Supplementary material

Tze Ho Elden Tse[1⋆], Zhongqun Zhang[1⋆], Kwang In Kim[2], Aleš Leonardis[1], Feng Zheng[3], and Hyung Jin Chang[1]

[1] University of Birmingham, UK
[2] UNIST, Korea
[3] SUSTech, China

In this supplemental document, we present:

1. implementation details of GCN-Contact (Section 1);
2. ablations on pseudo-labels (Section 2);
3. comparison with state-of-the-arts (Section 3);
4. additional qualitative examples (Section 4);
5. complete performances of different GCN-Contact design choices (Section 5);
6. complete table for computational analysis (Section 6).

Note that all the notation and abbreviations here are consistent with the main manuscript.

1 Implementation details of GCN-Contact

The network architecture is illustrated in Table 1.

2 Ablations on pseudo-labels

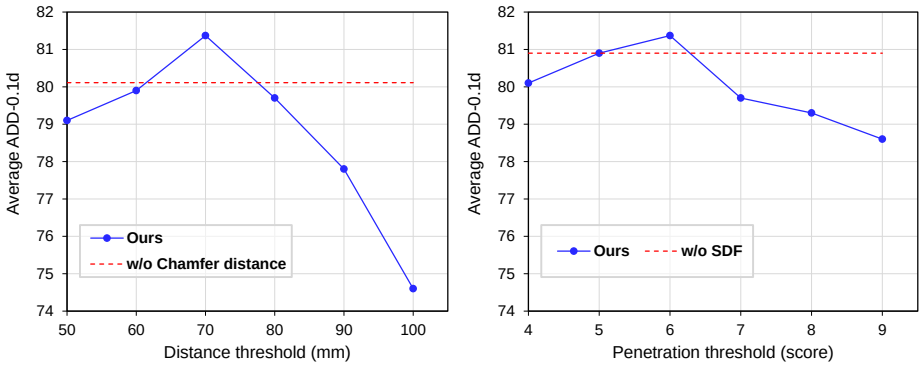

Fig. 1. Object performance with different geometric thresholds on *HO-3D*, distance threshold t_{dist} (*left*) and penetration threshold t_{pen} (*right*).

⋆ Equal contribution

Table 1. Architecture of GCN-Contact. B refers to batch size and log_softmax refers to softmax followed by a logarithm. Note that layers 2-5 are performed in parallel.

Layer	Operation	Dimensionality
	Input point clouds $\mathbf{P} = \{\mathbf{P}_{pos}, \mathbf{P}_F\}$	$B \times 28$
1	K-NN search on $\mathbf{P}_{pos}$	$2 \times 10B$
1	K-NN search on $\mathbf{P}_F$	$2 \times 10B$
2	Graph conv. on $\mathbf{P}_{pos}$	$B \times 64$
3	Graph conv. on $\mathbf{P}_{pos}$	$B \times 64$
4	Graph conv. on $\mathbf{P}_F$	$B \times 64$
5	Graph conv. on $\mathbf{P}_F$	$B \times 64$
6	Concat$(2, 3, 4, 5)$	$B \times 256$
7	MLP	$B \times 1024$
8	MLP	$B \times 256$
9	Dropout$(p = 0.5)$	$B \times 256$
10	MLP	$B \times 128$
11	Dropout$(p = 0.5)$	$B \times 128$
12	MLP	$B \times 10$
13	log_softmax(12)	$B \times 10$

2.1 Effects of varying thresholds for pseudo-labels generation

We study the effect of different thresholds of the pseudo-labels filtering mechanism on *HO-3D* in Figure 1 and Figure 2. As shown in Figure 1, the performance is higher than the method without Chamfer distance by large margins when $60 < t_{dist} \leq 80$. Note that the performance peaks at $t_{dist} = 70$. When $t_{dist} > 70$, further increasing t_{dist} led to a drastic drop in pseudo-label as hand and object are so far away from each other such that the resulting pseudo-labels are in low-quality. On the contrary, when $t_{dist} \leq 60$, there are less qualifying pseudo-labels leading to performance drop from insufficient training labels. We obtain similar observations for t_{pen} and t_{SSIM}.

2.2 Amounts of pseudo-labels

We analyse the effect of using different fractions of pseudo-labels in semi-supervised learning on *HO-3D* in Figure 3. We uniformly sample 20%, 40%, 60%, and 80% of the collected pseudo-labels for semi-supervised learning. As shown, the performance has been significantly improved after adding 20% of pseudo-labels. We observe that the more pseudo-labels used in training, the better the performance the model can achieve.

3 Comparison with state-of-the-arts

We compare against the state-of-the-art approaches [3,17,18] on *HO-3D* in Table 2. [3] is an optimisation-based method which leverages 2D image cues and

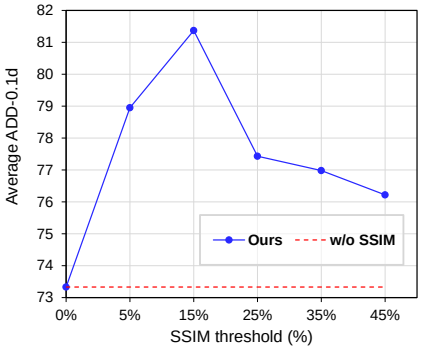

Fig. 2. Object performance with varying visual consistency constraint thresholds t_{SSIM} on *HO-3D*.

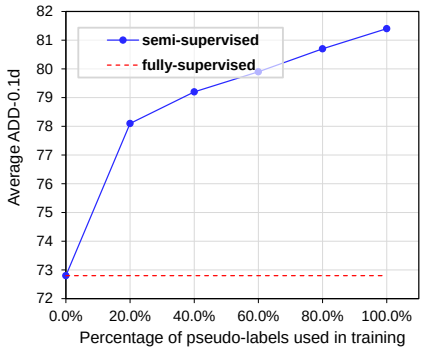

Fig. 3. Object performance with varying percentage of pseudo-labels on the *HO-3D*.

Table 2. Error rates on *HO-3D*. **cover** refers to intersection volume (cm^3) and **inter** refers to contact coverage (%).

Methods	Hand error joint ↓	Hand error mesh ↓	Object add-0.1d ↑	Contact cover ↑	Contact inter ↓
Cao *et al.* [3]	9.7	9.7	79.5	18.5	4.9±2.1
Hasson *et al.* [17]	11.1	11.0	74.5	4.4	15.3±21.1
Hasson *et al.* [18]	14.1	14.7	64.2	11.5	13.5±17.6
Ours	**8.7**	**8.9**	**81.4**	**19.2**	**3.5±1.8**

3D contact priors for reconstructing hand-object interactions. [17] uses a feed-forward neural network to predict 3D hand pose and object pose where its single-frame model with full 3D supervision. [18] follows a fitting-based approach which builds on estimates from neural network models for detection, object segmentation and 3D hand pose estimation trained with full supervision.

4 Additional qualitative examples

In this section, we show additional qualitative results. In Figure 4, we visualise the predictions of ContactOpt [13] and our method as well as the ground-truth. We can see that our method is able to better reconstruction hand and object with more accurate contact map estimations. Our method performs significantly better than previous approaches. In Figure 5, we show more examples of our method and [30] on *HO-3D*. Our method significantly improve the hand-object pose. The last row of Figure 5 shows a failure case where the region of hand-object contact was too small for the network to produce a good contact prediction.

We provide qualitative examples on out-of-domain objects in Figure 6.

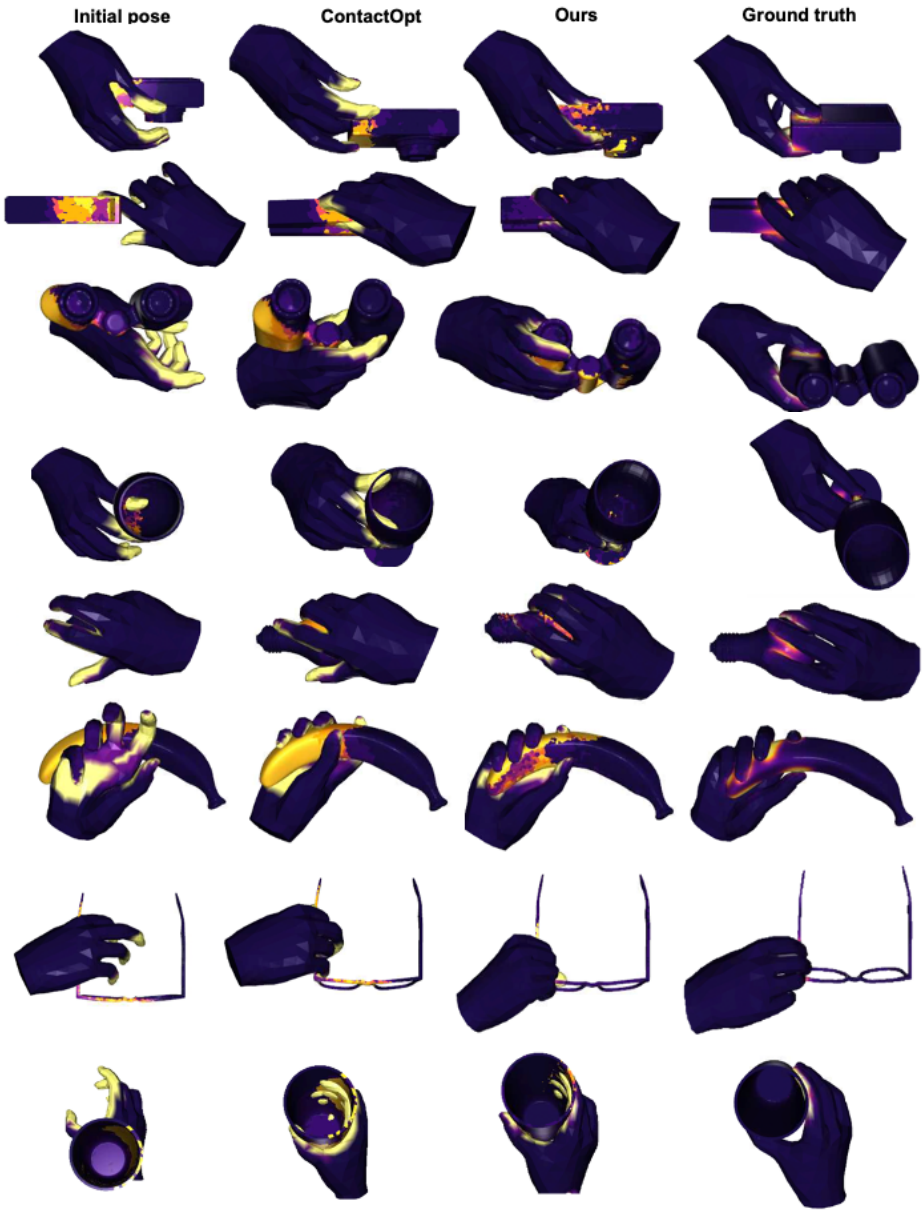

Fig. 4. Qualitative comparison with ContactOpt [13] on *ContactPose*. We observe that accurate contact map estimations allows our method to recover plausible grasps and significantly improves upon ContactOpt.

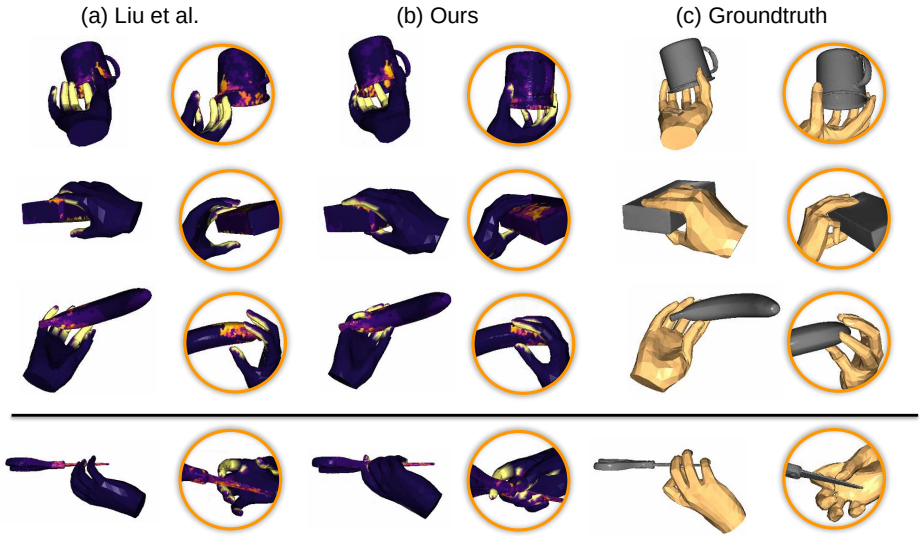

Fig. 5. Qualitative comparison with Liu *et al.* [30] on *HO-3D*.

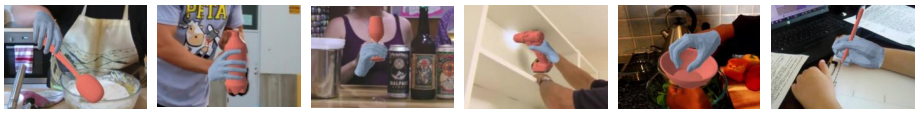

Fig. 6. Qualitative results on *EPIC-Kitchens* [1] and *100 Days of Hands* [2].

5 Performances of different GCN-Contact design choices

Complete results are reported in Table 3.

6 Computational analysis

Complete results are reported in Table 4.

References

1. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The EPIC-KITCHENS Dataset. In: ECCV (2018)
2. Shan, D., Geng, J., Shu, M., Fouhey, D.: Understanding human hands in contact at internet scale. In: CVPR (2020)
3. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. In: SIGGRAPH (2019)

Table 3. Performances of different GCN-Contact design choices on *ContactPose* and *HO-3D*. **semi** refers to semi-supervised learning. We experiment on (a) number of K neighbours without dilation, (b) size of dilation factor d with $K = 10$ and (c) combining K-NN computation (denoted with *) with $d = 4$.

		ContactPose Hand error		*HO-3D* w/o semi Hand error		Object error
	models	joint ↓	mesh ↓	joint ↓	mesh ↓	add-0.1d (↑)
(a)	$K = 5$	8.252	8.134	12.69	12.86	66.33
	$K = 10$	**6.691**	**6.562**	**11.81**	**11.91**	**68.71**
	$K = 15$	6.715	6.617	11.85	11.92	68.70
	$K = 20$	6.708	6.590	11.81	11.98	68.69
	$K = 25$	6.691	6.615	11.84	11.92	68.71
(b)	$d = 2$	5.959	5.865	10.91	10.86	70.25
	$d = 4$	**5.878**	**5.765**	**9.92**	**9.79**	**72.81**
	$d = 6$	5.911	5.805	9.95	9.79	72.81
	$d = 8$	5.899	5.812	9.94	9.85	72.80
	$d = 10$	5.889	5.776	9.93	9.79	72.78
(c)	$K^* = 5$	8.451	8.369	12.91	12.86	68.86
	$K^* = 10$	**8.359**	**8.251**	**11.55**	**11.97**	**69.10**
	$K^* = 25$	8.369	8.286	11.57	11.97	69.06

Table 4. Comparison on computational requirements of different networks.

	Baseline	DGCNN [3]	Ours
Parameters	1,424,138	530,442	587,658
GPU memory (GB)	13.3	21.1	10.4