

Supplemental Material: H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions

Bugra Tekin¹

Federica Bogo¹

Marc Pollefeys^{1,2}

¹ Microsoft

² ETH Zürich

In the supplemental material, we provide details on the network architecture and how the training images were prepared. We also present additional qualitative and quantitative results on the datasets we evaluate our approach on [2, 3].

Network Architecture. Our unified network that jointly predicts per-frame 3D hand poses, 6D object poses, object classes and activity categories is a single pass network that does not rely on external detection algorithms nor region proposals [4]. To facilitate reproducibility, we provide the full details of our network architecture in Table 1.

Training Images. As discussed in the main paper, while training on the synthetic hand data [3], we replace the chroma-keyed background with random images from the PASCAL VOC dataset [1]. This operation of using random backgrounds brings in robustness against different backgrounds and is essential to achieve proper generalization. In addition, we superimpose synthetic objects (cuboids) with known 6D poses on the training images. This allows us to

train a network for both hand and object pose estimation and gain robustness against object occlusions. Examples of such images, which are given as input to the network at training time are shown in Fig. 1.



Figure 1: We extract the foreground objects in our training images and composite them over random images from PASCAL VOC [1]. We also augment the training set by superimposing segmentation masks of objects of interest for which the 6D poses are known to be able to simultaneously predict the hand and object pose and gain robustness against object occlusions.

Layer	Type	Filters	Size/Stride	Input	Output
0	conv	32	$3 \times 3 / 1$	$416 \times 416 \times 3$	$416 \times 416 \times 32$
1	max		$2 \times 2 / 2$	$416 \times 416 \times 32$	$208 \times 208 \times 32$
2	conv	64	$3 \times 3 / 1$	$208 \times 208 \times 32$	$208 \times 208 \times 64$
3	max		$2 \times 2 / 2$	$208 \times 208 \times 64$	$104 \times 104 \times 64$
4	conv	128	$3 \times 3 / 1$	$104 \times 104 \times 64$	$104 \times 104 \times 128$
5	conv	64	$1 \times 1 / 1$	$104 \times 104 \times 128$	$104 \times 104 \times 64$
6	conv	128	$3 \times 3 / 1$	$104 \times 104 \times 64$	$104 \times 104 \times 128$
7	max		$2 \times 2 / 2$	$104 \times 104 \times 128$	$52 \times 52 \times 128$
8	conv	256	$3 \times 3 / 1$	$52 \times 52 \times 128$	$52 \times 52 \times 256$
9	conv	128	$1 \times 1 / 1$	$52 \times 52 \times 256$	$52 \times 52 \times 128$
10	conv	256	$3 \times 3 / 1$	$52 \times 52 \times 128$	$52 \times 52 \times 256$
11	max		$2 \times 2 / 2$	$52 \times 52 \times 256$	$26 \times 26 \times 256$
12	conv	512	$3 \times 3 / 1$	$26 \times 26 \times 256$	$26 \times 26 \times 512$
13	conv	256	$1 \times 1 / 1$	$26 \times 26 \times 512$	$26 \times 26 \times 256$
14	conv	512	$3 \times 3 / 1$	$26 \times 26 \times 256$	$26 \times 26 \times 512$
15	conv	256	$1 \times 1 / 1$	$26 \times 26 \times 512$	$26 \times 26 \times 256$
16	conv	512	$3 \times 3 / 1$	$26 \times 26 \times 256$	$26 \times 26 \times 512$
17	max		$2 \times 2 / 2$	$26 \times 26 \times 512$	$13 \times 13 \times 512$
18	conv	1024	$3 \times 3 / 1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
19	conv	512	$1 \times 1 / 1$	$13 \times 13 \times 1024$	$13 \times 13 \times 512$
20	conv	1024	$3 \times 3 / 1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
21	conv	512	$1 \times 1 / 1$	$13 \times 13 \times 1024$	$13 \times 13 \times 512$
22	conv	1024	$3 \times 3 / 1$	$13 \times 13 \times 512$	$13 \times 13 \times 1024$
23	conv	1024	$3 \times 3 / 1$	$13 \times 13 \times 1024$	$13 \times 13 \times 1024$
24	conv	1024	$3 \times 3 / 1$	$13 \times 13 \times 1024$	$13 \times 13 \times 1024$
25	route	16			
26	conv	64	$1 \times 1 / 1$	$26 \times 26 \times 512$	$26 \times 26 \times 64$
27	reorg		$/ 2$	$26 \times 26 \times 64$	$13 \times 13 \times 256$
28	route	27 24			
29	conv	1024	$3 \times 3 / 1$	$13 \times 13 \times 1280$	$13 \times 13 \times 1024$
30	conv	720	$1 \times 1 / 1$	$13 \times 13 \times 1024$	$13 \times 13 \times 10 \cdot (3 \times N_c + 1 + N_a + N_o)$
31	prediction				$13 \times 13 \times 5 \times 2 \times (3 \times N_c + 1 + N_a + N_o)$

Table 1: Network architecture

Qualitative Results. We show qualitative results of approach along with some failure cases on the datasets we evaluate our method on in Fig. 2. These examples show that our method is robust to severe occlusions, rotational ambiguities in appearance, reflections, viewpoint changes and scene clutter. We provide additional qualitative results in the accompanying video and demonstrate that our *Hand + Object* approach also results in temporally coherent estimates.

Recognition Accuracies Per Action. In Figure 3, we show action-specific recognition accuracies on the FPHA dataset. While some actions such as ‘sprinkle’, ‘give coin’ and ‘pour juice’ are easily identifiable, actions such as ‘open letter’ and ‘light candle’ are commonly confused, likely because hand poses are more subtle and dissimilar across different trials.

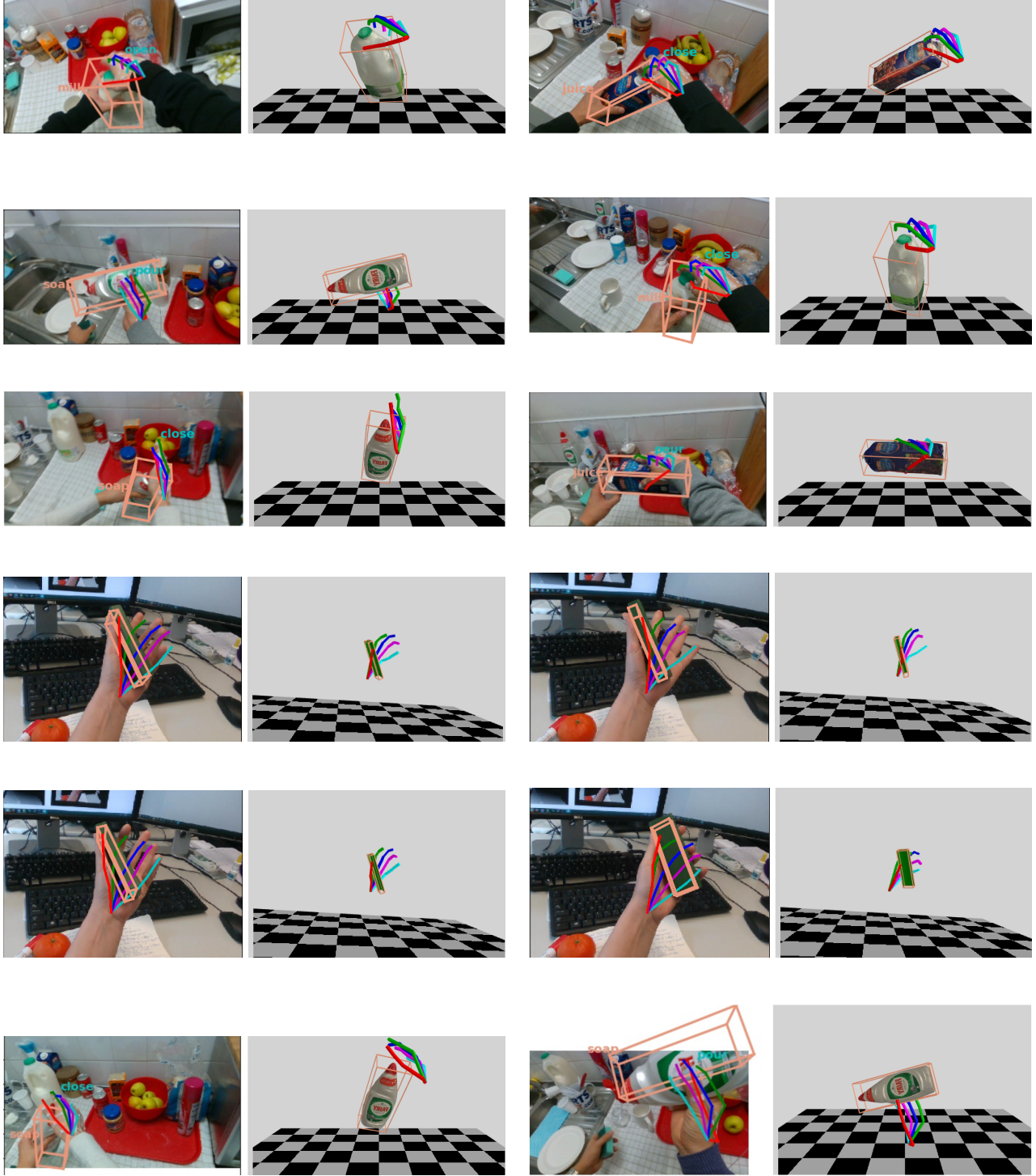


Figure 2: Qualitative results on the FPHA and EgoDexter dataset. We visualize the 3D hand pose estimates, 3D object bounding boxes which are transformed with the learned 6D object poses, and interaction classes. The proposed approach can handle motion blur, self-occlusions, clutter and complex articulations. We show in the last row failure cases due to an ambiguous action at the frame level (*e.g.* the “close” action is predicted, instead of the temporally symmetric “open” action) and the occlusion by the viewpoint (*e.g.* the object pose estimate is not very accurate as most of the object is out of the field of view). Best viewed in color.

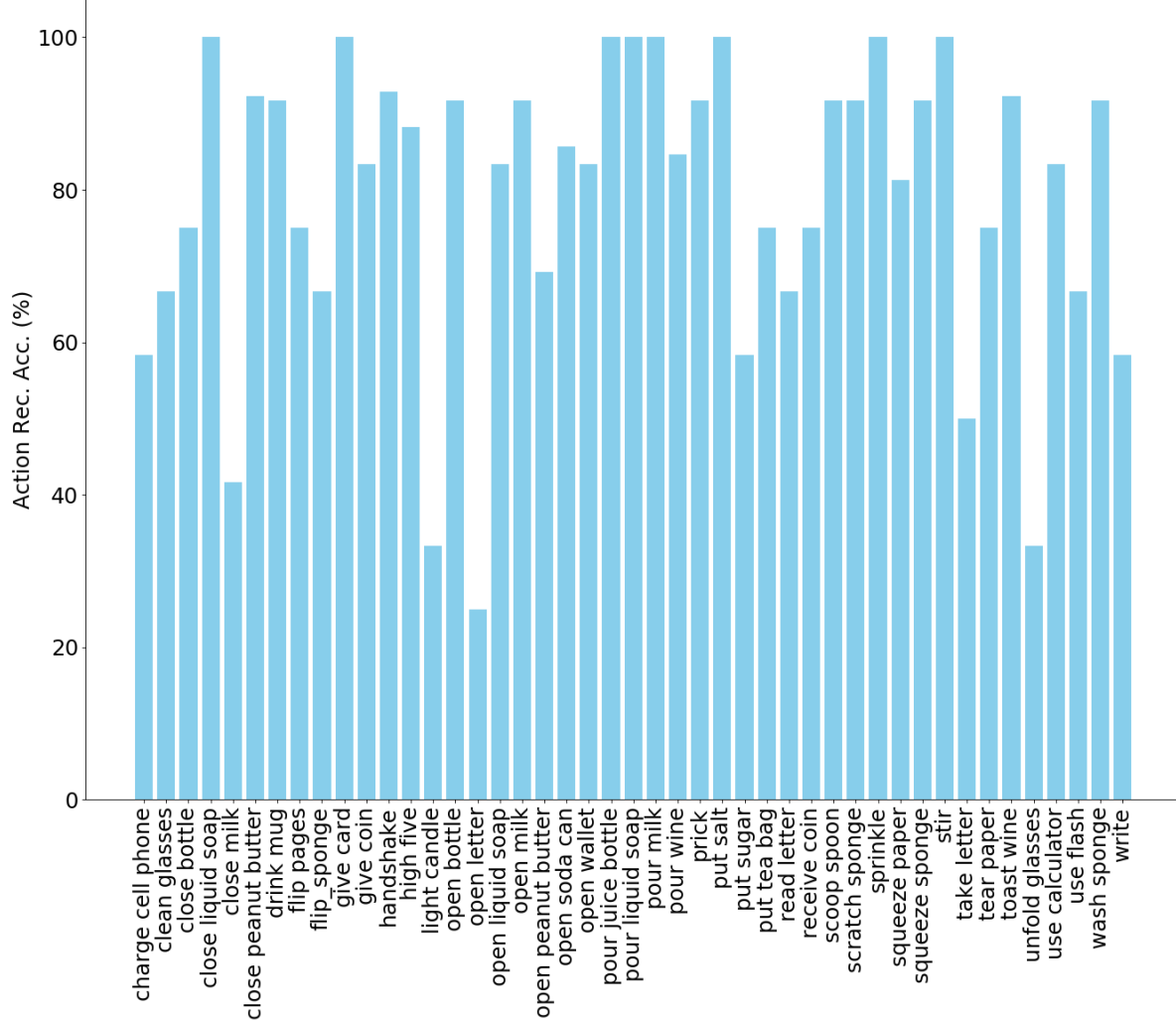


Figure 3: Action-specific recognition accuracies of our approach on the FPHA dataset.

References

- [1] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.
- [2] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations. In *CVPR*, 2018.
- [3] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor. In *ICCV*, 2017.
- [4] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *CVPR*, 2017.