Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos

Zhifan Zhu Dima Damen
School of Computer Science, University of Bristol, UK
https://zhifanzhu.github.io/getagrip

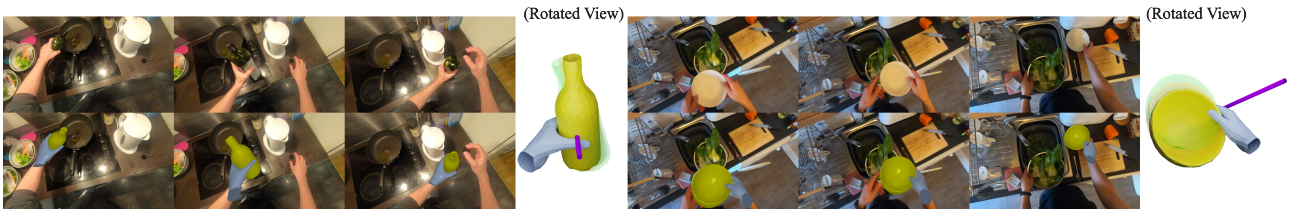

Figure 1. Two stable grasp sequences from EPIC-Grasps for a bottle (left) and bowl (right). We show sample frames (top) and reconstructions (bottom). Right: for each reconstruction, we show the rotated view, along with the latent 1-DoF axis.

Abstract

We propose the task of Hand-Object Stable Grasp Reconstruction (HO-SGR), the reconstruction of frames during which the hand is stably holding the object. We first develop the stable grasp definition based on the intuition that the in-contact area between the hand and object should remain stable. By analysing the 3D ARCTIC dataset, we identify stable grasp durations and showcase that objects in stable grasps move within a single degree of freedom (1-DoF). We thereby propose a method to jointly optimise all frames within a stable grasp, minimising object motions to a latent 1-DoF. Finally, we extend the knowledge to in-the-wild videos by labelling 2.4K clips of stable grasps. Our proposed EPIC-Grasps dataset includes 390 object instances of 9 categories, featuring stable grasps from videos of daily interactions in 141 environments. Without 3D ground truth, we use stable contact areas and 2D projection masks to assess the HO-SGR task in the wild. We evaluate relevant methods and our approach preserves significantly higher stable contact area, on both EPIC-Grasps and stable grasp sub-sequences from the ARCTIC dataset.

1. Introduction

In this work, we focus on the task of reconstruction in-the-wild on temporal periods of stable grasps. This work contains three components. **First,** we propose the task of Hand-Object Stable Grasp Reconstruction (HO-SGR) which jointly optimises the reconstructions across all frames within one stable grasp. We showcase that objects move within one degree of freedom (1-DoF), relative to the hand pose, throughout the stable grasp. **Second,** we accordingly propose a method that jointly reconstructs the hands and objects by minimising the object's motion, relative to the hand, to 1-DoF around a latent rotation axis, throughout the frames. We demonstrate our method outperforms baselines and alternative assumptions of object movement using 3D ground truth from the stable grasps within the egocentric views of the ARCTIC dataset [5]. **Third**, We label a sizeable dataset of 2.4K stable grasps clips from egocentric videos. Our EPIC-Grasps dataset is the first for hand-object reconstruction collected from unscripted activities, with individuals grasping 390 different objects by both hands. Similar to previous works [1, 2, 7, 8, 13], we restrict our evaluation to known category CAD models. Our dataset comes with pseudo-ground truth in the form of 2D segmentation masks available from [4], allowing to measure the 3D reconstruction's projection relative to this 2D ground truth.

2. Related Works

3D Hand Pose Estimation. Estimating 3D hand pose from RGB images has been proposed for both free hands and hands in-interactions, these methods [12, 14, 15, 17] are used as building blocks of hand-object reconstruction methods [1, 10, 13, 21, 22].

3D Hand-Object-Reconstruction. Methods are grouped into two categories. The first category, known-CAD methods [1, 10, 13, 18–20], assumes that object CAD models are given and fits 3D shapes into 2D observations. The second category, CAD-agnostic methods [3, 6, 9, 11, 21, 22], aims to estimate the hand and object poses without using explicit CAD models.

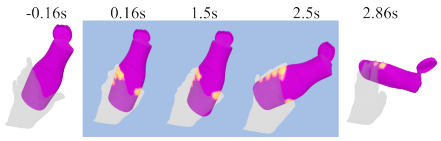

Figure 2. Sample hand-object mesh sequence from ARCTIC. Contact areas (in shiny yellow) are similar within the stable grasp (blue background). In -0.16s the hand has no contact with the object.

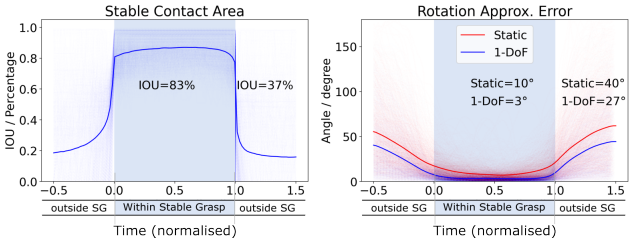

Figure 3. We compare within/outside grasps, analysing object in-contact area (left) and corresponding rotation errors of the static and 1-DoF rotation approximations (right), normalising all stable grasp duration for direct comparison (0 to 1 marked with blue background). While both the Static and 1-DoF assumptions result in low approximation error within stable grasps, the error of 1-DoF assumption is marginal (right).

3. Stable Grasp Reconstruction: Problem and Method

3.1. What is a *stable grasp (SG)*?

Definition. Formally, for any pair of frames i and j within an interaction sequence, we use S_i and S_j to denote the in-contact (by a hand) area on the object surface, and intersection-over-union $\text{IOU}(S_i, S_j)$ between in-contact areas. Following the intuition that the hand maintains a stable contact area with the object during the stable grasp, the duration of the stable grasp is defined as

$$[l^*, r^*] = \underset{l,r}{\text{argmax}}\,(r-l) \quad \text{s.t. IOU}(S_i, S_j) > \tau \;\; \forall l \le i < j \le r \quad (1)$$

where τ specifies the minimum IOU threshold. The $argmax(r-l)$ implies the longest duration representing the stable grasp, from its initiation to conclusion. Fig. 2 visualises an example of stable grasp.

Stable Grasp Study. We perform a study on the 3D-ground-truth dataset ARCTIC [5] to analyse different quantitative measures within/outside the temporal extent of the stable grasp. We use the threshold $\tau = 0.5$ (Eq 1) and automatically extract 1303 stable grasp sequences from a variety of objects and subjects throughout dataset.

We present the main finding in Fig 3. As anticipated, we show that the in-contact IOU, drops sharply outside the stable grasp – plotted in Fig 3 (left). That is the contact area remains stable only within the stable grasp's temporal extent. When we assess the **relative** object motion to the hand coordinate system within/outside the stable grasp, we note that the in-hand object motion can be approximated within single degree-of-freedom, plotted in Fig. 3 (right). Formally, we define a rotation axis ϕ around which the object can rotate. If an object is only allowed to rotate around this axis, the motion is restricted from its free 6-DoF to a single rotation angle around this axis – we thus refer to this as a 1-DoF motion. The object pose w.r.t. the hand would then be described as

$$T^n_{o2h} = \textbf{rot}\,(\omega^n, \phi) \circ T_{o2h} \quad (2)$$

i.e., we first apply one global object-to-hand transform T_{o2h} for all frames, followed by applying the per-frame rotation $\textbf{rot}\,(\omega^n, \phi)$ – the rotation of angle ω^n around the given rotation axis ϕ. Here $\circ$ denotes composition of transformations. In Fig. 3 (right) we plot the 1-DoF approximation error within/outside the temporal extent of stable grasp and compare that to the assumption that objects remain static (i.e. does not move relative to the hand). We show that the angle error when using the static assumption to be non-negligible (avg 10°). When we use the 1-DoF assumption, the error off the rotation axis ϕ is generally low (avg 3°).

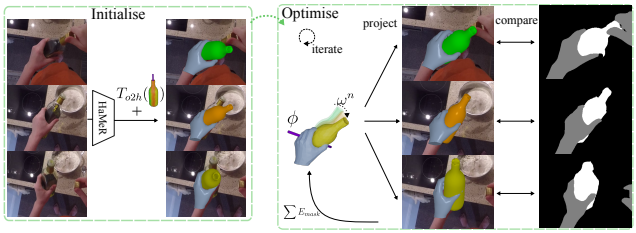

Figure 4. Our proposed reconstruction method. We show 3 frames within a stable grasp. HaMeR [14] produces the hand meshes (rendered in blue) from RGB, and we set the object-to-hand pose T^n_{o2h} to the same T_{o2h} initially. Then, during each iteration of the optimisation, the object's relative pose is optimised to 1-DoF and projected back to individual frames. These are compared with ground truth segmentation (right), jointly optimise for all frames. We ignore mask computation in hand occluded region (grey in the right figure). The physical terms are omitted in this figure.

3.2. Reconstructing Object Poses in a Stable Grasp

Given a start-end segment of a stable grasp, we aim to produce **consistent** hand object reconstructions across all N frames within the stable grasp. More specifically, we aim to produce, for every frame n, a pair of 3D meshes of the hand and the object w.r.t. the camera.

Following prior works [1, 10, 13], we use MANO [16] to represent the **hand mesh**. We use HaMeR [14], to obtain the finger articulations θ^n from individual frames. We thus have per-frame hand vertices $V^n_h = \text{MANO}(\theta^n)$. Additionally, the hand-to-camera (*h2c*) pose $T^n_{h2c} \in SE(3)$ which is

defined as the hand wrist pose, is produced by HaMeR by default. The T^n_{h2c} will be used to transform the hand and the object from the *hand coordinate system* to the *camera coordinate system* for each frame.

For the **object mesh**, we assume the category-level object CAD model is known, and denote the object vertices as $V_o \in \mathbb{R}^{|V_o| \times 3}$. We will reconstruct the object-to-hand ($o2h$) poses $T^n_{o2h} \in SE(3)$ and the scalar scale $s \in \mathbb{R}$, which transform object vertices to $V^n_{o:h}$ in the hand coordinate system for each frame. To render the object back to the images, we then use the hand-to-camera ($h2c$) pose T^n_{h2c} to transform $V^n_{o:h}$ to $V^n_{o:c}$ in the camera coordinate system:

$$V^n_{o:c} = T^n_{h2c}(T^n_{o2h}(s * V_o)) \tag{3}$$

We reconstruct the object-to-hand poses T^n_{o2h} with the *render-and-compare* approach, overviewed in Fig. 4. We propose to leverage our finding of 1-DoF motion from the stable grasp study (Sec. 3.1), to optimise the consistent object pose relative to the hand. As shown in Eq. (2), the 1-DoF object relative pose T^n_{o2h} at frame n is given by per-frame rotation ω^n, the rotation axis ϕ and the global base object-to-hand pose T_{o2h} that takes the object into the hand.

Our main objective function is given by:

$$E(\underbrace{\phi, \{\omega^n\}, T_{o2h}}_{\{T^n_{o2h}\}}, s; \{\theta^n\}, \{T^n_{h2c}\}) = \sum_{n=1}^{N} \lambda_1 E^n_{mask} + \lambda_2 E^n_{push} + \lambda_3 E^n_{pull} \tag{4}$$

where $\{\theta^n\}$ and $\{T^n_{h2c}\}$ are sets of outputs from HaMeR and are kept fixed, and s is the scalar object scale to be optimised.

We initialise the latent axis ϕ to the object's z-axis from the CAD model; $\{\omega^n\}$ are initialised to zero. We note the initialisations for T_{o2h} in implementation details. We then jointly optimise all parameters across frames, in particular the rotation axis ϕ and the per-frame rotation angles $\{\omega^n\}$.

We use three terms in the optimisation: $E^n_{mask}, E^n_{push}, E^n_{pull}$. The main term E_{mask} focuses on estimating a reconstruction that best matches the 2D projections of the object masks throughout the sequence. We measure the error via sum of pixel differences:

$$E^n_{mask} = |\mathcal{C}^n_o \otimes (\mathcal{M}^n_o - \Pi(V^n_{o:c}))|^2_2 \tag{5}$$

where $\mathcal{M}^n_o$ is the object mask which we use for supervision and $\Pi(\cdot)$ is the differentiable projection function. $\mathcal{C}^n_o$ is the occlusion-aware mask as in [10, 23] which only computes the error within regions of the object that are not occluded by the hand, set to 1 for the object and the background, and 0 for the hand. This masking is critical to avoid penalising the missing parts of the object mask due to in-hand occlusion.

We also employ two additional terms, used in previous works [1, 10, 13, 19]. We use the physical heuristics E_{push}, which pushes the object out of the penetrating region against the hand and a balancing loss E_{pull} which pulls the object to touch these contact regions.

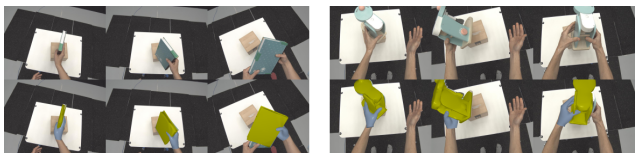

Figure 5. Two stable grasp reconstructions by 1-DoF method on ARCTIC-Grasps.

4. Results

We use 4 baselines to compare to: (i) **HOMan** [10], a common CAD-based baseline that progressively optimises the object pose relative to the hand; (ii) **Single Frame**, independent optimisation for each frame; (iii) **Static**, objects are not allowed any motion; (iv) **Dynamic**, objects are allowed to move freely within 6-DoF.

We propose quantitative metrics to measure the correctness of the predicted object poses within the stable grasp. **Average Distance (ADD %)** measures the distance of corresponding vertices between GT and predicted object vertices in the hand coordinate system. ADD is 1 for a sequence if the average distance is less than 10% of the object's diameter, and 0 otherwise.

Average Stable Contact Area at ADD Success (SCA-ADD %). When a pose is considered correct for a sequence, i.e. ADD is 1, we measure the stable contact area across the sequence, defined as the average IOU of in contact area between each pair of frames (Sec. 3.1). SCA-ADD is set to 0 when ADD is set to 0 (average distance below threshold). We average SCA-ADD over all examples.

Intersection-over-Union (IOU %). We use IOU as a proxy of pose accuracy when 3D GT is not available. We measure the IOU between the ground truth mask and the rendered mask for the object in camera view. We report average IOU across all frames.

Tab. 1 compares results on ARCTIC-Grasps, using all the metrics. We show qualitative examples in Fig. 5.

Tab. 2 compares results on EPIC-Grasps dataset using proxy metrics IOU and SCA-IOU at two thresholds (0.8 and 0.6). We compare 1-DoF to the best variation from ARCTIC-Grasps: Dynamic. 1-DoF achieves the best SCA-IOU metric for both thresholds for every object category and overall.

References

[1] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pages 12417–12426, 2021. 1, 2, 3

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021. 1

Category	Single Frame				HOMan [10]				Static				Dynamic				1-DoF			
	IOU	SCA-IOU	ADD	SCA-ADD	IOU	SCA-IOU	ADD	SCA-ADD	IOU	SCA-IOU	ADD	SCA-ADD	IOU	SCA-IOU	ADD	SCA-ADD	IOU	SCA-IOU	ADD	SCA-ADD
box	93.2	17.5	9.4	3.1	83.5	30.9	33.3	16.9	90.9	67.3	37.7	28.7	95.2	41.7	57.2	25.4	94.2	56.7	39.1	23.6
capsulemachine	79.3	17.8	8.4	3.3	84.0	34.4	42.1	24.5	79.8	34.9	48.4	34.1	88.8	41.6	53.7	29.2	86.1	51.4	62.1	41.2
espressomachine	82.8	19.1	22.8	8.3	84.2	36.9	44.6	28.8	82.6	49.2	48.5	36.2	89.6	43.1	66.3	34.5	87.4	54.9	54.5	35.6
ketchup	81.3	29.3	30.2	16.4	72.0	18.0	15.1	9.3	75.3	32.7	51.9	37.8	89.4	56.0	62.3	40.2	84.0	55.4	64.2	45.0
laptop	87.8	24.9	16.7	7.0	82.1	35.3	43.8	28.0	86.0	62.0	45.1	37.1	93.3	50.2	63.2	34.6	91.4	68.0	54.9	40.8
microwave	85.7	17.1	29.5	8.0	87.0	36.1	56.2	27.3	85.4	51.5	50.9	35.3	90.3	35.2	53.6	23.0	89.0	52.5	55.4	34.8
mixer	79.4	12.1	14.8	4.9	85.3	34.3	45.1	26.7	79.3	37.9	48.4	32.2	89.1	38.4	59.0	29.0	87.0	52.0	62.3	39.6
notebook	87.2	25.0	6.6	2.8	85.7	38.7	33.8	20.8	84.5	57.8	43.0	33.2	93.5	52.8	63.6	35.0	90.9	65.5	55.6	38.7
phone	82.9	25.7	15.1	6.8	80.9	39.5	28.1	19.9	77.8	36.7	39.7	29.2	91.1	58.8	37.7	23.5	88.1	61.9	54.1	37.6
scissors	48.5	0.0	14.0	9.3	40.9	5.2	7.0	5.2	48.2	0.0	47.4	36.3	73.9	18.2	56.1	38.9	62.4	2.7	68.4	51.8
waffleiron	85.7	13.1	3.1	0.9	86.5	36.3	45.0	26.2	84.7	48.3	59.5	39.2	92.7	44.3	81.7	38.1	91.2	55.9	69.5	43.7
All	83.3	19.5	15.0	6.0	81.4	33.1	37.1	22.0	81.4	46.6	46.9	34.2	**90.8**	45.5	**59.6**	31.5	88.1	**55.7**	57.3	**38.5**

Table 1. Results on ARCTIC-Grasps. Green shows the best performing method per metric and yellow shows the second best.

Category	HOMan [10]			Dynamic			1-DoF		
	IOU	SCA@0.8	SCA@0.6	IOU	SCA@0.8	SCA@0.6	IOU	SCA@0.8	SCA@0.6
bottle	56.7	3.4	5.8	75.5	21.8	36.6	72.1	26.1	51.3
bowl	54.4	1.1	1.6	58.0	9.3	19.8	56.2	11.1	25.7
can	48.3	3.4	5.1	56.0	13.7	20.1	54.1	16.5	26.6
cup	56.9	5.5	7.8	67.5	16.5	38.9	65.9	18.4	46.1
glass	55.4	3.0	4.2	65.2	14.8	30.1	62.5	15.7	37.8
mug	59.4	3.9	6.6	63.8	6.3	29.6	62.6	10.8	38.6
pan	48.3	0.5	1.9	48.9	4.3	12.8	49.1	7.0	20.4
plate	61.1	1.0	1.6	68.1	17.6	30.1	65.2	16.3	34.6
saucepan	51.1	0.4	2.4	57.5	5.4	25.7	56.8	8.5	34.6
All	54.9	1.9	3.3	**61.7**	12.2	25.3	59.9	**14.1**	**32.9**

Table 2. Results on EPIC-Grasps. Green for best and yellow shows the second best.

[3] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, pages 231–248, 2022. 1

[4] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *NeurIPS*, 2022. 1

[5] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 1, 2

[6] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *CVPR*, pages 494–504, 2024. 1

[7] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In *CVPR*, pages 409–419, 2018. 1

[8] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3193–3203, 2020. 1

[9] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *CVPR*, 2023. 1

[10] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668, 2021. 1, 2, 3, 4

[11] 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 *Proceedings of SIGGRAPH Asia 2022 Conference Papers*, 2022. 1

[12] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1

[13] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022. 1, 2, 3

[14] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. *arXiv preprint arXiv:2312.05251*, 2023. 1, 2

[15] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *ECCV*, 2024. 1

[16] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Trans. Graph*, 36:17, 2017. 2

[17] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshop*, pages 1749–1759, 2021. 1

[18] Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *WACV*, pages 5735–5745, 2023. 1

[19] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, pages 11097–11106, 2021. 3

[20] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *CVPR*, pages 2750–2760, 2022. 1

[21] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, pages 3895–3905, 2022. 1

[22] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, pages 19717–19728, 2023. 1

[23] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, pages 34–51, 2020. 3