In-Hand 3D Object Scanning from an RGB Sequence

Supplementary Material

Shreyas Hampali[1,3*] Tomas Hodan[1] Luan Tran[1]
Lingni Ma[1] Cem Keskin[1] Vincent Lepetit[2,3]

[1]Reality Labs at Meta
[2]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
[3]Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria

We provide more details and additional results of our method in this supplementary material and in the attached supplementary video. We discuss results on sequences captured with the Aria glasses, additional results on the HO-3D dataset, and describe the limitations of our approach.

1. In-hand Object Scanning with Aria [2]

The recently introduced Aria AR glasses [2] provide a first-person capture of the environment using cameras mounted on the glasses. A head-mounted camera provides an intuitive and simple way for scanning an object using both hands. Here we show that our method can be applied for reconstruction of unknown objects from sequences captured using the Aria glasses. Figure 1 shows the Aria glasses and an egocentric view of two hands manipulating an object from the YCB dataset.

We first linearize the fish-eye images from the Aria sequence and use Detic [4] to obtain the hand and unknown object masks in the images. The reconstruction result on the mustard bottle sequence captured using the Aria glasses shown in Figure 2 demonstrates that our proposed method can be applied to this in-hand scanning scenario as well.

Figure 1. **Aria glasses.** Aria glasses [2] provide egocentric views of the environment using cameras mounted on the glasses.

2. More Qualitative Results

We show more qualitative results from the HO-3D sequences in Figure 3. Corresponding quantitative results are provided in Tables 1-3 of the main paper. Our method can reconstruct partially-textured and texture-less objects such as the mustard bottle and mug in Figure 3. Fingers grasping the object are reconstructed on the mustard bottle sequence (second column of Figure 3) due to inaccurate hand masks and static grasp pose of the hand throughout the sequence. Parts of the object that are always occluded by the hand as in the mug sequence (third column of Figure 3) are also inaccurately reconstructed as we do not assume any prior on the object shape.

3. Shape Regularization Loss

Minimizing the shape regularization loss (Eq. 9 of the main paper) at $t = 0$ encourages the reconstructed object surface to be a plane parallel to the image plane. This can be seen by considering an orthographic projection of the rays and noting that for each object ray $\mathbf{r}$ that passes through an object pixel in the camera at $t = 0$,

$$\begin{aligned} \min \sum_k o_\theta(\mathbf{x}_k) \exp\left(\alpha \cdot \|\mathbf{x}_k\|_2\right) &= \min_k \exp\left(\alpha \cdot \|\mathbf{x}_k\|_2\right) \\ &= \exp \alpha \cdot \left(\min_k \|\mathbf{x}_k\|_2\right) \end{aligned}$$

where the first equality is due to $o_\theta(\mathbf{x}_k) = \{0, 1\} \forall k$. As we assume orthographic projection, the point on the object ray with the minimum euclidian distance to the origin lies in a plane that is parallel to the image plane and passing through the origin. Thus, the for all the object rays the reconstructed surface is on this plane.

[*] Work done as part of Shreyas's PhD thesis at TU Graz, Austria.

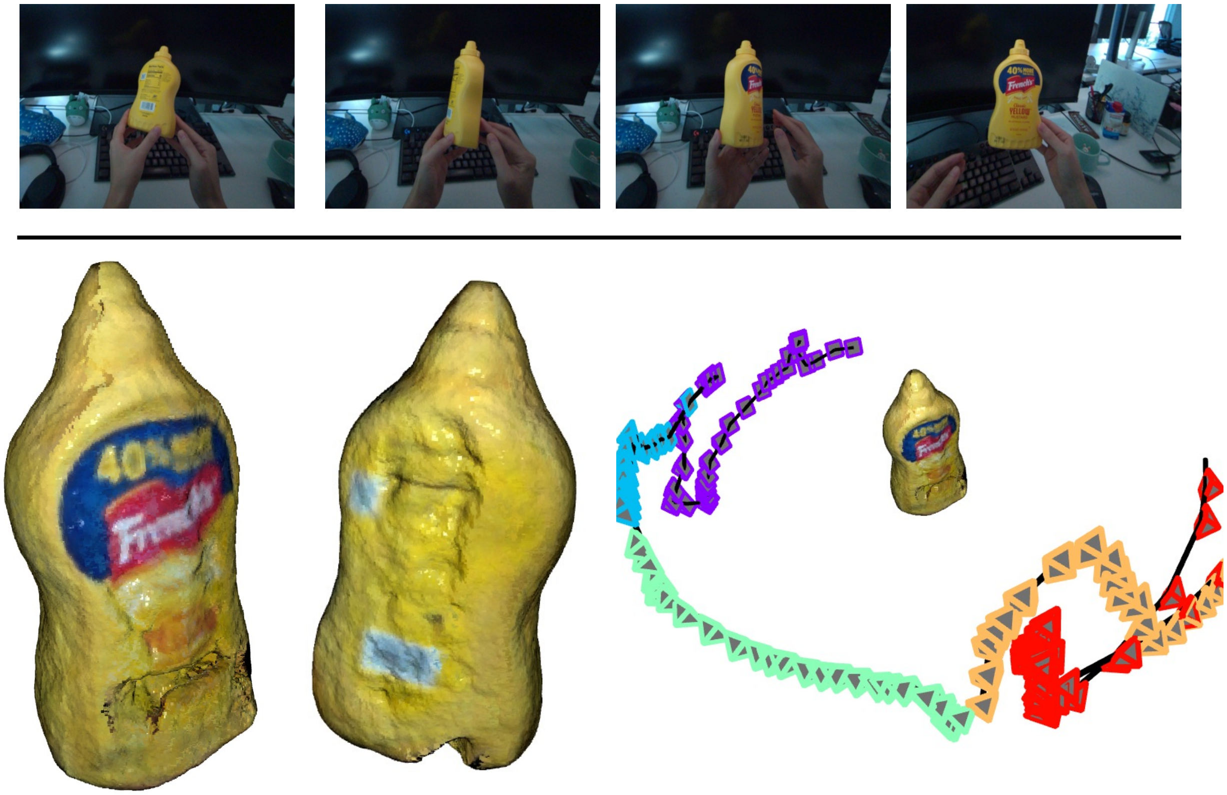


Figure 2. **In-hand object scanning with Aria glasses.** Our method can be applied to reconstruct (bottom left) objects and estimate its poses (bottom right) from RGB sequences captured using Aria glasses (top). The different segments created by our approach is color coded in the pose trajectory.

4. More Implementation Details

As discussed in Section 3.4.1 of the main paper, we divide the input RGB sequence into multiple overlapping segments, and incrementally reconstruct the object shape and estimate its pose in each segment. As reconstructing the object from every frame of the entire RGB sequence is not feasible, we first subsample the input RGB sequence and manually select the frame interval from the entire sequence on which we run our method. The interval is selected such that all parts of the object are visible during scanning. In Table 1, we provide the names of the sequences from the HO-3D dataset which are used for reconstruction, the chosen frame intervals, and the number of segments the RGB sequence is divided into by our method. Additionally, in Figure 6, we show the frame interval on which the reconstruction is performed for two objects in the HO-3D dataset along with the segment boundaries

Object	Sequence ID	Start Frame ID	End Frame ID	No. of Segments
3: cracker box	MC1	210	650	8
4: sugar box	ShSu12	276	1552	8
6: mustard	SM2	300	576	4
10: potted meat	GPMF12	70	334	4
21: bleach	ABF14	686	960	4
35: power drill	MDF14	270	772	3
19: pitcher base	AP13	60	370	4
11: banana	BB12	1100	1260	4
25: mug	SMu1	702	1236	5

Table 1. **Sequence IDs and frame intervals chosen for reconstruction from the HO-3D dataset, and number of segments created by our approach.**

5. Hand and Object Masks

We show some object and hand masks used by our method in Figure 4. We rely on the pre-trained network from [1] which segments dynamic foreground from static background and segments hand and object as one class. We then obtain hand-only masks from [3] and combine with foreground mask from [1] to obtain hand and object masks.

For the Aria sequences, as discussed in Section 1 which do not have static background, we use Detic [4] to obtain hand and object masks.

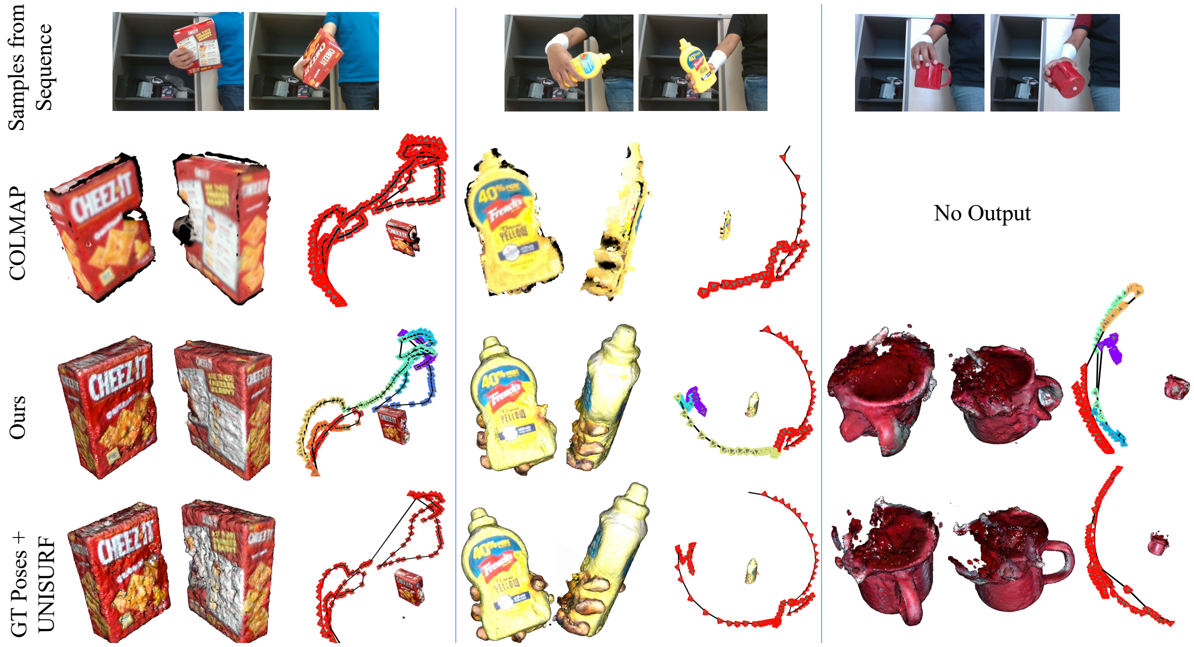
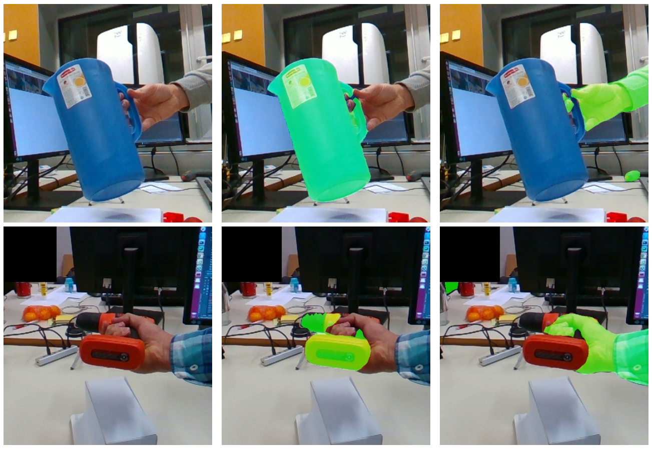


Figure 3. **More qualitative results (reconstructed models and pose trajectories) on HO3D dataset.** Left column: Our method and COLMAP obtain high quality reconstruction on textured objects. Middle column: Our method manages to return a complete reconstruction on this partially textured object, while COLMAP fails to reconstruct the back. The fingers reconstructed as part of the mustard bottle are due to inaccurate hand masks. Third column: We achieve reasonable results on this very challenging texture-less object, on which COLMAP fails completely. We could not reconstruct the parts of the object that are always occluded by the hand. The reconstruction quality of our method is similar to the quality obtained when using the ground truth poses.



(a) Image (b) Object mask (c) Hand mask

Figure 4. **Hand and object segmentation masks.** We obtain foreground masks from [1] and hand masks from [3].

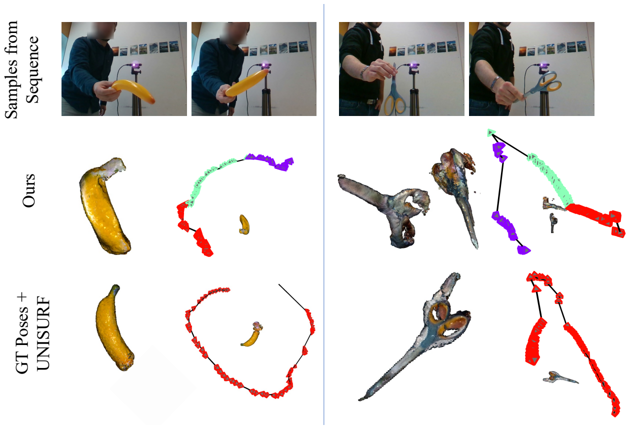
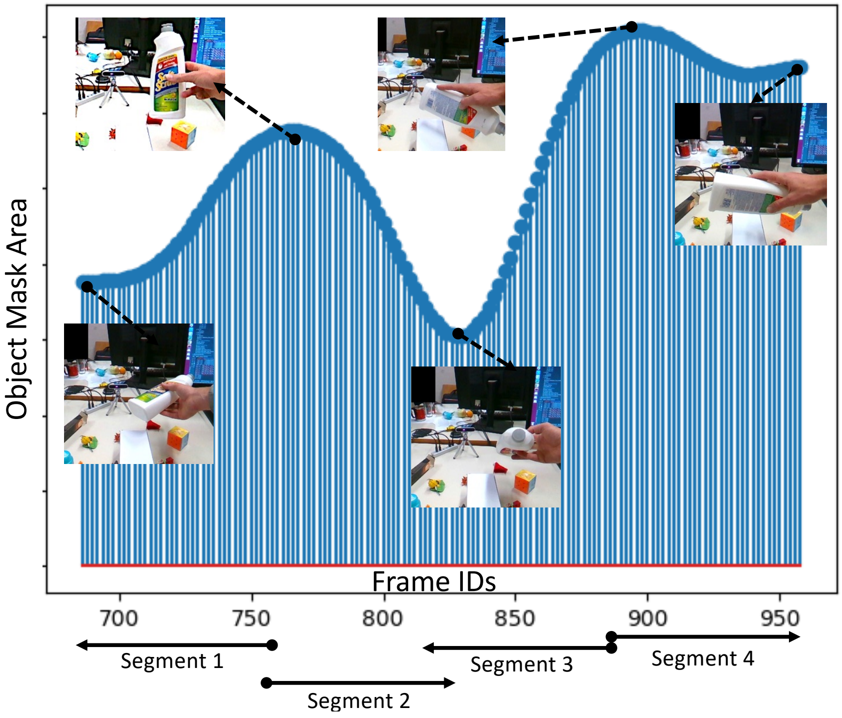


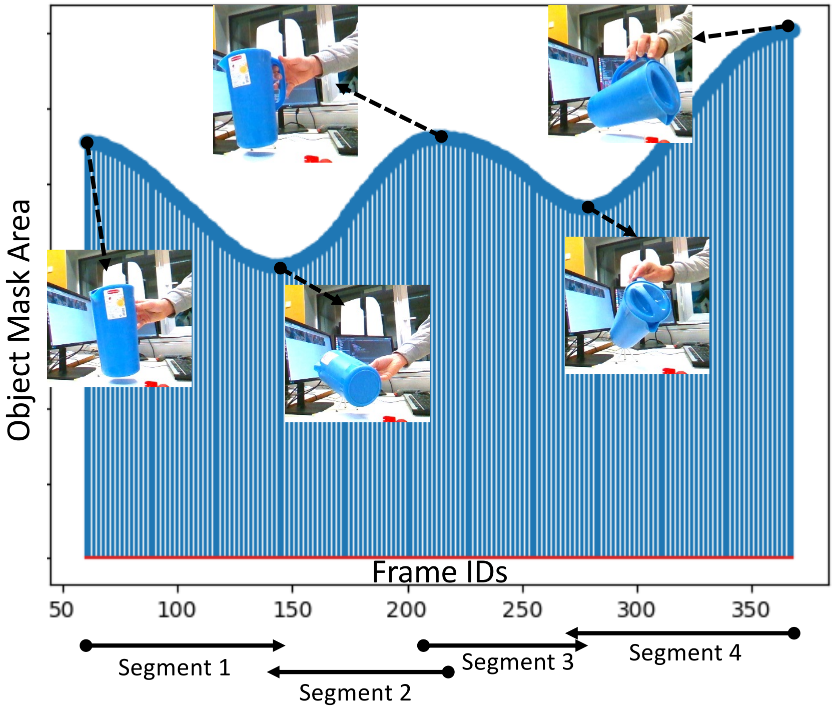
Figure 5. **Failure scenarios.** Our method fails to obtain poses and reconstruct texture-less symmetrical (left) and thin objects (right).

6. Limitations

Our method relies on the geometric or texture features on the object to incrementally reconstruct and estimate its pose within a segment.The proposed approach results in inaccurate pose estimates for texture-less and nearly symmetrical objects such as banana leading to erroneous reconstruction as shown in Figure 5. Our method also fails to estimate poses of thin objects such as scissors leading to inaccurate reconstructions as also shown in Figure 5. We believe hand pose information can provide additional cues to estimate the object poses during more challenging scenarios and is a potential future direction for our approach.



(a) Bleach bottle



(b) Pitcher base

Figure 6. **Object area curves and segment boundaries.** We show the segment boundaries for two objects (bleach bottle and pitcher base) which are calculated from the object area curves. In each segment, the incremental object reconstruction and pose tracking starts at the local maximum of the object area and ends at the local minimum.

References

[1] Wout Boerdijk, Martin Sundermeyer, Maximilian Durner, and Rudolph Triebel. What's This?" - Learning to Segment Unknown Objects from Manipulation Sequences. In *International Conference on Robotics and Automation*, 2021. 2, 3

[2] Zhaoyang Lv, Edward Miller, Jeff Meissner, Luis Pesqueira, Chris Sweeney, Jing Dong, Lingni Ma, Pratik Patel, Pierre Moulon, Kiran Somasundaram, Omkar Parkhi, Yuyang Zou, Nikhil Raina, Steve Saarinen, Yusuf M Mansour, Po-Kang Huang, Zijian Wang, Anton Troynikov, Raul Mur Artal, Daniel DeTone, Daniel Barnes, Elizabeth Argall, Andrey Lobanovskiy, David Jaeyun Kim, Philippe Bouttefroy, Julian Straub, Jakob Julian Engel, Prince Gupta, Mingfei Yan, Renzo De Nardi, and Richard Newcombe. Aria pilot dataset. https://about.facebook.com/realitylabs/projectaria/datasets, 2022. 1

[3] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. SeqFormer: A Frustratingly Simple Model for Video Instance Segmentation. In *arXiv Preprint*, 2021. 2, 3

[4] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In *European Conference on Computer Vision*, 2022. 1, 2