Research Article

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

Boran Wen[1,2], Dingbang Huang[1], Zichen Zhang[1], Jiahong Zhou[1], Jianbin Deng[1], Jingyu Gong[3], Yulong Chen[1], Lizhuang Ma[1], Yong-Lu Li[1,2]

1. Shanghai Jiao Tong University, Shanghai, China; 2. Shanghai Innovation Institute, China; 3. East China Normal University, China

Reconstructing human-object interactions (HOI) from single images is fundamental in computer vision. Existing methods are primarily trained and tested on indoor scenes due to the lack of 3D data, particularly constrained by the object variety, making it challenging to generalize to real-world scenes with a wide range of objects. The limitations of previous 3D HOI datasets were primarily due to the difficulty in acquiring 3D object assets. However, with the development of 3D reconstruction from single images, recently it has become possible to reconstruct various objects from 2D HOI images. We therefore propose a pipeline for annotating fine-grained 3D humans, objects, and their interactions from single images. We annotated 2.5k+ 3D HOI assets from existing 2D HOI datasets and built the first open-vocabulary in-the-wild 3D HOI dataset Open3DHOI, to serve as a future test set. Moreover, we design a novel Gaussian-HOI optimizer, which efficiently reconstructs the spatial interactions between humans and objects while learning the contact regions. Besides the 3D HOI reconstruction, we also propose several new tasks for 3D HOI understanding to pave the way for future work. Data and code will be publicly available at https://wenboran2002.github.io/3dhoi/.

1. Introduction

Human-Object Interaction (HOI) is an important area in action understanding, with numerous datasets and methods proposed. In the 2D HOI domain, large-scale image datasets such as HICO-DET[1] and HAKE[2] have been introduced. For 3D HOI, datasets like BEHAVE[3] and InterCap[4] have also been proposed to study human interactions with objects in 3D. Despite achieving promising results in open-vocabulary and in-the-wild scenarios within 2D HOI, the 3D HOI field faces challenges in generalizing existing methods to real-world images due to dataset limitations and the lack of 3D open-world HOI data.

 doi.org/10.32388/K7APQ4

In 3D HOI, many datasets have been introduced. Datasets such as BEHAVE[3], InterCap[4], ImHOI[5], and PROX-S[6] provide multi-view RGBD sequences and 3D annotations in indoor scenes. Though datasets like WildHOI and 3DIR[7][8] are constructed from in-the-wild images, they contain limited object categories and unreal CAD objects. To better understand 3D HOIs and apply them to the real world, we need to collect more realistic and diverse data on interactions with objects. Thus, in this work, we propose a novel 3D HOI annotation method for real-world images of any objects and interactions.

In detail, we built our annotation pipeline on two bases: 1) Existing 2D HOI datasets provide rich 2D annotations, including bounding boxes, and a wide variety of objects and actions. This diversity creates the potential for reconstructing 3D assets from 2D HOI images. 2) The development of existing image-based 3D object/human reconstruction techniques. We selected images with contact interactions from existing 2D HOI datasets, *e.g.*, HAKE[2] and SWIG-HOI[9]. Next, we used InstantMesh[10] and OSX[11] to reconstruct the objects and the human body respectively, and designed an algorithm for automatically reconstructing *rough* 3D interactions. Furthermore, we developed two annotation tools: one for filtering the reconstruction quality and the other for annotating 3D spatial positions. We manually annotated over **2.5k+** images to create an open-vocabulary, in-the-wild HOI dataset, as the test set for future 3D HOI studies and designed tasks and metrics to evaluate their performance. It consists of **370** 3D human-object pairs, **2,561** objects in **133**categories, and **3,671** interactions in **120** categories.

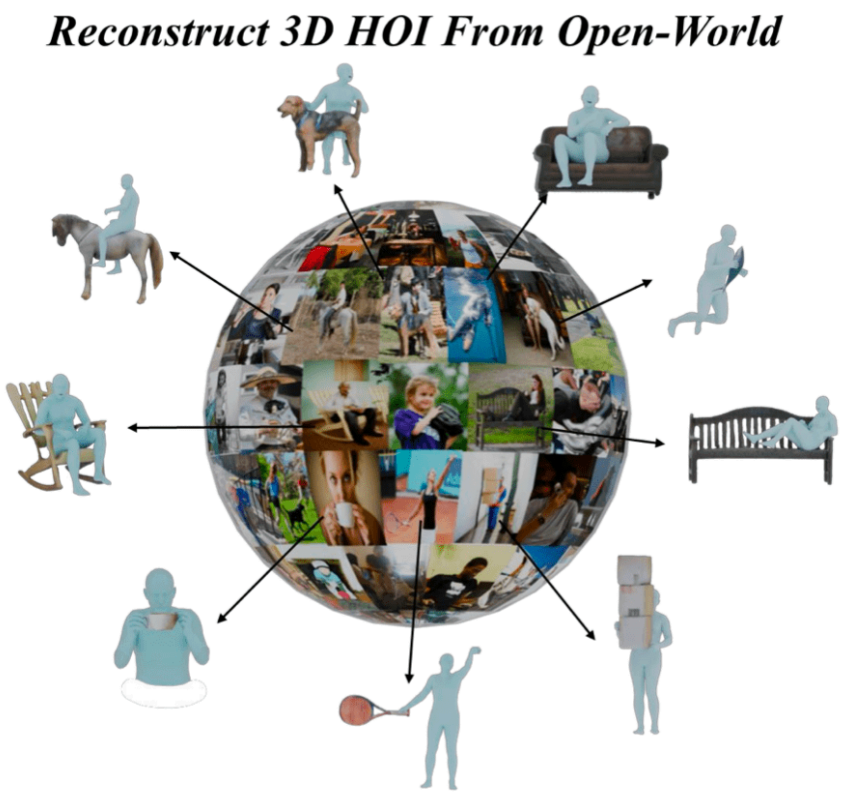

Figure 1. We aim to reconstruct 3D HOIs from arbitrary open-world images. We propose a pipeline for annotating fine-grained reconstructions to build a dataset. Additionally, we introduce a new optimizer suitable for reconstructing arbitrary objects.

Given the new dataset, we also proposed a *training-free* algorithm for reconstructing 3D HOIs from monocular images. On one hand, previous training-free algorithms do not require specific object categories or templates, but their optimization performance is generally limited, and they rely on manual annotations. On the other hand, previous training-based methods perform well for specific object categories and scenes but struggle to generalize to open-world environments. To this end, we leveraged the 3D Gaussian Splatting model to propose a novel Gaussian-HOI optimizer to improve the reconstruction quality. It utilizes Gaussian rendering capabilities to ensure that the reconstructed 3D assets are aligned with the image from the target view and takes advantage of the opacity attribute of Gaussians to identify contact regions, which makes the optimization of 3D interaction relationships more effective.

doi.org/10.32388/K7APQ4

Overall, our contributions are: 1) We utilized SOTA 3D reconstruction tools to develop a 3D HOI annotation method. 2) We built a new and extensive 3D HOI dataset Open3DHOI consisting of 2.5k+ images with rich 2D and 3D annotations. 3) We designed a 3D HOI optimizer based on 3D Gaussian Splatting to reconstruct the spatial interactions between humans and objects from single images.

2. Related Works

2.1. HOI Benchmarks

The development of 2D HOI benchmarks[2][12][13][9][14][15][16][17][18] has made our 3D HOI reconstruction approach possible. Datasets like HICO-DET[12] and HAKE[2] provide annotations for 80 object categories and 117 action categories. Additionally, open-vocabulary datasets such as SWIG-HOI[9] include annotations for over 1,000 object categories.

In contrast, 3D HOI datasets contain significantly fewer action and object categories, and most are recorded in fixed indoor environments. BEHAVE[3], as the earliest one, introduced a method for obtaining accurate 3D HOI annotations from multi-view videos, providing interaction data for 20 common objects. InterCap[4] further built upon it by offering more detailed hand interaction information. Recently, several benchmarks[8][7] for reconstructing 3D HOIs from real-world images have been proposed. However, they have notable limitations. Their object categories are focused on a few common types—such as balls, skateboards, and bicycles—and the number of instances is limited, with all objects derived from fixed 3D CAD models.

2.2. 3D Reconstruction

3D reconstruction has seen rapid advancements recently, both in humans and objects. After SMPL[19], parametric human body modeling has rapidly evolved. Currently, the SMPL-X model[20], which includes detailed hand and facial expression modeling, is widely used in the field[11][21][22][23]. We utilized a state-of-the-art one-stage model[11] for our human body reconstruction. What's more, image-to-3D has emerged as a rapidly advancing area in 3D vision recently. From SDS loss optimization methods[24][25] that leverage 2D diffusion priors to Multi-view Diffusion Models[26][27][28][29] and Large Reconstruction Models[30][31][10] based on large-scale data, the quality and efficiency of 3D generation from a single image have seen significant improvements. To generate our 3D HOI dataset, we need to

reconstruct a large volume of image data, requiring a *balance* between the quality and efficiency of existing 3D generation methods. Ultimately, we selected InstantMesh[10] as our reconstruction model.

2.3. 3D HOI Reconstruction

Reconstructing 3D HOIs from a single image[16][32] is a challenging task and important for many applications[33][34][35]. It requires maintaining consistency between the spatial positions of the human and the object within the image in the given camera view while ensuring that the spatial interactions are realistic and coherent. Kanazawa *et al.*[36] optimizes spatial interactions through predefined contact pairs, while Wang *et al.*[37] leverages GPT-3's prior knowledge to optimize spatial interactions. Xie *et al.*[38] learnt HOI spatial arrangement priors from the BEHAVE dataset. Wang *et al.*[7] learnt the prior distribution of the 2D human-object keypoint layout and viewports to tune the relative pose between the 3D human and object.

3. 3D HOI Annotation

In this section, we introduce our new pipeline for 3D HOI annotation from single-view images.

3.1. Coarse Reconstruction Annotation

First, we used a state-of-the-art human pose estimation method[11] to obtain the 3D representation of the human body and employed image-to-3D technique[10] to generate the 3D representation of the object.

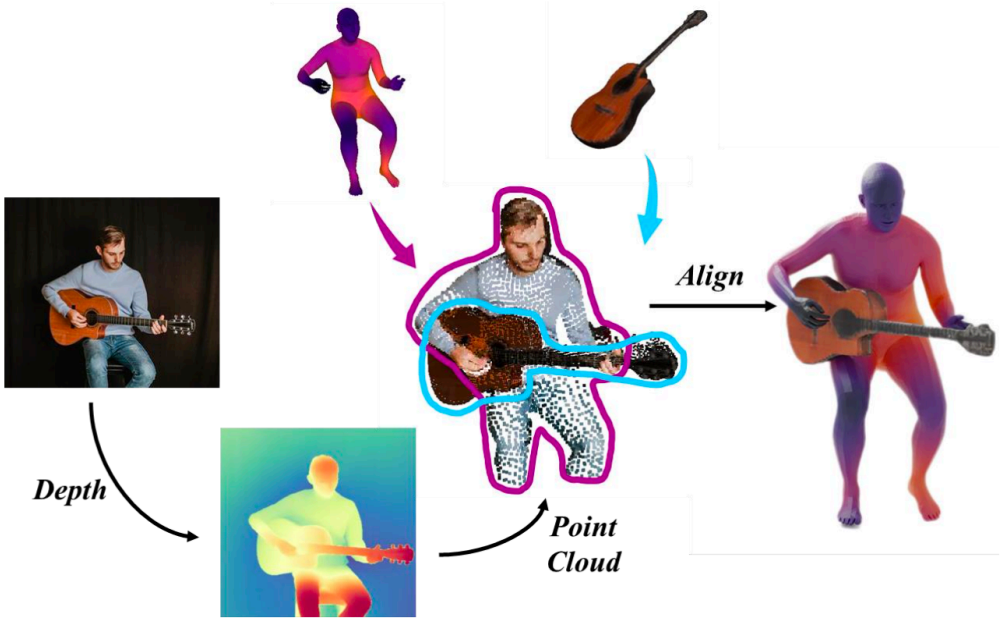

Figure 2. Coarse Reconstruction. We first obtain depth from the images and generate point clouds. Given masks, we extract the corresponding point clouds for the person (pink) and object (blue). We obtain a rough reconstruction by matching the MESH vertices of the person and the object with the depth point cloud.

Second, in-the-wild HOI images often involve significant occlusions between humans and objects and between different objects. To address the critical occlusion problem in object 3D reconstruction, we employed occlusion completion[39] and used Stable Diffusion 1.5[40] as the inpainting tool to obtain complete object images.

Finally, we applied our projection algorithm to estimate the rough spatial relationship, including the positions and sizes of the human and object. We used monocular image depth estimation[41] combined with human and object masks, to generate depth point clouds for both the human and the object. Then, we obtained a rough estimation of their spatial positions and size by matching the sampled point clouds from the human and object meshes with the depth point clouds as shown in Fig. 2.

The rough reconstruction can facilitate the subsequent manual annotation. Rough reconstruction usually has serious mesh collision issues and inaccurate scales and positions. Additionally, the object pose provided by InstantMesh[10] is also not accurate. Therefore, to obtain more precise 3D human-object interaction information, we manually performed further annotations.

3.2. Fine Reconstruction Annotation

We designed two annotation tools to facilitate manual annotating for obtaining more refined 3D HOI information. Fig. 3 shows the whole process.

Filtering Tool filters the initial reconstructions and annotates contact regions. The filtering consists of:

1. **Filtering the SMPL-X human reconstruction.** We project the reconstructed SMPL-X mesh onto the image to assess whether the pose estimation of key interaction joints is accurate. For joints not involved in the interaction, we relax the criteria. For example, if a person is drinking water and the target object is a cup, we focus on the accuracy of the hand pose reconstruction, while only ensuring that the lower body pose is reasonable.
2. **Filtering the object reconstruction.** Given the reconstructed object mesh, we project it from six viewpoints. Annotators evaluate the quality of the object mesh, retaining images with high-quality reconstructions, especially paying attention to the quality of the interaction area.
3. **Manual optimization of the inpainting mask.** The inpainting masks obtained from occlusion completion fail sometimes. Annotators can manually correct them using a brush, ensuring higher accuracy. In Fig. 3, (a) shows that the annotator manually drew a mask (green part) on the poorly reconstructed couch. After re-inpainting and reconstruction, it achieves a much better result. Considering that some objects may only partially appear in the image, we still allow for the completion of the object by manually editing the mask. Annotators can use a brush to fill in the parts of the object that are outside the image. We further divided the SMPL-X mesh into 34 human body regions, which are used for contact area annotations. Annotating contact areas at a fine-grained, vertex level would be costly and difficult to ensure high-quality results. Therefore, we opted for **part-level** contact annotations, which are sufficient for most interaction scenarios. In Fig. 3, (a) shows the annotation of the contact area, where the person is sitting on the couch, and the thighs and bottom are annotated (blue part).

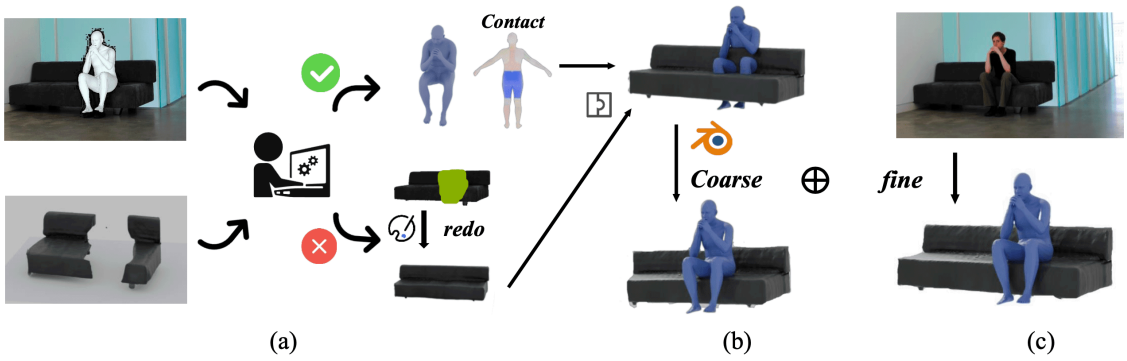

Figure 3. Annotation Pipeline. (a) Filtering. Given the reconstructed human and object meshes, annotators assess the quality. If the human reconstruction is eligible, the contact area is further annotated. If the object reconstruction fails, the mask is redrawn manually and the reconstruction is performed again. (b) Given the 3D human interaction through coarse reconstruction, we adjust the object position in Blender. For example, the rough annotation of the couch and the human body shows a mesh collision. We move the object to make sure the person is correctly seated on the couch. (c) We use a fine annotation tool to further align the annotated human and object with the image.

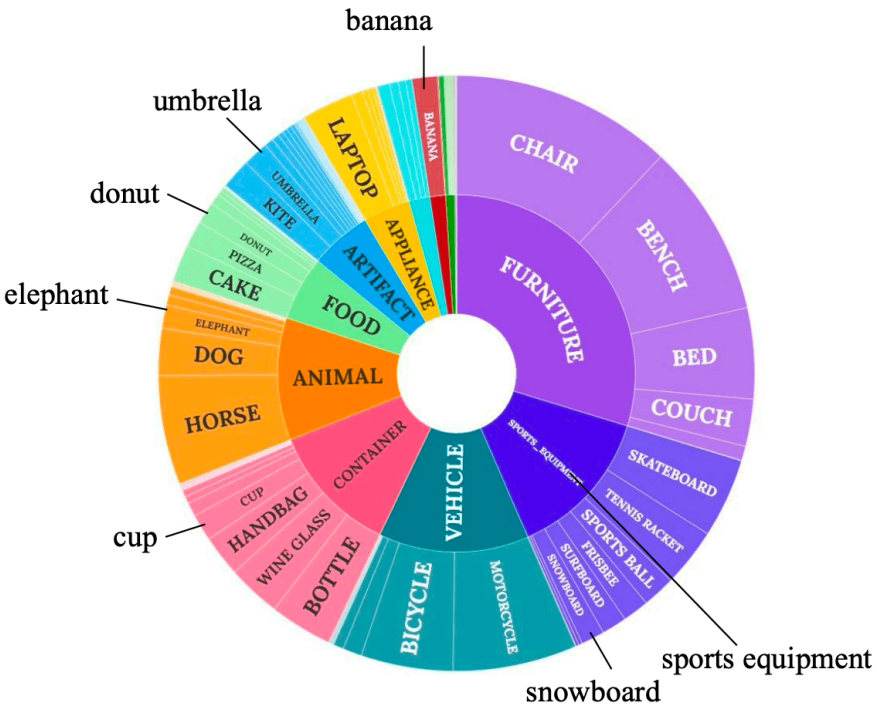

Figure 4. Object category distribution in Open3DHOI. It encompasses a wide range of object categories.

doi.org/10.32388/K7APQ4

3D Interaction Tool is designed for annotating the spatial interaction between humans and objects. We adopted a **coarse-to-fine** approach. Initially, we developed a 3D HOI annotation tool via Blender, where volunteers can adjust the objects' positions, rotations, and scales using a mouse. In Blender, the annotations cannot be compared pixel-by-pixel with the images, and the focus is more on the 3D interaction quality. Thus, we further performed detailed annotations for cases where there is a significant discrepancy when projected on the images. We built a website annotation tool based on ImageNet3D[42]. In this refine-annotation tool, volunteers can finetune the annotations by clicking buttons to translate, rotate, and scale the objects based on the projection of the object on the image. To ensure the accuracy of the 3D interactions, we provide multiple viewpoints projections of humans and objects, allowing them to more accurately evaluate and adjust the annotations.

3.3. Open3DHOI Dataset Construction

We selected 12k+ images from HAKE-Large and 3k+ images from SWIG-HOI, totaling 15k+ images as the database for our 3D reconstruction. We also collected some images from the website. After manual annotation and filtering, we obtained **2.5k+** images to form our dataset **Open3DHOI**. To our best knowledge, this is the **first in-the-wild, open-vocabulary 3D HOI dataset based on real-world images**.

Datasets	Objects	Action	Human	Object Pose	Contact	2D HOI
BEHAVE[3]	10	N/A	SMPL+H	✔	✔	✘
InterCAP[4]	10	N/A	SMPL-X	✔	✔	✘
WildHOI[7]	8	N/A	SMPL	✔	✘	✘
3DIR[8]	21	17	SMPL+H	✘	✔	✔
PROX-S[6]	40	17	SMPL-X	✔	✔	✘
Ours	**133**	**120**	SMPL-X	✔	✔	✔

Table 1. Dataset comparison between previous datasets and ours.

The 2D annotations include bounding boxes, HOI triplets, object labels, and masks obtained using

SAM[43]. The 3D annotations consist of object meshes, SMPL-X parameters for humans, the 6D poses of both objects and humans in space, as well as human contact regions. Our dataset includes **133** object categories and more than **120** interactions, significantly surpassing the current benchmarks in terms of semantic diversity. We referred to WordNet's classification of object categories and divided our object categories into several major categories, as shown in Fig. 4. It can be seen that our data includes a wide range of object categories, many of which were rarely attempted in previous 3D HOI datasets, such as food and animals.

We compared the differences between existing 3D HOI datasets and our dataset in Tab. 1. It can be seen that the object and action categories in the current benchmarks are much fewer than those in our dataset. At the same time, we also provide more detailed 2D and 3D annotations.

4. Method

4.1. 3D Gaussian Splatting

3D Gaussian Splatting[44] is advanced for rendering and reconstructing scenes by representing objects continuously, in a volumetric manner. Instead of relying on traditional mesh-based models, it leverages a collection of Gaussian kernels, each defined by its mean, covariance, and intensity, to describe the spatial distribution of objects in 3D. These Gaussian splats, which are soft, overlapping volumetric primitives, enable high-quality rendering while maintaining flexibility in representation. For a given pixel x, the depth of each overlapping 3D Gaussian is computed using the viewing transformation W, resulting in a depth-sorted list of Gaussians $\mathcal{N}$. The final color of the pixel is then determined using alpha compositing, expressed as:

$$C = \sum_{n=1}^{|\mathcal{N}|} c_n \alpha'_n \prod_{j=1}^{n-1} (1 - \alpha'_j), \tag{1}$$

where c_n denotes the color associated with the $n-th$ Gaussian. The effective opacity α'_j is got by multiplying the learned opacity α_n by a Gaussian weighting function:

$$\alpha'_n = \alpha_n \exp\left(-\frac{1}{2} (x' - \mu'_n)^\top \Sigma'^{-1}_n (x' - \mu'_n) \right). \tag{2}$$

Here, x' is the pixel's projected coordinate and μ'_n is the projected center of the $n-th$ Gaussian. Recent works[45][46] have applied 3D Gaussian Splatting to high-quality human reconstruction, which not only ensures rendering quality but also significantly improves reconstruction speed. For example, in GauGAN-

based approaches like Gauhuman[45], SMPL's vertices are used as the initial point clouds for 3D Gaussian splatting. This method simultaneously learns the SMPL pose and Linear Blend Skinning (LBS) parameters to optimize the human Gaussians, ultimately achieving high-quality reconstruction results. It successfully demonstrates the effectiveness of 3D Gaussian splatting for optimizing human reconstruction based on SMPL representations. Also, work like GS-pose[47] and 6D-GS[48] leverage 3D Gaussian splatting to optimize the 6D pose of objects. We reasonably infer that this approach can also be effectively applied to reconstruct human-object interactions from a single viewpoint, potentially replacing traditional silhouette-based optimizers[36].

4.2. HOI-Gaussian Optimizer

We developed the HOI-Gaussian optimizer specifically for 3D HOI reconstruction based on Gauhuman as shown in Fig. 5. We chose 3D Gaussian over other silhouette-based optimization methods as we believe it offers the following advantages:

1. Methods like Gauhuman have demonstrated that 3D Gaussian can be used to adjust human body parameters. Our HOI-Gaussian optimizer can simultaneously optimize object and human poses beyond traditional methods.
2. 3D Gaussian uses depth from point clouds better to align with the image, reducing cases where large pose discrepancies occur despite small silhouette losses.
3. We hope to use the features rendered from the 3D Gaussian point clouds to obtain potential contacts, thus reducing reliance on prior, such as manually annotated human-object parts pairs introduced by methods like PHOSA.

We used the vertices from the SMPL-X model to initialize the human 3D Gaussians g_h, and the vertices from the object mesh to initialize the object 3D Gaussians g_o. During optimizing, we follow GauHuman's pose refinement and LBS offset to learn the parameters of the human model, while introducing a learnable parameter W_{obj} to optimize the object's 6D pose. Our final interaction 3D Gaussian g_{hoi} is derived from g_h and g_o through

$$\begin{aligned} p^{hoi} &= p^h \oplus p^o, \quad \Sigma^{hoi} = \Sigma^h \oplus \Sigma^o, \\ p^h &= R^h p^{h'} + t^h, \quad p^o = R^o s^o p^{o'} + t^o, \\ \Sigma^h &= R^h \Sigma^{h'} R^{h\top}, \quad \Sigma^o = R^o \Sigma^{o'} R^{o\top}, \end{aligned} \tag{3}$$

where p represents the 3D position and Σ represents the covariance matrix. R^h and t^h are the rotation matrix and translation vector obtained from the SMPL-X model through pose and LBS parameters. R^o,

s^o, and t^o represent the object's rotation matrix, scale factor, and translation vector, respectively, and all three variables are learnable.

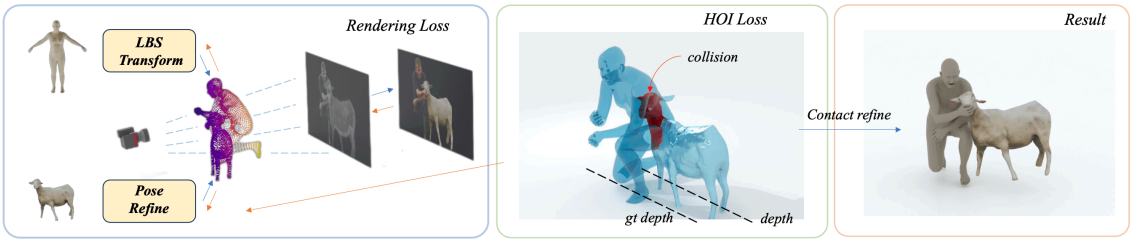

Figure 5. Our pipeline. The optimizer first converts the human and object into 3D Gaussian points, then calculates a rendering loss by comparing the Gaussian-rendered image with the ground truth image. This loss is backpropagated to update the object's pose parameters and the human's LBS parameters. We also calculate an HOI loss, which includes collision, depth and contact losses, the red overlapping areas between the human and object in the image represent collision regions and the dashed lines represent the ground truth depth and the depth during the optimization process. Finally, we refine the result by optimizing the contact regions.

4.3. Contact in Gaussian Model

The rendering capability of Gaussians ensures that the reconstructed human-object interactions are consistent with the images, while the depth information and rendering characteristics of Gaussians enable us to obtain potential contact areas. Given that in monocular images it is difficult to directly determine the contact area between the human and the object, however, we can identify areas where there is no interaction easily, *i.e.*, regions in the image where the human and object do not occlude each other. This allows us to infer potential contact areas between the human and object.

In the optimization of g^{hoi}, the Gaussian points where g^h and g^o occlude each other tend to have *lower opacity* α. In Eq. 1, the color of a pixel is influenced by the opacity of the Gaussian points. If the opacity is very low, the contribution of that Gaussian point to the pixel's color becomes less. While optimizing g^{hoi}, we also simultaneously optimize g^h and g^o. Given the original 2D image gt and human image gt^h (Fig. 6), the occluded areas between the human and the object appear as background in gt^h, so does the object image gt^o. Thus, the opacity of the Gaussian points projected in these areas will tend to decrease in rendering.

Additionally, even in regions without overlap in 2D, occlusion relations exist in 3D depth. Relative to the camera, we consider points in the back to be potential contact areas. Without constraints, both front and

doi.org/10.32388/K7APQ4

back points relative to the camera would participate in rendering. Thus, we set a very low opacity for points facing back from the camera by calculating normals in advance when initializing Gaussians. This approach is based on the assumption that the human pose will not undergo significant changes in optimization, effectively resolving the issue. Setting a low initial opacity allows points closer to the camera to be prioritized in rendering. As the Gaussian point scale increases, it naturally occludes points further back, preventing them from contributing to the rendering. In Fig. 6 (a) and (b), the *blue* area represents regions with high opacity, while the *red* area starts with relatively high opacity. As the optimization progresses, occluded parts of the human leg turn *blue*, designating them as potential contact regions. Meanwhile, the opacity of points behind the camera remains relatively stable.

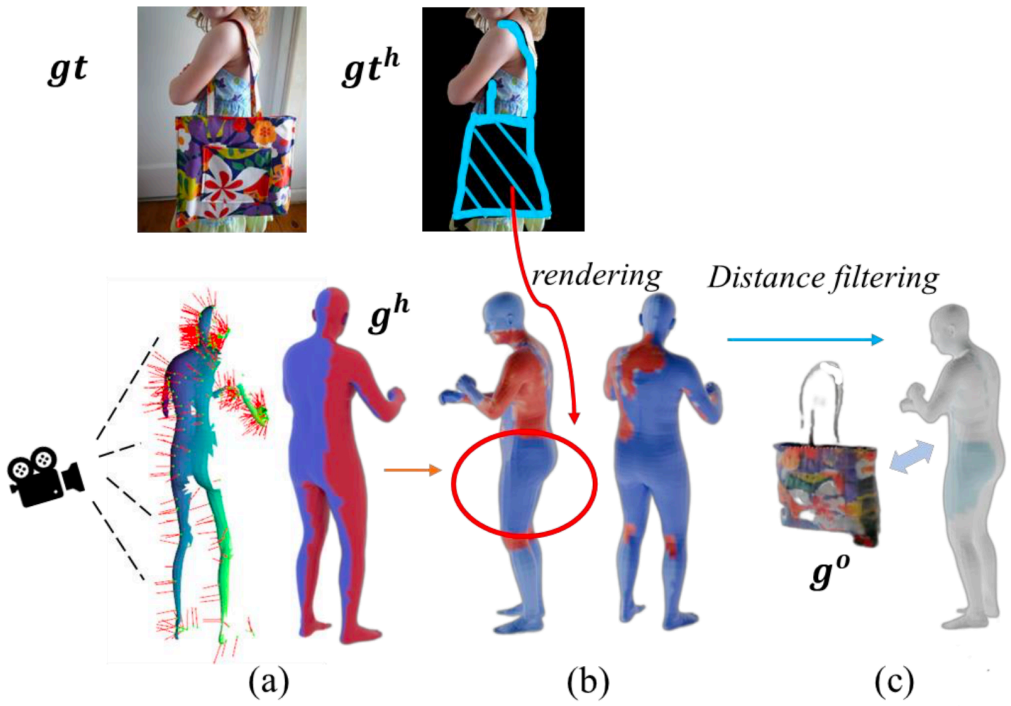

Figure 6. Contact region. (a) Opacity initialization using human normals. (b) The distribution of human body point cloud opacity scores is visualized to identify the *blue* region as a potential interaction area. (c) Based on the *approximate* distance between the human body and the object, the optimized contact region is further identified, shown in *light blue*.

Additionally, in the optimization, we will have a relatively stable positional relationship between the human and the object before we use the contact region to further optimize. By setting a distance threshold between human and object, we can further narrow down the potential contact area, shown in

Fig. 6 (c). Therefore, we introduce a new attribute c into g^h, which represents the contact interaction score of a Gaussian point. The calculation of c is given by

$$c = w_\alpha \cdot \mathrm{Norm}(\alpha^h) + w_d \cdot d_C(p^h, p^o)^h, \tag{4}$$

where $Norm$ means the normalization of the vector to a range between 0 and 1, and d_C means the Chamfer distance.

4.4. Loss Function

Our loss function is divided into two parts: Gaussian rendering loss which is used to optimize the 2D alignment, and HOI loss which is used to optimize the spatial interaction between the human and the object.

Rendering Loss. We adopt the training loss used in 3D Gaussian Splatting, including the L1 loss between the rendered image and GT image, the L2 loss between the rendered mask and GT mask, as well as SSIM (Structural SIMilarity index) loss and LPIPS (Learned Perceptual Image Patch Similarity) loss. To ensure the rendering quality of both the human and the object individually, as well as their combined rendering, we perform separate rendering and loss calculations for g^{hoi}, g^h, and g^o:

$$\mathcal{L}_r = w_h \mathcal{L}_r^h + w_o \mathcal{L}_r^o + w_{hoi} \mathcal{L}_r^{hoi}. \tag{5}$$

HOI Loss. We use the HOI loss to constrain the spatial interaction between the human and the object, ensuring its plausibility. First, we calculate the Chamfer distance between the human contact area, obtained through the Gaussian rendering process, and the object as a contact loss. At the same time, to ensure that the human and object meshes do not intersect, we follow [49] to add a collision loss $\mathcal{L}_{colli}$. We also follow [36] by adopting an Ordinal Depth loss $\mathcal{L}_{depth}$ to constrain the depth relationships. Our final HOI loss and the total training loss are:

$$\begin{aligned} \mathcal{L}_{hoi} &= \mathcal{L}_{cont} + \mathcal{L}_{colli} + \mathcal{L}_{depth}, \\ \mathcal{L} &= w_r \cdot \mathcal{L}_r + w_{hoi} \cdot \mathcal{L}_{hoi}. \end{aligned} \tag{6}$$

5. Experiments

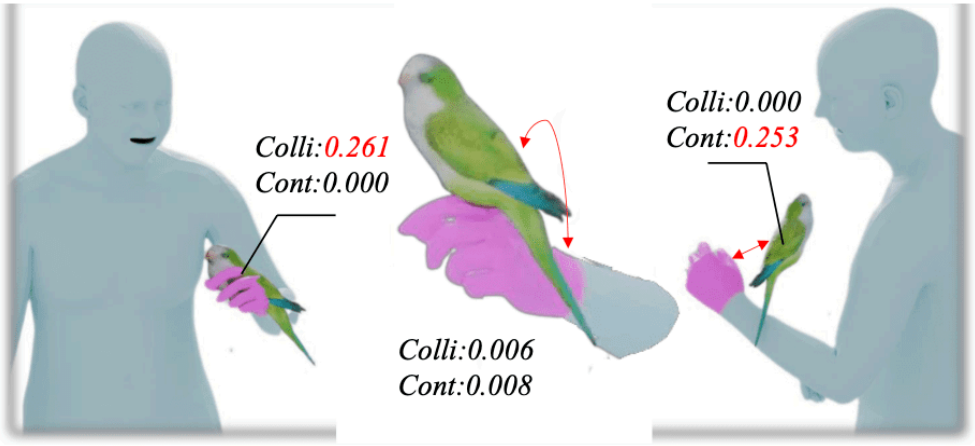

Figure 7. Visualization of our Co^2 Metric. For example, the Colli score focuses on the collision between the bird and the person, while the Cont score calculates the mean Chamfer Distance between the annotated contact part (left hand) and the bird.

5.1. Implementation Details

Experiments are conducted on our Open3DHOI test set and our method doesn't need training as an optimizer. Object meshes with pose before manual annotation are given. We adopt a staged optimization, optimizing each image for 160 iterations. For the first 100 iterations, only the rendering loss is optimized, and HOI loss is added in the subsequent iterations. Due to our dataset being open-vocabulary, it would be unfair to compare methods trained for specific objects on our dataset, so we use the same training-free method PHOSA[36] as our baseline. To ensure fairness, both our method and PHOSA initialize human body parameters using the parameters from our dataset.

doi.org/10.32388/K7APQ4

Methods	Scale↓	Translation (cm)↓	Rotation↓	Cf Distance(cm)
PHOSA[36]	0.39	77.79	0.95	49.1
Ours w/o HOI Loss	0.25	38.66	0.45	**16.9**
Ours	**0.16**	**38.44**	**0.41**	19.3

Table 2. Comparison on object pose metrics.

Methods	Co^2↓	Collision↓	Contact↓
PHOSA[36]	0.431	0.105	0.326
Coarse Recon	0.248	0.083	0.165
Ours Gs only	0.287	0.136	0.151
Gs& depth	0.216	0.080	0.136
Gs& colli	0.189	0.046	0.143
Gs & depth & colli	0.188	**0.045**	0.143
Gs & depth & colli &cont	**0.181**	0.053	**0.128**

Table 3. Comparison on collision and contact metrics.

score	Action		Object	
	obj	w/o obj	action	w/o action
Top-1	0.47	0.20	0.32	0.31

Table 4. Top-1 Accuracy under different prompts.

doi.org/10.32388/K7APQ4

5.2. Metrics

To more accurately evaluate the reconstruction quality, we used two metrics. First, we compared the reconstructions with the object pose in our annotated data, evaluating the differences in scale, translation, rotation and Chamfer Distance. The scale measures the difference between the predicted and GT objects' sizes. The translation is the distance between the predicted and GT objects in cm. Rotation is the norm of the predicted rotation matrix and the eye matrix. We designed an alternative metric that better evaluates 3D interaction quality. It combines the extent of collision between the human and the object with the Chamfer distance between the human and the object within our annotated contact regions. We call this metric as C_o^2 **(Collision-Contact) score:**

$$C_o^2 = \text{Sig}(\text{Colli}(h, o)) + \text{Sig}(\sum_{i}^{p}(\text{Cont}(i)/\text{Size})), \tag{7}$$

where Sig is sigmoid function, $Colli(h, o)$ is the collision between human mesh and object mesh. Cont calculates the Chamfer distance between each human body part and the object. $Size$ is the object mesh size.

5.3. Analysis

Gaussian Advantages. The results show that our method significantly outperforms PHOSA. Our method achieves a higher score in Rotation compared to PHOSA because PHOSA optimizes object pose solely through silhouette loss. In contrast, the 3D Gaussian approach can utilize color matching and richer features, reducing cases where there is minimal silhouette difference but significant disparity from the image. Additionally, our method with contact optimization improved the Co^2 score, particularly the Contact score. This indicates that the contact regions derived from Gaussian depth information effectively enhanced 3D human-object interaction quality.

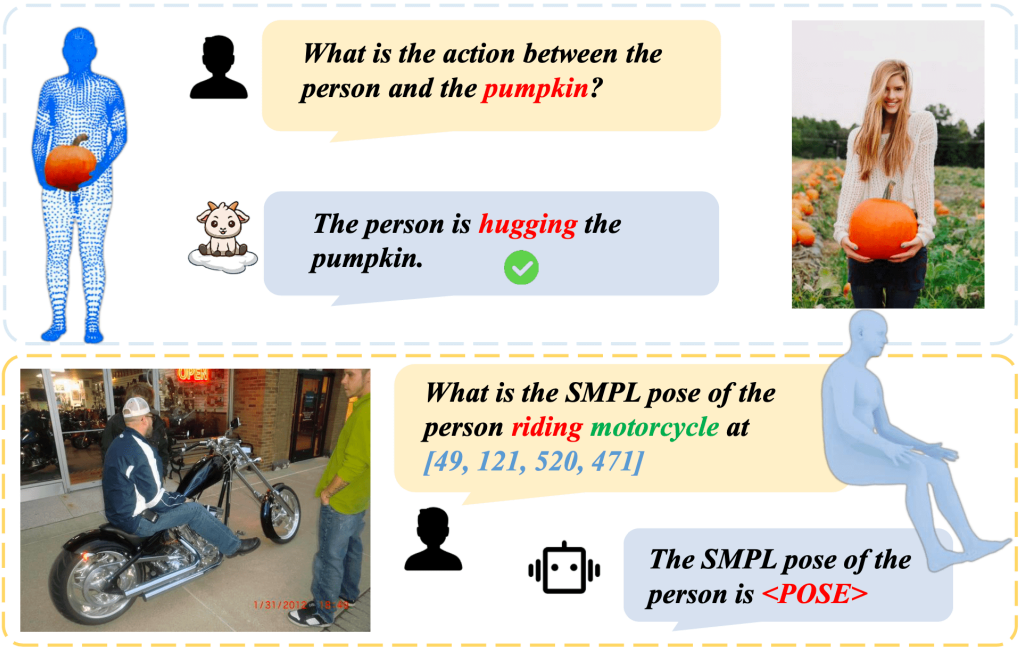

Figure 8. Visualization of HOI Understanding and HOI Pose Chat.

Ablation Study. Coarse Recon in Tab. 3 is the coarse reconstruction using depth and projection in B.1. The Co^2 score using only Gaussian optimization is lower than that of Coarse Recon. Since Gaussian optimization alone does not greatly enhance spatial interaction information, it mainly refines object pose according to the image. However, after adding depth, collision, and contact losses, the 3D score improves significantly, demonstrating that HOI losses are highly effective in optimizing interactions. After adding the contact loss, the collision score slightly decreased, but the contact score improved significantly. Because optimizing the contact area sometimes increases collision in certain images, as objects are moved closer to the intended contact regions. Co^2 metric is to better balance this trade-off.

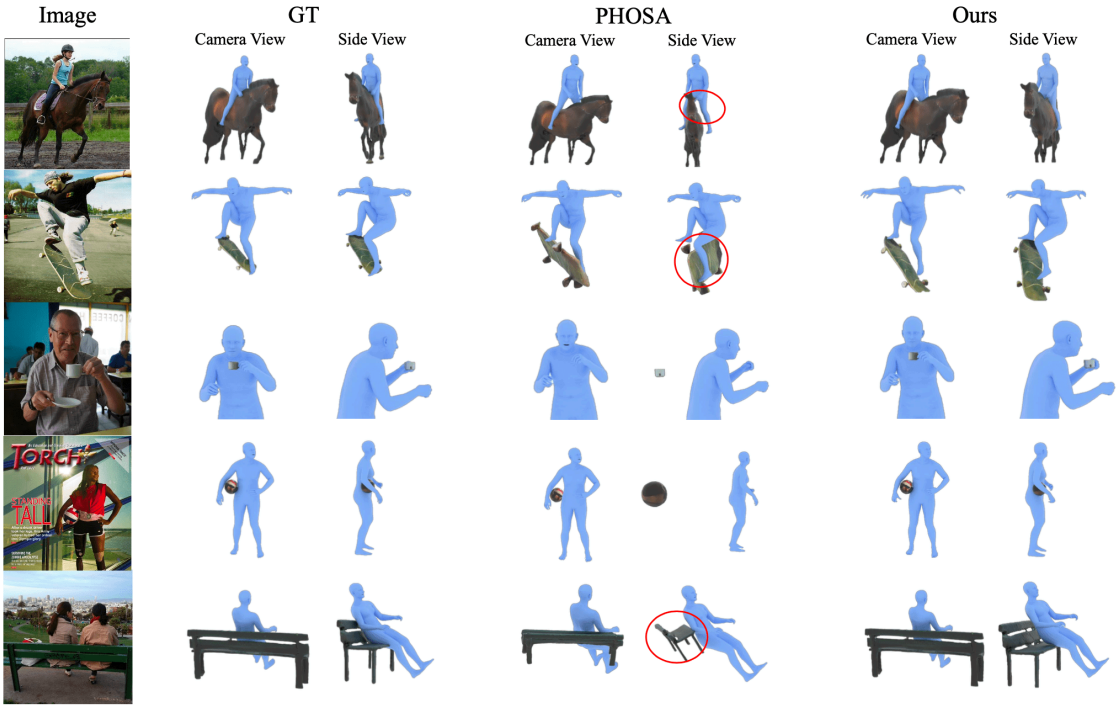

Figure 9. Visualized results comparison between GT, PHOSA, and Ours.

6. More Tasks

Our dataset, with its extensive 2D and 3D annotations, can be utilized for various other tasks. In this section, we propose two more tasks as shown in Fig. 8.

6.1. 3D HOI Understanding

Understanding 3D assets has been a long-standing area of interest, and recently some large models[50][51][52] have achieved impressive results in 3D object comprehension. We tested the current state-of-the-art point clouds understanding model PointLLM[50] on our 3D HOI data to evaluate its capability in action understanding. We provide the model with the point clouds of annotated human and object and ask it to answer the interaction verb between the human and the object. We then compared the output action with GT action annotation. We ask the LLM "What is the action between the person and the [object]?" and "What is the person [Interacting] with?", where [object] and [Interacting] can be replaced with specific verbs and object or not.

We used the Top-1 score as a metric, the results in Tab. 4 indicate that PointLLM demonstrates a certain level of understanding of human interaction point clouds, though it remains limited. Given object name significantly improves action answering performance because PointLLM will estimate the action according to the object category with common sense, but given action name will not improve object answering score, which indicates that PointLLM has limited ability to understand the interaction in the point cloud.

6.2. HOI Pose Chat

Recently, large models have focused on integrating semantics with 3D data. ChatPose[53] uses a framework with LLMs to understand and infer 3D human poses from images or textual descriptions. F-HOI[54] leverages large models to unify various HOI tasks. Since our dataset provides 2D annotations of HOI semantics, we tested the open-sourced large model ChatPose. To evaluate its HOI reasoning and pose generation, we selected cases from the dataset with more than one person, provided the model with the target object's location, and asked it to output the SMPL pose of the person interacting with that object. We then compared this generated SMPL pose with the GT SMPL pose from our dataset. We ask Chatpose "What is the SMPL pose of the person [Interacting] [object] at [Location]?".

The results in Tab. 5 indicate that ChatPose's ability to accurately locate the target human body and obtain the correct pose still needs improvement. In the future, we hope our dataset can help drive the development of more powerful models capable of better understanding images and simultaneously obtaining both human and object poses.

Prompt	MPJPE↓	MPVPE ↓
Action	103.6	131.2
w/o Action	105.2	133.5
Action + Object	**103.4**	**130.9**

Table 5. ChatPose performance results under different prompts.

doi.org/10.32388/K7APQ4

7. Discussion

We propose a real-world 3D HOI annotation pipeline that provides a paradigm for obtaining rich 3D human-object interaction data from unlimited 2D images. Our proposed annotation process relies on the ability of 3D human and object reconstruction tools. In the future, with more advanced 3D-AIGC tools, the annotation efficiency will be largely improved. Moreover, our LLM-based 3d testing tasks proved that existing 3D general models are poor at 3D HOI understanding. As understanding 3D HOI is an important task, it requires more fine-annotated data to drive more general and capable models in the future.

8. Conclusion

In this work, we propose a method for annotating 3D HOIs from open-world single-view images and create Open3DHOI. The rich annotations in our dataset can support various 3D action tasks. For 3D HOI reconstruction, we introduce a 3D Gaussian optimizer that surpasses baselines. Results of current methods reveal that they are not yet capable of understanding 3D HOIs well. We believe Open3DHOI will pave the way for future 3D HOI learning.

Appendix Overview

The contents of this supplementary material are:

- Sec. A: Characteristics of Open3DHOI.
- Sec. B: Method Details.
- Sec. 2: Additional Experiments.

Appendix A. Characteristics of Open3DHOI

A.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. 10 shows the co-occurrence between the major object categories and actions, and Tab. 6 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. 11, 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. 12 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. 13 shows some cases of our abnormal HOIs.

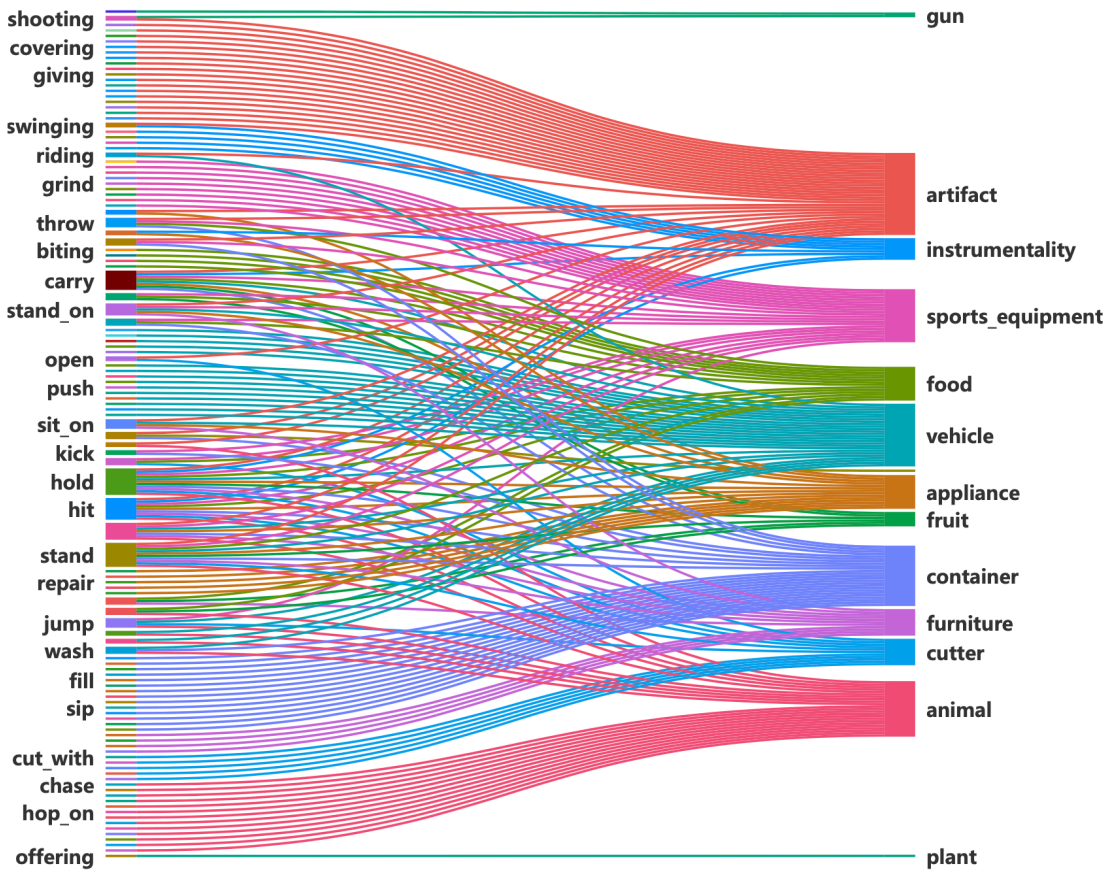

Figure 10. Co-occurence between major object category and actions in Open3DHOI.

doi.org/10.32388/K7APQ4

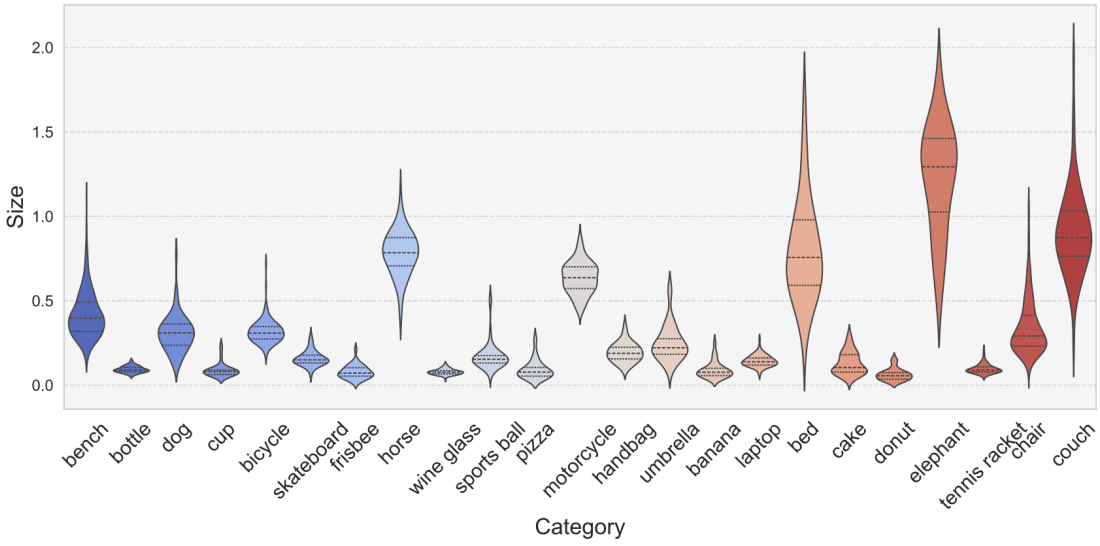

Figure 11. Object size distribution in different object categories.

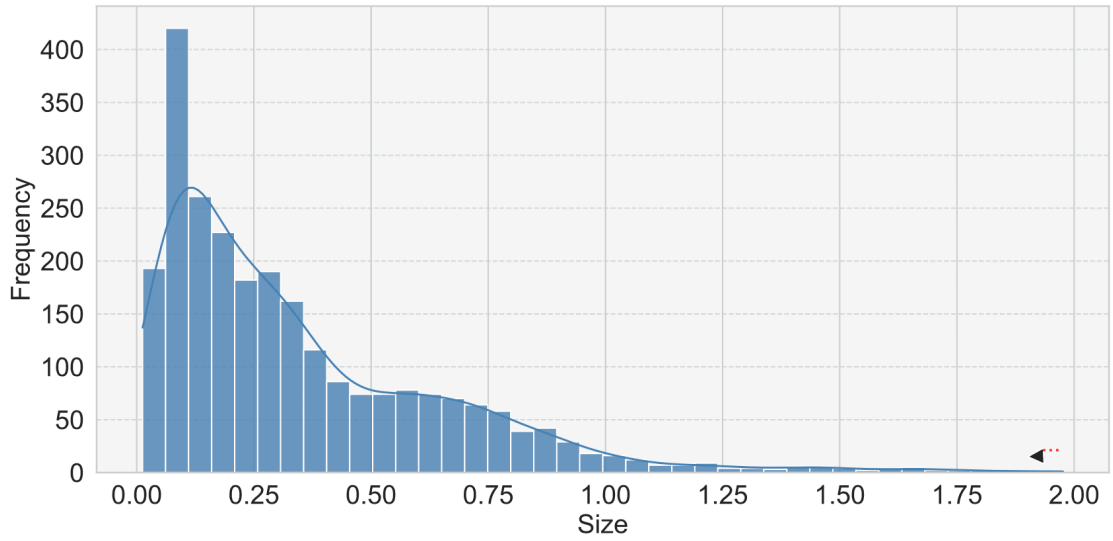

Figure 12. Object size distribution of all images.

doi.org/10.32388/K7APQ4

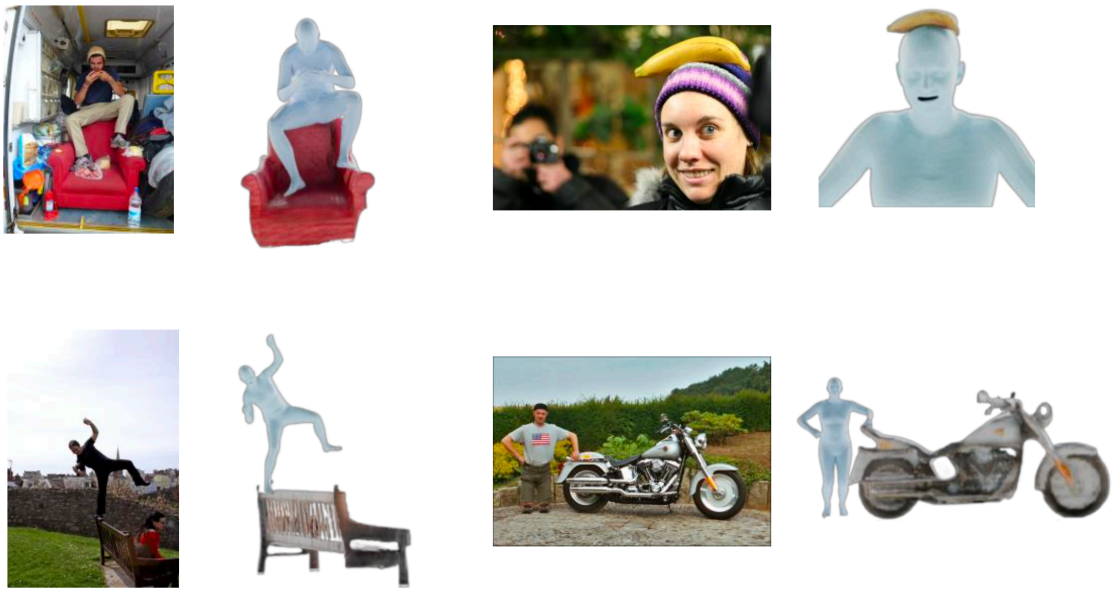

Figure 13. Abnormal HOIs in Open3DHOI.

doi.org/10.32388/K7APQ4

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

doi.org/10.32388/K7APQ4

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	pipage	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						

Table 6. Object categories in our Open3DHOI dataset.

A.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. 7. In Fig. 14, 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 involving other body regions.

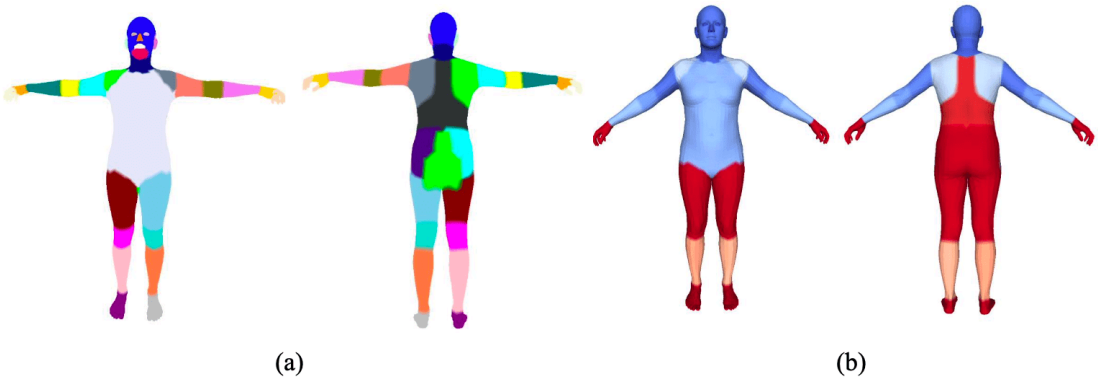

Figure 14. Body parts and annotation heat map.

doi.org/10.32388/K7APQ4

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

Table 7. Body part name and annotation number.

Appendix B. Method Details

B.1. Coarse Reconstruction

In paper Fig.2, we introduce the process of coarse reconstruction. In this section, we provide additional details about this process. After reconstructing human and object meshes, we use depth to initialize

doi.org/10.32388/K7APQ4

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.

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 $K_h = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix}$
$x_h^{proj} \leftarrow \frac{\overline{z}c_x}{f}, y_h^{proj} \leftarrow \frac{\overline{z}c_y}{f}$

2. $Index_h \leftarrow$ compare image and H_proj to obtain the indices of S corresponding to H_3D
$H\_3D \leftarrow H\_3D[argsort(H\_3D[:,2])//2]$ 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}
$translation \leftarrow mean(S_h) - mean(H\_3D * scale)$

4. $Obj\_3D \leftarrow$ operate Obj_{init} by $Scale * Obj_{init} + Translation$
return Obj_3D

Algorithm 1. Align object mesh with human mesh.

B.2. Annotation Tools

B.2.1. Filtering Tool

Fig. 16 (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. 17. 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. 15 shows cases with bad masks and with good masks but bad reconstructions.

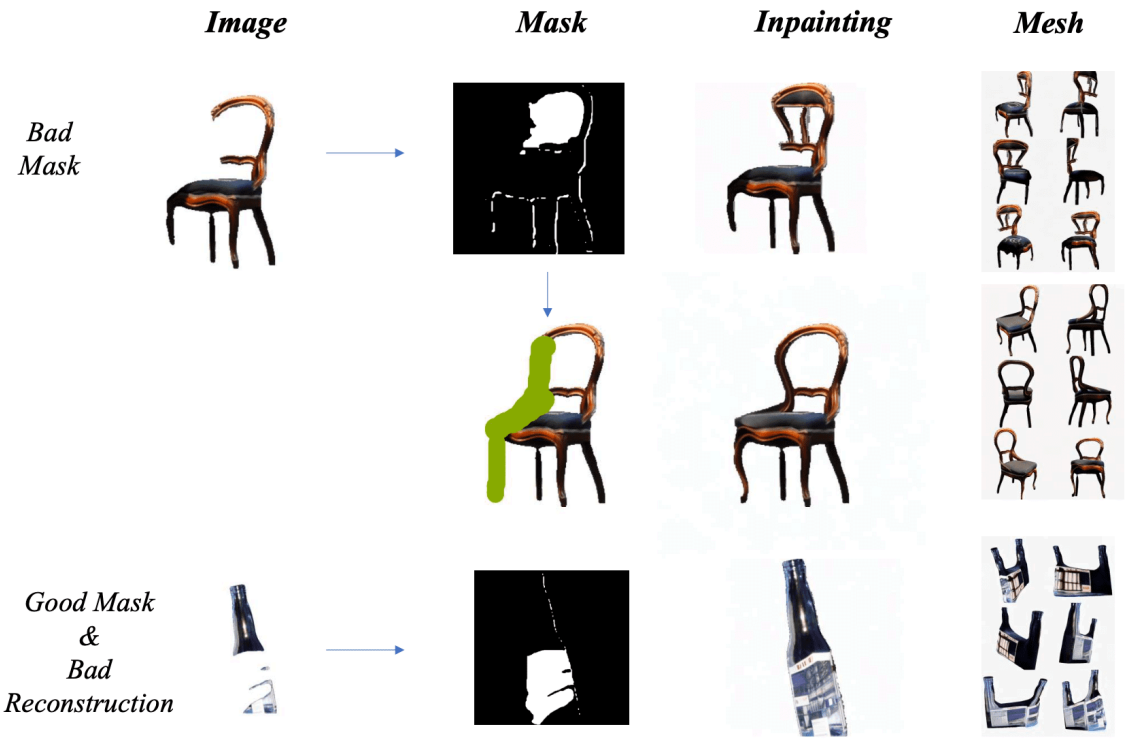

Figure 15. Special cases in filtering process.

doi.org/10.32388/K7APQ4

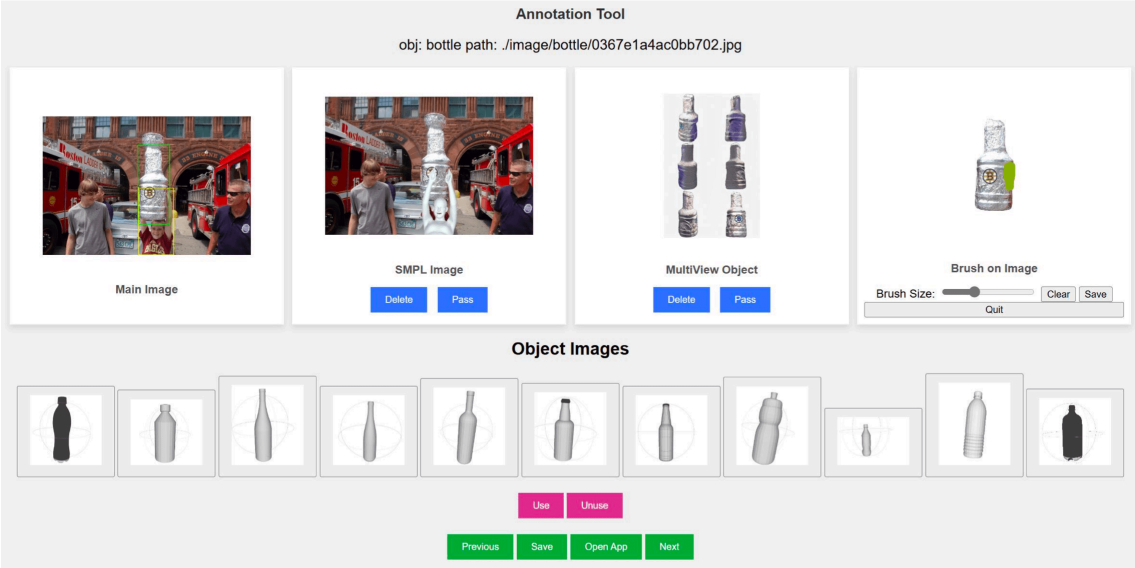

Figure 16. Filtering tool.

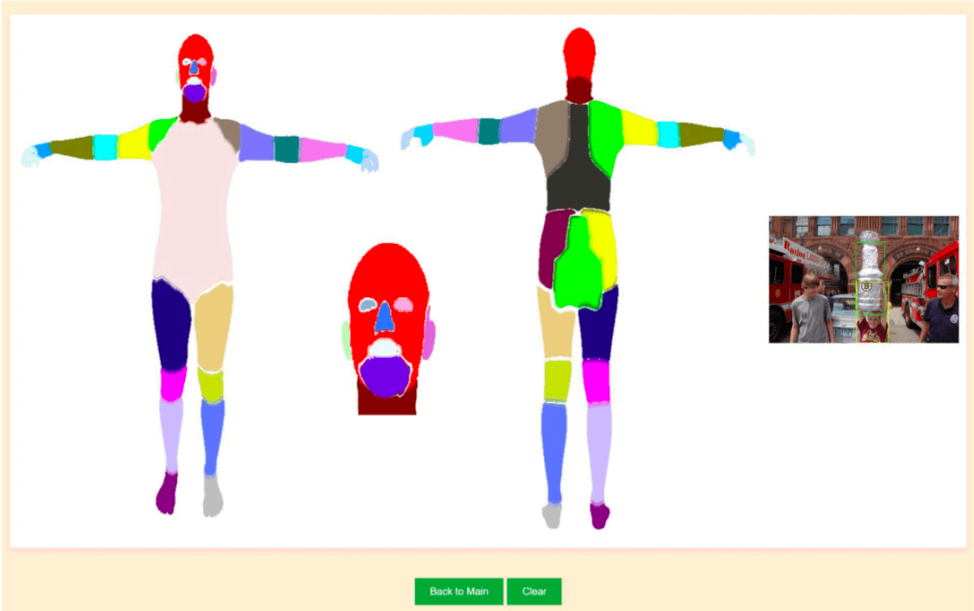

Figure 17. Contact part annotation tool.

B.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. B.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 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. 19 shows our 3D fine annotation tool based on ImageNet3D[42], 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.

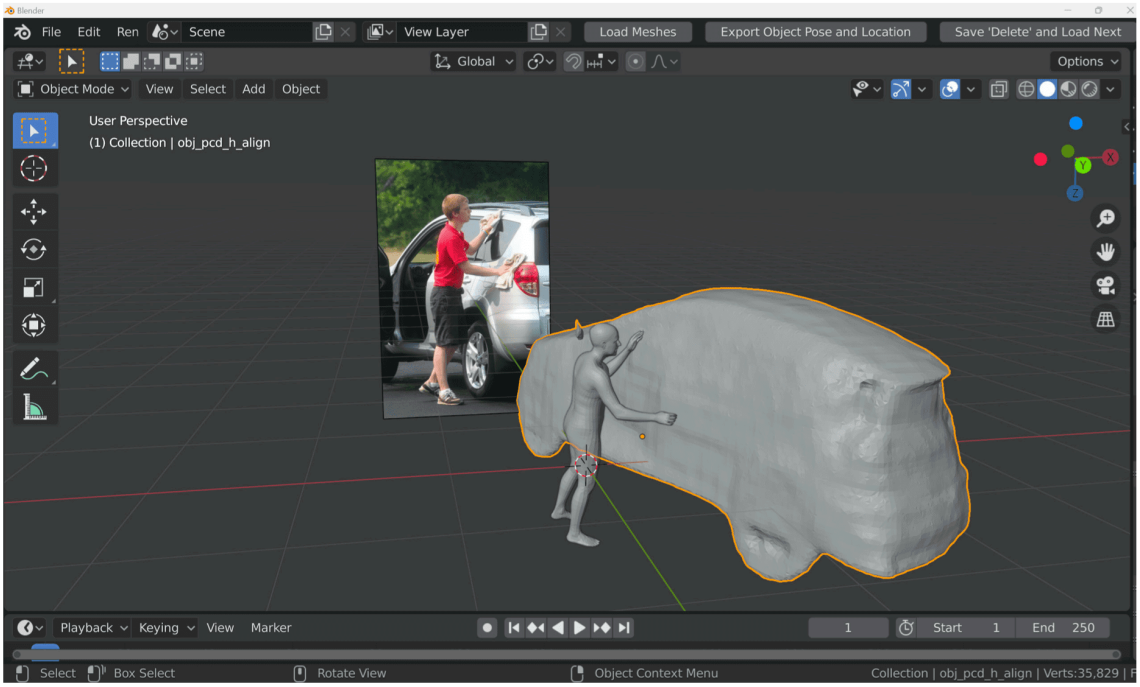

Figure 18.

doi.org/10.32388/K7APQ4

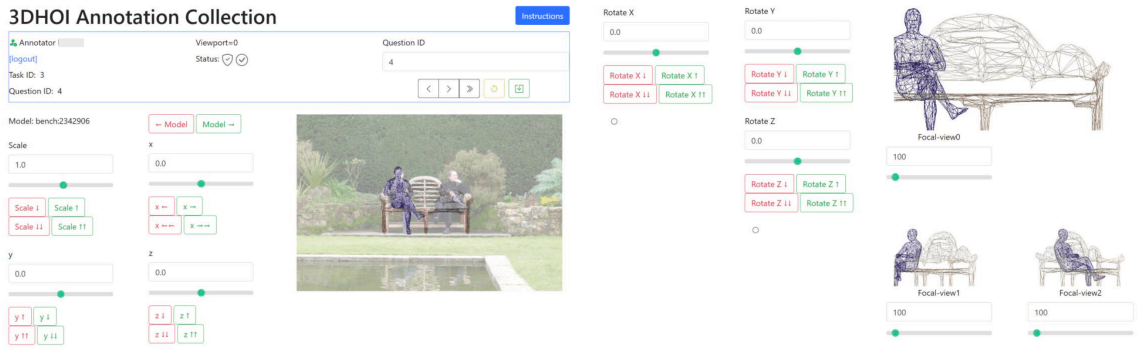

Figure 19. 3D fine annotation tool.

B.2.3. Dataset quality

1) Human-object penetration rate: we tested the penetration metric following[55] 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., 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.

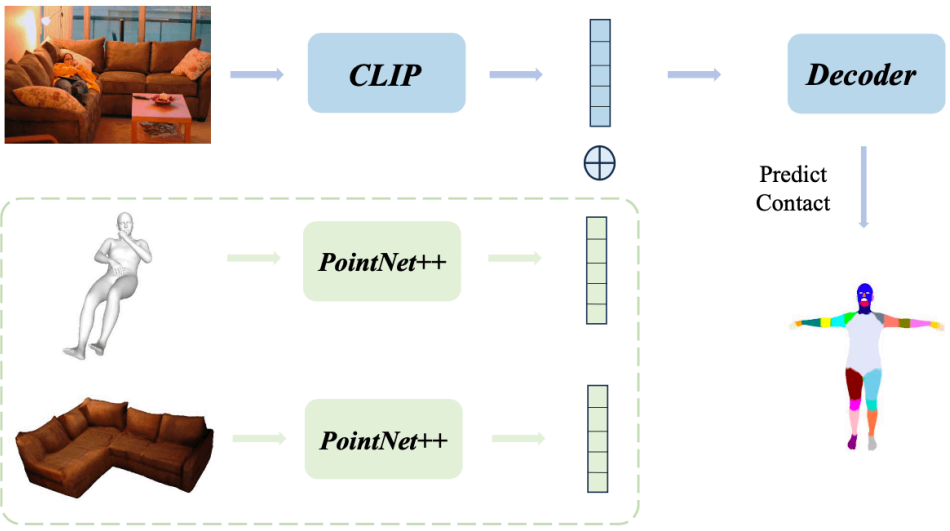

Figure 20. Our contact evaluation model.

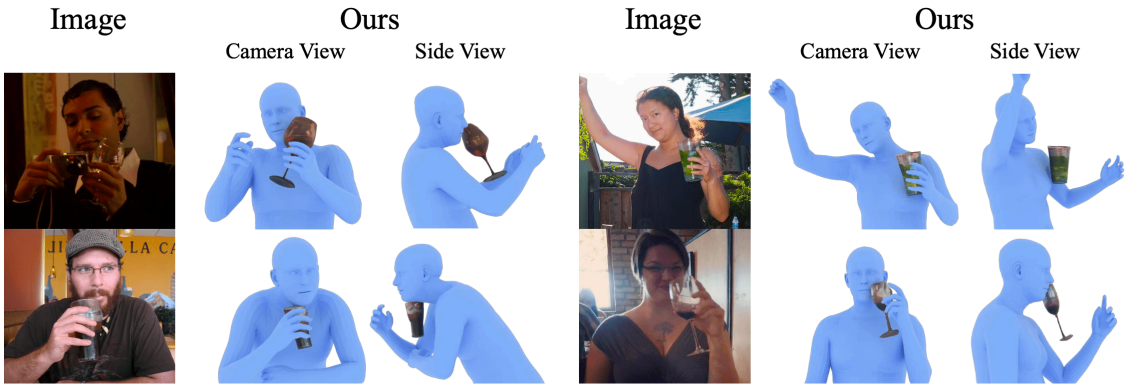

Figure 21. Failure cases of HOI-Gaussian.

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

Methods	Micro F1-Score ↑	Hamming Loss ↓	Jaccard Index ↑
2D	0.6118	0.0874	0.4303
2D&3D	**0.6207**	**0.0844**	**0.4561**

Table 8. Results of our contact evaluation task.

B.3. LLM Task Setting

B.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[56] to extract the exact word from our action and object list.

B.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[57] 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.

Appendix C. Additional Experiments

C.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. 20 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. 8 still indicate that 3D information is beneficial to the estimation of contact regions.

C.2. Failure Cases

Fig. 21 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.

C.3. More Results

Fig. 22 shows more results comparison between GT, PHOSA, and Ours.

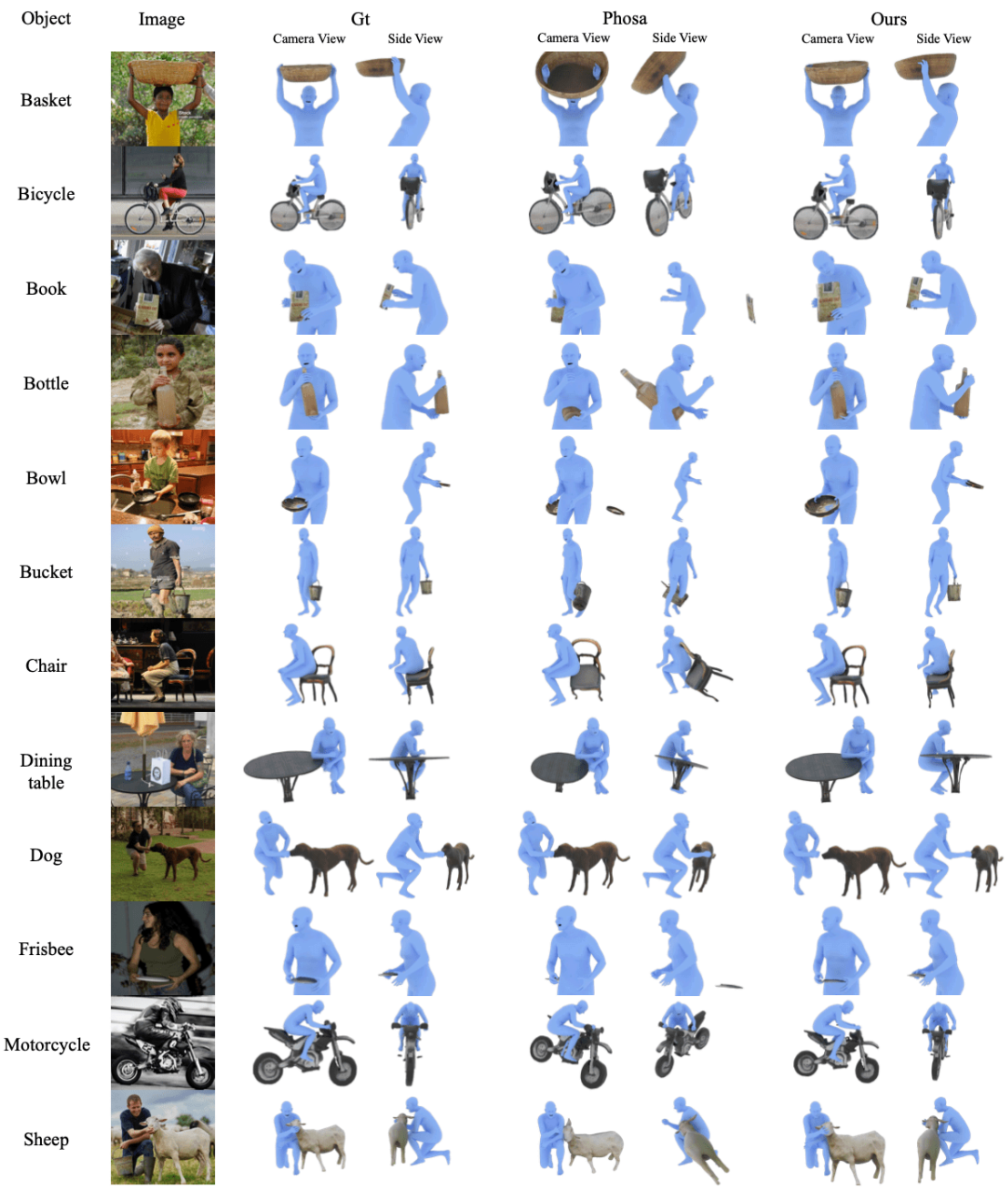

Figure 22. More results.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant No.62306175, 62302296, 72192821, 62472282.

References

1. ^Zhang FZ, Yuan Y, Campbell D, Zhong Z, Gould S. Exploring predicate visual context in detecting human–object interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023 Oct. p. 10411-10421.
2. a, b, c, dLi YL, Liu X, Wu X, Li Y, Qiu Z, Xu L, Xu Y, Fang HS, Lu C (2022). "Hake: a knowledge engine foundation for human activity understanding". IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. Published by IEEE.
3. a, b, c, dBhatnagar BL, Xie X, Petrov I, Sminchisescu C, Theobalt C, Pons-Moll G. "BEHAVE: Dataset and Method for Tracking Human Object Interactions." In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022.
4. a, b, c, dHuang Y, Taheri O, Black MJ, Tzionas D. "InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction." In: German Conference on Pattern Recognition (GCPR). Lecture Notes in Computer Science. Springer; 2022. p. 281-299.
5. ^Zhao C, Zhang J, Du J, Shan Z, Wang J, Yu J, Wang J, Xu L. "I'M HOI: Inertia-aware Monocular Capture of 3D Human-Object Interactions." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun. p. 729-741.
6. a, bZhao K, Wang S, Zhang Y, Beeler T, Tang S (2022). "Compositional Human-Scene Interaction Synthesis with Semantic Control". In: European conference on computer vision (ECCV).
7. a, b, c, dHuo C, Shi Y, Wang J. "Monocular Human-Object Reconstruction in the Wild." In: Proceedings of the 32nd ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery; 2024. p. 5547–5555. doi:10.1145/3664647.3681452.
8. a, b, cYang Y, Zhai W, Luo H, Cao Y, Zha ZJ (2023). "LEMON: Learning 3D Human-Object Interaction Relation from 2D Images". arXiv preprint arXiv:2312.08963. Available from: https://arxiv.org/abs/2312.08963.
9. a, b, cWang S, Yap KH, Ding H, Wu J, Yuan J, Tan YP (2021). "Discovering Human Interactions With Large-Vocabulary Objects via Query and Multi-Scale Detection". In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
10. a, b, c, d, eXu J, Cheng W, Gao Y, Wang X, Gao S, Shan Y (2024). "InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models". arXiv preprint arXiv:2404.07191.
11. a, b, c, dLin J, Zeng A, Wang H, Zhang L, Li Y (2023). "One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec

ognition. 2023: 21159–21168.

12. a, b*Chao YW, Liu Y, Liu X, Zeng H, Deng J (2018). "Learning to Detect Human-Object Interactions". In: Procee dings of the IEEE Winter Conference on Applications of Computer Vision.*
13. ^*Gupta S, Malik J (2015). "Visual Semantic Role Labeling". arXiv preprint arXiv:1505.04474.*
14. ^*Zhuang B, Wu Q, Shen C, Reid I, van den Hengel A (2017). "Care about you: towards large-scale human-cen tric visual relationship detection". arXiv preprint arXiv:1705.09892.*
15. ^*Liu X, Wen B, Liu X, Zhou Z, Fan H, Lu C, Ma L, Chen Y, Li YL. "Interacted Object Grounding in Spatio-Temp oral Human-Object Interactions." In: AAAI; 2025.*
16. a, b*Li YL, Liu X, Lu H, Wang S, Liu J, Li J, Lu C (2020). "Detailed 2D-3D Joint Representation for Human-Object Interaction". In: CVPR.*
17. ^*Li YL, Wu X, Liu X, Dou Y, Ji Y, Zhang J, Li Y, Tan J, Lu X, Lu C. From isolated islands to pangea: Unifying sem antic space for human action understanding. In: CVPR; 2024.*
18. ^*Li YL, Xu L, Liu X, Huang X, Xu Y, Wang S, Fang HS, Ma Z, Chen M, Lu C (2020). "PaStaNet: Toward Human Activity Knowledge Engine". In: CVPR.*
19. ^*Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015). "SMPL: A skinned multi-person linear mod el". ACM Transactions on Graphics, (Proc. SIGGRAPH Asia). **34** (6): 248:1–248:16.*
20. ^*Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AA, Tzionas D, Black MJ. "Expressive Body Capture: 3 D Hands, Face, and Body from a Single Image." In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR); 2019.*
21. ^*Cai Z, Yin W, Zeng A, Wei C, Sun Q, Yanjun W, Pang HE, Mei H, Zhang M, Zhang L, Loy CC, Yang L, Liu Z (20 23). "SMPLer-X: Scaling up expressive human pose and shape estimation." In: Advances in Neural Informati on Processing Systems.*
22. ^*Zhang H, Tian Y, Zhang Y, Li M, An L, Sun Z, Liu Y (2023). "PyMAF-X: Towards Well-aligned Full-body Mod el Regression from Monocular Images". IEEE Transactions on Pattern Analysis and Machine Intelligence.*
23. ^*Li Z, Liu J, Zhang Z, Xu S, Yan Y (2022). "CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation". In: ECCV.*
24. ^*Poole B, Jain A, Barron JT, Mildenhall B (2022). "DreamFusion: Text-to-3D using 2D Diffusion". arXiv.*
25. ^*Qian G, Mai J, Hamdi A, Ren J, Siarohin A, Li B, Lee H-Y, Skorokhodov I, Wonka P, Tulyakov S, Ghanem B. "Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors." In: Th e Twelfth International Conference on Learning Representations (ICLR); 2024. Available from: https://openr eview.net/forum?id=0jHkUDyEO9.*

26. ^Shi R, Chen H, Zhang Z, Liu M, Xu C, Wei X, Chen L, Zeng C, Su H (2023). "Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model". arXiv. arXiv:2310.15110.

27. ^Liu M, Xu C, Jin H, Chen L, Varma T M, Xu Z, Su H (2024). "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization". Advances in Neural Information Processing Systems. **36**.

28. ^Long X, Guo YC, Lin C, Liu Y, Dou Z, Liu L, Ma Y, Zhang SH, Habermann M, Theobalt C, et al. (2023). "Wonder3D: Single Image to 3D using Cross-Domain Diffusion". arXiv preprint arXiv:2310.15008. **2023**. Available from: arXiv:2310.15008.

29. ^Melas-Kyriazi L, Rupprecht C, Laina I, Vedaldi A. RealFusion: 360 Reconstruction of Any Object from a Single Image. In: CVPR; 2023. Available from: https://arxiv.org/abs/2302.10663.

30. ^Hong Y, Zhang K, Gu J, Bi S, Zhou Y, Liu D, Liu F, Sunkavalli K, Bui T, Tan H (2023). "Lrm: Large reconstruction model for single image to 3d". arXiv preprint arXiv:2311.04400. 2023.

31. ^Li J, Tan H, Zhang K, Xu Z, Luan F, Xu Y, Hong Y, Sunkavalli K, Shakhnarovich G, Bi S (2023). "Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model". arXiv.

32. ^Liu S, Li YL, Fang Z, Liu X, You Y, Lu C (2024). "Primitive-based 3D Human-Object Interaction Modelling and Programming". In: AAAI.

33. ^Xu X, Zhang Y, Li YL, Han L, Lu C (2024). "HumanVLA: Towards Vision-Language Directed Object Rearrangement by Physical Humanoid". In: NeurIPS.

34. ^Liu J, Liang X, Lin Z, Hou H, Li YL, Lu C (2025). "ImDy: Human Inverse Dynamics from Imitated Observations". In: ICLR.

35. ^Liu X, Li YL, Zeng A, Zhou Z, You Y, Lu C (2024). "Bridging the Gap between Human Motion and Action Semantics via Kinematic Phrases". In: ECCV.

36. a, b, c, d, e, fZhang JY, Pepose S, Joo H, Ramanan D, Malik J, Kanazawa A. "Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild." In: European Conference on Computer Vision (ECCV); 2020.

37. ^Wang X, Li G, Kuo YL, Kocabas M, Aksan E, Hilliges O (2022). "Reconstructing Action-Conditioned Human-Object Interactions Using Commonsense Knowledge Priors." In: 2022 International Conference on 3D Vision (3DV), pages 353-362. doi:10.1109/3DV57658.2022.00047.

38. ^Xie X, Bhatnagar BL, Pons-Moll G (2022). "CHORE: Contact, Human and Object REconstruction from a single RGB image". In: European Conference on Computer Vision (ECCV). {Springer}, October 2022.

39. ^Zhan G, Zheng C, Xie W, Zisserman A (2024). "Amodal Ground Truth and Completion in the Wild". CVPR.

40. ^Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. "High-Resolution Image Synthesis With Latent Diffusion Models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. p. 10684-10695.

41. ^Bhat SF, Birkl R, Wofk D, Wonka P, Müller M (2023). "ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth". arXiv. doi:10.48550/ARXIV.2302.12288. Available from: https://arxiv.org/abs/2302.12288.

42. a, bMa W, Zeng G, Zhang G, Liu Q, Zhang L, Kortylewski A, Liu Y, Yuille A (2024). "ImageNet3D: Towards General-Purpose Object-Level 3D Understanding". arXiv preprint arXiv:2406.09613.

43. ^Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R (2023). "Segment anything". arXiv:2304.02643.

44. ^Kerbl B, Kopanas G, Leimkühler T, Drettakis G (2023). "3D Gaussian Splatting for Real-Time Radiance Field Rendering". ACM Transactions on Graphics. **42** (4). Available from: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

45. a, bHu S, Liu Z (2023). "GauHuman: Articulated Gaussian Splatting from Monocular Human Videos". arXiv preprint arXiv:.

46. ^Lin W, Zheng C, Yong JH, Xu F (2022). "OcclusionFusion: Occlusion-aware Motion Estimation for Real-time Dynamic 3D Reconstruction". Conference on Computer Vision and Pattern Recognition (CVPR).

47. ^Cai D, Heikkilä J, Rahtu E (2024). "GS-Pose: Generalizable Segmentation-based 6D Object Pose Estimation with 3D Gaussian Splatting". arXiv preprint arXiv:2403.10683v2.

48. ^Bortolon M, Tsesmelis T, James S, Poiesi F, Del Bue A. "6DGS: 6D Pose Estimation from a Single Image and a 3D Gaussian Splatting Model." In: ECCV; 2024.

49. ^Jiang W, Kolotouros N, Pavlakos G, Zhou X, Daniilidis K (2020). "Coherent Reconstruction of Multiple Humans from a Single Image". In: CVPR.

50. a, bXu R, Wang X, Wang T, Chen Y, Pang J, Lin D (2024). "PointLLM: Empowering Large Language Models to Understand Point Clouds". In: ECCV.

51. ^Hong Y, Zhen H, Chen P, Zheng S, Du Y, Chen Z, Gan C (2023). "3D-LLM: Injecting the 3D World into Large Language Models". NeurIPS.

52. ^Qi Z, Fang Y, Sun Z, Wu X, Wu T, Wang J, Lin D, Zhao H (2024). "GPT4Point: A Unified Framework for Point-Language Understanding and Generation". In: CVPR.

53. ^Feng Y, Lin J, Dwivedi SK, Sun Y, Patel P, Black MJ. "ChatPose: Chatting about 3D Human Pose." In: CVPR, 2024.

54. ^Yang J, Niu X, Jiang N, Zhang R, Siyuan H (2024). "F-HOI: Toward Fine-grained Semantic-Aligned 3D Human-Object Interactions". European Conference on Computer Vision.

55. ^Jiang W, Kolotouros N, Pavlakos G, Zhou X, Daniilidis K (2020). "Coherent Reconstruction of Multiple Humans from a Single Image". In: CVPR.

56. ^Qwen Team (2024). "Qwen2.5: A Party of Foundation Models".

57. ^Wu Y, Kirillov A, Massa F, Lo WY, Girshick R. Detectron2. 2019. Available from: https://github.com/facebookresearch/detectron2.

Declarations

Funding: This work is supported in part by the National Natural Science Foundation of China under Grant No.62306175, 62302296, 72192821, 62472282.

Potential competing interests: No potential competing interests to declare.