Interacting Hand-Object Pose Estimation via Dense Mutual Attention Supplementary Material

Rong Wang Wei Mao Hongdong Li
The Australian National University
{rong.wang, wei.mao, hongdong.li}@anu.edu.au

In this supplementary document, we present:

- the details of the network architecture (Section 1);
- quantitative results on the additional object metric (Section 2);
- more qualitative results (Section 3);
- analysis of failure cases (Section 4).

1. Network Architectures

In this section, we describe the architecture of each module used in our network. At the initial stage, we use the ResNet-50 [3] network pre-trained on the ImageNet [8] for the image encoders $f_e^h(\cdot)$ and $f_e^o(\cdot)$. We denote the image features output from the final layer as $\mathbf{F}^h, \mathbf{F}^o \in \mathbb{R}^{8\times8\times2048}$. For the joint estimator $f_j^h(\cdot)$ and mesh estimators $f_m^h(\cdot)$, $f_m^o(\cdot)$, we use a similar architecture differ only on the output vertex numbers (*i.e.* 21, 778 and 1000 respectively). Take the hand joint estimator as an example, we first decode the image feature $\mathbf{F}^h$ using 4 transposed 2D convolutional layers to obtain a decoded feature $\mathbf{F}_J^h \in \mathbb{R}^{64\times64\times256}$. We then forward the features into 1D convolutional layers as:

$$\begin{aligned}\mathbf{j}_x &= \text{soft-argmax}(\text{Conv1D}(avg^x(\mathbf{F}_J^h)))\\ \mathbf{j}_y &= \text{soft-argmax}(\text{Conv1D}(avg^y(\mathbf{F}_J^h)))\\ \mathbf{j}_z &= \text{soft-argmax}(\text{Conv1D}(flat^{x,y}(\mathbf{F}^h)))\end{aligned} \tag{1}$$

where the positions of 21 hand joints can be found as $\hat{\mathbf{J}}^h = [\mathbf{j}_x, \mathbf{j}_y, \mathbf{j}_z] \in \mathbb{R}^{21\times3}$. avg^i denotes averaging along the i-th dimension (where we denote the feature shape as $\mathbb{R}^{x,y,c}$), and $flat^{x,y}$ denotes flattening the feature map in the spatial domain (so that the resulting feature has shape $\mathbb{R}^{64,2048}$). The independent convolutional layers Conv1D transform the feature channels to $L = 64$ matching the lixel definition. Compared to [9], we replace linear layers with convolutional layers to estimate the depth component, which achieves competitive results yet significantly reduces the model complexity. For the object mask estimator $f_s^o(\cdot)$, we follow [7] to adopt a U-Net architecture, where we use 4 2D convolutional layers as the encoder and 4 bilinear interpolation operation layers as the decoder.

At the refinement stage, we use $N = 4$ graph convolutional blocks, where the first and last blocks both contain 1 graph convolutional layer, and the rest blocks each contain 2 residually connected graph convolutional layers. To increase inference speed, we only include one mutual attention layer following [1] after the first block. We adopt a single 1D convolutional layer as the global feature fusion unit $f_g(\cdot)$ to obtain the unique global feature. In terms of mutual attention, we predict the keys, queries, and values with the same channel number $H = F = 64$. We fuse the aggregated features with the original node features using the value fusion unit $f_u(\cdot)$, which is implemented as another 1D convolutional layer.

2. More Results on the Object Metric

In order to compare the object pose estimation results with [5, 1], we additionally report the maximum symmetry-aware surface distance (MSSD) [4] on the HO3D v2 testing set in Table 1. The metric measures the maximum vertex distance between the predicted and ground truth object pose in all rotation invariants. We follow the same symmetry axes as described in [1] to define the aforementioned rotation invariants, and retrain the model with the symmetry-aware object corner loss proposed in [1] to reflect the symmetry. The results show that our method achieves competitive results with [5, 1] on the MSSD metric.

Table 1. **Quantitative comparison for the object pose estimation on the MSSD metric.** Best results are highlighted in **bold**.

Methods	MSSD $(cm)\downarrow$		
	mustard bottle	bleach cleanser	potted meat can
Artiboost [5]	**3.14**	5.72	6.36
Keypoint Trans [1]	4.41	6.03	9.08
Ours	4.30	**5.23**	**6.14**

3. More Qualitative Results

We provide more qualitative results on the HO3D and DexYCB testing sets in Figure 1. It can be seen that our

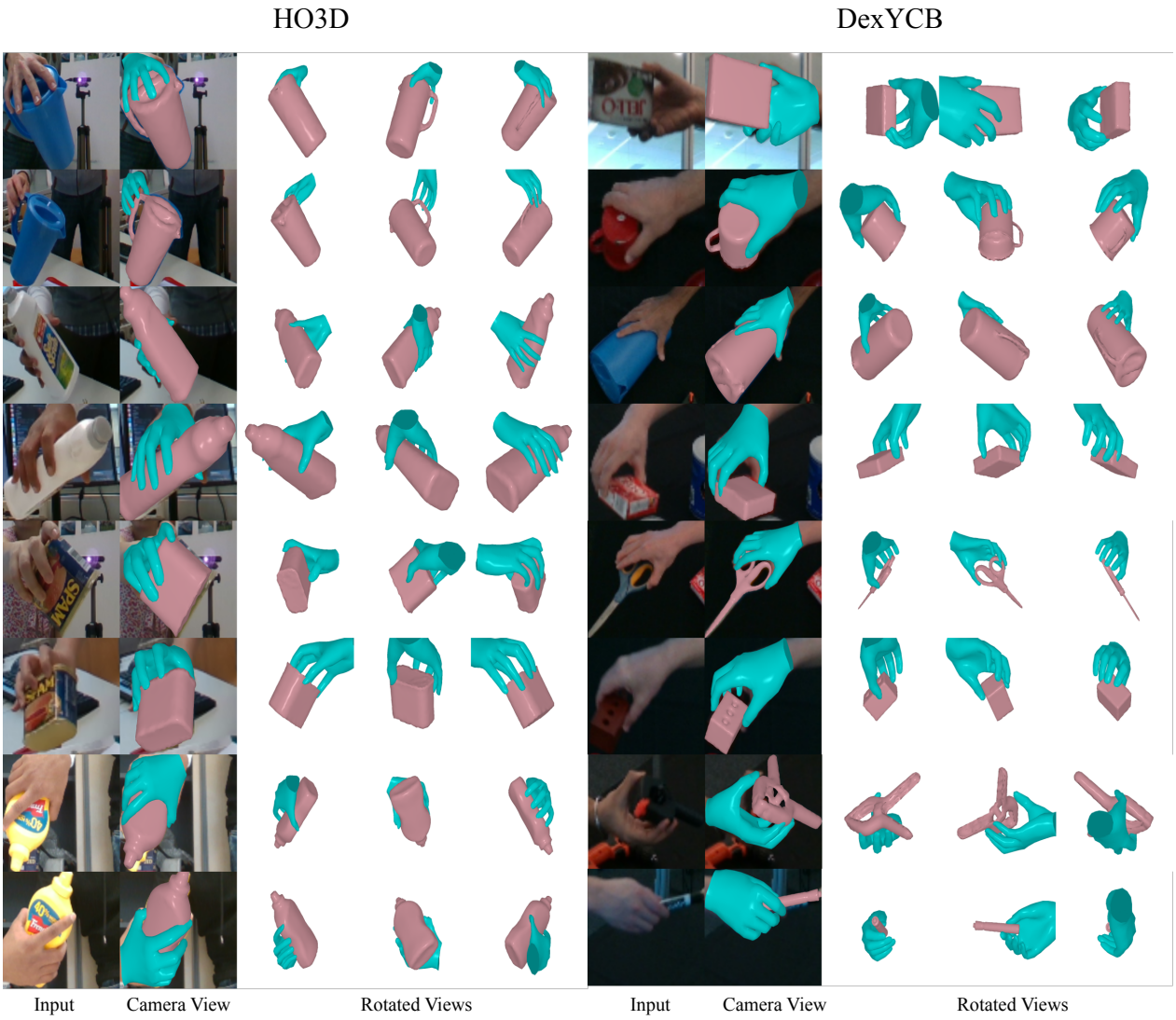

Figure 1. **More qualitative results on the HO3D and DexYCB testing sets.** Our method produces accurate hand-object poses on diverse object classes and grasping approaches. More views are shown in the supplementary video.

method produces accurate hand and object poses for various hand-object interaction scenes, including different grasping approaches and interacting objects. Moreover, we observe that valid contacts that satisfy physical constraints can be established in the estimated hand-object pose.

For completeness, we compare the estimated hand-object pose with the recent work [2] in Figure 2. We observe that our method produces much-improved results that better align with input images, showing the superiority of our proposed techniques.

4. Analysis of Failure Cases

We show the cases where our method fails to estimate the correct hand and object pose in Figure 3. In particular, when the hand is partially outside the camera's field of view, we observe that our method is unable to generate hand meshes with a well-defined shape. The corrupted hand features also impaired the estimation of the object pose. This happens since we rely on the lixel [6] representation for the mesh prediction. Recall that we quantize the image space to convert the pixel coordinates into lixels, therefore there are no valid correspondences for vertices that can not be projected onto the image space. Future works should consider more expressive mesh representations to handle the aforementioned issue.

References

[1] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identifica-

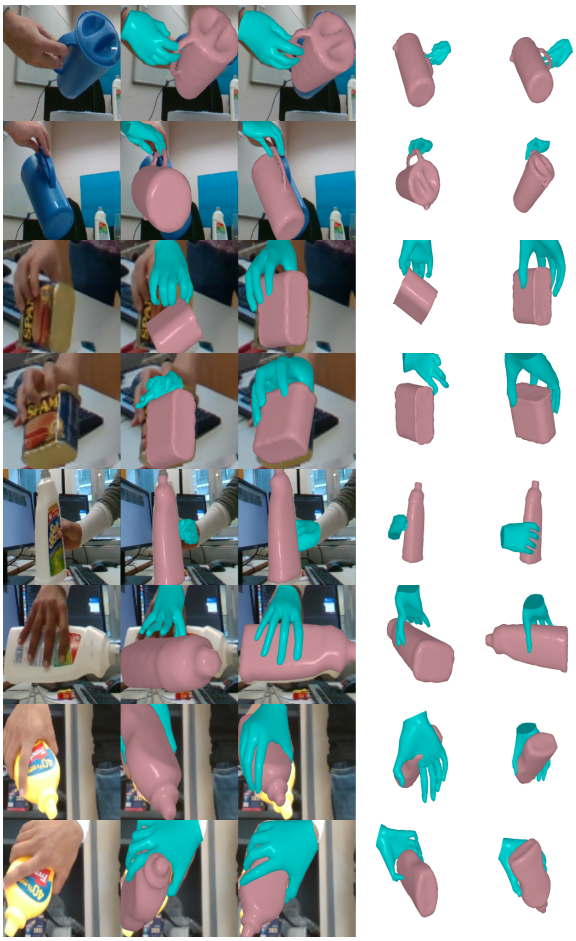

Figure 2. **Qualitative comparison with state-of-the-art method.** We compare the results of [2] in the camera view and a rotated view. Our method produces hand-object poses of higher quality with less penetration and more contact.

Figure 3. **Examples of failure cases.** Our methods can not recover a correct hand pose when the hand is partially outside the camera's field of view, as there is no valid lixel correspondence exists.

tion in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022.

[2] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *European Conference on Computer Vision*, pages 577–594. Springer, 2020.

[5] Kailin Li, Lixin Yang, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. *arXiv preprint arXiv:2109.05488*, 2021.

[6] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[9] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 11698–11707, 2021.