What's in your hands? 3D Reconstruction of Generic Objects in Hands Supplementary Material

A. Implementation Detail

A.1. Camera conversion

The off-the-shelf system predicts a weak perspective camera with a scale factor s and 2D translation t_x, t_y. One can transform the point via the global hand rotation and translation and then project it via the predicted camera s, t_x, t_y.

$$sT_{\theta_w}X + (t_x, t_y)$$

We found that a full perspective camera help to account for large perspective effect. Therefore, we convert the weak perspective camera to a full perspective one by translating the final mesh by an offset $(t_x, t_y, f/s)$. In summary, we project a query point in the wrist frame to the image by

$$\pi_{\theta_w}(X) = K[T_{\theta_w}X + (t_x, t_y, f/s)]$$

A.2. Coordinate Transformation

Our articulation embedder takes as input an articulation parameter θ_A and a point position in wrist frame X to output the articulation-aware encoding $\psi = h(X; \theta_A)$. The encoding is a concatenation of the coordinates relative to every joint. Given the articulation parameter θ_A, we run forward kinematics to derive transformation $T(\theta_A) : \mathbb{R}^3 \rightarrow \mathbb{R}^{45}$ that maps a point in wrist frame to each joint coordinate.

The transformation between wrist to one joint T_j is computed by forward kinematics chain. Consider one bone that connects joint j to its child i (*e.g.* index proximal phalanx). The transformation matrix from this joint frame to its child joint frame would be

$$T_{ji} = \begin{pmatrix} R(\theta_j) & t_{ji} \\ 0 & 1 \end{pmatrix}$$

where t_j is the bone length pre-defined in MANO models. Then the transformation from wrist to any joint is the product of every transformation in the kinematic chain $T_j = T_{wi} \cdot T_{ik} \cdot \ldots T_{lj}$. The coordinate of the queried point relative to the joint becomes ${}^jX = {}^j\ T_w^w X$.

A.3. Training

We train our model using Adam optimizer with learning rate $1e-4$ on 8 GPUs. The batch size is 64. We train our model on ObMan for 200 epochs and finetune it on HO3D and RHOI for 50k iterations respectively. The coefficient of eikonal term is 0.1.

A.4. HO3D dataset split

HO-3D [1] is a real-world video dataset consisting of 103k annotated images capturing 10 subjects interacting with 10 common YCB objects [4]. The original train-test splits are created by partitioning the interaction sequences. Sequences in the original test set involve only 4 objects of which three appear in train set (bleach cleanser, mustard bottle, meat can) and all of them are cuboidal shape. To test on more non-trivial shapes like power drill, we create a custom split by holding out one video sequence per object as test set. We list our sequences for train and test set in Table 1.

B. Qualitative Results

We provide more qualitative results rendered in the image frame and from another view in this PDF and video results when moving camera around the object in the zipped website.

Figure 1 visualizes reconstruction from our method and two baselines [2,3] on ObMan dataset from the image frame and a novel view.

Figure 2 visualizes reconstruction from our method and two baselines [2, 3] on HO3D dataset from the image frame and a novel view.

Figure 3, 4 visualizes reconstruction from our method and two baselines [2, 3] on RHOI dataset from the image frame and a novel view.

Figure 5 visualizes reconstruction of hand-held object with or without explicitly considering hand pose on ObMan, HO3D and RHOI.

Figure 6 visualizes reconstruction of hand-held object from our models that only trained on ObMan and RHOI datasets.

Figure 7 visualizes hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.

Figure 1. Visualizing reconstruction from our method and two baselines [2, 3] on ObMan dataset from the image frame and a novel view.

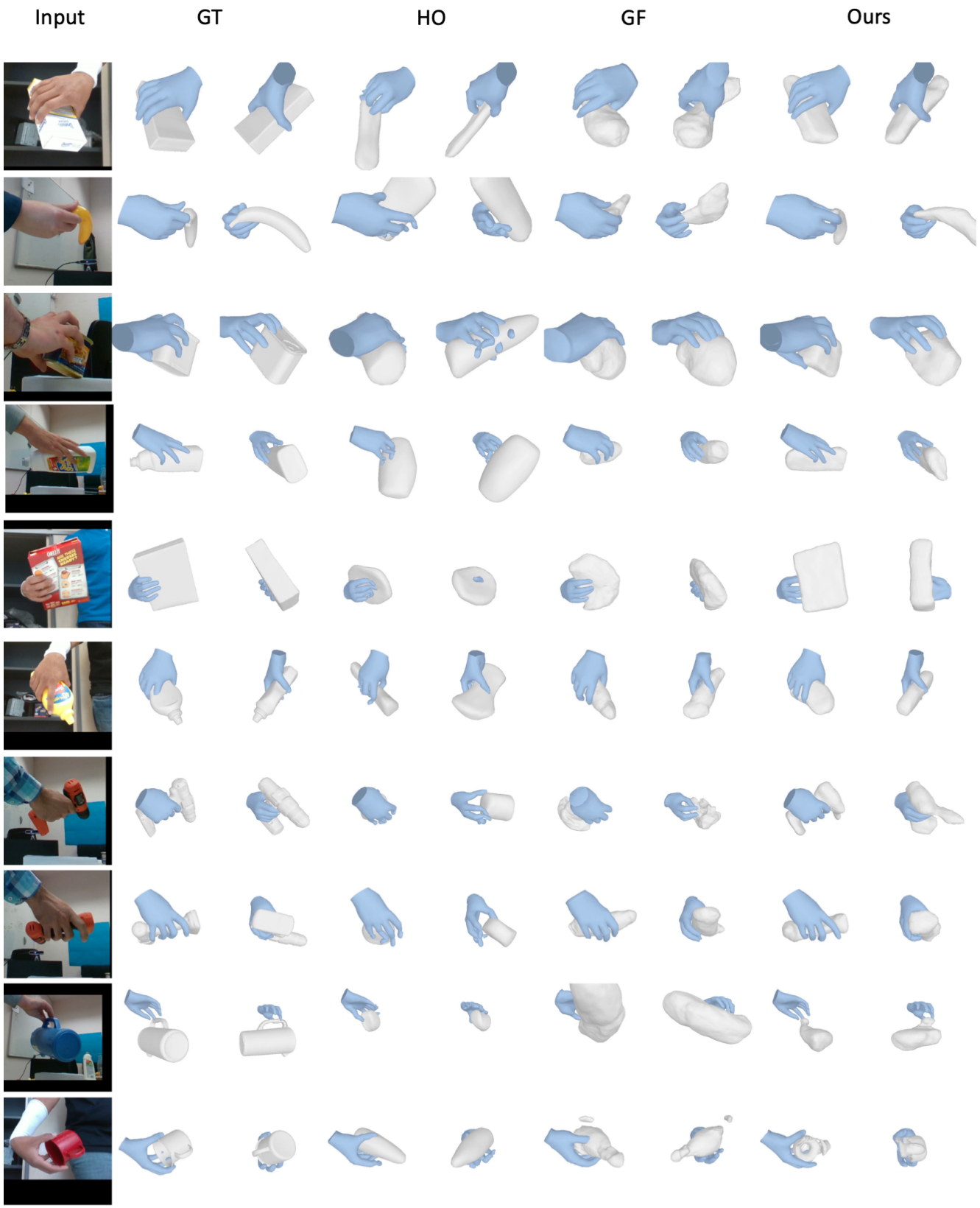

Figure 2. Visualizing reconstruction from our method and two baselines [2, 3] on HO3D dataset from the image frame and a novel view.

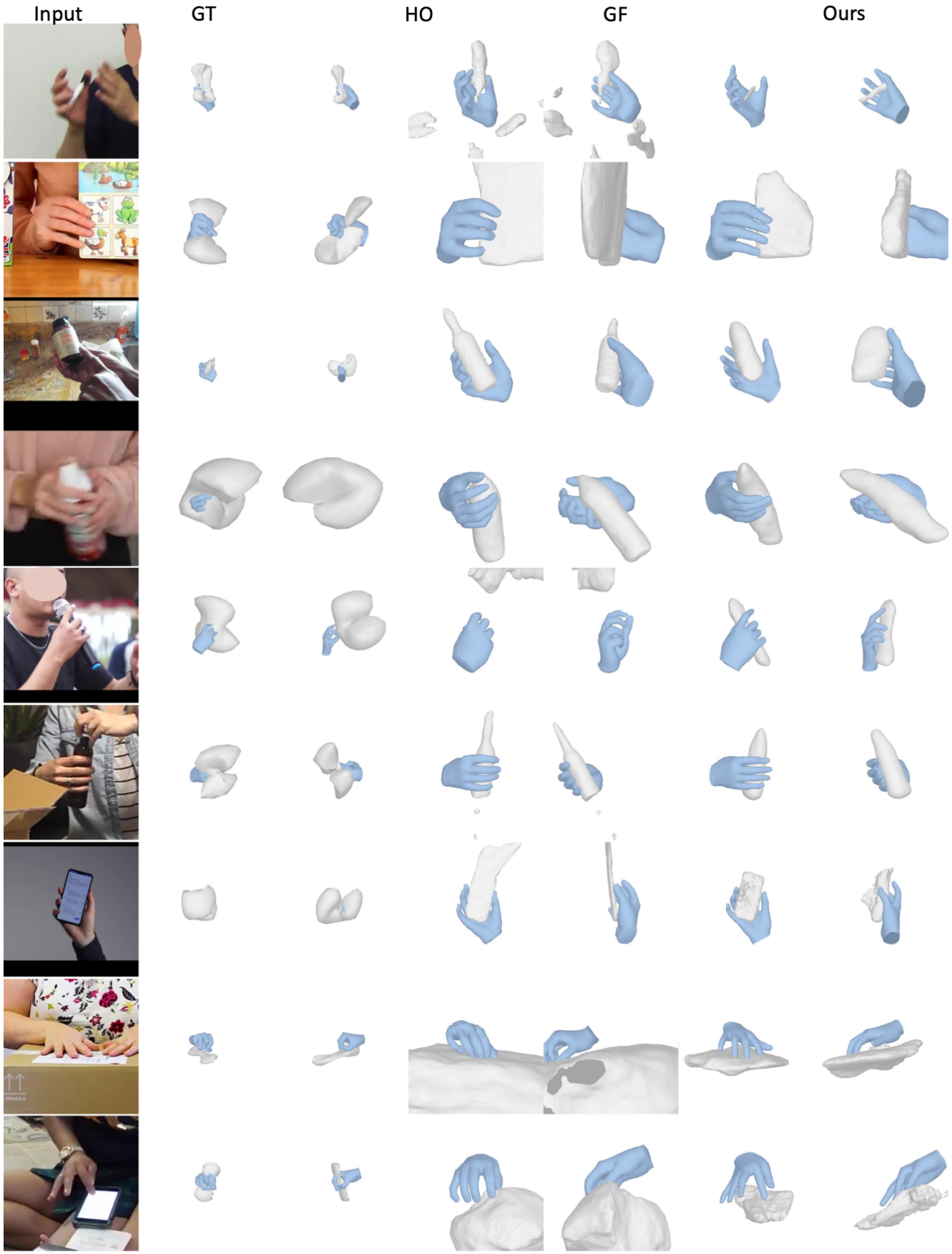

Figure 3. Visualizing reconstruction from our method and two baselines [2, 3] on RHOI dataset from the image frame and a novel view.

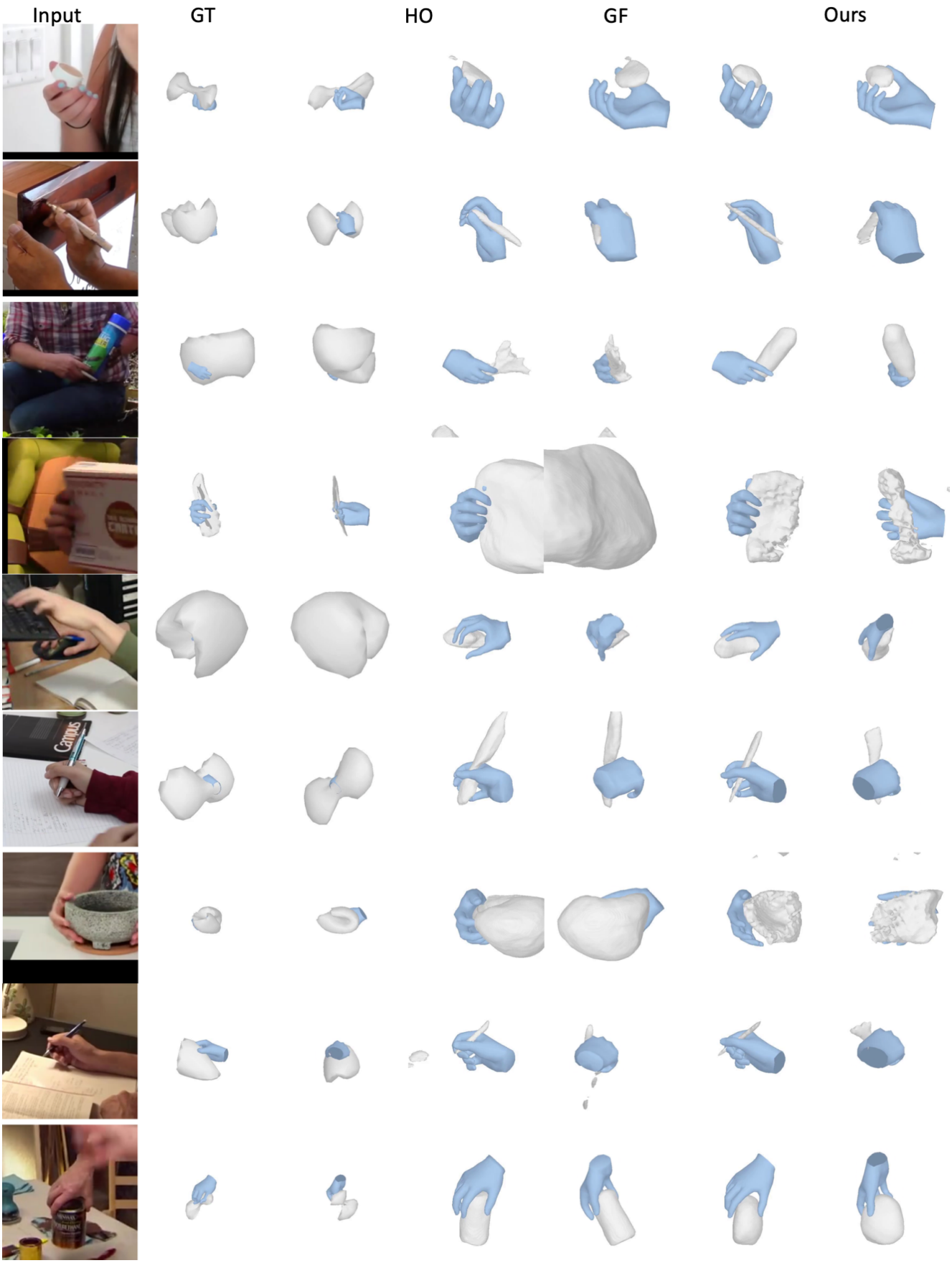

Figure 4. Visualizing reconstruction from our method and two baselines [2, 3] on RHOI dataset from the image frame and a novel view.

Objects	Test Sequences	Train Sequence
010_potted_meat_can	GPMF10	MPM14, GPMF13, MPM12, GPMF12, MPM11, GPMF11, MPM13, MPM10, GPMF14
021_bleach_cleanser	ABF10	SB11, SB12, ABF11, ABF13, SB10, ABF12, ABF14, SB13, SB14
019_pitcher_base	AP10	AP11, AP14, AP13, AP12
003_cracker_box	MC1	MC2, MC6, MC5, MC4
006_mustard_bottle	SM1	SM5, SM2, SM4, SM3
004_sugar_box	SS1	ShSu12, SiS1, SS2, ShSu14, ShSu13, SS3, ShSu10
035_power_drill	MDF10	MDF12, MDF14, MDF11, ND2, MDF13
011_banana	BB10	BB12, SiBF10, SiBF14, SiBF11, SiBF12, BB13, BB11, SiBF13, BB14
037_scissors	GSF10	GSF13, GSF12, GSF14, GSF11
025_mug	SMu1	SMu41, SMu42, SMu40

Table 1. Our customized split on HO3D dataset.

References

[1] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 1

[2] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2, 3, 4, 5

[3] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*. 1, 2, 3, 4, 5

[4] Berk Çalli, Arjun Singh, Aaron Walsman, Siddhartha S. Srinivasa, P. Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. *ICAR*, 2015. 1

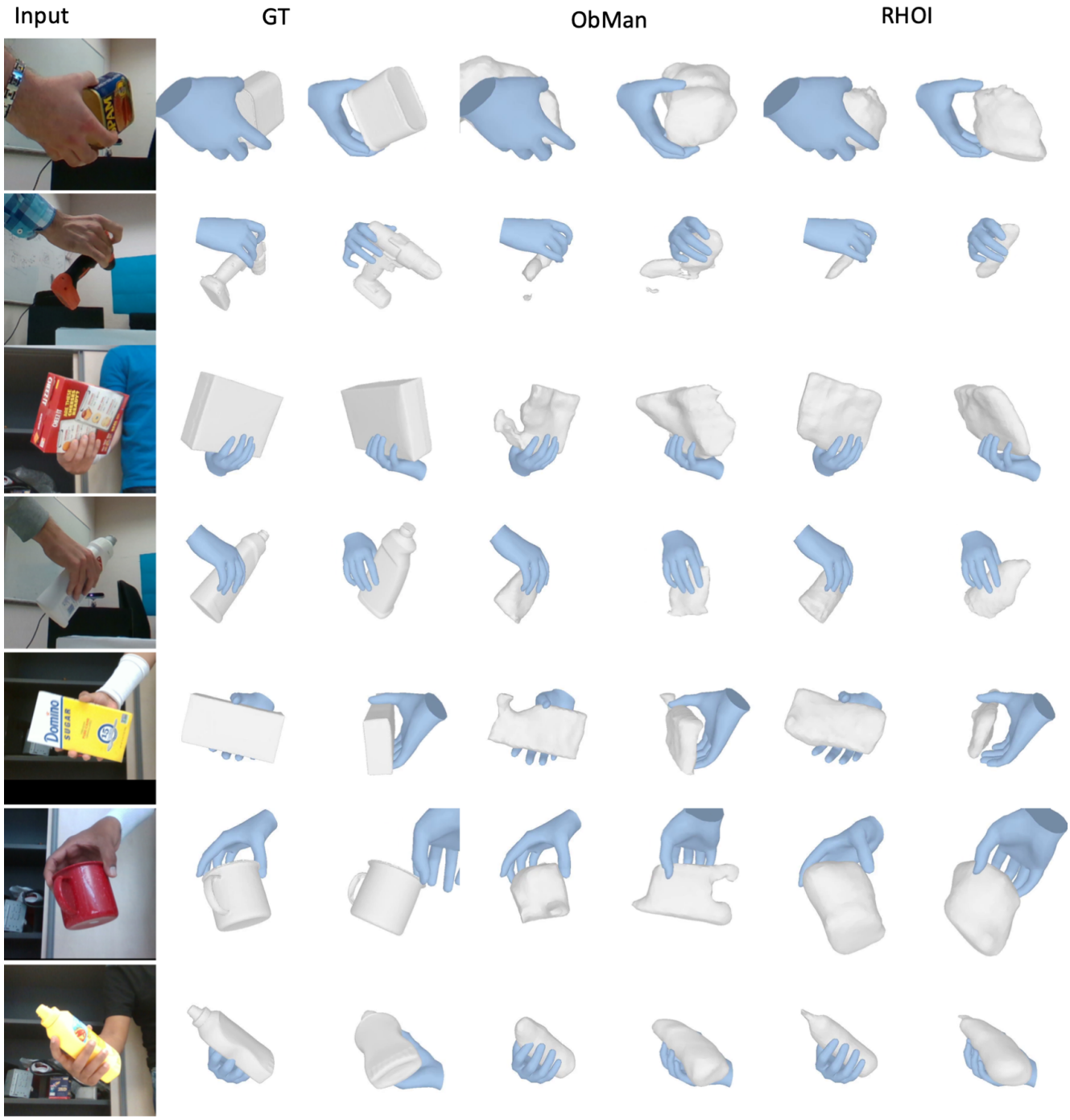

Figure 5. **Cross-dataset generalization:** we show quantitative results on HO3D for models pretrained on ObMan and RHOI.

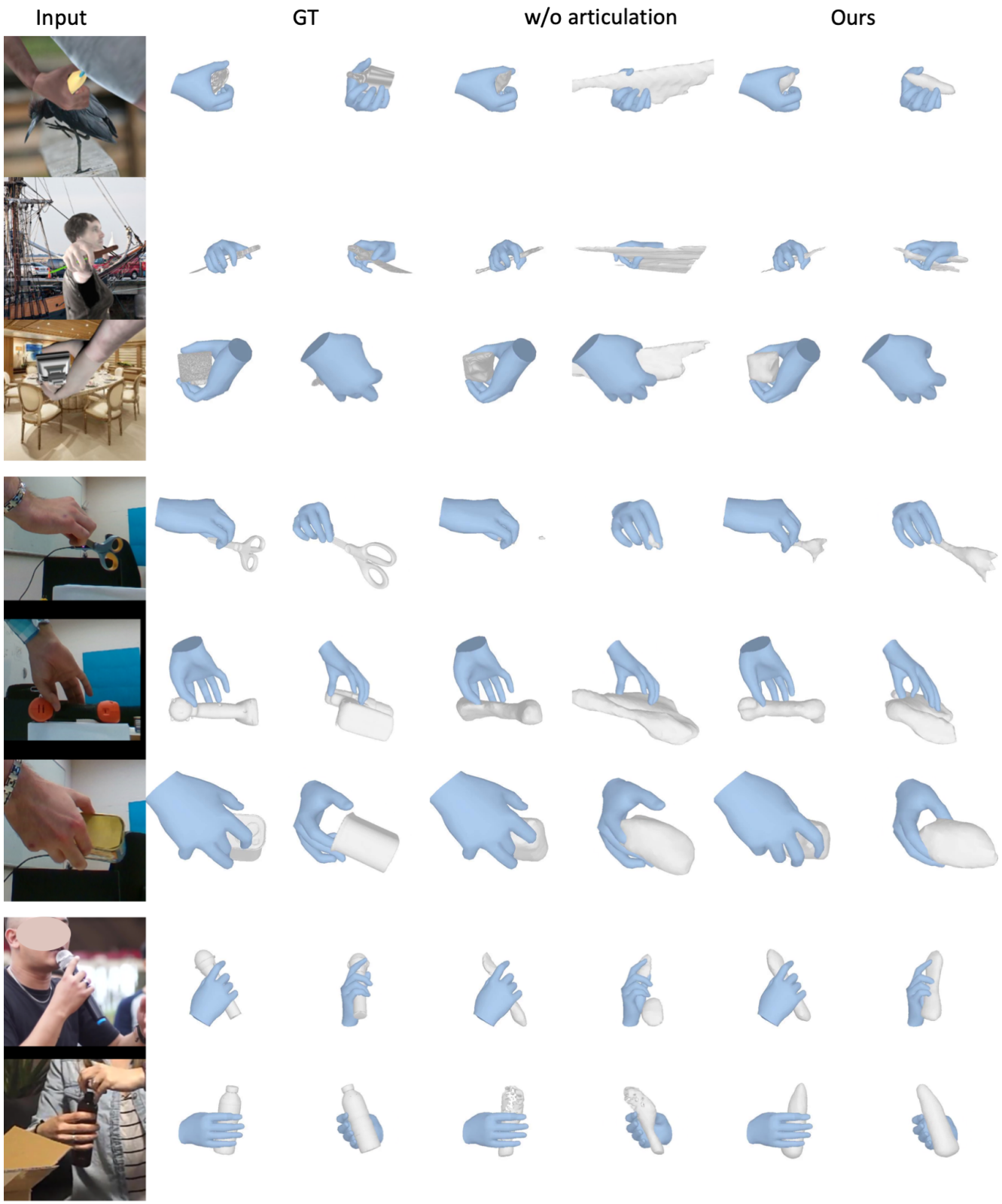

Figure 6. Visualizing reconstruction of hand-held object with or without explicitly considering hand pose on ObMan, HO3D and RHOI.

Figure 7. Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.