TOCH: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement - Supplemental

Keyang Zhou[1,2], Bharat Lal Bhatnagar[1,2], Jan Eric Lenssen[2], and Gerard Pons-Moll[1,2]

[1] University of Tübingen, Germany
{keyang.zhou,gerard.pons-moll}@uni-tuebingen.de
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{bbhatnag,jlenssen}@mpi-inf.mpg.de

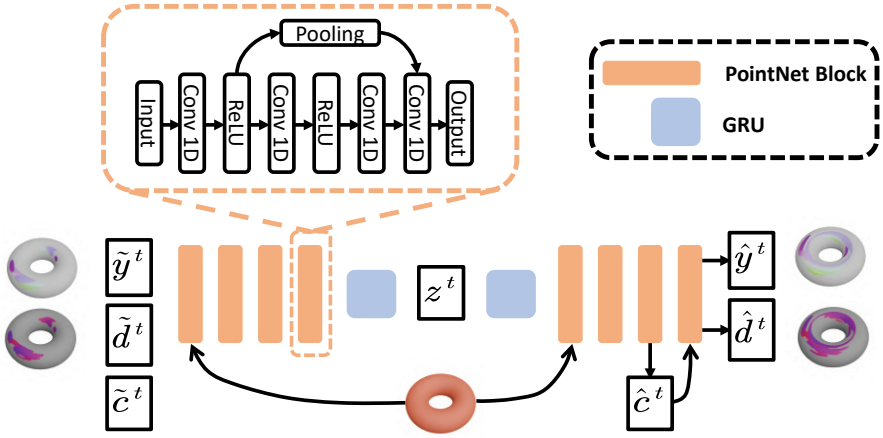

Fig. 1: Network architecture of TOCH. This diagram only shows the processing of one particular TOCH field at frame t, while in practice the network takes a temporal window of concatenated TOCH fields.

In this supplementary document we provide further details about our method such as network architecture and data processing. We also provide more qualitative results for hand-object tracking refinement.

1 Network Architecture

The network takes a concatenation of T frames as input. Since each frame is identically processed, we show the architecture for processing one frame in Figure 1. Specifically, $\boldsymbol{c}$ denotes binary correspondence mask, $\boldsymbol{y}$ denotes corresponding hand coordinates, and $\boldsymbol{d}$ denotes corresponding signed distances. The coordinates and normals of the conditioning object at frame t are also fed into the

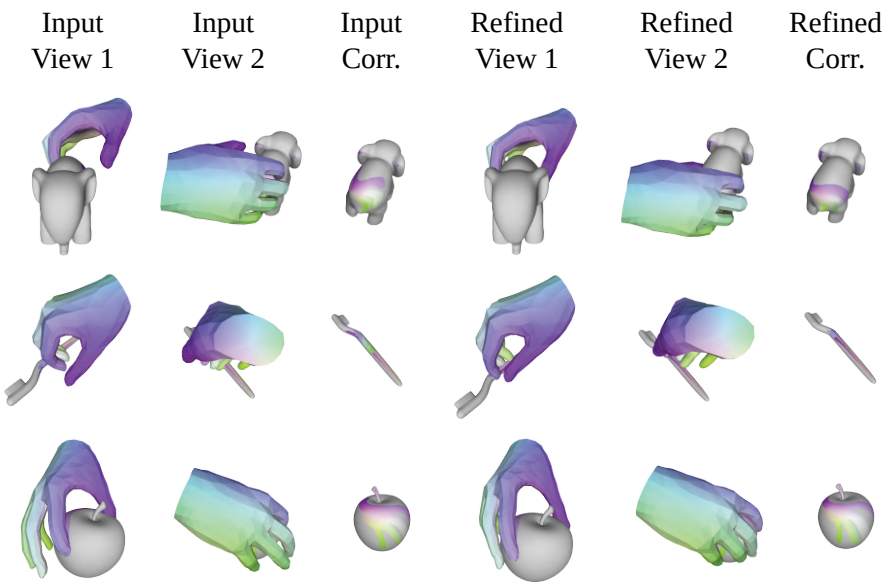

Fig. 2: Hand-object correspondence visualization for sample test frames from GRAB [2] dataset. Corresponding hand and object vertices share the same color.

first layer of encoder and decoder. Since the input features at each frame can be viewed as a point cloud, we use PointNet [1] to extract frame-wise features. Temporal information is propagated through a GRU layer.

2 Data Processing

For GRAB [2], we follow an object-based split. We reserve sequences involving [‘apple’, ‘toothbrush’, ‘elephant’, ‘hand’] for validation and [‘mug’, ‘wineglass’, ‘camera’, ‘binoculars’, ‘fryingpan’, ‘toothpaste’] for testing, and train on the rest. We sample 8000 points from each object mesh. At each training iteration, we use a random subset of 2000 points, which are centered and normalized to fit within a unit ball. We also apply a random SO(3) rotation to all point clouds within a temporal window to promote rotation invariance. For each TOCH field obtained from ray casting, we remove all corresponding hand-object points with correspondence distance larger than 10cm.

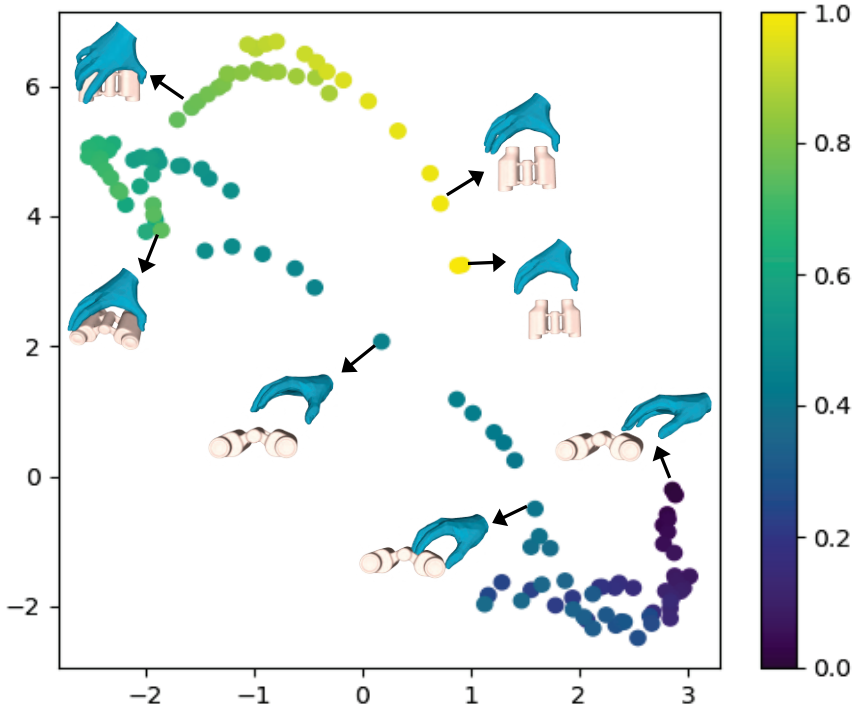

Fig. 3: Latent space visualization of a hand-object interaction sequence. Each dot represents the projected latent code of a frame. Dot color indicates time index (from 0 to 1). Arrows point to respective denoised hands.

3 Hand Fitting

Recall that the objective function for fitting hand to TOCH field is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}_H) = \sum_{i=1}^{T} w_{\text{corr}} \mathcal{L}_{\text{corr}}(\boldsymbol{\beta}, \boldsymbol{\theta}^i, \boldsymbol{t}_H^i) + \mathcal{L}_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\theta}) \tag{1}$$

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = w_1 \left\|\boldsymbol{\beta}\right\|^2 + w_2 \sum_{i=1}^{T} \left\|\boldsymbol{\theta}^i\right\|^2 + w_3 \sum_{i=1}^{T-1} \left\|\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\right\|^2 + w_4 \sum_{i=2}^{T-1} \sum_{k=1}^{J} \left\|\ddot{\boldsymbol{p}}_k^i\right\|, \tag{2}$$

We empirically set $w_{\text{corr}} = 20$, $w_1 = 0.001$, $w_2 = 0.0001$, $w_3 = 0.05$, and $w_4 = 1$. The training is divided into two stages. In the first stage, we only optimize hand orientation and translation with $\mathcal{L}_{\text{corr}}$. In the second stage we jointly optimize all the variables. Both stages are trained with an Adam optimizer. The first stage is optimized for 100 iterations with a learning rate of 0.1. The second stage is optimized for 2000 iterations with an initial learning rate of 0.1 for the first 1000

iterations, and a decayed learning rate of 0.05 for the rest. We parallelly optimize 5 random initialization of hand parameters and pick the best result.

4 TOCH Field Visualization

We visualize correspondence maps for three sets of hand-object meshes before and after refinement in Fig. 2.

5 Latent Space Visualization

Fig. 3 shows t-SNE plot of the model's latent space when encoding an hand-object interaction sequence. The sequence of latent codes clearly form a trajectory in latent space. Another interesting observation is that when the hand is in contact with the object, nearby dots form a small cluster. When the hand releases the object, distance between consecutive dots increases with hand-object distance.

6 Qualitative Results

Fig. 4 shows more qualitative results on hand-object tracking correction on GRAB. Please check the supplementary video for animated results.

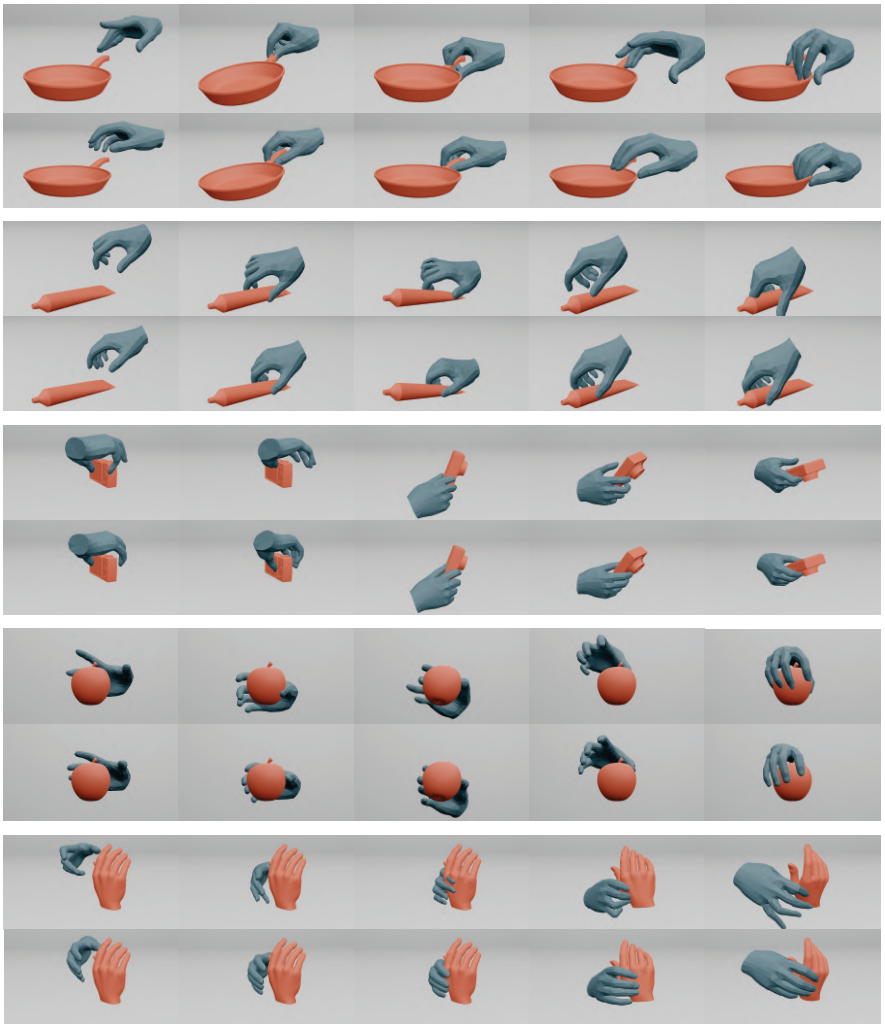

Fig. 4: Noisy sequence correction on GRAB. For each pair of rows, the first row shows snapshots from an erroneous tracking sequence and the bottom row shows results after TOCH correction.

References

1. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
2. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020), https://grab.is.tue.mpg.de