SHOWMe: Benchmarking Object-agnostic Hand-Object 3D Reconstruction
Supplementary Material

Anilkumar Swamy[1,2] Vincent Leroy[1] Philippe Weinzaepfel[1] Fabien Baradel[1] Salma Galaaoui[1]
Romain Brégier[1] Matthieu Armando[1] Jean-Sebastien Franco[2] Grégory Rogez[1]

[1]NAVER LABS Europe [2]Inria centre at the University Grenoble Alpes

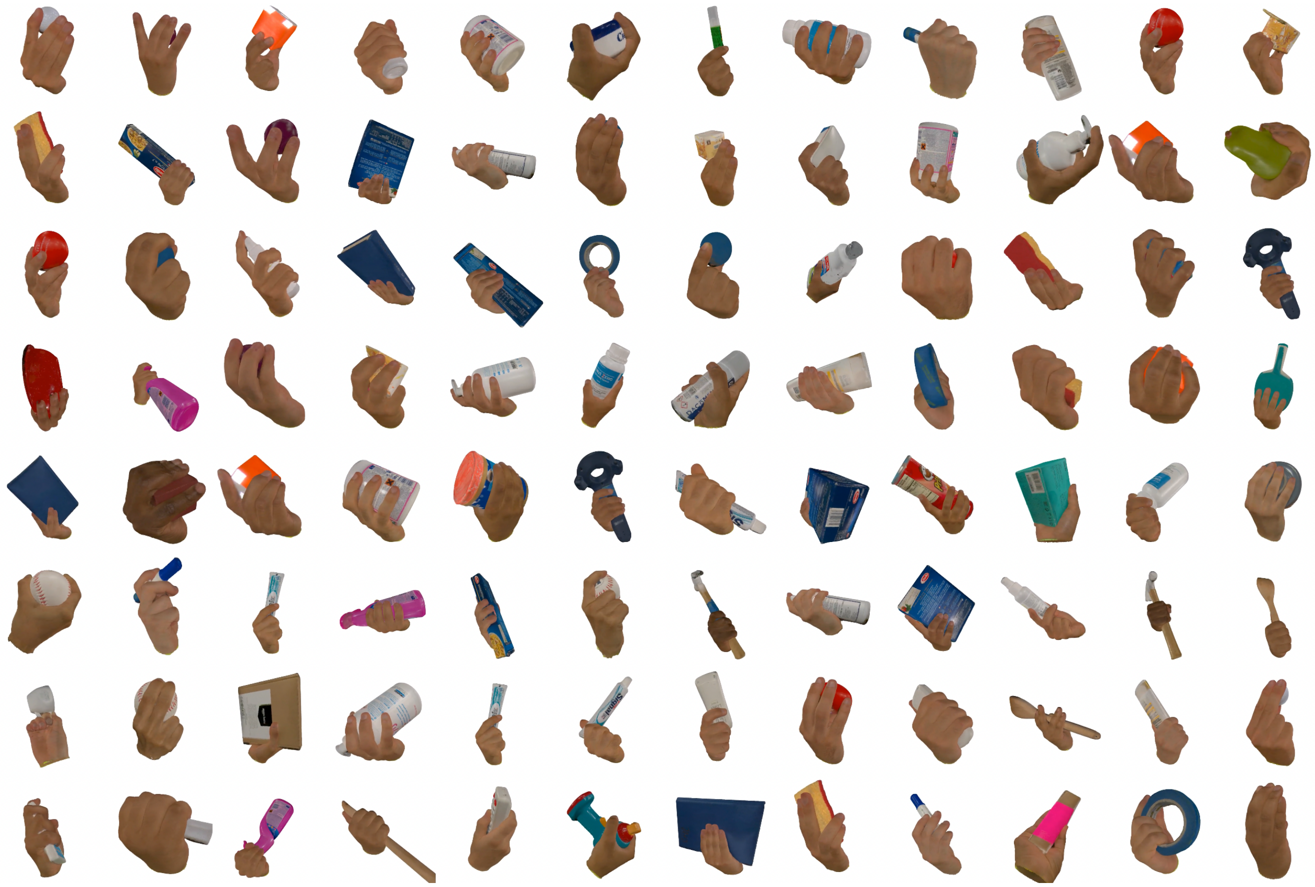


Figure 1: **Snapshot from the supplementary video of the SHOWMe dataset** with the ground-truth textured meshes it contains.

In this supplementary material, we provide more details about the introduced SHOWMe dataset in Section 1. Section 2 then provides a more detailed evaluation analysis of the results of our two-stage pipeline on the SHOWMe dataset, as well as the method based on aggregated single-view reconstruction from IHOI [8].

1. Additional Dataset Information

1.1. Qualitative Visualizations

The attached video, from which we show a snapshot in Figure 1, highlights the ground-truth 3D shapes with textures that our introduced SHOWMe dataset contains.

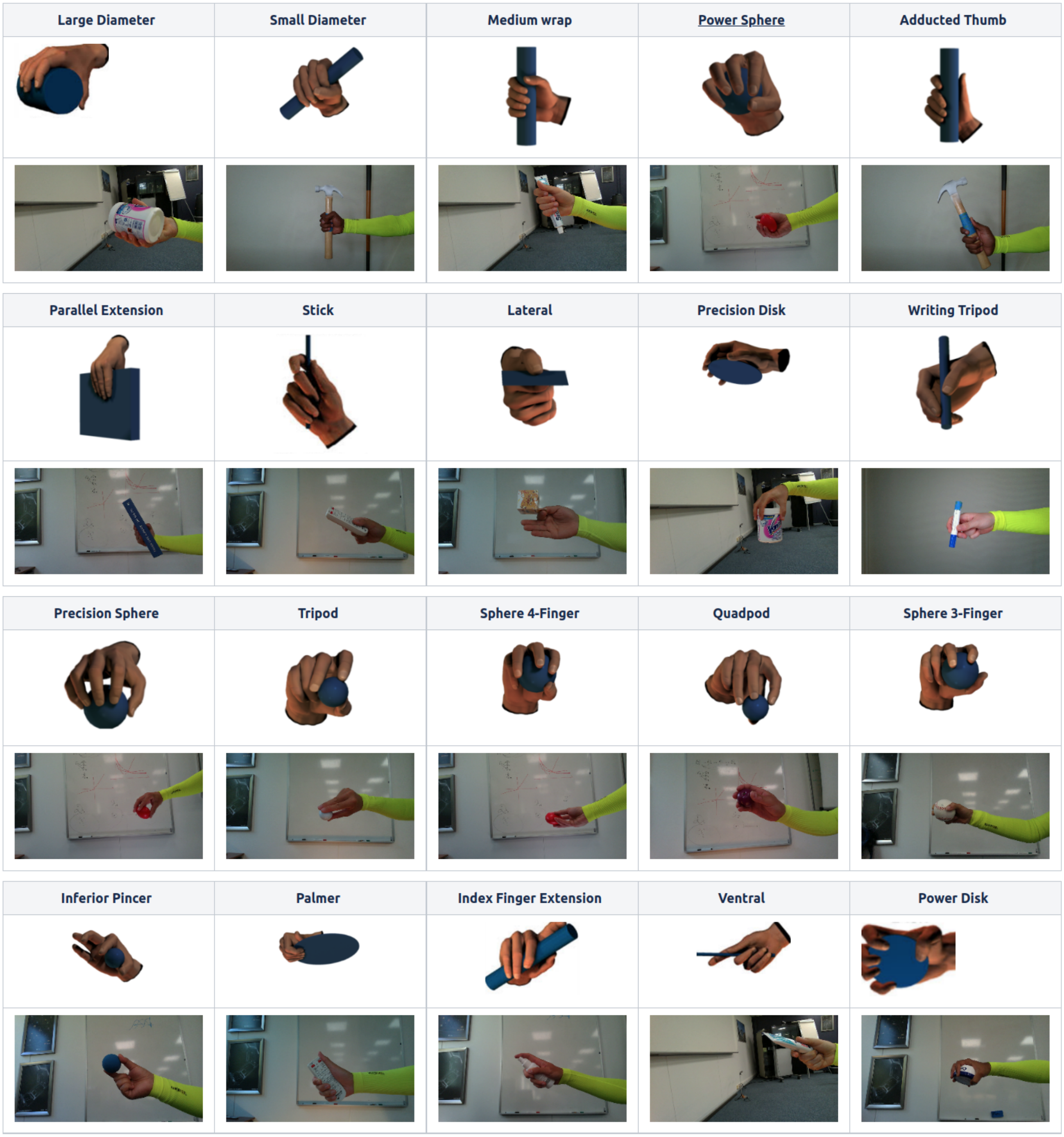


Figure 2: **Grasp categories within SHOWMe**.

With the scanner accuracy and resolution up to 0.1mm and 0.2mm respectively, the #vertices and #faces depend on the size and shape of the objects. Across all GT meshes, #vertices and #faces are in the range (4K to 263K) and (7.5K to 524K) respectively.

1.2. Grasp Variability

SHOWMe offers a large variability in terms of grasp types as depicted in Figure 2 where 20 of our hand-object interactions can be classified into different classes of the 33-grasp taxonomy introduced in [1].

1.3. Rigid transformation - Optimisation details

As explained in Section 3.2 of the main paper, we refine the camera poses such that the projection of the GT colored mesh matches the RGB observations for each frame and minimize a loss expressed in Equation (5) of the main paper.

We implement this optimization in PyTorch [4]. We represent the scene with a sparse voxel grid of resolution 128^3, that matches the bounding box of $\mathcal{M}$. We use the Adam optimizer [3] for 250 iterations, with $lr_{rgb} = 5.10^{-1}$ and $lr_{cam} = 5.10^{-3}$ the learning rates for the appearance and the camera poses respectively. Because it is not computationally tractable to fully render thousand of frames at every iteration, we randomly sample 500 rays from each camera at each iteration. We grid search various λ_{Smooth}, λ_{wd} weighting parameters and keep for each sequence the set of hyperparameters that gave the best (lowest) $\mathcal{L}_{RGB}$ value, considering that this metric is a relevant proxy for the quality of the pose annotations.

2. Detailed Evaluations of Baselines

In this section, we present in Sec. 2.1 additional qualitative results of the various baselines. We then provide in Sec. 2.2 more detailed and comprehensive experiments on the different steps of these baselines, before discussing single-image reconstruction results in Sec. 2.3.

2.1. Qualitative Results

We provide in Fig. 3 successful reconstructions obtained with VH, FDR and HHOR methods, using camera poses obtained from COLMAP. HHOR recovers much more details for both the hand and the object compared to VH and FDR.

2.2. Detailed analysis

In this section, we analyze the success and failure cases of our two baselines used for rigid transformation estimation which plays a crucial role in the final 3D reconstruction quality. COLMAP [6] camera pose estimation toolbox takes the set of overlapping images of the same object from different viewpoints and estimates the intrinsic and extrinsic camera parameters. This involves 2D images feature detection, extraction, and feature-matching steps. Figure 6 shows the example of detected features in two viewpoints and corresponding matched features. However, we observe that COLMAP fails to estimate the camera poses on roughly 20% of the sequences. An example of complete failure cases can be seen in Fig. 4 where COLMAP could not successfully output camera poses. This figure also shows the reconstruction of DOPE+FDR and DOPE +HHOR, and illustrates the robustness of a hand-based camera estimation approach, at the cost of a loss in reconstruction quality. In other cases, camera poses are estimated by COLMAP but are not sufficiently accurate to perform reasonable reconstruction as seen in Fig. 5.

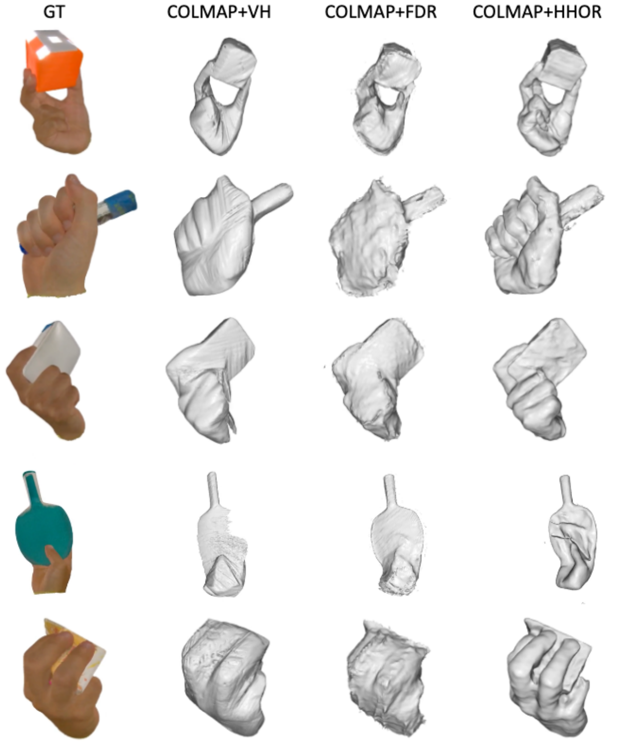


Figure 3: **Hand-Object Reconstructions** using COLMAP for camera pose estimation followed by VH, FDR and HHOR methods for reconstruction. *From left to right:* GT, COLMAP+VH, COLMAP+FDR, COLMAP+HHOR. Please see the video for more qualitative results.

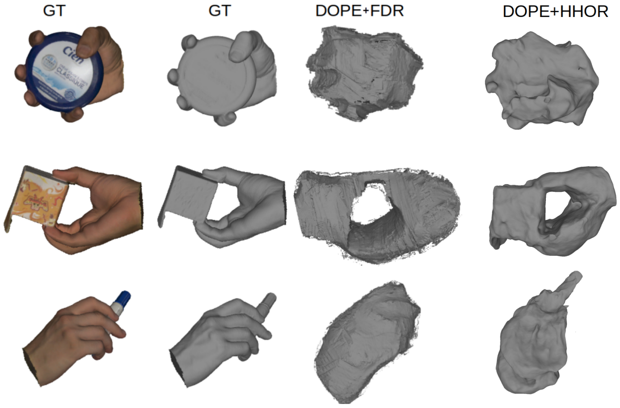


Figure 4: **Hand-Object Reconstruction using DOPE+FDR and DOPE+HHOR.**For these examples, COLMAP could not estimate any valid camera poses.

Object Size Analysis. Upon careful analysis in the main paper, we found that reconstructions based on COLMAP failed or performed poorly on objects of small size compared to larger-size objects while using DOPE led to better performance for small objects. To corroborate this, we

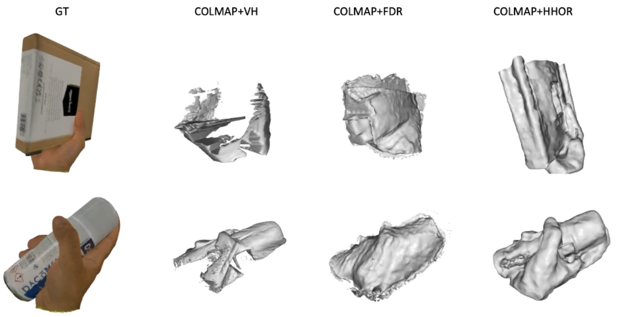
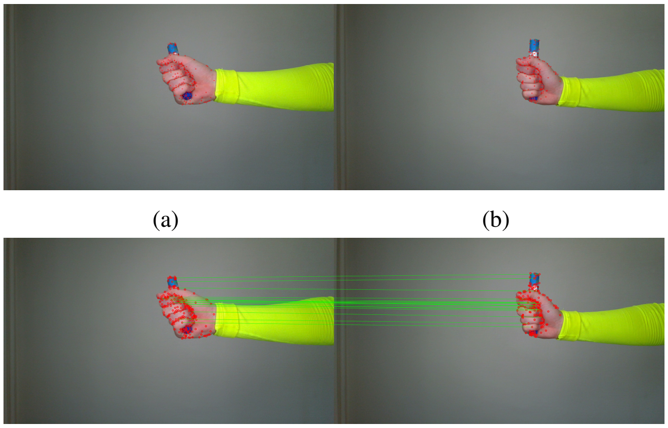


Figure 5: **Hand-Object Reconstruction failure examples** obtained using COLMAP as pose estimator. The poses estimated are not accurate enough to allow for reasonable reconstruction results.



(a) (b)

(c)

Figure 6: **Visualization of COLMAP detected features** on (a) frame 1 and (b) frame 264. We show the corresponding matched features in (c).

manually label each object as small or large and compute rigid transformation estimation qualities and reconstruction errors on the two sets (small size set and larger size set) of objects. Table 1 and 2 shows the rigid transformation accuracy averaged across sequences within SHOWMe for three different level of camera pose precision (*i.e.*, after binning camera estimates with respect to translation and rotation errors). COLMAP shows comparable average performance for both types of objects but a much higher variance in camera pose quality for small objects, meaning that it completely fails for some sequences. On the other hand, DOPE clearly provides more accurate camera estimates for small objects. Using HHOR as a reconstruction method does not compensate for mediocre camera estimates and the same observation can be made (Table 3).

object size	Per-Frame Relative Pose Quality (%) ↑		
	@(2cm&4°)	@(5cm&10°)	@(10cm&20°)
small	9.96±15.60	36.90±33.57	74.50±31.20
larger	6.58±5.17	28.10±20.85	69.50±25.21

Table 1: **Rigid transformation evaluation obtained from COLMAP [6]** with the per-frame relative pose quality, *i.e.* the percentages of frames where the error is below a given threshold, averaged across sequences.

object size	Per-Frame Relative Pose Quality (%) ↑		
	@(2cm&4°)	@(5cm&10°)	@(10cm&20°)
small	2.06±2.13	14.20±9.59	40.10±19.70
larger	1.70±1.82	9.75±6.66	30.80±17.50

Table 2: **Rigid transformation evaluation obtained from hand joints estimates from DOPE [7]** with the per-frame relative pose quality, *i.e.* the percentages of frames where the error is below a given threshold, averaged across sequences.

Object Texture Analysis. We also analyze the success and failure cases of our baselines used for rigid transformation estimation and 3D reconstruction with respect to the texture quality of the objects in the sequence. We have empirically observed that COLMAP camera pose estimation quality reduces drastically for less textured objects, *i.e.*, with a rather uniform texture. In order to verify this observation, we performed an experiment where we manually categorize all the hand-object sequences into two types "*textured*" and "*less-textured*" objects and evaluate the rigid transformations estimated with COLMAP in Table 4. COLMAP shows comparable average performance for both types of objects but a much higher variance in camera pose quality for less-textured objects, meaning that it completely fails for some sequences. We also evaluate the reconstruction baseline COLMAP+HHOR for the two sets and found that acc. ratio, comp. ratio and Fscore all degrade by roughly 15% as shown in Table 5.

Additional Analysis. In Figure 7, we show qualitative results of the DOPE hand keypoint detector, in both successful (left) and failure cases (right). We can see that in the case of a frame without large occlusion, the hand pose is well estimated. However, when large occlusions occur, the detector wrongly estimates the hand pose, which impacts the camera pose estimation.

Finally, we provide a detailed study of the F-Score over the whole dataset for various setups in Fig. 8. We can see that both MVS methods, VH and FDR attain good Fscore using the annotated poses (GT) while IHOI tends to lack accurate and complete reconstructions with the 5mm threshold. Also, DOPE-based rigid transformation estimations tend to provide worse reconstruction scores but are much more reliable than COLMAP, which fails completely on more sequences.

method	object size	acc. ratio @5mm (%) ↑	comp. ratio @5mm (%) ↑	Fscore @5mm (%) ↑
COLMAP+HHOR	small	32.87	33.57	33.07
	larger	**54.35**	**56.14**	**55.00**
DOPE+HHOR	small	**40.87**	**42.48**	**40.96**
	larger	37.67	41.03	38.73

Table 3: **Hand-Object Reconstruction evaluation of HHOR [2] vs object size**.

texture type	Per-Frame Relative Pose Quality ↑		
	@(2cm&4°)	@(5cm&10°)	@(10cm&20°)
textured	6.69±4.6	29.2±21.1	71.5±25.9
less textured	9.92±15.8	33.7±30.9	73.1±28.5

Table 4: **Rigid transformation evaluation obtained from COLMAP [6]** with the per-frame relative pose quality, *i.e.* the percentages of frames where the error is below a given threshold, averaged across sequences.

texture type	acc. ratio @5mm (%) ↑	comp. ratio @5mm (%) ↑	Fscore @5mm (%) ↑
textured	**51.87**	**47.64**	**49.47**
less-textured	36.19	33.06	34.46

Table 5: **Hand-Object Reconstruction evaluation of COLMAP + HHOR [2] vs object texture**.

2.3. Single-image HO reconstruction

We provide in Fig. 9 some qualitative reconstruction results on a few objects of the dataset. In Section 5.2 of the main paper, we compare the single-view reconstruction method of Ye *et al*. [8] (IHOI) to multiview baselines that take the whole video sequence as input. For a fairer comparison, we slightly modify IHOI for video processing: the method is run independently on multiple, evenly-distributed frames of the sequence. From these single-image reconstructions, we extract the raw object-SDFs (signed distance functions), prior to the meshing step, and average them directly in the canonical reconstruction space, centered around the wrist. Figure 10 shows the effect of this temporal aggregation of SDFs on different objects. In Table 6, we show a quantitative comparison of this modified approach, dubbed IHOI+temp, with baseline IHOI, following the evaluation setup presented in Section 5.2 (and Table 3) of the main paper.

3. Limitations

As stated in the main paper, we assume that the hand remains static with respect to the object i.e. the hand pose remains the same with respect to the object throughout the sequence. This assumption enabled to do category-agnostic hand-object reconstruction. However, this limits the reconstruction baselines to dynamic hand-object manipulation scenarios. In summary, this assumption is reasonable in terms of application and an important step towards dynamic object-agnostic HO reconstruction.

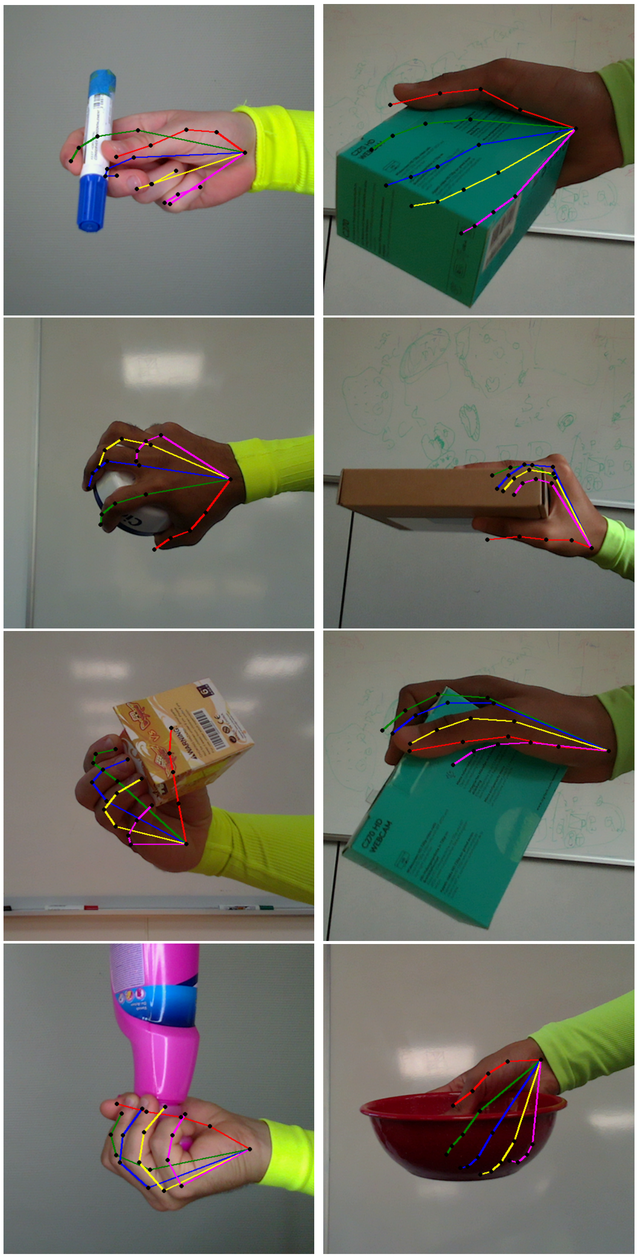


Figure 7: **2D hand keypoints results from DOPE.** The left column shows successful examples, typically when there is little occlusion by the object, while the right column shows overall failure cases which occur when the objects largely occlude the hand.

References

[1] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic. A comprehensive grasp taxonomy. In *Robotics, Sci-*

Reconstruction Method	Acc. (cm) ↓	Comp. (cm) ↓	Acc. ratio @5mm (%) ↑	Comp. ratio @5mm (%) ↑	Fscore @5mm (%) ↑
IHOI [8]	0.798	**1.344**	41.77	**37.80**	**39.32**
IHOI+temp	**0.757**	1.580	**42.88**	34.93	38.08

Table 6: **Comparison of baseline IHOI [8] with our temporal extension (IHOI+temp).** On average, combining SDFs from multiple frames improves accuracy but degrades completeness.

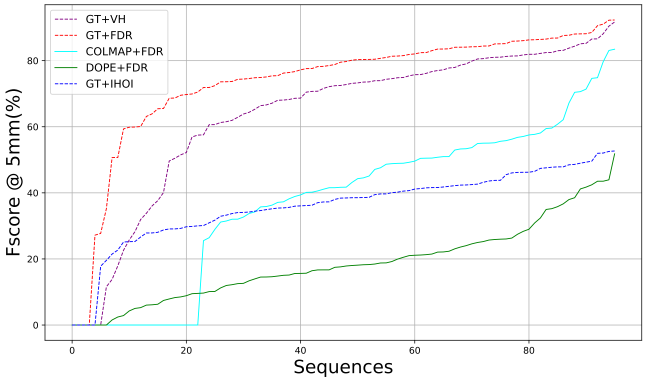


Figure 8: **Fscore@5mm over the dataset sorted by increasing order for various methods.** The closer the curve is to the top-left, the better the method.

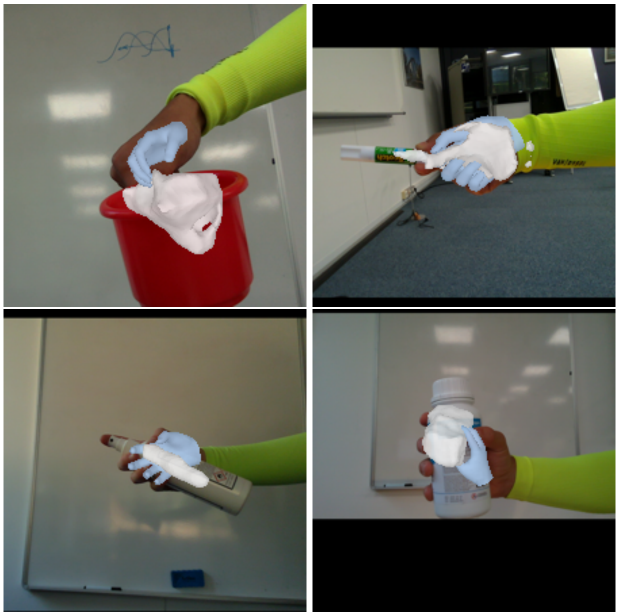


Figure 9: **Failure cases of single-image reconstructions** using FrankMocap [5] and [8]. The method of [8] struggles with unusual objects or grasps, such as an object held at the tip of the fingers or large objects.

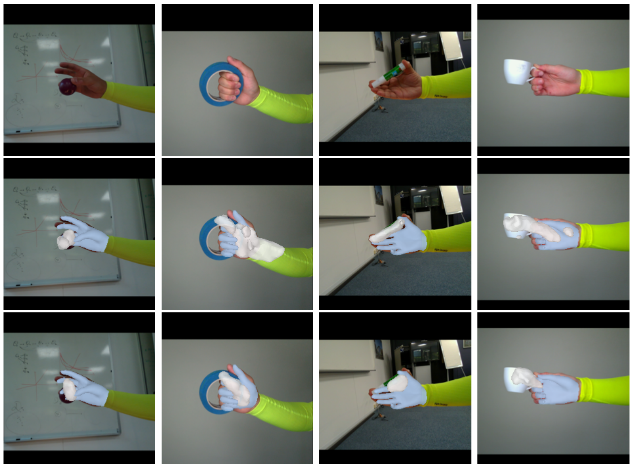


Figure 10: **Fusing single-image reconstructions** contributes to smoothing the results, but sometimes degrades the reconstruction. *Top:* input image. *Middle:* result of a single-image reconstruction method [8]. *Bottom:* result from fusing SDFs produced by [8] on several frames of a given sequence. SDFs are estimated in a canonical hand space (with a fixed wrist position), and fused directly in that space.

ence and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation, 2009. 2

[2] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia*, 2022. 5

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019. 3

[5] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021. 6

[6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 3, 4, 5

[7] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild. In *ECCV*, 2020. 4

[8] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 1, 5, 6