

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

Boran Wen^{1,2} Dingbang Huang¹ Zichen Zhang¹ Jiahong Zhou¹ Jianbin Deng¹
Jingyu Gong³ Yulong Chen^{1*} Lizhuang Ma^{1*} Yong-Lu Li^{1,2*}

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

{wenboran, 2331970490, zhoujiahong, 13610214637, deng4a42}@sjtu.edu.cn

{llong-c, lzma, yonglu.li}@sjtu.edu.cn, jygong@cs.ecnu.edu.cn

Abstract

*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 [49] and HAKE [20] have been

introduced. For 3D HOI, datasets like BEHAVE [2] and InterCap [12] 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.

In 3D HOI, many datasets have been introduced. Datasets such as BEHAVE [2], InterCap [12], ImHOI [52], and PROX-S [53] provide multi-view RGBD sequences and 3D annotations in *indoor* scenes. Though datasets like WildHOI and 3DIR [13, 47] 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 [20] and SWIG-HOI [40]. Next, we used InstantMesh [43] and OSX [23] 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.

Given the new dataset, we also proposed a *training-free*

*Corresponding authors.

¹Boran Wen and Yong-Lu Li are with the School of Artificial Intelligence, Shanghai Jiao Tong University, and the Zhiyuan College, Shanghai Jiao Tong University.

²Zichen Zhang, Jiahong Zhou, Jianbin Deng, Yulong Chen, and Lizhuang Ma are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University.

Reconstruct 3D HOI From Open-World



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.

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.

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 [6, 8, 18–21, 29, 40, 54] has made our 3D HOI reconstruction approach possible. Datasets like HICO-DET [6] and HAKE [20] provide annotations for 80 object categories and 117 action categories. Additionally, open-vocabulary datasets such as SWIG-HOI [40] 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 [2], 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 [12] further built upon it by offering more detailed hand interaction information. Recently, several benchmarks [13, 47] 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 [31], parametric human body modeling has rapidly evolved. Currently, the SMPL-X model [34], which includes detailed hand and facial expression modeling, is widely used in the field [5, 22, 23, 50]. We utilized a state-of-the-art one-stage model [23] 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 [35, 37] that leverage 2D diffusion priors to Multi-view Diffusion Models [26, 30, 33, 39] and Large Reconstruction Models [9, 17, 43] 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 [43] as our reconstruction model.

2.3. 3D HOI Reconstruction

Reconstructing 3D HOIs from a single image [18, 27] is a challenging task and important for many applications [25, 28, 45]. 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.* [51] optimizes spatial interactions through predefined contact pairs, while Wang *et al.* [41] leverages GPT-3’s prior knowledge to optimize spatial interactions. Xie *et al.* [42] learned HOI spatial arrangement priors from the BEHAVE dataset. Wang *et al.* [13] learned the prior distribution of the 2D human-object keypoint layout and viewpoints to tune the relative pose between the 3D human and the object.

3. 3D HOI Annotation

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

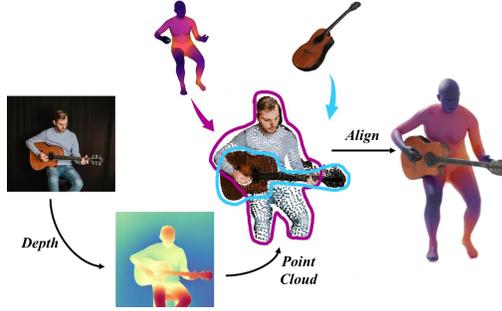


Figure 2. Coarse Reconstruction. We first obtain depth from the images and generate point clouds. Given masks, we extract the corresponding point clouds for the person (pink) and object (blue). We obtain a rough reconstruction by matching the MESH vertices of the person and the object with the depth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

3.1. Coarse Reconstruction Annotation

First, we used a state-of-the-art human pose estimation method [23] to obtain the 3D representation of the human body and employed image-to-3D technique [43] to generate the 3D representation of the object.

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 [48] and used Stable Diffusion 1.5 [38] 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 [1] 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 [43] 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).

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 [32]. 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**.

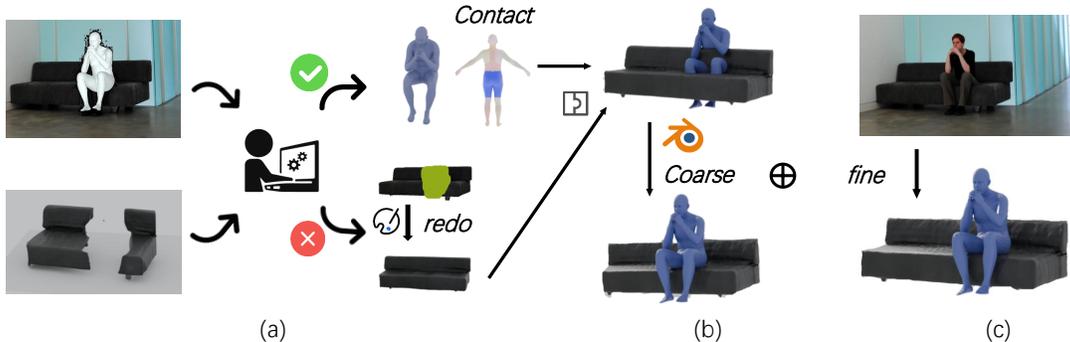


Figure 3. Annotation Pipeline. (a) Filtering. Given the reconstructed human and object meshes, annotators assess the quality. If the human reconstruction is eligible, the contact area is further annotated. If the object reconstruction fails, the mask is redrawn manually and the reconstruction is performed 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.

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 [2]	10	N/A	SMPL+H	✓	✓	✗
InterCAP [12]	10	N/A	SMPL-X	✓	✓	✗
WildHOI [13]	8	N/A	SMPL	✓	✗	✗
3DIR [47]	21	17	SMPL+H	✗	✓	✓
PROX-S [53]	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 [16]. 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 [15] 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 N . The final color of the pixel is then determined using alpha compositing, expressed as:

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

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:

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

Here, \mathbf{x}' is the pixel’s projected coordinate and $\boldsymbol{\mu}'_n$ is the projected center of the n -th Gaussian. Recent

works [11, 24] 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 [11], 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 [4] and 6D-GS [3] 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 [51].

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

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

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 repre-

sent the object’s rotation matrix, scale factor, and translation vector, respectively, and all three variables are learnable.

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 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.

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

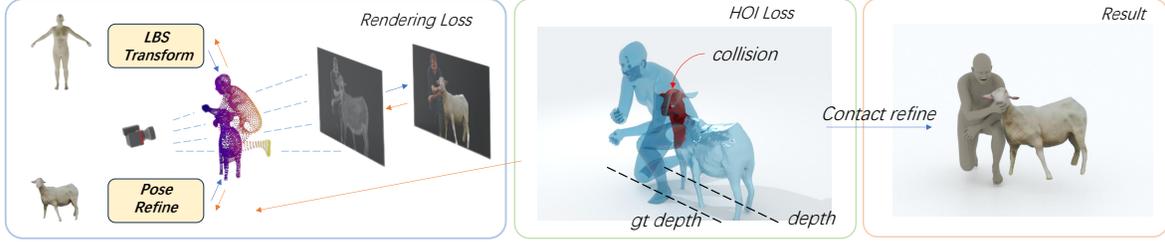


Figure 5. Our pipeline. The optimizer first converts the human and object into 3D Gaussian points, then calculates a rendering loss by comparing the Gaussian-rendered image with the ground truth image. This loss is backpropagated to update the object’s pose parameters and the human’s LBS parameters. 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.

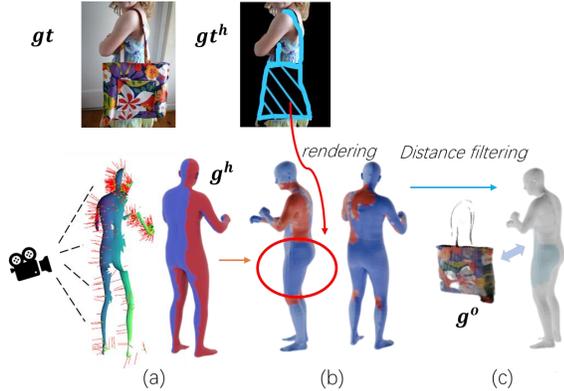


Figure 6. Contact region. (a) Opacity initialization using human normals. (b) The distribution of human body point cloud opacity scores is visualized to identify the *blue* region as a potential interaction area. (c) Based on the approximate distance between the human body and the object, the optimized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

score of a Gaussian point. The calculation of c is given by

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

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 :

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

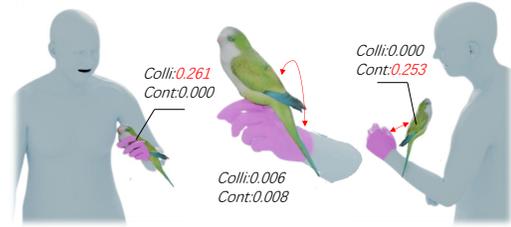


Figure 7. Visualization of our Co^2 Metric. For example, the Colli score focuses on the collision between the bird and the person, while the Cont score calculates the mean Chamfer Distance between the annotated contact part (left hand) and the bird.

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 [14] to add a collision loss \mathcal{L}_{colli} . We also follow [51] by adopting an Ordinal Depth loss \mathcal{L}_{depth} to constrain the depth relationships. Our final HOI loss and the total training loss are:

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

5. Experiments

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 the 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 adopted, 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 [51] as our baseline.

Methods	Scale↓	Translation (cm)↓	Rotation↓	Cf Distance(cm)
PHOSA[51]	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 [51]	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.

To ensure fairness, both our method and PHOSA initialize human body parameters using the parameters from our dataset.

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 in Tab. 2. 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 Co^2 (**Collision-Contact**) score:

$$Co^2 = \text{Sig}(\text{Colli}(h, o)) + \text{Sig}\left(\sum_i^p (\text{Cont}(i)/\text{Size})\right), \quad (7)$$

where *Sig* is sigmoid function, $\text{Colli}(h, o)$ 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

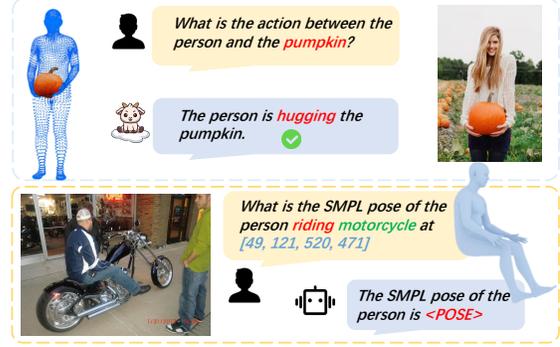


Figure 8. Visualization of HOI Understanding and HOI Pose Chat.

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.

Ablation Study. Coarse Recon in Tab. 3 is the coarse reconstruction using depth and projection in 3.1. 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.

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 [10, 36, 44] have achieved impressive results in 3D object comprehension. We tested the current state-of-the-art point clouds understanding model PointLLM [44] 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 un-

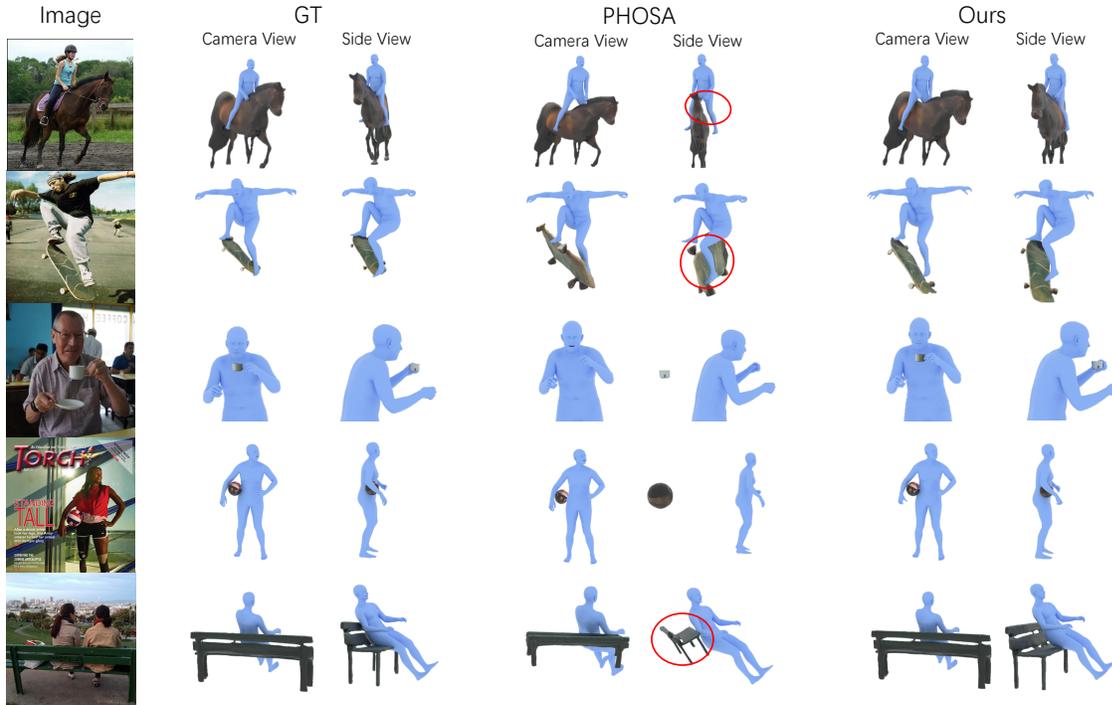


Figure 9. Visualized results comparison between GT, PHOSA, and Ours.

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 [7] uses a framework with LLMs to understand and infer 3D human poses from images or textual descriptions. F-HOI [46] 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.

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.

9. Acknowledgment

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

References

- [1] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 3
- [2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 1, 2, 4
- [3] Matteo Bortolon, Theodore Tsesmelis, Stuart James, Fabio Poiesi, and Alessio Del Bue. 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. In *ECCV*, 2024. 5
- [4] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Gspose: Generalizable segmentation-based 6d object pose estimation with 3d gaussian splatting. 2024. 5
- [5] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023. 2
- [6] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 2
- [7] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 8
- [8] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2
- [9] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2
- [10] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023. 7
- [11] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:*, 2023. 5
- [12] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 1, 2, 4
- [13] Chaofan Huo, Ye Shi, and Jingya Wang. Monocular human-object reconstruction in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 5547–5555, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 4
- [14] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 6
- [15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 4
- [16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4
- [17] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. <https://arxiv.org/abs/2311.06214>, 2023. 2
- [18] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2
- [19] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020.
- [20] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: a knowledge engine foundation for human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2
- [21] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Jingru Tan, Xudong Lu, and Cewu Lu. From isolated islands to pangea: Unifying semantic space for human action understanding. In *CVPR*, 2024. 2
- [22] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2
- [23] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with

- component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 1, 2, 3
- [24] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. 2022. 5
- [25] Junxuan Liu, Xinpengand Liang, Zili Lin, Haowen Hou, Yong-Lu Li, and Cewu Lu. Imdy: Human inverse dynamics from imitated observations. In *ICLR*, 2025. 2
- [26] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [27] Siqi Liu, Yong-Lu Li, Zhou Fang, Xinpeng Liu, Yang You, and Cewu Lu. Primitive-based 3d human-object interaction modelling and programming. In *AAAI*, 2024. 2
- [28] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. Bridging the gap between human motion and action semantics via kinematic phrases. In *ECCV*, 2024. 2
- [29] Xiaoyang Liu, Boran Wen, Xinpeng Liu, Zizheng Zhou, Hongwei Fan, Cewu Lu, Lizhuang Ma, Yulong Chen, and Yong-Lu Li. Interacted object grounding in spatio-temporal human-object interactions. In *AAAI*, 2025. 2
- [30] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2
- [31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2
- [32] 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. 3
- [33] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 2
- [34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [35] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2
- [36] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *CVPR*, 2024. 7
- [37] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3
- [39] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2
- [40] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. 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. 1, 2
- [41] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *2022 International Conference on 3D Vision (3DV)*, pages 353–362, 2022. 2
- [42] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2
- [43] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 2, 3
- [44] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024. 7

- [45] Xinyu Xu, Yizheng Zhang, Yong-Lu Li, Lei Han, and Cewu Lu. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. In *NeurIPS*, 2024. 2
- [46] Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Huang Siyuan. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. *European Conference on Computer Vision*, 2024. 8
- [47] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. *arXiv preprint arXiv:2312.08963*, 2023. 1, 2, 4
- [48] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. *CVPR*, 2024. 3
- [49] Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10411–10421, 2023. 1
- [50] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [51] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6, 7
- [52] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. 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)*, pages 729–741, 2024. 1
- [53] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European conference on computer vision (ECCV)*, 2022. 1, 4
- [54] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Care about you: towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv:1705.09892*, 2017. 2