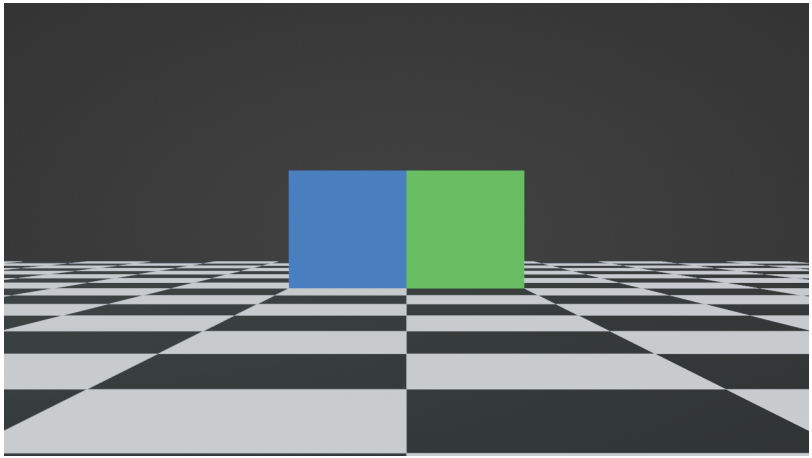


Outline

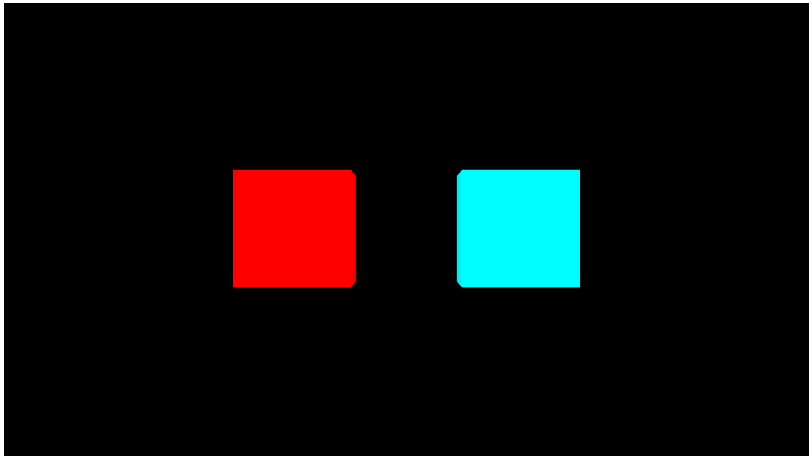
Frame-based optical flow

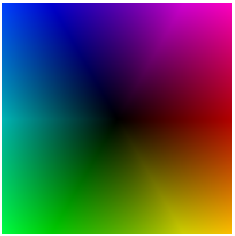


Frame-based optical flow



Frame-based optical flow



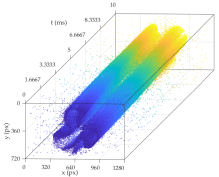


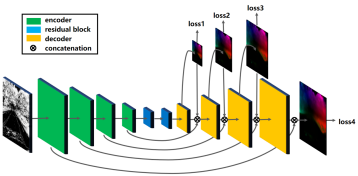
Frame-based optical flow

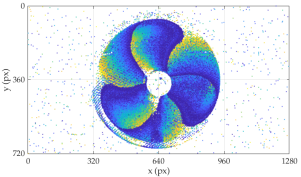
- 2D motion field of the observed scene
- 1981: Horn-Schunck [Horn et al. 1981], Lucas-Kanade [Lucas et al. 1981]
  2022: GMFlow [Xu et al. 2022]
- Still an open area of research for event cameras

Event-based optical flow

- Purely event-based
  [Benosman et al. 2014; Gallego et al. 2018]
- Learning-based
  [Zhu et al. 2018a; Gehrig et al. 2021c; Liu et al. 2023]
- Frame-based
  [Almatrafi et al. 2020; Nagata et al. 2021]







Real-time event-based optical flow

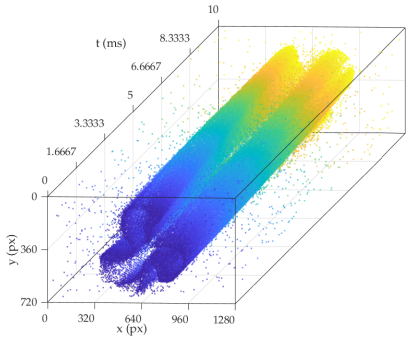
- **Real-time** event-based optical flow

Real-time event-based optical flow

- **Real-time** event-based optical flow
- Even with **high-definition** event cameras
  - Up to millions of events per second

Real-time event-based optical flow

- Purely event-based
  - Slow, as each event is considered individually
  - Hard to optimize on current hardware



Real-time event-based optical flow

- Learning-based
  - Large networks are accurate but slow
  - Small networks are fast but lack accuracy

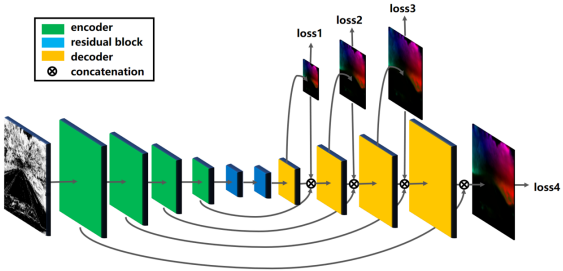
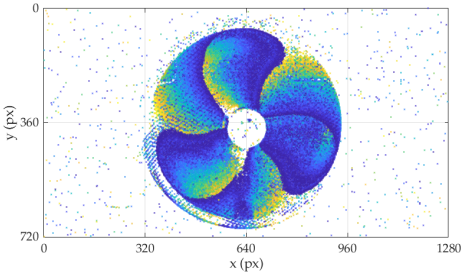


Figure: EV-FlowNet [Zhu et al. 2018a]

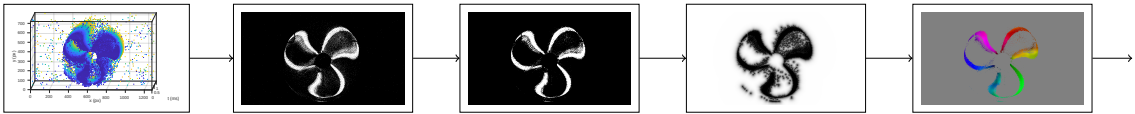
Real-time event-based optical flow

- Frame-based
  - Only the asynchrony is lost
    - Still low reaction times, HDR, no under-/over-exposure
  - Reuse of state-of-the-art frame-based methods
  - GPUs
  - No learning process



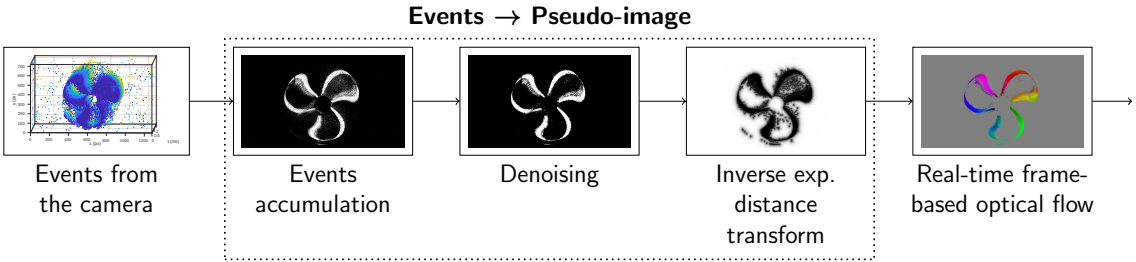
Proposed architecture: RTEF

- Pipeline architecture (parallelism of the tasks)
- Transformation from a sparse input to a frame-like representation, on which optical flow is computed



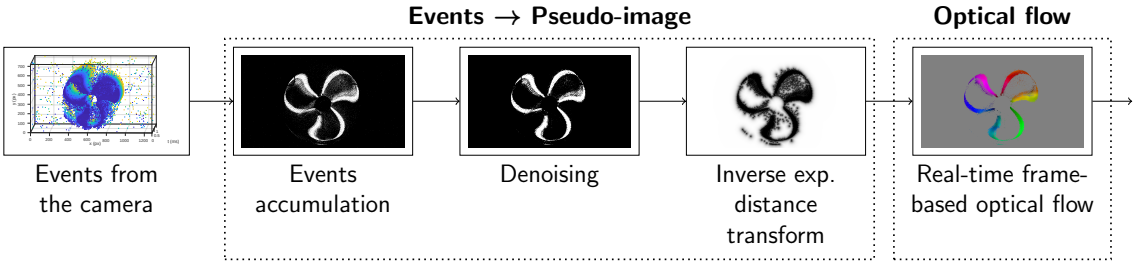
Proposed architecture: RTEF

- Pipeline architecture (parallelism of the tasks)
- Transformation from a sparse input to a frame-like representation, on which optical flow is computed



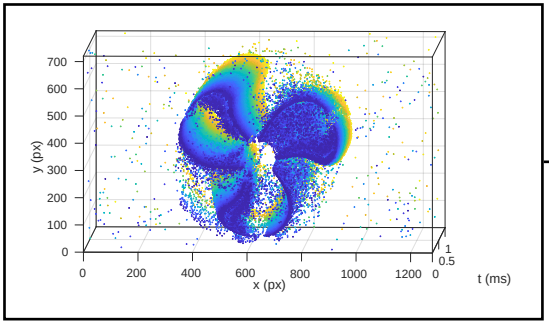
Proposed architecture: RTEF

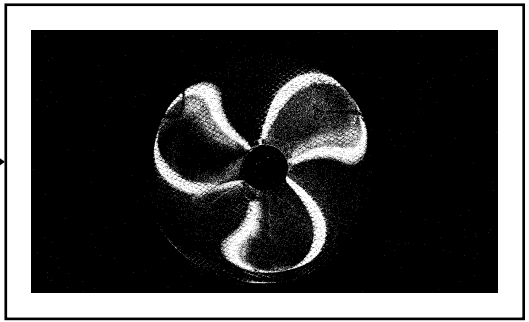
- Pipeline architecture (parallelism of the tasks)
- Transformation from a sparse input to a frame-like representation, on which optical flow is computed



Step 1: Events accumulation

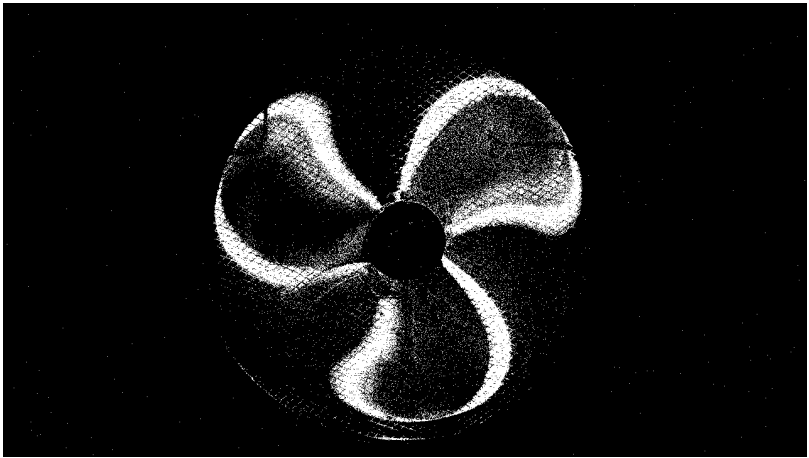
- Accumulation over a fixed short time window Δt [4ms – 50ms]
- Conversion from sparse events into a first 2D representation





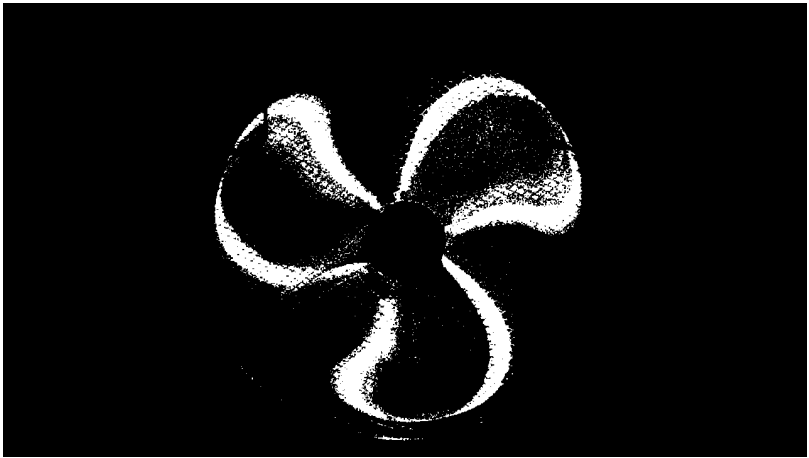
Step 2: Denoising

- Noisy binary image



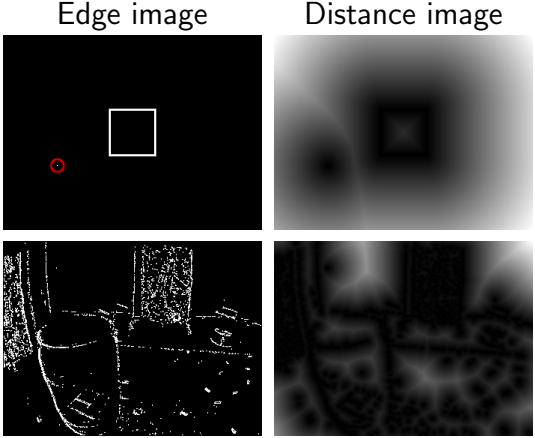
Step 2: Denoising

- Stabilization of the appearance, by only keeping predominant edges



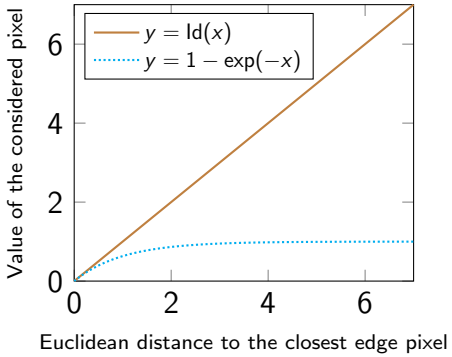
Step 3: Distance transform

- Transformation proposed by [Almatrafi et al. 2020]

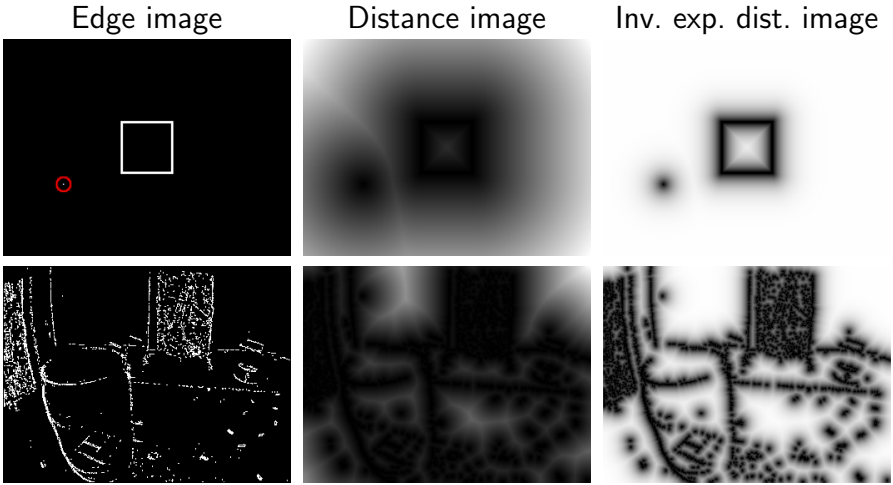


Step 3: Inverse exponential distance transform

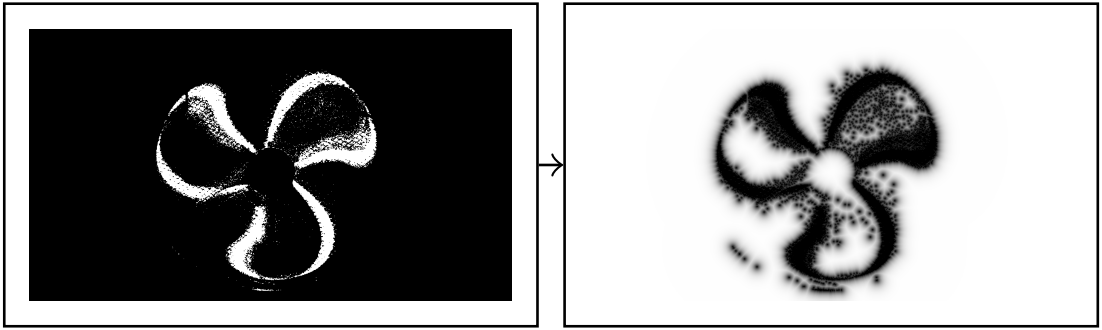
- $d_{\text{exp}} \doteq 1 - \exp(-d/\alpha)$



Step 3: Inverse exponential distance transform

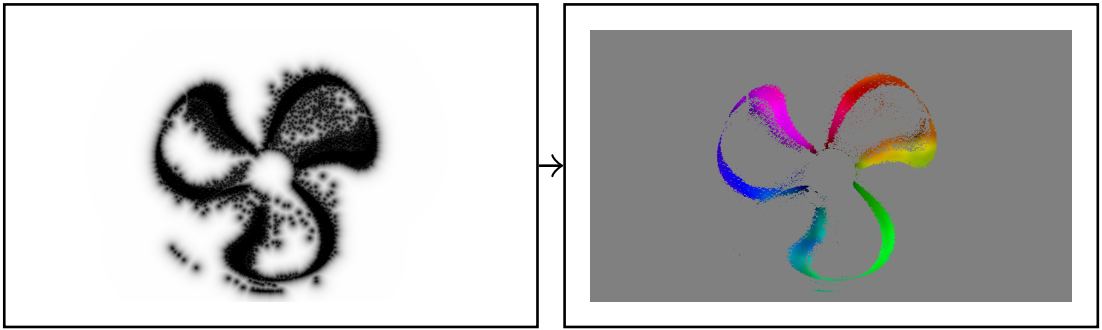


Step 3: Inverse exponential distance transform



Step 4: Real-time frame-based optical flow

- Use of a state-of-the-art method: [Adarve et al. 2016]



Results on the MVSEC dataset (346×260)

	Indoor flying 1		Indoor flying 2		Indoor flying 3		Outdoor day 1	
	AEE (px) ↓	% outliers ↓	AEE (px) ↓	% outliers ↓	AEE (px) ↓	% outliers ↓	AEE (px) ↓	% outliers ↓
RTEF (ours)	0.52	**0.1**	0.98	5.5	0.71	2.1	0.53	0.2
MultiCM [Shiba et al. 2022]	**0.42**	**0.1**	**0.60**	**0.6**	**0.50**	**0.3**	**0.30**	**0.1**
FireFlowNet [Paredes-Vallés et al. 2021]	0.97	2.6	1.67	15.3	1.43	11.0	1.06	6.6

Table: Raw errors on the optical flow

RTEF (ours)	MultiCM [Shiba et al. 2022]	FireFlowNet [Paredes-Vallés et al. 2021]
250Hz	0.03Hz	262Hz

Table: Execution speed

Results on a 20-minute-long driving sequence (1280×720)

Still high inference rate: 83Hz

Conclusion

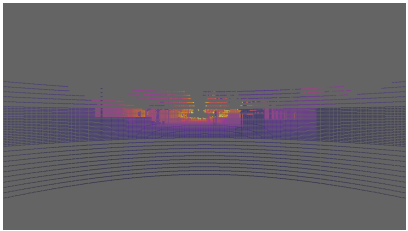
- Contributions
  - First real-time event-based optical flow method, even with high-resolution cameras
  - Good accuracy
- Limitations
  - Accuracy limited by the real-time constraint

Outline

LiDAR and event camera

LiDAR

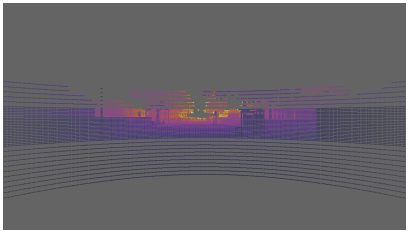
- Accurate depth data
- Limitations:
  - Low frequency (10–20Hz)
  - Sparse point cloud
  - Limited vertical range



LiDAR and event camera

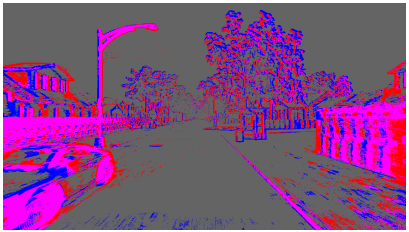
LiDAR

- Accurate depth data
- Limitations:
  - Low frequency (10–20Hz)
  - Sparse point cloud
  - Limited vertical range



Event camera

- Sparse but rich spatio-temporal data
- Under motion, highlights the edges of objects
- Could complement well the LiDAR data



Dual objective

LiDAR densification, using the events as a guide

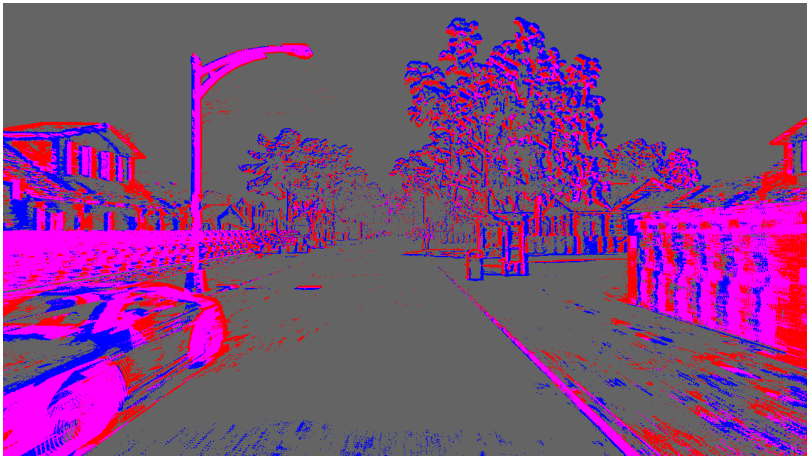
Giving a depth to each event, allowing for their 3D reprojection

Two depths per event

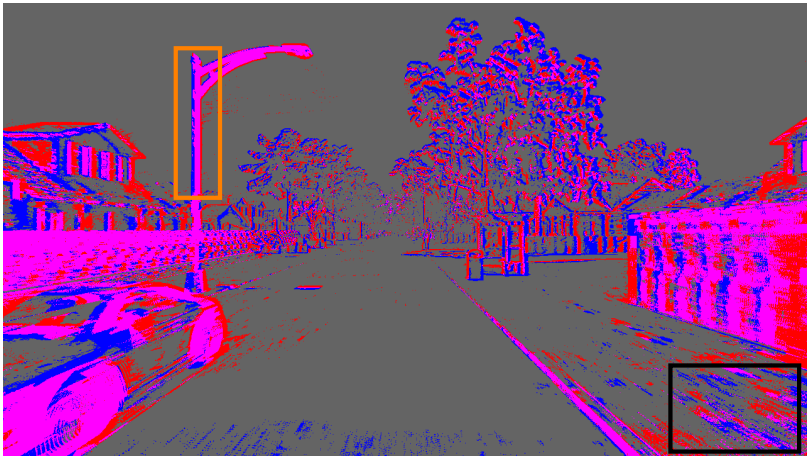
- An event originates from a **change of brightness** ΔL

$$\begin{aligned}\Delta L(\mathbf{x}, t) &\doteq L(\mathbf{x}, t) - L(\mathbf{x}, t - \Delta t) \\ &\doteq L_t - L_{t-\Delta t}\end{aligned}$$

Two depths per event



Two depths per event



Two depths per event

Two depths per event

- An event originates from a **change of brightness** ΔL

$$\begin{aligned}\Delta L(\mathbf{x}, t) &\doteq L(\mathbf{x}, t) - L(\mathbf{x}, t - \Delta t) \\ &\doteq L_t - L_{t-\Delta t}\end{aligned}$$

- Under motion, this change of brightness might be due to a **change of depth** Δd

$$\begin{aligned}\Delta d &\doteq d_t - d_{t-\Delta t} \\ &\doteq d_{\text{af}} - d_{\text{bf}}\end{aligned}$$

- So, each event should be associated with **two depths**:
  - one **after** the event: d_{af}
  - one **before** the event: d_{bf}

Issues

Data fusion between

- two very different modalities,
- with different output rates,
- not covering the same parts of the image,
- and with potential noise.

Traditional 2D and 3D geometry

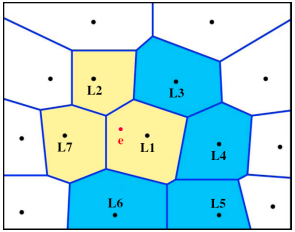


Figure: 2D method of [Li et al. 2021]

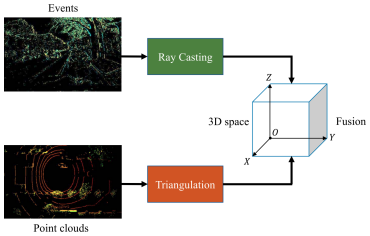


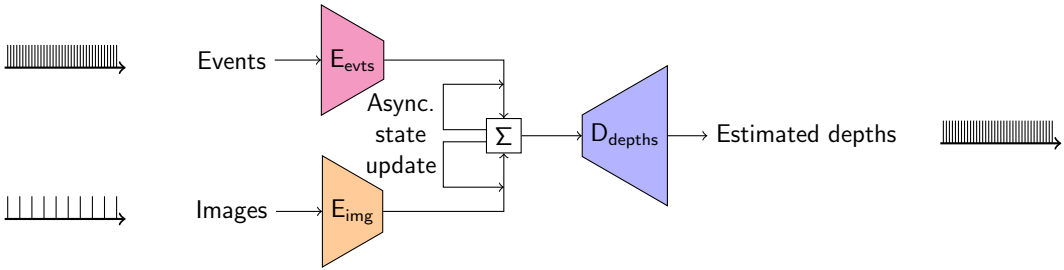
Figure: 3D method of [Cui et al. 2022]

Limitations of these methods:

- Only work for areas with LiDAR data
- Restricted to the frequency of the LiDAR
- Sensitive to noise

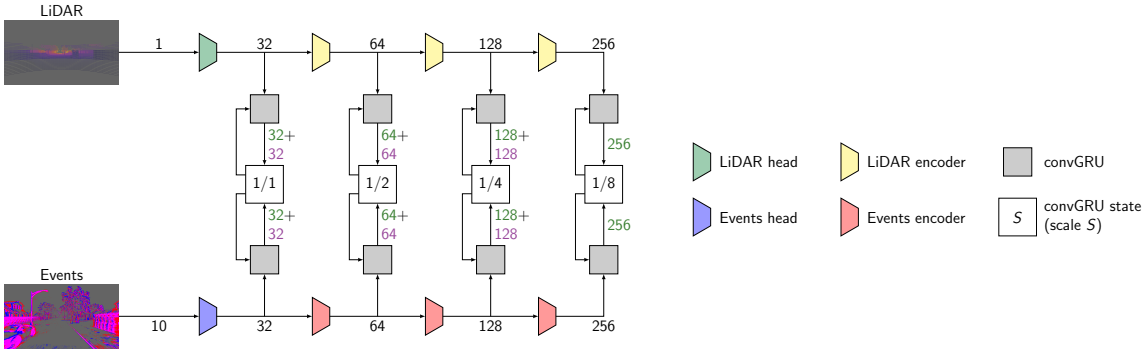
Learning-based

Fusion of events and images: RAMNet [Gehrig et al. 2021a]

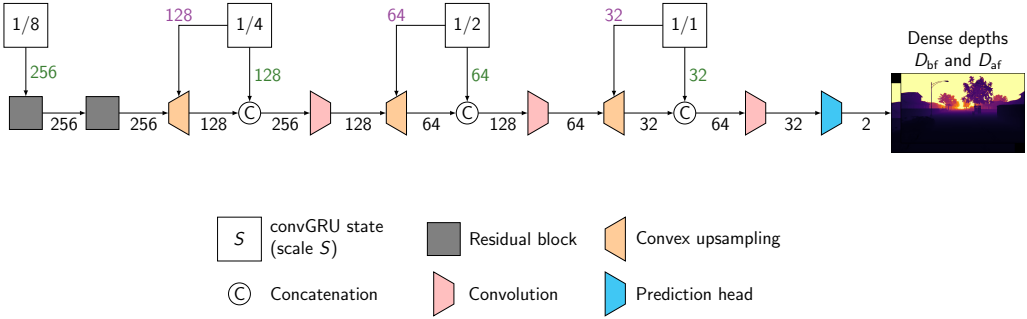


Our network: ALED (encoding part)

ALED: Asynchronous LiDAR and Events Depths densification network

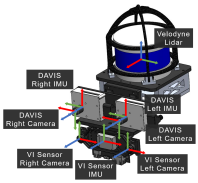


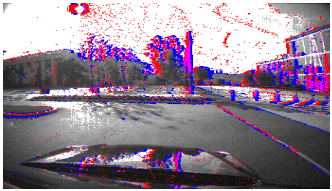
Our network: ALED (decoding part)

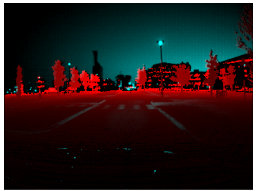


The MVSEC dataset [Zhu et al. 2018b]

- Limitations
  - Low resolution (346×260)
  - Incorrect ground truth for moving objects
  - Approximate synchronization and calibration



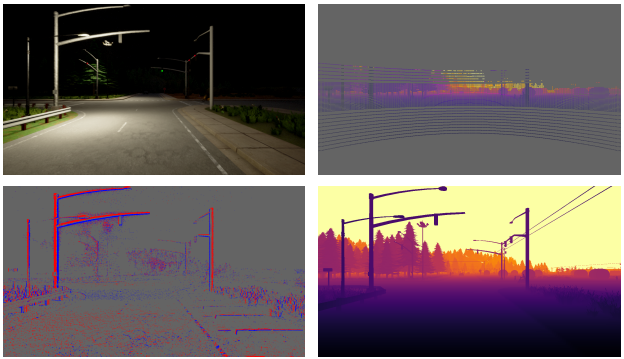




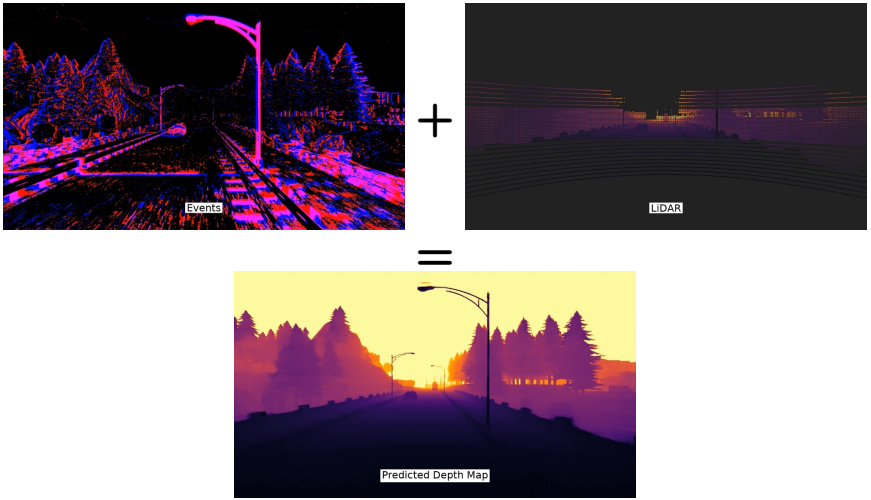
Our open dataset: SLED

SLED: Synthetic LiDAR Events Depths dataset

- CARLA [Dosovitskiy et al. 2017]
  - 160 sequences of 10s each
  - Diverse environments
  - Train/val/test sets
- Advantages
  - High definition (1280×720)
  - Perfect ground truth
  - Perfect synchronization and calibration



Results on our SLED dataset



Results on the MVSEC dataset

	Cutoff	Events (stereo)	Events & Images	Events & LiDAR		
		StereoSpike [Rançon et al. 2021]	EvT$^+$ [Sabater et al. 2022]	3D geometry [Cui et al. 2022]	**ALED$_{\text{MVSEC}}$**	**ALED$_{\text{SLED}\to\text{MVSEC}}$**
Outdoor day 1	10m	**0.79**	1.24	1.24	0.91	
	20m	1.47	1.91	1.28	**1.22**	
	30m	1.92	2.36	4.87	**1.43**	
	50m	—	—	—	**1.67**	
	100m	3.17	—	—	**1.96**	
Outdoor night 1	10m	**1.38**	1.45	2.26	1.75	
	20m	2.26	**2.10**	2.19	**2.10**	
	30m	2.97	2.88	4.50	**2.25**	
	50m	—	—	—	**2.44**	
	100m	4.82	—	—	**2.73**	
Outdoor night 2	10m	—	1.48	1.88	**1.19**	
	20m	—	2.13	2.14	**1.65**	
	30m	—	2.90	4.67	**1.81**	
	50m	—	—	—	**1.95**	
	100m	—	—	—	**2.11**	
Outdoor night 3	10m	—	1.38	1.78	**0.85**	
	20m	—	2.03	1.93	**1.25**	
	30m	—	2.77	4.55	**1.42**	
	50m	—	—	—	**1.57**	
	100m	—	—	—	**1.73**	

Table: Average depth errors (in meters)

Results on the MVSEC dataset

	Cutoff	Events (stereo)	Events & Images	Events & LiDAR		
		StereoSpike [Rançon et al. 2021]	EvT$^+$ [Sabater et al. 2022]	3D geometry [Cui et al. 2022]	**ALED$_{\text{MVSEC}}$**	**ALED$_{\text{SLED}\to\text{MVSEC}}$**
Outdoor day 1	10m	0.79	1.24	1.24	0.91	**0.50**
	20m	1.47	1.91	1.28	1.22	**0.80**
	30m	1.92	2.36	4.87	1.43	**1.02**
	50m	—	—	—	1.67	**1.31**
	100m	3.17	—	—	1.96	**1.60**
Outdoor night 1	10m	**1.38**	1.45	2.26	1.75	1.52
	20m	2.26	2.10	2.19	2.10	**1.81**
	30m	2.97	2.88	4.50	2.25	**1.95**
	50m	—	—	—	2.44	**2.20**
	100m	4.82	—	—	2.73	**2.54**
Outdoor night 2	10m	—	1.48	1.88	1.19	**1.09**
	20m	—	2.13	2.14	1.65	**1.49**
	30m	—	2.90	4.67	1.81	**1.64**
	50m	—	—	—	1.95	**1.80**
	100m	—	—	—	2.11	**1.97**
Outdoor night 3	10m	—	1.38	1.78	0.85	**0.81**
	20m	—	2.03	1.93	1.25	**1.16**
	30m	—	2.77	4.55	1.42	**1.33**
	50m	—	—	—	1.57	**1.51**
	100m	—	—	—	1.73	**1.66**

Table: Average depth errors (in meters)

Conclusion

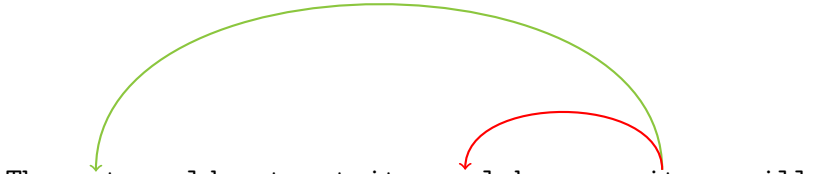
- Contributions
  - First asynchronous LiDAR and event fusion
  - Improvement of up to 79.1% compared to the current state of the art
  - Novel SLED dataset, to improve training and evaluation
- Main limitation
  - Better accuracy still needed for short ranges

Outline

Concept of attention

The cat could not eat its meal because it was ill.

Concept of attention

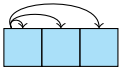


Concept of attention

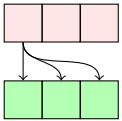
- Attention: assign a score for each pair of elements, based on how they are related
- The concept of attention is not new [Nadaraya 1964; Watson 1964]
- Made popular again by Transformers [Vaswani et al. 2017]

Self- and cross-attention

- Self-attention



- Cross-attention

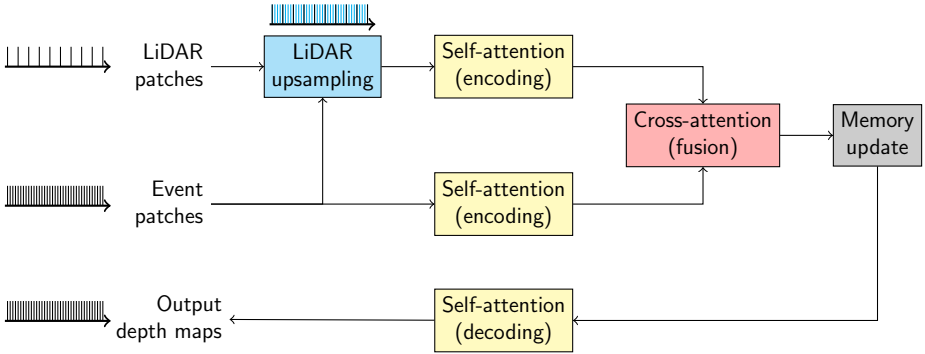


Attention for images

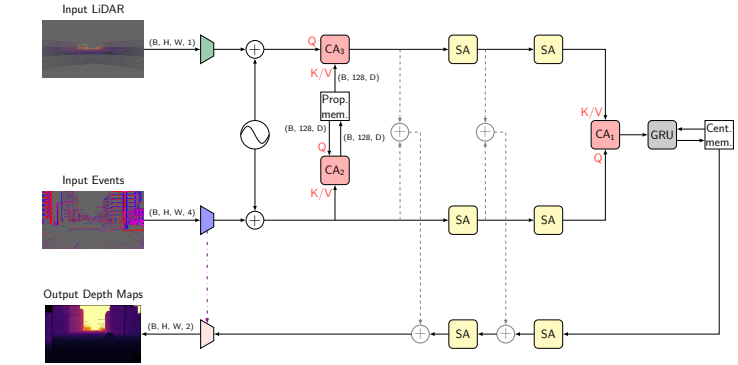
Figure: Image patches [Dosovitskiy et al. 2021]

Overview of our network: DELTA

DELTA: Dense depths from Events and LiDAR using Transformer's Attention



Our network: DELTA

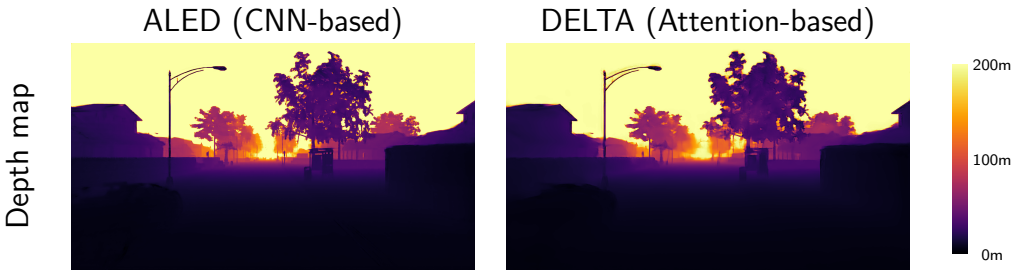


Results on our SLED dataset

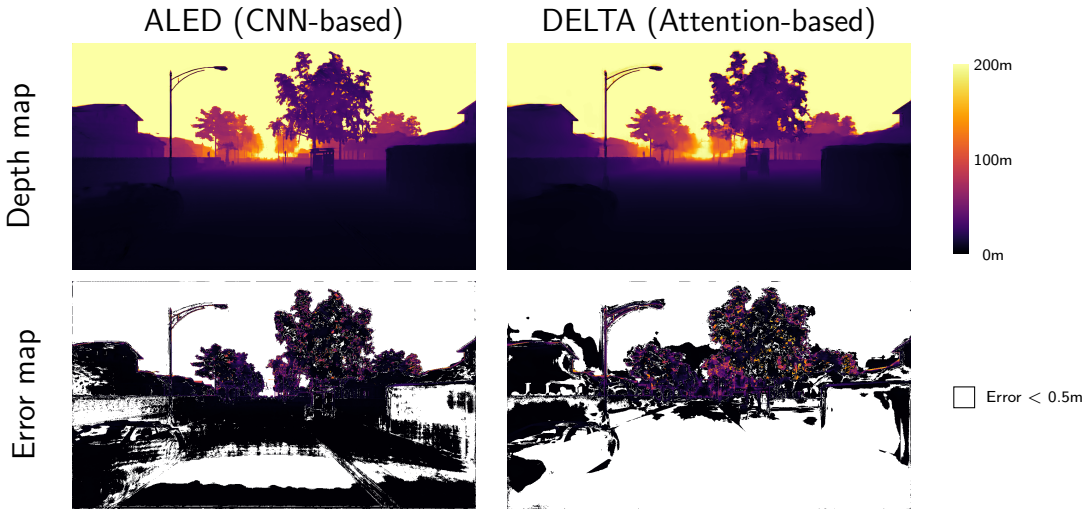
	Cutoff	On D_{bf}		On D_{af}	
		ALED	**DELTA**	**ALED**	**DELTA**
Town01	10m	1.24	**0.64**	1.37	**0.67**
	20m	2.08	**1.45**	2.27	**1.50**
	30m	2.72	**2.11**	2.92	**2.17**
	100m	4.25	**3.80**	4.51	**3.89**
	200m	**4.53**	5.37	**4.81**	5.42
Town03	10m	2.00	**0.49**	2.09	**0.50**
	20m	2.85	**1.15**	2.97	**1.18**
	30m	3.33	**1.72**	3.45	**1.77**
	100m	4.60	**3.12**	4.77	**3.18**
	200m	4.86	**4.81**	5.03	**4.81**

Table: Average depth errors (in meters)

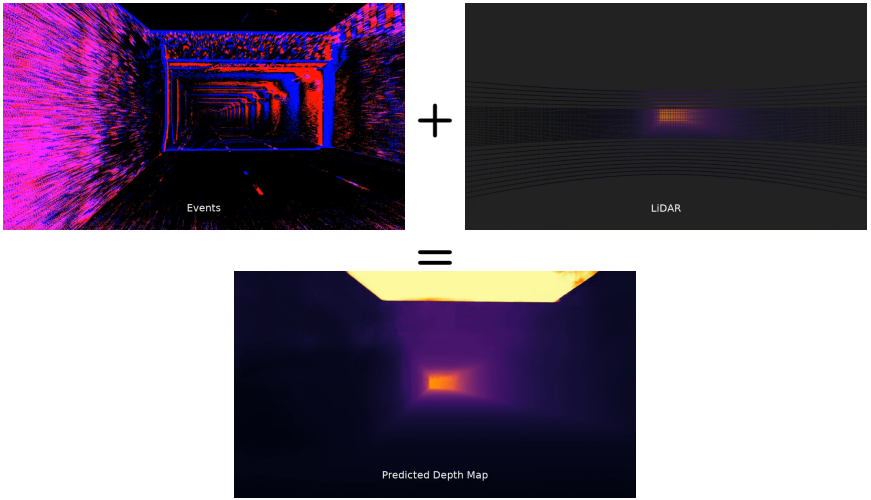
Results on our SLED dataset



Results on our SLED dataset



Results on our SLED dataset



Conclusion

- Contributions
  - First attention-based network for fusing event and LiDAR data
  - Average error reduced up to four times for short ranges when compared to ALED
- Limitations
  - Slightly less accurate than ALED for long ranges
  - Large network (ALED: 26 million of parameters, DELTA: 180 million)

Outline

Contributions

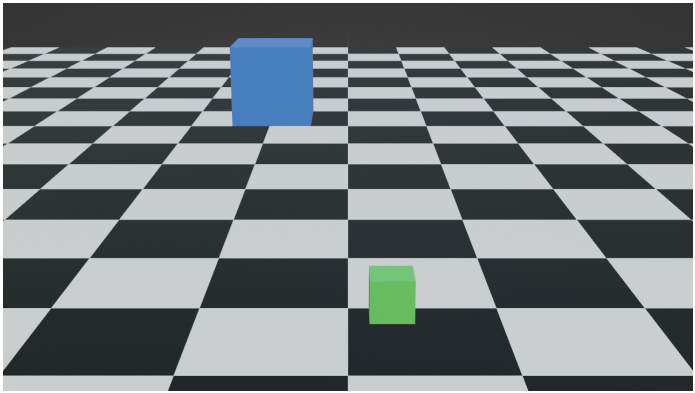
- Estimation of **motion**
  - Focus set on real-time compatibility and accuracy (geometry-based)
  - Proposition of a novel pipeline-based architecture: RTEF
- Estimation of **depth**
  - Focus set on accuracy (learning-based)
  - Proposition of "two depths per event"
  - Proposition of two networks: ALED (CNN-based) and DELTA (Attention-based)
  - Proposition of a novel dataset: SLED

Discussions

1. Scene flow: combination of optical flow and depth maps

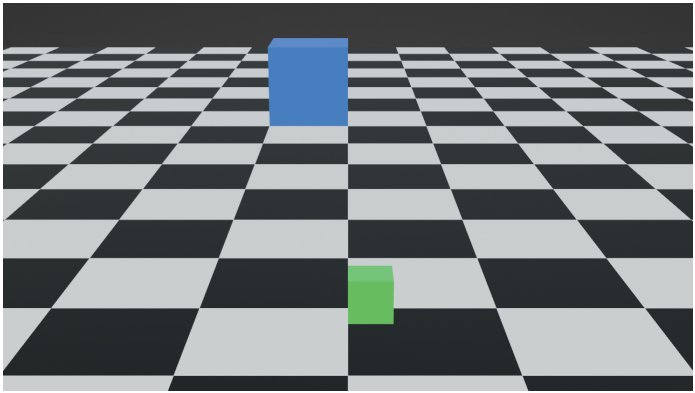
Discussions

Scene flow: 3D estimation of motion



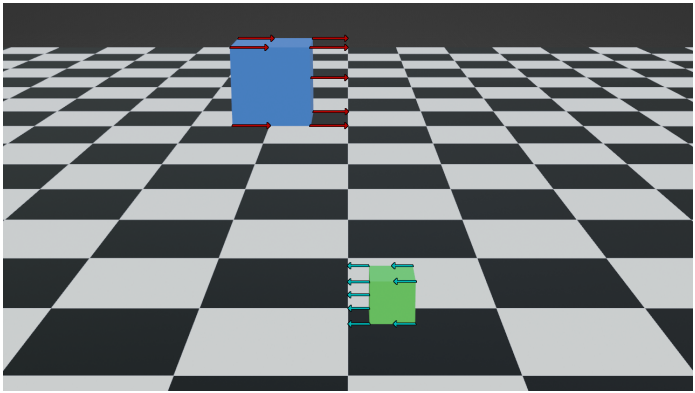
Discussions

Scene flow: 3D estimation of motion



Discussions

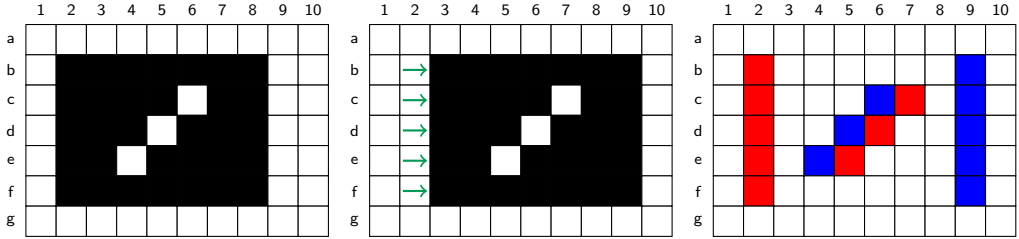
Scene flow: 3D estimation of motion



Discussions

1. Scene flow: combination of optical flow and depth maps
2. We made the assumption that events are stable edges (spatial derivative)

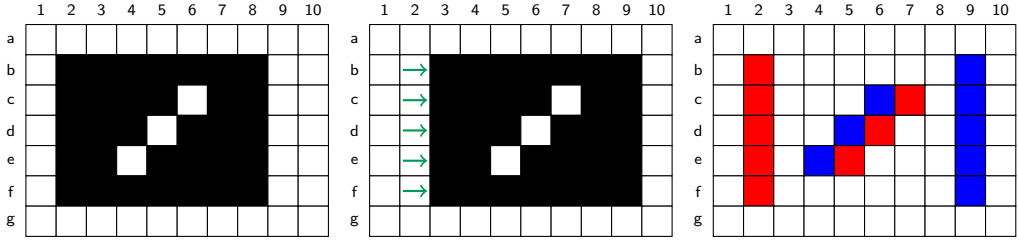
Discussions



Key property 1

Under motion, events are only produced for **edges** and **textures** of objects

Discussions



Key property 1

Under motion, events are only produced for **edges** and **textures** of objects

Key property 2

Events are a **temporal** derivative, not a **spatial** derivative

Discussions

1. Scene flow: combination of optical flow and depth maps
2. We made the assumption that events are stable edges (spatial derivative)
3. Recording of a real-world dataset

Publications

- Optical flow
  - Two articles published: IEEE T-ITS & RFIAP 2022
  - Paper, code, videos: https://vbrebion.github.io/RTEF
- CNN-based depth estimation
  - Article published for SCIA 2023
  - Paper, code, dataset, videos: https://vbrebion.github.io/ALED
- Attention-based depth estimation
  - Article submitted to CVPR 2024





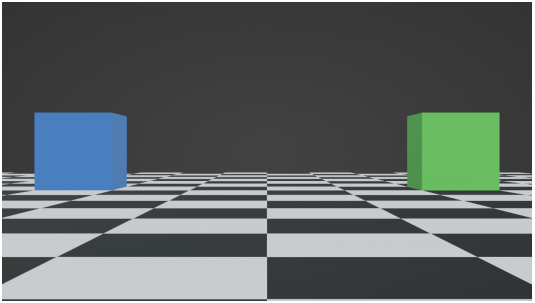
Thank you for your attention!

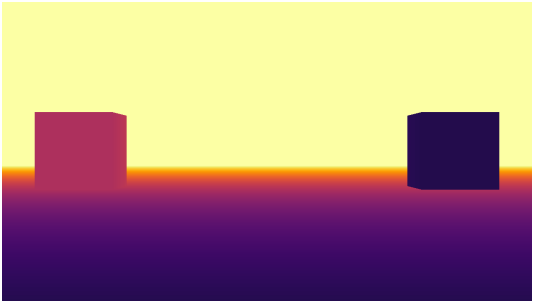


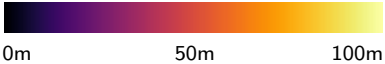


Vannes (France), Sirkka (Finland), Vancouver (Canada)

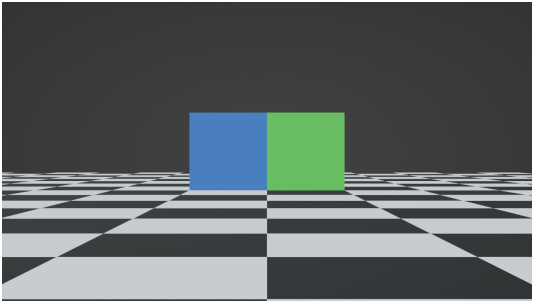
Importance of depth

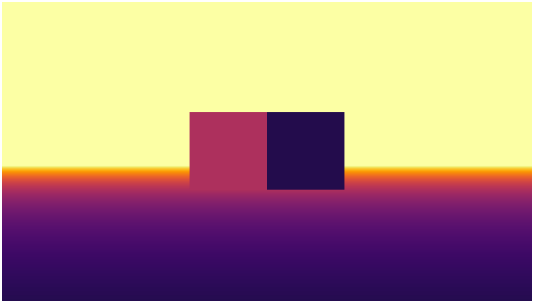


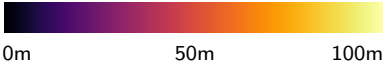




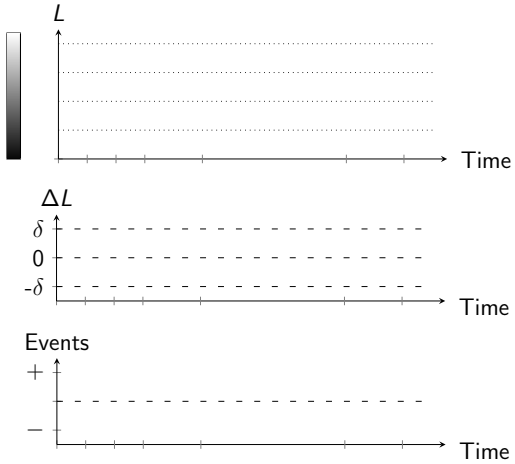
Importance of depth





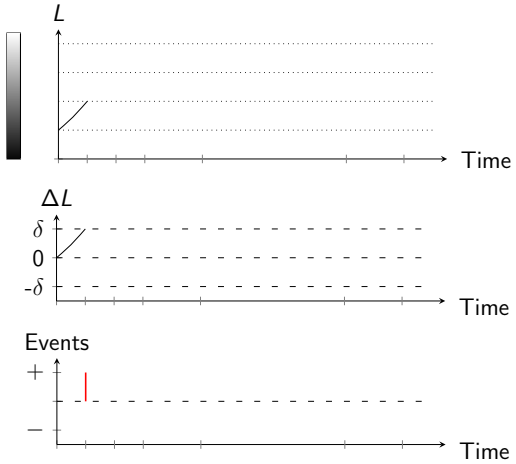


Events for a single pixel



Output of this pixel:

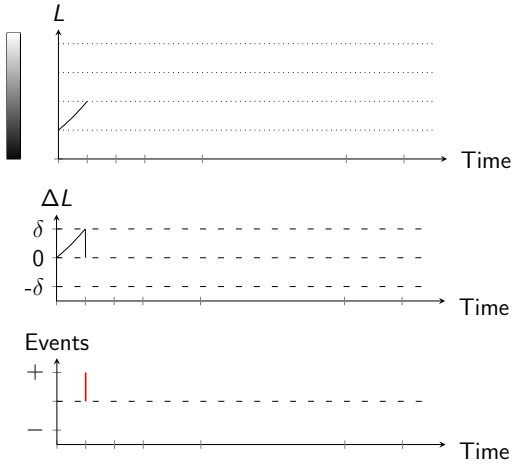
Events for a single pixel



Output of this pixel:

```
x = 266
y = 14
t = 1
p = +
```

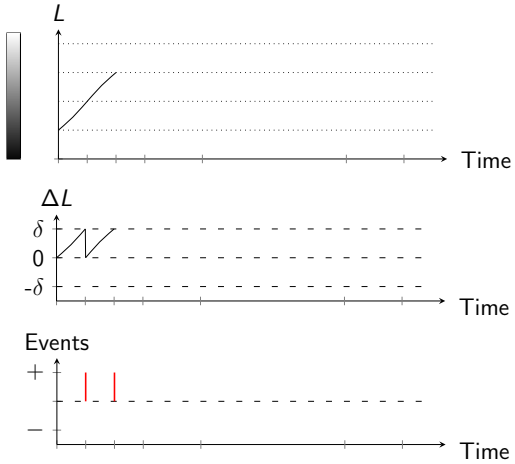
Events for a single pixel



Output of this pixel:

```
x = 266
y = 14
t = 1
p = +
```

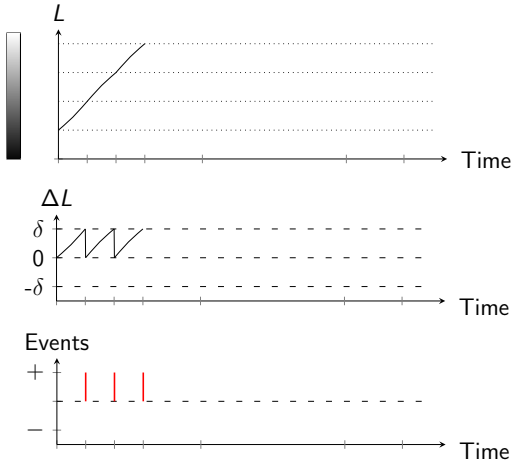
Events for a single pixel



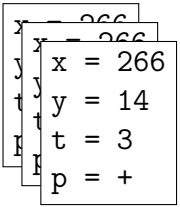
Output of this pixel:

```
x = 266
y = 14
t = 2
p = +
```

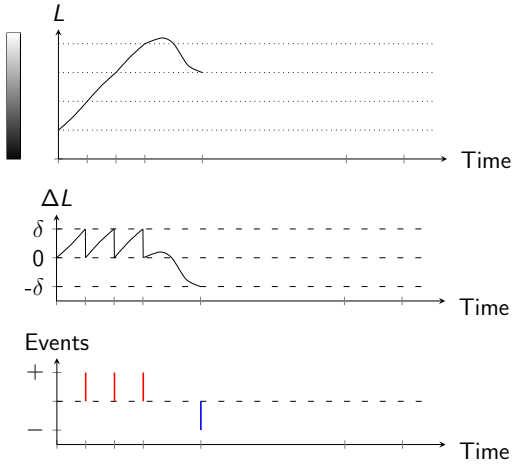
Events for a single pixel



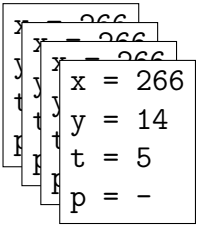
Output of this pixel:



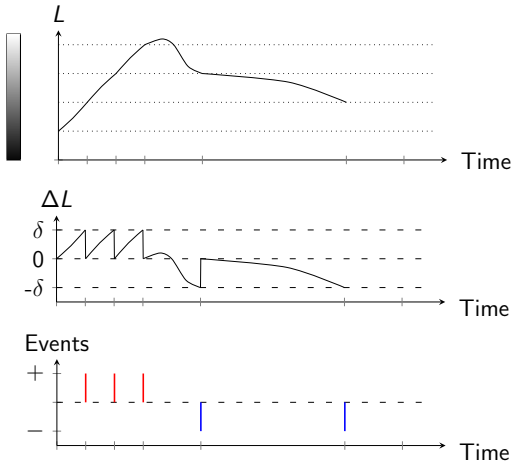
Events for a single pixel



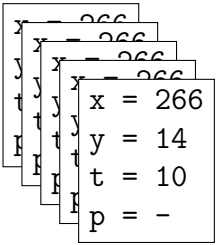
Output of this pixel:



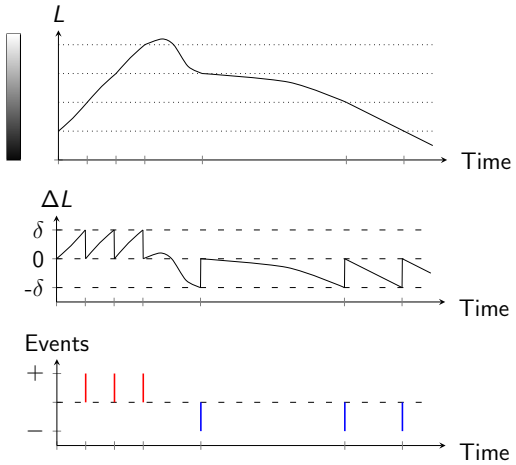
Events for a single pixel



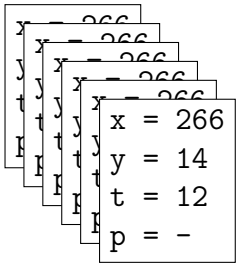
Output of this pixel:



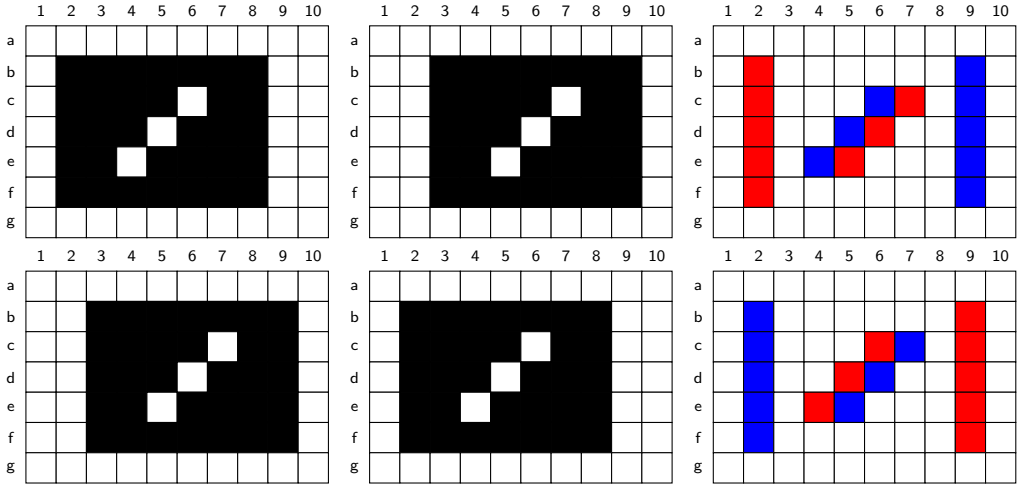
Events for a single pixel



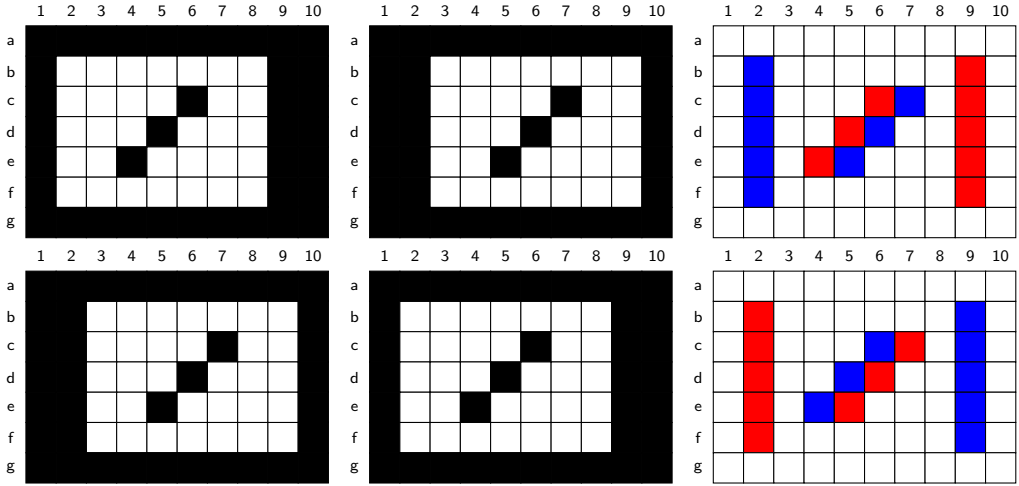
Output of this pixel:



Triggering events with motion (all cases I)



Triggering events with motion (all cases II)



Advantages and challenges of using event cameras

Advantages:

- Asynchrony
  - Very low reaction times
  - No motion blur
  - No under-/over-exposure
- High Dynamic Range (HDR)
- Low energy consumption

Challenges:

- Change of paradigm
- Noise
- Output rates
- Lack of absolute brightness values

Datasets of the state of the art used in this thesis

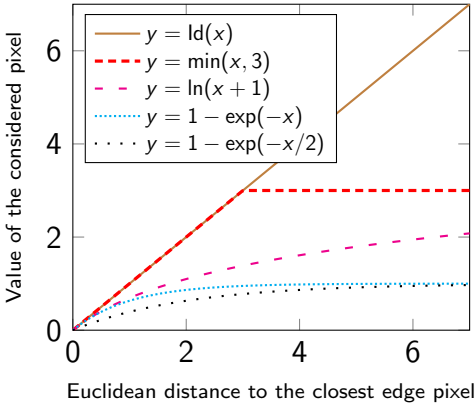
	MVSEC [Zhu et al. 2018b]	**DSEC** [Gehrig et al. 2021b]	**M3ED** [Chaney et al. 2023]
Year	2018	2021	2023
Event camera (resolution)	DAVIS346 (346×260)	Prophesee Gen3.1 (640×480)	Prophesee Gen4 (1280×720)
Application(s)	Motion + Depth	Motion only	Depth only

Definition of "real-time" with event cameras

Definition

A process using events is real-time if:
(1) no event is discarded, even under high input loads, and
(2) if events are accumulated over Δt, they should be processed in less than Δt

All tested distance transform formulations



Flow Warping Loss (FWL)

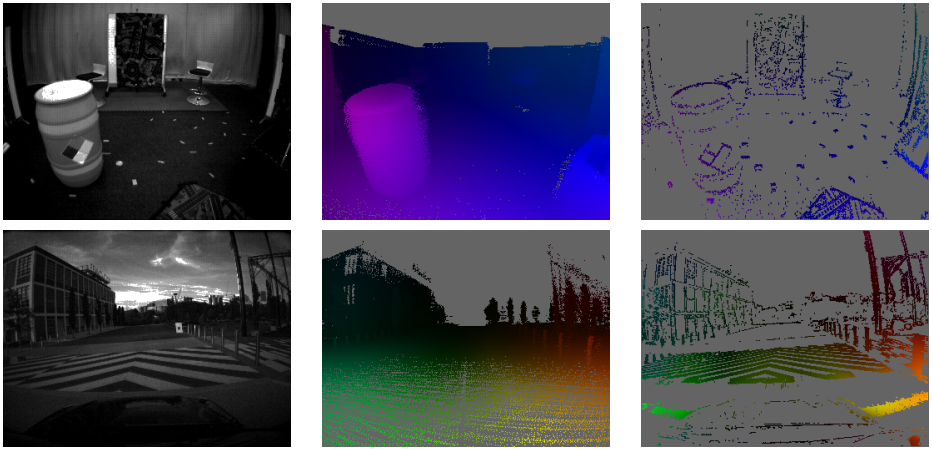
- Proposed by [Stoffregen et al. 2020]:

$$\text{FWL} = \frac{\sigma^2(I_{\text{comp}})}{\sigma^2(I_{\text{uncomp}})}$$

- Evaluates the compensation of motion
- Favors event collapse

Optical flow results on the MVSEC dataset

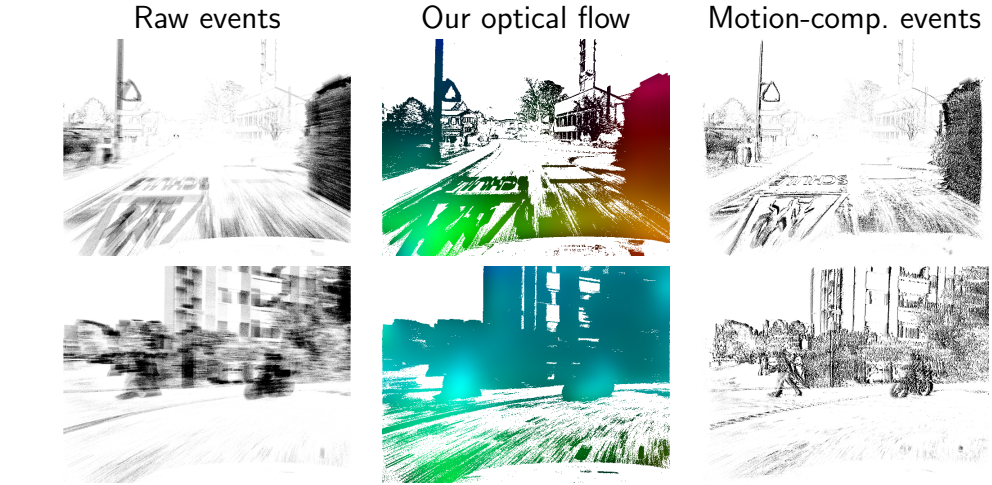
Reference image | Ground truth | Our optical flow



Quantitative comparison of the distance transforms (MVSEC)

	Indoor flying 1		Indoor flying 2		Indoor flying 3		Outdoor day 1		Outdoor day 2		Outdoor night 1		Outdoor night 2		Outdoor night 3	
	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓	AEE ↓	% outl. ↓
RTEF	**0.52**	**0.1**	**0.98**	**5.5**	**0.71**	2.1	**0.53**	**0.2**	**0.74**	**1.2**	**2.91**	**30.6**	**3.45**	**39.1**	**3.62**	**39.8**
$RTEF_{DS,L}$	1.81	16.4	2.54	26.4	1.95	18.2	2.12	21.7	1.30	8.7	4.04	45.8	4.78	55.6	5.10	58.7
$RTEF_{DS,LB}$	0.62	0.3	1.02	5.6	0.79	**1.8**	0.64	0.5	0.79	1.3	3.05	32.5	3.68	41.8	3.90	43.4
$RTEF_{DS,Log}$	0.70	1.4	1.07	6.5	0.82	2.4	0.69	1.6	0.79	1.9	3.08	33.0	3.70	42.1	3.93	43.8

Optical flow results on the DSEC dataset (640×480)

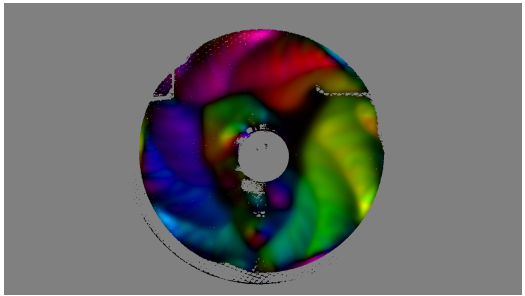


FWL results on the 20-minute-long driving sequence

Village (0'00 - 4'00)	Side Road (4'00 - 7'00)	Highway (7'00 - 11'00)	Suburban (11'00 - 14'30)	Urban (14'30 - 20'45)	*Full sequence* (0'00 - 20'45)
1.70	1.45	1.67	1.62	1.43	1.56

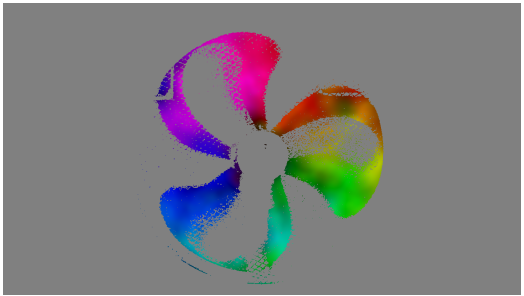
Optical flow results for the "Fan" sequence

$\Delta t = 15$ms



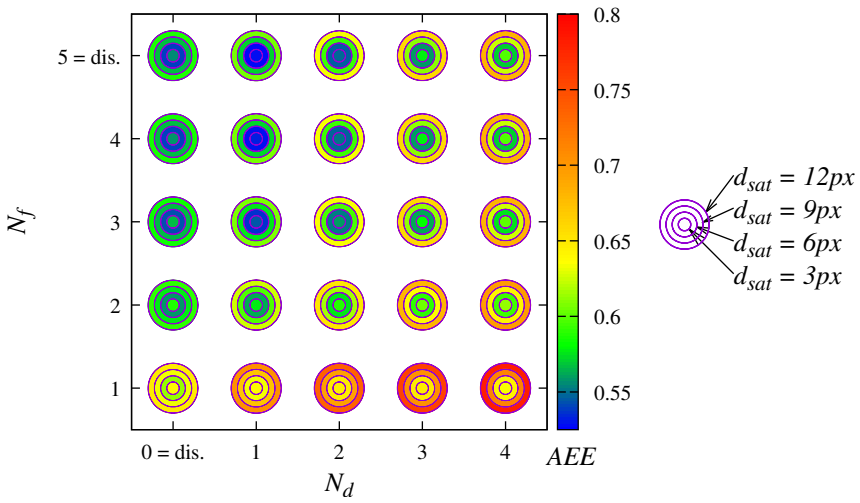
FWL $= 1.05$

$\Delta t = 5$ms



FWL $= 1.53$

Sensitivity analysis of the parameters for the optical flow



Detailed optical flow computation times

	Edge image (after accumulation)*	Denoising & filling	Inv. exp. distance transform	Optical flow†	Total
Low-resolution (346 × 260)					
CPU-only	0.07±0.02	0.76±0.16	1.62±0.85	*2.81±0.08*	5.27±0.93
CPU & GPU	0.15±0.05	0.42±0.02	1.29±0.08	3.46±0.12	5.31±0.20
Mid-resolution (640 × 480)					
CPU-only	0.82±1.25	0.30±0.03	13.00±7.46	*16.41±3.27*	30.59±9.31
CPU & GPU	0.42±0.16	0.33±0.57	1.15±0.79	15.38±0.94	17.28±2.02
High-resolution (1280 × 720)					
CPU-only	0.64±0.43	5.00±0.63	22.54±3.03	*12.77±0.62*	40.96±3.27
CPU & GPU	0.47±0.04	0.66±0.06	2.69±0.45	10.41±0.85	14.21±1.22

*CPU-only

†GPU-only (due to the use of the library of [Adarve et al. 2016])

Extending the optical flow

Classification of events based on the depth difference

Colors:

- $d_{\text{af}} - d_{\text{bf}} < -1\text{m}$
- $d_{\text{af}} - d_{\text{bf}} \in [-1\text{m}, +1\text{m}]$
- $d_{\text{af}} - d_{\text{bf}} > +1\text{m}$

Losses for ALED and DELTA

$$\mathcal{L}_{\text{pw}} = \sum_{\mathbf{x}} \left\| D(\mathbf{x}) - \hat{D}(\mathbf{x}) \right\|_1$$

$$\mathcal{L}_{\text{msg}} = \sum_{h \in \{1,2,4,8,16\}} \sum_{\mathbf{x}} \left\| \mathbf{g}[D](\mathbf{x}, h) - \mathbf{g}[\hat{D}](\mathbf{x}, h) \right\|_2$$

$$\mathbf{g}[f](x, y, h) = \begin{pmatrix} f(x+h, y) - f(x, y) \\ f(x, y+h) - f(x, y) \end{pmatrix}$$

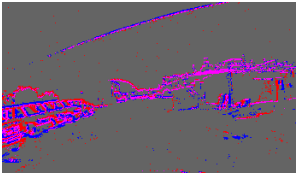
$$\mathcal{L} = \sum_{t=1}^{T} \sum_{\text{bf,af}} (\mathcal{L}^t_{\text{pw}} + \alpha \mathcal{L}^t_{\text{msg}})$$

Content of our SLED dataset

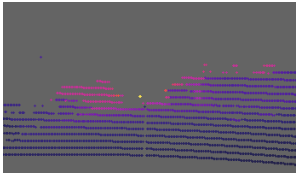
	Set	Environment	Features	Night seq.	Day seq.
Town01	Test	Town	Small buildings, bridges, distant forests and mountains, palm trees	4	16
Town02	Train	Town	Small buildings, plazas, forest road	4	16
Town03	Test	City	Tall and small buildings, roundabouts, tunnel, aerial railway	4	16
Town04	Val.	Town	Small buildings, highway, parking, lake, forests and mountains	4	16
Town05	Train	City	Tall buildings, parking, aerial beltway and railway	4	16
Town06	Train	Suburban	Small buildings, U-turns, distant hills	4	16
Town07	Train	Countryside	Barns, grain silos, fields, mountain road	4	16
Town10	Train	City	Buildings, monuments and sculptures, playgrounds, seaside	4	16

ALED results on the MVSEC dataset

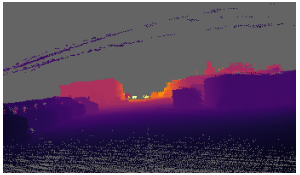
(a) Grayscale image



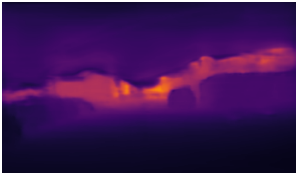
(b) Events



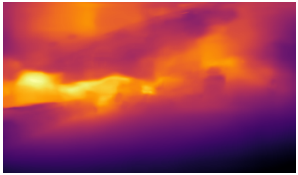
(c) LiDAR



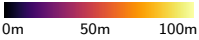
(d) Ground truth



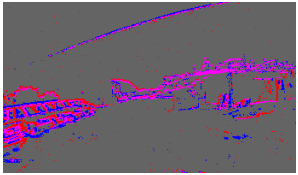
(e) $\text{ALED}_{\text{SLED}\rightarrow\text{MVSEC}}$



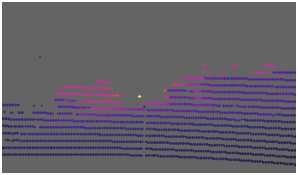
(f) RAMNet [Gehrig et al. 2021a]



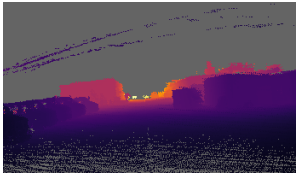
ALED results on the MVSEC dataset



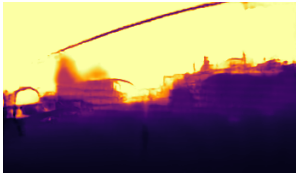
(a) Events



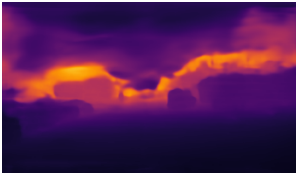
(b) LiDAR



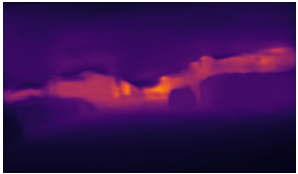
(c) Ground truth



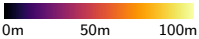
(d) $\text{ALED}_{\text{SLED}}$



(e) $\text{ALED}_{\text{MVSEC}}$

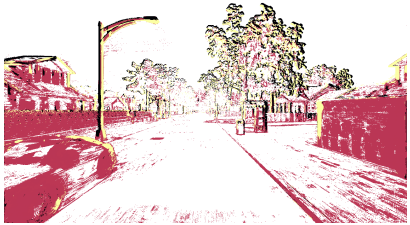


(f) $\text{ALED}_{\text{SLED}\rightarrow\text{MVSEC}}$



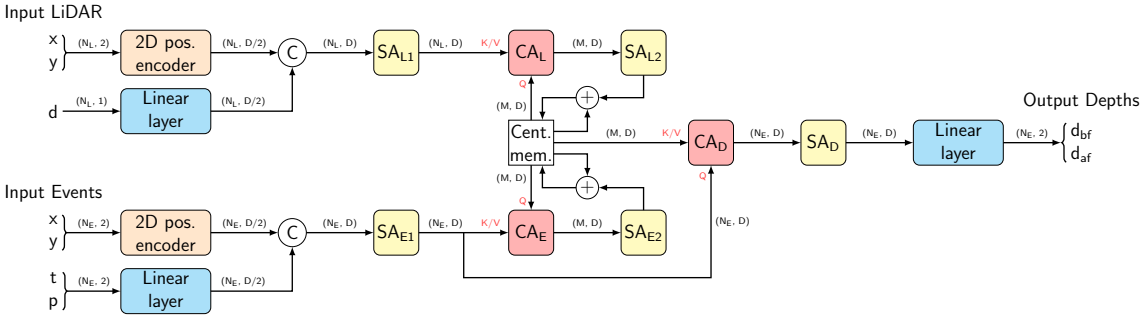
Depth change maps predicted by ALED

Predicted depth change map

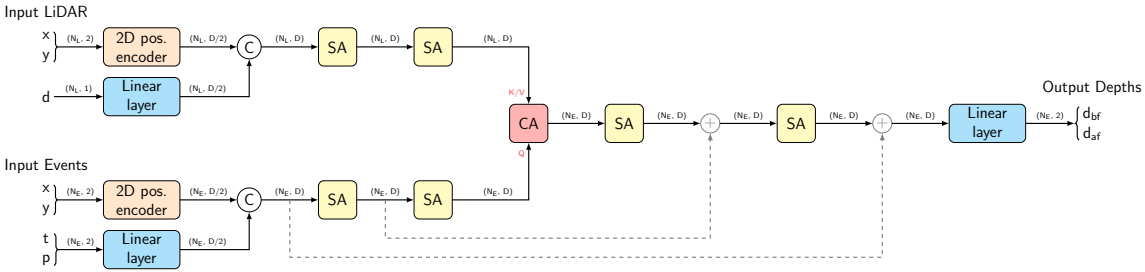


Ground truth

Sparse attention-based network (v1)



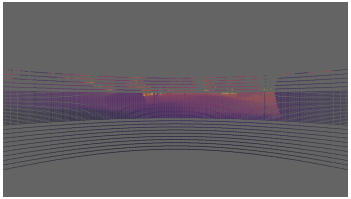
Sparse attention-based network (v2)



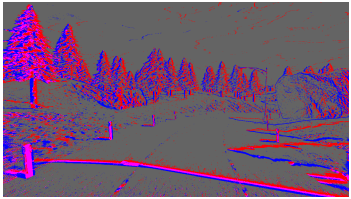
Issues with sparse attention-based networks

- Memory usage
- Granularity of data
- Lack of structure
- Nearest-neighbor-like behavior

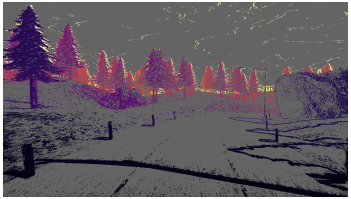
Nearest-neighbor-like behavior



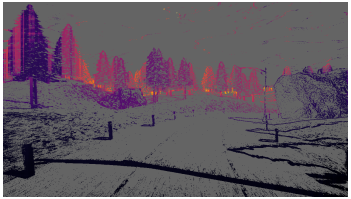
(a) LiDAR projection



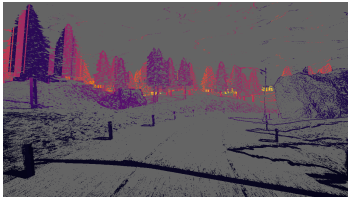
(b) Events



(c) Ground truth



(d) Prediction of the network



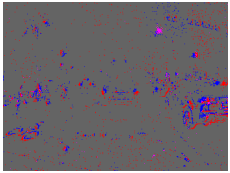
(e) Prediction of a NN method

Quantitative comparison between ALED and DELTA on SLED

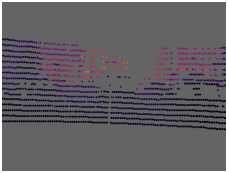
		Cutoff	Dense depths errors				Sparse depths errors				Depth change map errors	
			On D_{bf}		On D_{af}		On D_{bf}		On D_{af}		Abs. (m)	Correctly classified events (%)
			Abs. (m)	Rel. (%)	Abs. (m)	Rel. (%)	NN (m)	$ALED_{SL}$ (m)	NN (m)	$ALED_{SL}$ (m)		(with a threshold of ±1m)
$ALED_{SL}$	Town01	10m	1.24	20.99	1.37	23.60	1.32	1.46	2.24	1.79	**2.11**	90.27
		20m	2.08	23.06	2.27	25.48	1.51	1.84	2.53	2.15	3.18	85.07
		30m	2.72	23.76	2.92	26.03	1.71	2.37	2.83	2.67	**3.88**	81.68
		100m	4.25	24.01	4.51	26.07	2.40	3.48	3.91	3.95	**5.12**	77.48
		200m	**4.53**	17.20	**4.81**	18.66	7.86	**5.44**	9.76	**6.23**	**7.36**	75.54
	Town03	10m	2.00	28.91	2.09	30.11	0.47	0.56	0.67	0.66	1.14	93.70
		20m	2.85	29.91	2.97	31.15	0.64	0.75	1.12	0.87	2.54	87.16
		30m	3.33	29.10	3.45	30.24	0.92	1.11	1.61	1.26	3.23	83.71
		100m	4.60	27.37	4.77	28.42	1.88	2.55	3.17	2.88	4.47	78.50
		200m	4.86	21.50	5.03	22.33	4.43	3.60	5.93	4.10	**6.20**	77.23

		Cutoff	Dense depths errors				Sparse depths errors				Depth change map errors	
			On D_{bf}		On D_{af}		On D_{bf}		On D_{af}		Abs. (m)	Correctly classified events (%)
			Abs. (m)	Rel. (%)	Abs. (m)	Rel. (%)	NN (m)	$DELTA_{SL}$ (m)	NN (m)	$DELTA_{SL}$ (m)		(with a threshold of ±1m)
$DELTA_{SL}$	Town01	10m	**0.64**	**10.11**	**0.67**	**10.55**	1.32	**1.14**	2.24	**1.25**	2.19	**91.81**
		20m	**1.45**	**13.84**	**1.50**	**14.38**	1.51	**1.62**	2.53	**1.74**	**3.17**	**87.81**
		30m	**2.11**	**15.52**	**2.17**	**16.06**	1.71	**2.03**	2.83	**2.15**	**3.88**	**84.45**
		100m	**3.80**	**17.36**	**3.89**	**17.88**	2.40	**3.05**	3.91	**3.24**	5.14	**79.86**
		200m	5.37	**13.27**	5.42	**13.63**	7.86	6.04	9.76	6.24	7.94	**77.47**
	Town03	10m	**0.49**	**8.09**	**0.50**	**8.17**	0.47	**0.36**	0.56	**0.40**	**0.76**	**97.35**
		20m	**1.15**	**10.70**	**1.18**	**10.95**	0.64	**0.58**	1.12	**0.64**	**2.31**	**91.84**
		30m	**1.72**	**12.12**	**1.77**	**12.42**	0.92	**0.90**	1.61	**0.98**	**3.06**	**88.35**
		100m	**3.12**	**13.33**	**3.18**	**13.64**	1.88	**2.25**	3.17	**2.38**	**4.30**	**82.88**
		200m	**4.81**	**11.42**	**4.81**	**11.62**	4.43	**3.54**	5.93	**3.72**	6.69	**81.18**

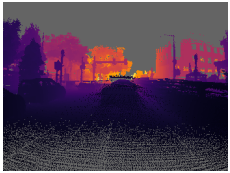
Qualitative comparison between ALED and DELTA on MVSEC



(a) Events

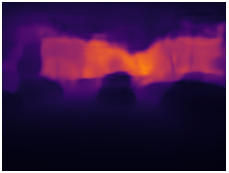


(b) LiDAR



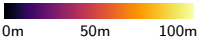
(c) Ground truth

(d) $\text{ALED}_{\text{SLED}\rightarrow\text{MVSEC}}$

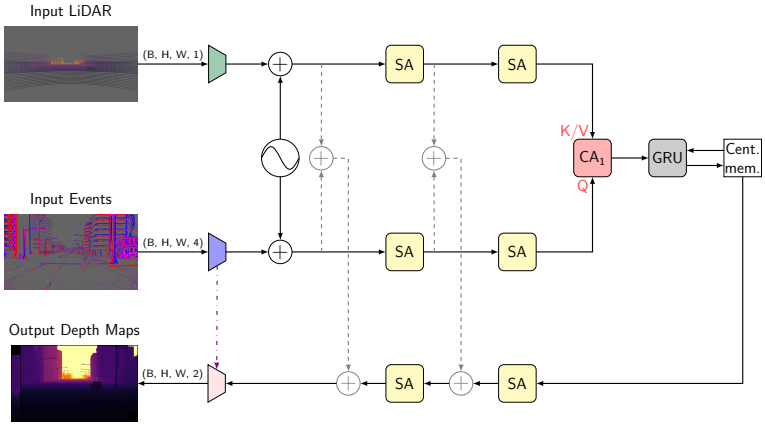


(e) $\text{DELTA}_{\text{SLED}\rightarrow\text{MVSEC}}$

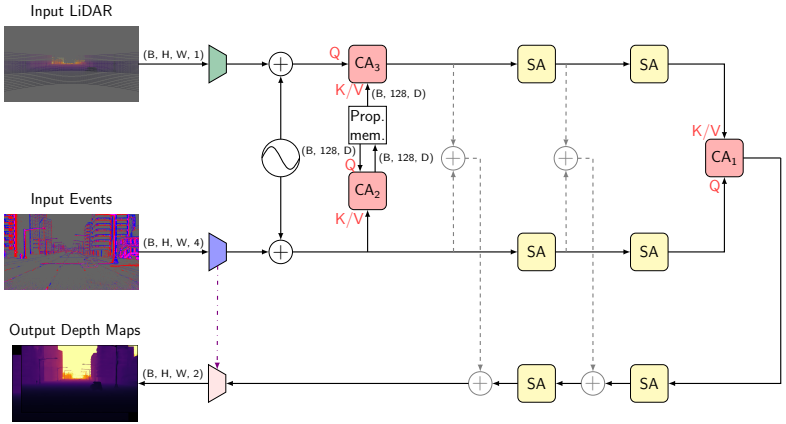
(f) 3D method of [Cui et al. 2022]



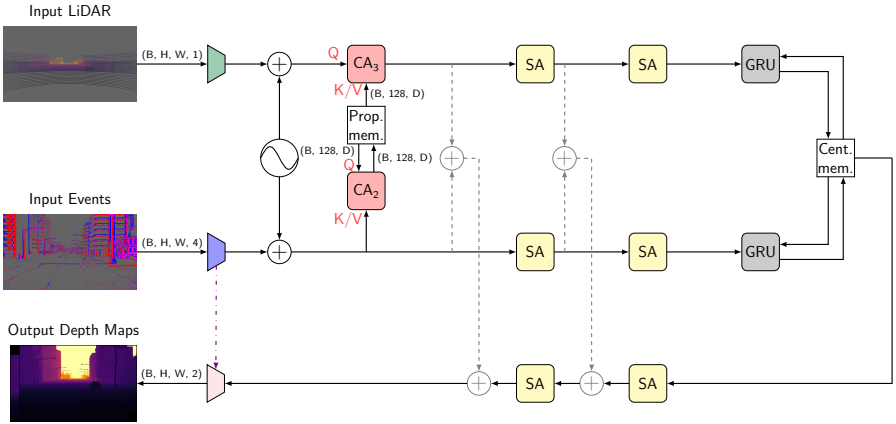
No Propagation Memory (DELTA^{NPM})



No Central Memory (DELTA^{NCM})



No Cross Attention (DELTA^{NCA})



Evaluation of alternative versions of DELTA on SLED

Cutoff	DELTA		DELTA^{NPM}		DELTA^{NCM}		DELTA^{NCA}	
	D_{bf}	D_{af}	D_{bf} (rel.)	D_{af} (rel.)	D_{bf} (rel.)	D_{af} (rel.)	D_{bf} (rel.)	D_{af} (rel.)
10m	**0.57**	**0.58**	0.85 (+0.28)	0.86 (+0.28)	0.91 (+0.34)	0.91 (+0.33)	0.95 (+0.38)	0.94 (+0.36)
20m	**1.29**	**1.33**	1.64 (+0.35)	1.66 (+0.33)	1.71 (+0.42)	1.70 (+0.37)	1.80 (+0.51)	1.79 (+0.46)
30m	**1.91**	**1.96**	2.24 (+0.33)	2.28 (+0.32)	2.21 (+0.30)	2.22 (+0.26)	2.40 (+0.49)	2.41 (+0.45)
100m	**3.44**	**3.52**	3.73 (+0.29)	3.78 (+0.26)	3.53 (+0.09)	3.56 (+0.04)	3.85 (+0.41)	3.89 (+0.37)
200m	5.09	5.11	5.01 (-0.08)	5.02 (-0.09)	**4.49** (-0.60)	**4.52** (-0.59)	5.03 (-0.06)	5.10 (-0.01)

Table: Average absolute and relative depth errors (in meters)

References I

Adarve, J. and R. Mahony (2016). "A Filter Formulation for Computing Real Time Optical Flow". In: *IEEE Robotics and Automation Letters (RA-L)* 1, pp. 1192–1199.

Almatrafi, M. et al. (2020). "Distance Surface for Event-Based Optical Flow". In: *IEEE TPAMI* 42, pp. 1547–1556.

Benosman, R. et al. (2014). "Event-Based Visual Flow". In: *IEEE Transactions on Neural Networks and Learning Systems* 25, pp. 407–417.

Chaney, Kenneth et al. (2023). "M3ED: Multi-Robot, Multi-Sensor, Multi-Environment Event Dataset". In: pp. 4016–4023.

Cui, Mingyue et al. (2022). "Dense Depth-Map Estimation Based on Fusion of Event Camera and Sparse LiDAR". In: *IEEE Transactions on Instrumentation and Measurement* 71, pp. 1–11.

Dosovitskiy, Alexey et al. (2017). "CARLA: An Open Urban Driving Simulator". In: *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16.

Dosovitskiy, Alexey et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*.

Gallego, Guillermo, Henri Rebecq, and D. Scaramuzza (2018). "A Unifying Contrast Maximization Framework for Event Cameras, with Applications to Motion, Depth, and Optical Flow Estimation". In: *CVPR*, pp. 3867–3876.

Gehrig, Daniel et al. (2021a). "Combining Events and Frames Using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction". In: *IEEE Robotics and Automation Letters* 6, pp. 2822–2829.

Gehrig, Mathias et al. (2021b). "DSEC: A Stereo Event Camera Dataset for Driving Scenarios". In: *IEEE Robotics and Automation Letters* 6, pp. 4947–4954.

Gehrig, Mathias et al. (2021c). "E-RAFT: Dense Optical Flow from Event Cameras". In: *International Conference on 3D Vision (3DV)*.

Horn, Berthold K. P. and Brian G. Schunck (1981). "Determining Optical Flow". In: *Artificial Intelligence* 17, pp. 185–203.

Li, Boyang et al. (2021). "Enhancing 3-D LiDAR Point Clouds With Event-Based Camera". In: *IEEE Transactions on Instrumentation and Measurement* 70, pp. 1–12.

Liu, Haotian et al. (2023). "TMA: Temporal Motion Aggregation for Event-based Optical Flow". In: *ICCV*.

Lucas, Bruce D. and Takeo Kanade (1981). "An Iterative Image Registration Technique with an Application to Stereo Vision". In: *International Joint Conference on Artificial Intelligence*.

Nadaraya, Elizbar (1964). "On Estimating Regression". In: *Theory of Probability and Its Applications* 9, pp. 141–142.

Nagata, Jun, Yusuke Sekikawa, and Y. Aoki (2021). "Optical Flow Estimation by Matching Time Surface with Event-Based Cameras". In: *Sensors* 21.

Paredes-Vallés, F. and G. D. Croon (2021). "Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3446–3455.

References II

Rançon, Ulysse et al. (2021). "StereoSpike: Depth Learning With a Spiking Neural Network". In: *IEEE Access* 10, pp. 127428–127439.

Sabater, Alberto, Luis Montesano, and Ana Cristina Murillo (2022). "Event Transformer+. A multi-purpose solution for efficient event data processing". In: *ArXiv* abs/2211.12222.

Shiba, Shintaro, Yoshimitsu Aoki, and Guillermo Gallego (2022). "Secrets of Event-Based Optical Flow". In: *European Conference on Computer Vision*.

Stoffregen, Timo et al. (2020). "Reducing the Sim-to-Real Gap for Event Cameras". In: *ECCV*.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Neural Information Processing Systems*.

Watson, Geoffrey Stuart (1964). "Smooth regression analysis". In: *Sankhyā: The Indian Journal of Statistics*. Vol. 26. 4, pp. 359–372.

Wikimedia, Commons (2021). *File:Perseverance rover on Mars.jpg — Wikimedia Commons, the free media repository*. URL: https://commons.wikimedia.org/w/index.php?title=File:Perseverance_rover_on_Mars.jpg (visited on 12/05/2023).

— (2022). *File:Xiaomi robot vacuum.jpg — Wikimedia Commons, the free media repository*. URL: https://commons.wikimedia.org/w/index.php?title=File:Xiaomi_robot_vacuum.jpg (visited on 12/05/2023).

Xu, Haofei et al. (2022). "Unifying Flow, Stereo and Depth Estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, pp. 13941–13958.

Zhu, A. Z. et al. (2018a). "EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras". In: *Proceedings of Robotics: Science and Systems*.

Zhu, A. Z. et al. (2018b). "The Multivehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception". In: *IEEE Robotics and Automation Letters (RA-L)* 3, pp. 2032–2039.