MouseSIS: A Frames-and-Events Dataset for Space-Time Instance Segmentation of Mice

Friedhelm Hamann, Hanxiong Li, Paul Mieske,
Lars Lewejohann, and Guillermo Gallego

Summary

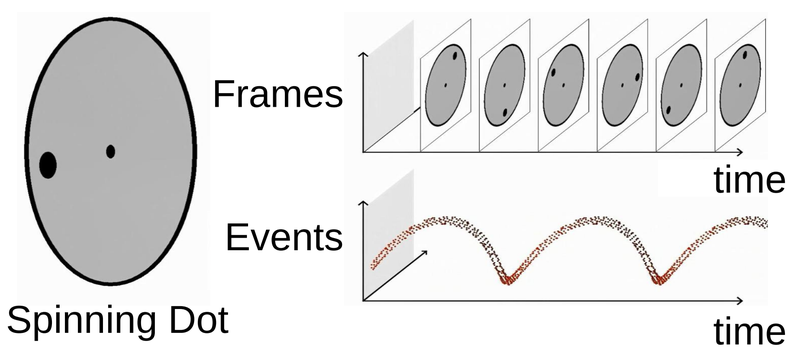
- **Input**: events and/or frames.
- **Output**: pixel-level masks for the mice class with consistent IDs.
- We term the task **Space-time Instance Segmentation (SIS),** analogous to Video Instance Segmentation but extended to time-continuous events.

Contributions

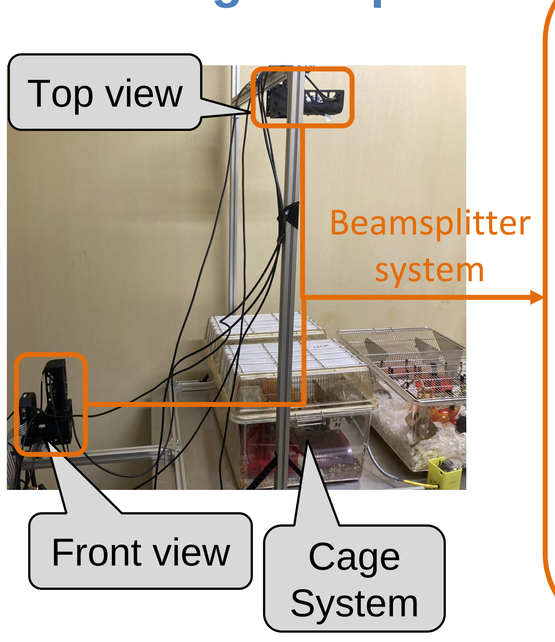
- **SIS Dataset** with pixel-level aligned frames and events from 2 views (~640s of annotated data with 157 spatio-temporal instances yielding a total of 75000 binary masks of top view)
- **Two methods** for our SIS task: A tracking-by-detection-based method, and an end-to-end learned transformer-based model.
- Extensive **evaluation** of our dataset with the two introduced methods

Event Cameras

- Output brightness changes **asynchronously** instead of frames.
- **Advantages:** low-power consumption, high dynamic range, high temporal resolution.

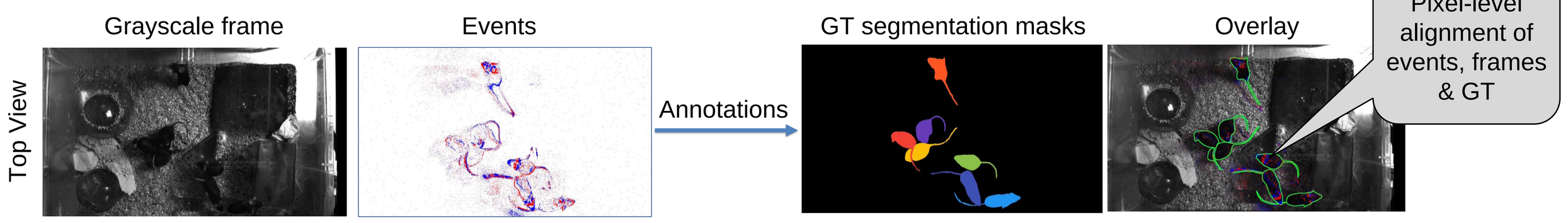


Recording Setup



- Complementary data from both **top** and **front** camera views.
- **Beamsplitter system** ensuring precise alignment and synchronization of both optical axes and timestamps.
- Two connected **type 4 cage systems**

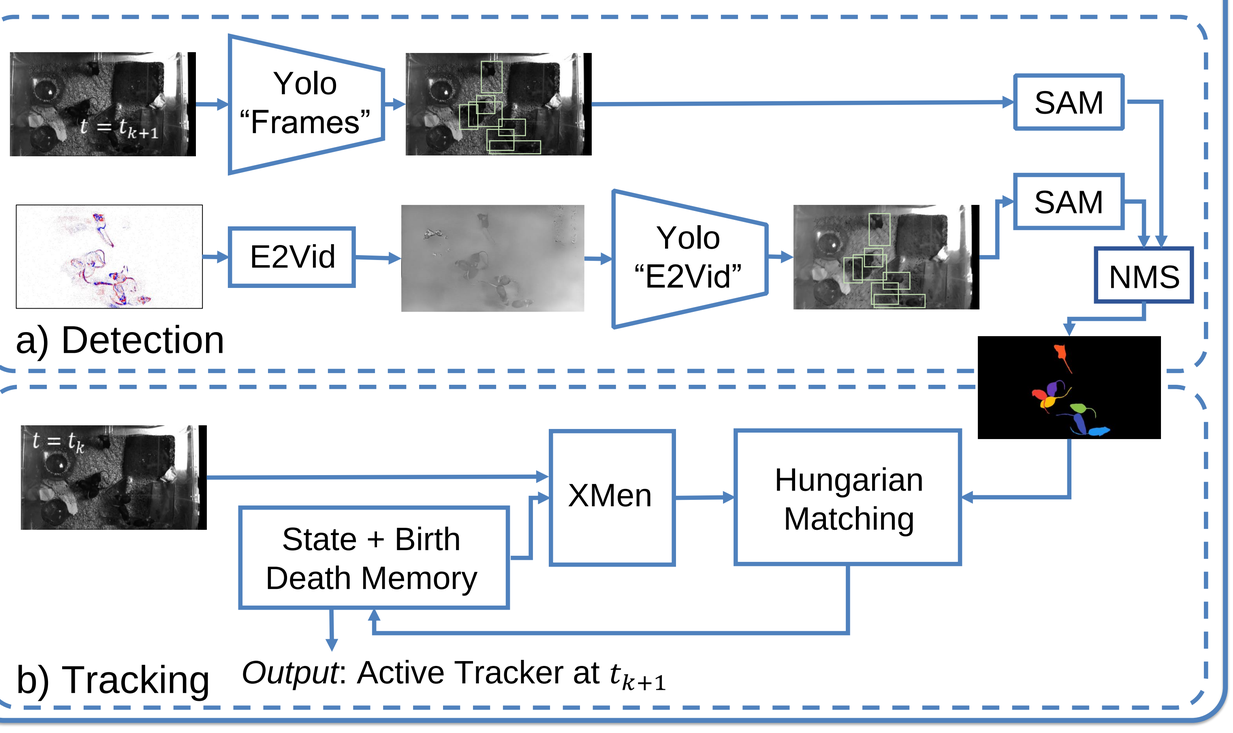
MouseSIS Dataset



- A total of 637 seconds of synchronized video and event data, divided into 33 x 20s sequences.
- Captured and calibrated from both top and front views.
- Pixel-level annotations for 637 seconds of mice, with 75,000 binary masks and over 12,000 labeled instances for the top view.

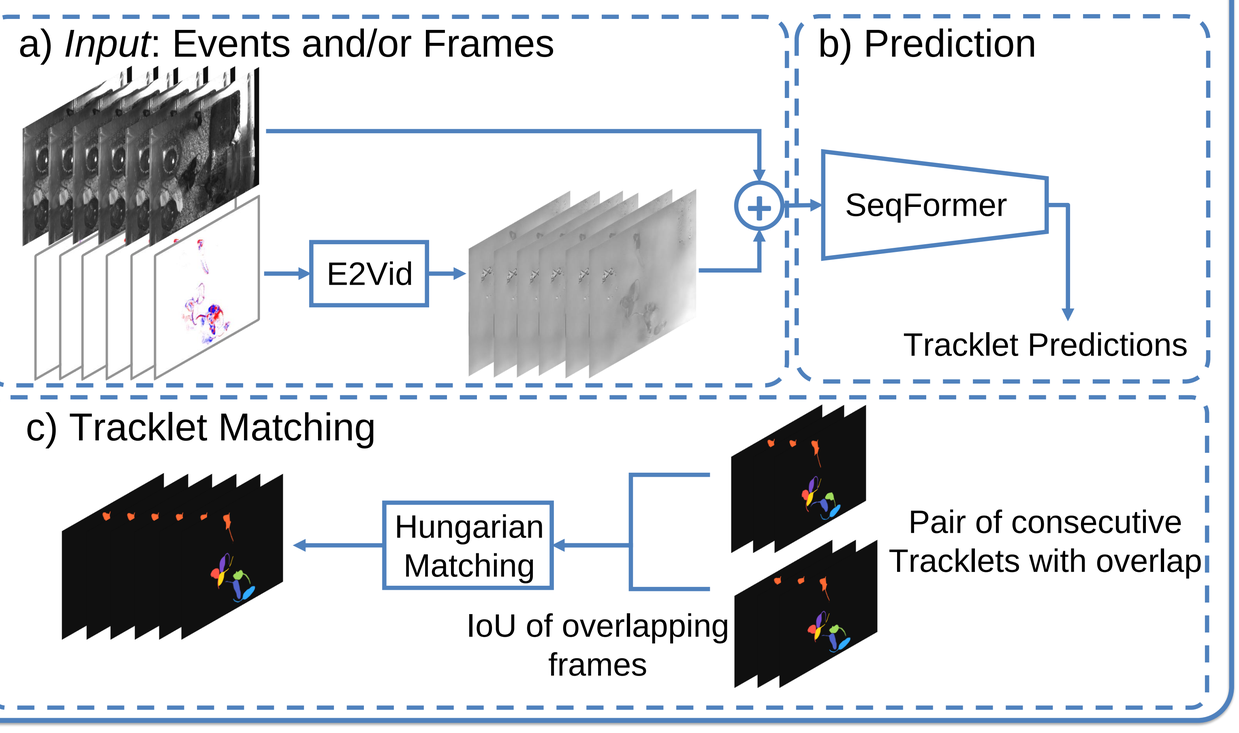
Method 1: ModelMixSort

Tracking-by-detection



Method 2: EventSeqFormer

End-to-end learned



Tracking Results

Method	Frames	Events	Mota↑	IDF1↑	HOTA↑	DetA↑	AssA↑
ModelMixSort	✓	✗	34.42	45.41	41.83	46.47	38.45
	✗	✓	32.13	40.06	33.68	33.58	34.07
	✓	✓	**54.94**	**65.17**	**54.19**	**53.69**	55.91
SeqFormer	✓	✗	40.22	61.42	53.07	47.57	**60.27**
EventSeqFormer (E2VID)	✓	✓	-16.34	34.82	30.52	24.26	38.58
	✗	✓	39.45	56.12	47.36	45.66	49.57
EventSeqFormer (Voxel)	✓	✓	40.72	60.14	47.82	44.23	52.41
	✗	✓	−67.98	24.91	23.14	16.49	32.63

MOTA: Multi-object-tracking accuracy; **IDF1**: Identification F1 score; **HOTA**: Higher Order Tracking Accuracy; **DetA**: Detection Accuracy; **AssA**: Association Accuracy

- Models using event data **consistently outperform** frame-only models, particularly in the ModelMixSort method.
- SORT-based methods excel in detection accuracy, while SeqFormer shows better association accuracy due to its ability to process longer context windows.
- E2VID models struggle with sequences recorded at low contrast settings, resulting in poor performance for event-only models in such cases.