2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SMARTPOSE: Development of a Sample-efficient, Model-Agnostic, Robust, Two-stage POSE Estimator for Unknown Satellites

Yash Kishorbhai Joshi[1] and Suresh Sundaram[2]

Abstract— **Accurate relative pose estimation is critical for autonomous close-proximity satellite operations, such as on-orbit servicing and debris removal. However, this task remains highly challenging for unknown, non-cooperative targets due to the unavailability of geometric priors, scarce training data, and the extreme variations in lighting and backgrounds inherent to the space environment. Existing approaches for pose estimation can be categorized into: 1. model-based methods, which achieve higher accuracy but assume access to 3D CAD models of target satellites; and 2. model-free methods, which can be used for unknown satellites but typically suffer from reduced accuracy and robustness. To bridge this gap, this paper introduces a sample-efficient, model-free pose estimation framework that achieves high accuracy and robustness for unknown satellites while demonstrating strong generalization under the uncertain conditions inherent to the space environment. The proposed method utilizes a novel appearance aware 3D reconstruction to generate satellite model from images accounting for different lighting conditions during the training. This model is then used to generate a large, diverse dataset to train a pose predictor network (stage 1). The predicted pose is refined using the 3D reconstruction by utilizing the appearance information of the target image along with differentiable rendering (stage 2). Evaluated across SPEED+ and URSO Soyuz datasets, our approach achieves state-of-the-art accuracy and proves highly robust to test-time domain shifts, notably reducing rotation error by 80% on the challenging URSO Soyuz dataset.**

I. INTRODUCTION

Increasing congestion in Low Earth Orbit (LEO), driven by mega-constellations, inactive satellites, and debris, has made autonomous spacecraft navigation and proximity operations a priority and a challenge. As of May 2025, over 14,000 active satellites and 120 million debris fragments are tracked in orbit, posing critical risks to both commercial and defence missions [1]. Reliable onboard pose estimation of target satellites is required for tasks such as debris removal, docking, and on-orbit servicing, particularly in uncooperative scenarios where no prior geometric models are available. Given the stringent mass and power constraints of space missions, monocular vision offers a compelling solution compared to more resource-intensive systems like LiDAR or radar. The feasibility of visual navigation in space has been demonstrated by NASA's Seeker mission [2]. However, inferring 6-DoF pose from a 2D image is an inherently ill-posed problem. Furthermore, space-based images exhibit high variability in lighting and backgrounds due to changing orbital conditions. Additionally, acquiring sufficient training data for unknown satellites can be particularly difficult, as it often requires capturing multiple images while orbiting the satellite, resulting in limited datasets, making the task of building a robust perception framework for accurate pose estimation especially challenging.

The existing works on satellite pose estimation can be grouped into model-based and model-free methods, distinguished by their reliance on prior geometric information about the target. Model-based methods leverage known 3D CAD model of the target satellite to achieve high accuracy. D'Amico [3] utilized edge detection to establish a 2D-3D correspondence between the image and the wireframe model of the satellite. Another approach, proposed by Sharma et al. [4], employed feature detection in conjunction with Perspective-n-Point (PnP) using known keypoint locations on the satellite geometry. However, these classical approaches perform poorly under the variable lighting and backgrounds of real-world orbital operations.

The advent of deep learning has significantly improved the robustness of model-based techniques. Spacecraft Pose Network (SPN) [5] was the first to employ a CNN based pose estimation method which could be trained using paired image, pose, and bounding box data. Keypoint Regression Network (KRN) identified 11 keypoints on the satellite and predicted the location of the keypoints using a CNN based network. The 2D location of keypoints and 3D ground truth locations were used to derive the satellite pose using PnP. Building upon these approaches, SPNv2 [6] introduced a multi-task network that predicts the pose along with other geometric priors, such as bounding box, keypoint locations, and segmentation masks. By having access to these geometric priors, the network is grounded to learn the satellite geometry effectively and generalizes well across domains.

In contrast to this, model-free methods use only the paired image-pose data for training without requiring prior knowledge of the spacecraft geometry or any other geometric priors. These approaches learn mapping from learnt image features to pose prediction. Notable examples include URSONet [7], FilterFormerPose [8], and the method proposed by Shukla et al. [9]. These methods, however, typically exhibit lower accuracy and robustness, as learning the complex mapping from 2D images to 6-DoF pose without any geometric grounding requires vast and diverse data. This presents a critical dilemma for autonomous operations with unknown targets: model-based methods are rendered unusable by the lack of a geometric prior, while the performance of existing model-free methods is often insufficient for the precision

*We acknowledge the financial support form ARTPARK for this research
[1]Yash Kishorbhai Joshi was an M. Tech. student with the Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India joshi.yash.235@gmail.com
[2]Suresh Sundaram is with the Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India vssuresh@iisc.ac.in

required by on-orbit tasks.

Advances in 3D reconstruction such as NeRF [10] and 3D Gaussian Splatting [11] present a promising path to bridge the gap between model-free and model-based methods. Recent studies leverage 3D reconstruction to enhance satellite pose estimation. Wang et al. [12] utilized neural surface representations (NeuS) to extract component level segmentation masks, Legrand et al. [13] utilized a NeRF based method to generate diverse synthetic views. However, both approaches rely on CAD models or fail to leverage reconstruction as a direct supervisory signal for pose estimation.

A key limitation of standard 3D reconstruction techniques is handling reflective surfaces and drastic variations in lighting conditions, which can result in floating artefacts. Several "in-the-wild" reconstruction techniques have been developed to manage appearance changes and occlusions in unconstrained photo collections. For example, [14] proposed a per image learnt appearance embedding and occlusion to perform reconstruction. GS-W [15] utilizes a complex UNet based architecture to extract appearance features and visibility map from images. These methods are typically benchmarked on datasets like Photo Tourism [16], where the primary challenge is handling occlusions from street-level views. Consequently, they are not optimized for the distinct and more severe problem of high-frequency lighting changes and reflective surfaces that dominate satellite imaging.

In this work, we propose SMARTPOSE: a sample-efficient, model-agnostic, robust, two-stage pose estimation framework that eliminates the dependence on prior knowledge of the spacecraft model while achieving high accuracy and robustness in pose estimation from monocular images.

Our key contributions are as follows:

- We propose a two-stage predictor-corrector framework for pose estimation, built upon a novel appearance aware 3D reconstruction (AAR) module that creates high-fidelity 3D models from limited images under extreme lighting variations.
- In stage 1, the trained AAR model acts as a synthetic data engine to train a robust CNN-based predictor, generating a diverse dataset of novel poses and appearances
- In stage 2, the corrector refines the initial estimate by first conditioning the AAR model on the target image's specific appearance. It then leverages the model's differentiable nature to iteratively update the predicted pose by minimizing the photometric error between the rendered and the target image.

By leveraging AAR for robust scene representation and integrating it with a two-stage framework, SMARTPOSE establishes a new direction in satellite pose estimation, capable of operating in unknown, adversarial, or degraded orbital conditions. We validate our approach and analyze the contribution of each component through extensive ablation studies on two public datasets, demonstrating the framework's effectiveness and key design choices.

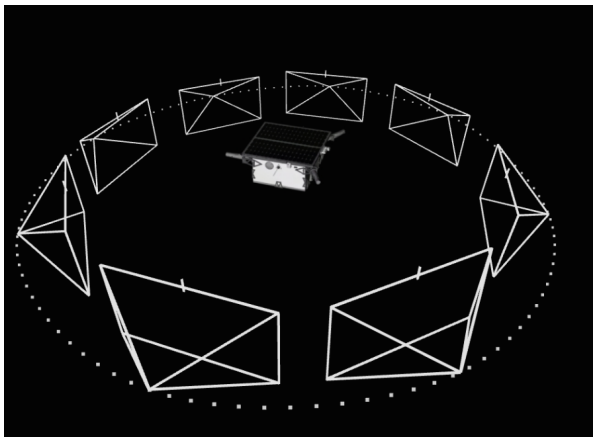


Fig. 1: Schematic of the data acquisition scenario. A chaser satellite (camera poses shown as frustums) performs an initial fly-around maneuver to capture a limited set of images of an unknown target satellite from various viewpoints

II. Development of SMARTPOSE

A. Problem Statement

For a chaser satellite to autonomously perform critical on-orbit tasks like servicing or debris removal, it must estimate the unknown, non-cooperative target's pose amidst a challenging environment of extreme lighting changes and diverse visual backgrounds with accuracy and robustness. Let the target's body frame be denoted by B and the camera's frame by C. The pose is defined as:

- **A translation vector** $t \in \mathbb{R}^3$ representing the position of the origin of frame B with respect to frame C.
- **A rotation**, represented by a quaternion $q \in \mathbb{R}^4$ describing the orientation of frame B relative to frame C.

Given a target image I_t, the problem is to find a mapping that predicts the pose $(\hat{q},\hat{t})$ that is as close as possible to the ground truth pose (q_{gt},t_{gt}).

This work addresses the problem under the fundamental constraint that no prior 3D CAD model of the target is available. Instead, we assume that a small, uncurated set of initial image-pose pairs, $S = \{(I_i,(q_i,t_i))\}_{i=1}^{N}$, can be acquired during an initial observation or fly-around phase of a mission. as depicted in Fig. 1. This limited dataset, where N is on the order of a few thousand images, constitutes the only target-specific information available. The core challenge is to leverage this sparse data to create a robust system for accurate, real-time pose estimation.

B. Framework

The SMARTPOSE approach first involves a model preparation phase for a specific target. In this phase, our Appearance-Aware Reconstruction (AAR) module learns a 3D model from a small, initial set of images. This learned model is then used to generate a large, synthetically labelled dataset to train a fast, CNN-based pose predictor. The system then proceeds to the inference phase, which operates in two stages: the trained predictor provides a rapid initial

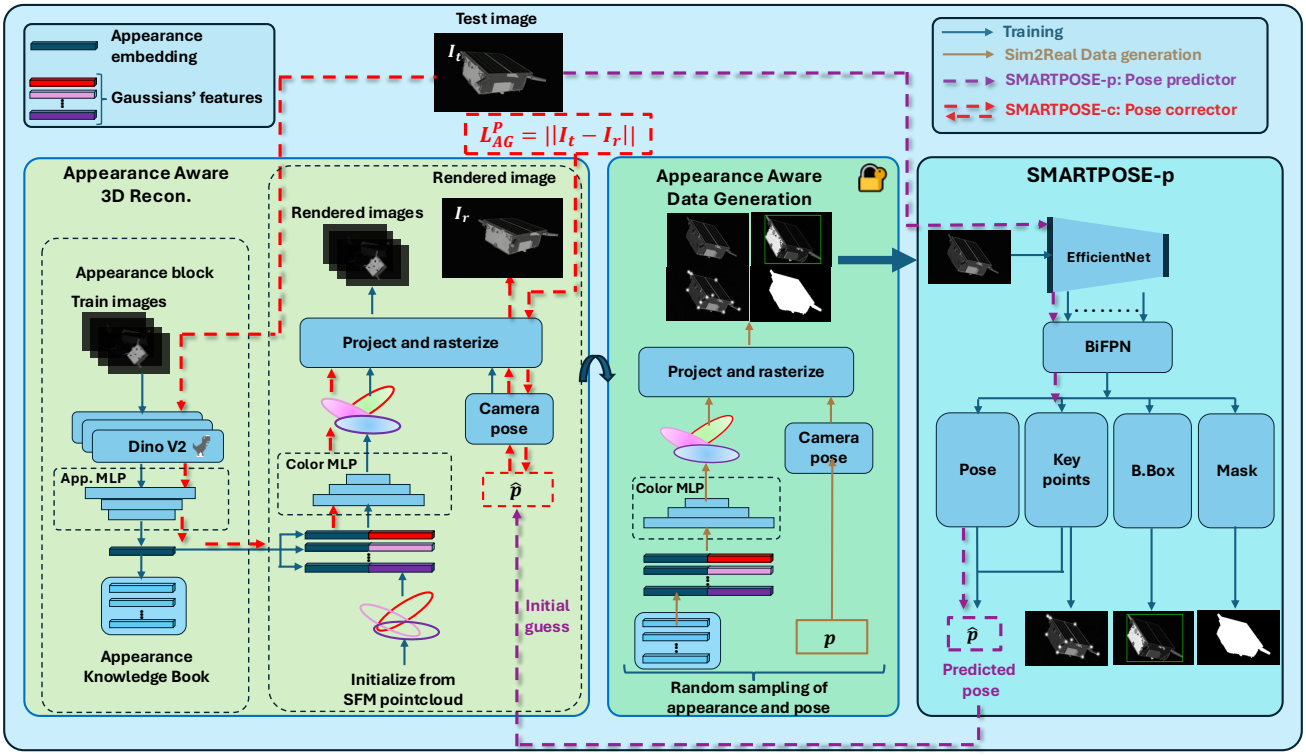


Fig. 2: Overview of the SMARTPOSE framework. During the Training Phase, our Appearance Aware Reconstruction module (left) learns a 3D model from a few images, which then serves as a synthetic data generator (center) to train the SMARTPOSE-p predictor (right). The Appearance Aware Reconstruction extracts and stores an appearance embedding for each training image in an Appearance Knowledge Book. During training, the appearance embedding corresponding to train image is concatenated with features of all Gaussians and passed through a color MLP to obtain the color of the Gaussians conditioned on the appearance. The knowledge book is again used for data generation stage to sample random appearance and pose combinations. At inference, a two-Stage process is used: the predictor provides a fast initial guess (Stage 1, pink flow), which is then refined by the corrector (Stage 2, red dashed flow). The corrector iteratively minimizes photometric error against renderings from the 3D model, which is conditioned on the target image's specific appearance.

pose estimate, which is then passed to a corrector stage for iterative, high-accuracy refinement against the AAR model. The overall framework is shown in Fig. 2.

1) Appearance Aware 3D Reconstruction (AAR): Our method builds upon 3D Gaussian Splatting (3DGS) proposed by Kerbl et al. [11] that represents 3D scenes as a collection of 3D Gaussians. Each Gaussian primitive in the scene is defined by its physical properties: position, shape, color, and opacity. Each Gaussian has a mean and a covariance matrix which describe its position and shape respectively. The position is described by a mean vector, μ, and the shape by a covariance matrix, Σ. The influence of a single Gaussian at any 3D point, x, is calculated as:

$$G(x) = \exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right) \tag{1}$$

During rendering, the 3D Gaussians are projected onto the camera plane. The covariance matrix in camera coordinate frame can be written as:

$$\Sigma' = JW\Sigma W^{T}J^{T} \tag{2}$$

where, W is the viewing transformation for the camera plane and J is the Jacobian of affine approximation of projective transformation,

In the baseline implementation, view dependent radiance field is represented by spherical harmonics. Spherical harmonics combined with view direction gives the color of the Gaussian. The projected Gaussians are sorted based on their distances and α blending of the projected Gaussians is carried out to obtain the resulting color at any pixel:

$$C = \Sigma_{i\in N} c_i \alpha_i \Pi_{j=1}^{i-1}(1-\alpha_i) \tag{3}$$

where, c_i is the color of each Gaussian and α_i is the opacity of the Gaussian (G_i) at a particular pixel location.

Instead of storing the coefficients of spherical harmonics for each Gaussian, we assign a learnable feature vector to each Gaussian, as the available images of our target dataset are likely to have significant variation in lighting conditions (also referred to as variation in appearance). We create an appearance network by combining a pre-trained Dino V2 network [17] with an MLP (called appearance MLP). The appearance feature vector is concatenated with

the feature vector of each Gaussian to form a combined appearance feature vector. The appearance feature vector is passed through a small MLP (called color MLP) to obtain the color of the Gaussian. Finally, α-blending is carried out to render an image (Eq. 3). The appearance network is trained along with the Gaussians. The appearance vectors of the training images are stored in a knowledge book for use at test time. Once the color of each Gaussian is calculated based on the appearance embedding, the rest of the process is similar to baseline 3DGS, i.e. project and rasterize Gaussians according to camera pose, take the loss between rendered and training image, and back-propagate to train the model.

This formulation allows the Gaussians to adapt their color across lighting conditions and viewing contexts based on the input image, reducing artefacts like floaters. The network is trained using a weighted sum of L_1 and D-SSIM losses, while geometric densification (cloning, splitting, pruning) is applied iteratively to improve spatial accuracy.

Once the appearance aware reconstruction is trained, it is frozen and used as a synthetic data generator. Novel images are rendered by sampling combinations of camera poses and appearance embeddings from the knowledge book, including interpolation of the learned embeddings using spherical linear interpolation (SLERP). The corresponding ground truths for bounding boxes, segmentation masks, and projected keypoints are simultaneously extracted. Segmentation is obtained by applying a threshold on opacity α values, and keypoints are extracted by marking the coordinates of relevant points (such as corners, extrema, etc.) in world coordinate frame and then projected onto camera plane corresponding to different synthetically generated images.

Hyperparameters for the appearance aware 3D reconstruction are arrived at through extensive experiments. The Appearance Aware 3D Reconstruction (AAR) model uses a `DINOv2_ViT-L` backbone followed by a three-layer MLP to generate a 128-D appearance vector for each image. This is combined with 128-D Gaussian specific feature vectors, and a final two-layer MLP maps the resulting 256-D vector to an RGB color. The model is trained for 200,000 iterations. The DINO backbone is fine-tuned with a low initial learning rate of 1×10^{-7}, while the appearance and color MLPs use higher initial rates of 5×10^{-4} and 2.5×10^{-4}, respectively. All learning rates undergo exponential decay for the first 70,000 steps before being held constant. To refine the satellite's geometry, an adaptive geometric densification strategy is applied every 100 iterations, which involves cloning, splitting, and pruning the 3D Gaussians. This process is frozen after 16,000 iterations to allow the training to focus solely on fine-tuning the appearance and feature weights.

2) Pose estimation: The pose predictor (SMARTPOSE-p) follows the architecture of SPNv2 [6], a multi-task convolutional network trained on image-pose pairs and additional geometric priors such as segmentation masks and keypoint locations. It uses EfficientNet as a feature extractor, followed by a bidirectional feature pyramid network (BiFPN) for multi-scale fusion. Three heads predict bounding boxes, keypoint locations on camera plane, and segmentation masks. Pose is predicted via regression in addition to performing PnP on predicted keypoints.

The total training loss combines IoU and focal losses for bounding boxes, pixel-wise MSE loss for keypoints, binary cross-entropy for segmentation, and total pose loss for orientation and translation. The pose loss ($\mathscr{L}_{pose}$) minimizes the difference between the predicted pose $(\hat{q},\hat{t})$ and the ground-truth pose (q,t). It is defined as the combination of rotational error E_R normalized translational error E_t:

$$E_R = 2\cos^{-1}(|\hat{q} \cdot q|) \tag{4}$$

$$E_t = \|\hat{t} - t\|_2 \tag{5}$$

$$\mathscr{L}_{pose} = E_R + \frac{E_t}{\|t\|_2} \tag{6}$$

To improve robustness, extensive data augmentations like neural style transfer and random solar flares are utilized.

The corrector refines the predictor's initial estimate at test time by "inverting" the AAR model's differentiable rendering process (inspired from [19]). First, the AAR model is conditioned using the appearance vector extracted from the target image (I_t). Then, starting with the initial pose guess from the predictor, an image (I_r) is rendered. The pixel-wise L_2 difference between the rendered image and the input image is termed as geometry and appearance aware loss ($\mathscr{L}^P_{AG}$) as it depends on the geometry from 3D reconstruction and on the appearance of the target image. This loss is backpropagated through the Gaussians to update the pose estimate $(\hat{q},\hat{t})$. This optimization continues until the pose converges, leveraging the model's geometric and appearance information to achieve a highly accurate final estimate.

$$\mathscr{L}^P_{AG} = ||I_t - I_r(\hat{q},\hat{t},I_t)||_2 \tag{7}$$

SMARTPOSE thus combines model-free adaptability with model-based precision, leveraging geometry, appearance, and synthetic supervision in a unified, end-to-end framework.

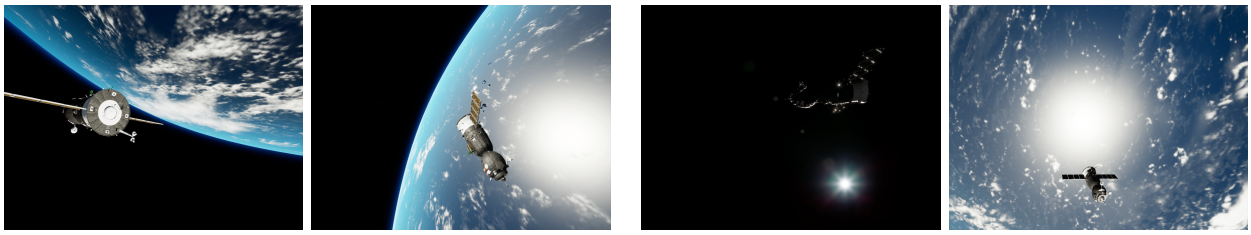
III. Evaluation Setup

We demonstrate that SMARTPOSE bridges the gap between model-free generalization and model-based accuracy. We evaluate SMARTPOSE on two publicly available satellite pose estimation datasets: URSO Soyuz [7] and SPEED+ [20], each presenting unique structural and visual challenges.

A. Datasets:

The URSO Soyuz dataset contains 4000 training and 500 test images each for two domains: *Soyuz easy* and *Soyuz hard*. It consists of rendered images of the Soyuz spacecraft with complex illumination, but lacks geometric priors like keypoints or segmentation masks. The dataset is simulated using Unreal Engine. *Soyuz hard* introduces broader variations in lighting and camera distance (10–40 m) compared to *Soyuz easy* (10–20 m). Fig. 3 shows representative images from both domains.

The SPEED+ dataset consists of over 50,000 images of the Tango satellite captured across *synthetic*, *lightbox*,



(a) Soyuz easy (b) Soyuz hard

Fig. 3: Sample images from URSO Soyuz dataset in easy and hard domains [7]

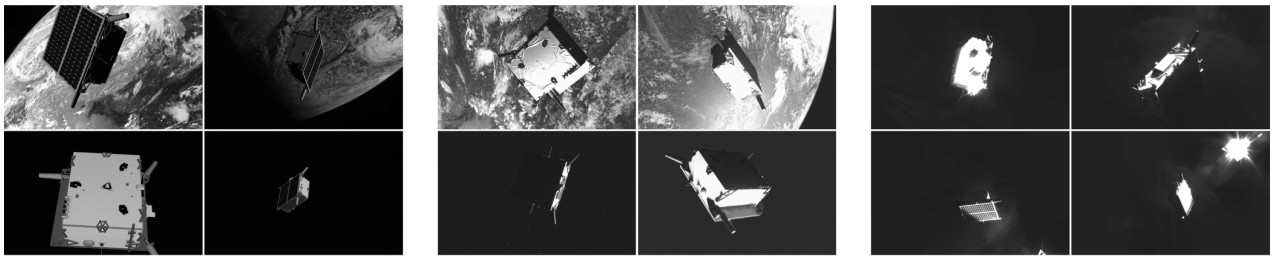


Fig. 4: Sample images from SPEED+ dataset in synthetic, lightbox, and sunlamp domain [18]

and *sunlamp* domains. It includes accurate camera intrinsics, 3D keypoint annotations, and segmentation masks. The Tango satellite is a cuboid-shaped object with three cylindrical appendages. The dataset offers significant variation in lighting and viewpoint, including Hardware-in-the-Loop (HIL) imagery designed to simulate the domain gap between simulation and real-world conditions. The *lightbox* domain simulates lighting from Earth and Moon albedo, while the *sunlamp* domain simulates extreme glare from direct sunlight. Sample images are shown in Fig. 4.

B. Performance Metrics.

Pose estimation accuracy is measured by the rotational error E_R, translational error E_t, and the combined pose error E_{pose}, as defined in the previous section. The quality of the 3D reconstruction is measured using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [21] on held-out validation images. PSNR measures pixel intensity change across the reconstructed and test images, a high PSNR indicates agreement between the pixel intensity values of rendered and validation image. SSIM compares the two images in terms of their luminance, contrast and structural content. An SSIM value closer to 1 indicates a higher degree of structural similarity, serving as a reliable indicator of human perceived quality. These benchmarks allow us to assess both pose prediction accuracy and the visual fidelity of our reconstructions under domain shifts. SMARTPOSE requires 20 hours of training NVIDIA RTX 4090 GPU for a given dataset. At inference, SMARTPOSE-p and SMARTPOSE-c achieve processing speeds of 10 fps and 0.5 fps, respectively.

IV. Performance Evaluation

We evaluated SMARTPOSE on the URSO Soyuz and SPEED+ datasets to demonstrate its key characteristics: sample efficiency, model-free operation, and robust pose estimation under significant domain shifts. We begin with the URSO Soyuz dataset, which is particularly challenging due to its limited data and lack of geometric priors.

A. Results for URSO Soyuz

The URSO Soyuz dataset consists of *easy* and *hard* domains, with the latter featuring more extreme lighting and distance variations. Due to the limited data and lack of geometric priors, prior model-free methods have only reported results by training and testing within the same domain. In contrast, we demonstrate SMARTPOSE's robust generalization by training it on the simpler *easy* domain and testing it on both, a more challenging and practical scenario.

Appearance aware reconstruction (AAR) model is trained using just 2000 images from the *Soyuz easy* training set, achieving a PSNR of 30.78 and an SSIM of 0.9787. Using this AAR model, we generated a large dataset of $\sim$50,000 synthetic images to train the SMARTPOSE-p (predictor). To ensure coverage of both test domains, we sampled camera poses from 10m to 40m and synthesized novel lighting conditions by interpolating appearance embeddings via SLERP. Fig. 5 shows sample generated images from the 3D reconstruction using different appearance embeddings. The Earth backgrounds are added to the generated images in a similar fashion as the URSO dataset in post processing. Fig. 6 shows the different geometric ground truths along with the corresponding generated image. This synthesised dataset is used to train the SMARTPOSE-p network.

The pose estimation results for URSO Soyuz dataset are

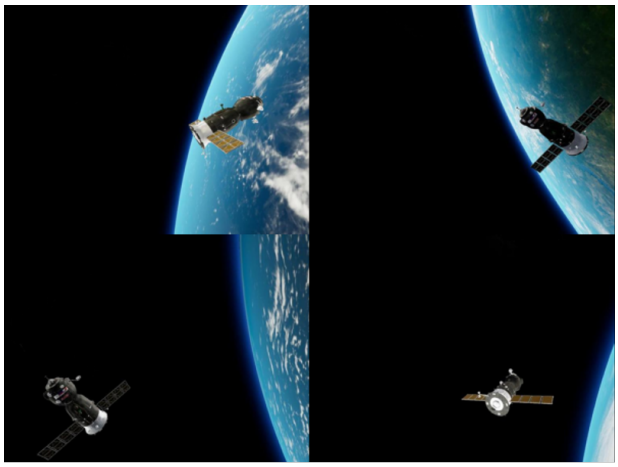


Fig. 5: Generated images from the appearance aware reconstruction created using 2000 images from *Soyuz easy* domain with the background added post generation.

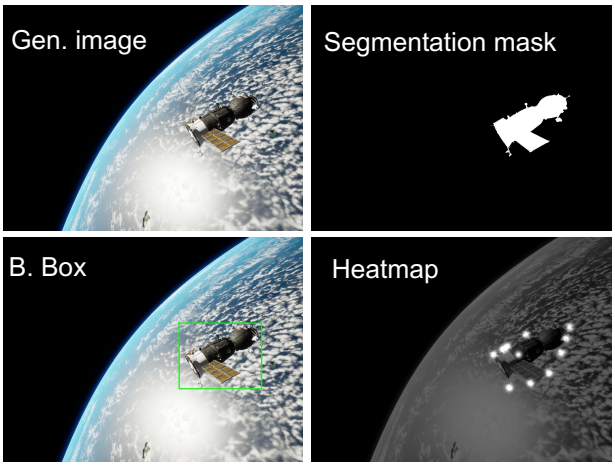


Fig. 6: A sample generated image along with its corresponding geometric priors (segmentation mask, bounding box, and keypoint locations) obtained from the 3D reconstruction.

given in Tab. I. Performance of SMARTPOSE is compared with previous methods in terms of the rotational error (E_R) and translational error (E_T). The predictor stage (SMARTPOSE-p) can provide fast estimates for pose as it is a feed-forward network. It can be seen that SMARTPOSE-p significantly outperforms all existing model-free methods. It reduces the translational error by $\sim$60% and rotational error by $\sim$80% as compared to the best prior method, FilterFormerPose. Its strong generalization is particularly noteworthy; when tested on the unseen and more challenging *Soyuz hard* domain, the predictor (trained only on *Soyuz easy* domain) achieves a rotation error of 2.93° and translation error of 0.44 m which represent an improvement of $\sim$50% over the best prior method, FilterFormerPose, which was specifically trained and tested on the *Soyuz hard* data.

Building on this robust initial estimate, the iterative second stage, SMARTPOSE-c, consistently refines these predictions to achieve even higher accuracy. This corrector stage refines

Model	Trained	Easy		Hard	
		E_T (m)	E_R (°)	E_T (m)	E_R (°)
URSONet [7]	Easy	0.8	7.7	-	-
	Hard	-	-	0.9	13.9
Ye et al. [8]	Easy	0.49	4.61	-	-
	Hard	-	-	0.75	5.20
Shukla et al. [9]	Easy	0.69	6.74	-	-
	Hard	-	-	1.32	12.7
SMARTPOSE-p	Easy	0.17	1.04	0.44	2.93
SMARTPOSE-c	Easy	**0.08**	**0.40**	**0.36**	**2.30**

TABLE I: Performance comparison on the URSO Soyuz dataset against prior model-free methods across two different domains. Both stages of SMARTPOSE significantly outperform existing works and demonstrate strong domain generalization. Best performance is highlighted in bold.

the pose estimate and generalizes effectively to the unseen domain. The final pose errors are exceptionally low, and in the *Soyuz easy* case, are nearly an order of magnitude lower than the previous state-of-the-art (0.40° vs. 4.61°).

B. Results for SPEED+ dataset

To evaluate the sample efficiency of SMARTPOSE, we utilized only a small fraction of the SPEED+ dataset. From the full synthetic training set of approximately 48,000 images, we used only 2,000 randomly selected images ($\sim$4%) that did not feature the Earth in the background. Despite training without any geometric priors, the AAR module achieved a high-fidelity reconstruction on a 250-image validation set, reaching a PSNR of 35.22 and an SSIM of 0.9533 as compared to PSNR of 30.35 and SSIM of 0.9356 for the 3DGS. This result confirms the module's ability to learn the geometry and appearance information robust manner from a remarkably small and diverse dataset. The synthetic domain in SPEED+ dataset contains images with deep space and the Earth as background in $\sim$50:50 ratio. Hence, background of the Earth are added to half of the generated images.

Using the generated 3D model, a dataset of 48,000 images is synthesised to train SMARTPOSE-p, the predictor stage. The corresponding geometric priors such as bounding boxes, segmentation masks, and keypoint locations are generated for all images. The corrector stage, SMARTPOSE-c, then refines these initial predictions from SMARTPOSE-p. Tab. II compares our performance against SPNv2, which uses the same predictor architecture but was trained on the full dataset. Despite being trained on data synthesized from a fraction of the original images, SMARTPOSE-p achieves comparable or superior pose estimation accuracy across all domains. This result highlights the powerful sample efficiency of the AAR-driven data generation approach. The corrector stage, SMARTPOSE-c, demonstrates its effectiveness in the synthetic domain by further reducing the overall pose error by $\sim$65%. As is common with iterative refinement techniques, SMARTPOSE-c requires a reasonably accurate initial guess to ensure convergence. For the challenging lightbox and sunlamp domains, the initial pose errors from the predictor

Model	Synthetic			Lightbox			Sunlamp		
	E_T (m)	E_R (°)	E_{pose}	E_T (m)	E_R (°)	E_{pose}	E_T (m)	E_R (°)	E_{pose}
SPNv2 [6]	0.047	0.99	0.025	**0.198**	8.12	0.17	0.25	14.94	0.30
SMARTPOSE-p	0.044	0.88	0.023	0.222	**7.82**	**0.17**	**0.23**	**13.68**	**0.28**
SMARTPOSE-c	**0.013**	**0.34**	**0.008**	-	-	-	-	-	-

TABLE II: Pose estimation performance on the SPEED+ dataset. SMARTPOSE demonstrates superior or comparable performance to the baseline while using significantly less source data. The corrector stage (SMARTPOSE-c) dramatically improves accuracy further.

Model	Trained on	Soyuz Easy		Soyuz Hard	
		E_T (m)	E_R (°)	E_T (m)	E_R (°)
SMARTPOSE-c	Easy	**0.08**	**0.40**	0.36	2.30
	Hard	**0.08**	0.52	**0.30**	**1.56**

TABLE III: Ablation study on cross-domain training for the URSO Soyuz dataset. SMARTPOSE generalizes effectively in both directions (easy ↔ hard).

were larger than 7°, falling outside the convergence basin of the corrector. Consequently, we report corrector results for the synthetic domain where it serves its intended role: high-precision refinement.

C. Ablation studies

To further analyse the model's robustness, we conducted a series of ablation studies. The following studies show the cross domain generalization capability of SMARTPOSE on both the datasets as well as effect of auxiliary tasks and data augmentation from AAR.

1) Cross domain generalization: We test cross-domain generalization on the URSO Soyuz dataset by training on the *hard* domain and testing on the *easy* domain. The results, shown in Tab. III, demonstrate strong bidirectional generalization. Notably, the performance when training on the *Soyuz hard domain* and testing on the *Soyuz easy* domain is excellent, with a final rotation error of 0.52° as compared to 0.41° when trained on *Soyuz easy* domain. However, we consider the primary experiment (training on *Soyuz easy* and testing on *Soyuz hard*) to be a more difficult and realistic test of generalization, as it requires the model to adapt from simpler to more complex and varied visual conditions.

Similarly for SPEED+, we trained SMARTPOSE on 2000 images randomly sampled from the more challenging HIL lightbox domain. We compare our results to the work of [13], who trained on the lightbox domain. However, there are key differences in methodology: 1) Their NeRF-based approach synthesized images but still relied on the dataset's geometric ground truths for training, whereas our method is fully model-free and infers all necessary geometric constraints internally. 2) They used a manually curated subset of high-quality images, while we use a random, uncurated sample, which is more representative of a real-world scenario.

As shown in Tab. IV, while their approach yields slightly lower in-domain translational error, SMARTPOSE-p shows improvement for rotational error and demonstrates significantly stronger generalization, achieving a 37% improvement in pose error on the unseen sunlamp domain. This highlights superior robustness of SMARTPOSE to test-time domain shifts.

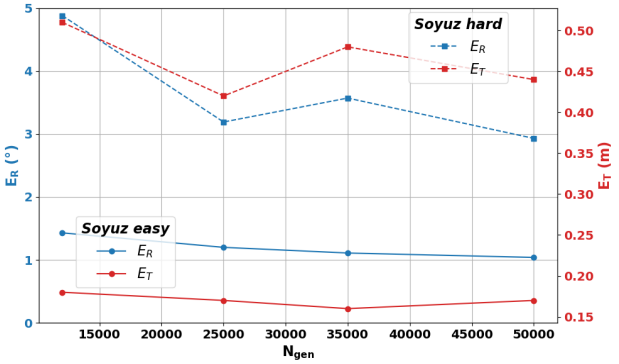


Fig. 7: Change in rotation error (E_R) and translation error (E_T) with increasing N_{gen} for SMARTPOSE-p across both *Soyuz easy* and *hard* domains. The AAR was trained using 2000 images from *easy* domain and then different number of images were synthesised for training SMARTPOSE-p.

2) Auxiliary tasks and data augmentation: In addition, we also analysed the effect of different components of SMARTPOSE-p architecture. As shown in Tab. V, training on the 4000 original *Soyuz easy* images with generated geometric ground truths shows that including auxiliary tasks like segmentation and keypoint detection improves performance, with keypoint regression providing the most significant gain. Furthermore, training on a large dataset (50,000 images) synthesized by our AAR model drastically improves accuracy and domain generalization compared to training on the original limited dataset, confirming the value of our data generation approach. We also found that performance gains saturated around 35,000-50,000 generated images (Fig. 7).

V. Conclusion

This paper introduced SMARTPOSE, a sample-efficient, two-stage framework for 6-DoF pose estimation of unknown satellites from monocular images, eliminating the need for prior CAD models. The core contribution is a novel Appearance Aware 3D Reconstruction (AAR) method that builds a high-fidelity, differentiable 3D model from a small, uncurated set of images. By synthesizing a large labelled dataset from this model, we train a robust predictor network (SMARTPOSE-p) for fast pose estimation, which is then refined by a geometry-aware corrector (SMARTPOSE-c).

Model	Lightbox			Sunlamp		
	E_T (m)	E_R (°)	E_{pose}	E_T (m)	E_R (°)	E_{pose}
Legrand et al. [13]	**0.2**	7.2	**0.158**	0.37	15.3	0.375
SMARTPOSE-p (lightbox)	0.278	**6.81**	0.170	**0.22**	**10.97**	**0.231**

TABLE IV: Training on 2000 lightbox images, SMARTPOSE shows stronger generalization to the unseen sunlamp domain.

SMARTPOSE-p			Trained on	Soyuz easy		Soyuz hard	
Pose	Segmentation	Keypoints		E_T (m)	E_R (°)	E_T (m)	E_R (°)
✓			Soyuz easy (N = 4000)	1.61	91.32	8.39	98.66
✓	✓			1.22	47.80	8.02	79.08
✓		✓		0.25	2.02	1.71	26.09
✓	✓	✓		0.23	1.92	1.94	24.00
✓	✓	✓	Generated images (N = 50000)	**0.17**	**1.04**	**0.44**	**2.93**

TABLE V: Effect of auxiliary tasks and data augmentation on SMARTPOSE-p performance for the URSO dataset. Including keypoint regression provides the most significant performance gain. The last row shows the results with generated images from AAR to train SMARTPOSE-p. Training on this large synthetically generated dataset yields the best performance as well as robustness to test-time domain gap.

SMARTPOSE is evaluated on the challenging SPEED+ and URSO Soyuz datasets. On the data-scarce URSO Soyuz dataset, SMARTPOSE demonstrated remarkable generalization, outperforming all prior methods by reducing rotation error by up to 75% on an unseen domain. On SPEED+, SMARTPOSE achieved performance comparable to state-of-the-art model-based methods while using only 4% of the source training data. These results confirm that SMARTPOSE effectively bridges the performance gap between model-based and model-free approaches, paving the way for scalable and autonomous perception systems for on-orbit servicing and space debris removal. Future work should focus on making the 3D reconstruction pipeline more physics-aware by explicitly incorporating illumination parameters such as sun angle, reflectance modelling, and spacecraft material properties.

REFERENCES

[1] European Space Agency. (2025) Space debris by the numbers. European Space Agency. [Online]. Available: https://www.esa.int/Space_Safety/Space_Debris/Space_debris_by_the_numbers

[2] S. Pedrotty, J. Sullivan, E. Gambone, and T. Kirven, "Seeker free-flying inspector gnc system overview," in *American Astronautical Society Annual Guidance and Control Conference (AAS GNC 2019)*, no. JSC-E-DAA-TN64849, 2019.

[3] S. D'Amico, M. Benn, and J. L. Jørgensen, "Pose estimation of an uncooperative spacecraft from actual space imagery," *International Journal of Space Science and Engineering 5*, vol. 2, no. 2, pp. 171–189, 2014.

[4] S. Sharma and S. D'Amico, "Pose estimation of uncooperative spacecraft using monocular vision," *Perspective*, vol. 3, p. 2D, 2014.

[5] S. Sharma and S. D'Amico, "Pose estimation for non-cooperative rendezvous using neural networks," *arXiv preprint arXiv:1906.09868*, 2019.

[6] T. H. Park and S. D'Amico, "Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap," *Advances in Space Research*, vol. 73, no. 11, pp. 5726–5740, 2024.

[7] P. Proencca and Y. Gao, "Deep learning for spacecraft pose estimation from photorealistic rendering," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6007–6013.

[8] R. Ye, L. Wang, Y. Ren, Y. Wang, X. Chen, and Y. Liu, "Filterformer-pose: satellite pose estimation using filterformer," *Sensors*, vol. 23, no. 20, p. 8633, 2023.

[9] S. Shukla, R. Srivastava, R. Lima, and T. Bera, "Satellite-model-free deep learning based pose estimation of non-cooperative satellite and tracking using navigation filter," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 4548–4555.

[10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2022.

[11] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023.

[12] Z. Wang, M. Chen, Y. Guo, Z. Li, and Q. Yu, "Bridging the domain gap in satellite pose estimation: A self-training approach based on geometrical constraints," *IEEE transactions on aerospace and electronic systems*, vol. 60, no. 3, pp. 2500–2514, 2023.

[13] A. Legrand, R. Detry, and C. De Vleeschouwer, "Leveraging neural radiance fields for pose estimation of an unknown space object during proximity operations," *arXiv preprint arXiv:2405.12728*, 2024.

[14] H. Dahmani, M. Bennehar, N. Piasco, L. Roldao, and D. Tsishkou, "Swag: Splatting in the wild images with appearance-conditioned gaussians," in *European Conference on Computer Vision*. Springer, 2024, pp. 325–340.

[15] D. Zhang, C. Wang, W. Wang, P. Li, M. Qin, and H. Wang, "Gaussian in the wild: 3d gaussian splatting for unconstrained image collections," in *European Conference on Computer Vision*. Springer, 2024, pp. 341–359.

[16] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM SIGGRAPH 2006 Papers*. ACM, 2006, pp. 835–846.

[17] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[18] T. H. Park, M. Märtens, G. Lecuyer, D. Izzo, and S. D'Amico, "Speed+: Next-generation dataset for spacecraft pose estimation across domain gap," in *2022 IEEE aerospace conference (AERO)*. IEEE, 2022, pp. 1–15.

[19] A. Rahman, R. Gupta, and S. Kim, "iComMa: Iterative camera pose refinement via differentiable gaussian splatting," *arXiv preprint arXiv:2312.01234*, 2023.

[20] T.-H. Park, M. Märtens, G. Lecuyer, D. Izzo, and S. D'Amico, "Speed+: Next-generation dataset for spacecraft pose estimation across domain gap," in *2022 IEEE aerospace conference (AERO)*. IEEE, 2022, pp. 1–15.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.