SUPPLEMENTARY MATERIAL

1. IMPLEMENTATION DETAILS

We use ResNet-18 [1] to extract the input image feature. We follow the MANO layer network from [2, 3] to recover hand mesh. We empirically set the hyper-parameter to be $\lambda_1 = 10, \lambda_2 = 10, \lambda_3 = 0.1, \lambda_4 = 0.01, \lambda_5 = 0.1, \lambda_6 = 10, \lambda_7 = 10$ and$\lambda_8 = 10$. The dimension of the encoder latent feature is 512. We train all parts of the network using the AdamW optimizer for 150 epoches with a batch size of 64. We start with an initial learning rate of 10^{-4} for all training settings and lower it by a factor of 10 at the 50th and 100th epoches. We follow [2] and use a learning rate of $1e-6$ and a physical contact loss to refine the hand-object interaction after the 10th epoch.

2. DATASET DETAILS

DexYCB [4] is the latest large-scale RGB-based-dataset which contains 582k samples of hands interacting with different objects and we evaluate using the official "S0" split. The hand-object images in this dataset contain 10 objects modeled from YCB objects [5] and we consider each object as a single category.

Ho3D provides 3D annotations for multiple multi-view sequences of hands and objects. The images in this dataset contain 10 objects modeled also from YCB objects [5]. We use their default training and test split to evaluate our hand-object interaction reconstruction.

HOI4D [6] is the first category-level hand-object interaction dataset which contains 2.4M RGB-D egocentric video frames over 4000 sequences. We utilize their rigid object frames since we focus on hand interacting with rigid objects. This dataset consists of 7 category-level objects with over 50 similar object models for each category.

3. ABLATION STUDY

GAN Module. Our baseline models (Ours w/o GAN) results are presented in Table 1. Our model (w/o GAN) is a pipeline without utilizing the GAN module to learn the object shape offset. It is obvious that by using the GAN module, our pipeline can effectively capture the difference between the input object and the initial template (see Fig. 1). Furthermore, the GAN module significantly reduces the outliers in the estimated mesh model, validating the effectiveness of our proposed conditional GAN module.

Template Initialization. We also compare different template initialization methods in Table 1. Our model (w/o Shape Prior) represents a pipeline without utilizing an initialization template but relies on a sphere to learn the object shape. It is evident that using object initialization allows our pipeline can learn the more accurately the object's shape and pose.

Depth Maps. We show our baseline models (Ours w/o Depth) results in Table 1. Our model (w/o Depth) represents a pipeline without using the depth map as supervision during the training stage. Based on the experimental results, depth information is important for the 3D hand and object shape reconstruction. This is because directly fitting the RGB input to the mask map is challenging, as the mask map fails to capture the detailed geometry of the hand or object.

4. CONTACT MAP

To enhance the quality of hand-object interaction results, we incorporate a contact map during the refinement stage. By following [7], we establish the ground truth contact map by assuming that the closest distance between the hand and the object is under 4mm. During training, we use PointNet to determine whether the point cloud is in contact or not. In the inference stage, this network is then used to estimate the contact region and refine the penetration region. We provide some examples visualization results of contact map estimation Figure 2.

5. INTERACTION REFINEMENT

We provide qualitative examples for interaction refinement in Fig. 3. We show that our model consistently demonstrates superior performance across all comparisons. By incorporating the MANO model, we further refine hand-object interactions through the application of physical contact loss, leading to substantial improvements in interaction refinement. This underscores the efficacy of our proposed hand and object interaction refinement and highlights the critical role of physical contact loss in enhancing interaction fidelity.

Dataset	HOI4D				DexYCB			
Method	**MPJPE**	**MPVPE**	**CD**	**ADD**	**MPJPE**	**MPVPE**	**CD**	**ADD**
w/o GAN	17.31	17.40	11.90	10.6	11.20	11.28	6.8	6.1
w/o Shape Prior	17.35	17.39	10.10	9.7	10.68	11.61	7.1	6.9
w/o Depth	22.41	22.60	14.17	13.7	17.64	17.70	12.1	11.8
Ours (Only Depth)	20.14	20.16	12.48	13.12	13.05	13.49	8.5	9.1
Ours (Full)	17.30	17.38	5.70	5.90	11.15	11.20	5.0	4.6

Table 1. Ablation study on HOI4D and DexYCB test sets.

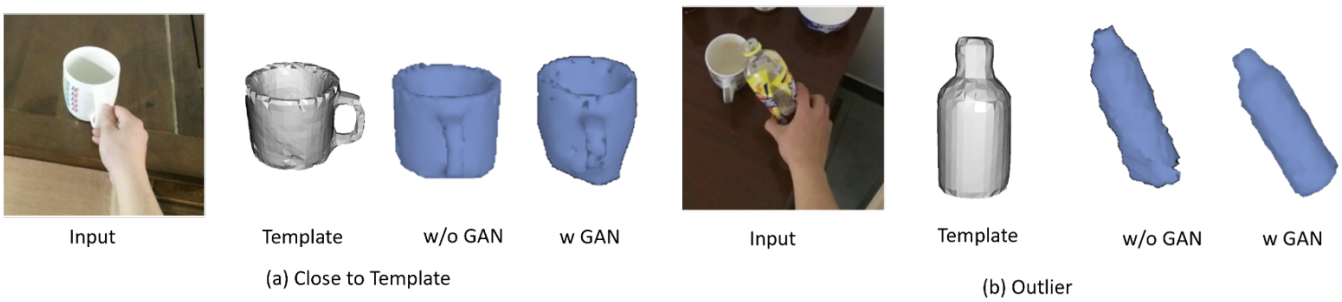

Fig. 1. Overview of our proposed GAN module results. Predicted object shape with correspondence initial template shape demonstrates that the GAN module aids the model in learning the difference between the template and the object mesh (a), and illustrates the GAN module boost the model to reduce the outliers (b). For each row, left to right columns correspond to RGB input, template mesh model, the estimated mesh without the GAN module and our estimated mesh with the GAN module.

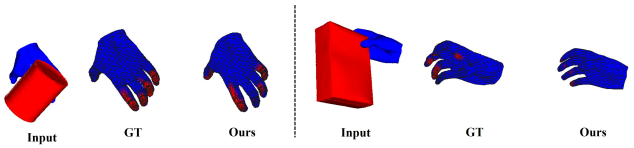

Fig. 2. Contact map results. From right to left, the input hand-object point cloud, the ground truth contact region and the estimated contact regions.

6. MORE QUALITATIVE RESULTS

In this section, we present additional qualitative results for our 3D hand reconstruction in Figure 6. We provide both the multi-view and key points results of the hand reconstruction. As shown in Figure 6, our predicted key points and hand meshes align well with both the hand joints and surface in the images. Moreover, the multi-view results highlight our model's ability to accurately estimate the invisible areas using single-view RGB input. We also present the full hand-object shape reconstruction qualitative comparisons in Figure 4.

7. LIMITATIONS

In our hand-object interaction pipeline, the hand pose reconstruction pipeline utilizes the RGB image as input, with the mask and depth maps as supervision. However, heavy object occlusions impose limitations on our results when compared to the ground truth (see Fig. 7). Even though our initial hand meshes are not well aligned, the output meshes closely approximate the ground truth. To address these limitations, we could explore incorporating additional sources of information, such as sequence or multi-view information, to provide extra supervision and enhance the accuracy of hand mesh reconstruction.

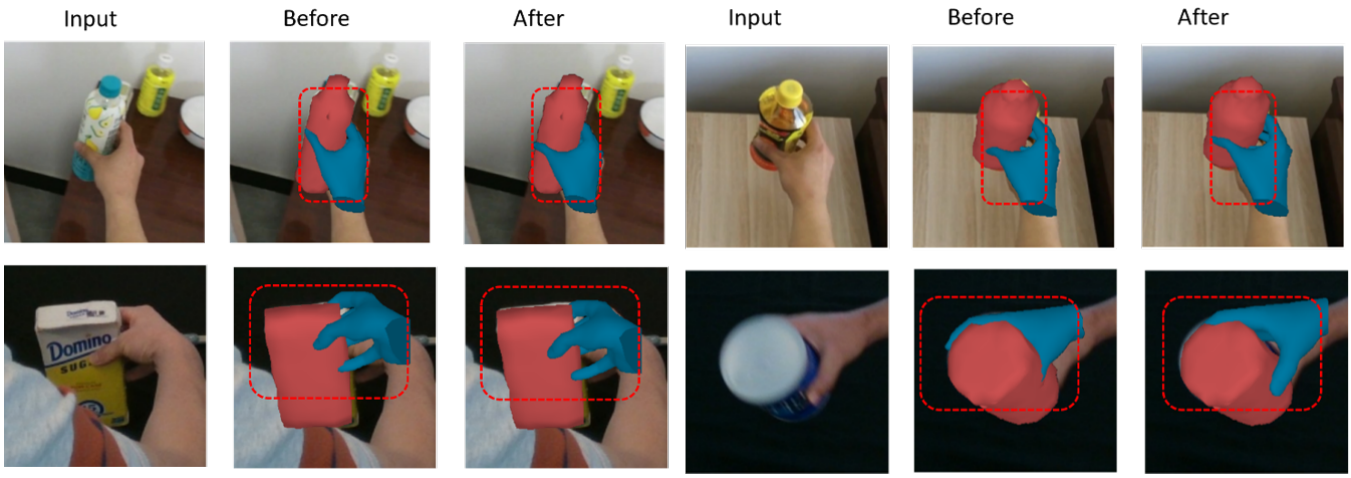

Fig. 3. Interaction refinement results. For each triplet, left to right columns correspond to the RGB input, our meshes before and after interaction refinement. Red boxes highlight the interaction refinement regions.

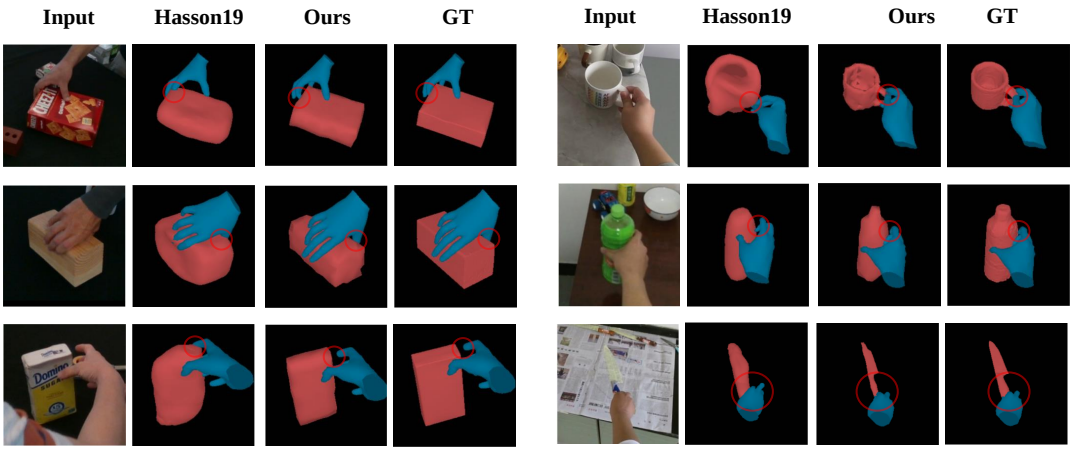

Fig. 4. Hand-object shape reconstruction results. For each row, left to right columns correspond to RGB input, template-free based method [2], template-based method [3], our method and ground truth in camera view.

8. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[2] 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.

[3] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev, "Towards unconstrained joint hand-object reconstruction from rgb videos," in *3DV*, 2021.

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox, "DexYCB: A benchmark for capturing hand grasping of objects," in *CVPR*, 2021.

[5] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[6] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *CVPR*, 2022.

[7] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *CVPR*, 2020.

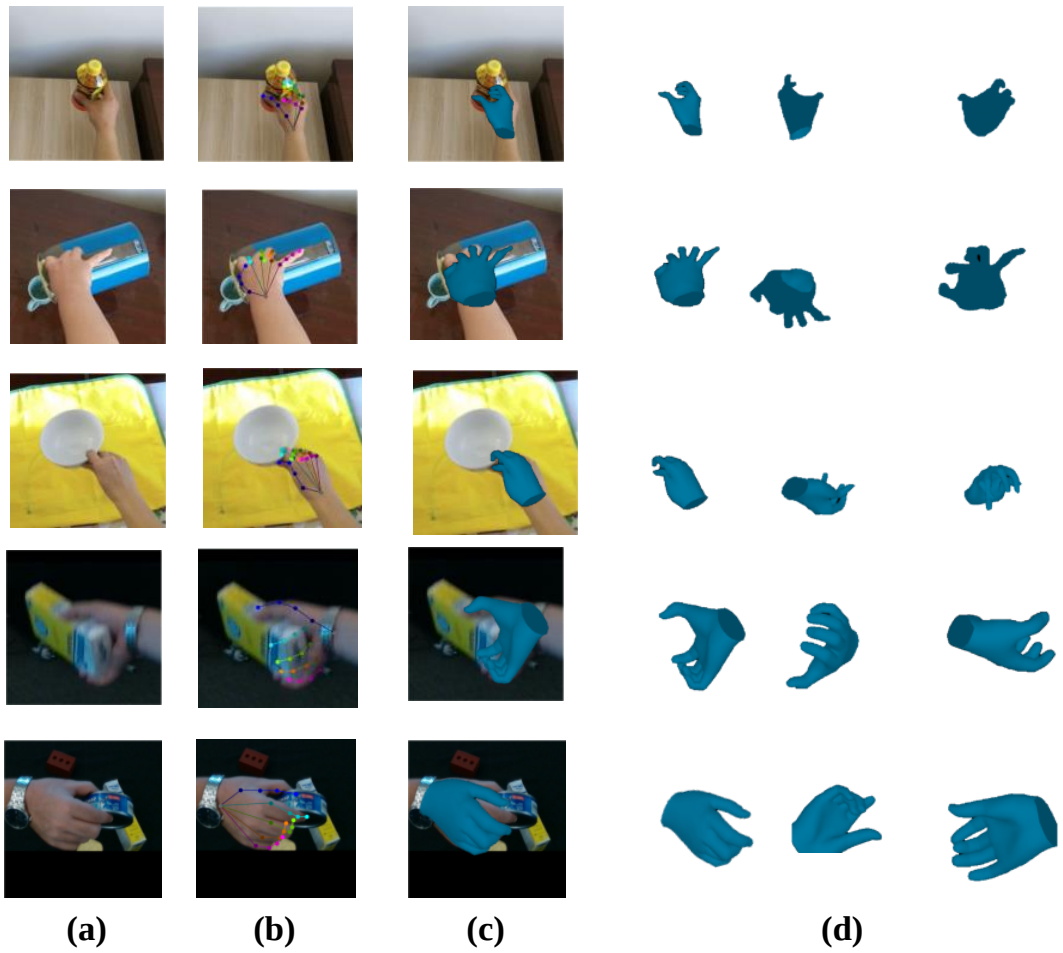

Fig. 5. Additional qualitative examples. From left to right: (a) RGB input, (b) 2D key points results, (c) projection of the reconstructed mesh on the original image, (d) and the multi-view visualization of reconstructed 3D meshes. We show that our pipeline yields highly accurate and plausible 3D hand meshes.

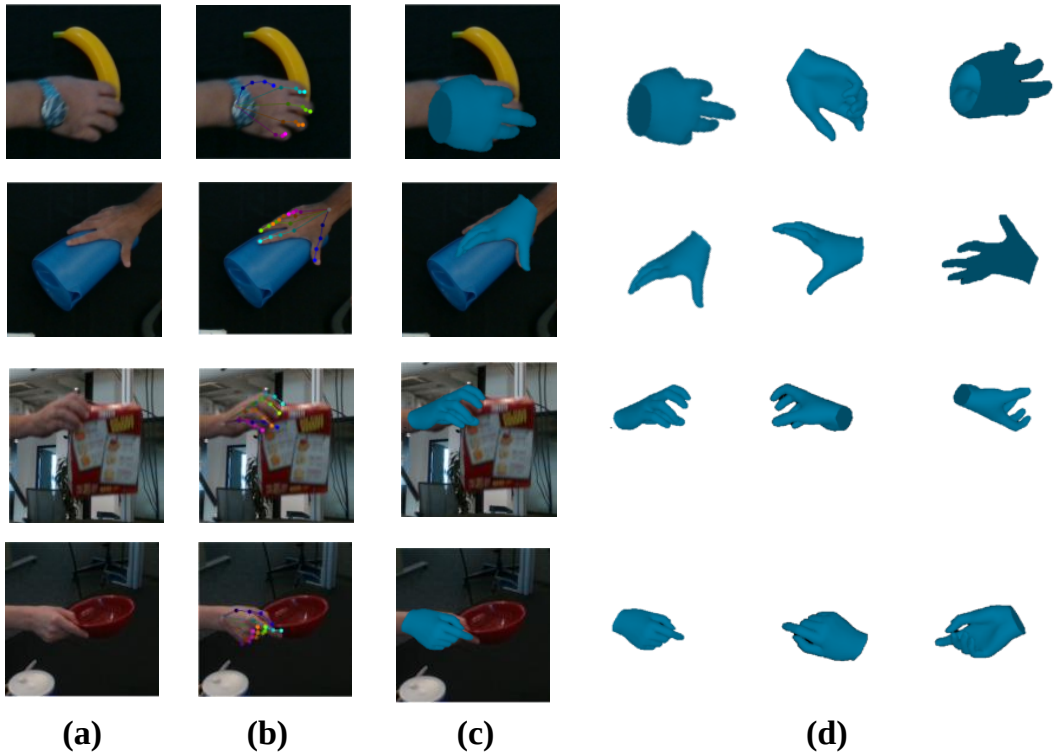

Fig. 6. Additional qualitative examples. From left to right: (a) RGB input, (b) 2D key points results, (c) projection of the reconstructed mesh on the original image, (d) and the multi-view visualization of reconstructed 3D meshes. We show that our pipeline yields highly accurate and plausible 3D hand meshes.

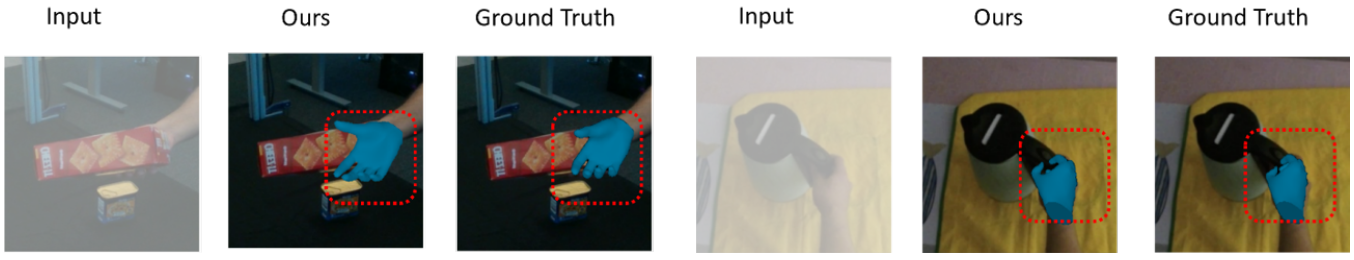

Fig. 7. Limitation results. For each row, left to right columns correspond to RGB input, our hand mesh and ground truth. We are limited by heavy object occlusions. Red boxes highlight the non-aligned regions.