Event-Driven Spacecraft Pose Estimation via CNN Direct Regression from Voxel Grid

Santhosh SRS, Computer Science
Mentor: Bharatesh Chakravarthi, Assistant Teaching Professor
School of Computing and Augmented Intelligence

Research Question

Can a deep learning model trained only on synthetic event-camera data accurately estimate the 6-DoF pose (position + orientation) of a spacecraft when tested on real-world recordings and what is the source of the synthetic-to-real performance gap?

Methodology

Event Representation:

- Collect all events in a 100 ms window, Split into 3 sub-windows (~33 ms each).
- w = polarity × exp(-(t_end - t) / 30ms).
- Accumulate into a 3-channel Voxel grid.

Network Architecture:

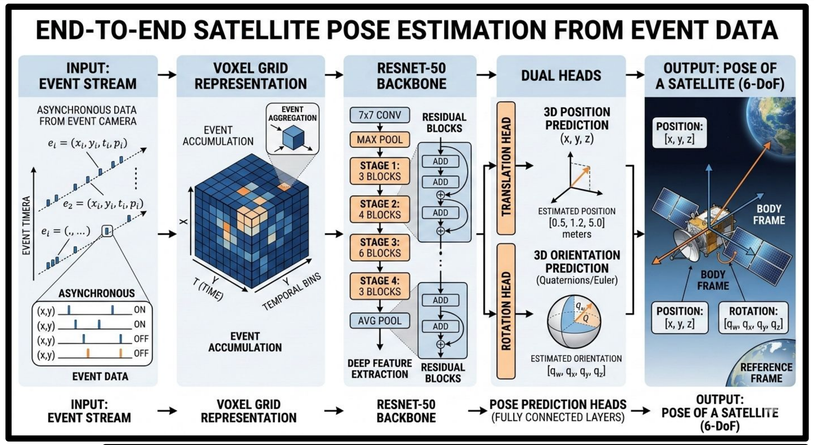
- ResNet-50 backbone, Global average pooling.
- Translation head:
  - Predicts [x, y, z] in metres.
- Rotation head:
  - unit quaternion [Qw, Qx, Qy, Qz].

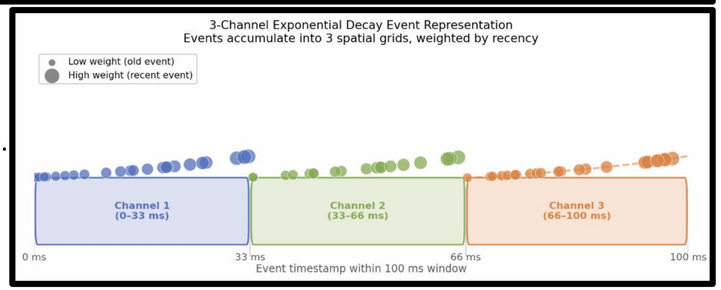
Geodesic Rotation Loss:

- $L = \lambda_t \cdot \mathrm{MSE}(\hat{t}, t) + \lambda_r \cdot 2\arccos(|\hat{q} \cdot q|)$.
- Final weights: $\lambda_t = 2.0$, $\lambda_r = 1.0$.

Two-Phase Training:

- Phase 1: 25% data, 20 epochs.
- Phase 2: Full dataset, 100 epochs.





Background & Motivation

What is an event camera?

- Event cameras report per-pixel brightness changes asynchronously at microsecond resolution.
- Benefits: no motion blur, high dynamic range, low power, and output: (x, y, polarity, timestamp).

Why spacecraft pose estimation?

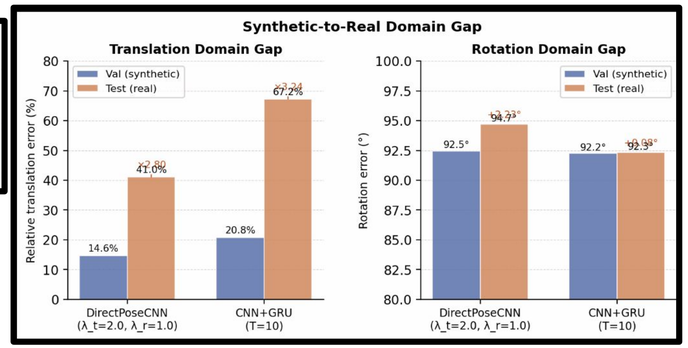
- Autonomous rendezvous and on-orbit servicing all require knowing a satellite's 6-DoF pose.
- Standard cameras fail under the extreme lighting condition and fast motion in space.

The SPADES Dataset (Rathinam et al., 2023)

- 300 synthetic training sequences + 31 real test sequences. 1280×720 Prophesee event camera.
- Labels: 3D position (metres) + unit quaternion orientation at 10 Hz.

Results & Analysis

Method	E_t	E_r (rad)	E
SPADES Direct [12]	0.0513	1.416	1.47
SPADES Hybrid [12]	0.0334	1.378	1.41
CNN+GRU, T=10 (ours)	0.6722	1.6113	2.2835
DirectPoseCNN, λ_t=2.0, λ_r=1.0 (ours)	0.4105	1.6526	**2.063**



Key findings:

- The source of domain shift is a 100× timestamp unit mismatch between synthetic (100 µs) and real (1 µs) data which compresses the voxel bin distributions at test time.
- Translation error amplifies 2.77× from synthetic val to real test, rotation is nearly unchanged (+2.23°).

Conclusion

- Frame-by-frame CNN outperforms temporal GRU at 10 Hz, making temporal aggregation redundant.
- Rotation error is robust to domain shift while translation error is not.
- Root cause of domain gap identified and has a clear fix.

Future Work

- Normalized timestamp voxel construction to eliminate unit mismatch.
- KeypointHeatmapNet + PnP architecture.
- Event rate normalization across domains.

References

1. Rathinam et al., "SPADES," arXiv:2311.05310, 2023
2. Gallego et al., "Event-based vision: A survey," IEEE TPAMI, 2022
3. He et al., "Deep Residual Learning," CVPR 2016
4. Ganin et al., "Domain-Adversarial Training," JMLR 2016

Acknowledgements

I would like to express my sincere gratitude to Dr. Bharatesh Chakravarthi for mentoring me throughout the process. I would also like to thank Kashyap Kota Hegde and Pruthvi Janga for their immense support and expreties.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University