CroSpace6D: Leveraging Geometric and Motion Cues for High-Precision Cross-Domain 6DoF Pose Estimation for Non-Cooperative Spacecrafts

Jianhong Zuo[1,3], Shengyang Zhang[2,3], Qianyu Zhang[2,3], Yutao Zhao[2,3],
Baichuan Liu[2,3], Aodi Wu[2,3], Xue Wan[3], Leizheng Shu[3], Guohua Kang[1]

[1] Nanjing University of Aeronautics and Astronautics
{jianhongzuo,kanggh}@nuaa.edu.cn
[2] University of Chinese Academy of Sciences
{zhangshengyang22,zhangqianyu22,zhaoyutao22,liubaichuan23,wuaodi20}@mails.ucas.ac.cn
[3] Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences
{wanxue,shuleizheng}@csu.ac.cn

Abstract

The utilization of monocular vision for non-cooperative spacecraft pose estimation has been significantly researched in space target monitoring, on-orbit servicing, and satellite maintenance. The challenge lies in addressing the cross-domain variations in shape, texture, lighting, and motion patterns between simulated and real captured images. To tackle this issue, a novel domain adaptation 6DoF pose estimation algorithm is proposed to extract the geometric and semantic consistency between cross-domain training and testing datasets. Experimental results demonstrate that our pose estimation method achieves state-of-the-art performance on the SPARK2024 dataset.

1. Introduction

The enhancement of Space Situational Awareness (SSA) is imperative for managing the densely populated near-Earth orbital environment[10]. Monocular pose estimation represents a milestone, underscoring the pivotal influence of Artificial Intelligence (AI) in aerospace, crucial for the effective gathering of data for orbital monitoring.

Advances in non-cooperative spacecraft pose estimation has been notable, particularly with the utilization of monocular sensors—integral for space object monitoring, on-orbit servicing, and satellite maintenance. HRNet[14] highlights keypoint detection methodologies that identify distinct 3D model points in images, then use Perspective-n-Point (PnP) algorithms to infer object depth. While promising, these methods over-rely on annotated data and falter in the presence of noise and occlusions, compromising their effectiveness in real-world and cross-domain applications. CNN-based techniques[5] streamline the determination of camera positions directly from images. These approaches provide a quick option for ascertaining spacecraft orientation and position but their accuracy declines when translating simulations to actual textures, predominantly owing to reliance on simulated data. Furthermore, these methods inadequately capitalize on the geometric consistency of targets. Moreover, Model-based pipelines[6] excelled in controlled settings, evidenced by their performance in the BOP competition[4], despite difficult lighting, lack of object textures, clutter, and obstructions. Nevertheless, these pipelines are principally suited for indoor contexts and struggle with the complex backgrounds characteristic of in-space environments, particularly on-orbit. They also neglect temporal information, which is essential for refining pose predictions. The persistence of specific, unresolved issues in existing techniques underscores the need for enhanced non-cooperative spacecraft pose estimation: enhancing robustness against noisy backgrounds and fluctuating lighting, bridging the gap between simulation and reality, and refining the pose estimation by utilizing temporal information.

This paper introduces an innovative framework Crospace6D that not only addresses these enduring challenges but also augments the current field by focusing on the geometric stability and motion dynamics of spacecraft, culminating in precise, cross-domain pose estimation. This paper makes the following contributions to the field of non-cooperative spacecraft pose estimation: (i) Addressing noisy background interference by using semantic segmentation for background removal and tracking algorithms for subject consistency. (ii) Despite the variability observed

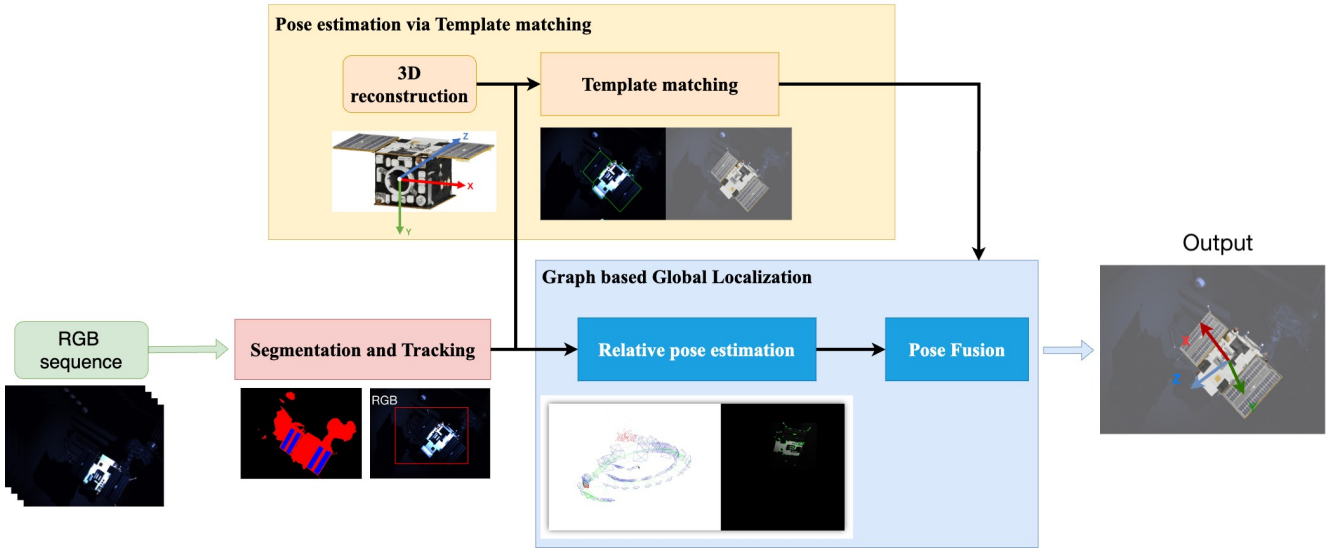

Figure 1. The structure of Crospace6D pipeline.

in 2D imagery, the underlying 3D shape remains invariant across both virtual and real-world scenarios. Exploiting this geometric stability, the paper reconstructs the 3D mesh of non-cooperative spacecraft from simulated images. By integrating the reconstruction model with template-based pose estimation, consistent and domain-agnostic transformation of the spacecraft's geometric information is achieved. (iii) Moreover, a graph-based optimization approach that uses visual odometry association is proposed. Alongside the enforcement of motion consistency constraints using visual odometry to capture target geometric information. Experimental results validate the state-of-the-art performance of the Crospace6D method on the SPARK2024 dataset[11]. The pipeline is shown in Fig.1.

The following is the paper's outline: Section 2 reviews previous research on pose estimate techniques, Section 3 details our suggested architecture, Crospace6D, Section 4 contains the comparative analysis and ablation study, and Section 5 summarizes the whole paper.

2. Relate Work

6D pose estimation is a fundamental task in the field of computer vision, aiming to accurately determine the position and orientation in 3D space based on given images. According to different principles underlying 6D pose estimation technology, it can be categorized into three distinct methods: keypoint-based, template matching-based, and SLAM-based.

2.1. Keypoint based 6d pose estimation

The keypoint-based methods employ a detection network to precisely locate the specific keypoints of an object in the image. These image keypoints are then matched with corresponding points in the prior CAD model, and the 6D pose of the object is estimated using the PnP (Perspective-n-Point) algorithm. Keypoints can be represented as either coordinates or heat maps. DeepPose[13] introduced the concept of keypoint location regression by utilizing convolutional neural networks to predict keypoint coordinates based on learned image features. HRNet[14] is a classical approach for estimating keypoint heat maps, which incorporates parallel networks with different resolutions to maintain high-resolution feature maps throughout the entire network while preserving local details and global context information, resulting in outstanding performance in keypoint detection. Although partially addressing occlusion issues, keypoint-based methods require a sufficient number of reliable keypoints for accurate pose estimation.

2.2. Template matching based 6d pose estimation

The template matching-based methods utilize a prior CAD model to generate an extensive library of templates representing diverse object poses, enabling estimation of the 6D pose by comparing the similarity between the image target region and these predefined templates. TemplatePose[9] offers a solution for recognizing and estimating the pose of new objects, even in partially occluded scenarios, without necessitating training on new objects. It learns local object representations from a small set of training objects and subsequently matches test images with rendered images de-

rived from CAD models to obtain their poses. MegaPose[6] is also suitable for pose estimation of new objects and it introduces a novel coarse pose estimation approach along with a pose refiner under rendering and comparison strategies. The training on large-scale synthetic datasets makes it more generalizable to new objects. The template matching-based methods are simple and intuitive, but they may not provide smooth enough poses when processing continuous image sequences.

2.3. SLAM based 6d pose estimation

SLAM-based methods leverage a continuous sequence of images to simultaneously estimate camera motion and the structure of the surrounding environment. The methods can be broadly classified into direct and indirect approaches. LSD-SLAM[3] is an optical flow-based direct SLAM that focuses on pixels in high gradient regions of the image, estimating the relative pose by minimizing the photometric error to achieve semi-dense mapping. ORB-SLAM[8], on the other hand, is a feature-based indirect SLAM that extracts and matches ORB features from images, estimating relative pose by minimizing reprojection error for sparse mapping. The ORB-SLAM3[1] exhibits a significant improvement over the previous version, enhancing system flexibility and scope while enabling accurate 3D reconstruction and real-time localization in complex environments. Unlike the aforementioned methods, SLAM does not rely on prior information and can operate effectively in unknown environments. However, it is limited to calculating only the 6D pose between frames and can not accomplish absolute object localization. Besides, when estimating poses using a moving target as the reference frame, background interference needs to be eliminated[17].

3. Proposed Method

3.1. Spacecraft semantic segmentation and tracking

Semantic segmentation. To mitigate the impact of background interference, such as robotic arms, this paper initially extracts the foreground of the satellite and subsequently performs pose estimation. The foreground extraction is accomplished using the Mask2former[2] semantic segmentation algorithm, which is based on the Transformer architecture. In Mask2Former[2], the input image undergoes a preprocessing stage to generate a series of feature maps. These feature maps are then passed through the Pixel Decoder module to enhance them into high-resolution feature maps. Finally, these feature maps are utilized to generate masks, thereby achieving the image segmentation task.

Tracking. Despite Mask2former strong image segmentation ability and interactivity, its performance in consistent image sequence segmentation falls short. Thus, we use Track Anything[16] in this work, designed to achieve high-performance segmentation and produces a bounding box in image sequences.

3.2. Pose estimation via template matching

To overcome variations in shape, texture, lighting, and motion patterns between simulated and real captured images, the crux lies in exploiting the target's three-dimensional information—a cross-domain stable characteristic. Three-dimensional information remains consistent under varying lighting and motion conditions and is insensitive to occlusions. Consequently, this method begins by reconstructing the three-dimensional model of the non-cooperative spacecraft using virtual imagery, subsequently applying a template matching-based pose estimation approach.

3D reconstruction. This paper utilized the current training dataset images for the 3D reconstruction of the satellite. Addressing the shortcomings of conventional explicit 3D reconstruction methods, which include pronounced noise, a lack of geometric detail, and incomplete models, our research adopts NeuS[15], a neural scene representation technique that employs a signed distance field for the implicit 3D reconstruction of the target satellite Proba, drawing on deep learning methodologies. To infuse the model with texture information, the color values of the model mesh vertices are retrieved from the neural signed distance field (SDF).

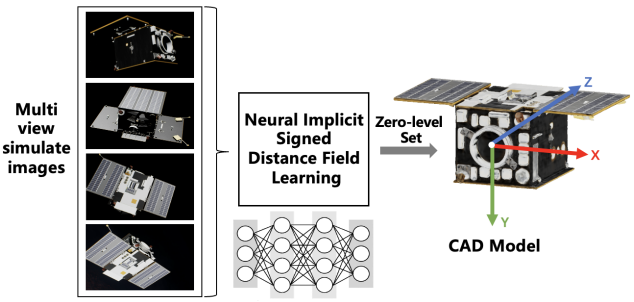

Figure 2. NeuS based 3D reconstruction using training images.

Template matching. Our objective is to estimate the i time absolute pose of the target spacecraft $T_{c_0}^{s_i} \in \mathbf{SE}(3)$, with the camera serving as the reference system. This estimation is based on the input RGB sequences and the region containing the target, after reconstructing the spacecraft mesh through 3D reconstruction. We selected the Template-Matching Monocular Pose Estimation algorithm MegaPose[6] as the baseline for non-cooperated spacecraft pose estimation. MegaPose employs 3D reconstruction results to determine the target's pose, thereby mitigating image underexposure and target symmetry issues to some extent. The result is shown in Fig 3.

While this method performs well on most data, single-frame approaches like this inevitably exhibit significant

fluctuations and discontinuities when applied to continuous motion pose estimation. To address this issue, the next step Graph-base Global Localization is needed.

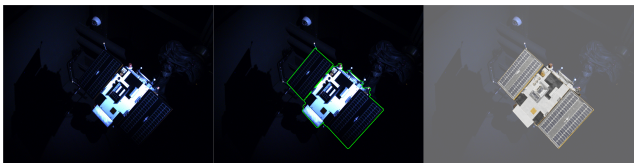

Figure 3. Template matching result, left: origin image, middle: contour overlay, right: mesh overlay.

3.3. Graph-based Global Localization

To increase the accuracy of pose estimation, the key approach is to fuse absolute pose estimation with relative pose estimation. In contrast to the single-frame estimation of the absolute poses that mentioned in Section 3.2, the simultaneous localization and mapping (SLAM) approach benefits from estimating the inter-frame relative pose has higher smoothness. However, due to the lack of depth information by the monocular camera, relative poses estimated by SLAM have incorrect scales. Moreover, SLAM estimated the pose of each moment relative to the pose of the initial motion. Thus, this method uses graph-based pose fusion after every relative pose estimation with the corresponding absolute pose.

Relative pose estimation. SLAM's estimation of the relative pose is based on static environment assumptions. In order to estimate the camera pose via ORB-SLAM3 with the target spacecraft as the static environment, other backgrounds such as the Earth and the Moon need to be masked by the method of Section 3.1. The relative pose at time i can be represented as $T_{c_0}^{c_i} \in \mathbf{SE}(3)$.

Pose fusion. The relative pose estimated by monocular SLAM is obtained in a normalized scale and may exhibit scale discrepancies compared to the true pose. Furthermore, the coordinate system of the relative pose is based on the camera's pose at the beginning of motion, lacking the transformation relationship with the target spacecraft's coordinate system. Therefore, it is possible to estimate a similarity transformation $A_{s_0}^{c_0} \in \mathbf{Sim}(3)$ to transform the relative pose as follows:

$$\hat{T_{s_0}^{c_i}} = A_{s_0}^{c_0} T_{c_0}^{c_i} \tag{1}$$

$\hat{T_{s_0}^{c_i}}$ is i time camera pose in target spacecraft frame.Transforming the reference system from the target spacecraft to the camera:

$$\hat{T_{c_0}^{s_i}} = \hat{T_{s_0}^{c_i}}^{-1} \tag{2}$$

$\hat{T_{c_0}^{s_i}}$ is the absolute position estimate of the target spacecraft at moment i, using the camera as the reference system.

The residual equation corresponding to $T_{c_0}^{s_i}$ as the observed value can be constructed as follows:

$$e = \log(T_{c_0}^{s_i} \hat{T_{c_0}^{s_i}}^{-1})^\vee = \log(T_{c_0}^{s_i} A_{s_0}^{c_0} T_{c_0}^{c_i})^\vee = (\phi^T, \tau^T) \tag{3}$$

where $(\phi^T, \tau^T) \in \mathfrak{se}(3)$.

The Jacobian matrix corresponding to $A_{s_0}^{c_0}$ is:

$$\begin{aligned} &\frac{\partial}{\partial \epsilon} log(T_{c_0}^{s_i} \exp(\epsilon^\wedge) A_{s_0}^{c_0} T_{c_0}^{c_i})^\vee \\ &\approx \left(I + \tfrac{1}{2} \cdot \begin{bmatrix} -\phi^\wedge & 0 \\ -\tau^\wedge & -\phi^\wedge \end{bmatrix} \right) \cdot \begin{bmatrix} R_{c_0}^{s_i} & 0 & 0 \\ t_{c_0}^{s_i \wedge} R_{c_0}^{s_i} & R_{c_0}^{s_i} & -t_{c_0}^{s_i} \end{bmatrix} \end{aligned} \tag{4}$$

By employing g2o as the graph optimization algorithm, we construct the following graph structure. As shown in Fig.4, the blue triangular nodes represent the relative poses, which are fixed during the optimization process. The red square nodes represent the similarity transformations to be estimated. The green edges utilize the residual equations mentioned above, with the observed values being the absolute pose estimates. The objective is to minimize the residuals through continuous optimization until convergence. This process yields the similarity transformation matrix, enabling the fusion of absolute pose estimation and relative pose estimation. Consequently, a smoother estimation in the temporal domain can be achieved, providing the absolute pose of the target spacecraft at i time with respect to the camera as the reference system.

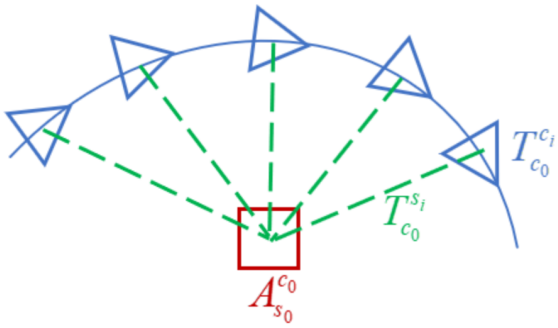

Figure 4. Pose graph.

4. Experiments

4.1. Experiments Settings

Datasets. This work is evaluated on the SPARK2024 Stream2 - Spacecraft Trajectory Estimation dataset[11], which is collected from the Zero-Gravity Laboratory (Zero-G Lab) facility, at SnT–Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg. The

SPARK2024 the dataset was simulated with a state-of-the-art rendering engine (Unity3D), which aims to design data-driven approaches for spacecraft semantic segmentation and trajectory estimation.

Implementation Details. Benefiting from the temporal continuity of RT001, RT002 and RT003, RT000 is constructed by splicing them together.Considering that initialization in monocular mode with ORB-SLAM3 requires some time, in order to estimate the complete trajectory, RT000 and RT004 are extended in reverse order. The final pose estimation is then obtained by taking the poses of the original sequence length in reverse order of the results. All experiments reported in this paper were carried out on a computer with an Intel Core i9-13900K CPU @ 5.80GHz, 64 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU.

4.2. Comparative Analysis

4.2.1 Semantic segmentation

The semantic segmentation results of FCN[7], Mask2Former and Track Anything are shown in the image below. It can be observed that both algorithms of FCN and Mask2former segment some background robotic arm regions such as the spacecraft body or solar panels. However, Mask2Former demonstrates higher segmentation accuracy and robustness in the solar panel area compared to FCN. Therefore, this paper utilizes the segmentation results of Mask2Former in the solar panel region as the Track Anything initial values, to achieve stable segmentation of the solar panel and the regions between surfboards.

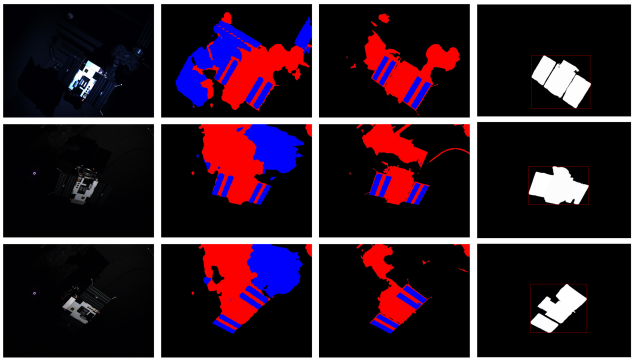

Figure 5. Semantic segmentation and tracking.

4.2.2 Pose estimation via template matching

To demonstrate the performance of NeuS, an implicit 3D reconstruction method, over the traditional explicit 3D reconstruction method, this work employed the NeuS and SfM algorithm[12] provided by Colmap software to reconstruct identical images from the training dataset. A qualitative comparison was conducted between the reconstruction results of both methods, as illustrated in Fig 6. It can be seen the SfM reconstruction model shows a sparse and incomplete representation, characterized by significant noise that results in indistinct geometric details. In contrast, the NeuS reconstruction model demonstrates clear geometric edges, enhanced completeness, and smoothness, as well as the ability to reconstruct detailed components with richer texture information. It is better suited for template matching in subsequent pose estimation tasks. Therefore, the neural scene representation method offers a more efficient and reliable solution for high-precision 3D reconstruction in space application scenes.

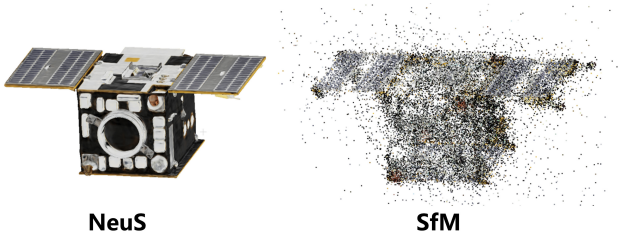

Figure 6. Comparison of NeuS and SfM reconstruction.

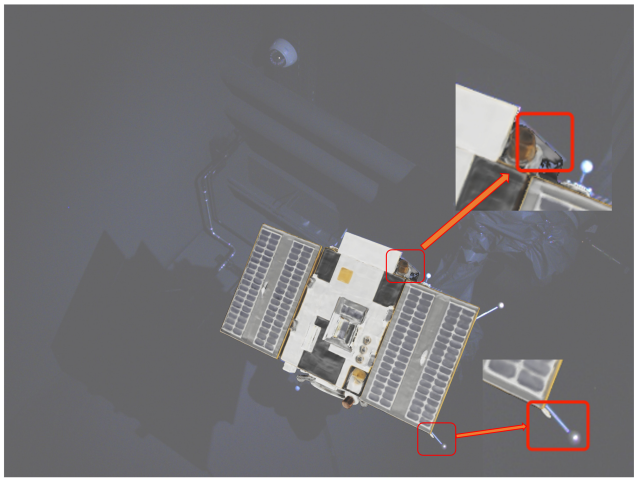

Figure 7. Analyzing pose estimation accuracy through mesh reprojection, discrepancies highlighted within enlarged red box.

In this paper, by inputting both the reconstruction results and the image segmentation tracking results into the megapose algorithm, we achieved template-based target pose estimation. Utilizing the camera's intrinsic parameters, the 3D reconstruction results were reprojected onto the original image according to the target pose, with the reprojection results shown in Fig 7. It can be observed that the projection of the model aligns perfectly with the solar panel part of the original image. Only within the enlarged red box area can minor errors be detected: a slight shift in the antenna located

Method	Translation error	Orientation error	Pose error
Absolute poses estimation + Filter	0.0276	0.0593	0.0663
Fusion poses estimation	0.0252	0.0187	0.0252

Table 1. Evaluated on the SPARK2024 dataset.

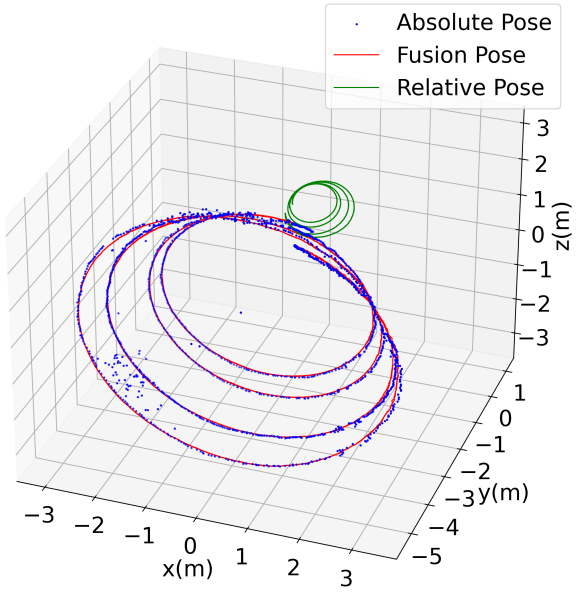

Figure 8. The trajectories of poses estimations.

at the bottom right and a subtle pose error in the upper right main body. These results validate the effectiveness of using simulated data to recover target geometric information for pose estimation, thereby proving that the consistency of target geometry between simulated and real data is key to solving such cross-domain issues.

4.2.3 Global Localization and Mapping

For the RT000 sequence, the relative pose, absolute pose, and fusion pose are estimated separately, and all the trajectories are shown in Fig.8. It can be seen that the pose fusion transforms the relative poses to the same coordinate system as the absolute poses. This study compares the performance of the conventional approach of optimizing absolute pose estimation using a filter and fusion pose method. The outcomes are illustrated in the accompanying Fig.9. The results demonstrate that the method significantly improved the precision of the absolute pose by leveraging the relative pose. Compared to the absolute pose, the fused pose exhibited higher smoothness and continuity. According to the SPARK2024 competition evaluation metrics, Table 1 presents the experimental results of the complete test set.

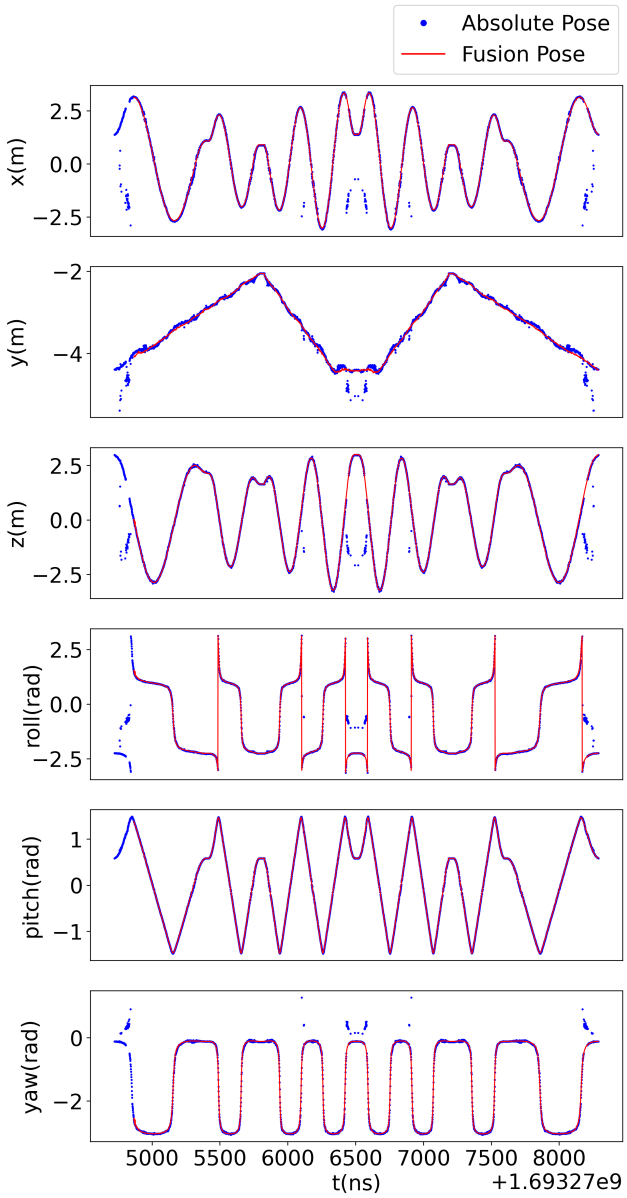

Figure 9. The comparison of absolute pose optimization using filter and fusion method.

5. Conclusion

In conclusion, this paper presents a significant advancement in the field of non-cooperative spacecraft pose estima-

tion using monocular vision. By addressing the challenges posed by cross-domain scenes in both simulated and real captured images, the framework CroSpace6D we proposed achieves high precision and demonstrates outstanding performance on the SPARK2024 dataset.

References

[1] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 3

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 3

[3] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 3

[4] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects. *arXiv preprint arXiv:2403.09799*, 2024. 1

[5] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization, 2016. 1

[6] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare. In *CoRL*. 1, 3

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 5

[8] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 3

[9] Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[10] Leo Pauly, Wassim Rharbaoui, Carl Shneider, Arunkumar Rathinam, Vincent Gaudillière, and Djamila Aouada. A survey on deep learning-based monocular spacecraft pose estimation: Current state, limitations and prospects. *Acta Astronautica*, 2023. 1

[11] Arunkumar Rathinam, Mohamed Adel Mohamed Ali, Vincent Gaudilliere, and Djamila Aouada. SPARK 2024: Datasets for Spacecraft Semantic Segmentation and Spacecraft Trajectory Estimation, 2024. 2, 4

[12] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[13] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[14] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 2

[15] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[16] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 3

[17] Shengyang Zhang, Aodi Wu, Quan Ye, Jianhong Zuo, Yadong Shao, Leizheng Shu, and Xue Wan. Space non-cooperative target visual navigation fused with object detection under complex backgrounds. In *2022 International Conference on Service Robotics (ICoSR)*, pages 145–151, 2022. 3