Collaborative Learning for Hand and Object Reconstruction with Attention-guided Graph Convolution

Tze Ho Elden Tse[1] Kwang In Kim[2] Aleš Leonardis[1] Hyung Jin Chang[1]
[1]University of Birmingham [2]UNIST
txt994@student.bham.ac.uk, kimki@unist.ac.kr, {a.leonardis, h.j.chang}@bham.ac.uk

In this supplemental document, we present:

- analysis on the choice of PCA hand pose components for MANO (Sec. 1.1);
- analysis on hand shape regularisation (Sec. 1.2);
- mathematical formulation of graph convolutions we used for comparisons (Sec. 2);
- analysis on multi-head attention mechanism (Sec. 3);
- illustration of naïve collaborative learning baseline 4;
- summary of dataset statistics (Sec. 5);
- additional reconstruction examples (Sec. 6).

1. Hand mesh estimation

1.1. MANO pose representation

As described in Section 3.1 in the main paper, our hand branch outputs a 45-dimensional vector to represent the hand. We experiment with different dimensionality for the latent hand representation and summarise our findings in Table 1. We observe low-dimensionality fails to capture some poses present in the datasets and full 45-dimensional vector is required to produce the best result. Therefore, we use this value for all experiments in the main paper.

Table 1. We report the mean end-point error (mm) on *FHB*$^-$ and *ObMan* to study the effect of the number of PCA hand pose components for the latent MANO representation.

PCA components	15	30	45
ObMan	11.7	9.6	**9.2**
FHB$^-$	28.2	26.1	**25.3**

1.2. MANO shape regularisation

As described in Section 3.1 in the main paper, we observe that hand reconstruction performance increases with a larger saturated hand shape value than when it is trained with hand shape regularisation. We experiment with the loss on 3D joints ($\mathcal{L}_J$) and shape regularisation ($\mathcal{L}_\beta$). Table 2 shows that the hand reconstruction performance increases without shape regularisation ($\mathcal{L}_\beta$). As dense vertex supervision is not available in the real dataset *FHB*$^-$ [3], we omit experimenting on vertex loss $\mathcal{L}_V$.

Table 2. The mean end-point errors (mm) of two versions of our system that use 1) only the 3D joint loss ($\mathcal{L}_J$) and 2) a combination of the joint loss and shape regularisation ($\mathcal{L}_J + \mathcal{L}_\beta$). For both *ObMan* [5] and *FHB*$^-$ [3], low errors are measured when shape regularisation is disabled.

	ObMan	*FHB*$^-$
$\mathcal{L}_J$	**9.2**	**25.3**
$\mathcal{L}_J + \mathcal{L}_\beta$	10.3	27.5

2. Graph convolution

In the following, we provide the mathematical formulation of GCN [6] and spiral mesh convolution [4, 7] which were used for comparisons in the main paper.

For GCN, we followed [2] and used the adaptive graph convolution:

$$Y = \sigma(\tilde{A}XW), \tag{1}$$

where Y is the output feature with N nodes and l output features for each node, σ is the activation function, $W \in \mathbb{R}^{k\times l}$ with k input features for each node, $X \in \mathbb{R}^{N\times k}$ is the matrix of input features, and $\tilde{A} \in \mathbb{R}^{N\times N}$ is the row-normalised adjacency of the graph.

For spiral mesh convolution, we adopted the most basic form of mesh convolution using spiral neighbourhoods from [4, 7]. Mathematically, spiral neighbourhoods $S(i, l)$ of vertex i with length l can be defined as follows [8]:

$$\begin{aligned} 0\text{-ring}(i) &= \{i\}, \\ l\text{-disk}(i) &= \cup_{v=0,\dots,l} v\text{-ring}(i), \\ (l+1)\text{-ring}(i) &= \mathcal{N}(l\text{-ring}(i)) \setminus l\text{-disk}(i), \\ S(i,l) &\subset (0\text{-ring}(i), \dots, k\text{-ring}(i)), \end{aligned} \tag{2}$$

where $\mathcal{N}(V)$ is the set of all vertices adjacent to any vertex in set V. Finally, we arrive at the basic spiral mesh convolution:

Table 3. Performances of EdgeConv [9] on *FHB*$^-$. We experiment on network iterations P and associative loss $\mathcal{L}_{asso}$. We kept $k = 20$ for k-NN neighbourhood construction.

Method	w $\mathcal{L}_{asso}$		w/o $\mathcal{L}_{asso}$	
	Hand Error	Object Error	Hand Error	Object Error
$P = 1$	26.5	1583.8	27.2	1629.7
$P = 2$	26.7	1582.2	27.5	1629.5

$$\mathbf{v_i^{(k)}} = \text{MLP}\big(\|_{u \in S(i,l)} \mathbf{v_u^{(k-1)}}\big) \quad (3)$$

where the aggregating function $\|$ is a concatenation of spiral neighbourhood and MLP as update function. We used spiral length $l = 10$ for all experiments.

In addition, we experiment with a popular k-nearest neighbours (k-NN) based graph convolution, EdgeConv [9]. They dynamically construct k-NN neighbours $\mathcal{N}_{k-nn}(V_i)$ and can be described as:

$$\mathbf{h}_i^{t+1} = \max_{j \in \mathcal{N}_{k-nn}(V_i)} \Big(\text{ReLU}(\text{MLP}(\mathbf{V}_j - \mathbf{V}_i, \mathbf{V}_i))\Big) \quad (4)$$

Similar to the two above graph convolution operators, Table 3 shows that k-NN like approaches suffer from local neighbourhood aggregation as the incoming mesh are 3D positions.

3. Multi-head attention

As described in Section 3.3 in the main paper, we found multi-head attention to be beneficial. We experiment with different number of heads and summarise our findings in Table 4. We used multi-head attention $K = 3$ in all experiments as it provides the best performance.

Table 4. We report the mean end-point error (mm) on *FHB*$^-$ and *ObMan* to study the effect of the number of multi-head attention mechanism.

#heads	1	2	3	4	5
ObMan	12.4	11.4	**9.2**	9.3	9.6
FHB$^-$	28.7	26.8	**25.3**	25.5	25.4

4. Naïve collaborative learning baseline

To motivate our design choices, we experiment on a naïve collaborative learning baseline as shown in Fig. 1. This design framework directly predicts embeddings ϕ_θ and reconstruct meshes $\mathbf{m}_\theta$ at the final stage. The key difference between Fig. 1 and the final design is the attention-guided graph convolution which is proposed to tackle the **mutual occlusion** problem in hand-object interactions. Our experiments demonstrate that our attention-guided graph convolution combined with collaborative learning enables better mesh quality as well as more accurate pose estimation.

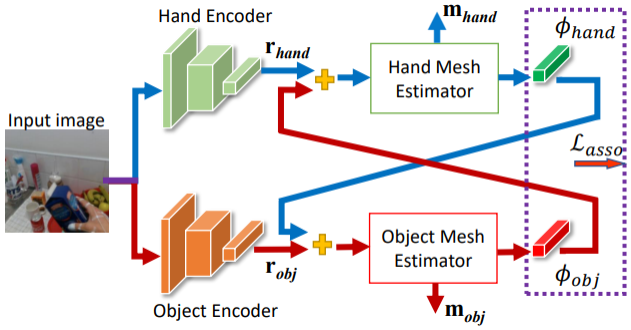

Figure 1. Simple collaborative learning framework design. Note that the yellow cross sign refers to addition.

5. Details of dataset

Table 5 summarised the statistics for each datasets [5].

Table 5. Dataset details for train/test splits.

	ObMan	*FHB*	*FHB*$^-$
#frames	141K/6K	8,420/9,103	5,077/5,657
#video sequence	-	115/127	76/88
#object instances	1,947/411	4	3

6. Reconstruction examples

We provide additional qualitative comparisons with Hasson *et al*. [5] on the synthetic dataset *ObMan*. Fig. 2 demonstrates that our method is able to produce more physically plausible hand reconstruction than [5] without physical constraints. In particular, for the first two rows of Fig. 2, ObMan's hand reconstruction contains over-bending fingers which is infeasible for humans. The bottom three rows of Fig. 2 shows that our method is able to produce a more refined and accurate hand reconstruction.

In addition to synthetic dataset *ObMan*, we also provide additional reconstruction examples for real dataset *DexYCB* [1] in the supplementary video. Our method is able to accurately reconstruct hand and object mesh across various hand poses and object class. In particular, we demonstrate the importance of our attention-guided graph convolution in collaborative learning by directly comparing with a naïve collaborative learning baseline (shown in Fig. 1). Also, we provide reconstruction examples on pre-grasp stages as our method is not restricted by contact loss terms [5].

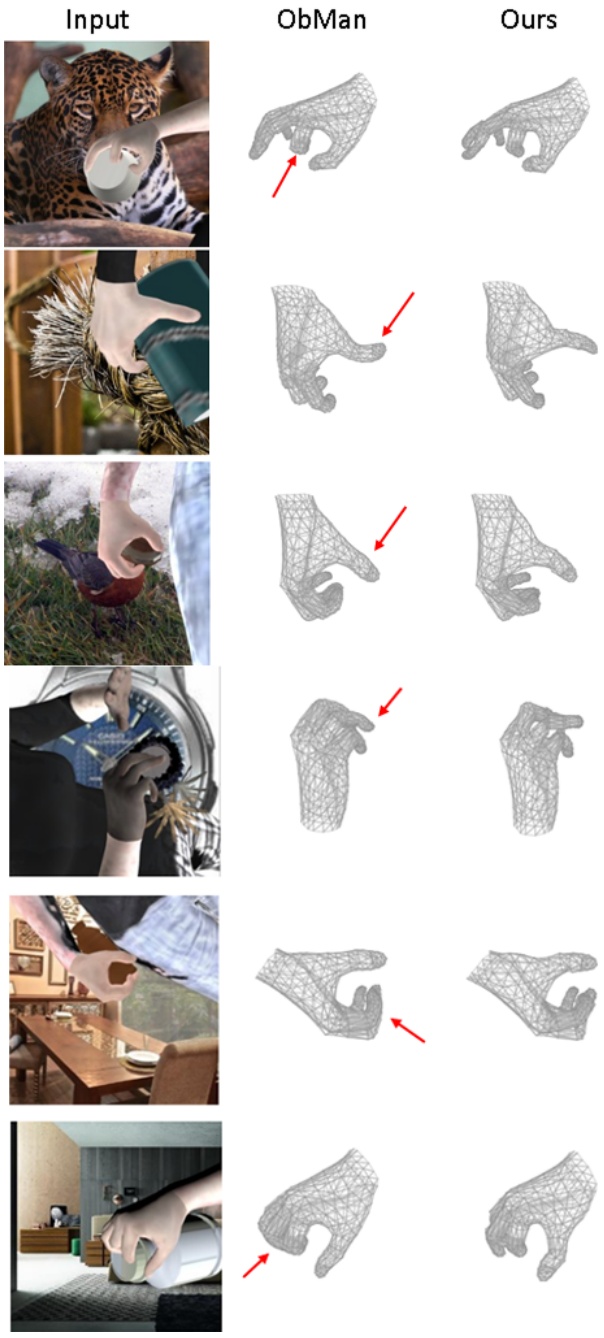

Figure 2. Qualitative comparison with Hasson *et al*. [5] on synthetic dataset *ObMan*. Our method is able to reconstruct physically plausible hand mesh without physical constraints.

References

[1] 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*, 2021. 2

[2] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. HOPE-Net: A graph-based model for hand-object pose estimation. In *CVPR*, 2020. 1

[3] 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*, 2018. 1

[4] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. SpiralNet++: A fast and highly efficient mesh convolution operator. In *ICCV Workshops*, 2019. 1

[5] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2, 3

[6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 1

[7] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 1

[8] Isaak Lim, Alexander Dielen, Marcel Campen, and Leif Kobbelt. A simple approach to intrinsic correspondence learning on unstructured 3d meshes. In *ECCV Workshops*, 2018. 1

[9] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph CNN for learning on point clouds. In *SIGGRAPH*, 2019. 2