Supplementary Material:

H2ONet: Hand-Occlusion-and-Orientation-aware Network for Real-time 3D Hand Mesh Reconstruction

Hao Xu[1,2] Tianyu Wang[1] Xiao Tang[1] Chi-Wing Fu[1,2,3]
[1]Department of Computer Science and Engineering [2]Institute of Medical Intelligence and XR
[3]Shun Hing Institute of Advanced Engineering
The Chinese University of Hong Kong
{xuhao,wangty,cwfu}@cse.cuhk.edu.hk, xtang@link.cuhk.edu.hk

There are two parts in this supplementary material.

Part 1 presents additional results to analyze the distribution of hand orientations and frame gaps in the Dex-YCB and HO3D-v2 datasets.

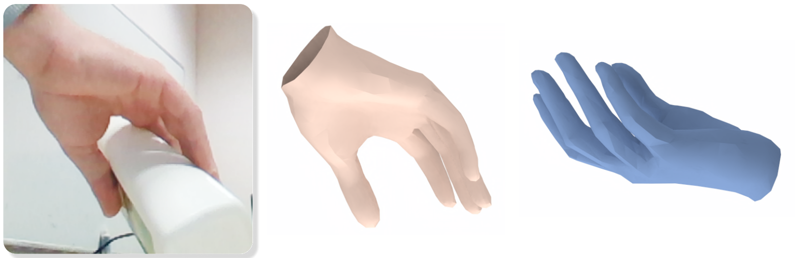
Part 2 presents more quantitative and qualitative results.

Part 1: Dataset Analysis

We provide visualization results to analyze the hand orientation distribution and frame gaps in both the Dex-YCB [1] and HO3D-v2 [2] datasets.

Part 1.1: Canonical Pose

To have a better intuition, we provide a comparison between the normal and canonical poses, as shown in Fig. 1.



(a) Input image (b) Normal pose (c) Canonical pose

Figure 1. Comparison between the normal pose and canonical pose.

Specifically, the canonical pose is defined by setting the angle and axis of the hand orientation to zero and unit vector, respectively.

Part 1.2: Multi-frame Selection

Next, we plot 3D vertex error and Euler angle error for different frame gap of the Dex-YCB and HO3D-v2 datasets in Fig. 2.

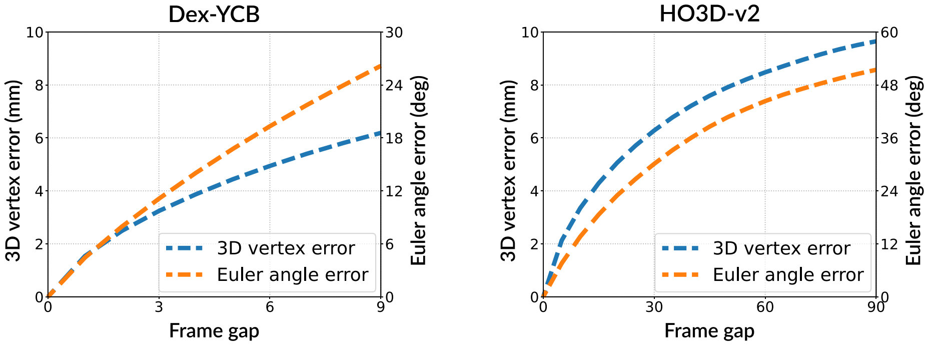


Figure 2. Frame gaps of different datasets.

We compute the average error of the 3D vertices and Euler angles of the hand orientations between input frames $\mathbf{I}_t$ and $\mathbf{I}_{t-k}$ in the training set, where k ranges from 1 to 9 for the Dex-YCB dataset and from 1 to 90 for the HO3D-v2 dataset. We can observe from Fig. 2 that the 3D vertex error is within 10 mm even though the rotation angles between two frames become relatively large, implying that the transformation among nearby frames is mostly rigid.

Besides, we set the frame gap k among the input multi-frames for the Dex-YCB dataset to 5 and for the HO3D-v2 dataset to 30. For the former, as the videos are already frame-drawn and each sequence has $\sim$70 frames in length, a relatively small frame interval ensures that sufficient multi-frame samples are generated from each video. For the latter, the movements of subjects are relatively slow and each sequence has $\sim$1,000 frames in length, so a larger frame gap allows some redundant frames to be filtered out and ensures the differences among frames to be large enough to provide more information about different parts of the hand. For the case where the frame number is less than the frame gap at the

Models	PA-J-PE ↓	PA-V-PE ↓	J-PE ↓	V-PE ↓
Baseline	5.65	5.45	14.02	13.03
$k = 1$	5.54	5.36	13.87	12.91
$k = 7$	5.70	5.52	14.10	13.13
$k = 5$	**5.30**	**5.19**	**13.68**	**12.70**

Table 1. Results on the Dex-YCB dataset with different frame gaps. The best results are marked in **bold** for a better comparison.

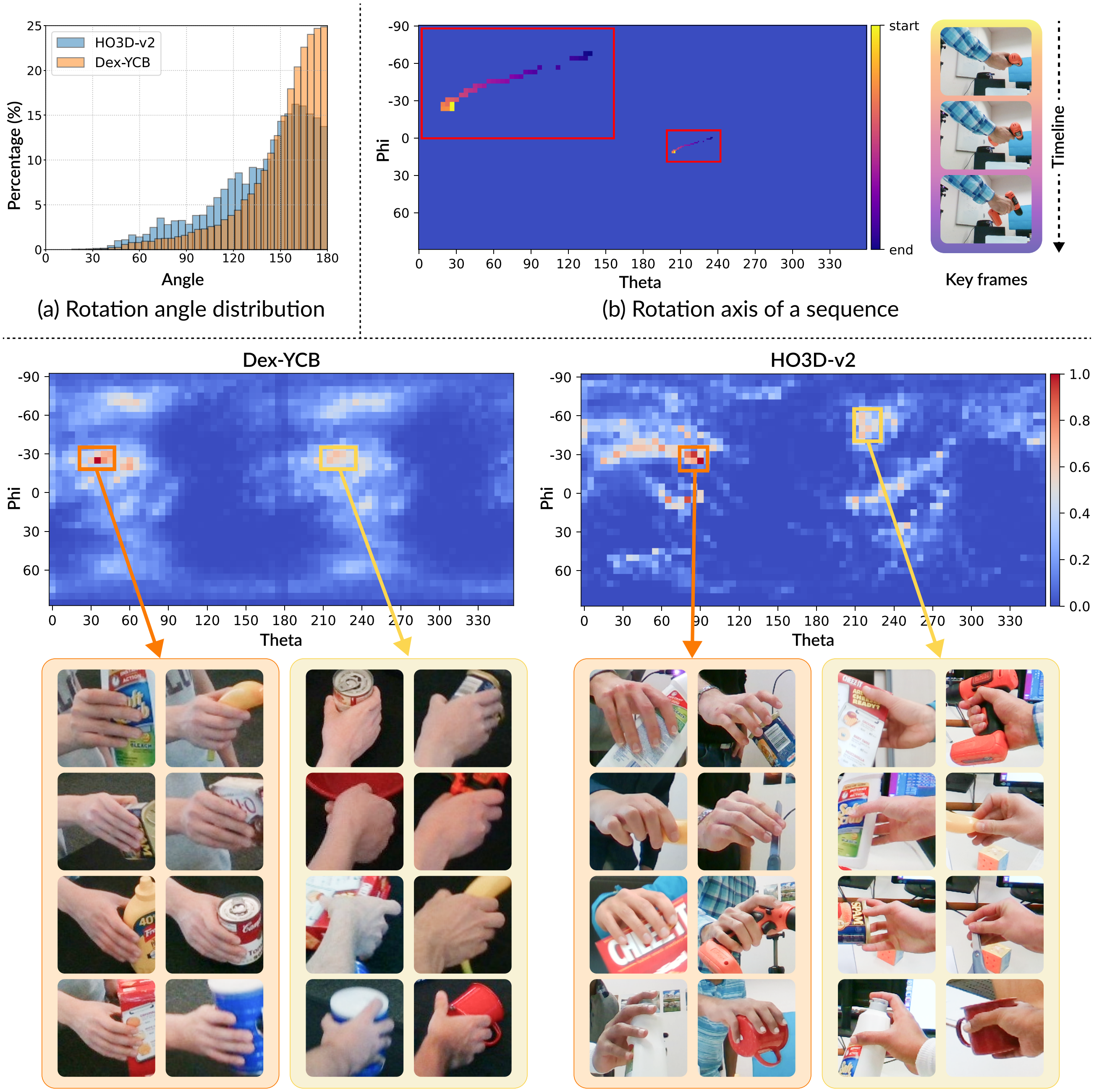
beginning of each sequence, we take three frames at equal gaps. Note that for the first frame, we make three copies of itself as input. Moreover, it can be found that the 3D rotation differences among frames in the HO3D-v2 dataset are less even, as shown in Fig. 2. To facilitate the robustness to different speeds of the hand motion, we apply the augmentation on the frame selection during training on the HO3D-v2 dataset, where the previous two frames are randomly chosen from $[t-k, t-k+5]$ and $[t-2k, t-2k+5]$, respectively. Furthermore, we set different frame gaps, conduct experiments on the Dex-YCB dataset, and use the single-frame version of our method as a baseline, as shown in Table 1. We can see that adopting consecutive frames as input brings a slight improvement, while an excessive value leads to performance degradation. Therefore, we choose a middle ground $k = 5$ for our experiments on the Dex-YCB dataset.

Part 1.3: Orientation Distribution Comparison

In this section, we conduct analysis on the hand orientation distribution in the Dex-YCB and HO3D-v2 datasets, as shown in Fig. 3. Fig. 3(a) presents the rotation angle distribution, and we can see that the distribution of the Dex-YCB dataset is smoother than that of the HO3D-v2 dataset due to its large number of sequences, which are more uniformly captured in more different camera views.

For a better understanding of the rotation axis distribution, we visualize the varying process of the axis for a short sequence in Fig. 3(b). The rotation axis is first normalized to a unit vector and then plotted on the surface of a unit sphere. Then, to show all the rotation axes distributed on the sphere at the same time, we unfold the sphere into a rectangle along one of the meridians. We can find that the position of the rotation axis changes continuously on the sphere as the video plays.

In addition, we present the rotation axes of all the training samples in these two datasets, as shown in Fig. 3(c). For a better understanding, some example images corresponding to the most common axes of rotation in the two datasets are shown below. Comparing the distributions of these two datasets, it can be found that the distribution of the HO3D-v2 dataset is more concentrated around some common poses due to the limited number of video sequences and less evenly distributed camera views. We find in experiments that our hand orientation regression module can easily overfit the HO3D-v2 dataset while performing better than previous methods on the Dex-YCB dataset, so we choose to pre-train one epoch on the Dex-YCB dataset to alleviate the overfitting problem.



(a) Rotation angle distribution

(b) Rotation axis of a sequence

(c) Rotation axis distribution

Figure 3. Some visual results of the hand orientation distribution.

Part 2: More Experimental Results

In this section, we conduct more experiments and present more results to demonstrate the effectiveness of our method.

Part 2.1: Finger-level Occlusion Classifier

First, we evaluate the performance of different network designs to show the effectiveness of our finger-level occlusion classifier architecture. Since some 2D joints may lie too close to each other, applying a shared network to predict occlusions only from the feature of individual fingers may lead to confusion. We compare the performance of adopting shared MLP and individual MLP (Ours), as shown in Table 2.

Models	Thumb	Index	Middle	Ring	Little
Dex-YCB test set					
Shared MLP	83.6	83.8	83.6	82.6	85.4
Ours	**91.3**	**91.1**	**91.1**	**91.2**	**91.5**
HO3D-v2 validation set					
Shared MLP	88.4	92.8	88.8	87.0	88.2
Ours	**97.9**	**98.1**	**97.7**	**97.9**	**97.2**

Table 2. Accuracy comparison of different designs of finger-level occlusion classifier.

It can be found that the individual structure can improve the classification accuracy of all fingers on both the Dex-YCB test set and the HO3D-v2 validation set, showing that the non-shared design can alleviate the confusion issue. Note that the ground truths of the HO3D-v2 test set are not publicly available, so we split 5% samples from the training set uniformly as the validation set.

Part 2.2: More Qualitative Results

We also provide more qualitative results on the Dex-YCB and HO3D-v2 datasets, as illustrated in Fig. 5 and Fig. 4. Most of them contain severe occlusions.

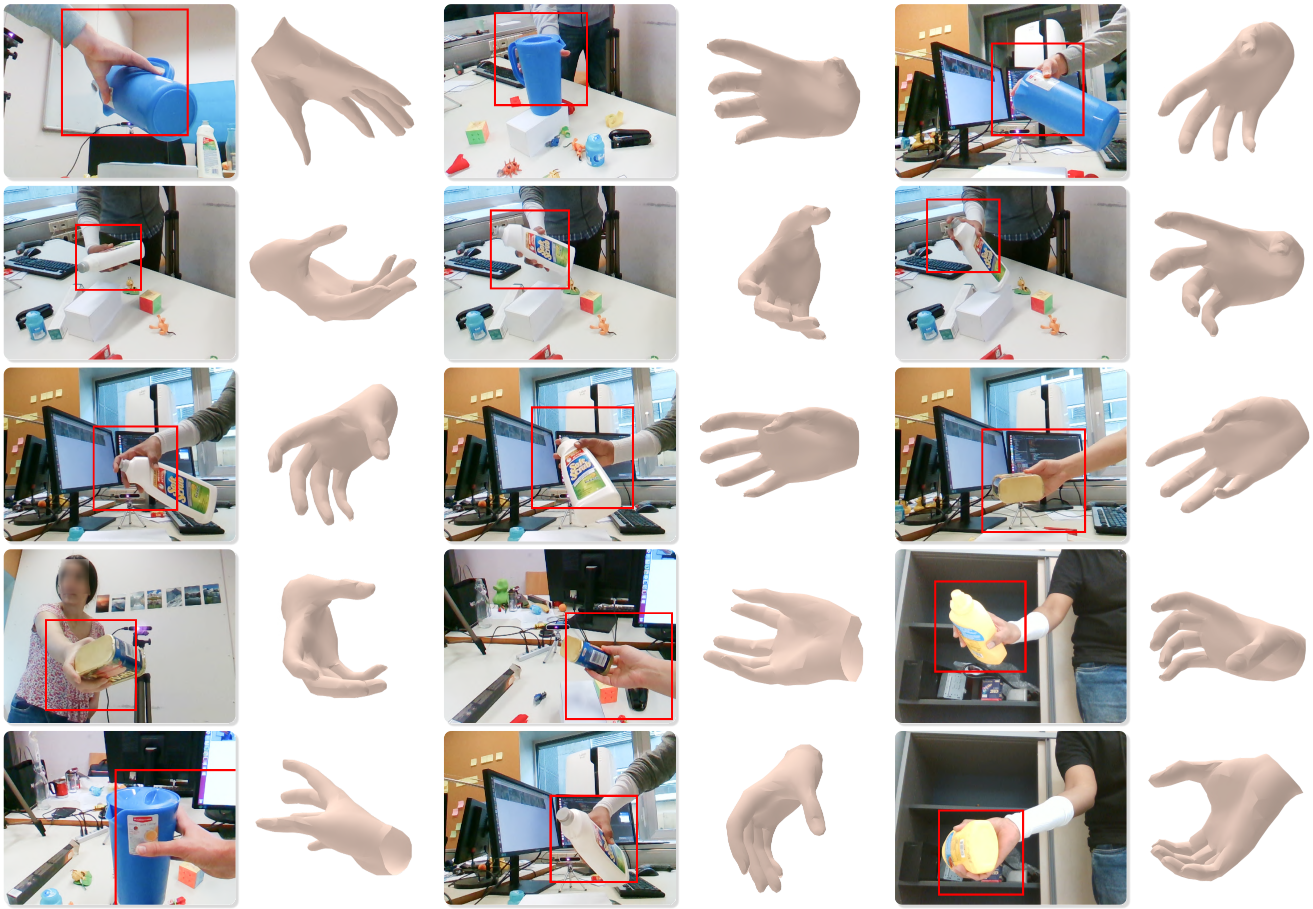


Figure 4. More qualitative results of our method on samples containing severe occlusions in the HO3D-v2 dataset.

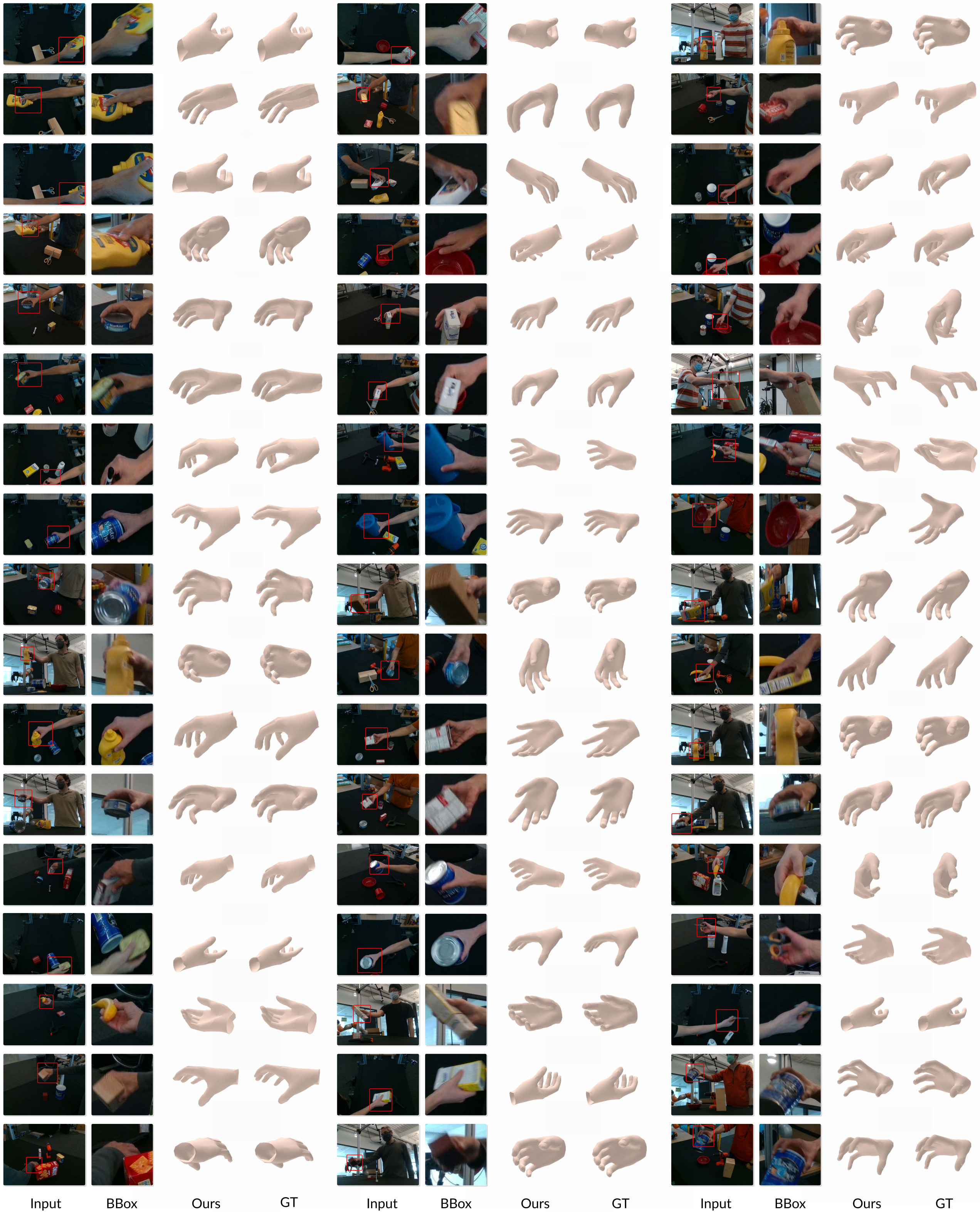


Figure 5. More qualitative results of our method on samples containing severe occlusions in the Dex-YCB dataset.

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*, pages 9044–9053, 2021. 3

[2] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 3