

Supplementary Material for: Dynamic Reconstruction of Hand-Object Interaction with Distributed Force-aware Contact Representation

Zhenjun Yu^{*}, Wenqiang Xu^{*}, Pengfei Xie, Yutong Li, Cewu Lu[§]
Shanghai Jiao Tong University

{jeffson-yu, vinjohn, pf.xie, davidliyutong, lucewu}@sjtu.edu.cn

1. Supplementary Video

In the supplementary material, we provide a video that visually demonstrates several key information in our paper. The video is organized as follows:

- **0:00–0:17:** Qualitative results on DexYCB Dataset for comparing our method and gSDF [1].
- **0:17–0:30:** Qualitative results on HOT Dataset for comparing our method with CPF [3], TOCH [4] and ViTaM [2].
- **0:30–0:42:** Hardware setup for our real-world experiments.
- **0:42–1:00:** Real-world experiment results on two stuffed toys.

The video is intended to complement our demonstration in the main paper by providing dynamic illustrations of our experimental results.

2. Architecture of Flow Prediction Module

We introduce the detailed architecture of our proposed flow prediction module $\mathcal{F}_f(\cdot)$. Assuming at frame t , the per-point features F_t, F_{t-1} is extracted from \mathcal{P}_t and \mathcal{P}_{t-1} , the flow prediction module takes the two set of point clouds and their feature to predict the point cloud flow between two frames:

$$f_{t-1 \rightarrow t} = \mathcal{F}_f(F_t, F_{t-1}, \mathcal{P}_t, \mathcal{P}_{t-1}). \quad (1)$$

We first perform a Cartesian product of the two extracted features, yielding a tensor of size $N \times N \times 2d$. This tensor is fed into a 3-layer MLP to obtain \mathcal{C}_v , which is then used in two ways. First, it passes through a 2D convolutional layer for downsampling to obtain $p_c \in \mathbb{R}^{N \times N}$, representing the matching probability of each point between two frames. Second, \mathcal{C}_v is sent through a softmax function and a one-layer MLP to downsample, resulting in $\mathcal{C}_c \in \mathbb{R}^N$, indicating whether the points in the first frame are matched in the second frame. Thus, the final matching probability matrix $p_m \in \mathbb{R}^{N \times N}$, which describes the correspondence likelihood between the two point sets, can be computed as:

$$p_m = p_c \times \mathcal{C}_c \quad (2)$$

After estimating the matching probability, we compute the disparity of two point cloud sets $\mathcal{D} \in \mathbb{R}^{N \times N \times 3}$, with $\mathcal{D}_{ij} = \mathcal{P}_t^i - \mathcal{P}_{t-1}^j$, and concatenate the disparity with p_m . The concatenated tensor is then fed into four 2D convolutional layers with batch normalization and a softmax function to obtain the disparity feature $F_d \in \mathbb{R}^{N \times d'}$. Finally, we use PointNet++ to regress the flow $f_{t-1 \rightarrow t} \in \mathbb{R}^{N \times 3}$.

In practice, the point cloud feature size is $d = 128$, and the disparity feature size is $d' = 64$.

3. Flow Prediction Result

To validate the accuracy of our proposed flow prediction module, we report the Chamfer distance error on the DexYCB and HOT datasets in Tab. 1. The relatively low Chamfer distance demonstrates the efficacy of our network. The slightly better results on the DexYCB dataset are likely due to the larger hand-object movements and more challenging object deformations in the HOT dataset.

We also present some qualitative results in Fig. 1. The top section compares our predicted flow added to the last frame's point cloud with the ground truth of the current frame's point cloud. The near overlap of the two point clouds indicates high prediction accuracy. The bottom section shows a sequence of our estimation results, illustrating our method's ability to track hand movements and object deformations in whole sequences.

| Dataset | DexYCB | | HOT | |
|----------|--------|--------|--------|------------|
| Category | Box | Bottle | Sponge | Plasticine |
| CD(mm)↓ | 8.7 | 9.1 | 10.3 | 12.2 |

Table 1. Quantitative results for flow predictions in visual dynamic tracking net.

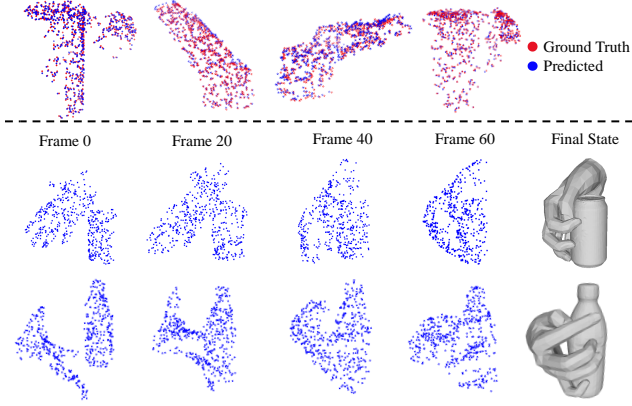


Figure 1. Qualitative results on flow prediction using the Flow Prediction Module.

4. Ablation Study, Extended

4.1. Importance of flow prediction module

To demonstrate the importance of our flow prediction module, we conducted an experiment by directly fusing the visual features extracted from the input point clouds using the transformer fusion layer. The quantitative results are shown in Tab. 2. Introducing flow prediction module significantly improves all scores, validating our feature fusion approach that incorporates the 3D static information of the current frame and the correspondence feature from the flow.

| Metrics | IoU \uparrow | CD \downarrow | MPJPE \downarrow |
|----------------|----------------|-----------------|--------------------|
| Dataset | DexYCB | | |
| w/o Flow Pred. | 84.3 | 15.7 | 17.1 |
| w. Flow Pred. | 90.1 | 9.6 | 13.2 |
| Dataset | HOT | | |
| w/o Flow Pred. | 64.7 | 20.2 | 16.3 |
| w. Flow Pred. | 80.5 | 10.9 | 13.6 |

Table 2. Quantitative results on DexYCB and HOT Dataset for whether using the flow prediction module. "Flow Pred." stands for flow prediction module, "w/o" indicates without.

4.2. Different point pair establishment strategies

This section discusses the influence of two point pair establishment strategies: using **keypoints** or **all-hand vertices**. When considering all hand vertices, we establish point pairs between them and nearby object vertices, treating the force exerted by the hand as the reading from the nearest sensor. The quantitative results of these two methods are shown in Tab. 3. While penetration depth improves slightly, both contact IoU and MPJPE decrease. This is likely because considering all sensors leads to conflicting optimization directions for the same joint, as sensors within the same re-

| Metrics | MPJPE \downarrow | PD \downarrow | CIoU \uparrow | Iter. Time(s) \downarrow |
|-----------|--------------------|-----------------|-----------------|---------------------------------|
| Keypoint | 11.3 | 7.3 | 40.3 | 3.5 \pm 0.5 |
| All Verts | 14.5 | 6.9 | 25.6 | 37 \pm 3 |

Table 3. Quantitative results on evaluating point pair establishment on key points and on all-hand vertices.

gions may experience different contact situations. Additionally, the iteration time increases about tenfold compared to our setting.

4.3. Tactile feature fusion in visual dynamic tracking net.

To assess the impact of introducing distributed tactile arrays in visual dynamic tracking net, we first use a 3-layer MLP to encode the tactile features of each region. By estimating hand pose, we fuse these regional features to the sample points, adding the encoded tactile feature to the point-wise feature of each sample position. We train visual dynamic tracking net with fused tactile information on our HOT dataset, and the results are shown in Tab. 4. Quantitative results show no significant improvements, likely because the tactile data are much more sparse than the visual inputs, causing feature misalignment. Therefore, we implement DF-Field to convert force readings into contact states for hand-pose optimization.

| Metrics | IoU \uparrow | CD \downarrow |
|--------------------|----------------|-----------------|
| w/o Tactile Fusion | 81.0 | 10.9 |
| w. Tactile Fusion | 81.3 | 11.5 |

Table 4. Quantitative results for whether or not fusing tactile information in visual dynamic tracking net. "w/o" indicates without.

4.4. Effectiveness of Energy and Loss Terms for Force-aware Optimization

To better understand the contribution of each component in our force-aware optimization, we conduct ablation studies on the HOT dataset, evaluating the effects of removing barrier energy B_{ij} , relative potential energy E_{ij} , and loss terms \mathcal{L}_r and \mathcal{L}_o from our framework.

The results are reported in Table 5, demonstrating the importance of each term in maintaining physical plausibility and reconstruction accuracy. Removing the Barrier Function leads to significant interpenetration between the hand and object, and weakens contact quality due to the unbalanced attraction from the Relative Potential Energy. Conversely, eliminating the Relative Potential Energy causes the Barrier Function to over-separate the hand from the object, resulting in poor contact recovery despite low penetration.

Without the regularization term \mathcal{L}_r , fingers may exhibit unnatural poses, leading to increased joint error. Omitting

\mathcal{L}_o causes the reconstructed hand pose to deviate notably from the initial prediction, as evidenced by a large increase in MPJPE (20.2 mm).

These results validate the necessity of combining all proposed energy and loss terms for physically consistent and accurate dynamic hand-object reconstruction.

| Config. | MPJPE(mm)↓ | PD(mm)↓ | CIoU(%)↑ |
|---------------------|-------------|------------|-------------|
| Full Opt. | 11.3 | 7.3 | 40.3 |
| w/o B_{ij} | 13.8 | 14.5 | 25.0 |
| w/o E_{ij} | 17.2 | 2.5 | 12.0 |
| w/o \mathcal{L}_r | 15.7 | 8.4 | 33.2 |
| w/o \mathcal{L}_o | 20.2 | 8.1 | 35.1 |

Table 5. Ablation study of energy and loss terms on the HOT dataset.

5. Limitation

Our current approach mainly focuses on single-hand interaction scenarios and leverages depth input for reliable geometry observation rather than RGB settings. While effective, the method assumes a reasonably reconstructed object mesh; in rare cases where object geometry is significantly degraded, force-aware optimization may be affected. Extending the framework to handle multi-hand interactions, further improving object reconstruction robustness, and refinement for object meshes based on force optimization, remain valuable directions for future work.

References

- [1] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12890–12900, 2023. 1
- [2] Chunpeng Jiang, Wenqiang Xu, Yutong Li, Zhenjun Yu, Longchun Wang, Xiaotong Hu, Zhengyi Xie, Qingkun Liu, Bin Yang, Xiaolin Wang, et al. Capturing forceful interaction with deformable objects using a deep learning-powered stretchable tactile array. *Nature Communications*, 15(1):9513, 2024. 1
- [3] 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 IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 1
- [4] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 1