

6-DOF Relative Pose Estimation For Non-cooperative Spacecraft Using Event Streams

Wang yishi^{*}, Michele Maestrini^{*}, Mauro Massari^{*}, Zhang Zexu^{*†}, Pierluigi Di Lizia^{*}

^{*}*Department of Deep Space Exploration Research Center, Harbin Institute of Technology
No.2, Yikuang Street, Nangang District, Harbin, 150080, PR China*

^{*}*Department of Aerospace Science and Technology, Politecnico di Milano
Via La Masa 34, 20156 Milano, Italy*

wangyishi@hit.edu.cn · michele.maestrini@polimi.it · mauro.massari@polimi.it
zexuzhang@hit.edu.cn · pierluigi.dilizia@polimi.it

[†]Corresponding author

Abstract

Accurate relative pose estimation of non-cooperative spacecraft is essential for on-orbit servicing and active debris removal, serving as a critical component of autonomous navigation by providing position and orientation measurements for navigation filters. Convolutional neural networks (CNNs) have shown superior performance in monocular pose estimation over traditional methods. However, CNNs-based methods are limited in capturing global image dependencies, and low illumination, texture-less, and motion blur in the space environment present significant challenges for those monocular optical camera-based approaches. To improve performance in challenging space environments, we propose a novel 6D pose estimation method leveraging event streams and Transformer model. Compared with the traditional optical sensors, event cameras offer advantages such as microsecond-level temporal resolution, high dynamic range, and low power consumption, making them superior in motion-related tasks and robust to lighting variations. Specifically, we design an architecture combining CNNs and Transformer to estimate keypoint heatmaps. The pipeline first converts event streams into event images, extracts local features via a CNN backbone, and encodes long-range dependencies among spatial features of the events using a Transformer encoder. The multi-head self-attention layers built in Transformer enable our model to capture global contextual information and reveal what dependencies the predicted keypoints rely on. This is followed by a keypoint heatmaps regression head, which estimates the target keypoints at the sub-pixel level based on Gaussian peaks. Finally, our method estimates the 6D pose by solving the Perspective-n-Point problem. To support this research, we generated an Event-Monocular-Spacecraft (EMS) dataset, a publicly available synthetic event monocular dataset, containing RGB images under varied lighting conditions, corresponding event streams, and precise pose labels. The proposed method's performance is evaluated on the E-M-S dataset. Extensive experiments demonstrate that the proposed method achieves a high-precision estimation rate of 84%, representing an 8% improvement over traditional optical models. Furthermore, it outperforms other state-of-the-art event-based methods, achieving superior pose estimation accuracy and maintaining robust performance under extreme lighting conditions.

1. Introduction

Recent advancements in aerospace engineering have enabled a wide range of space missions, including active debris removal [1], robotic servicing of defunct satellites [2], and autonomous rendezvous and docking maneuvers [3]. In such missions, the servicing spacecraft must often interact with non-cooperative targets, which typically have known structural characteristics but lack fiducial markers and may exhibit unpredictable rotational motion. Successfully executing proximity operations in such scenarios relies on the development of advanced autonomous navigation techniques. In particular, accurate estimation of the relative six degrees of freedom (6-DOF) pose and other motion parameters is essential for enabling reliable autonomous navigation and control systems.

In close-range proximity operations, 6-DOF relative pose estimation serves as the front-end observation module in autonomous navigation, providing high-precision relative position and attitude observations for downstream relative navigation filtering systems [4]. For non-cooperative targets, the spacecraft lacks an active communication link, and

ground-based motion estimation is often subject to significant uncertainties. As a result, it cannot meet the real-time and high-precision requirements of close-range operations, rendering ground-based relative navigation infeasible. Therefore, onboard—particularly visual sensors—must be utilized by the servicing spacecraft to estimate the relative pose of the target.

Various visual sensors have been explored for 6-DOF pose estimation, including monocular cameras, LiDAR, and depth-sensing cameras. Depth cameras are primarily used indoors due to their limited measurement range, making them unsuitable for space applications [5]. The accuracy of binocular cameras is heavily dependent on baseline length, restricting their usage in space [6]. LiDAR, while effective, consumes more energy and adds significant weight [7]. In contrast, monocular camera systems are preferred in space due to their low cost, lightweight design, and energy efficiency [8], as seen in the RemoveDEBRIS mission by Surrey Space Centre [9], AVANTI experiment conducted by German Aerospace Center [10], SEEKER-1 mission by NASA [11].

Monocular pose estimation methods are generally based on traditional approaches using handcrafted features [12] or deep learning approaches. Traditional methods are not very robust in challenging space environments. Due to the significant success of deep learning in computer vision tasks, an increasing number of researchers are applying learning-based methods to 6D pose estimation of space targets [13]. The European Space Agency’s Advanced Concepts Team (EST) and Stanford’s Space Rendezvous Laboratory (SLAB) introduced the Satellite Pose Estimation Challenge (SPEC) [14], which compared the strengths and limitations of various pose estimation methods, highlighting the potential of learning-based approaches. In the SPEC competition, learning-based pose estimation methods include end-to-end and hybrid modular approaches. End-to-end methods [15] train a deep learning model to estimate the spacecraft pose directly from images without depending on intermediate stages. In contrast, hybrid modular methods [16] utilize neural networks to predict the 2D keypoints of the target, and then calculate the relative pose based on geometric optimization methods [6]. Hybrid modular methods that combine deep learning models with traditional Perspective-n-Point (PnP) algorithms [13] have been shown to achieve significantly higher accuracy [6]. However, despite progress in monocular pose estimation, challenges remain due to the low signal-to-noise ratio (SNR) and complex lighting conditions for monocular source images in space. Study [17] provides substantial experimental evidence that extreme lighting conditions negatively impact image-based pose estimation. Both overexposure and insufficient lighting significantly reduce estimation accuracy, with extremely low lighting causing the greatest decline. These difficulties make robust pose estimation still an open problem in extreme space environments.

Recently, neuromorphic cameras, also known as event cameras, have gained significant attention in visual tasks such as object recognition and depth estimation [18]. Event cameras are bio-inspired sensors that, in contrast to traditional cameras, do not capture intensity frames at a fixed framerate but record each pixel’s brightness change as an asynchronous event stream, including a timestamp, pixel position, and the polarity of the brightness change [19]. Based on the novel design concept, event cameras offer advantages like microsecond-level temporal resolution, a high dynamic range (HDR) of 140 dB, and low power consumption, making them superior in motion-related tasks [20]. These sensors are also robust to lighting variations and function effectively under extreme illumination, such as overexposure or low light [21]. Furthermore, event cameras can handle motion blur, a limitation of traditional cameras. These characteristics make event sensors particularly well-suited for addressing various visual tasks, especially in challenging space scenes.

This study proposes a method to estimate the pose of non-cooperative targets using event streams, leveraging the high dynamic range and asynchronous change detection of event sensors to perceive spacecrafts in a way that is insensitive to space extreme illuminations. The main contributions of this work are summarized under the following four aspects. First, a novel event-based method for solving the pose estimation problem of non-cooperative spacecrafts is proposed. Then, this study introduces a hybrid modular pipeline based on deep learning and PnP algorithm, presenting a keypoints regression network architecture that combines Convolutional Neural Networks (CNNs) and Transformers, enabling CNNs to extract detailed local features while Transformers handle global contextual information. Our network framework leverages the self-attention mechanism to capture long-range dependencies efficiently and reveal what dependencies the predicted keypoints rely on. Next, this study develops a publicly available synthetic Event-Monocular Spacecraft (E-M-S) images dataset, simulating various space illumination. This dataset includes a large number of event streams and corresponding optical images which can be used for 6D pose estimation modal training and evaluation, also contribute to other space vision research. Finally, extensive experiments demonstrate that the proposed approach significantly outperforms the methods based on monocular images and other alternative state-of-the-art event-based methods, achieving higher accuracy and maintaining robust performance under extreme lighting conditions.

The remainder of this paper is organized as follows. After a discussion of related work, Section 3 presents the core concepts and architecture of the pose estimation method. Section 4 introduces the Event-Monocular Spacecrafts (E-M-S) dataset. Section 5 describes extensive experiments conducted on simulated datasets to evaluate the performance of the proposed model. Finally, the conclusions and future works of this paper are presented in Section 6.

2. Related works

2.1 Learning-based Pose Estimation of Spacecraft

Traditional approaches rely on hand-crafted features and geometry knowledge to solve the spacecraft pose estimation problem [22]. But such methods usually show deteriorated performance for poorly textured targets in complex changeable backgrounds. Recently, deep learning-based methods have achieved success in satellite pose estimation, thanks to the powerful feature representation ability of deep models. Albert et al. [23] explores a novel deep learning-based methodology that employs Convolutional Neural Networks (CNNs) to directly regress the poses of uncooperative spacecraft without requiring any prior 3D information. Sharma et al. [24] introduced a novel deep learning-based pose estimation framework Spacecraft Pose Network (SPN). The approach utilized Faster R-CNN [25] for satellite detection, and estimated pose by optimizing the geometric constraints of the perspective transformation. Similar to SPN, other works [13] addressed the pose estimation problem by decoupling it into two networks: object detection and pose estimation. The pose estimation is performed by regressing the 2D locations of the spacecraft's keypoints then solving the Perspective-n-Point (PnP) problem.

Although these methods rely on monocular images and CNNs, which have demonstrated strong performance in controlled tests, their effectiveness diminishes significantly in complex real-world scenarios. This decline stems from the inherent architectural limitations of CNNs. While CNNs excel at extracting local features, they struggle to model long-range dependencies, resulting in a bias towards image textures and a failure to capture global features effectively. Furthermore, under extreme illumination conditions, monocular images experience reduced target texture visibility, which hinders generalization. Consequently, combining the illumination robustness of event images with Transformer-based architectures provides an effective solution to the unique challenges of spacecraft perception in space.

2.2 Transformer models application to pose estimation

Transformer architecture [26] was initially developed for Natural Language Processing (NLP) tasks, addressing the challenge of learning complex dependencies in long sequences. Unlike CNNs, which excel in extracting local features, Transformers capture global dependencies through multi-head self-attention [26]. Vision Transformer (ViT) [27] was the first to successfully apply Transformer architecture to the visual domain, achieving strong results in image classification tasks. This breakthrough demonstrated the potential of Transformers for computer vision applications, inspiring further research into their use across a variety of tasks.

Wang et al. [12] devised a transformer-based keypoint predictor for predicting the semantic keypoints of the target in monocular images, and then estimated the pose by solving the perspective-n-point (PnP) problem. Wang et al. [6] applied the Transformer architecture to non-cooperative target pose estimation using two-stage methods. They used the Detection Transformer (DETR) framework for keypoints prediction, demonstrating that Transformer-based predictors offer a better trade-off between speed and accuracy compared to CNN-based approaches. Liu et al. [28] proposed the Deformable Transformer-based Single Stage End-to-End SpaceNet (DTSE-SpaceNet). They applied the deformable transformer to monocular spacecraft pose estimation, achieving both multiscale fusion and global modeling without increasing computational complexity.

The aforementioned Transformer-based pose estimation methods have demonstrated improved performance over CNN-based approaches. However, they are all designed for conventional monocular optical images and still suffer from performance degradation under extreme lighting conditions, such as low illumination, failing to effectively address the challenges posed by harsh space environments.

2.3 Event-based vision applications

Event sensors offer several advantages over standard optical sensors [29], are gradually becoming more popular in the computer vision and robotics community. Many recent works such as object recognition [30] and simultaneous localization and mapping (SLAM) [31] have been actively investigated. For pose estimation tasks, Elias et al. [32] presented the first onboard perception system for 6-DoF pose tracking during high-speed maneuvers using event cameras. Nguyen et al. [19] introduced a method to localize the 6DoF pose of an event camera solely. They first created the event images from a list of event streams, then a Stacked Spatial LSTM Network is used to estimate the pose. Building upon [19], Tabia et al. [33] proposed a new method for estimating the 6DOF pose of an event camera using the fully convolutional network. However, the aforementioned methods have primarily been applied in ground scenarios, and they focus on determining the pose of the camera rather than the pose of the target. In the space domain, the application of event sensors on spacecrafts is still in its infancy. Jawaaid et al. [34] explored the application of event cameras for pose estimation of satellites, employing classical learning-based methods to detect the 2D landmarks of satellites, and

utilizing a PnP solver to determine the pose. However, their approach relies solely on convolutional networks, which are limited in capturing the global semantic information of event images. As a result, the model struggles to learn long-range dependencies within event streams. Liu et al. [20] proposed a line-based pose estimation method for spacecraft using an event camera. This approach estimates the pose by directly extracting line features from the event data and matching them to the 3D model. While robust for high-speed moving targets, the accuracy of this method relies heavily on handcrafted designed line feature extraction and matching, requiring complex initialization to establish 2D-3D correspondences.

The application of event sensors in spacecraft perception is still in its early stages, with only a few studies exploring their use in the aerospace domain. This research area remains promising but also presents significant challenges.

3. Event-based pose estimation method

This section introduces a novel framework for estimating the 6D pose of non-cooperative spacecraft using event streams. Firstly, section 3.1 details the methods for converting event streams into event images. Section 3.2 provides an overview of the designed network architecture, describes the associated loss functions, and details the method used for 6D pose estimation.

3.1 Event-to-Frame conversion

Event cameras do not capture image frames in the conventional sense; instead, their output is a stream of asynchronous events with no notion of frames. Each pixel u_i operates independently and responds to changes in brightness. When the log brightness change exceeds a threshold c , a new event $\mathbf{e}_i = (x_i, y_i, t_i, p_i)$ is generated at pixel $u_i = (x_i, y_i)$. The polarity p_i is either -1 or +1 (negative or positive), depending on the brightness change direction, and t_i denotes the event's timestamp.

To leverage existing CNN architectures, the continuous asynchronous event streams are generally converted into synchronous frames with fixed tensors. An event stream $\mathbf{E}$ is represented as $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots\}$, where each event $\mathbf{e}_i = \{x_i, y_i, t_i, p_i\}$. This study converts $\mathbf{E}$ into event frames $\mathbf{F} = \{I_1, I_2, \dots\}$, as illustrated in Fig. 1 with examples of event images. Following [19], this study assumes that within a specific time window W_j , n events will maintain the same camera pose. This assumption is based on the fact that event cameras capture numerous events in a short interval τ , during which the camera pose is assumed to remain relatively constant. With the assumption of instantaneous pose consistency, the conversion process can ignore the noise events introduced by differential changes in illumination caused by relative instantaneous motion.

For the specific event-to-frame (E2F) method, this study accumulates events $\mathbf{e}_i\{x_i, y_i, t_i, p_i\}$ over a given time interval τ in a pixel-wise manner, resulting in 2D histograms of events with a resolution of height $\times$ width as the camera. Similar to [36] and [34], the event-streams are converted to image-like representation $I \in \mathbb{R}^{4 \times H \times W}$ with four channels. The value of the first two channels based on the polarity p illustrated in Eq.1 and Eq.2. The polarity distributions on the frame can effectively describe the spatial characteristics of the event-stream.

$$I_{\text{positive}}(x_i, y_i) = \begin{cases} 1, & \text{if } p_i = 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$I_{\text{negative}}(x_i, y_i) = \begin{cases} 1, & \text{if } p_i = -1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The values of other two channels are constructed from the timestamps of the positive and negative events, respectively. These values represent ratios that describe the temporal characteristics. The ratio at pixel (x_i, y_i) is defined as:

$$I(x_i, y_i) = \frac{t_{i,j} - t_{\text{begin}}}{t_{\text{end}} - t_{\text{begin}}}, \quad (3)$$

where $t_{i,j}$ is the timestamp of the positive (negative) event at pixel (i, j) , t_{begin} represents the timestamp of first positive (negative) event and t_{end} is the last positive (negative) event of the whole stream.

3.2 Pose estimation network architecture

The network architecture shown in Fig. 2 consists of three main components: a CNN backbone for feature extraction, a Transformer encoder for capturing global feature dependencies, and a Multilayer Perceptron (MLP) prediction head

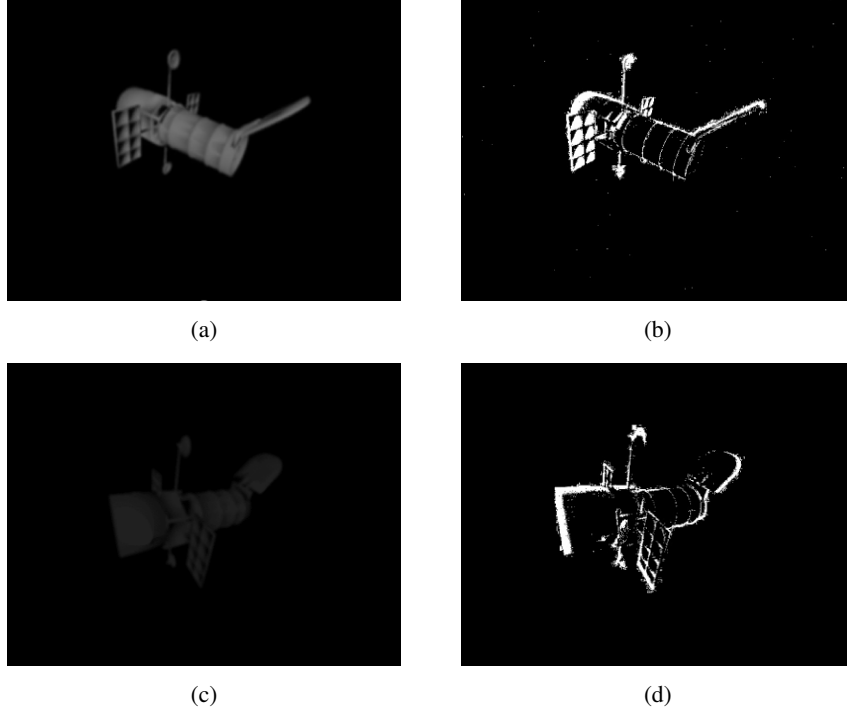


Figure 1: (a) Rendered frame of a satellite with normal lighting.(b) Event frame corresponding to (a), generated using V2E [35].(c) Rendered frame of a satellite with low lighting.(d) Event frame corresponding to (c).The rendered frames are preprocessed with Gaussian blur.

for estimating keypoint heatmaps. First, the event stream is converted into event images and passed through a feature extraction branch to capture local spatial features F_{event} , which encode edge structures and motion-induced contours. These features are then processed by a Transformer encoder to generate global contextual features F_{global} . Subsequently, the global features are upsampled to a higher resolution, and an MLP prediction head is employed to estimate the keypoint heatmaps. Finally, the 6-DOF pose is recovered by solving the Perspective-n-Point (PnP) problem using the predicted 2D -3D correspondences.

CNN-based backbone: In this work High-Resolution Network (HRNet) [37] was adopted as the CNN backbone for feature extraction. Traditional convolutional networks, such as ResNet[38], progressively downsample feature maps to a low resolution (typically 1/32) during feature extraction, resulting in the compression of spatial information. To recover spatial detail, these architectures typically rely on multi-layer upsampling operations, which may introduce interpolation errors and reduce localization accuracy. In contrast, HRNet maintains a high-resolution branch (e.g., 1/4 resolution) throughout the entire network and continuously exchanges information with parallel medium-resolution and low-resolution branches (1/8, 1/16). This design preserves spatial details more effectively and produces more accurate and reliable feature representations, especially for tasks requiring precise localization such as keypoint detection and pose estimation.

HRNet-W32 [37], pre-trained on the ImageNet dataset [39], was employed as the feature extractor. However, unlike the approach in [13], our network does not output feature maps at the original resolution $H_0 \times W_0$; instead, it produces feature maps at a downsampled resolution of $H, W = H_0/r, W_0/r$, where $r = 8$. This is because the computational complexity of the self-attention mechanism in the Transformer is $O((H \times W)^2 \cdot d)$. Feeding full-resolution feature maps of size $H_0 \times W_0$ into the Transformer encoder significantly increases both memory usage and computational cost. Using a 1/8 resolution effectively balances the computational cost of the Transformer module with the preservation of spatial detail in feature maps, providing a favorable trade-off between performance and efficiency. Given event images $I_{\text{event}} \in \mathbb{R}^{C_0 \times H_0 \times W_0}$, the CNNs-based backbone generates feature maps $F_{\text{event}} \in \mathbb{R}^{C \times H \times W}$, $C = 64$ and $H, W = H_0/8, W_0/8$. Then, each feature map goes through a 1×1 convolution to transform the feature dimension to d . This event feature map $F_{\text{event}} \in \mathbb{R}^{d \times H \times W}$ is then be fed to the Transformer encoder.

Transformer Encoder for Global Context Modeling: First, the event feature map F_{event} is flattened into a sequence $F_{\text{event}} \in \mathbb{R}^{l \times d}$ to meet the input requirements of the Transformer. This sequence is then fed into the Transformer encoder. As shown in Fig. 2, the Transformer encoder follows the standard architecture, comprising a multi-head self-attention module and a fully connected feed-forward network. Following [40], the proposed method applies 2D

EVENT-BASED 6D POSE ESTIMATION FOR NON-COOPERATIVE SPACECRAFT

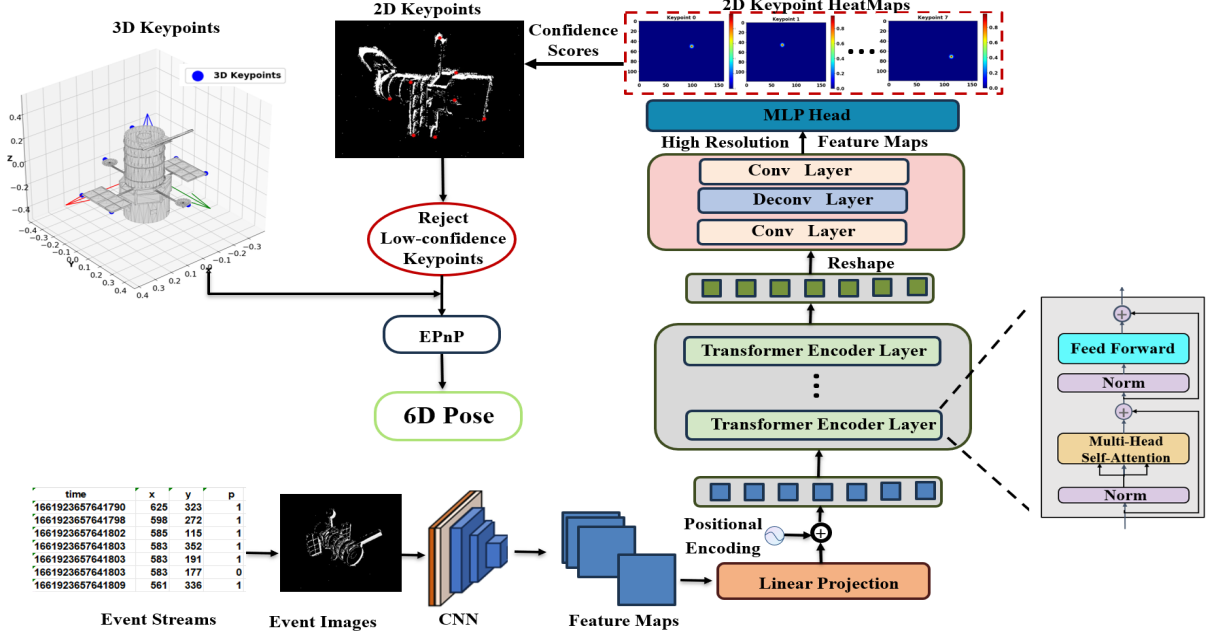
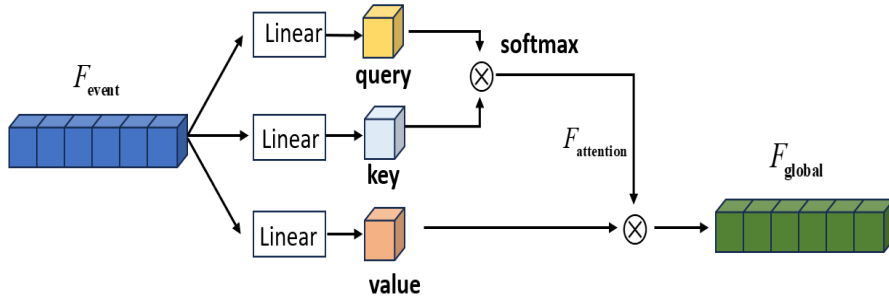


Figure 2: Overview of the proposed Event-based pose estimation approach.

sine positional embeddings to provide order to the input sequence.

The self-attention mechanism is the key component of the Transformer architecture. It enables each spatial feature (i.e., token) to dynamically attend to all other positions within the feature map, allowing the network to effectively capture global context. This mechanism effectively aggregates contextually relevant features while suppressing irrelevant or noisy activations through attention weighting. This contributes to the downstream modules' accurate estimation of the keypoint heatmaps. Fig.3 illustrates the process of generating global features. The input sequence F_{event} is first projected into the query $\mathbf{Q} \in \mathbb{R}^{l \times d}$, key $\mathbf{K} \in \mathbb{R}^{l \times d}$, and value $\mathbf{V} \in \mathbb{R}^{l \times d}$ matrices using three trainable weight matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$. The attention map $F_{\text{attention}}$ is then computed according to Eq.5, representing the importance of each feature in the overall feature representation. Then, the global feature F_{global} is derived by performing element-wise multiplication of $F_{\text{attention}}$ and $\mathbf{V}$.

Figure 3: Diagram of Self-Attention Mechanism used to generate the self-attention feature $F_{\text{attention}}$ from event images.

$$\begin{cases} \mathbf{Q} = F_{\text{event}} \mathbf{W}_q \\ \mathbf{K} = F_{\text{event}} \mathbf{W}_k \\ \mathbf{V} = F_{\text{event}} \mathbf{W}_v \end{cases} \quad (4)$$

$$F_{\text{attention}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right), \quad (5)$$

$$F_{\text{global}} = F_{\text{attention}} \otimes \mathbf{V} \quad (6)$$

Keypoints regression and pose estimation: The d dimensional global feature $F_{\text{global}} \in \mathbb{R}^{l \times d}$ is first reshaped to $\mathbb{R}^{d \times H \times W}$. Then, the feature maps are passed through a 3×3 convolutional layer, followed by upsampling to $1/4$

of the original resolution using either deconvolution or bilinear interpolation. The upsampled feature maps are subsequently processed by a 3×3 convolutional layer to enhance spatial structure re-aggregation. An MLP head is then applied to generate K keypoint heatmaps, with each keypoint corresponding to an independently regressed heatmap. The output is a prediction tensor $\mathbf{P} \in \mathbb{R}^{K \times H^* \times W^*}$ where $H^*, W^* = H_0/4, W_0/4$. These heatmaps are modeled as 2D Gaussian distributions centered at the keypoint locations. The response values in the Gaussian heatmaps are treated as confidence scores, and the final keypoint coordinates are inferred by identifying the peak value in each heatmap. Since the predicted heatmaps are generated at $1/4$ of the original image resolution, the extracted 2D keypoint coordinates are scaled accordingly to match the full-resolution image space. Additionally, to handle images that only capture partial target, we label the visibility v_i of each keypoint $\mathbf{p}_i$.

$$v_i = \begin{cases} 1, & \text{if } \mathbf{p}_i \text{ is inside image frame,} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

The loss function of the keypoint heatmaps is defined as the mean squared error between the estimated heatmap $h(\mathbf{p}_i)$ and the corresponding ground truth $h(\mathbf{p}_i^*)$, averaged over all the K keypoints lying inside the image.

$$\mathcal{L} = \frac{1}{K} \sum_{i=1}^K v_i (h(\mathbf{p}_i) - h(\mathbf{p}_i^*))^2, \quad (8)$$

where $h(\cdot)$ denotes a heatmap representation of a keypoint. The ground truth heatmaps are generated as 2D Gaussian distributions, and standard deviations of 2 pixel. The loss function $\mathcal{L}$ is defined based on a single event image. In a mini batch, $\mathcal{L}$ is simply averaged. After predicting the final 2D keypoints, their associated confidence scores are used to filter out unreliable points that could otherwise cause the pose solver to diverge or yield inaccurate estimates. Finally, the 6D pose is estimated by solving the Perspective-n-Point (PnP) problem[13] with Random Sample Consensus (RANSAC), based on known 3D-2D correspondences.

4. Event-Monocular synthetic dataset for spacecrafts pose estimation

Compared to monocular images, the High Dynamic Range (HDR) and asynchronous response of event sensors make them less sensitive to illumination variations, enhancing target perception in challenging environments. The SEENIC dataset [41] introduced in [34] is notable as the first and currently only event-sensing dataset specifically for spacecraft pose estimation. However, it presents limitations: restricted diversity of pose labels, limited event frames for training, and the absence of corresponding optical images, which hinders the performance comparison and validation across different modalities. To address these issues, this study created a new synthetic Event-Monocular-Spacecraft dataset (E-M-S) [42], incorporating a wider distribution of pose labels, a greater number of event streams, and corresponding monocular RGB images. The dataset simulates varied lighting conditions, including low-light scenarios common in space when targets enter penumbra areas. The dataset is publicly available at <https://doi.org/10.5281/zenodo.14774652>.

First, this study generates a monocular RGB images dataset of the spacecraft. Using pyrender with OpenGL bindings, our simulator renders synthetic images with physically based materials and illumination. The weak-texture 3D model of the Hubble Space Telescope (HST) [43] was selected due to its complex structure with protruding, non-uniform elements that create self-occlusions and shadows. The simulator uses directional light to simulate sunlight conditions in space. By adjusting the light's angle relative to the target and its intensity, this study simulates both standard and low illumination scenarios. Then, images are rendered at a resolution of 640×480 pixels by a virtual camera with a 85° horizontal FOV, as illustrated in Fig.1. To make the generated images more realistic, the synthetic images are processed with speckle noise and Gaussian blur, as shown in Fig.4. To ease the adoption of the dataset for other users, the monocular images are published without noise so that researchers can easily add different types of noise depending on their tasks.

The relative motion trajectory between the camera and the target is generated concurrently with image capture. Similar to the SEENIC dataset [41], this study sets the maximum diameter of the target CAD model at 0.6 meters, and the movement distance of the camera relative to the target is between 0.5-1.2 meters. Unlike SEENIC [41], which views objects only from one side, our setup includes 6 discrete viewing spaces, allowing object observation from multiple perspectives and generating a diverse set of pose labels. In each discrete space, 80 initial cameras and initial pose are generated, ensuring that the target remains within the camera's field of view, as shown in Fig.5 (a).

Each initial camera generates a continuous motion trajectory with 100 frames, as shown in Fig.5 (b). The motion trajectory is based on the target body coordinate system as the reference frame. Sequences with varying relative velocities, distances, and viewing angles are rendered. In the translation vector, T_x and T_y are constrained by the camera's FOV. Continuous image sequences are essential to create event streams, which rely on brightness changes

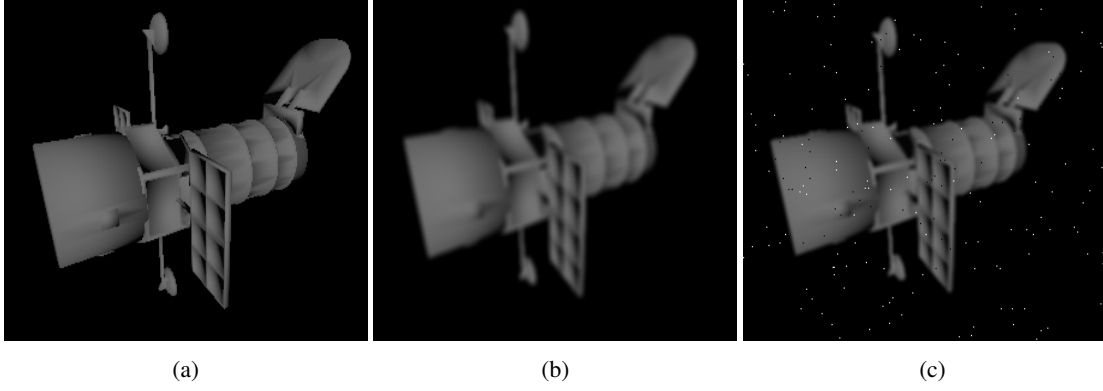


Figure 4: Images augmentation: (a) Ideal images. (b) Introduction of Gaussian blur. (c) Addition of noise.

between consecutive frames. For training, 480 sequences were generated, each containing 100 RGB frames, totaling 48,000 frames (24,000 images with norm lighting and 24,000 images with low lighting) and their corresponding 6D poses. Following the same method, this study produced 4,800 frames for testing and an additional 4,800 frames for validation in the same scene. The distributions of pose labels are shown in Fig. 6. Figure 6 (a) visualizes the distribution of relative positional components (X, Y, Z). Figure 6 (b) displays the distribution of rotation, parameterized by Euler angles.

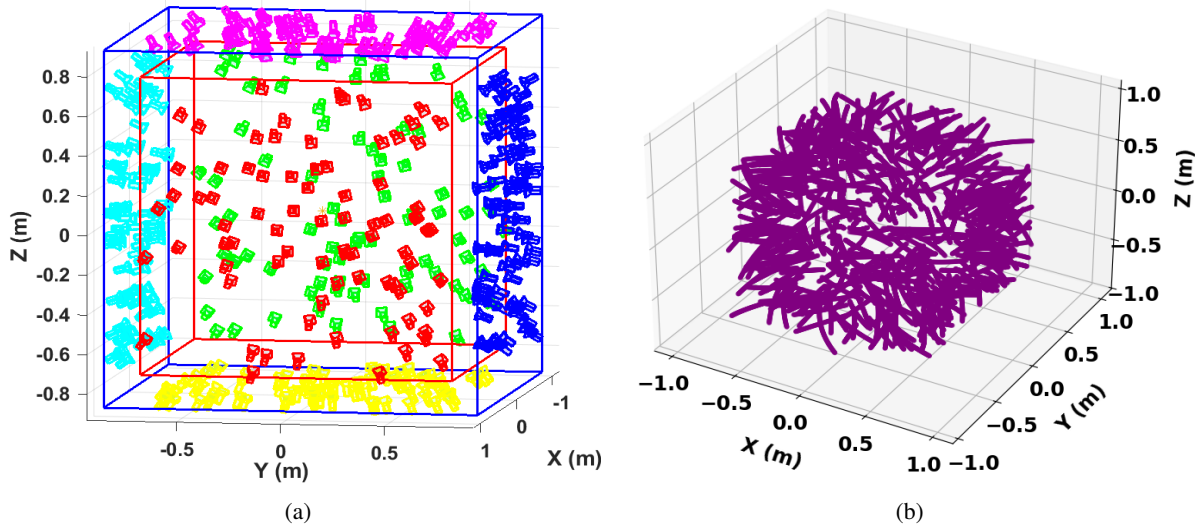


Figure 5: (a) represents the initial camera's pose space discretization ; (b) represents the continuous motion trajectories of cameras.

The event stream dataset is generated using the V2E event simulator [35], which converts linear optical frames (0-255 intensity) to log intensity. This simulator provides a realistic event sensor simulation by accurately modeling pixel-level behaviors, incorporating latency, noise, frame event duration, and other relevant characteristics. This study sets the Dynamic Vision Sensor (DVS) exposure duration time (event accumulation time) to 0.01s, the threshold for positive/negative event triggering to 0.1, and add noise to the event streams. From the generated event streams, this study produces event frames following the E2F procedure in Sec.3.1. Samples of generated synthetic event images are depicted in Fig.7.

To train the Gaussian keypoint heatmap regression network, ground truth heatmaps are generated for supervision. A set of $K = 8$ 3D keypoints is selected from the target's known 3D model structure. These 3D keypoints are projected onto the 2D image plane using the corresponding virtual camera intrinsics and ground-truth 6D poses. For each projected 2D keypoint $\mathbf{p}_i$, a heatmap is generated based on a 2D Gaussian distribution centered at $\mathbf{p}_i$, following the Eq.9 This process is applied to every event image to create per-keypoint supervision. The generated ground truth

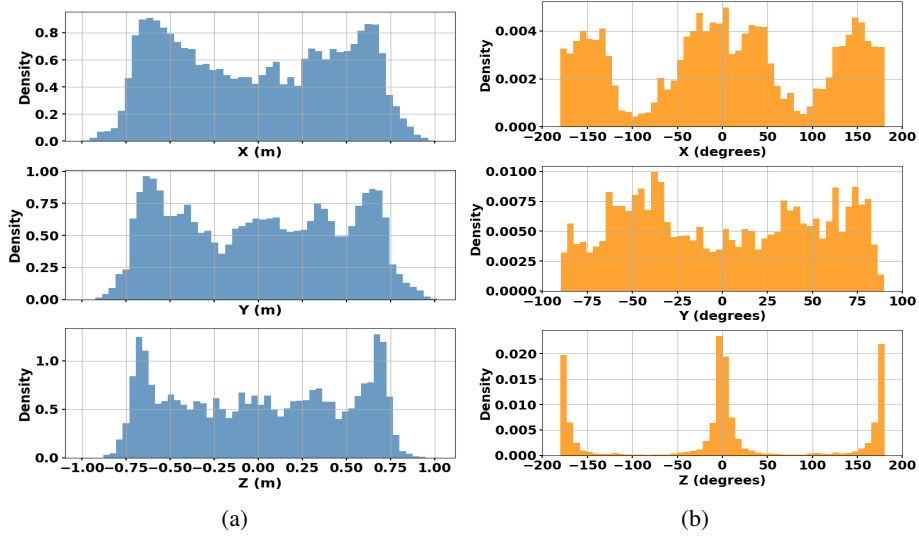


Figure 6: The distributions of pose:(a) Histograms of the camera positions distribution;(b) Histograms of the camera attitude distribution.

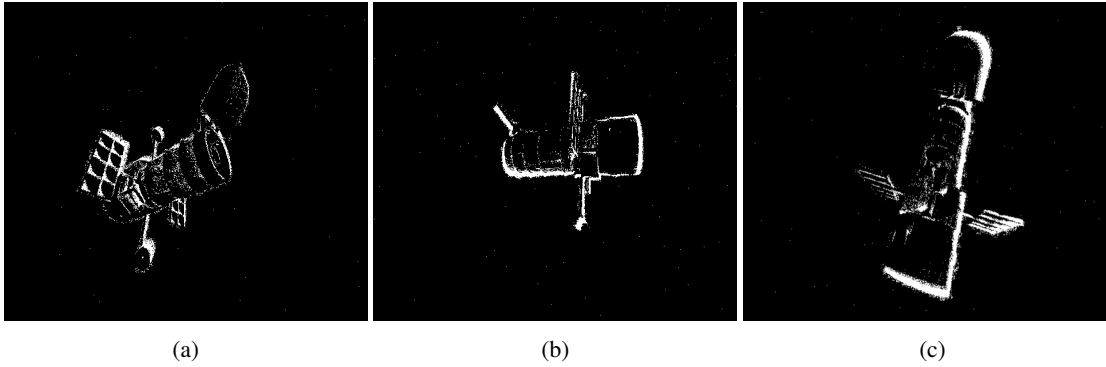


Figure 7: Example event images from the generated dataset

Gaussian keypoint heatmaps are illustrated in Fig.8.

$$h(\mathbf{p}(x_i, y_i)) = \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right) \quad (9)$$

5. Experiments

In this section, we evaluate the proposed method using our generated Event-Monocular Spacecraft dataset, detailed in Sec.4. To comprehensively evaluate the performance advantage of event data over monocular images, we conducted experiments using the proposed method on each modality separately. Additionally, we compared the proposed method with other state-of-the-art event-based spacecraft pose estimation approaches [34, 38] to highlight its superior performance. Specifically, the designed network pipeline predicts the 2D coordinates of each 3D keypoint of the target on the image plane to establish 2D-3D correspondences. The 6D pose is then estimated using the Efficient Perspective-n-Point (EPnP) algorithm combined with Random Sample Consensus (RANSAC).

5.1 Evaluation metrics

The pose estimation is evaluated using translation error ϕ and rotation error ψ . The translation error ϕ is the $L2$ norm of the difference between the predicted value $\hat{\mathbf{t}}$ and the ground truth value $\mathbf{t}$.

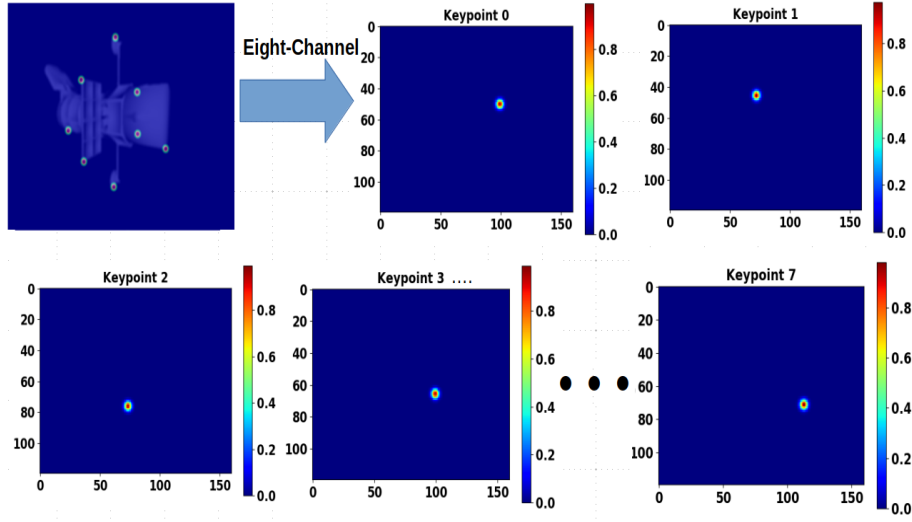


Figure 8: Gaussian keypoint heatmaps

$$\phi = \|\mathbf{t} - \hat{\mathbf{t}}\|^2 \quad (10)$$

The rotation error ψ is defined as the angle between the predicted quaternion $\hat{\mathbf{q}}$ and the ground truth value $\mathbf{q}$. The quaternion $\hat{\mathbf{q}}$ and $\mathbf{q}$ are obtained by converting the calculated rotation matrix $\hat{\mathbf{R}}$ and the ground truth rotation matrix $\mathbf{R}$, respectively.

$$\psi = 2 \arccos(|\langle \mathbf{q}, \hat{\mathbf{q}} \rangle|) \quad (11)$$

We report the average pose error for each sequence ℓ .

$$\Phi = \sqrt{\frac{1}{\ell-1} \sum_{i=1}^{\ell-1} \phi_i^2}, \quad \Psi = \frac{1}{\ell-1} \sum_{i=1}^{\ell-1} \psi_i \quad (12)$$

5.2 Implementation details

Our model is implemented using the PyTorch framework, and all experiments are conducted on 4 NVIDIA RTX A6000 graphics cards. We use the AdamW optimizer with an initial learning rate of $1e-5$, regulated by the MultiStepLR scheduler, which reduces the learning rate by a factor of 0.1 after 20%, 40%, and 60% of the total epochs. Training is conducted over 10 epochs with a batch size of 16. The Transformer encoder is configured with 4 layers. We apply L2 regularization by setting the weight decay of AdamW to $1e-4$ and a dropout rate of 0.2 to reduce the risk of overfitting.

5.3 Evaluation on monocular images

This section evaluates the performance of the proposed model using the self-generated optical monocular image dataset. The experiment has two objectives: first to establish a performance baseline for comparison with the event model, and secondly to identify the limitations of monocular RGB, especially in challenging scenarios like low-light conditions and weakly textured objects. The model was trained on 48,000 monocular images and evaluated on a separate test set comprising 4,800 monocular images (48 sequences) under varying lighting conditions. The position and attitude error distributions are presented for each image sequence as errorbars in Fig.9. Each bar represent the mean and standard deviation of pose errors. The figure indicates that the maximum average translation error is 8.5 cm, while the maximum average orientation error is 7.9° across all sequences. To assess precision, the percentages of samples meeting the following criteria were calculated: translation error below 6 cm (10% of the target maximum diameter), average orientation error below 5° . The statistical threshold for orientation error was appropriately relaxed to account for challenging low-light and low-texture images in the test set. The quantitative evaluation results are reported in Table 1.

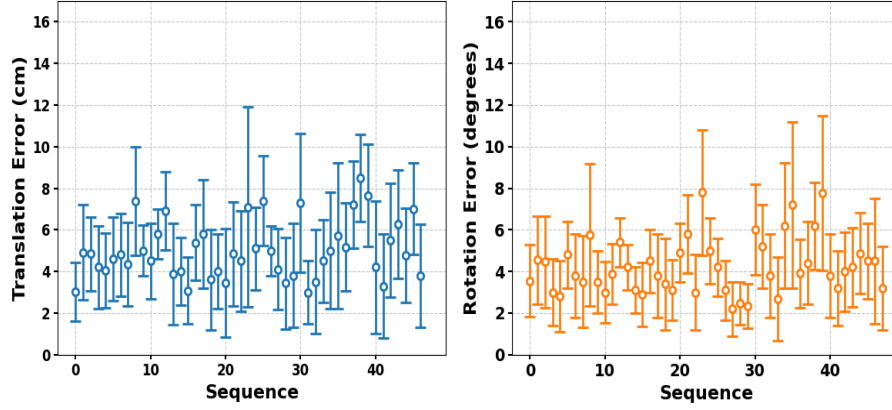


Figure 9: Optical images-based error distribution of the pose estimation results.

This study analyzes a sequence of 100 frames captured under low-light conditions. Figure 10 presents sample images annotated with the predicted 2D keypoints and the corresponding translation and angular errors. The average translation error is 6.2 cm, and the average angular error is 5.8° . The limited precision is primarily caused by low-light conditions, where optical RGB images may lose texture, edges, and other object details due to insufficient exposure, thereby impairing the deep learning model’s ability to extract meaningful features.

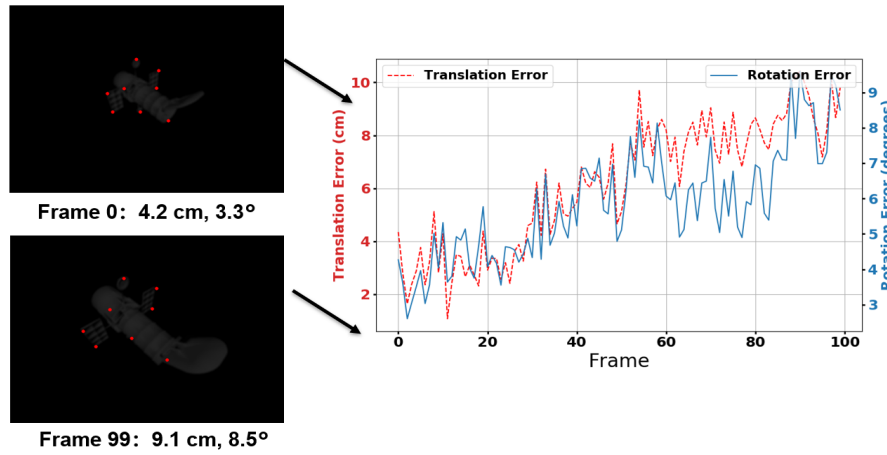


Figure 10: Relative pose errors (ψ and ϕ) over optical frames for one sequence from our dataset

5.4 Evaluation on Event images

This section evaluates the performance of the proposed model on the self-generated event dataset and compares it with the monocular optical-based model described in Sec.5.3. The model was trained on 48,000 event streams and tested on a separate set of 4,800 event streams. The distributions of position and attitude errors are illustrated in Fig.11. Most sequences exhibit translation errors between 2 and 6 cm and attitude errors ranging from 2° to 5° , even in the challenging lighting conditions. Compared with the monocular-based experiments in Sec.5.3, the event-based model demonstrates significantly improved performance. Similarly, various pose precision metrics are computed and compared with those reported in Sec.5.3. The quantitative results are summarized in Table 1. The proposed event-based approach achieves an 84% high-precision estimation rate, with a 8% improvement over the monocular-based model.

We also analyze a specific event sequence for the same conditions as in sec.5.3. Figure 12 presents two different modality images captured under identical conditions. Under low illumination, the optical RGB image loses texture details and exhibits blurred edge geometry. In contrast, the event image clearly preserves the target’s contour structure despite the extreme lighting, providing stable spatial structure information that aids the network in localizing keypoints. Figure 13 shows that its average translation error is 3.6 cm and its average attitude error is 3.2° . Under identical low-light conditions, the event-based method significantly outperforms the traditional monocular RGB-based approach in pose estimation accuracy and robustness.

EVENT-BASED 6D POSE ESTIMATION FOR NON-COOPERATIVE SPACECRAFT

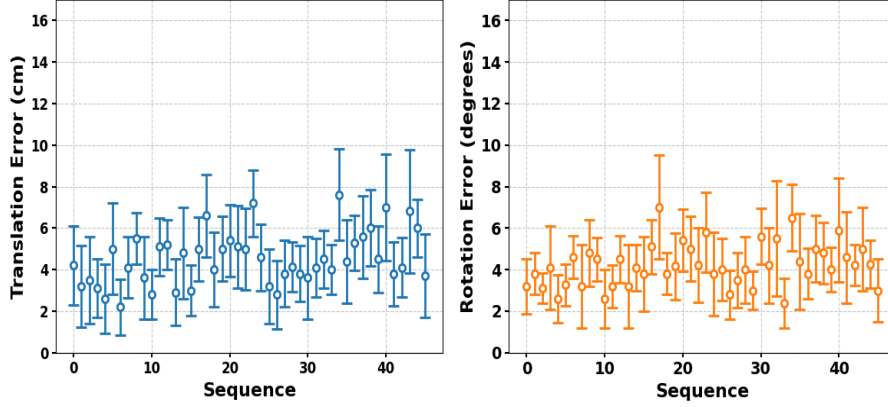


Figure 11: Event images-based error distribution of the pose estimation results.

Table 1: Quantitative evaluation of the proposed Event-based method and comparison with optical-based method. Correct estimation of all groups: rotation error $\leq 5^\circ$, translation error $\leq 6cm$

Method	$R \leq 5^\circ$	$T \leq 6cm$	$R \leq 5^\circ$ and $T \leq 6cm$
Optical-based	80.5%	78.0%	76.0%
Event-based	86.0%	87.5%	84.0%
Δ Pose	+5.5%	+9.5%	+8%

Figure 14 illustrates the predicted 2D keypoints. The results show that the proposed CNN-Transformer hybrid architecture accurately localizes target keypoints in event images. Moreover, the event-based approach maintains robust performance under extreme lighting conditions, demonstrating the effectiveness and reliability of our method for estimating the 6D pose of non-cooperative space targets.

To further validate the superiority of the proposed method, we compare it against two state-of-the-art event-based spacecraft pose estimation approaches: [34] and [38]. Method [34] adopts a two-stage CNN-based pipeline. Specifically, it uses Faster R-CNN [25] for target region detection, HRNet [37] for 2D keypoints regression from event images, and then computes the 6D pose using 2D-3D correspondences via the Perspective-n-Point (PnP) algorithm. Since our dataset does not contain complex backgrounds such as the Earth, object detection modules are not integrated into our proposed pipeline. For a fair comparison, we removed the object recognition module from the framework in [34] and retained the remaining components to perform comparative experiments. The model was trained on our event dataset using the hyperparameters reported in [34], and the pose estimation results were analyzed. Method [38] proposes an end-to-end framework that integrates CNNs and Transformers. It uses ResNet-34 as the feature extractor for event images, followed by a Transformer encoder to model global contextual information, and directly regresses the 6D pose via a multi-layer perceptron (MLP). The model was trained using the hyperparameters reported in [38] on our generated event dataset. Finally, several event sequences with varying motion trajectories and lighting conditions were selected from the test set to compare and analyze the three methods. The comparison results for pose estimation accuracy are summarized in Table 2, which reports the average error for each sequence (100 frames).

As shown in Table 2, the method proposed in this study achieves superior pose estimation accuracy compared to the other two approaches. In contrast to the pure convolutional network approach in [34], the proposed CNN-Transformer hybrid architecture effectively integrates the local feature extraction capability of CNNs with the global contextual modeling strengths of Transformers. This design enables more effective learning of the spatial-temporal characteristics of event streams, resulting in higher accuracy than CNN-only methods. Although the framework in [38] also employed a hybrid CNN-Transformer architecture, its end-to-end design that directly regresses 6D poses exhibited inferior performance compared to two-stage methods that integrate neural networks with geometric optimization. Moreover, in the satellite pose estimation challenge, those two-stage methods based on PnP demonstrated significantly higher accuracy [6].

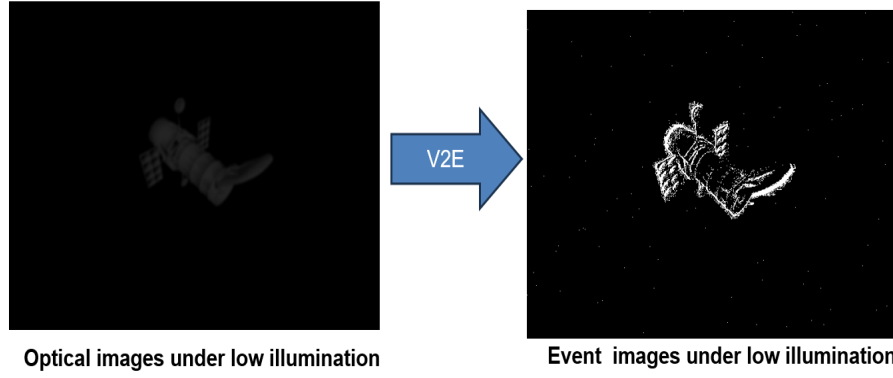


Figure 12: Comparison of optical and event images under low illumination conditions.

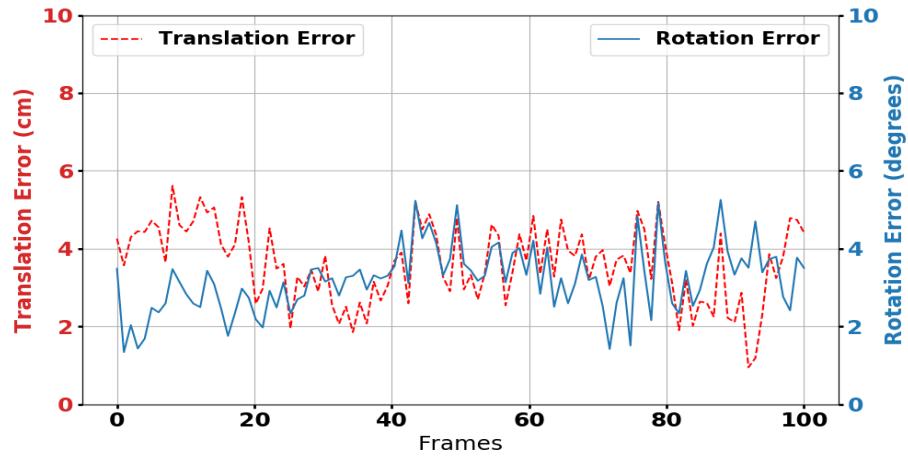
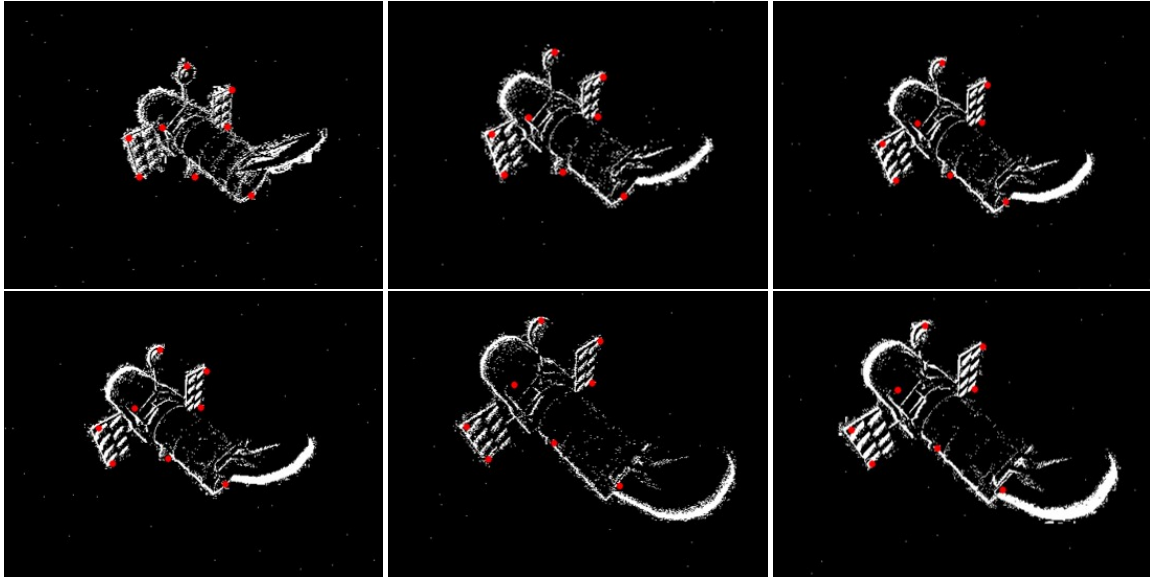
Figure 13: Relative pose errors (ψ and ϕ) over event frames for one sequence from our dataset

Figure 14: Sample of event frames with the predicted 2D keypoints

6. Conclusions and future works

This study proposes a novel 6D pose estimation method for non-cooperative space targets using event cameras, aiming to improve accuracy and robustness under extreme lighting conditions commonly encountered in space. An hybrid

Table 2: Quantitative evaluation of our event-based model on the generated dataset and comparison with other state-of-the-art methods.

Trajectory	Scene	Method [34]		Method[38]		Our Method	
		Φ (cm)	Ψ ($^{\circ}$)	Φ (cm)	Ψ ($^{\circ}$)	Φ (cm)	ψ ($^{\circ}$)
approach	normal-light	5.8	2.2	3.3	3.7	2.2	2.5
	low-light	6.6	3.2	3.5	4.2	3.8	3.4
orbit	normal-light	6.3	3.0	4.2	3.8	3.2	3.0
	low-light	7.0	3.5	4.8	4.5	4.1	3.9
Average		6.425	2.975	3.950	4.050	3.315	3.200

network architecture is designed to regress Gaussian keypoints heatmaps, combining CNNs and Transformers, enabling CNNs to extract detailed local features while Transformers handle global contextual information. Our network framework leverages the self-attention mechanism to capture long-range dependencies of event streams efficiently. Furthermore, the proposed framework effectively integrates the feature learning capabilities of deep neural networks with the strengths of geometric algorithms for accurate pose estimation. To support this research, the E-M-S dataset is developed as the publicly available synthetic event-monocular image dataset for spacecraft. It includes optical RGB images under varied lighting conditions, corresponding event streams, and precise pose labels.

Through a series of comprehensive experiments, this study validated the performance and advantages of the proposed method. The baseline experiments reveal the limited performance of optical image in low-light conditions, while the proposed event-based method shows significant improvements, particularly under challenging lighting scenarios. On the E-M-S dataset, the proposed method achieves a high-precision rate of 84%, representing a 8% improvement over the optical-based model. Even for highly challenging targets with weak textures, most test results achieve translation errors below 6 cm and angular errors under 5° . Compared to other state-of-the-art event-based spacecraft pose estimation methods, the proposed approach achieves superior performance even under extreme lighting conditions, with an average translational error of 3.3 cm and an average angular error of 3.2° . These results demonstrate that the proposed event camera-based approach and hybrid network architecture can accurately and robustly estimate the pose of non-cooperative targets.

This research focuses on exploring the performance of event streams for pose estimation in challenging space environments. Despite its promising performance, several limitations of the current framework remain. The experimental scenarios in this study excluded complex backgrounds, such as Earth, which could potentially interfere with pose estimation. Future work could introduce object recognition networks into our pipeline for target segmentation. Moreover, this study does not incorporate additional pose refinement methods for optimizing the estimated pose. Future work will focus on integrating pose refinement techniques to further improve pose accuracy. Another promising direction is to combine the proposed pipeline with filtering-based navigation algorithms (e.g., Kalman filters), which could directly benefit the development of autonomous navigation in space applications.

References

- [1] Thomas O Dixon, Stephen A Giles, and Alex A Gorodetsky. Predicting uncertainty in vision-based satellite pose estimation using deep evidential regression. *Aerospace Science and Technology*, page 110055, 2025.
- [2] Cuicui Jiang and Qinglei Hu. Constrained kalman filter for uncooperative spacecraft estimation by stereovision. *Aerospace Science and Technology*, 106:106133, 2020.
- [3] Yishi Wang, Zexu Zhang, Yefei Huang, and Yu Su. High precision pose estimation for uncooperative targets based on monocular vision and 1d laser fusion. *The Journal of the Astronautical Sciences*, 71(5):43, 2024.
- [4] Roberto Opromolla, Giancarmine Fasano, Giancarlo Rufino, and Michele Grassi. A review of cooperative and uncooperative spacecraft pose determination techniques for close-proximity operations. *Progress in Aerospace Sciences*, 93:53–72, 2017.
- [5] Zhuang Zhang, Rujin Zhao, Enhai Liu, Kun Yan, Yuebo Ma, and Yunze Xu. A fusion method of 1d laser and vision based on depth estimation for pose estimation and reconstruction. *Robotics and Autonomous Systems*, 116:181–191, 2019.

- [6] Zi Wang, Zhuo Zhang, Xiaoliang Sun, Zhang Li, and Qifeng Yu. Revisiting monocular satellite pose estimation with transformer. *IEEE Transactions on Aerospace and Electronic Systems*, 58(5):4279–4294, 2022.
- [7] Ao Xiang, Lin Zhang, and Li Fan. Shadow removal of spacecraft images with multi-illumination angles image fusion. *Aerospace Science and Technology*, 140:108453, 2023.
- [8] Yue Liu, Shijie Zhang, Niccolo Faraco, Michele Maestrini, Chao Zhang, and Pierluigi Di Lizia. A factor-ignored module for easy detection of uncooperative spacecraft using small training samples. *IEEE Aerospace and Electronic Systems Magazine*, 2024.
- [9] Jason L Forshaw, Guglielmo S Aglietti, Nimal Navarathinam, Haval Kadhém, Thierry Salmon, Aurélien Pisseloup, Eric Joffre, Thomas Chabot, Ingo Retat, Robert Axthelm, et al. Removedebbris: An in-orbit active debris removal demonstration mission. *Acta Astronautica*, 127:448–463, 2016.
- [10] Gabriella Gaias and J-S Ardaens. Design challenges and safety concept for the avanti experiment. *Acta Astronautica*, 123:409–419, 2016.
- [11] Siddarth Kaki and Maruthi R Akella. Pose terrier–image-based pose estimation and filtering for spacecraft applications. In *2021 AAS/AIAA Astrodynamics Specialist Conference*, 2021.
- [12] Jinghao Wang, Yang Li, Zhang Li, Zi Wang, and Qifeng Yu. Monocular satellite pose estimation based on uncertainty estimation and self-assessment. *IEEE Transactions on Aerospace and Electronic Systems*, 2024.
- [13] Massimo Piazza, Michele Maestrini, and Pierluigi Di Lizia. Monocular relative pose estimation pipeline for uncooperative resident space objects. *Journal of Aerospace Information Systems*, 19(9):613–632, 2022.
- [14] Tae Ha Park, Marcus Mörtens, Mohsi Jawaid, Zi Wang, Bo Chen, Tat-Jun Chin, Dario Izzo, and Simone D’Amico. Satellite pose estimation competition 2021: Results and analyses. *Acta Astronautica*, 204:640–665, 2023.
- [15] Thaweerath Phisannupawong, Patcharin Kamsing, Peerapong Torteeka, Sittiporn Channumsin, Utane Sawangwit, Warunyu Hematulin, Tanatthep Jarawan, Thanaporn Somjit, Soemsak Yooyen, Daniel Delahaye, et al. Vision-based spacecraft pose estimation via a deep convolutional neural network for noncooperative docking operations. *Aerospace*, 7(9):126, 2020.
- [16] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and nonlinear pose refinement. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
- [17] Lorenzo Pasqualetto Cassinis, Alessandra Menicucci, Eberhard Gill, Ingo Ahrns, and Manuel Sanchez-Gestido. On-ground validation of a cnn-based monocular pose estimation system for uncooperative spacecraft: Bridging domain shift in rendezvous scenarios. *Acta Astronautica*, 196:123–138, 2022.
- [18] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pages 534–542. IEEE, 2020.
- [19] Anh Nguyen, Thanh-Toan Do, Darwin G Caldwell, and Nikos G Tsagarakis. Real-time 6dof pose relocalization for event cameras with stacked spatial lstm networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [20] Zibin Liu, Banglei Guan, Yang Shang, Qifeng Yu, and Laurent Kneip. Line-based 6-dof object pose estimation and tracking with an event camera. *IEEE Transactions on Image Processing*, 2024.
- [21] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.
- [22] Yefei Huang, Zexu Zhang, Hutao Cui, and Liang Zhang. A low-dimensional binary-based descriptor for unknown satellite relative pose estimation. *Acta Astronautica*, 181:427–438, 2021.
- [23] Albert Garcia, Mohamed Adel Musallam, Vincent Gaudilliere, Enjie Ghorbel, Kassem Al Ismaeil, Marcos Perez, and Djamila Aouada. Lspnet: A 2d localization-oriented spacecraft pose estimation neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2048–2056, 2021.

- [24] Sumant Sharma and Simone D' Amico. Neural network-based pose estimation for noncooperative spacecraft rendezvous. *IEEE Transactions on Aerospace and Electronic Systems*, 56(6):4638–4658, 2020.
- [25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- [26] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [27] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [28] Fengyi Liu, Zhujun Zhang, and Sijue Li. Dtse-spacenet: Deformable-transformer-based single-stage end-to-end network for 6-d pose estimation in space. *IEEE Transactions on Aerospace and Electronic Systems*, 60(3):2555–2571, 2023.
- [29] Pavel Rojtborg and Thomas Pöllabauer. Ycb-ev: Event-vision dataset for 6dof object pose estimation. *arXiv preprint arXiv:2309.08482*, 2023.
- [30] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023.
- [31] William Chamorro, Joan Sola, and Juan Andrade-Cetto. Event-based line slam in real-time. *IEEE Robotics and Automation Letters*, 7(3):8146–8153, 2022.
- [32] Elias Mueggler, Basil Huber, and Davide Scaramuzza. Event-based, 6-dof pose tracking for high-speed maneuvers. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2761–2768. IEEE, 2014.
- [33] Ahmed Tabia, Fabien Bonardi, and Samia Bouchafa-Bruneau. Fully convolutional neural network for event camera pose estimation. In *VISIGRAPP (4: VISAPP)*, pages 594–599, 2023.
- [34] Mohsi Jawaaid, Ethan Elms, Yasir Latif, and Tat-Jun Chin. Towards bridging the space domain gap for satellite pose estimation using event sensing. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11866–11873. IEEE, 2023.
- [35] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1312–1321, June 2021.
- [36] Yanxiang Wang, Xian Zhang, Yiran Shen, Bowen Du, Guangrong Zhao, Lizhen Cui, and Hongkai Wen. Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3436–3449, 2021.
- [37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.
- [38] Wang Yishi, Michele Maestrini, Zhang Zexu, Mauro Massari, and Pierluigi Di Lizia. Cross-modal fusion of monocular images and neuromorphic streams for 6d pose estimation of non-cooperative targets. *Aerospace Science and Technology*, page 110338, 2025.
- [39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [40] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [41] Ethan Elms, Mohsi Jawaaid, Yasir Latif, and Tat-Jun Chin. Seenic: dataset for spacecraft pose estimation with neuromorphic vision, October 2022.
- [42] Y. Wang, M. Michele, M. Mauro, D. L. Pierluigi, and Z. Zhang. E-m-s: Event-monocular-spacecraft dataset for 6d pose estimation of non-cooperative targets, 2025.
- [43] E. Summers and M. Horn. Hubble space telescope. <https://nasa3d.arc.nasa.gov/detail/HST>.