

Reconstructing In-the-Wild Open-Vocabulary Human-Object Interactions

Supplementary Material

Overview

The contents of this supplementary material are:

- Sec. 1: Characteristics of Open3DHOI.
- Sec. 2: Method Details.
- Sec. 2.2.3: Additional Experiments.

1. Characteristics of Open3DHOI

1.1. Image Selection for Open3DHOI

Considering the complexity and difficulty of the 3D HOI annotation process, we only select images with single-person annotation from the existing 2D HOI dataset, HAKE, and SWIG-HOI. There are 63 images in our final dataset that have multiple objects interacting with one person. For these images, we split the annotation to keep one image having one HOI pair.

Interaction. Notice that we have 3,671 interactions, more than our image number, 2,561, because one person can interact with an object with multiple actions, like drinking with and holding a bottle at the same time. Fig. 9 shows the co-occurrence between the major object categories and actions, and Tab. 1 shows the object list in our Open3DHOI dataset.

Object size. The object size in our dataset varies significantly across different categories, and even within the same category, there is also a variation in size. In Fig. 2, we chose object categories with more than 30 images and draw the size distribution in each category. We use the volume function from Trimesh to compute the volume of each object mesh, then take the cube root to obtain the size. We can see that object like elephants has larger sizes and bottles has smaller sizes. What's more, for objects like wine glasses, the size variation within the category is minimal, while for objects like couches, the variation is much larger. Fig. 3 shows the size distribution of all images.

Abnormal HOI. Because our dataset is created from 2D HOI datasets, which have many abnormal HOIs like standing on a chair, our dataset also contains many abnormal interactions. Fig. 4 shows some cases of our abnormal HOIs.

1.2. Contact Annotation

In our manual annotation process, we annotate the contact regions for images with qualified reconstruction. We split the human SMPL-X body into 34 parts and counted the number of annotations for each body part in Tab. 2. In Fig. 5, we show body parts on SMPL-X mesh and annotation heat map. It can be observed that interactions involving the hands, feet, and legs occur more frequently than those

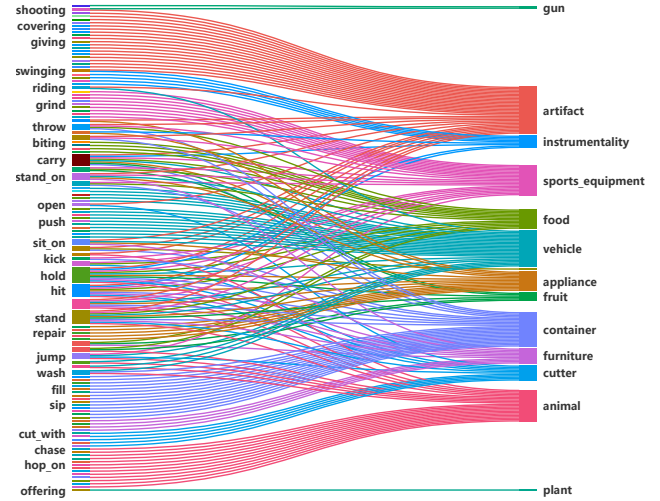


Figure 1. Co-occurrence between major object category and actions in Open3DHOI.

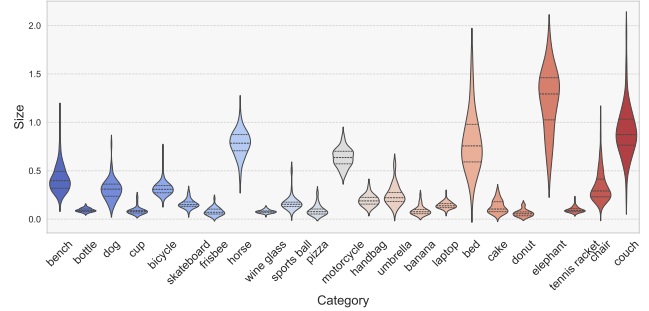


Figure 2. Object size distribution in different object categories.

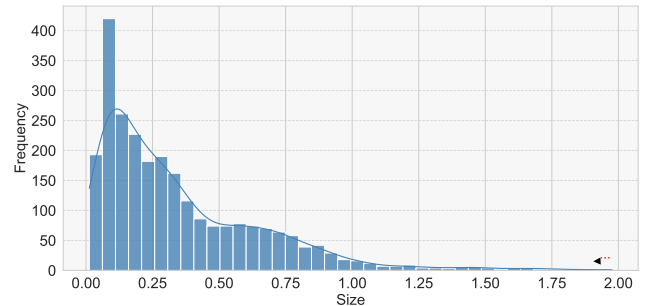


Figure 3. Object size distribution of all images.

involving other body regions.

Table 1. Object categories in our Open3DHOI dataset.

Object Id	Object Class	Object Id	Object Class	Object Id	Object Class	Object Id	Object Class
0	bird	1	television	2	surfboard	3	dining table
4	mug	5	bench	6	goat	7	Gallus gallus
8	fish	9	eggs	10	torch	11	rose
12	award	13	guitar	14	pistol	15	ashcan
16	baseball glove	17	bowl	18	shovel	19	bottle
20	cookie	21	piano	22	home plate	23	furniture
24	barrow	25	dog	26	boot	27	pot
28	handcart	29	cell phone	30	donkey	31	hair drier
32	basket	33	airplane	34	chain	35	oven
36	box	37	cup	38	truck	39	bicycle
40	snowboard	41	bucket	42	cat	43	pump
44	hammock	45	skateboard	46	stone	47	sniper rifle
48	cattle	49	tiger	50	power drill	51	mouse
52	frisbee	53	helmet	54	violin	55	hobby
56	car	57	book	58	horse	59	camel
60	fire hydrant	61	backpack	62	backhoe	63	wine glass
64	sports ball	65	clock	66	scissors	67	pizza
68	raft	69	motorcycle	70	hammer	71	loaf of bread
72	handbag	73	teddy bear	74	suitcase	75	vacuum cleaner
76	pitcher	77	tie	78	vase	79	keyboard
80	pumpkin	81	ice cream	82	boat	83	kite
84	tarpaulin	85	umbrella	86	dinghy	87	package
88	coffee cup	89	banana	90	laptop	91	knife
92	mortar	93	hot dog	94	hairbrush	95	bed
96	float	97	spoon	98	cow	99	cake
100	sandwich	101	pen	102	bouquet	103	hoe
104	jeep	105	lion	106	donut	107	apple
108	whip	109	toilet	110	elephant	111	wrench
112	tennis racket	113	liquor	114	hand glass	115	tricycle
116	remote	117	bullet	118	pipe	119	baggage
120	toothbrush	121	skis	122	chair	123	couch
124	sculpture	125	fork	126	air cushion	127	light bulb
128	sheep	129	pottery	130	carrot	131	barrel
132	fire extinguisher						



Figure 4. Abnormal HOIs in Open3DHOI.

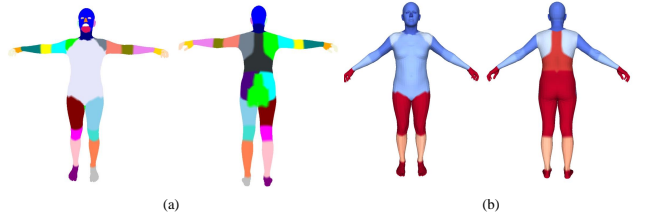


Figure 5. Body parts and annotation heat map.

2. Method Details

2.1. Coarse Reconstruction

In paper Fig.2, we introduce the process of coarse reconstruction. In this section, we provide additional details

Table 2. Body part name and annotation number.

Body Part	Number	Body Part	Number
bottom	974	head	31
left elbow	55	left foot	335
left palm	689	left hip	435
left knee	383	left lower arm	87
left lower leg	27	left shoulder	112
left upper arm	28	left upper leg	801
neck	15	right elbow	58
right foot	332	right palm	849
right hip	417	right knee	361
right lower arm	81	right lower leg	195
right shoulder	118	right upper arm	37
right upper leg	781	torso	76
left eye	1	right eye	1
left fingers	866	right fingers	1065
left ear	1	right ear	0
jaw	22	nose	0
mouse	32	back	270

about this process. After reconstructing human and object meshes, we use depth to initialize coarse spatial alignment. We use Zoedepth to estimate depth information for each image and convert the depth to a point cloud S . We use an object mask to segment points of objects and place the object mesh to the point cloud center as Obj_{init} . Next, we use Algorithm 1 to align the object mesh with the human mesh.

2.2. Annotation Tools

2.2.1. Filtering Tool

Fig. 7 (a) shows our filtering tool. First, we judge whether human reconstruction is qualified using the rendered image. There are two buttons, “Delete” and “Pass”, if human reconstruction is bad, we click on the “Delete” button to delete this image otherwise we click on the “Pass” button and go to the next procedure to judge object reconstruction quality. According to the six-view rendering, we choose to keep the image and not. If the reconstruction is bad because the mask completion doesn’t work well, we will ask the volunteer to correct the mask in the last column using a mouse brush. If the mask completion is not bad but the reconstruction is still terrible, or if the occlusion is too serious to reconstruct, we choose to click on the “Delete” button to delete this image. If the volunteer clicks on the “Pass” button for both human and object, then he needs to click on the “Open App” button on the bottom to open the contact annotation app in Fig. 8. Each body part in the app is clickable for volunteers to choose the contact part. After selection, the volunteer needs to go back to the main page and save the final annotation result. Fig. 6 shows cases with bad masks

Algorithm 1 Align object mesh with human mesh.

Input: Points cloud of scene S , 3D human points H_{3D} , camera intrinsic parameter K_h , 3D object model Obj_{init}

Output: 3D points of objects Obj_{3D}

1. $H_{proj} \leftarrow$ project H_{3D} by $K_h = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix}$
 $x_h^{proj} \leftarrow \frac{\bar{z}c_x}{f}, y_h^{proj} \leftarrow \frac{\bar{z}c_y}{f}$
 2. $Index_h \leftarrow$ compare image and H_{proj} to obtain the indices of S corresponding to H_{3D}
 $H_{3D} \leftarrow H_{3D}[\text{argsort}(H_{3D}[:, 2])/2]$ get half of human points by depth
 $S_h \leftarrow$ extract points belong to human in S by $Index_h$
 3. $Scale, Translation \leftarrow$ compare S_h and H_{3D} to get 3D transformation
 $scale \leftarrow$ using $\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \|p_i - p_j\|_2$ to get scale of S_h and H_{3D} , and scale is $s_{H_{3D}}/s_{S_h}$
 $translation \leftarrow \text{mean}(S_h) - \text{mean}(H_{3D} * scale)$
 4. $Obj_{3D} \leftarrow$ operate Obj_{init} by $Scale * Obj_{init} + Translation$
- return** Obj_{3D}

and with good masks but bad reconstructions.

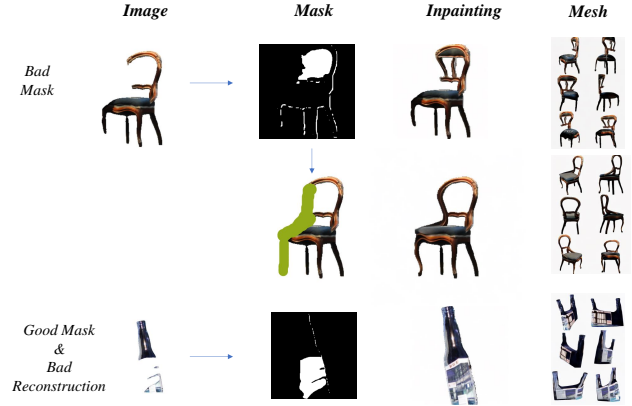


Figure 6. Special cases in filtering process.

2.2.2. 3D Interaction Tool

Blender Annotation Tool. When we have filtered human and object meshes, then we use the coarse reconstruction method in Sec. 2.1 to initialize 3D HOI. We designed a blender add-on for 3D HOI annotation. There are three buttons on the top, “Load Meshes”, “Export Object Pose and Location” and “Save Delete and Load Next”. The first is to load human and object meshes and image references. Volunteers need to adjust the objects’ positions, rotations, and

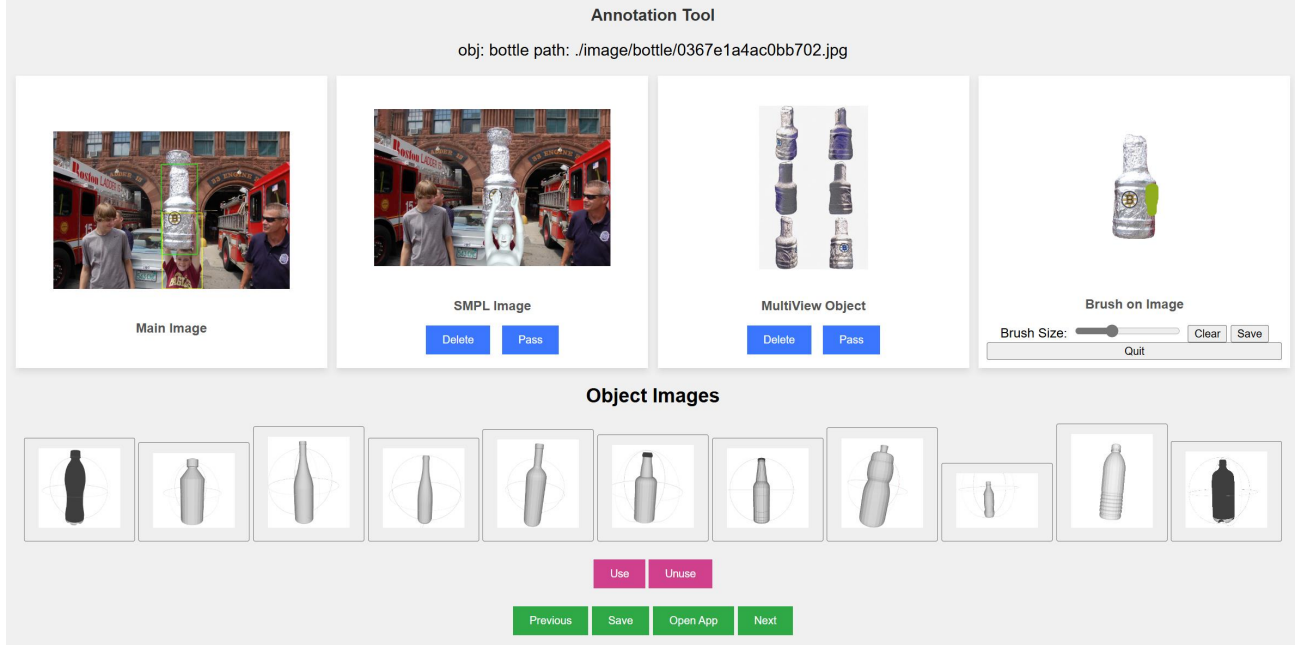


Figure 7. Filtering tool.

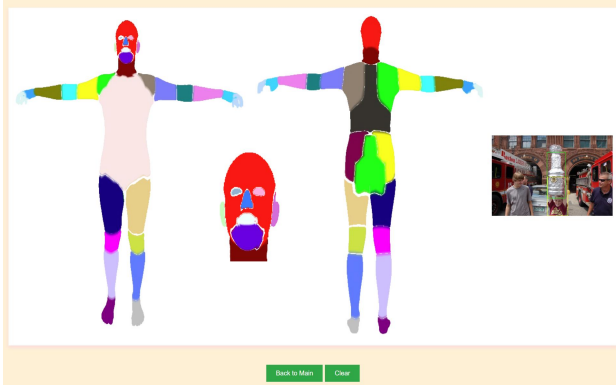


Figure 8. Contact part annotation tool.

scales using a mouse, while the human is fixed. After annotating, volunteers can use the second button to save the result and load the next image, or choose to use the third button to delete this image if it is hard to annotate.

Fine Annotation Tool. During the annotation process in Blender, the images were used as references without precise alignment. Although we ensured reasonable 3D interaction during the Blender annotation process, some objects’ poses still exhibit discrepancies compared to the images. Fig. 10 shows our 3D fine annotation tool based on ImageNet3D [1], to optimize the results from previous annotation. We select 581 images with IoU between human-object projection and mask lower than 0.5 and project a line set of

meshes on the image. To ensure 3D interaction accuracy, we also project the meshes from three novel views. Volunteers need to click on the buttons to move, rescale, and rotate the object until it is aligned with the image.

2.2.3. Dataset quality

1) Human-object penetration rate: we tested the penetration metric following[?] by adding human-in-object penetration and object-in-human penetration together, which is 3.26 while PHOSA is 4.26. Notice that only considering penetration is not fair because in some cases where objects and humans are far away from each other also have zero penetration. Since we annotated human contact parts, so we also tested the distance between the annotated contact part and the object divided by object size, the score of GT after normalization is 0.058, and PHOSA is 0.326. **2) Human and object projection error:** the human projection IoU is 0.621, the object projection error is 0.384, and the H+O projection error is 0.634. Notice that there is a significant occlusion of objects and humans in wild images, especially for objects, so this score can only serve as a reference. **3) Reconstruction quality:** since there is no GT object in our dataset, it is difficult to evaluate the quality of object reconstruction using traditional metrics. We use the inpainted GT object images and the projections of the annotated object mesh to compute SSIM and LPIPS for evaluation. Due to discrepancies between the object pose and the GT image, as well as the inherent differences between real images and mesh projections, including lighting, noise, etc.,

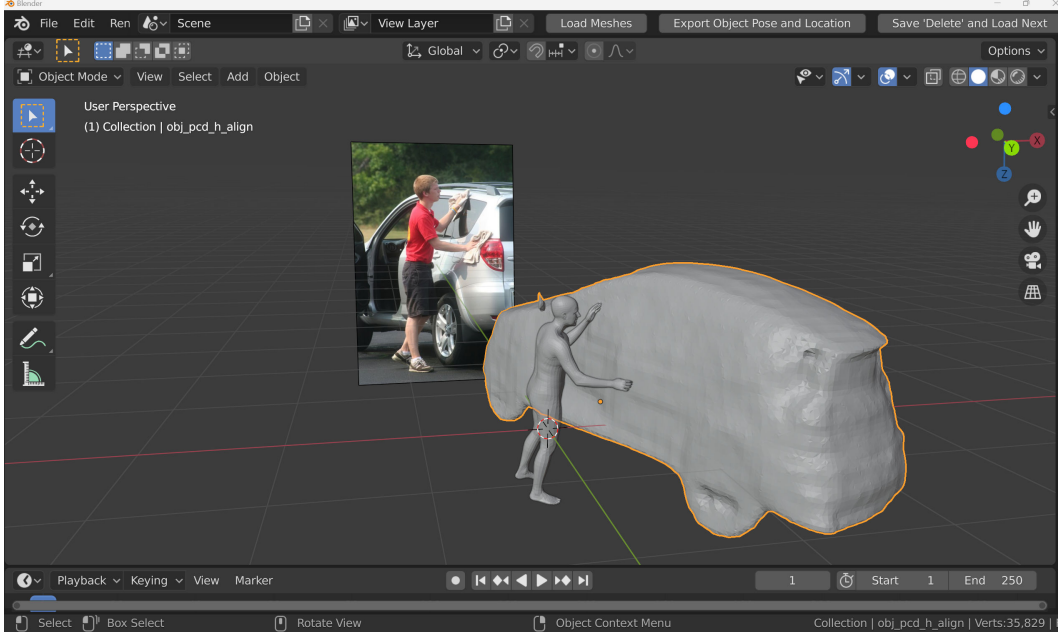


Figure 9

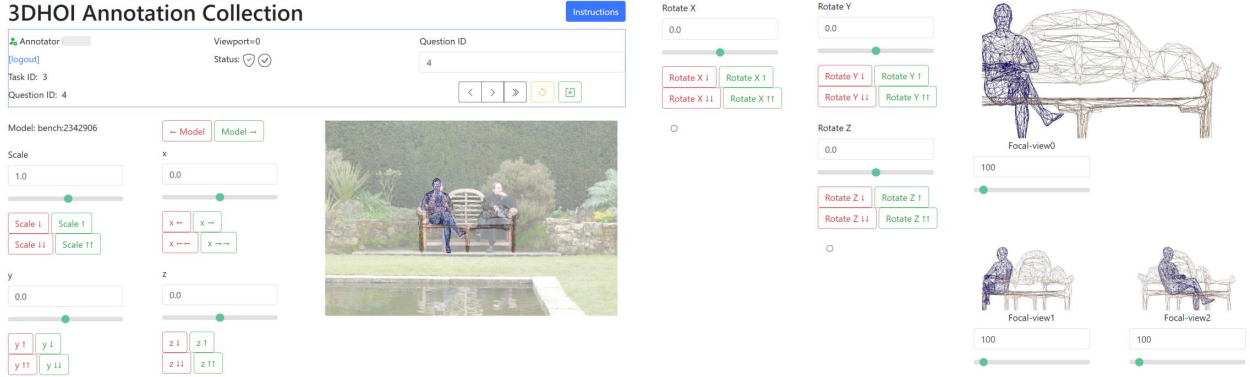


Figure 10. 3D fine annotation tool.

this evaluation is not entirely fair. However, our reconstruction scores still reached an LPIPS of 0.714 and an SSIM of 0.294, demonstrating that the quality of our reconstruction is high.

2.2.4. Discussion

Throughout the whole annotation process, we collected 2.5k+ images from 15k source images, resulting in a pass rate of 17%, which indicates that most 2D HOI images are hard to reconstruct 3D representations. In the future, the filtering process can be accelerated by training a model to judge the reconstruction result, and volunteers only need to filter based on the predictions. Our annotation process has provided enough data to train a judge model.

At the bottom of our filtering app, there are many object

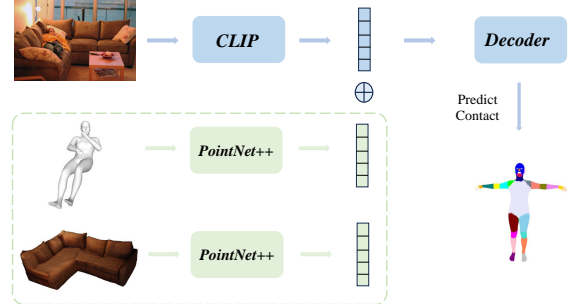


Figure 11. Our contact evaluation model.

template buttons, which are designed to assign corresponding templates for images that closely match the template

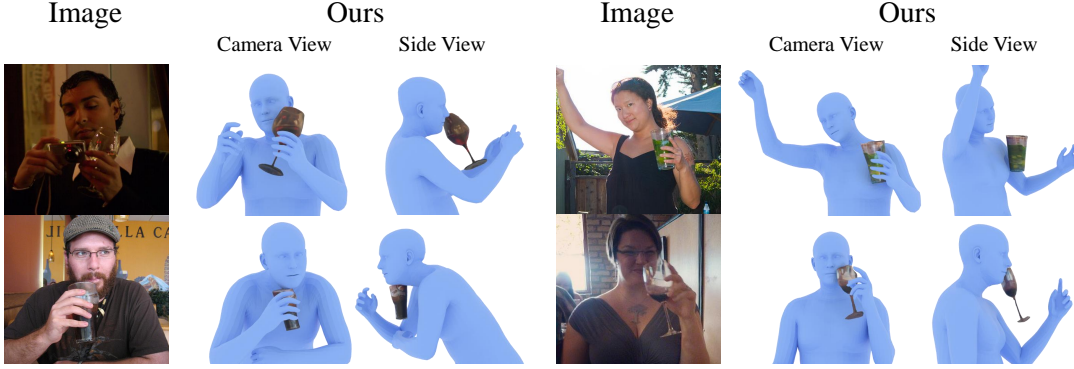


Figure 12. Failure cases of HOI-Gaussian.

Table 3. Results of our contact evaluation task.

Methods	Micro F1-Score \uparrow	Hamming Loss \downarrow	Jaccard Index \uparrow
2D	0.6118	0.0874	0.4303
2D&3D	0.6207	0.0844	0.4561

but have poor reconstruction quality. We build a template library for **58** object categories and totally **212** templates. Although we didn’t use this library to build our Open3DHOI dataset, it is still very useful for future work.

After our 3D fine annotation process, the IoU between human-object projection and mask increased from 0.48 to 0.57, and the IoU between object mesh projection and object mask increased from 0.32 to 0.48, which indicates that our fine annotation tool indeed improved the pose alignment.

2.3. LLM Task Setting

2.3.1. PointLLM

We used PointLLM-7B as a test model, and input our annotated human and object mesh vertices. Object vertices have colors and human vertices are colored black. When asking, we will tell PointLLM that “The point cloud is a person interacting with an object. The person is black.” first and then asks specific questions. To decrease the difficulty, we ask PointLM to generate a description first and use Qwen2.5 [2] to extract the exact word from our action and object list.

2.3.2. ChatPose

In Sec.6.2, we state that we select images with multiple images. Although our dataset only contains single-person annotation, there are still many images with more than one person, we used Detectron2 [3] to detect these images for our testing. Our task is to ask ChatPose to locate the specific person interacting with the specific object according to its understanding of the interaction in the image. The pose it answered has no root pose and location, so we compare the prediction with GT using the same root pose, zero pose at zero location. The metrics we used are MPJPE (Mean

Per Joint Position Error) and MPVPE (Mean Per Vertex Position Error), which are common metrics in human pose estimation.

3. Additional Experiments

3.1. Contact Evaluation

Since our dataset contains contact annotations, we want to evaluate whether 3D information would be conducive to estimating contact regions compared to image only. Therefore, we design a simple pipeline to estimate the contact regions. As Fig. 11 shows, we use clip-ViT-B/32 to encode image and pointnet++ to encode normalized human point clouds and object point clouds respectively. Image features and point clouds features of human and object are fused and put into an MLP decoder. We treat this problem as a multi-label classification task and use Micro F1 Score, Hamming Loss and Jaccard Index to evaluate the accuracy. The Micro F1 Score calculates precision and recall globally across all labels. Hamming Loss measures the fraction of incorrect label predictions over the total number of labels. Jaccard Index evaluates the similarity between the predicted and true label sets for each sample. Our current implementation simply concatenates 3D and 2D features and is trained on only 2,000 samples. However, our results over multiple metrics in Tab. 3 still indicate that 3D information is beneficial to the estimation of contact regions.

3.2. Failure Cases

Fig. 12 shows some failure cases of our HOI-Gaussian optimizer. In these cases, human body parts occlude each other severely, and the object happens to be located between the occluded areas, which becomes challenging to determine which body part the object should contact with.

3.3. More Results

Fig. 13 shows more results comparison between GT, PHOSA, and Ours.

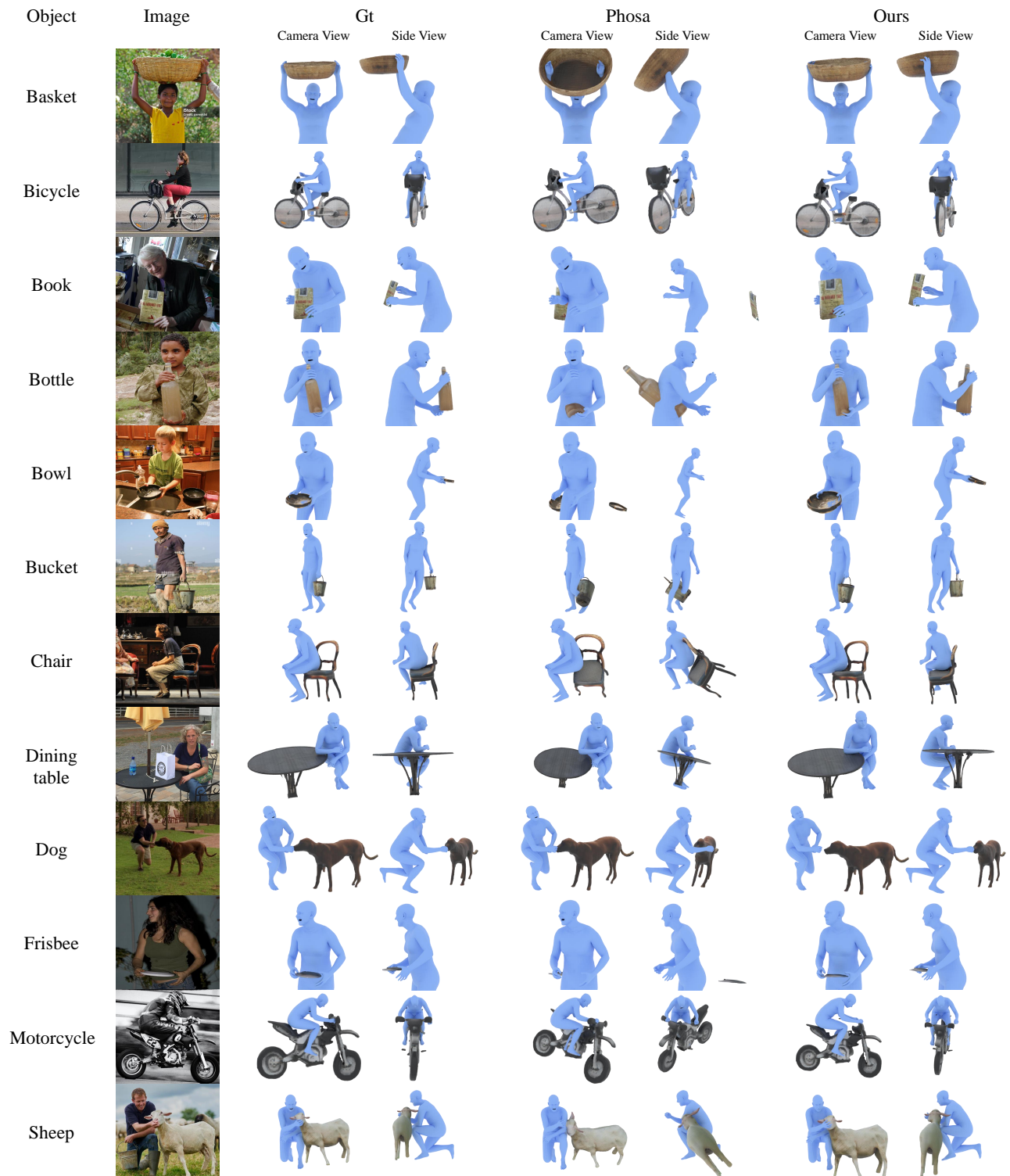


Figure 13. More results.

References

- [1] Wufei Ma, Guanning Zeng, Guofeng Zhang, Qihao Liu, Letian Zhang, Adam Kortylewski, Yaoyao Liu, and Alan

Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *arXiv preprint arXiv:2406.09613*, 2024. 4

- [2] Qwen Team. Qwen2.5: A party of foundation models, 2024. 6
- [3] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6