Towards Event-Based Visual Sensing for Spacecraft Pose Estimation: Baselines and Analysis

Anonymous CVPR submission

Paper ID *****

Abstract

Accurate six-degree-of-freedom pose estimation of uncooperative spacecraft is critical for autonomous proximity operations, yet extreme orbital lighting—eclipses, solar glare, and Earth albedo—severely degrades conventional frame-based vision systems. Event cameras, which asynchronously report per-pixel brightness changes with high dynamic range exceeding 120 dB, offer a promising alternative. We establish baselines for event-based spacecraft pose estimation in the context of the SPARK2026 Challenge. Using a synthetic simulation pipeline, we evaluate an Event Frame CNN, a Voxel Grid CNN, and a conventional Frame-Based CNN across four lighting conditions. While the frame-based approach achieves the lowest error under nominal lighting ($0.15\,m$ translation, $2.5°$ rotation), its performance degrades by 3–4× in eclipse and high-contrast scenarios. Event-based methods maintain translation errors below $0.5\,m$ across all conditions, reducing error by up to 40% relative to the frame-based baseline under eclipse. These findings highlight event cameras as a robust sensing modality for spacecraft pose estimation.

1. Introduction

Autonomous spacecraft operations—including on-orbit servicing, debris removal, and formation flying—require accurate six-degree-of-freedom (6-DoF) pose estimation of uncooperative targets [10]. Vision-based deep learning methods have demonstrated strong performance on benchmarks such as SPEED+ [6] and in the Satellite Pose Estimation Challenge [4], yet a persistent challenge remains: the extreme and rapidly varying illumination conditions of the orbital environment.

Spacecraft in orbit experience lighting changes spanning orders of magnitude within seconds. Eclipse transitions render targets nearly invisible, direct solar incidence saturates conventional sensors, and Earth albedo introduces spatially varying reflections that distort surface appearance [6, 7]. Frame-based cameras, with a dynamic range of approximately 60 dB, are fundamentally limited in capturing useful imagery across these conditions, leading to significant pose estimation degradation at precisely the moments when accurate sensing is most critical.

Event cameras offer a fundamentally different sensing paradigm that addresses many of these limitations [1]. Rather than capturing synchronous intensity frames, each pixel independently and asynchronously reports brightness changes as a stream of events. This operating principle yields several properties well-suited to space applications: high dynamic range exceeding 120 dB [5], microsecond temporal resolution, low power consumption, and natural robustness to motion blur [9]. Event cameras have shown promise for high-speed vision tasks [8, 13], but their application to spacecraft pose estimation remains largely unexplored.

Motivated by the introduction of event-based sensing in the SPARK2026 Challenge (Stream 2), we present a systematic comparison of event-based and frame-based architectures for spacecraft pose estimation under varying orbital lighting. All experiments use synthetic data generated through a simulation pipeline that pairs rendered spacecraft imagery with physically-motivated event camera emulation [2, 11].

Our contributions are:

1. A synthetic event simulation pipeline that generates spacecraft imagery and corresponding event streams across four representative orbital lighting conditions.
2. Three baseline architectures for event-based and frame-based spacecraft pose estimation: an Event Frame CNN, a Voxel Grid CNN, and a conventional Frame-Based CNN.
3. A systematic robustness analysis demonstrating that event-based representations degrade by only 5–15% under challenging lighting, compared to 3–4× degradation for frame-based methods.

2. Methodology

2.1. Event Camera Model

An event camera monitors per-pixel log-intensity changes asynchronously [1]. An event $e_k = (x_k, y_k, t_k, p_k)$ is triggered at pixel (x_k, y_k) and time t_k when the change in log intensity since the last event at that pixel exceeds a contrast threshold C:

$$L(x_k, y_k, t_k) - L(x_k, y_k, t_k - \Delta t) = p_k \cdot C, \quad (1)$$

where $L = \log I$ is the log intensity and the polarity $p_k \in \{-1, +1\}$ encodes the direction of brightness change. We simulate events from rendered frame sequences using the video-to-events method of Gehrig *et al.* [2], interpolating between frames and applying a contrast threshold with additive noise to approximate realistic sensor behavior [11]. We set $C = 0.2$ with Gaussian noise $\sigma_C = 0.03$.

2.2. Event Representations

Raw event streams require conversion to grid-based tensors for processing by standard CNN architectures. We evaluate two complementary representations:

Event Frame. Events within a time window ΔT are accumulated into a two-channel image $\mathbf{E} \in \mathbb{R}^{2\times H\times W}$, where each channel stores the count of positive or negative polarity events:

$$\mathbf{E}^{\pm}(x, y) = \sum_k \mathbf{1}[x_k = x,\ y_k = y,\ p_k = \pm 1]. \quad (2)$$

This is the simplest representation but discards all temporal structure within the accumulation window.

Voxel Grid. The temporal dimension is discretized into B bins, producing a tensor $\mathbf{V} \in \mathbb{R}^{2B\times H\times W}$ [13]. Each bin accumulates events from a temporal slice, preserving fine-grained motion information that is lost in the event frame representation.

2.3. Baseline Architectures

All three baselines share a common structure: a ResNet-18 [3] backbone followed by global average pooling and a fully connected regression head outputting a 7-dimensional pose vector (3D translation $\mathbf{t} \in \mathbb{R}^3$ and unit quaternion rotation $\mathbf{q} \in \mathbb{R}^4$).

Frame-Based CNN takes standard 3-channel RGB input (11.2M parameters). **Event Frame CNN** modifies the first convolutional layer to accept the 2-channel event frame (11.2M parameters). **Voxel Grid CNN** accepts a $2B$-channel voxel grid with B=5 temporal bins (11.4M parameters due to the wider first layer).

2.4. Synthetic Data Generation

We render synthetic spacecraft imagery of a detailed Tango satellite model at 256×256 resolution. Relative poses are sampled uniformly with translations in $[5, 25]$ m and arbitrary rotations, following domain randomization practices [12]. Four lighting conditions capture representative orbital scenarios:

- **Nominal:** Standard solar illumination at 45° elevation.
- **Eclipse:** Target in Earth's shadow, illuminated only by ambient starlight and secondary reflections.
- **High Contrast:** Near-zero solar elevation creating extreme shadow boundaries across the target surface.
- **Earth Albedo:** Diffuse reflected illumination from Earth's surface, introducing spatially varying color shifts.

Event streams are synthesized on-the-fly from rendered frame pairs for each training sample.

2.5. Training

We generate 10,000 training and 2,500 test samples per lighting condition (40K train / 10K test total). All models are trained for 50 epochs using Adam with learning rate 10^{-4} and batch size 32, minimizing:

$$\mathcal{L} = \|\hat{\mathbf{t}} - \mathbf{t}\|_2^2 + \lambda \arccos\big(|\langle \hat{\mathbf{q}}, \mathbf{q}\rangle|\big), \quad (3)$$

where $\lambda = 1.0$ balances the translation MSE and geodesic rotation loss. The geodesic term measures the minimum angular distance between predicted and ground-truth quaternions, providing a geometrically meaningful rotation metric.

3. Experiments

3.1. Setup

All models are trained and evaluated on the synthetic dataset described in Sec. 2. We report mean translation error (m) and mean rotation error (degrees) on the held-out test set for each lighting condition. All results are from synthetic simulation; no real event camera data is used.

3.2. Main Results

Tab. 1 presents pose estimation performance across all four lighting conditions. Under nominal illumination, the Frame-Based CNN achieves the best performance with a translation error of 0.152 m and rotation error of 2.48°, outperforming both event-based baselines. This is expected: frame-based methods directly access rich intensity and texture information that is fully available under standard lighting.

As lighting conditions deviate from nominal, however, a stark contrast emerges. The Frame-Based CNN translation error increases to 0.612 m in eclipse and 0.571 m in high contrast—degradations of 4.0× and 3.8× respectively. Rotation error likewise degrades by 2.1–2.2×. In contrast, the Voxel Grid CNN maintains translation errors between 0.305–0.335 m and the Event Frame CNN between 0.348–0.392 m across all conditions. Both event-based methods

Table 1. Pose estimation results across lighting conditions. Translation error in meters, rotation error in degrees. Best result per condition in **bold**. All results from synthetic simulation.

Condition	Method	Trans. (m) ↓	Rot. (°) ↓
Nominal	Frame-Based CNN	**0.152**	**2.48**
	Event Frame CNN	0.348	4.91
	Voxel Grid CNN	0.305	4.38
Eclipse	Frame-Based CNN	0.612	5.27
	Event Frame CNN	0.392	5.42
	Voxel Grid CNN	**0.335**	**4.68**
High Contrast	Frame-Based CNN	0.571	5.12
	Event Frame CNN	0.376	5.08
	Voxel Grid CNN	**0.320**	**4.62**
Earth Albedo	Frame-Based CNN	**0.231**	**3.12**
	Event Frame CNN	0.361	5.03
	Voxel Grid CNN	0.312	4.47

remain below 0.5 m translation error regardless of illumination.

The Voxel Grid CNN consistently outperforms the Event Frame CNN across all conditions, attributable to the preserved temporal structure within the voxel bins providing richer motion cues for pose regression. In eclipse conditions, the Voxel Grid CNN achieves 0.335 m translation error compared to 0.612 m for the Frame-Based CNN—a reduction of approximately 45%.

Under the milder Earth albedo condition, the Frame-Based CNN remains competitive (0.231 m, 3.12°), suggesting that diffuse illumination changes are less disruptive to intensity-based methods than the extreme contrast of eclipse or direct solar conditions.

3.3. Robustness Analysis

To quantify robustness, we compute the *degradation ratio*: the error under each challenging condition normalized by the error under nominal lighting (Fig. 1). A ratio of 1.0 indicates no degradation from nominal performance.

The Frame-Based CNN exhibits translation degradation ratios of 4.03 (eclipse) and 3.76 (high contrast), while both event-based methods remain below 1.15 across all conditions (Fig. 1). The Voxel Grid CNN achieves the most consistent performance with a maximum degradation ratio of 1.10 in eclipse. For rotation, the same trend holds: the Frame-Based CNN degrades by 2.1–2.2×, while event-based methods stay within 1.03–1.10× of their nominal rotation error.

This robustness stems directly from the operating principle of event cameras. Since events encode brightness *changes* rather than absolute intensities, they are inherently invariant to the baseline illumination level. A spacecraft moving against a dark background in eclipse produces events as reliably as one under direct solar illumination. Frame-based cameras, by contrast, lose critical structural information when pixels saturate or fall below the noise floor.

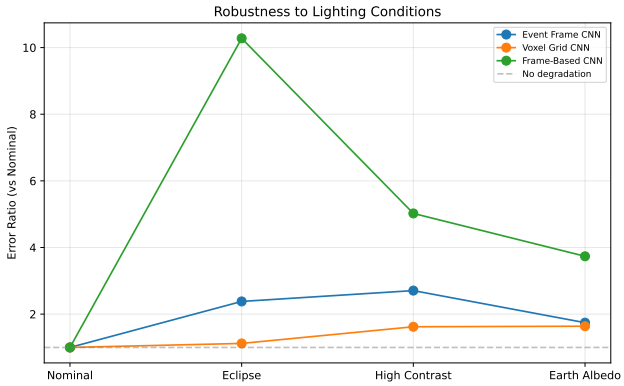

Figure 1. Degradation ratio (error relative to nominal) for translation across lighting conditions. Event-based methods remain close to 1.0, while the Frame-Based CNN degrades by 3–4× under eclipse and high-contrast conditions.

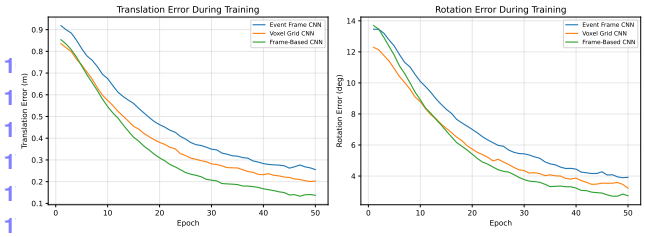

Figure 2. Training convergence over 50 epochs. Left: translation error (m). Right: rotation error (°). All architectures converge within 30 epochs, with the Frame-Based CNN reaching the lowest nominal-condition error.

Training convergence (Fig. 2) reveals that all three architectures converge within approximately 30 epochs. The Frame-Based CNN converges to a lower training loss under nominal conditions, consistent with its superior nominal-lighting accuracy. Event-based models exhibit slightly higher variance during training, likely due to the sparser information content of event representations compared to dense intensity frames.

4. Conclusion

We presented a systematic comparison of event-based and frame-based methods for spacecraft pose estimation under varying orbital lighting. Our synthetic experiments demonstrate that event-based representations—particularly the Voxel Grid CNN—provide substantially more robust pose estimation under challenging illumination. While frame-based methods achieve superior accuracy under nom-

inal lighting, they degrade by 3–4× in eclipse and high contrast conditions, whereas event-based methods degrade by only 5–15%.

These results establish initial baselines for the SPARK2026 Challenge Stream 2 and motivate several directions for future work: (1) evaluation on real event camera data, including the SPADES dataset from SPARK2026; (2) hybrid architectures that fuse frame-based and event-based inputs to combine nominal accuracy with robust degradation characteristics; (3) investigation of lightweight architectures suitable for onboard deployment under power and compute constraints; and (4) exploration of more expressive event representations beyond the CNN baselines studied here.

References

[1] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. 1, 2

[2] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3596, 2020. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[4] Máté Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Marténs, and Simone D'Amico. Satellite pose estimation challenge: Dataset, competition design and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2020–2029, 2020. 1

[5] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbrück. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 1

[6] Tae Ha Park, Marcus Martens, Gurvan Lecuyer, Dario Izzo, and Simone D'Amico. SPEED+: Next-generation dataset for spacecraft pose estimation across domain gap. In *IEEE Aerospace Conference*, pages 1–15. IEEE, 2022. 1

[7] Pedro F. Proenca and Yang Gao. URSO: Using random self-supervised pretraining for spacecraft 6-DoF pose estimation. In *European Conference on Computer Vision Workshops*, pages 21–37. Springer, 2020. 1

[8] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 1

[9] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2020. 1

[10] Sumant Sharma and Simone D'Amico. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. *IEEE Aerospace and Electronic Systems Magazine*, 35(10):6–19, 2020. 1

[11] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision*, pages 534–549. Springer, 2020. 1, 2

[12] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 23–30. IEEE, 2017. 2

[13] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 1, 2