Getting Close: Multi-Body Tracking Enables Camera-based Close Proximity Robotic Manipulation for On-Orbit Servicing

Anne Elisabeth Reichert[1], Maximilian Ulmer[1,3], Rudolph Triebel[1,2] and Maximilian Durner[1]

***Abstract*— Reliable 6D pose tracking of uncooperative targets is a prerequisite for autonomous on-orbit servicing and space debris removal. While terrestrial trackers have reached high levels of maturity, the orbital environment introduces unique challenges: extreme high-contrast lighting, grayscale sensor data, and highly symmetric geometries. This work extends the state-of-the-art M3T framework to address these specific failure modes. Previously, we introduced a linear motion prior to constrain the optimization space and a Bayesian fusion approach that integrates learned segmentation masks with classical intensity histograms. In this work, we present our ongoing research on multi-object configuration to address the challenges of close-range satellite images. We validate our method using synthetic path-tracing data and real-world hardware-in-the-loop trajectories, demonstrating higher resilience compared to classical region-based baselines.**

I. Introduction

The expansion of orbital infrastructure requires robotic systems capable of maintenance, refueling, and repair. Central to these operations is visual perception, specifically 6D pose tracking, which provides the precise relative position and orientation over time required for safe rendezvous and docking.

However, standard trackers are often designed for terrestrial RGB-D scenarios. In space, sensors are predominantly grayscale, and the lighting ranges from total darkness to blinding specular reflections. To this end, we adapt the Multi-Body Multi-Modality Multi-Camera 3D Tracker (M3T) framework [1] to accommodate these constraints. Building on our line of work, we focus on improving the robustness of region-based tracking by:

1) Using motion prior to mitigate pose ambiguities caused by satellite symmetries.
2) Developing a Bayesian fusion method to incorporate learned segmentation features (here created by Segment Anything Model Version 2.1 (SAMv2) [2], [3]), handling the silhouette degradation typical in high-contrast orbital imagery.
3) Differentiating between different areas of the tracked satellite model to improve robustness in close-range scenarios with only partial object visibility.
4) Evaluating the framework on specialized sequential datasets, bridging the gap left by non-sequential benchmarks like SPEED [4].

*This work was not supported by any organization

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR) firstname.lastname@dlr.de

[2]Department of Informatics, Karlsruhe Institute of Technology (KIT)

[3]Department of Computer Science, Technical University of Munich (TUM)

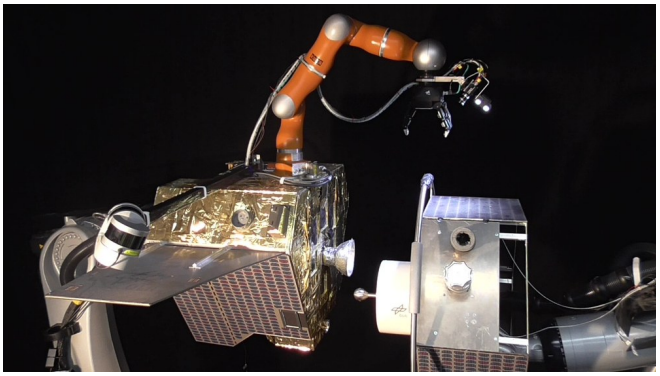


Fig. 1: Image of the On-Orbit Servicing Simulator (OOS-SIM) testbed at German Aerospace Center (DLR) holding the Compact Satellite Mockup (CSM) satellite in a close-range rendez-vous scenario used to record the hardware-in-the-loop data.

II. Related Work

Satellite pose estimation has seen significant progress with datasets such as SPEED+ [4], [5] and advances in synthetic data generation [6], [7], [8]. While these facilitate the training of deep learning models [9], [10], they lack sequential images, which are essential for evaluating the robustness of temporal tracking. Alternative sensors such as event-based cameras [11] and time-of-flight sensors [12] show promise, yet monochrome grayscale cameras remain the industry standard for space missions.

Classical object tracking approaches on grayscale data usually obtain the 6D pose from the object's geometry, which is provided in the form of a mesh. While simply relying on edges and corners, as [13] proposes, other works, such as [14], describe a more complex multi-stage processing pipeline to prevent drift and the loss of feature points by combining multiple feature correspondences.

Learned approaches like CroSpace6D [15] also incorporate geometric information to perform template matching with the object's 3D model. Additionally, the approach leverages TrackAnything [16], which utilizes SAM [2] to perform sequential image segmentation for its image processing step and incorporates motion cues, achieving high accuracy on the SPARKS2024 [17] dataset.

Our work builds upon the efficiency of M3T [18] while adding the global robustness of foundation models like SAM2 [3].

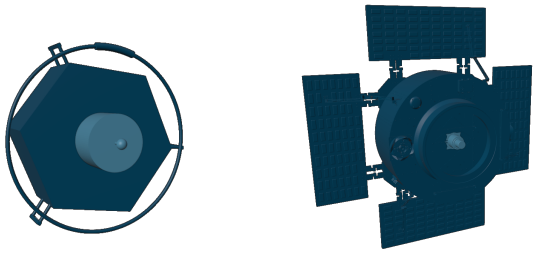


Fig. 2: Meshes of the CSM satellite on the left and of the EU:CROPIS satellite on the right. The base mesh is colored in dark blue, while the lighter blue indicates the secondary mesh.

III. Methodology

We use the motion prior and image segmentation extension of the Multi-Body Multi-Modality Multi-Camera 3D Tracker (M3T) formulation [1] presented in [19] and propose the multi-object configuration for the CSM and EU:CROPIS satellite.

A. Motion Priors and Soft Constraints

In the presence of symmetric geometries (e.g., solar panels or cylindrical bodies), visual data alone can lead to local minima or "flips" in the estimated pose. We incorporate a linear motion model that assumes constant velocity between high-frequency frames. Following the soft constraint formulation in [18], the optimization cost function is modified to:

$$E(\mathbf{p}) = E_{\text{visual}}(\mathbf{p}) + \lambda E_{\text{prior}}(\mathbf{p}|\hat{\mathbf{p}}), \tag{1}$$

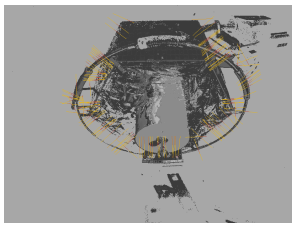
where $\hat{\mathbf{p}}$ is the predicted pose. In our experiments, we define a deviation threshold of $1\,\text{cm}$ for translation and 5° for rotation to prevent the prior from over-constraining the visual tracking under high-confidence conditions.

B. Bayesian Fusion of SAM2 Segmentation

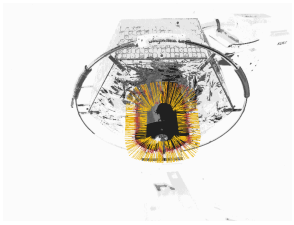
Region-based tracking relies on the ability to distinguish the objects' foreground pixels from the background using a probability $P(M|y)$, where y is the pixel intensity. In space, color histograms often fail due to harsh illumination variations, such as limited discriminability between foreground and background. To address this, we leverage the strengths of learning-based segmentation methods. Specifically, we integrate SAMv2 masks [2] by fusing their output with classical intensity-based probabilities. This allows the tracker to leverage SAM2's zero-shot segmentation capabilities [3], effectively "guiding" the contour-based optimization even when the satellite silhouette is partially obscured by shadow.

C. Multi-Object Modeling

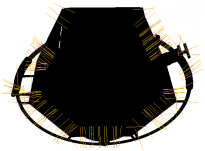
In close-range scenarios, the satellite's contour is not necessarily fully visible, which significantly degrades tracking performance. To address this challenge, we propose to differentiate between different parts of the satellite and introduce sub-regions into the tracking optimization. For this, we split our satellite meshes of the CSM and EU:CROPIS appearing in the Hardware-in-the-Loop (HIL) dataset into a base and secondary mesh as shown in Fig. 2.



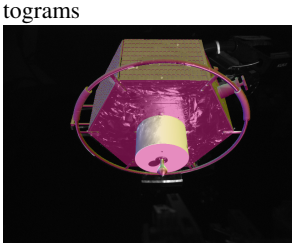
(a) Base mesh image segmentation using intensity histograms.



(b) Secondary mesh image segmentation using intensity histograms



(c) Base mesh image segmentation with fused masks from SAMv2



(d) Rendered overlay of the combined pose estimation of the base mesh constrained to the pose of the secondary mesh.

Fig. 3: Multi-object configuration approach for the CSM satellite model on HIL data.

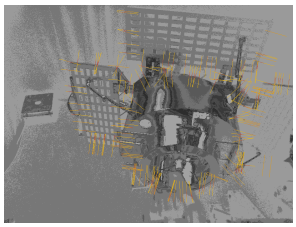
While the base mesh can still be tracked using our introduced fusion of segmentation masks, our secondary mesh relies entirely on the foreground-background segmentation based on the intensity histogram. Fig. 3 and Fig. 4 show how the base and secondary mesh of the CSM and EU:CROPIS optimize on different visual features using correspondences (drawn as yellow lines) in the image. In Fig. 3b, we observe that the foreground and background using intensity histograms can provide additional information to investigate our idea of using multi-object modeling. Here, the image is clearly divided into foreground (black) and background (white). In contrast, the segmentation for the secondary mesh of the EU:CROPIS satellite in Fig. 4b is less conclusive and shows more grayish areas. Motivated by this insight, we will segment the satellite's secondary parts and apply the provided multi-region approach to model its various regions. For now, we combine both tracked objects using a simple *Constraint* from the tracking framework M3T on all 6 Degrees of Freedom (DoF) of the optimized pose.

IV. Experimental Evaluation

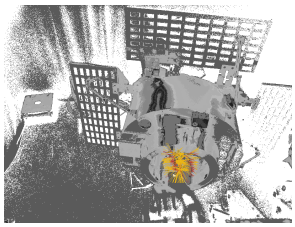
A. Datasets

We utilize two primary data sources:

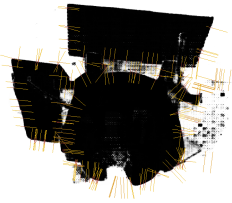
- **Synthetic Dataset:** Rendered using a high-fidelity path-tracing simulator for three satellites: CSM, Eu:CROPIS, and CALIPSO.
- **HIL Data:** Collected at the OOS-SIM facility at DLR shown in Fig. 1, providing real sensor noise and hardware-in-the-loop dynamics.



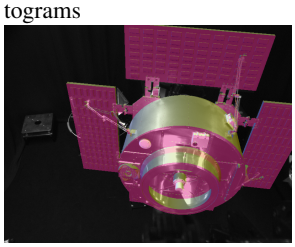
(a) Base mesh image segmentation using intensity histograms.



(b) Secondary mesh image segmentation using intensity histograms



(c) Base mesh image segmentation with fused masks from SAMv2



(d) Rendered overlay of the combined pose estimation of the base mesh constrained to the pose of the secondary mesh.

Fig. 4: Multi-object configuration approach for the Eu:CROPIS satellite model on HIL data.

extension A	extension B	extension C	CSM ADD	CSM ADD-S	Eu:CROPIS ADD	Eu:CROPIS ADD-S
	GT		71.58	86.27	94.39	97.92
	GT	✓	**83.44**	**91.29**	**94.76**	**98.13**
		✓	46.50	71.55	40.09	71.27
✓		✓	**53.47**	76.42	45.69	79.97
	✓	✓	38.15	75.50	**64.75**	84.93
✓	✓		43.77	72.67	64.12	82.19
✓	✓	✓	43.58	**78.95**	63.91	**85.80**

TABLE I: Comparison of the impact of the proposed extensions to M3T on accuracy using HIL data: motion priors (A), Bayesian fusion of SAM2 segmentation (B), and multi-object modeling (C). For extension B, we additionally report an upper-bound baseline using ground-truth masks.

B. Results and Discussion

While the results on the synthetic data reported in [19] show improved performance with our proposed extensions, during the crucial rendezvous and interaction phase, which is primarily represented in the HIL dataset, the performance does not significantly increase even if the fused segmentation masks are obtained from the ground truth. This is probably due to the close-range images that partially occlude the satellite's contour. Therefore, we compare our single-object approach to the proposed multi-object approach, including all previous extensions (fused segmentation masks and motion priors), on the HIL dataset. Preliminary results are reported in Table I. Here we can observe that applying all the suggested method extensions A-C results in the best ADD-S score. Additionally, the evaluation of Method B using the ground truth segmentation masks shows that the accuracy of the used segmentation masks remains crucial to the tracking performance, and including additional visual features from a more differentiated object modeling in the multi-object configuration further improves its potential.

The positive effect of dividing the satellite model is also shown in Fig. 5a and Fig. 5b, where two exemplary image sequences from the HIL dataset are provided. However, splitting the optimization of the object and simply constraining the process is not the ideal solution. First, the poses can still diverge if the defined constraints cannot be satisfied during optimization. This leads to decreased performance, since the resulting pose is incorrect for both. Additionally, tracking two constrained objects increases the computational load. A better solution might be to leverage the multi-region functionality of the M3T framework. Here, instead of using two separate image segmentations for the different parts of the mesh, the multi-region image segmentation considers previously defined sub-regions. Therefore, no additional constraint is needed, and tracking and optimization are performed once again for only one object. To do this, however, the fused segmentation mask must also recognize and differentiate the chosen sub-regions.

V. Conclusion

This paper presented a robust extension of region-based tracking for the space domain and provided insight into our current approach to address the identified challenges in HIL data. Combining classical optimization with learning-based segmentation, motion priors, and leveraging the satellite's structure improved tracking stability in critical, close-range scenarios. Future work will further investigate the promising multi-object approach by, for example, directly modeling multiple regions [20] to handle complex substructures such as antennas and docking ports, rather than relying on the currently implemented constraints. For this, the segmentation masks need to be extended to differentiate between different satellite parts.

References

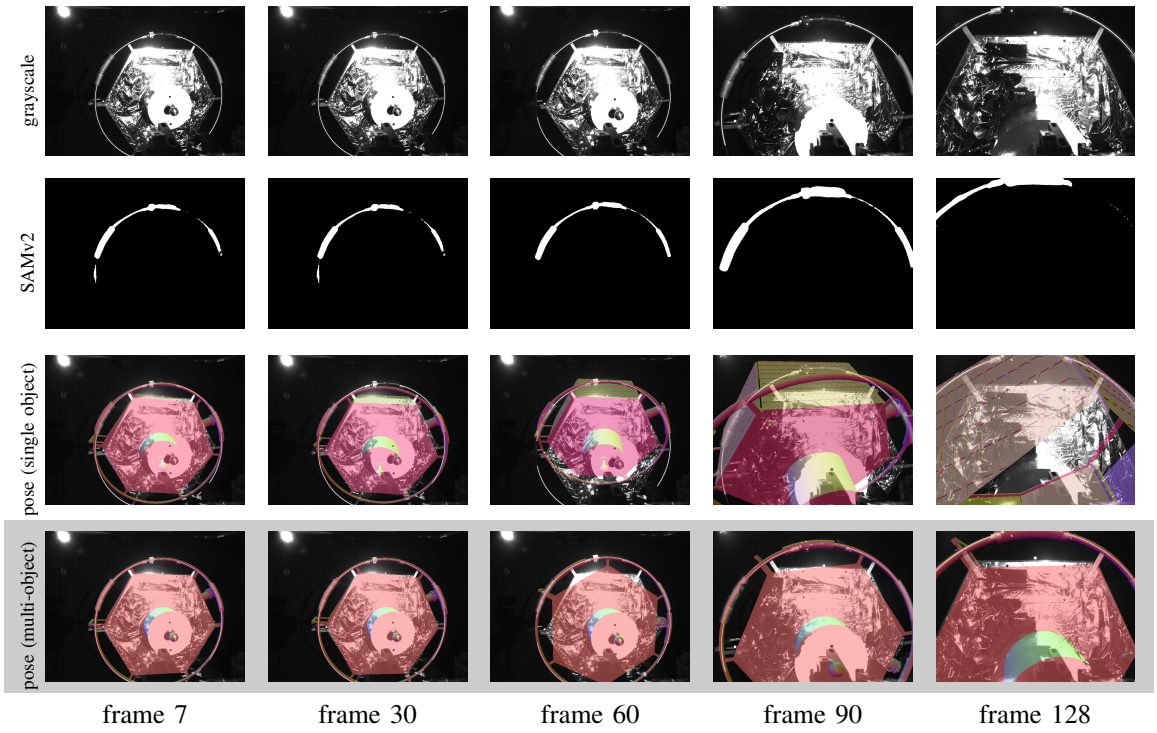
[1] M. Stoiber, M. Pfanne, K. H. Strobl, R. Triebel, and A. Albu-Schaeffer, "A sparse gaussian approach to region-based 6dof object tracking," in *Asian Conf. on Computer Vision*, 2020, pp. 666–682.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

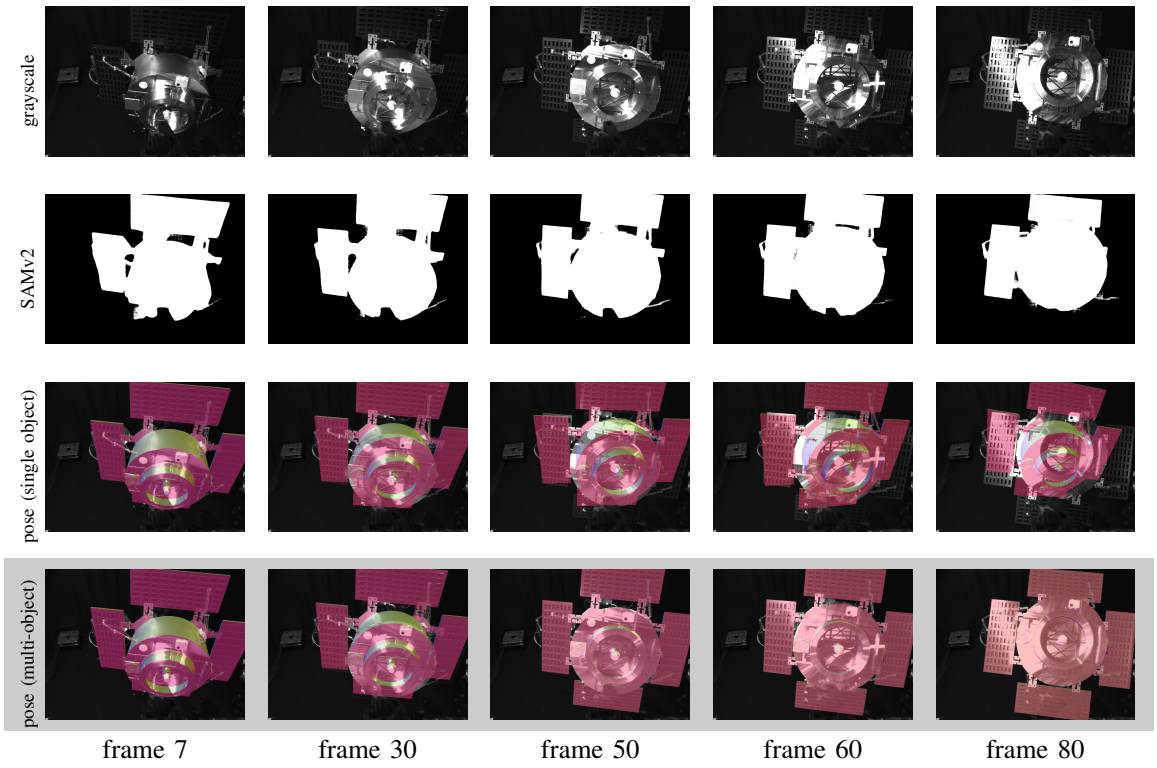
[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[4] M. Kisantal, S. Sharma, T. H. Park, D. Izzo, M. Märtens, and S. D'Amico, "Satellite pose estimation challenge: Dataset, competition design, and results," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 5, pp. 4083–4098, 2020.

[5] T. H. Park, M. Märtens, G. Lecuyer, D. Izzo, and S. D'Amico, "Next generation spacecraft pose estimation dataset (speed+)," Oct. 2021. [Online]. Available: https://doi.org/10.25740/wv398fc4383



(a) Sequence evaluation for the CSM satellite model.



(b) Sequence evaluation for the EU:CROPIS satellite model.

Fig. 5: Selected images of the HIL data showing the EU:CROPIS and CSM satellites. The first row shows the grayscale input image, and the second row shows the foreground and background segmentation by SAMv2. The third row shows tracking performed with the improved M3T, using SAMv2 masks and motion models, where the poses are rendered as an overlay on the image. The last row displays the same M3T setup but with the currently researched multi-object configuration, which, in this example, outperforms the single-object configuration.

[6] M. G. Müller, M. Durner, A. Gawel, W. Stürzl, R. Triebel, and R. Siegwart, "A Photorealistic Terrain Simulation Pipeline for Unstructured Outdoor Environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

[7] M. G. Müller, W. Boerdijk, M. Ulmer, W. Stürzl, A. Gawel, R. Siegwart, R. Triebel, and M. Durner, "Vision beyond earth: Synthetic satellite data for neural perception in orbit," in *2026 IEEE Aerospace Conference*. IEEE, 2026.

[8] D. Schenk, J. Wulkop, M. Ulmer, A. Gerndt, and G. C. Albuquerque Richers, "Picture your satellite in space: A hybrid rendering framework for physically based space images," in *2026 IEEE Aerospace Conference*. IEEE, 2026.

[9] M. Ulmer, M. Durner, M. Sundermeyer, M. Stoiber, and R. Triebel, "6d object pose estimation from approximate 3d models for orbital robotics," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 10 749–10 756.

[10] M. Ulmer, L. Klüpfel, M. Durner, and R. Triebel, "How important are data augmentations to close the domain gap for object detection in orbit?" in *2025 IEEE Aerospace Conference*. IEEE, 2025, pp. 1–12.

[11] S. S. Malik, M. Moshrefizadeh, O. Tahri, X. Bai, E. Blasch, V. Sagan, and H. AliAkbarpour, "Evsat3d: Satellite pose estimation and 3d reconstruction with event camera," *IEEE Access*, vol. 13, pp. 130 340–130 352, 2025.

[12] J. Hu, S. Li, and M. Xin, "Real-time pose determination of ultraclose noncooperative satellite using time-of-flight camera," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 6, pp. 8239–8254, 2024.

[13] N. W. Oumer, "Monocular 3d pose tracking of a specular object," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 3, 2014, pp. 458–465.

[14] 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.

[15] J. Zuo, S. Zhang, Q. Zhang, Y. Zhao, B. Liu, A. Wu, X. Wan, L. Shu, and G. Kang, "Crospace6d: Leveraging geometric and motion cues for high-precision cross-domain 6dof pose estimation for non-cooperative spacecrafts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 6857–6863.

[16] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv preprint arXiv:2304.11968*, 2023.

[17] A. Rathinam, M. A. Mohamed Ali, V. Gaudilliere, and D. Aouada, "Spark 2024: Datasets for spacecraft semantic segmentation and spacecraft trajectory estimation," Feb. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10908215

[18] M. Stoiber, M. Sundermeyer, and R. Triebel, "Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 6855–6865.

[19] A. Reichert, M. Ulmer, M. Piccinin, D. Eklund, H. Haglund, D. Schenk, M. Durner, and R. Triebel, "Towards robust visual tracking for on-orbit-servicing with learned segmentation and motion priors," in *2026 IEEE Aerospace Conference*. IEEE, 2026, pp. 1–12.

[20] M. Stoiber, M. Elsayed, A. E. Reichert, F. Steidle, D. Lee, and R. Triebel, "Fusing visual appearance and geometry for multi-modality 6dof object tracking," in *2023 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS 2023*. IEEE, 2023. [Online]. Available: https://elib.dlr.de/195723/