HOGSA: Bimanual Hand-Object Interaction Understanding with 3D Gaussian Splatting Based Data Augmentation

Wentian Qu[1,2], Jiahe Li[1,2], Jian Cheng[1,2], Jian Shi[3,2], Chenyu Meng[1,2], Cuixia Ma[1,2], Hongan Wang[1,2], Xiaoming Deng[1,2*], Yinda Zhang[4*]

[1]Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Institute of Automation, Chinese Academy of Sciences
[4]Google
{wentian2019, lijiahe2021, chengjian, cuixia, hongan,xiaoming}@iscas.ac.cn,
jian.shi@ia.ac.cn, mengchenyu21@mails.ucas.ac.cn, yindaz@google.com

Abstract

Understanding of bimanual hand-object interaction plays an important role in robotics and virtual reality. However, due to significant occlusions between hands and object as well as the high degree-of-freedom motions, it is challenging to collect and annotate a high-quality, large-scale dataset, which prevents further improvement of bimanual hand-object interaction-related baselines. In this work, we propose a new 3D Gaussian Splatting based data augmentation framework for bimanual hand-object interaction, which is capable of augmenting existing dataset to large-scale photorealistic data with various hand-object pose and viewpoints. First, we use mesh-based 3DGS to model objects and hands, and to deal with the rendering blur problem due to multi-resolution input images used, we design a super-resolution module. Second, we extend the single hand grasping pose optimization module for the bimanual hand object to generate various poses of bimanual hand-object interaction, which can significantly expand the pose distribution of the dataset. Third, we conduct an analysis for the impact of different aspects of the proposed data augmentation on the understanding of the bimanual hand-object interaction. We perform our data augmentation on two benchmarks, H2O and Arctic, and verify that our method can improve the performance of the baselines.

Project Page — https://iscas3dv.github.io/HOGSA/

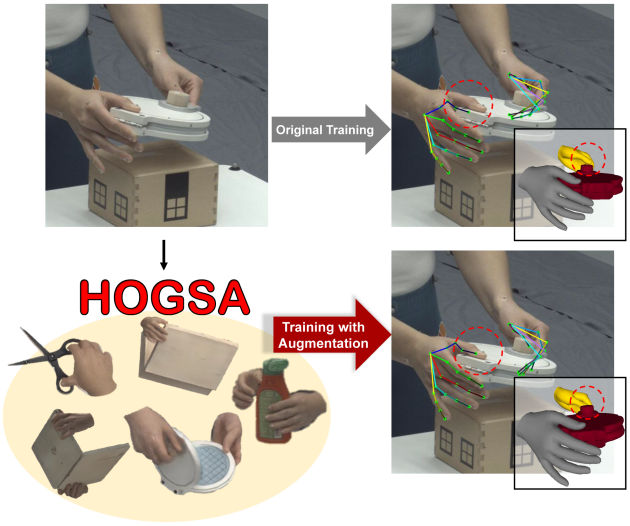

Figure 1: We propose a new 3DGS-based data augmentation framework for bimanual hand-object interaction to augment existing dataset with various hand-object pose and viewpoints. Our method can improve the performance of the baselines, and achieve more accurate pose and contact.

Introduction

Understanding of the bimanual hand-object interaction (Fan et al. 2023; Kwon et al. 2021), especially the estimation of the pose and the contact relationship, plays an increasingly important role in robotics and virtual reality applications. One of the most popular approaches to address this problem is deep learning methods, which require large-scale bimanual hand-object interaction dataset with rich annotations. However, due to significant occlusions and high-degree-of-freedom motions of the interaction, it is still challenging to collect and annotate a high-quality dataset, which prevents further improvement of the task.

To address the challenges of data scarcity and inaccurate 3D annotations, existing work has explored data augmentation methods with synthetic data under conventional rendering pipelines (Corona et al. 2020; Jian et al. 2023; Yang et al. 2022). However, these approaches usually require complex and time-consuming 3D scanning and post-processing to capture high-quality shapes and texture maps of the hand and object, and additional blending weights of hand is necessary for augmentation, which requires substantial expertise (Romero, Tzionas, and Black 2022; Deng et al. 2021). Moreover, capturing a realistic texture map (i.e. subtle details and natural appearance) from observed images is difficult, thus it often results in rendering results that lack realism (Qian et al. 2020). Recently, benefitting from the scene representation ability, neural rendering methods such as NeRF (Mildenhall et al. 2021) and

*indicates corresponding author.

3DGS (Kerbl et al. 2023) enable high-quality data augmentation by synthesizing novel views (Feldmann et al. 2024) or novel hand poses (Qu et al. 2023). NeRFmentation (Feldmann et al. 2024) uses NeRF to perform data augmentation for the monocular depth estimation task in static scenes, but it cannot break the accuracy bottleneck when the scene changes significantly. HO-NeRF (Qu et al. 2023) builds a pose-driven NeRF for hand-object interaction scenarios and demonstrates the potential to generate diverse data, yet it requires offline modeling of hands and objects and a time-costing rendering process, making it infeasible for data augmentation of large-scale dataset. Although these neural rendering methods can support realistic novel view synthesis (Wang et al. 2021) and are potentially useful for data augmentation, they still suffer from image blur due to multi-resolution image input and inaccurate annotations (Yu et al. 2024; Barron et al. 2021). Unrealistic images cannot resolve the gap between real images and synthetic images, which will lead to a degradation of model performance. Another key factor that affects the baseline performance is the diversity of poses (Deng et al. 2021). It is necessary to establish a data augmentation approach of bimanual hand-object interaction that enables efficient rendering, various feasible hand-object poses, and photorealistic rendering images.

In this paper, we propose a 3DGS-based data augmentation framework **H**and-**O**bject **G**aussian **S**platting **A**ugmentation (**HOGSA**) for bimanual hand-object interaction understanding. First, we use mesh-based 3DGS to model the hand and object based on the interaction images, which can efficiently synthesize interaction images with the input hand-object pose and viewpoints. Second, in order to enhance the pose diversity of the dataset, we use the pose optimization module to generate diverse poses of two hands and object to drive the hand-object Gaussian splatting model to render images of novel interaction poses. Third, in order to ensure the realism of the rendered images, we design the super-resolution module to improve the rendering quality of the coarse images generated by 3DGS. Finally, we combine our augmented dataset with the original dataset to refine the baseline of bimanual hand-object interaction, and conduct a systematic analysis of different aspects that affect interaction understanding accuracy in the augmented dataset. We evaluate our method on two main benchmarks H2O (Kwon et al. 2021) and Arctic (Fan et al. 2023), and the baseline performances are improved with our augmented dataset.

The contributions of our method can be summarized as follows: 1) A 3DGS-based data augmentation framework for bimanual hand-object interaction understanding; 2) a super-resolution module and a pose optimization module to improve the realism and pose diversity in the data augmentation; 3) We provide fine-tuning models using our HOGSA that achieve state-of-the-art results on H2O and Arctic benchmarks, and make a systematic analysis of the impact for augmented data on accuracy.

Related Work

Hand-Object Interaction Understanding. Hand-object interaction understanding focuses on 3D reconstruction, pose and contact estimation. Early approaches mainly rely on predefined templates to estimate hand and object poses (Cao et al. 2021; Corona et al. 2020; Liu et al. 2021; Tekin, Bogo, and Pollefeys 2019; Yang et al. 2024; Qi et al. 2024). These template-based methods face significant challenges in realism gap with real-world. Recent advancements have transitioned towards template-free approaches that leverage large-scale 3D hand-object interaction datasets (Chen et al. 2023; Ye, Gupta, and Tulsiani 2022; Fan et al. 2024; Zhang et al. 2024a). Nonetheless, the limited diversity and quantity of 3D data restricts models' ability to generalize effectively across various scenarios. Several datasets with hand-object contact annotations have advanced contact estimation (Taheri et al. 2020; Brahmbhatt et al. 2020; Fan et al. 2023; Grady et al. 2021). Some studies (Narasimhaswamy, Nguyen, and Nguyen 2020; Shan et al. 2020) infer 2D bounding boxes of hands with contact from RGB images, while others (Rogez, Supancic, and Ramanan 2015; Fan et al. 2023) explore 3D contact inference. These data-driven approaches are suffer from the diverse poses and viewpoints in dataset. In this paper, we tackle with the limitation of data collection, and propose a 3DGS-based data augmentation framework to enhance the performance of the baselines.

Data Augmentation for Hand-Object Interaction. Data augmentation is essential for enhancing model performance in hand-object interaction understanding. Existing methods can be divided into rendering-based and generative-based methods. Rendering-based methods generate plausible hand-object interaction images by designing specific pipelines (Corona et al. 2020; Jian et al. 2023; Yang et al. 2022; Gao et al. 2022; Li et al. 2023) or utilizing off-the-shelf tools such as GraspTTA (Jiang et al. 2021) and UniDexGrasp (Xu et al. 2023). However, the rendering-based methods often lack realism. Generative-based methods will create more realistic hand-object interaction images. HOGAN (Hu et al. 2022) synthesizes novel views using target poses as guidance. Several approaches employ conditional diffusion models to generate hand grasping. Affordance Diffusion (Ye et al. 2023) generates hand-object interaction images, conditioned on a hand orientation mask. HandBooster (Xu et al. 2024) and HOIDiffusion (Zhang et al. 2024b) synthesizes realistic hand-object images with diverse appearances, poses, views, and backgrounds. MANUS (Pokhariya et al. 2023) utilizes 3DGS to model hand and object respectively and combine them to form a data set. However, the augmentation approaches are mostly focus on interacting between single hand and object. These setups are more serious mutual occlusion when migrating to bimanual hand-object interacting tasks, making the model difficult to learn effective pose priors.

In this work, to achieve high realism of augumented data, we enforce the generated bimanual hand-object interaction results to meet geometric constraints and overcome the realism issues of 3DGS, which enables our data augmentation method to improve the performance of the baseline.

Method

In this work, we propose a data augmentation framework for bimanual hand-object interaction to improve the accuracy of

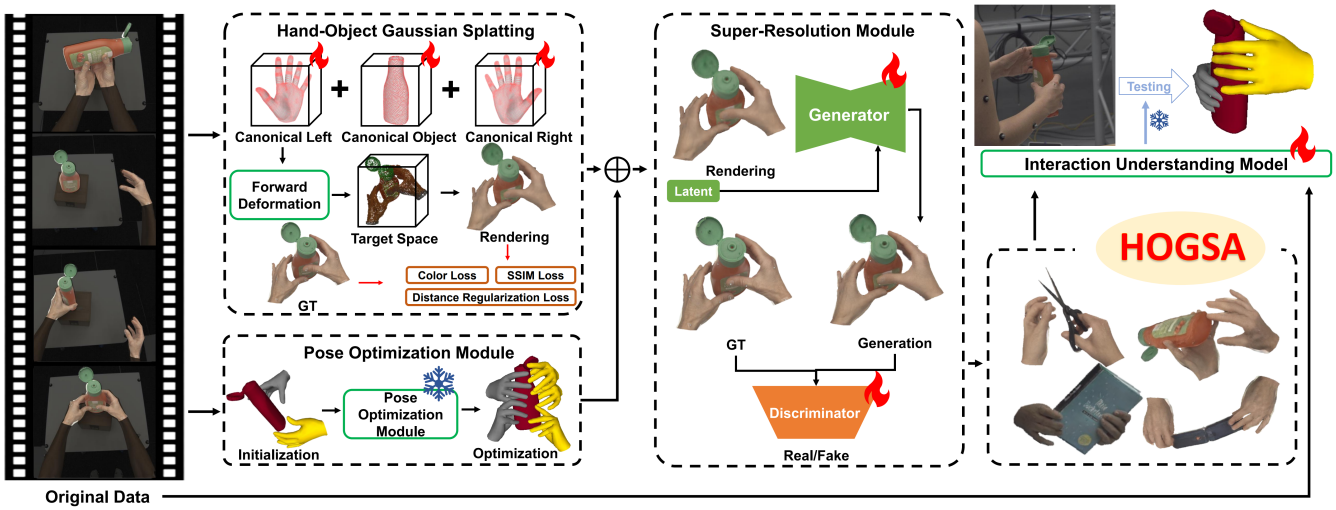

Figure 2: Overview of our data augmentation framework for bimanual hand-object interaction. Based on the original dataset, we first establish mesh-based 3DGS models and input the original poses to pose optimization module to expand the diversity of interaction. The novel pose and 3DGS can be combined to render the low-quality image, which is then fed into the super-resolution module to further enhance the realism. Based on the above modules, we can automatically build an expanded dataset and support model fine-tuning for the interaction understanding baseline to improve performance.

the baselines for interaction pose estimation (Fig. 2). Our data augmentation method HOGSA contains diverse poses and realistic rendering images, and can achieve end-to-end generation based on novel hand-object and camera poses. We introduce the pipeline of our data augmentation in first section and detail the implementation in the resting sections. We first use the mesh-based 3DGS method to model the left hand, right hand and object models based on the input hand-object interaction images respectively. Then we use the Pose Optimization Module (POM) to optimize the novel poses of the bimanual hand-object interaction, ensuring the realism of the grasp while expanding the diversity of the pose. Combining Hand-Object Gaussian Splatting (HOGS) and Pose Optimization Module, we render low-quality images, and design the Super-Resolution Module (SRM) to improve the rendering quality. Finally, we use both the augmentation and the original dataset to refine the baseline to improve the accuracy of the interaction pose estimation task.

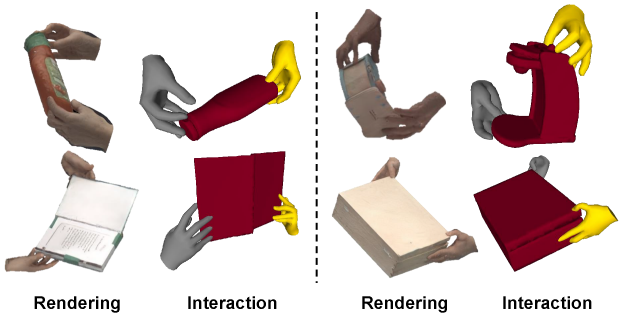

Figure 3: Examples of our HOGSA, which contains diverse interactive poses and ensures the realism of the images.

Hand-Object Gaussian Splatting Augmentation

Based on the original dataset, we first build a HOGS model for each sequence. Second, based on the original poses, we use POM to optimize and expand the diversity of poses of the bimanual hand-object interaction. Then we use the HOGS model with these novel poses of hand-object interaction and camera poses to render new images as our new data augmentation process, and feed the new image into SRM to improve the image quality. The above three modules provide an image rendering pipeline, pose diversity, and rendering realism, thus achieving an efficient data augmentation solution. We use the bimanual hand-object interaction dataset Arctic (Fan et al. 2023) and H2O (Kwon et al. 2021) to generate synthetic data HOGSA respectively, and the results are shown in Fig. 3. To organize the data used to train the baseline, we are inspired by GANerated Hands (Mueller et al. 2018) and add COCO2017 (Lin et al. 2014) images as background to fuse the generated image, and crop it to 224×224 resolution (as shown in Fig. 4). Finally, we combine HOGSA and the original data to improve the performance of the baseline.

Hand-Object Gaussian Splatting

Our method utilizes color images as input to represent the complex dynamic scene using mesh-based 3D Gaussian Splatting, accomplishing rendering under novel poses and novel views. For initialization, we define the Gaussian kernel on MANO-HD (Chen, Wang, and Shum 2023) template hand mesh and object mesh, respectively, which are named canonical space. The standard 3D Gaussian Splatting (Kerbl et al. 2023) can be defined as:

$$\mathbf{G}(\mathbf{x}) = e^{-(\mathbf{x}-\mathbf{x}_c)^\top \Sigma^{-1} (\mathbf{x}-\mathbf{x}_c)}, \quad (1)$$

Figure 4: The augmented data we used to train the baseline. Compared with the original data, our images ensure realism and have various poses.

where $\mathbf{x}_c \in \mathbb{R}^3$ represents the Gaussian center and $\Sigma \in \mathbb{R}^{3\times3}$ represents the covariance matrix, which is parameterized by rotation matrix $\mathbf{R}$ and scaling matrix $\mathbf{S}$ as $\Sigma = \mathbf{RSS}^\top\mathbf{R}^\top$. Inspired by GaMeS (Waczyńska et al. 2024), we convert the Gaussian kernels into mesh surfaces and the Gaussian center can be defined as:

$$\mathbf{x}_c = \beta\mathbf{V} = \beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2 + \beta_3\mathbf{v}_3, \tag{2}$$

where $\beta = \{\beta_i\}_{i=1}^3$ are the trainable parameters that satisfies $\beta_1 + \beta_2 + \beta_3 = 1$ and $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^3$ are the vertices from the mesh face. For the definition of the rotation matrix $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^3$, $\mathbf{r}_1$ is the surface normal, $\mathbf{r}_2$ is the vector from the center of the surface $\mathbf{m} = mean(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ to the vertex $\mathbf{v}_1$, and $\mathbf{r}_3$ is obtained by orthonormalizing the vector for $\mathbf{r}_1, \mathbf{r}_2$. For the scaling matrix $\mathbf{S} = \{s_i\}_{i=1}^3$, we define $s_1 = s_2 = \|\mathbf{m} - \mathbf{v}_2\|, s_3 = \langle \mathbf{v}_2, \mathbf{R}_3 \rangle$.

Hand Model. To model the hand motion, we use the blending weights from MANO-HD to initialize the hand Gaussian skinning weights $\mathbf{w}$, where $\mathbf{w} = (w_1, w_2, ..., w_n)$ with n joints. For the target frame, we use bone transformation matrix $\mathbf{B}$ and blending weights to deform the Gaussian kernels in canonical space to the target space as:

$$\mathbf{x}_t^h = (\sum_{i=1}^{n} w_i\mathbf{B}_i) \cdot \mathbf{x}_c^h, \tag{3}$$

where $\mathbf{x}_c^h$ is the hand canonical space point, $\mathbf{x}_t^h$ is the corresponding point in target space.

Object Model. We use 6D rigid pose $\mathbf{T} \in SE(3)$ to perform rigid transformation from canonical space to target space:

$$\mathbf{x}_t^o = \mathbf{T} \times \mathbf{x}_c^o, \tag{4}$$

where $\mathbf{x}_c^o$ is the object canonical space point, $\mathbf{x}_t^o$ represents the mapping point in target space.

Alpha Blending. The 3D Gaussian points in target space are projected to 2D Gaussian, then sorted based on depth and the color value of the pixel can be calculated by:

$$\mathbf{c} = \sum_{i=1}^{N} \mathbf{c}_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{5}$$

where $\mathbf{c}$ represents the color value of the pixel, $\mathbf{c}_i$ and α_i represent the color and pixel translucency of the i-th Gaussian kernel respectively.

Training. For the Gaussian kernels defined in left-hand, right-hand and object model, the optimized parameters include mesh vertex positions $\mathbf{V}$, patch weights α, scale factor s, spherical harmonics coefficients and opacity. The loss function of our HOGS model can be expressed as:

$$L_{HOGS} = (1-\lambda_{SSIM})L_1 + \lambda_{SSIM}L_{SSIM} + \lambda_R L_R, \tag{6}$$

where L1 loss L_1 and SSIM loss L_{SSIM} are used to minimizes the difference between the input and rendered images, the distance regularization loss L_R constrains the vertices in the canonical space not to deviate far from the initial mesh, $\lambda_{SSIM}, \lambda_R$ are hyperparameters.

We split the datasets for training and testing our HOGS. For Arctic, we select subjects except '$s03$' and '$s05$' to build the HOGS model to meet the data division of the interaction understanding task. We select the sequences '$grab\_01$' and 'use_01' to train HOGS, and cropped the original images to a resolution of 1400×1000. For H2O, we select subjects except '$subject4$' to build the HOGS model. We select sequences except '$o2$' scene for training HOGS and keep the original image resolution 1280×720.

Pose Optimization Module

In order to ensure the diversity of pose in augmentation, we need to synthesize more bimanual hand-object interaction results that meet geometric constraints. Inspired by GraspTTA (Jiang et al. 2021) with a fitting strategy, we extend their solution of single-hand grasping to bimanual hand object interaction. We discard the 'GraspCVAE' model used to generate the initial hand pose and only fixed the parameters of the 'ContactNet' model for contact map inference. We randomly perturb the hands extracted from the original dataset a small distance away from the object as the initial pose. Based on the point cloud of the initial pose, a contact map Ω can be calculated. At the same time, the point cloud can be input into 'ContactNet' to predict a contact map $\hat{\Omega}$ under prior knowledge. We leverage a self-supervised consistency loss $L_C = \|\Omega - \hat{\Omega}\|_2^2$ to enforce a reasonable contact. We also use hand-centric loss L_H and penetration loss L_P to ensure the physically plausible interaction. The loss function is formulated as follows:

$$L_{POM} = \sum_{i\in\{l,r\}} (\lambda_C L_C^i + \lambda_H L_H^i + \lambda_P L_P^i), \tag{7}$$

where l and r represent the left and right hands, $\lambda_C, \lambda_H, \lambda_P$ are hyperparameters. For an initial input, we fix the object pose and optimize the pose of both hands separately with 200 iterations to get the optimized pose. Then we input the optimized pose and camera parameters into HOGS model to synthesize a bimanual hand-object interaction image.

For the initial input pose, we generate a transformation matrix with random rotation between $[0, 20^\circ]$ around the x, y, and z axis and apply the rotation to the original pose. We then perturb the distance by which the hands are farther away from the object. We calculate the distance and direction of the hand root joint relative to the object, then translate it 5% of the relative distance away from the object, and then add a perturbation of $[0, 6cm]$ to the position of the hand and object, respectively. Note that for the Arctic dataset, we impose an angle of $[0.01\pi, 0.2\pi]$ on the one-dimensional rotation of the articulated object.

Super-Resolution Module

We notice that the different resolutions of the images and the deviation in pose annotation to train 3DGS can result in blurry rendered images with artifacts to train 3DGS . Therefore, we use CNN to improve the rendering quality in 3DGS. Inspired by StyleAvatar (Wang et al. 2023), we can use the encoder-decoder backbone to learn the local and global features of the image and integrate it into the GAN framework to improve the rendering quality. We input the coarse image $\mathbf{I}_C$ rendered by 3DGS into 'StyleUNet' to obtain the refined image $\mathbf{I}_R$ and use the ground truth image to constrain its generation quality. The loss function is defined as follows:

$$L_{SRM} = \lambda_1 L_1 + \lambda_{VGG} L_{VGG} + L_{GAN}, \tag{8}$$

where L_1 represents L1 loss, L_{VGG} represents VGG loss, L_{GAN} represents the loss used in adversarial learning, and λ_1, λ_{VGG} are hyperparameters. In this module, we follow the data split used in HOGS. We leverage HOGS to render images based on the training set and then pair them with the ground truth to train the SRM model.

Experiment

Datasets, Baseline and Evaluation Metric

Datasets. We evaluate our method on Arctic (Fan et al. 2023) and H2O (Kwon et al. 2021). Arctic is a dataset for dexterous bimanual hand-object interaction. It consists of 10 humans, with manipulating 11 articulated objects. We follow allocentric validation split in Arctic to train and test the baseline. H2O is a dataset for two hands manipulating objects. It consists of 4 subjects, with 6 scenarios manipulating 8 rigid objects. We select the first 3 subjects for augmentation and baseline training, and use the last subject for testing.
Augmentation. In the construction of HOGSA for each benchmark, we consider the interaction of the same subject with the same object as a sequence. For the Arctic dataset, we build a total of 82 HOGS models and generate 1.7M images using the novel poses. Compared to the original 1.5M training images, we have automatically expanded the data by almost 113%. For the original H2O dataset with 0.4M training data, we generate 0.7M images with 24 HOGS models, which has automatically expanded the data by almost 175%.
Baseline. *Interaction Understanding*: We follow Arctic (Fan et al. 2023) to define two tasks including 'Consistent Motion Reconstruction' and 'Interaction Field Estimation', and select 'ArcticNet-SF' and 'InterField-SF' in Arctic as the baselines, respectively. *Augmentation*: We compare with HOIDiffusion (Zhang et al. 2024b) ('Arctic+HOID') and replace the mesh-based 3DGS in the HOGS module with original 3DGS (Kerbl et al. 2023) ('Arctic+3DGS').
Evaluation Metric. To evaluate the quality of the bimanual hand-object interaction understanding, we follow Arctic (Fan et al. 2023) to use Contact Deviation (CDev, mm), Motion Deviation (MDev, mm) and Acceleration Error (ACC, m/s^2) to measure the accurate hand-object contact, stable move, and smooth motion, respectively. We use the mean per-joint position error (MPJPE, mm), the average articulation error (AAE, n°), the success rate (SR, $n\%$) to measure the pose accuracy, use the mean relative-root position error (MRRPE, mm) to calculate hand-object relative translations and use average distance error to measure the interaction field. To evaluate the quality of novel view synthesis in ablation study, we follow 3DGS (Kerbl et al. 2023) to adopt PSNR, SSIM and LIPIPS as metrics.
Implementation Details Our method is trained on a Nvidia RTX4090 GPU. We train HOGS and SRM modules from scratch. We train HOGS with 50,000 iterations, which costs an average of 10 GB of memory. We train SRM with a total of 150,000 iterations and cost 8 GB of memory. For the POM, each initial input needs to be iterated 200 times. We set hyperparameters $\lambda_{SSIM}, \lambda_R, \lambda_C, \lambda_H, \lambda_P, \lambda_1, \lambda_{VGG}$ to 0.2, 0.5, 1, 1, 17, 5, 0.03 respectively.

Comparison with Baseline

Consistent Motion Reconstruction. We evaluate this task over the baseline 'ArcticNet-SF' on Arctic and H2O benchmarks, and the results are shown in Tab. 1. After refining the baseline by adding our expanded dataset to the original training set, we find that all metrics are improved. HOIDiffusion deals with the data augmentation for right hand-object interacting scene, and some augmentation images are ambiguous, which will make the baseline difficult to learn effective pose priors under severe occlusion caused by the interaction between two hands and objects, and slightly improves the baseline. Our method ensures the diversity of bimanual hand-object pose as well as realism of rendered images, and can effectively improve the performance of baseline. Our method takes 0.06s to generate an augmented image, which is also more efficient than HOIDiffusion's 4.5s. We show qualitative results in Fig. 5 and find that after adding our data augmentation, the estimation of novel poses and contacts near occlusions are more accurate. This is because our augmented dataset provides richer viewpoint and pose priors. Compared with the original point cloud based 3DGS, our mesh-based 3DGS avoid losing valid information due to points straying too far from the instance based on mesh constraints, which results in higher rendering quality. The original 3DGS rendered images have serious artifacts, which reduces the effect of data augmentation.
Interaction Field Estimation. The interaction field estimation measures the relative spatial relations between hands and the object. We show the results using the 'InterField-SF' in Fig. 5 (b) and Tab. 2. After adding augmentation, the baseline is less affected by occlusion, and the rich pose prior makes its estimation more accurate. The baseline refined by HOGSA promotes stable mutual movement between the

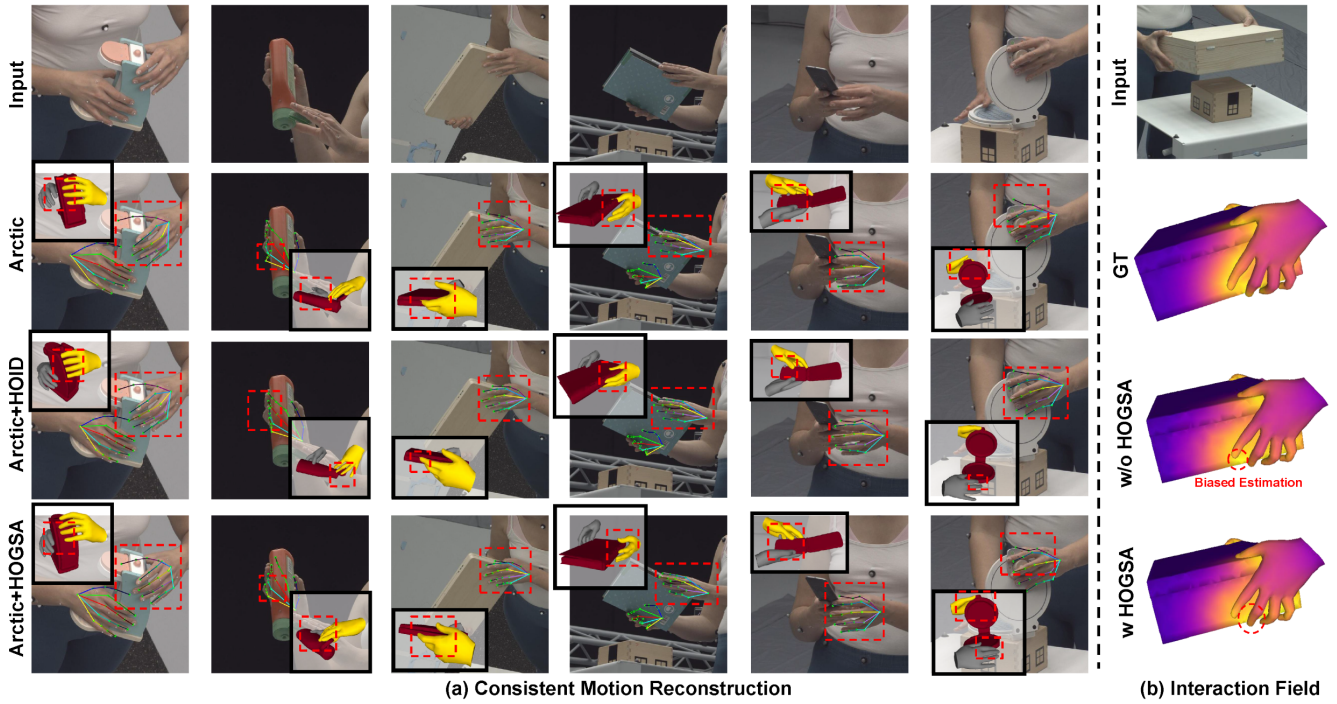

Figure 5: Qualitative results of our data augmentation method HOGSA on the baseline. After optimization, the model can cover a wider range of interactive poses and achieve a more accurate estimation of the pose and contact area.

		Contact and Relative Position		Motion		Hand	Object	
Method	Dataset	CDev$_{ho}$ ↓	MRRPE$_{rl/ro}$ ↓	MDev$_{ho}$ ↓	ACC$_{h/o}$ ↓	MPJPE$_{h}$ ↓	AAE ↓	SR↑
ArcticNet-SF	Arctic	41.35	50.14/37.59	10.46	6.63/8.80	23.01	5.85	71.77
	Arctic+HOID	39.19	46.57/36.93	9.75	**6.36**/8.38	22.44	5.88	73.56
	Arctic+3DGS	36.07	45.17/34.68	8.89	6.41/7.59	21.78	5.76	77.06
	Arctic+HOGSA	**35.23**	**43.69/33.48**	**8.77**	**6.36/7.54**	**20.96**	**5.67**	**77.85**
	H2O	35.74	57.65/47.53	5.31	3.80/6.05	34.38	-	39.80
	H2O+HOGSA	**32.27**	**54.63/43.93**	**5.16**	**3.79/6.02**	**32.24**	-	**45.27**

Table 1: Quantitative results of our data augmentation method on the baseline. Our data augmentation method can be automatically applied to different datasets and support improving the performance of different baselines.

Dataset	HOGSA	Average Distance Error ↓	ACC ↓
Arctic	w/o	9.63/9.91	3.01/2.95
	w	**9.30/8.98**	**2.98/2.79**
H2O	w/o	7.75/10.86	1.84/1.89
	w	**7.54/9.87**	**1.82/1.82**

Table 2: After the interaction field estimation baseline is fine-tuned using HOGSA, the performance is improved.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Arctic	w/o SRM	35.07	**0.988**	0.0194
	w SRM	**36.53**	**0.988**	**0.0163**
H2O	w/o SRM	32.32	**0.987**	0.0133
	w SRM	**32.44**	0.986	**0.0117**

Table 3: Our SRM explores the combination of 3DGS and CNN which aggregates pixel and image local semantic information to further improve the realism of image.

hand and the object and effectively reduces pose jitter. Based on the various pose and viewpoints of our augmentation, the occluded right hand position can predict a more accurate contact area without noise (Fig. 5 (b)). Therefore, our method can be applied to the baseline of various interaction understanding tasks to improve performance.

Ablation Study

Effect of Super-Resolution Module. We compare the rendering quality of images generated by HOGS with and without SRM. The results are shown in Fig. 6 and Tab. 3. The rendering quality is significantly improved after adding SRM, especially in the texture details of objects and the wrinkles on the hands. The local semantic feature infor-

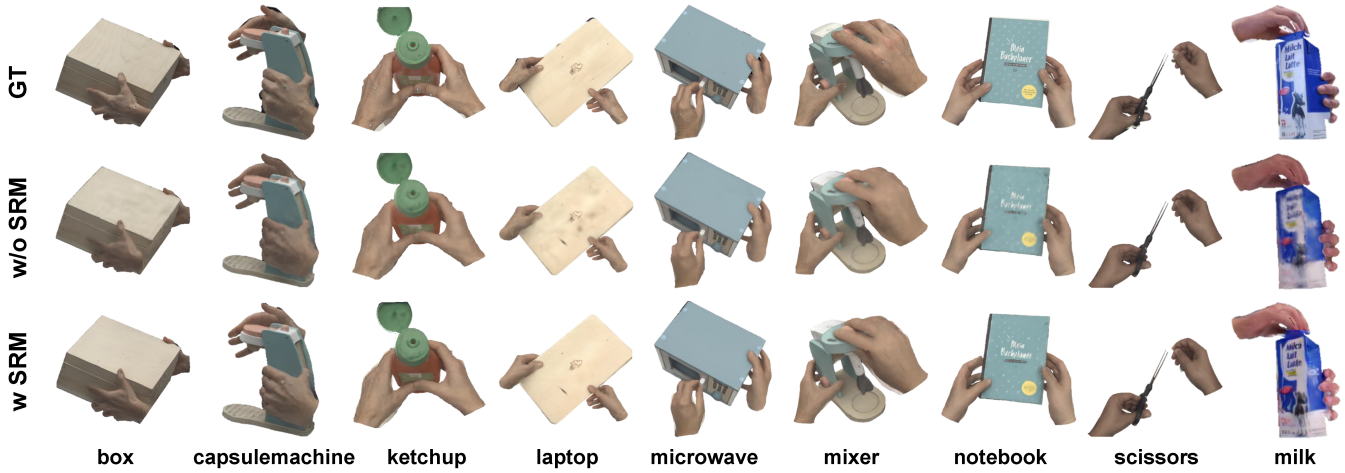

Figure 6: Ablation study on SRM. After using SRM, the realism of the image rendered by HOGS is significantly improved, especially the texture details of the object. This greatly reduces the gap between synthetic and real data.

		Contact and Relative Position		Motion		Hand	Object	
Dataset	Method	CDev$_{ho}$ ↓	MRRPE$_{rl/ro}$ ↓	MDev$_{ho}$ ↓	ACC$_{h/o}$ ↓	MPJPE$_{h}$ ↓	AAE ↓	SR↑
Arctic	w/o SRM	36.85	45.40/34.97	9.39	6.51/8.05	21.75	5.93	75.56
	w/o POM	37.67	45.35/35.78	9.22	6.42/7.90	21.95	6.08	75.10
	Full	**35.23**	**43.69/33.48**	**8.77**	**6.36/7.54**	**20.96**	**5.67**	**77.85**

Table 4: Ablation study on our SRM and POM used for motion reconstruction. We evaluate our method using 'ArcticNet-SF' baseline on Arctic dataset, and the expanded dataset can effectively improve the accuracy of the baseline.

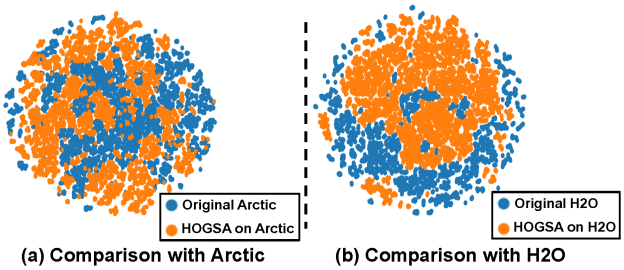

Figure 7: We used the joints of both hands to perform T-SNE and evaluate the distribution of poses before and after data augmentation. The poses of our augmented dataset can further enrich the diversity of the original data.

mation learned by CNN compensates for artifacts and blur caused by images of different resolutions and pose bias. It is noted that the SSIM is comparable to the method without SRM. The reason is that the geometry of the image is mainly affected by HOGS, while SRM mainly improves the texture quality of the image. We compare the impact of augmented datasets with different rendering qualities on the accuracy of consistent motion reconstruction, and the results are shown in the first and third rows of Tab. 4. The SRM further reduces the gap between real and synthetic data, and more realistic images can better improve the performance of the baseline.

Effect of Pose Optimization Module. We compare the pose diversity in the original data and the data enhanced by POM. We encode the hand joints and then use the T-SNE clustering (Van der Maaten and Hinton 2008) to show in Fig. 7. After adding the POM module, the orange points distributed around the blue points complement the diversity of the entire data pose distribution, which shows a reinforcement of the original data. In order to evaluate the effect of POM, we remove the POM to generate augmented data (w/o POM) and combine them with original data to train the baseline, and the results are shown in the second and third rows of Tab. 4. We find that while perspective augmentation improves the performance of the model, the prediction accuracy of novel poses is still not significant. Since the POM can enhance the distribution of poses, it can improve the interaction understanding ability of the model.

Conclusion

In this work, we propose a new 3DGS-based data augmentation framework for bimanual hand-object interaction, which can augment the existing benchmark to large-scale photorealistic data with various hand-object pose and viewpoints. The expanded dataset can further improve the performance of the existing baseline. We propose the pose optimization module to generate various physically feasible poses of bimanual hand-object interaction and the super-resolution module to improve the realism of rendered images using 3DGS. We perform our data augmentation on two benchmarks, H2O and Arctic, and verify that our method can improve the performance of the baselines.

Acknowledgments

This work was supported in part by National Science and Technology Major Project (2022ZD0119404), National Natural Science Foundation of China (62473356,62373061), Beijing Natural Science Foundation (L232028), CAS Major Project (RCJJ-145-24-14), Science and Technology Innovation Key R&D Program of Chongqing (CSTB2023TIAD-STX0027), and Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support No. ZLRK202330.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5855–5864.

Brahmbhatt, S.; Tang, C.; Twigg, C. D.; Kemp, C. C.; and Hays, J. 2020. ContactPose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 361–378. Springer.

Cao, Z.; Radosavovic, I.; Kanazawa, A.; and Malik, J. 2021. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12417–12426.

Chen, X.; Wang, B.; and Shum, H.-Y. 2023. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8683–8693.

Chen, Z.; Chen, S.; Schmid, C.; and Laptev, I. 2023. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12890–12900.

Corona, E.; Pumarola, A.; Alenya, G.; Moreno-Noguer, F.; and Rogez, G. 2020. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5031–5041.

Deng, X.; Zhang, Y.; Shi, J.; Zhu, Y.; Cheng, D.; Zuo, D.; Cui, Z.; Tan, P.; Chang, L.; and Wang, H. 2021. Hand pose understanding with large-scale photo-realistic rendering dataset. *IEEE Transactions on Image Processing*, 30: 4275–4290.

Fan, Z.; Parelli, M.; Kadoglou, M. E.; Chen, X.; Kocabas, M.; Black, M. J.; and Hilliges, O. 2024. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 494–504.

Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12943–12954.

Feldmann, C.; Siegenheim, N.; Hars, N.; Rabuzin, L.; Ertugrul, M.; Wolfart, L.; Pollefeys, M.; Bauer, Z.; and Oswald, M. R. 2024. NeRFmentation: NeRF-based Augmentation for Monocular Depth Estimation. *arXiv preprint arXiv:2401.03771*.

Gao, D.; Xiu, Y.; Li, K.; Yang, L.; Wang, F.; Zhang, P.; Zhang, B.; Lu, C.; and Tan, P. 2022. DART: Articulated hand model with diverse accessories and rich textures. *Advances in Neural Information Processing Systems*, 35: 37055–37067.

Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1471–1481.

Hu, H.; Wang, W.; Zhou, W.; and Li, H. 2022. Hand-object interaction image generation. *Advances in Neural Information Processing Systems*, 35: 23805–23817.

Jian, J.; Liu, X.; Li, M.; Hu, R.; and Liu, J. 2023. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14713–14724.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11107–11116.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kwon, T.; Tekin, B.; Stühmer, J.; Bogo, F.; and Pollefeys, M. 2021. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10138–10148.

Li, K.; Yang, L.; Zhen, H.; Lin, Z.; Zhan, X.; Zhong, L.; Xu, J.; Wu, K.; and Lu, C. 2023. Chord: Category-level hand-held object reconstruction via shape deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9444–9454.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Jiang, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14687–14697.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; and Theobalt, C. 2018. Ganerated hands

for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 49–59.

Narasimhaswamy, S.; Nguyen, T.; and Nguyen, M. H. 2020. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 33: 7841–7851.

Pokhariya, C.; Shah, I. N.; Xing, A.; Li, Z.; Chen, K.; Sharma, A.; and Sridhar, S. 2023. MANUS: Markerless Grasp Capture using Articulated 3D Gaussians. *arXiv preprint arXiv:2312.02137*.

Qi, H.; Zhao, C.; Salzmann, M.; and Mathis, A. 2024. HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields: Processed data and trained models. In *The 2024 IEEE/CVF Computer Vision and Pattern Recognition Conference*. Zenodo.

Qian, N.; Wang, J.; Mueller, F.; Bernard, F.; Golyanik, V.; and Theobalt, C. 2020. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 54–71. Springer.

Qu, W.; Cui, Z.; Zhang, Y.; Meng, C.; Ma, C.; Deng, X.; and Wang, H. 2023. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15100–15111.

Rogez, G.; Supancic, J. S.; and Ramanan, D. 2015. Understanding everyday hands in action from rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, 3889–3897.

Romero, J.; Tzionas, D.; and Black, M. J. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

Shan, D.; Geng, J.; Shu, M.; and Fouhey, D. F. 2020. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9869–9878.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 581–600. Springer.

Tekin, B.; Bogo, F.; and Pollefeys, M. 2019. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4511–4520.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Waczyńska, J.; Borycki, P.; Tadeja, S.; Tabor, J.; and Spurek, P. 2024. Games: Mesh-based adapting and modification of gaussian splatting. *arXiv preprint arXiv:2402.01459*.

Wang, L.; Zhao, X.; Sun, J.; Zhang, Y.; Zhang, H.; Yu, T.; and Liu, Y. 2023. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–10.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Xu, H.; Li, H.; Wang, Y.; Liu, S.; and Fu, C.-W. 2024. HandBooster: Boosting 3D Hand-Mesh Reconstruction by Conditional Synthesis and Sampling of Hand-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10159–10169.

Xu, Y.; Wan, W.; Zhang, J.; Liu, H.; Shan, Z.; Shen, H.; Wang, R.; Geng, H.; Weng, Y.; Chen, J.; et al. 2023. UniDexGrasp: Universal Robotic Dexterous Grasping via Learning Diverse Proposal Generation and Goal-Conditioned Policy. *arXiv preprint arXiv:2303.00938*.

Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2750–2760.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Zhang, J.; Li, J.; and Lu, C. 2024. Learning a contact potential field for modeling the hand-object interaction. *IEEE transactions on pattern analysis and machine intelligence*.

Ye, Y.; Gupta, A.; and Tulsiani, S. 2022. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3895–3905.

Ye, Y.; Li, X.; Gupta, A.; De Mello, S.; Birchfield, S.; Song, J.; Tulsiani, S.; and Liu, S. 2023. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22479–22489.

Yu, Z.; Chen, A.; Huang, B.; Sattler, T.; and Geiger, A. 2024. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19447–19456.

Zhang, C.; Jiao, G.; Di, Y.; Wang, G.; Huang, Z.; Zhang, R.; Manhardt, F.; Fu, B.; Tombari, F.; and Ji, X. 2024a. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9992–10002.

Zhang, M.; Fu, Y.; Ding, Z.; Liu, S.; Tu, Z.; and Wang, X. 2024b. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8521–8531.