DERD-Net: Learning Depth from Event-based Ray Densities

Diego Hitzges*, Suman Ghosh*, Guillermo Gallego

Technische Universität Berlin · science of intelligence · Robotics Institute Germany · Einstein Center Digital Future

Neural Information Processing Systems Spotlight

Summary

Motivation:

Learning from events is challenging as their asynchronous and continuous data is inherently incompatible with conventional deep-learning approaches.

Contribution:

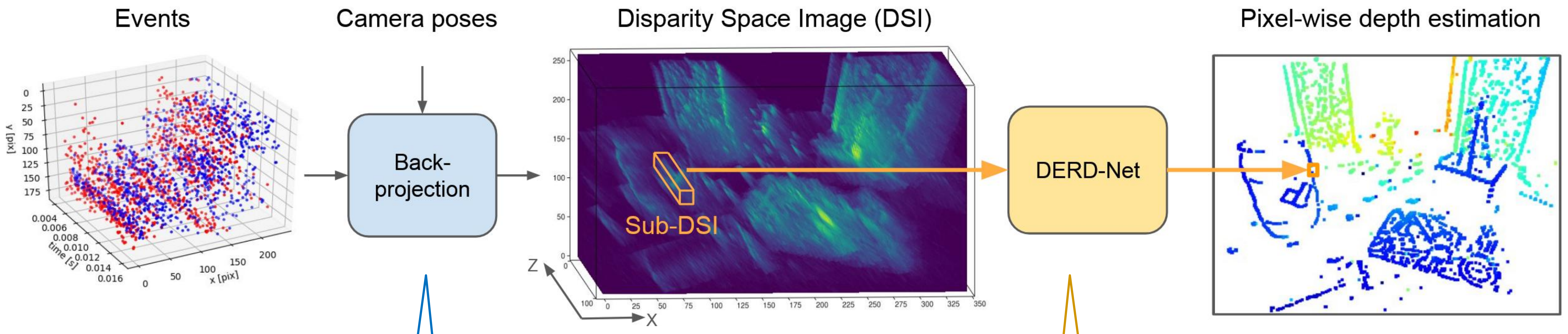
We propose the **1st event-based deep multi-view stereo method for 3D reconstruction**.

Large performance improvement over SOTA methods on MVSEC and DSEC benchmarks.

Approach:

Transform event streams into a 3D-geometric representation and feed small local subregions of this representation to a compact artificial neural network.

Methodology

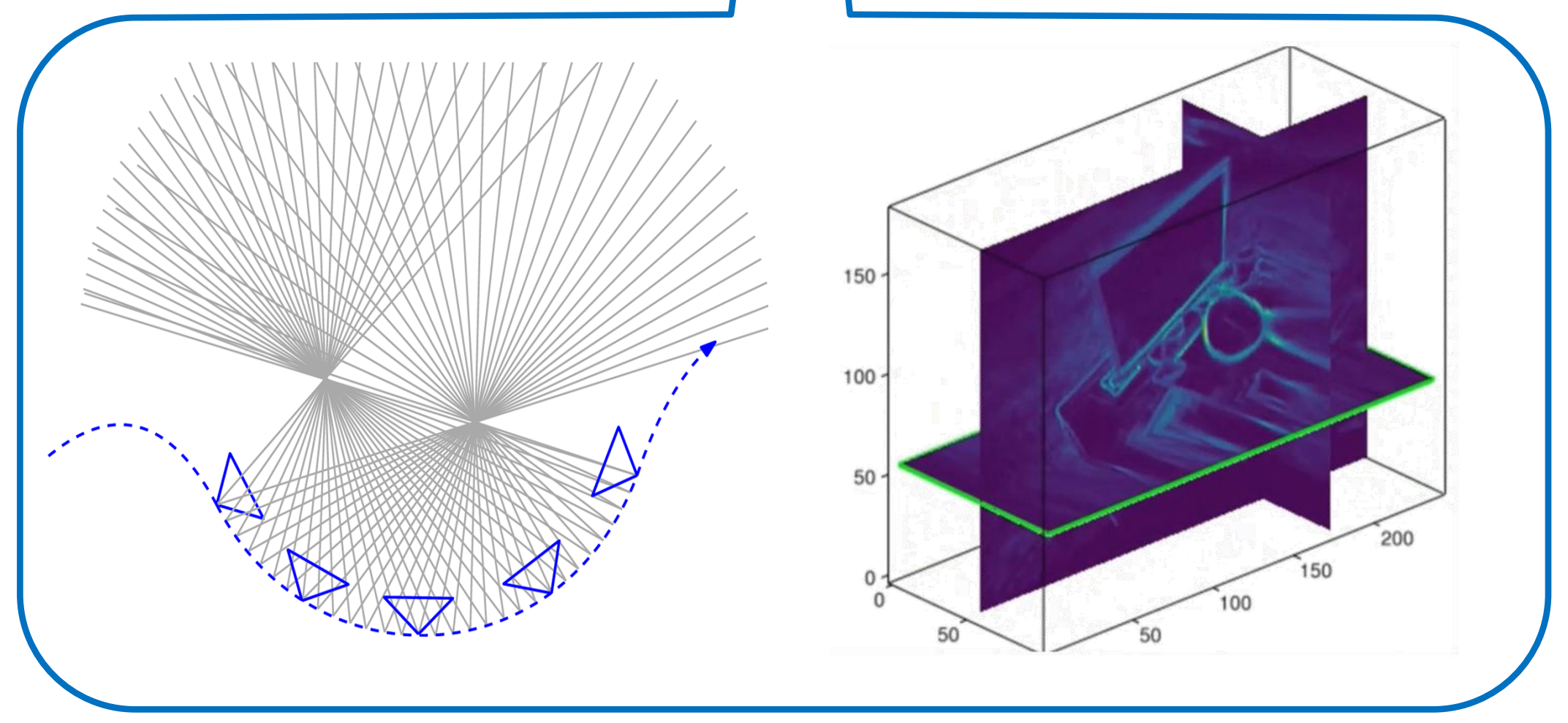


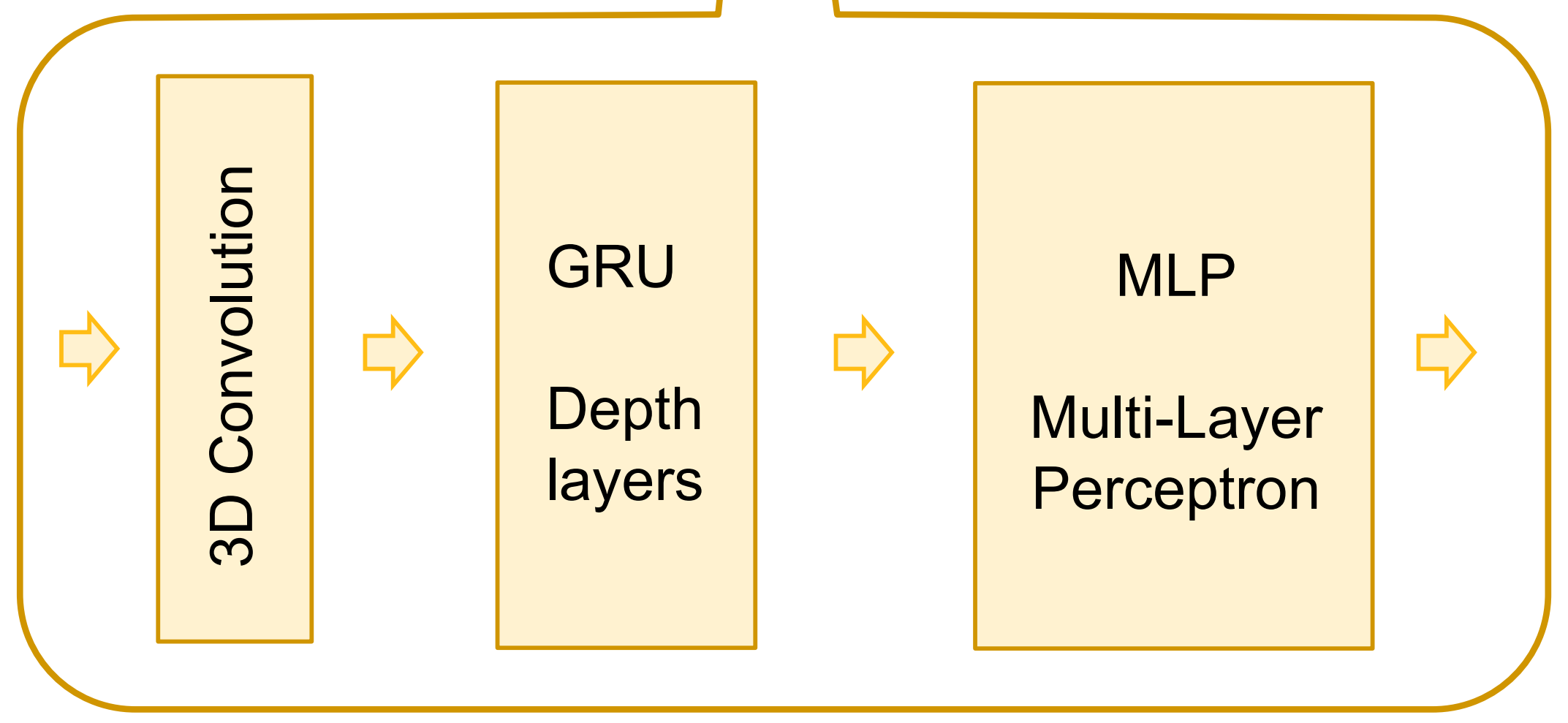
Geometric Event Representation: Events are converted into DSIs using known camera motion.

Local Processing: Small local subregions (Sub-DSIs) around selected pixels are extracted and processed individually by a neural network.

Depth Map: Pixel-wise estimation produces SLAM-compatible semi-dense depth maps.

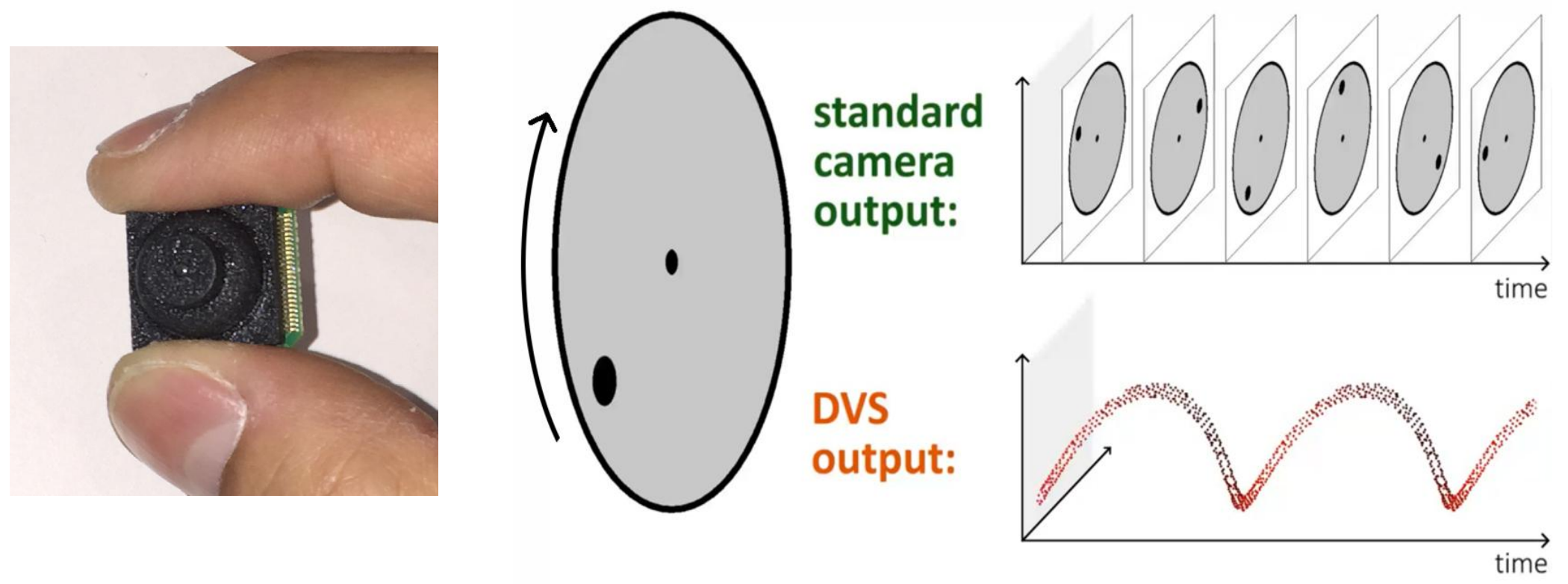
Ray Density



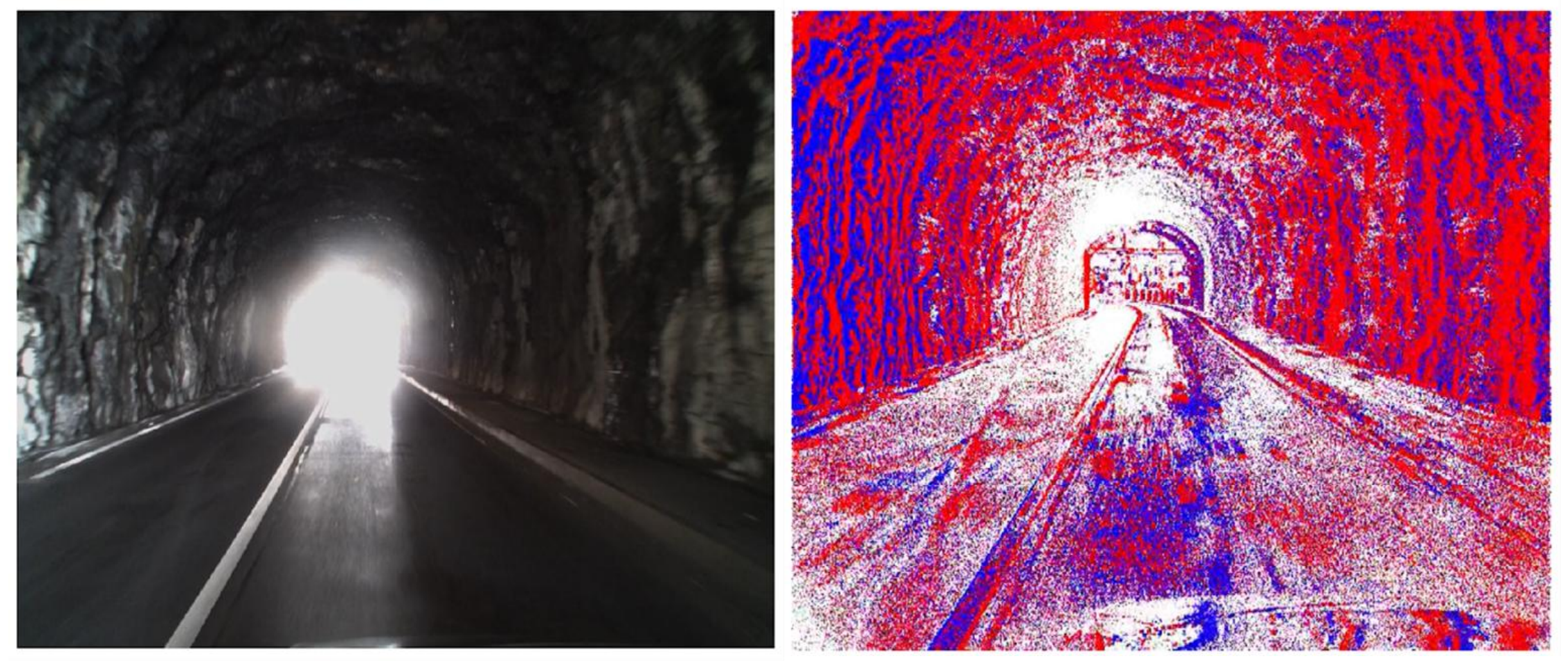


Network Architecture

What is an Event Camera?



- Only transmits **brightness changes**.
- Output is a stream of **asynchronous events**.
- **Advantages:** low latency, no motion blur, very high dynamic range (HDR), low power.



Frames Events

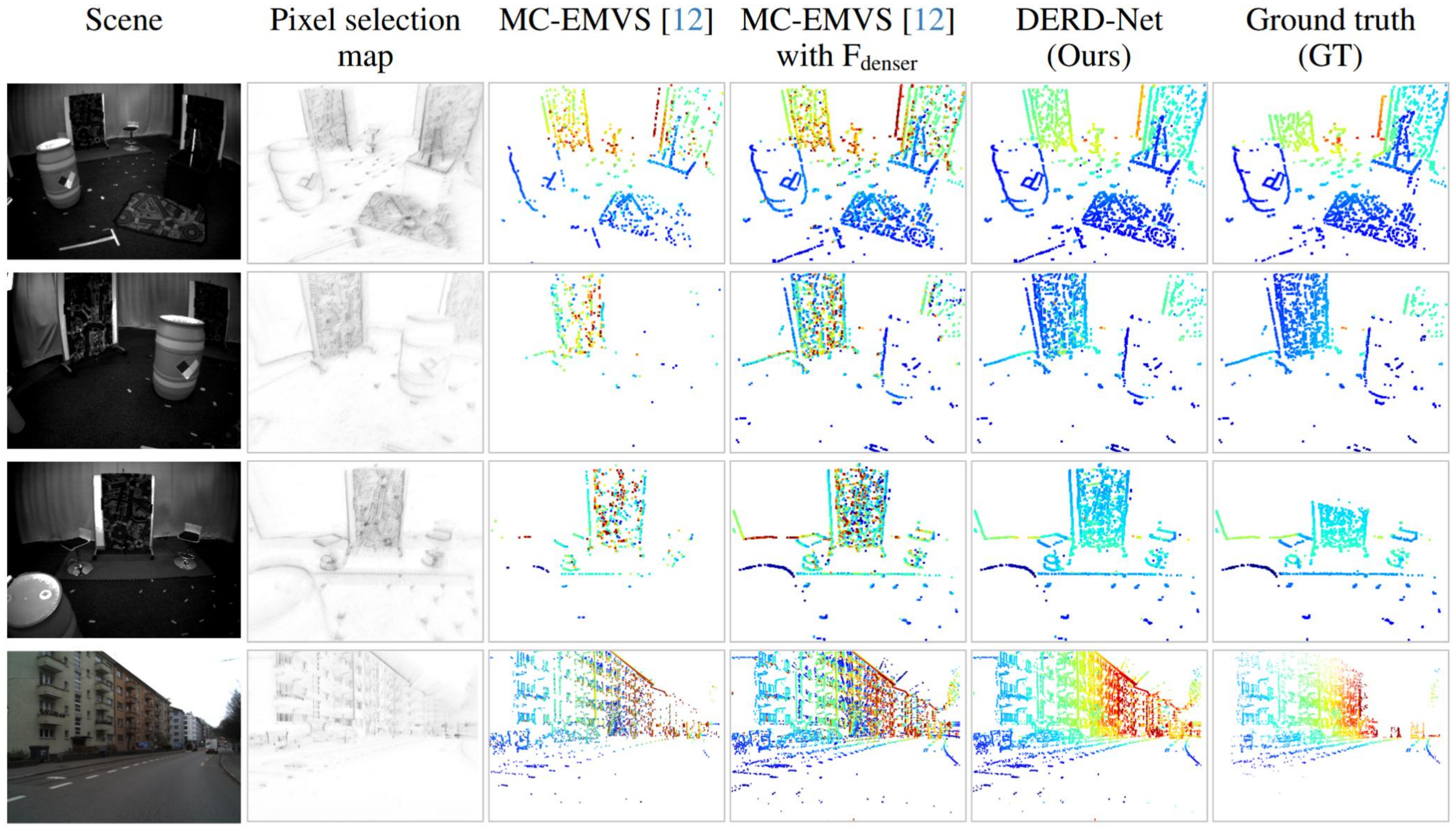
References

- MC-EMVS: Ghosh et al. *Multi-Event-Camera Depth Estimation and Outlier Rejection by Refocused Events Fusion*, Adv. Intell. Syst. 2022.
- Zhou et al, ESVO: Event-based Stereo Visual Odometry, IEEE T-RO, 2021.
- Ghosh et al., Event-based Stereo Depth Estimation: A Survey, IEEE T-PAMI 2025.

Design Properties

- **Scalability:** Independent of pixel resolution and DSI depth.
- **Flexibility**: Supports both monocular and stereo without requiring event simultaneity.
- **Implicit Data Augmentation:** Drastic increase in effective dataset size due to pixel-wise approach.
- **Ultra-leightweight:** Network has ~70k parameters and is <1 MB size.
- **Efficiency:** Leverages event-data sparsity and enables full parallel processing of inputs.
- **Robustness**: Processing only small local subregions reduces overfitting to specific scenes.

Depth Estimation Results



	Method		MVSEC				DSEC			
	Algorithm	Modality	Mean Err [cm] ↓	Median Err [cm] ↓	bad-pix [%] ↓	#Points [million]↑	Mean Err [m] ↓	Median Err [m] ↓	bad-pix [%] ↓	#Points [million]↑
SOTA	EMVS	monocular	33.78	14.35	3.84	1.27	5.64	2.52	13.68	1.31
	ESVO	stereo	22.70	9.83	2.83	1.56	3.93	1.62	10.54	9.40
	MC-EMVS	stereo	20.07	9.53	1.35	0.81	3.27	0.90	10.75	1.25
Ours	DERD-Net	monocular + F_{orig}	23.68	11.55	2.78	1.21	3.12	1.60	5.50	2.10
	DERD-Net	stereo + F_{orig}	**11.69**	**5.50**	**0.89**	0.79	**1.61**	**0.46**	**4.12**	1.67
	DERD-Net	stereo + F_{denser}	15.24	6.68	1.70	2.77	1.80	0.54	5.04	4.64
	DERD-Net (multi-pixel)	stereo + F_{denser}	15.68	6.73	1.74	**11.33**	1.79	0.54	4.61	**14.74**

- **stereo:** strong improvement over SOTA, mean absolute error reduced by at least 42%
- **monocular:** comparable to SOTA *stereo.*
- **depth completeness:** more than 3-fold while still reducing median absolute error by at least 30%
- **sensitivity:** robust to noise in camera poses.