SPACE-HOP: Spacecraft 6-DoF Pose Estimation using Embedding-Predictive Pretraining and Hopf Map

Anonymous CVPR submission

Paper ID 36

Abstract

Vision-based methods capable of determining the ***6-Degrees of Freedom (6-DoF) pose of a spacecraft*** *are crucial for rendezvous, proximity operations, and docking. These operations involve a series of orbital maneuvers required to achieve orbital synchronization and precise orientation alignment. Despite significant progress, existing methods exhibit two major limitations. First, reconstruction-driven objectives often prioritize pixel-level fidelity, neglecting the physically meaningful geometric structure embedded in semantic representations. Second, deterministic pose regression methods are inherently sensitive to noise and prone to convergence instability. Although continuous probabilistic modeling such as the Matrix Fisher distribution captures spatial uncertainty, its computational complexity is impractical.*

To address these challenges, we propose SPACE-HOP which leverages masking-based self-supervised embedding-predictive method for spacecraft pose estimation. Our method pre-trains an encoder-predictor Vision Transformer to learn geometry-aware representations. We then present a novel method of estimating rotation as a classification task over a discrete distribution of the $SO(3)$ manifold, and further reduce quantization error to mimic a continuous manifold. The core mechanism involves classifying the orientation across a uniform Hopf fibration grid of $SO(3)$ anchors and achieving the associated accurate offsets. We demonstrate that our method despite being a foundational study, performs comparable with SOTA regression models, without extensive external pretraining or Test-Time Adaptation.

1. Introduction

The next generation of scalable space infrastructure demands autonomous spacecraft capable of performing rendezvous and terminal docking without continuous ground intervention. Traditional ground-teleoperated rendezvous becomes increasingly impractical in deep-space and large-scale orbital operations due to communication latency, and signal occultation. Autonomous close-range operations are therefore essential for emerging applications such as Active Debris Removal (ADR), On-Orbit Servicing, Assembly, and Manufacturing (OSAM), and autonomous propellant transfer. A core requirement for this autonomy is the ability of a spacecraft to precisely calculate the 6-DoF pose of an uncooperative target using only monocular vision.

The existing deep learning approaches (like, voting-based methods [13]) for 6-DoF pose estimation achieve suitable performance but rely on computationally expensive and non-differentiable Perspective-n-Point (PnP) solvers. These iterative geometric post-processing steps limit deployability on edge hardware. Template-matching and reference-view approaches [7] fail under extreme illumination and self-shadowing of space due to their dependence on strict photometric consistency. Similarly, shape-prior frameworks [18] struggle to generalize to out-of-distribution (OOD) geometries. Meanwhile, direct feedforward regression offers a low-latency alternative, yet standard training objectives are often reconstruction-driven, encouraging pixel-level fidelity rather than learning geometry-aware latent structure [16].

To this end, we propose SPACE-HOP, a geometry-aware framework for 6-DoF spacecraft pose estimation. Our method combines masking-based joint-embedding predictive pretraining [3] with a structured discretization of the $SO(3)$ rotational manifold via the Hopf fibration [8]. Instead of regressing rotation parameters directly, we reformulate rotation estimation as classification over a uniformly constructed grid of anchor rotations, followed by continuous refinement. This discrete-to-continuous formulation introduces bounded optimization, stabilizes training, and preserves manifold structure without incurring the computational burden of probabilistic continuous models. Our contributions are summarized as follows:

- We introduce **joint-embedding predictive pretraining into the spacecraft pose estimation domain**, demonstrating its suitability for geometry-centric representation learning under orbital imaging conditions.

- **A keypoint-anchored pretraining strategy** that uses geometrically guided masking. By enforcing the network to predict the latent embeddings of key structural features (like antennas and corners), the model learns 3D geometry instead of 2D pixel textures.
- A **Classification-and-Refinement Rotation Head** that avoids the numerical instability of direct 6D coordinate regression. It classifies the pose across a predefined Hopf Fibration grid and refines it using continuous Lie algebra offsets for sub-degree accuracy.
- An **Inherently Low-Latency Architectur**e optimized for the edge. By bypassing iterative geometric solvers like PnP and relying purely on a single-pass feedforward ViT with lightweight linear heads, SPACE-HOP guarantees fast inference speeds suitable for real-time orbital deployment. With a ViT latency of 18.0 ± 3.8 **ms** under FP32 configuration on a **Jetson Orin**.

We evaluate our method on the SPEED+ [10] dataset. Despite being a foundational study trained entirely from scratch, SPACE-HOP achieves accuracy on par with state-of-the-art direct regression models and remains competitive with computationally intensive PnP-based solvers.

2. Related Work

2.1. Deep Learning for Satellite Pose Estimation

Existing architectures for (6-DoF) pose estimation are generally categorized into two distinct paradigms based on how they bridge the gap between 2D image observations and 3D spatial coordinates:

Geometric Solver Methods [11, 15]: These hybrid approaches utilize deep neural networks (DNNs) as feature extractors to localize semantic keypoints in the 2D image plane. These are then mapped to a known 3D satellite model using classical geometric solvers, such as the Perspective-n-Point (PnP) algorithm, often wrapped within a RANSAC loop to handle outliers. While highly accurate, these methods rely heavily on the availability of a precise 3D CAD model and offer high latency. **Direct Regression Methods** [1]: These frameworks leverage a unified network to directly map input images to pose parameters, typically represented as a translation vector $\mathbf{t} \in \mathbb{R}^3$ and an orientation represented as a unit quaternion $\mathbf{q} \in \mathbb{H}$. By bypassing the need for explicit 2D-3D correspondences or intermediate geometric optimization, these methods significantly reduce computational overhead, hence more suited for satellite pose estimation.

2.2. Vision Transformers for Pose Estimation

Vision Transformers are generally used as a pretrained backbone for the hybrid solvers, or more often used in direct methods as end to end pipelines [1] [6] where the learned embeddings are mapped by regression heads to estimate the rigid body transformation from camera to rigid body frame

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathrm{SE}(3), \tag{1}$$

with $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$.

While Vision Transformers (ViTs) excel at extracting high-level semantic features, they are inherently designed for spatial invariance. This presents a challenge for 6-DoF pose estimation, which requires SE(3) equivariance—the ability for internal representations to transform proportionally with object movement.

2.3. Joint Embedding Predictive Architecture

Joint-Embedding Predictive Architectures (JEPA) provide a self-supervised learning framework that learns to predict latent embeddings. Image-based JEPA (or I-JEPA) [2] focuses on learning predictive representations directly in latent space without image reconstruction. Given an image I, a context encoder f_θ processes a visible subset of patches I_c to produce a contextual representation $\mathbf{z}_c = f_\theta(I_c)$, while a target encoder f_ξ encodes a masked target region I_t to obtain $\mathbf{z}_t = f_\xi(I_t)$. A predictor g_ϕ is trained to estimate $\mathbf{z}_t$ from $\mathbf{z}_c$ by minimizing a feature-space discrepancy $\mathcal{L} = d(g_\phi(\mathbf{z}_c), \mathbf{z}_t)$. By operating purely in embedding space, I-JEPA encourages the model to capture high-level semantic and structural regularities rather than low-level texture statistics, producing representations that exhibit improved abstraction, robustness, and transferability to downstream tasks.

RoboPEPP [4] extends I-JEPA to robotic perception by learning geometry-aware predictive representations conditioned on both visual observations and robot state. To incorporate robot kinematics, the model additionally conditions on joint configuration $\mathbf{q} \in \mathbb{R}^n$, where n denotes the number of joints, This joint visual-kinematic modeling encourages embeddings that are both semantically meaningful and physically grounded for 6-DoF pose estimation. We extend this idea to Satellite Pose Estimation, where we deal with lack of kinematics metadata, occlusion, motion blur, light reflections, symmetrical objects, etc. JEPA architecture is suitable in such conditions, because it does not concern itself with the task of pixel reconstruction and ignores low level textures.

2.4. Rotation Representations and Probabilistic Estimation

Mapping visual features to the $SO(3)$ manifold introduces significant geometric challenges. Traditional Euler angles and unit quaternions suffer from gimbal lock and antipodal ambiguity, while continuous 6D representations [20] risk exploding gradients during Gram-Schmidt orthogonalization if predicted vectors are near-collinear.For handling spatial uncertainty, the theoretically ideal approach is the

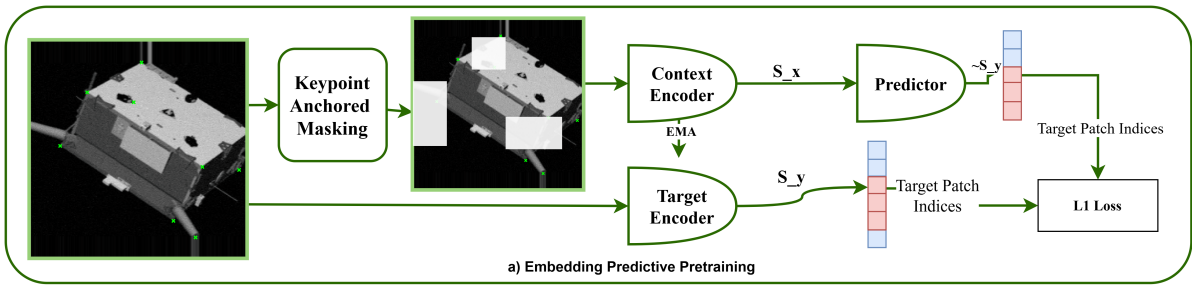


Figure 1. **Embedding-Predictive Pretraining with Keypoint-Anchored Masking**: Given a spacecraft image, target patches are selected around projected structural keypoints (e.g., corners, antennas), while the remaining visible patches form the context. A Context Encoder processes unmasked patches, and a Target Encoder extracts latent features for masked patches. A lightweight transformer predicts masked embeddings from context embeddings and mask tokens. Training uses Smooth-L1 loss in embedding space, promoting geometry-aware learning without pixel reconstruction.

Matrix Fisher distribution, which parameterizes anisotropic confidence via a 3×3 matrix. However, the eigenvalue decompositions required to compute its normalizing constants during the forward pass render it computationally prohibitive for edge deployment.While adapting Matrix Fisher frameworks for low-latency orbital hardware remains an area of our ongoing research interest, approximating uncertainty over a discrete $SO(3)$ grid currently serves as the most viable practical alternative [5, 9]. To construct this grid, we discretize the $SO(3)$ manifold using the Hopf fibration [19], which rigorously maps S^3 to S^2 with S^1 fibers. This enables a perfectly uniform sampling of the rotation space that prevents polar clustering and entirely circumvents the numerical vulnerabilities of continuous regression.

3. Methodology

Problem Description: Given a monocular RGB image capturing a rigid satellite, we treat the target as a rigid body defined by a set of N semantic keypoints, $\mathcal{P} = \{p_i\}_{i=1}^N \subset \mathbb{R}^3$, in canonical frame of the satellite. Our objective is to estimate the relative pose, represented by the transformation matrix $T_S^C \in SE(3)$, which maps the satellite frame to the camera coordinate system. The transformation consists of a rotation $R \in SO(3)$ and a translation $t \in \mathbb{R}^3$. We assume the camera intrinsic matrix $K \in \mathbb{R}^{3\times3}$ is calibrated and known. **Method Overview:** We leverage the joint-embedding predictive architecture as a self-supervised pretraining paradigm for training the encoder for monocular satellite pose estimation. We also introduce an architecture using Hopf fibrations, for SO(3) Representation, to stream down the regression task into an n-class classification task, which we refine using offsets on SO(3) lie group.

3.1. Encoder Pretraining

The foundational representation of our model is learned through a self-supervised task following the work of I-JEPA [2]. Given a context block, the objective is to predict the representations of various target blocks of the same image. We train the encoder on the SPEED+ dataset, constituting of 59,960 images and using visual and spatial augmentations.

3.1.1. Image Partitioning and Geometric Masking

We partition the input image $I \in \mathbb{R}^{H\times W\times3}$ into a set of N non-overlapping patches $\mathcal{U} = \{u_i\}_{i=1}^N$, where each patch $u_i \in \mathbb{R}^{p\times p\times3}$ and $p = 16$ denotes the patch size. The masking operation is defined by an index set $\mathcal{M} \subset \{1, \dots, N\}$, which identifies the target patches to be predicted. We denote the collection of these target patches as $\mathcal{B} = \{u_j \mid j \in \mathcal{M}\}$. The complement set, $\mathcal{O} = \{1, \dots, N\} \setminus \mathcal{M}$, defines the indices of the unmasked context patches, denoted as $\mathcal{V} = \{u_k \mid k \in \mathcal{O}\}$.

In our proposed framework, the selection of the index set $\mathcal{M}$ is not stochastic; rather, it is geometrically guided by the spacecraft's topology, a principle recently demonstrated to significantly enhance physical model understanding in articulated robotics [4]. Given the set of K projected keypoint coordinates $\mathcal{K} = \{k_n\}_{n=1}^K$ in the image plane, we define the target index set $\mathcal{M}$ based on the spatial intersection of keypoints and patches. For each selected keypoint $k_n = (x_n, y_n)$, we sample a target block of size (h, w) constrained by:

$$y_{top} \sim \mathcal{U}(\max(0, \lfloor\frac{y_n}{p}\rfloor - h), \min(H_{grid} - h, \lfloor\frac{y_n}{p}\rfloor + 1)) \tag{2}$$

$$x_{left} \sim \mathcal{U}(\max(0, \lfloor\frac{x_n}{p}\rfloor - w), \min(W_{grid} - w, \lfloor\frac{x_n}{p}\rfloor + 1)) \tag{3}$$

where H_{grid} and W_{grid} are the dimensions of the patch grid. By constructing masking protocol such that $\mathcal{B} \cap \mathcal{V} = \emptyset$, we force the predictive architecture to reconstruct the latent embeddings of the structural vertices $\mathcal{B}$ solely from the spatially distant context $\mathcal{V}$. This constraint encourages the

model to internalize the underlying rigid-body geometry of the satellite rather than relying on local pixel correlations.

3.1.2. Predictive Latent Architecture

The framework utilizes three specific modules:

a. **Context Encoder (E_θ):** Uses context patches $\mathcal{V}$ to produce context embeddings z_{ctxt}.

b. **Target Encoder (E_ϕ):** Extracts target embeddings $\bar{v}_i \in \mathbb{R}^d$ for $i \in \mathcal{M}$ from the original image. The weights ϕ are updated via an Exponential Moving Average (EMA) of the context encoder weights

$$\theta : \phi_t = m_t \phi_{t-1} + (1 - m_t)\theta_t \tag{4}$$

where the momentum coefficient m_t linearly increases from 0.996 to 1.0 during training.

c. **Predictor ViT (P_ψ):** A modified vision transformer that accepts z_{ctxt} and learnable mask tokens to predict the representations v_i for all masked patches.

3.1.3. Pretraining Objective

The model is supervised by a Smooth-L_1 loss in the latent representation space. Embeddings of the masked patches represented as $\bar{\mathcal{M}}$ are used to compute the error:

$$\mathcal{L}_{\text{pre-train}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|v_i - \bar{v}_i\|_1 \tag{5}$$

Using this keypoint-anchored approach, we shift the focus of the standard ViT Encoder from the whole satellite to the specific keypoint locations critical for inferring the pose of the rigid satellite, such as the corners and antennas Fig. 2. This mimics how humans perceive objects while performing tasks of picking or moving an object, prioritizing structural vertices over surface texture.

3.2. 6D Pose over a Discrete $SO(3)$ Manifold

We formulate rotation estimation as a classification-and-refinement task over a discretized SO(3) manifold (Fig. 3). To address the discontinuity of a discrete space, which would not allow pose predictions to be decimal-accurate as the error will never be able to converge within the gaps, we train an offset refinement to cover these gaps.

3.2.1. SO(3) Manifold Discretization

To ensure a uniform distribution of orientations across the rotation space, we generate a predefined grid of K anchor rotations $R_k \in SO(3)$ using the Hopf fibration principle [19]. We first distribute N points on a (S^2) sphere using a Fibonacci lattice to define base pointing directions. For each direction, we construct an orthonormal frame and apply M discrete in-plane rotations around the z-axis. This approach provides a dense, unbiased sampling of the manifold, avoiding the gimbal lock and non-uniformity.

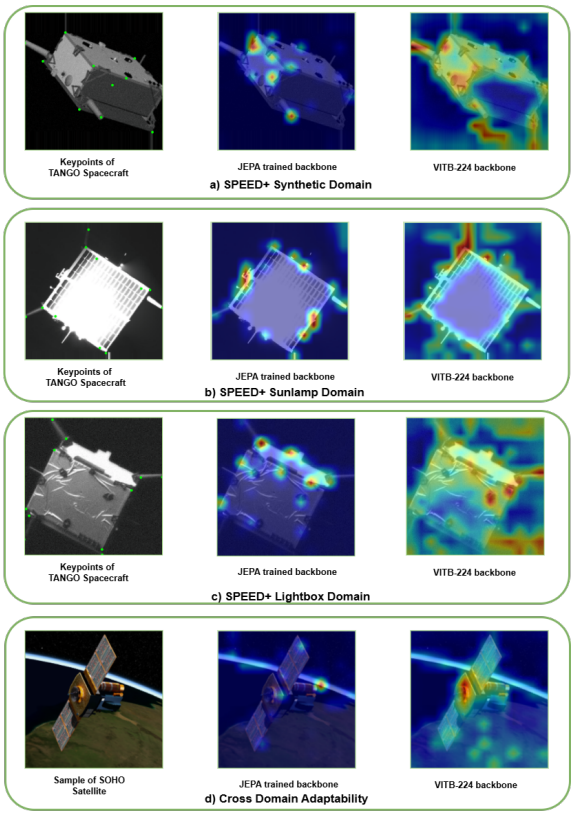


Figure 2. **Comparison of attention maps** produced by a JEPA-pretrained backbone and a supervised ViT-B/224 backbone across multiple domains of the SPEED+ dataset. From left to right in each panel: input image with projected keypoints, attention map from the JEPA-trained encoder, and attention map from the supervised ViT backbone. The JEPA-pretrained model exhibits stronger concentration around structurally meaningful regions (e.g., corners, antennas, and solar panels), while the supervised backbone displays more texture-driven activations. In (d) while the supervised ViT-B focuses on the whole satellite as well as the background planet, our model tries to locate the potential keypoints and geometric features.

3.2.2. Parallel Linear Heads

The embedding predictive pre-trained Vision Transformer Fig. 1 divides an image into non-overlapping patches, projects the patches into embeddings called "tokens," and prepends a learnable [CLS] token to aggregate global information via self-attention. We extract this [CLS] token and pass it into three parallel linear heads:

- **Translation Head:** This linear layer maps the [CLS] to a normalized 3D translation vector, $\mathbf{u} \in \mathbb{R}^3$. To recover the final translation $\mathbf{T} \in \mathbb{R}^3$ we process $\mathbf{u}$ through an Inverse Perspective Mapping Sec. 3.3, scaling it against the known camera intrinsic matrix $\mathbf{K}$ and the 2D bounding box crop parameters.

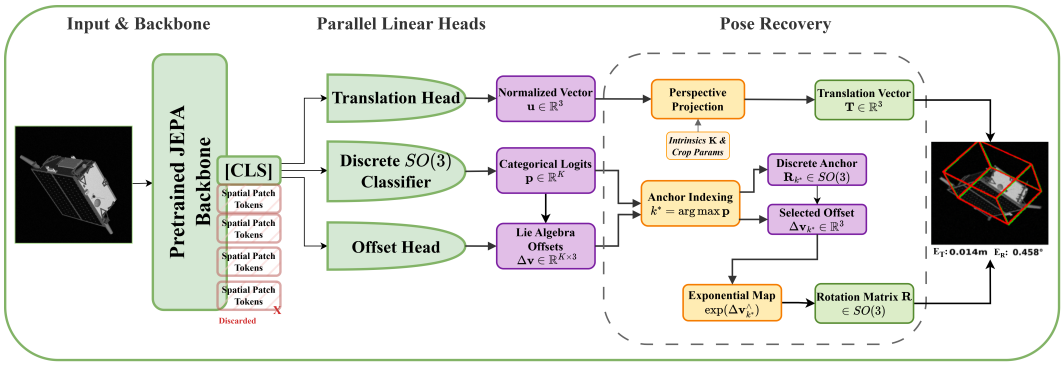


Figure 3. SPACE-HOP Diagram detailing the Input & Backbone, Parallel Linear Heads, and Pose Recovery stages.

- **Discrete $SO(3)$ Classifier:** To avoid the optimization challenges of continuous rotation spaces, this head projects the token into a categorical distribution, yielding logits $\mathbf{p} \in \mathbb{R}^K$ over the K predefined Hopf fibration anchors. The anchor corresponding to the highest probability, $k^* = \arg\max \mathbf{p}$, is selected as the discrete base rotation, $\mathbf{R}_{k^*} \in SO(3)$.
- **Tangent-Space Offset Refinement:** To cover the discretization gap and achieve sub-degree precision, we derive a set of continuous Lie algebra offsets, $\Delta\mathbf{v} \in \mathbb{R}^{K\times 3}$. Rather than attempting to map the overall target rotation, we estimates local 3D offsets along the tangent plane of the manifold. We index the specific offset corresponding to the selected anchor, $\Delta\mathbf{v}_{k^*}$, and project it back onto the curved manifold via the exponential map [17], $\exp(\Delta\mathbf{v}^{\wedge}_{k^*})$. The final continuous rotation matrix is:

$$\mathbf{R} = \mathbf{R}_{k^*} \exp(\Delta\mathbf{v}^{\wedge}_{k^*}) \tag{6}$$

We also experimented with deeper MLPs and GRN, but found no consistent improvement over the simple linear projection. This suggests that the JEPA pretrained ViT backbone already encodes sufficient global information.

3.3. Translation using Normalization of coordinates

Directly predicting translation coordinates $\mathbf{T} = [X, Y, Z]^\top$ caused scale ambiguity. We solve this by separating the depth estimation from the lateral shifts by mapping the translation to a normalized coordinate space, yielding the target vector $\mathbf{u} = [U_x, U_y, U_z]^\top$.

3.3.1. Depth Normalization

Given a 2D bounding box defined by its center and dimensions (b_x, b_y, b_w, b_h) relative to the original image dimensions (W_{full}, H_{full}), we calculate the scaling factors. To account for bounding box padding during the cropping phase, we introduce an empirical scaling scalar $\alpha = 1.6$:

$$s_x = \frac{W_{full}}{\alpha \cdot b_w}, \quad s_y = \frac{H_{full}}{b_h} \tag{7}$$

The absolute depth Z is normalized into U_z by averaging these scaling factors:

$$U_z = \frac{1}{2}\left(\frac{1}{s_x} + \frac{1}{s_y}\right) Z \tag{8}$$

(U_x, U_y) represent the normalized spatial offset between the projected 3D center of the spacecraft and the 2D bounding box center, scaled by the cropped box dimensions and the camera intrinsic focal lengths (f_x, f_y) and optical centers (c_x, c_y):

$$U_x = \frac{1}{b_w}\left(\frac{f_x X}{Z} + c_x - b_x\right) \tag{9}$$

$$U_y = \frac{1}{b_h}\left(\frac{f_y Y}{Z} + c_y - b_y\right) \tag{10}$$

This our one of our learnable targets.

3.3.2. Lateral Target

2D projection of the satellite's 3D center (x, y) rarely aligns perfectly with the geometric center of the cropped bounding box (b_x, b_y), so we must calculate

$$(\Delta x, \Delta y) = (x - b_x, y - b_y) \tag{11}$$

To make these independent of specific crop dimensions, we normalize them by the bounding box width b_w and height b_h, with the camera focal lengths (f_x, f_y) and optical centers (c_x, c_y):

$$U_x = \frac{\Delta x}{b_w} = \frac{1}{b_w}\left(\frac{f_x X}{Z} + c_x - b_x\right) \tag{12}$$

$$U_y = \frac{\Delta y}{b_h} = \frac{1}{b_h}\left(\frac{f_y Y}{Z} + c_y - b_y\right) \tag{13}$$

The Translation Head is trained to predict the scale-invariant vector $\mathbf{u}_{pred}$.

3.4. Objective Functions

Our objective function jointly supervises the discrete global orientation, the continuous local refinement, and the normalized translation.

3.4.1. Discrete SO(3) Classification

To supervise the discrete $SO(3)$ classifier, we must first identify the closest geometric anchor k^* to the target rotation $\mathbf{R}_{gt}$. This is achieved efficiently by maximizing the trace of the relative rotation matrix via Einstein summation across the entire Hopf fibration grid $\mathcal{H}$:

$$k^* = \arg\max_{k \in \mathcal{H}} \mathrm{Tr}(\mathbf{R}_{gt}\mathbf{R}_k^\top) \tag{14}$$

The classification objective is then enforced using standard Cross-Entropy over the predicted logits $\mathbf{p}$:

$$\mathcal{L}_{coarse} = -\log\left(\frac{\exp(p_{k^*})}{\sum_{i=1}^{K}\exp(p_i)}\right) \tag{15}$$

3.4.2. Tangent-Space Offset Refinement

To supervise the offset, we isolate the predicted 3D Lie algebra offset corresponding exclusively to the target bin, $\Delta\mathbf{v}_{pred,k^*}$. The ground-truth relative transformation from the selected anchor to the rotation is computed as

$$\mathbf{R}_{rel} = \mathbf{R}_{k^*}^\top \mathbf{R}_{gt} \tag{16}$$

which is subsequently mapped to its axis-angle vector representation $\Delta\mathbf{v}_{gt}$.

To manage outlier predictions and prevent exploding gradients of $SO(3)$ boundary conditions, we use a Smooth L1 (Huber) Loss with a strict $\beta = 0.01$ threshold. Let the offset prediction error be defined as $\mathbf{e} = \Delta\mathbf{v}_{pred,k^*} - \Delta\mathbf{v}_{gt}$. The fine loss is computed as:

$$\mathcal{L}_{fine} = \begin{cases} \frac{0.5}{\beta}\|\mathbf{e}\|_2^2 & \text{if } \|\mathbf{e}\|_1 < \beta \\ \|\mathbf{e}\|_1 - 0.5\beta & \text{otherwise} \end{cases} \tag{17}$$

3.4.3. Translation Estimation

The normalized translation vector $\mathbf{u}$ is trained via the squared Euclidean distance between the predicted coordinates and the ground truth:

$$\mathcal{L}_{trans} = \frac{1}{B}\sum_{i=1}^{B}\|\mathbf{u}_{pred} - \mathbf{u}_{gt}\|_2^2 \tag{18}$$

3.4.4. Total Multi-Task Objective

The final training objective is the unweighted sum of the translation and coarse classification errors, Regularized by weighting factor $\lambda_{fine} = 10.0$ applied to the offset:

$$\mathcal{L}_{total} = \mathcal{L}_{trans} + \mathcal{L}_{coarse} + \lambda_{fine}\mathcal{L}_{fine} \tag{19}$$

4. Experiments and Results

4.1. Evaluation Metrics

Evaluation Metrics. We use the official metrics from the ESA Pose Estimation Challenge. The translation error, E_T, is the Euclidean distance between the predicted and ground-truth translation vectors (T_{pred} and T_{gt}), i.e.,

$$E_T = \|T_{pred} - T_{gt}\|_2. \tag{20}$$

The rotation error, E_R, is the geodesic distance between the predicted and ground-truth orientations. For this metric, the rotations are represented as quaternions (q), and the error is calculated as

$$E_R = 2 \cdot \arccos\left(|\langle q_{pred}, q_{gt}\rangle|\right). \tag{21}$$

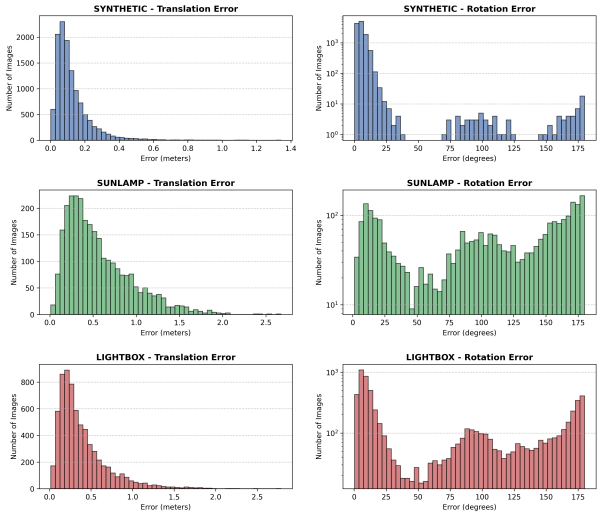


Figure 4. **Errors Distribution Histogram Across Domains**: Translation error (meters) remains tightly concentrated below 0.2 m in the Synthetic domain and largely below 0.5 m under real lighting, indicating robust spatial localization despite domain shift and severe photometric degradation. Rotation error (degrees, log-scale y-axis) exhibits a dominant peak near 0°, confirming effective coarse-to-fine refinement, and a secondary peak at 180°, reflecting the intrinsic axial symmetry ambiguity of the Tango spacecraft rather than geometric failure. Increased mid-range errors in Sunlamp highlight the impact of harsh directional illumination and shadow-induced ambiguity on initial anchor selection

4.2. Comparison with the State of The Art

We compare our models with previous published works [1, 10, 11, 15] on SPEED + [12], covering both PnP and Non PnP methods. SPEED+ consists of synthetic training data of 59,960 images and real captured test data, with a high domain shift from the synthetic data. Our baseline which is trained from scratch on SPEED+ using JEPA (Joint Embedding Predictive Architecture), performs on par with SOTA regression based methods which use pretrained pipelines of ViT-B, trained on more data.

SPACE-HOP thus provides evidence that predictive pretraining scales effectively to 6-DoF pose estimation.

Table 1. Result Comparison with SoTA on SPEED+ Lightbox (Sorted by E_T)

Method	Category	E_T [m]	E_R [°]
VPU [14]	PnP	0.076	**2.728**
SPNv2 [10]	PnP	0.161	9.788
SPACE-HOP-224 (Ours)	**Regression**	**0.302**	**9.015**
FastPose-ViT-224 [1]	Regression	0.460	28.017
SPN [15]	PnP	0.45	65.12
KRN [11]	PnP	2.25	44.53

Table 2. Comparison with state-of-the-art methods on SPEED+ Sunlamp.

Method	Category	E_T [m]	E_R [°]
VPU [14]	PnP	0.131	**4.577**
SPNv2 [10]	PnP	0.150	5.577
SPACE-HOP-224 (Ours)	**Regression**	**0.282**	**26.646**
FastPose-ViT-224 [1]	Regression	0.305	14.668
SPN [15]	PnP	0.65	92.95
KRN [11]	PnP	14.64	80.95

4.3. Ablation Study

4.3.1. Impact of JEPA Pretraining

Model Variant	E_T	E_R
JEPA-ViT-224	0.096	3.995
ViT-B-224 (scratch)	0.125	6.143

Table 3. Comparison of different ViT Backbones on SPEED+ Synthetic Val Domain

We pretrained a standard ViT-B from scratch on the SPEED+ dataset for reconstruction objective and compare it with the JEPA objective pretrained ViT. The JEPA pretrained ViT outperforms ViT-B on the same SPACE-HOP pipeline, as observed in Tab. 3.

The following shows that the geometric masking and the embedding predictive pretraining, results in a better result in the exact same scenario. The experiments do demonstrate the potential of Embedding Predictive Pretraining.

4.3.2. Impact of binned classification and offset refinement

We compare our method of formulating rotation estimation as a classification-and-refinement task to direct regression method by training both pipelines:

a. Direct Regression of SE(3) pose
b. SPACE-HOP's binned classification and offset refinement

over our pre-trained JEPA-ViT baseline Fig. 1 on SPEED+[10] dataset, keeping all augmentation strategies alike.

Method	E_T	E_R
SPACE-HOP	0.096	3.995
Direct Regression	0.173	12.324

Table 4. Comparison of regression to binned classification and refinement method on SPEED+ Synthetic Val Domain

4.4. Symmetry-Induced Instability and Jensen Gain Analysis

Spacecraft structures often exhibit strong geometric symmetries, particularly under in-plane rotations. In Fig. 4, we observe that the rotation error distribution forms a characteristic **W-shaped pattern**, with prominent peaks near 0° and 180°. This bimodal behavior indicates that the model frequently predicts orientations that are visually symmetry-consistent yet incorrect in absolute pose. Such failure modes arise when multiple rotations produce nearly indistinguishable image evidence, forcing the network to rely on weak asymmetries or spurious cues.

To quantify this instability, we introduce **Jensen Gain** as a diagnostic measure of symmetry breaking in the learned representation. Let $f(X)$ denote the predicted rotation for input image X. We generate an orbit of symmetry-preserving transformations (e.g., $K = 16$ in-plane rotations), predict the pose for each transformed input, and map all predictions back to a common canonical reference frame. Jensen Gain is defined as:

$$G_J = \mathbb{E}[f(X)] - f(\mathbb{E}[X]). \tag{22}$$

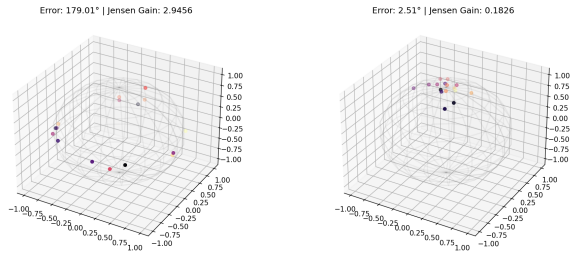
Intuitively, G_J measures the dispersion of canonicalized predictions across a symmetry orbit. A perfectly symmetry-consistent model would produce identical outputs for all transformations, yielding $G_J = 0$. In contrast, $G_J > 0$ indicates orientation-dependent variation in latent representation, revealing symmetry breaking within the backbone.

Table 5 reports the average Jensen Gain across domains of the SPEED+ dataset. Figure 5 further visualizes this orbit wobble. In a symmetry-breaking case ($E_R = 179.6°$), canonicalized predictions exhibit bimodal dispersion on the unit sphere. Conversely, a stable sample ($E_R = 3.51°$) shows tightly clustered predictions with negligible wobble.

These results indicate that while SPACE-HOP improves rotational stability through structured $SO(3)$ discretization and refinement, the underlying JEPA representation remains susceptible to symmetry-induced ambiguity, particularly under challenging lighting conditions.

Domain	Average Jensen Gain
Synthetic-Val	2.287
Sunlamp	2.819
Lightbox	2.831

Table 5. Average Jensen Gain of SPACE-HOP across different domains of the SPEED+ dataset. Higher values observed in Sunlamp and Lightbox domains suggest increased symmetry instability under illumination variation.



(a) Representation of Recovered Pose in sample where $E_R = 179.6$

(b) Representation of Recovered Pose in sample where $E_R = 3.51$

Figure 5. Orbit wobble visualisation obtained by rotating the input image $K = 16$ times, predicting the pose for each rotation, and un-rotating all predictions back to a common canonical frame. Each dot represents the predicted rotation (third column of the rotation matrix) after alignment. The x, y, and z axes denote the components of this unit direction vector on the sphere.Colours indicate the applied rotation angle. **Left**: a symmetry-breaking case ($E_R = 179.6°$, $G_J = 2.95$) showing bimodal dispersion. **Right**: a stable case ($E_R = 3.51°$, $G_J = 0.18$) where predictions collapse to a tight cluster.

5. Conclusions

This paper presented **SPACE-HOP**, an assessment of **Joint-Embedding Predictive Architecture (JEPA)** for **spacecraft 6-DoF pose estimation**. Our findings demonstrate the following:

- Latent-space feature prediction effectively captures the rigid-body geometry required for orbital pose tasks without needing pixel-level reconstruction.
- The architecture maintains high attention on structural keypoints across Synthetic, Sunlamp, and Lightbox domains, proving robust against extreme illumination and domain shifts, Also the model generalises well to out of domain samples of other satellites.
- Sampling the target masks based on spacecraft topology forces the model to prioritize critical components like antennas and corners over surface textures.
- Formulating orientation as a classification-and-refinement task over a discretized **Hopf Fibration** grid reduces sensitivity to noise and effectively handles axial symmetry ambiguities.

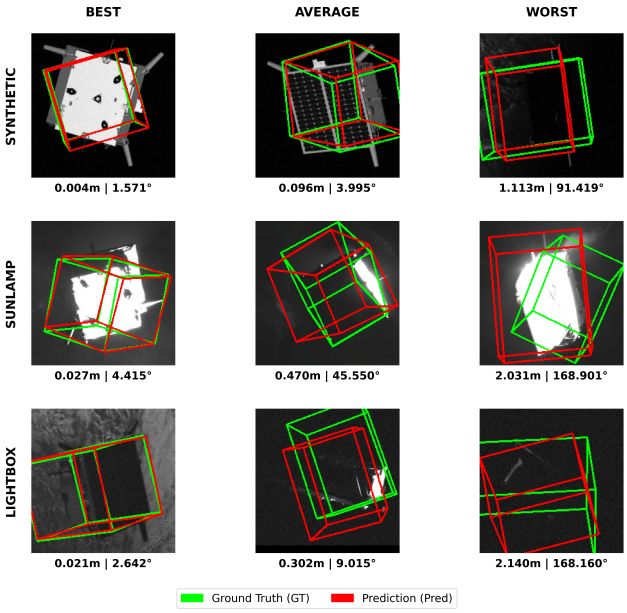


Figure 6. **Qualitative pose predictions across domains**

In summary, this assessment establishes embedding-predictive pretraining as a scalable, physically grounded alternative to generative or supervised pipelines in uncooperative space environments.

6. Future Work

With evidence that a pretrained single-level JEPA is effective, we aim to transition to a multi-target Hierarchical Joint-Embedding Predictive Architecture (H-JEPA). Leveraging multiple latent spaces and distinct representations, this hierarchy improves structural understanding by learning fine-grained details alongside geometric regularities. Consequently, it reduces position-dependent shortcuts that violate physical symmetries, establishing a stable foundation for orientation recognition while minimizing the description-length tax of enforcing invariance.

The Jensen Gain analysis reveals opportunities for strengthening symmetry consistency in learned representations. First, Jensen Gain may be incorporated as a regularization term to penalize prediction variance across symmetry orbits during training. Second, standard 2D positional embeddings could be replaced with geometry-aware encodings aligned with underlying SE(3) transformations.

References

[1] Pierre Ancey, Andrew Price, Saqib Javed, and Mathieu Salzmann. Fastpose-vit: A vision transformer for real-time spacecraft pose estimation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. 2, 6, 7

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. 2, 3

[3] Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104011, 2024. 1

[4] Raktim Gautam Goswami, Prashanth Krishnamurthy, Yann LeCun, and Farshad Khorrami. Robopepp: Vision-based robot pose and joint angle estimation through embedding predictive pre-training. 2025. 2, 3

[5] Rasmus L. Haugaard, Frederik Hagelskjær, and Thorbjørn M. Iversen. Spyropose: Se(3) pyramids for object pose distribution estimation. 2023. 3

[6] Thomas Jantos, Mohamed Amin Hamdad, Wolfgang Granig, Stephan Weiss, and Jan Steinbrener. Poet: Pose estimation transformer for single-view, multi-object 6d pose estimation, 2022. 2

[7] Yi Li, Gu Wang Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. 2018. 1

[8] David W. Lyons. An elementary introduction to the hopf fibration, 2022. 1

[9] Kieran Murphy, Varun Jampani, Sreethama Ramalingam, and Ameesh Makadia. Implicit-pdf: Non-parametric representation of probability distributions on the rotation group so(3). 2021. 3

[10] Tae Ha Park and Simone D'Amico. Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap. *Advances in Space Research*, 73(11): 5726–5740, 2024. 2, 6, 7

[11] Tae Ha Park, Sumant Sharma, and Simone D'Amico. Towards robust learning-based pose estimation of noncooperative spacecraft. In *2019 AAS/AIAA Astrodynamics Specialist Conference, Portland, Maine*, 2019. 2, 6, 7

[12] 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 *2022 IEEE Aerospace Conference (AERO)*, page 1–15. IEEE, 2022. 6

[13] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. 2019. 1

[14] Juan Ignacio Bravo Pérez-Villar, Álvaro García-Martín, and Jesús Bescós. Spacecraft pose estimation based on unsupervised domain adaptation and on a 3d-guided loss combination, 2022. 7

[15] Sumant Sharma and Simone D'Amico. Pose estimation for non-cooperative spacecraft rendezvous using neural networks. In *2019 AAS/AIAA Space Flight Mechanics Meeting, Ka'anapali, Maui, HI*, 2019. 2, 6, 7

[16] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. 2018. 1

[17] George Terzakis, Manolis Lourakis, and Djamel Ait-Boudaoud. A consistently convergent exponential map for so(3) regression. 2018. 5

[18] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. 2019. 1

[19] Anna Yershova, Steven M. LaValle, and Steven M. Jain. Generating uniform incremental grids on so(3) using the hopf fibration. *The International Journal of Robotics Research*, 2010. 3, 4

[20] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. 2019. 2