CHORUS: Learning Canonicalized 3D Human-Object Spatial Relations from Unbounded Synthesized Images

Sookwan Han   Hanbyul Joo
Seoul National University
{jellyheadandrew, hbjoo}@snu.ac.kr

Figure 1. Given an object category, our method, **CHORUS**, automatically learns human-object 3D spatial arrangements in a self-supervised way. Our method leverages a generative model to synthesize an unbounded number of images to reason about the 3D spatial relationship. As output, CHORUS produces the 3D spatial distribution in the canonical space, which can be deformed for any posed space.

Abstract

We present a method for teaching machines to understand and model the underlying spatial common sense of diverse human-object interactions in 3D in a self-supervised way. This is a challenging task, as there exist specific manifolds of the interactions that can be considered human-like and natural, but the human pose and the geometry of objects can vary even for similar interactions. Such diversity makes the annotating task of 3D interactions difficult and hard to scale, which limits the potential to reason about that in a supervised way. One way of learning the 3D spatial relationship between humans and objects during interaction is by showing multiple 2D images captured from different viewpoints when humans interact with the same type of objects. The core idea of our method is to leverage a generative model that produces high-quality 2D images from an arbitrary text prompt input as an "unbounded" data generator with effective controllability and view diversity. Despite its imperfection of the image quality over real images, we demonstrate that the synthesized images are sufficient to learn the 3D human-object spatial relations. We present multiple strategies to leverage the synthesized images, including (1) the first method to leverage a generative image model for 3D human-object spatial relation learning; (2) a framework to reason about the 3D spatial relations from inconsistent 2D cues in a self-supervised manner via 3D occupancy reasoning with pose canonicalization; (3) semantic clustering to disambiguate different types of interactions with the same object types; and (4) a novel metric to assess the quality of 3D spatial learning of interaction.

1. Introduction

Humans interact with objects in specific ways. We wear shoes on our feet, a hat on our heads, and ride a bike by holding handles and putting two feet on the pedals. While this common sense regarding the 3D arrangements of the way we interact with objects is known to us, teaching such things to robots and machines is challenging, requiring numerous variations in diverse human-object interactions in 3D.

As providing 3D supervision by manually annotating various cases is hard to scale, an alternative way of teaching such things is by showing 2D photos from the Internet containing the interaction with the same object from many viewpoints. A text-based image retrieval (e.g., Google Image Search) can be an option to crawl many images from a text description similar to NEIL [8]. However, this approach fundamentally suffers from several obstacles such as (other than the challenges in 3D spatial relation reasoning): (1) viewpoint variations are insufficient and hard to control; (2) the number of related images decreases drastically as the compositionality of the prompt increases; and (3) the images are often biased due to commercial websites.

Project Page: https://jellyheadandrew.github.io/projects/chorus

Figure 2. **Internet-crawled images vs. synthesized images.** Under the prompt "A person is riding a bicycle, top view", synthesized images exhibit superior fidelity to the intended "top view" perspective compared to internet-crawled images. The correct viewpoint image is indicated by the green bounding box.

In this paper, we present a novel idea of leveraging a text-conditional generative model [52] as a controllable data producer in synthesizing "unbounded", "multi-view", "diverse" images to learn the 3D human-object spatial relations in a self-supervised manner. Despite its imperfectness in quality, we observe that the synthesized images from generative models are more suitable for our objective, as the generative model effectively links desired semantics of human-object interaction (HOI) described in natural language. Our synthesis-based approach allows better controllability in obtaining images for spatial relation learning, providing more relevant images from diverse viewpoints. See the examples in Fig. 2.

Nonetheless, inferring 3D spatial knowledge in human-object interaction from in-the-wild 2D image collections is still a non-trivial problem due to the inconsistency and "wildness" among synthesized images. In particular, a method should handle the following challenges: (1) Semantic variation: given a category, there can be various semantic situations of human-object interaction; thus, the spatial distribution of object may vary significantly as semantics vary; (2) Human pose variation: even assuming the same semantic situation, human postures can vary as diverse actions and pose are available in the same situation; (3) Intra-class variation: object can exist in various forms (even if same category); and (4) Visual variance: when learning from 2D cues, the visual properties (e.g., illumination, camera) may vary for even the same 3D arrangement, making it difficult to localize and extract 2D cues.

To this end, we present a self-supervised method to learn the spatial common sense of diverse human-object interactions in 3D for arbitrary object categories without any human annotations. We present several novel components to achieve the goal, including (1) automatic prompt generation for diverse view and semantic variations via chatGPT [45], (2) outlier filtering strategies, (3) automatic camera view calibrations by using a human as an anchor, (4) accumulating spatial interaction cues from inconsistent multi-view 2D knowledge with varying human pose, object geometry, and (5) clustering for various semantic human-object interaction types. The output of our method can be considered as *possible occupancy distributions* of the category-specified 3D object

relative to the human body in a canonical person-centric space regarding the intra-class object variation. We demonstrate the efficacy of our method on various object categories and human-object interaction types, as shown in Fig. 1 and Fig. 7. As the first in this direction, we also introduce a new metric, namely *Projective Average Precision*, to quantify the quality of 3D spatial inference outputs. Our contributions are summarized as follows: (1) the first method to leverage a generative text/image model for 3D human-object spatial relation learning, including automatic prompt generation, outlier filtering, and 3D viewpoint estimation via estimated 3D humans; (2) a framework to reason about the 3D spatial relations from inconsistent 2D cues in a self-supervised manner via 3D occupancy reasoning with pose canonicalization; (3) semantic clustering to disambiguate different types of interactions with the same object types; (4) a novel metric to assess the quality of 3D spatial learning of interaction.

2. Related Work

Text-to-Image Synthesis with Diffusion Models. Recent approaches in text-to-image synthesis show great performance by leveraging emerging diffusion models [31, 44, 52, 58–61]. Diffusion models are a class of generative models that "noises" the data from training distribution and learn to "denoise" the perturbed images at arbitrary noise scale. Diffusion models are likelihood-based generative models and are known to show high mode coverage and generate high-fidelity images. One drawback is the slow inference time due to multiple iterations of the denoising process, which can be mitigated by recent approaches [31, 36, 52, 59]. Diffusion models enable text-conditional image generation via utilizing CLIP [49] embeddings [43, 50, 52] or large-language-model encodings [54]. In practice, classifier guidance [14] or classifier-free guidance [23] is applied at inference steps further to enhance the quality and text-coherency in trade-off with diversity.

Generative Models as Data Producer. Similar to our approaches, there have been extensive approaches to leverage generative models as data generators. Prior approaches utilize GANs [17, 30, 48] to synthesize data for tasks including classification [1, 7, 39, 65, 68], 3D vision [19, 46, 57, 78], 2D segmentation [40, 67, 79], dense visual alignment [47]; or diffusion models [24, 52] for data augmentation [66] or as synthetic data for few-shot learning [22]. One major challenge of such methods is that generative models output "free" and "uncontrolled" images. Ali *et al.* [26] presents a method to generate multi-view images for representation learning by applying transformations on the latent vectors of the generative models, similar to our approach.

Learning 3D Spatial Arrangement between Human and Object. Prior work models 3D human-object interaction for each specified interaction type and input. Approaches include inferring hand-object interaction from
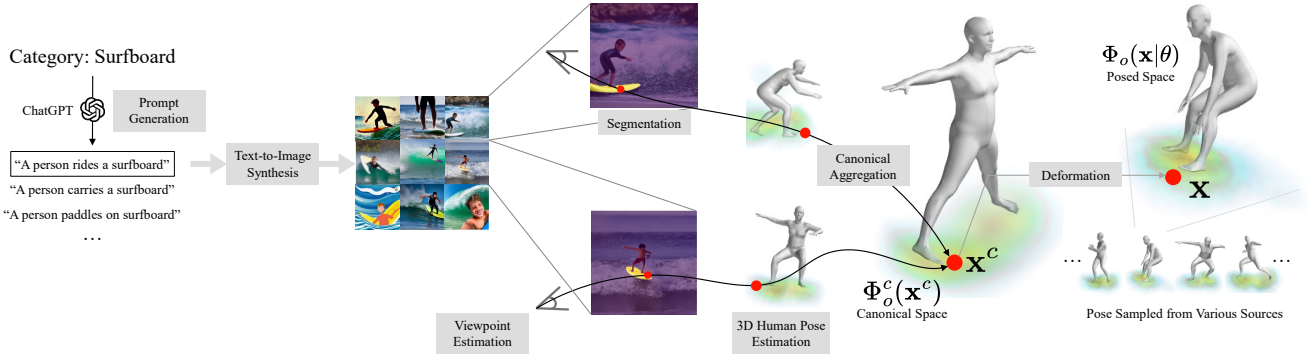
Figure 3. **Method overview.** Our method starts with generating prompts for human-object relationships within a specific object category. These prompts produce a multitude of images, incorporating HOI semantics. After applying filtering strategies to eliminate outliers, the chosen images are aggregated in canonical 3D space using a canonicalization approach. The resulting distribution can be flexibly adapted to different human postures.

3D [4, 64, 80], 2.5D such as heatmap [3, 5], or provided images [13, 15, 21, 32, 71]. These methods only model hand-object relationships and cannot predict the full body. The methods that model full-body interactions include generating/reconstructing 3D humans from 3D scene constraints [20, 56, 76, 77], capturing interaction from multi-view camera [2, 28, 62], or reconstructing 3D scene from human-scene interaction type [72], or human-human interactions [16]. Also, there exist approaches to model contacts [6, 11, 25, 34, 41, 51]. PHOSA [75] reconstructs 3D human and object from a single image; however, it requires manual labeling for the human-object interaction region, which limits the generalizability of the approach. CHORE [70] additionally models contact via distance between human and object by leveraging implicit surface learning to fit parametric SMPL [37] human and template object mesh. Unlike previous methods, our method does not require any object templates or supervision.

3. Method

3.1. Overview

Our method extracts 3D spatial knowledge of human-object interactions (HOI) by modeling object location as an occupancy probability distribution Φ_o when the category of the object o is given. The spatial relationship during HOI can vary even for the same object category due to the variation of human state (e.g., pose, body height and body shapes) and types of interactions (e.g., we can ride or hold a surfboard). Thus, we model the 3D spatial probability distribution in pose-conditioned and type-conditioned manners:

$$\Phi_o(\mathbf{x}|\theta, s) \in [0, 1], \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ is a 3D location of object occurrence in the "pose-deformed" space, θ is the 3D state of a person in terms of pose and shape variation, and s is a type of human-object

interaction among possible discrete variations $\mathbf{S}$. We represent Φ_o in an explicit voxel space with 48^3 resolutions, where the probability shows the likelihood that the specific location is occupied by the object o during interactions. The human state $\theta \in \mathbb{R}^{24 \times 3}$ is parameterized by SMPL [37] pose parameters for 24 joints to represent the condition of 3D human pose variations. $\mathbf{S}$ represents plausible types of interactions for an object o, which is automatically computed via our semantics clustering module described in Sec. 3.4. Specifically, $\mathbf{S}$ considers the semantics described by (1) natural language prompts and (2) contacted parts between humans and objects. For example, holding a surfboard and riding a surfboard should be considered as different interactions, despite the same object category, which is the objective to have $\mathbf{S}$ as the input for our Φ_o.

As a way to learn the 3D spatial distribution from inconsistent 2D image data, our method aggregates the HOI cues in a canonical space:

$$\Phi_o^c(\mathbf{x}^c|s) \in [0, 1], \tag{2}$$

where the $\mathbf{x}^c$ is a 3D location at the canonical space. The canonical space is defined by the rest-pose of SMPL, as shown in Fig. 3, where we put zero rotations for most joints except hips; $\pm\pi/6$ z-axis rotation on the left and right hips, respectively; following the approach proposed by SNARF [9] (we empirically find that it is advantageous to keep a sufficient distance between legs). Note that this canonical distribution Φ_o^c is independent of the 3D human pose state θ, in contrast to Eq. 1. Our 3D spatial reasoning in the canonical space is inspired by the recently emerging approaches to building animatable 3D avatars [9, 10, 27, 55], where cues from multiple 3D scans or views from different postures are aggregated in the canonical space. Yet, our framework is much simpler without requiring precise linear-blend skinning estimation since our target is reasoning the approximate object locations w.r.t. human rather than high-fidelity 3D human surface estimation. By reasoning in the

canonical space, we can handle the inconsistency and variation of synthesized multiview images in HOIs. To this end, our framework provides a warping function to convert back and forth between Φ_o^c and Φ_o, and our learned 3D spatial distribution can be applied to any 3D human postures to guess potential object locations as shown in Fig. 7. As an interesting key idea for generating multi-view image collections for 3D spatial relation learning, we use a SOTA diffusion model [52] that can synthesize realistic images from text prompt inputs. We refer readers to Fig. 3 for an overview of our method, and Supp. Mat. A.1~A.8 for more details on each component.

3.2. Synthesizing Images

Prompt Generation. We aim to produce various text prompts that describe the diverse semantics of human-object interaction on a target object o, which a text-conditioned diffusion model [52] can then take for image generation. In particular, we want to obtain visual cues on various human-object interactions with the target object o with many viewpoints for 3D reasoning.

Towards an automatic process, we present a solution by leveraging the ChatGPT [45], an instruction-tuned large language model, to produce a set of plausible prompts automatically. Specifically, we give the input query sentence as shown in Fig. 4, where we can simply replace the object category (e.g., skis as shown in a blue box) for any target objects. As shown, ChatGPT produces various related output prompts. We empirically observe that our solution with ChatGPT is much more efficient with higher quality, compared to the possible alternative way of directly generating prompt sentences via exhaustive combinations of a set of subjects and verbs, which may produce awkward expressions (e.g., "wear a bike"). Furthermore, to encourage similar scenes from diverse viewpoints as much as possible, we also present a strategy to control viewpoints of the synthesized images by augmenting view conditions on each prompt produced from ChatGPT, such as "back view", "side view", and "far view". As shown in Fig. 4, we demonstrate this viewpoint augmentation on prompts is advantageous in producing scenes from diverse views, while it is only partially guaranteed that the outputs follow the instructed viewpoints due to the imperfection of generative models. We generate 3 to 20 prompts per category, where each is augmented with 22 different auxiliary prompts to control viewpoints.

Synthesizing Text-Conditioned Images via Diffusion. Given a set of produced prompts regarding a human-object interaction, we link these natural language descriptions to visual entities by synthesizing images via an off-the-shelf latent diffusion model [52], which has great power in high-fidelity image generation with high mode coverage. Examples are shown in Fig. 4. Our strategy enables us to produce an unbounded number of diverse scenes containing the de-
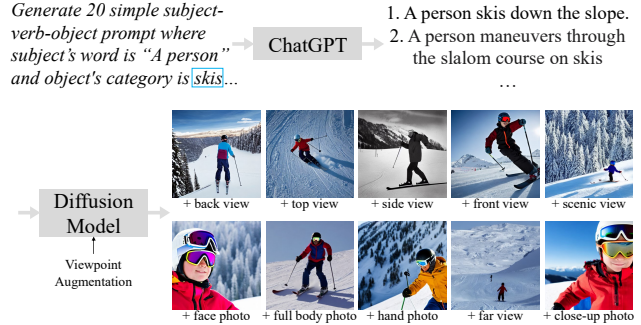

Figure 4. **Viewpoint augmentation.** Examples of prompts produced by ChatGPT and images synthesized via diffusion with viewpoint augmentation.

sired human object interaction with the target object category. Notably, the same prompt can be used multiple times by simply changing the initial random latent, resulting in different outputs. We create $5000 \sim 90000$ images per object category, evenly distributed for each given prompt within the category.

Filtering. Despite the efficacy of using the diffusion model for an image producer, the collected images are still in a "wild" status, making it challenging to use them directly for 3D spatial HOI learning. Thus, we first apply a series of filtering strategies to retain useful images only. Notably, the unbounded nature of our data generation allows us to use tough thresholds to filter out useless image samples aggressively. We use the following criteria in determining "valid" image samples: (1) the image contains a single human for efficient localization without ambiguity; (2) the image contains the target object; (3) the human and the target object should be close enough with sufficient overlapping in the bounding boxes (with a certain threshold); and (4) the whole torso part of the human should be visible. We use an off-the-shelf object detector [33, 69] for the bounding box detection and instance segmentation. To check the torso's existence, we use a 2D keypoint detector [12, 63, 74] and check whether shoulders and hips are within the image boundary and their detection confidences are above certain thresholds.

3.3. Pose Canonicalization

A large collection of images obtained from our approach can provide cues for 3D spatial HOI reasoning. As its first step, we estimate camera viewpoint by using the estimated 3D human orientation as an anchor. Then, we aggregate 2D object mask cues from each view to estimate 3D occupancy distribution, where we apply pose canonicalization to handle pose variations among synthesized images.

Viewpoint Estimation via 3D Human Pose Estimation. The orientation of the human body in the scene can provide a clue to estimate the relative camera viewpoint w.r.t the human. For this purpose, we use an off-the-shelf monocular 3D human pose estimator [53], which outputs the 3D global

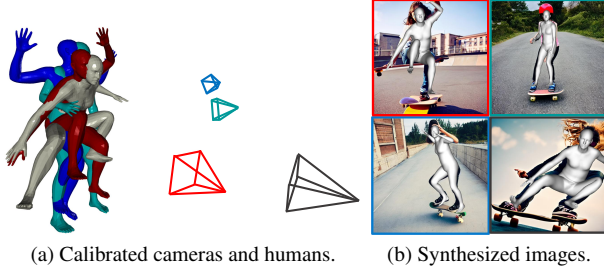(a) Calibrated cameras and humans.   (b) Synthesized images.

Figure 5. **Human-anchored camera calibration.** Predicted humans and calibrated cameras for synthesized images.

orientation of humans along with 3D joint orientations in SMPL [37] parameterzation. Specifically, given an image $\mathbf{I}$ and a human bounding box $\mathbf{B} \in \mathbb{R}^4$, a 3D pose regressor $\mathcal{F}_{\text{pose3d}}$ outputs:

$$\{\phi, \theta, \beta, \pi, \mathbf{j}\} = \mathcal{F}_{\text{pose3d}}(\mathbf{I}, \mathbf{B}), \tag{3}$$

where $\phi \in \mathbb{R}^3$ is the global orientation of the human defined in the camera-centric coordinate, $\theta \in \mathbb{R}^{23 \times 3}$ is 3D joint angles (excluding the pelvis root) in angle axis, $\beta \in \mathbb{R}^{10}$ is shape parameters, and $\pi \in \mathbb{R}^3$ is the weak-perspective camera parameters, and $\mathbf{j} \in \mathbb{R}^{24 \times 2}$ is the projected 2D joint locations on image space. We then convert π and ϕ into a perspective camera model Π in the person-centric coordinate system by minimizing the distance between projected SMPL joints via π and Π. To this point, all images are "calibrated" in a common 3D space, where the SMPL humans are aligned and centered at the origin (i.e., pelvis is set as origin), as depicted in Fig. 5.

3D Occupancy Estimation via Human Pose Canonicalization. We represent spatial 3D HOI reasoning via the occupancy probability field Φ_o^c as defined in Eq. 2. For brevity, we may first consider a holistic distribution $\Phi_o^c(\mathbf{x}^c) \in [0, 1]$, which does not consider semantics and returns a marginalized distribution for the most probable HOI. Given the virtually calibrated multi-view images we processed above, we compute Φ_o^c via visual hull reconstruction using the 2D object segmentation masks obtained in the filtering stage. However, as an issue, the 3D human poses may not be consistent across views, as shown in Fig. 5, making it challenging to apply visual hull reconstruction directly.

To handle this issue, we consider a canonical space defined via the rest pose of the SMPL model and compute the conversion between this canonical space and the pose-deformed space corresponding to each view. This process is inspired by the deformable object reconstruction [42] or the animatable human reconstruction pipelines [9, 10, 27, 55], where SMPL body posture provides the guidance to convert between two spaces. Unlike the 3D human modeling approaches that use MLPs for neural skinning [9, 10, 55], our goal is to enable the 3D voting to the corresponding 3D canonical volume space by warping into pose-deformed volume space and checking the means of 2D mask occupancies.

Thus, we take a simple strategy to find the mapping between the canonical and posed spaces. Specifically, we define Linear Blending Skinning (LBS) weights for 3D point $\mathbf{x}^c$ in the canonical space (we use 48^3 voxel grid) by averaging the ones from $\mathbf{k}$-nearest neighbor SMPL vertices with inverse distance weights, similar to SelfRecon [27]. That is:

$$\omega(\mathbf{x}^c) = \frac{\sum\limits_{i \in \mathbf{N_k}(\mathbf{x}^c)} w_i \, / \, \|\mathbf{x}^c - \mathbf{v}_i\|}{\sum\limits_{i \in \mathbf{N_k}(\mathbf{x}^c)} 1 \, / \, \|\mathbf{x}^c - \mathbf{v}_i\|} \tag{4}$$

where $\mathbf{N_k}(\mathbf{x}^c)$ is a set of $\mathbf{k}$-nearest neighbor vertex indices on SMPL mesh, $w_i \in \mathbb{R}^{24}$ and $\mathbf{v}_i \in \mathbb{R}^3$ are each associated LBS skinning weights and location of i-th vertex in SMPL. Intuitively, the 3D space is transformed by following the motion of the closest SMPL vertices. Additionally, assuming the monotonic decrease of the effect of bone transformation as $\mathbf{x}^c$ is far from the SMPL mesh model, we encourage "zero deformations" for the far-away points by decreasing the effect of LBS with skinning weight adjustments by simply applying weighted sum between $\omega(\mathbf{x}^c)$ and LBS skinning weights for pelvis $\mathbf{e}_1 \in \mathbb{R}^{24}$ which has no effect on deformation. Refer to Supp. Mat. A.5 for more details.

Given computed LBS weights, the warping from the canonical space $\mathbf{x}^c$ to a pose-deformed space $\mathbf{x}$ with SMPL pose θ can be performed by following the forward-kinematics pipeline of SMPL skeletal hierarchy:

$$\mathbf{x} = \mathcal{W}(\mathbf{x}^c) = \sum_{j=1}^{n_b} \omega_j(\mathbf{x}^c) \cdot \mathbf{B}_j(\theta_j) \cdot \mathbf{x}^c, \tag{5}$$

where ω_j denotes the j-th LBS weights, $\mathbf{B}_j(\theta_j) \in \mathrm{SE}(3)$ represent j-th bone's global 3D transformations.

Finally, the 3D occupancy aggregation can be performed by warping the 3D canonical points into the deformed space of each view and checking the means of the 2D object mask occupancies when projected:

$$\Phi_o^c(\mathbf{x}^c) = \frac{\sum\limits_{k=1}^{|\mathbf{G}|} r_k \mathcal{M}_k(\Pi_k(\mathcal{W}(\mathbf{x}^c)))}{\sum\limits_{k=1}^{|\mathbf{G}|} r_k \mathcal{I}_k(\Pi_k(\mathcal{W}(\mathbf{x}^c)))} \tag{6}$$

where $\mathbf{G}$ is the set of generated images, r_k is accumulation score for image k based on predicted camera distribution, $\mathcal{M}_k, \mathcal{I}_k$ are each mask and image operator for k-th image which returns 1 if the provided 2D value lies within mask/image space else 0, and Π_k is perspective projection for k-th image view.

Uniform View Sampling. Despite our viewpoint augmentation when prompting, the resulting images still may have biases toward specific viewpoints, making 3D reasoning difficult. Thus, we enforce uniform view sampling (similar to

importance sampling) by dividing the azimuth region into a fixed number of bins and setting the accumulation score r_k in Eq. 6 as the inverse of camera numbers in the bin.

Inference for Posed Space. At inference, we can transform our occupancy probability distribution $\Phi_o^c(\mathbf{x}^c)$ defined in the canonical space into the pose-deformed space $\Phi_o(\mathbf{x}|\theta)$ by using backward skinning, unlike the forward skinning (refer to Eq. 5) during training. Specifically, we set $\Phi_o(\mathbf{x}|\theta)$ by warping $\mathbf{x}$ in pose-deformed space to $\mathbf{x}^c$ in canonical space and retrieving the learned occupancy probability $\Phi_o^c(\mathbf{x}^c)$. To this end, we compute the LBS weights in the *pose-deformed* space (in contrast to training) and apply inverse transformation $\mathcal{W}^{-1}$.

3.4. Selective Aggregation via Semantic Clustering

The canonical distribution $\Phi_o^c(\mathbf{x}^c)$ from Eq. 6 does not take into account the different semantics. However, the form of human-object interaction can differ even within the same object category, requiring corresponding variations of $\Phi_o^c(\mathbf{x}^c)$, as defined in Eq. 2. To formulate this, we define the interaction type $s \in \mathbf{S}$ as a pair of prompt p and body part $\mathbf{a}$ (optional) in contact with the object:

$$\mathbf{S} = \{(p, \mathbf{a}) \mid p \in \mathbf{P}, \mathbf{a} \in \mathbf{A}\}, \tag{7}$$

where $\mathbf{P}$ is the set of entire prompts produced and $\mathbf{A}$ is the set of body part segments, given as a part of SMPL mesh vertices. Intuitively, a prompt represents a semantic (e.g., "playing with a ball"), and the body segment can further specify the interaction type (e.g., foot→"kicking").

We compute $\Phi_o^c(\mathbf{x}^c|s)$ for each interaction type $s = (p, \mathbf{a})$ by aggregating a semantic cluster of images that depict such interactions, selected from $\mathbf{G}$. Specifically, we utilize images generated from the single prompt p. Body part $\mathbf{a}$ is used as a proximity cue to retain relevant image samples, where we consider the image irrelevant if the 3D rays from the object mask do not intersect with the interaction region of $\mathbf{a}$. Our method of "selective aggregation" enhances spatial HOI reasoning for multimodal scenarios.

4. Experiments

We provide both quantitative and qualitative comparisons to verify our method. In Sec. 4.2, we present a new metric, namely *Projective Average Precision (PAP)*. We provide a brief explanation of PAP and detailed protocols in Supp. Mat. B.2. In Sec. 4.3, we quantitatively compare our method with various ablations to provide justifications for our design choices. We also compare our method trained on synthesized images with the one trained on internet-crawled images, and show that synthesized images are more suitable for learning 3D HOI. In Sec. 4.4, we show qualitative results. We demonstrate the quality of the learned HOI spatial distribution when various human pose is given for a diverse set of

object categories. We then explore the effects of semantic types on the distribution by changing HOI prompts and body part specifications. We also analyze the effects of canonicalization via a comparison with an ablation. In Sec. 4.5, we exemplify an application of our method to the downstream task: *3D Human-Object Reconstruction from a Single-view Image*. Refer to Supp. Mat. B.1∼B.2 for more details on each part, and Supp. Mat. C for extensive qualitative results.

4.1. Dataset

Generated Dataset. Since our method is a fully-autonomous and self-supervised approach, no external dataset is required for training. We test our method for 19 object categories from COCO [35] (e.g., bicycle, chair) and 5 categories from LVIS [18] (e.g., hat, sweater), where we generate $5000 \sim 90000$ images per category.

Image Search Dataset. We provide a quantitative comparison between the results of our method using image search data and using synthesized images in Sec. 4.3. To this end, we use AutoCrawler [73] to collect images for category *motorcycle* from the internet, using the same prompts used for preparing generated dataset.

Extended COCO-EFT Dataset for Testing. To quantitatively evaluate our learned HOI distributions, we use COCO dataset [35], which provides GT 2D object masks. In particular, we use COCO-EFT (val) [29], where pseudo GTs for 3D human poses are provided in SMPL [37] format. Among all samples in COCO-EFT, we perform a filtering process to keep only the samples where human-object interactions happen by retaining the samples with a single human and a single object with sufficient overlaps between them. After the filtering process, we compute perspective camera parameters similar to Sec. 3.3 from the weak-perspective cameras provided in the dataset.

4.2. Projective Average Precision

As no standard metrics exist for this task, we define a new evaluation metric: *Projective Average Precision (PAP)*. Refer to Fig. 6 for an overview of the PAP evaluation protocols.

Evaluation Protocols. We utilize the COCO-EFT [29] dataset with pseudo-GT 3D human pose and object masks to compare the projection of estimated 3D distribution with GT object mask. Specifically, we first deform the canonical 3D distribution into pose-deformed space and apply threshold for discretization. We project discretized occupancy values to 2D using the perspective camera and compute *pixel-wise* precision and recall between ground-truth object mask and projected occupancy. By setting multiple threshold values, we enable quantifying the validity of distribution regarding the intra-class object variation in terms of geometry. Note that we compute precision and recall for evaluation as our method aims to infer the *object distribution*, not to reconstruct the exact geometry of the target. The calculated pre-
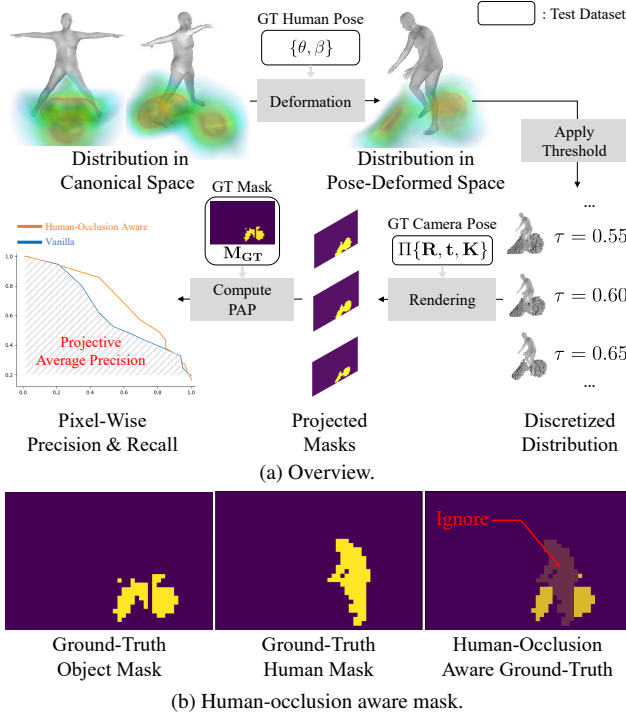
(a) Overview.

Ground-Truth Object Mask | Ground-Truth Human Mask | Human-Occlusion Aware Ground-Truth

(b) Human-occlusion aware mask.

Figure 6. **Overview of evaluation protocols for PAP.**

Table 1. **Image Search vs. Synthesized Images: Statistics.**

Method	# Images	# Images after Filtering	Rejection-rate (%)↓	Camera Entropy (bits)↑
Image Search$_{1k}$	1201	391	67.44	10.04
Image Search$_{10k}$	9408	2526	73.15	12.25
Synthetic$_{1k}$ (Ours)	1201	456	62.03	12.21
Synthetic$_{10k}$ (Ours)	9408	3626	**61.46**	**14.63**

Table 2. **Quantitative Evaluation Results.** (up) Results for ablation studies on COCO-EFT categories. (down) Results for quantitative comparison between image search and synthesize images on single category *motorcycle*.

Method	Vanilla		Human-Occlusion Aware	
	mPAP↑	mPAP$_{strict}$↑	mPAP↑	mPAP$_{strict}$↑
Ours$_{w/o\ Keypoint\ Filtering}$	18.31	15.90	19.97	16.52
Ours$_{w/o\ Canonicalization}$	17.52	15.12	19.56	15.85
Ours$_{1\%}$	14.98	13.81	15.21	13.73
Ours$_{10\%}$	17.90	16.13	18.86	15.82
Ours$_{full}$	**19.86**	**17.18**	**20.28**	**16.80**
Image Search$_{1k}$	53.64	52.15	60.08	56.86
Image Search$_{10k}$	55.97	54.59	63.62	61.11
Synthetic$_{1k}$ (Ours)	56.33	55.11	63.91	61.85
Synthetic$_{10k}$ (Ours)	**57.14**	**55.19**	**64.82**	**62.11**

- *w/o Canonicalization*: We compute the HOI distribution without pose canonicalization. Notably, we find that removing canonicalization damages the metrics by a large margin. This is an expected result, as removing canonicalization will disperse the 2D cues and limit the precise aggregation due to pose variation.

- *10%, 1% Trained*: To see the effect of the number of generated data used for training (i.e., training time), we reduce the images by randomly selecting 10% or 1% of filtered images per prompt and run the 3D aggregation. The results imply that using more synthesized images for aggregation is beneficial for the quality of learned distribution, and the potential of our method for scalability.

Image Search vs. Synthesized Images. To validate the use of synthesized images, we compare our method's output for *motorcycle* category across diverse datasets. We opt for *motorcycle* due to the abundance of accessible and common images, which ensures a fair and rigorous comparison. We assess four settings, altering (1) data source (image search or synthesis) and (2) image count (1k or 10k). To ensure fairness, we randomly select images from the generated dataset, equivalent to the size of image search dataset. We follow the same procedures described in Sec. 3 for aggregation. We report the dataset's retrieval statistics including post-filtering rejection rate, and camera distribution entropy for filtered images in Tab. 1. For the computation of camera distribution entropy, we fit Gaussian normals with a standard deviation of $\sigma = 0.01$ to each camera's location on the unit sphere, using geodesic metrics.

We observe a higher and increasing rejection rate in the image search dataset compared to the stable, relatively low rate in the generated dataset. We attribute this to the word-occurrence-based search algorithm, which may retrieve im-
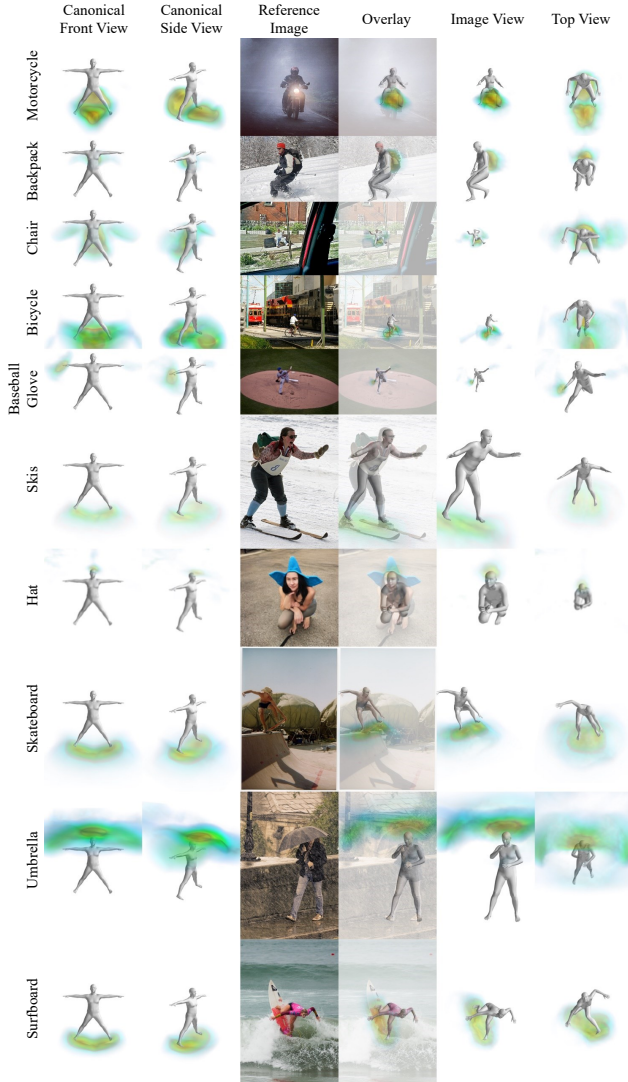
cision and recall values are then employed to determine the average precision (AP) using two distinct methods (vanilla, strict). Averaging the AP values across all images within the category yields the *Projective Average Precision (PAP)*. Further insights and comprehensive details can be found in Supp. Mat. B.2.

Human-Occlusion-Aware PAP. We observe that human masks frequently overlap with objects, potentially causing inaccuracies in evaluating rendered 3D occupancy against partially-removed ground-truth masks as shown in Fig. 6b. To address this issue, we exclude precision and recall calculations within the 2D region of the human projection, considering the likelihood of occlusions.

4.3. Quantitative Evaluation

Ablation Studies. To justify our design choice, we quantitatively compare our method with various ablations. We use mPAP (average of PAP for all categories) metric for comparison, and we report the results in Tab. 2. We list the baselines used for comparison and discussion for each result below. Unless specified, all baselines follow the same implementation details provided in Sec. 3 and Supp. Mat. A.

- *w/o Keypoint Filtering*: We deactivate keypoint filtering and conduct experiments using an equivalent number of images as in the case with keypoint filtering. We randomly sample from a combination of unfiltered and filtered images. We report a drop in performance when keypoint filters are not applied, which is attributed to erroneous 3D human predictions.

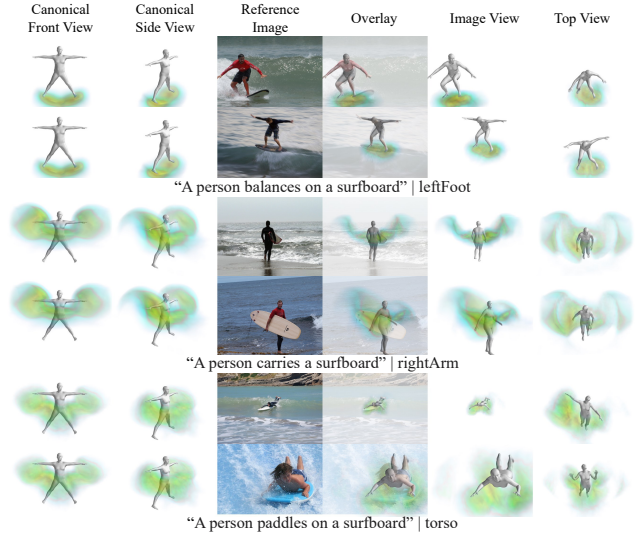Figure 7. **Qualitative Results for Various Categories.**

Figure 8. **Effects of semantic type condition on object distribution.** As semantic types vary, the canonical and pose-deformed object distribution changes accordingly.

ages without HOI (e.g., commercials). Thus, we argue that the generated dataset offers enhanced scalability for HOI learning. Furthermore, the superior camera distribution entropy in synthesized images implies a richer diversity of camera poses than the image search dataset. Such diversity provides more informative content for 3D aggregation in the generated dataset. While our method addresses imbalanced camera distribution during aggregation through uniform view sampling (Sec. 3.3), results in Tab. 2 consistently demonstrate the synthetic dataset's superior performance across metrics. This reaffirms the effectiveness of our data generation approach for HOI learning, even after minimizing information differences from camera imbalance.

4.4. Qualitative Evaluation

Various Categories. We demonstrate that our method can learn spatial human-object relationships for a diverse set of

object categories in Fig. 7. We use SMPL pose and reference image sampled from; the COCO-EFT [29] dataset for COCO [35] categories and generated dataset for LVIS [18] categories; to deform the distribution in canonical space to pose-deformed space. We visualize the semantic cluster that best matches the reference image. Refer to Supp. Mat. C for extensive results.

Effects of Semantic Type Condition. We find that our set of learned distributions conveys the human-object interaction effectively. We report the examples of object distributions (category: *surfboard*) in canonical and pose-deformed space for different semantic types $s = (p, \mathbf{a})$ in Fig. 8. Notably, the object distribution of the surfboard for the given prompt "A person carries a surfboard" shows the mode near the human torso and arms: which is a plausible location for holding a surfboard when carrying around. We also report results when semantic body part $\mathbf{a}$ vary for the same semantic prompt description. In Fig. 9, we observe that body parts act important as a condition when the provided semantic prompt contains little or less information on HOI; hence, ambiguous. For example, the semantics for "A man playing with a sports ball" can be different in terms of sports categories. In Fig. 10, we can see the effects of prompt change when body parts are the same. While the semantic type with prompt "The cyclist pedaled the bicycle" return a shape similar to the original bicycle as expected, we observe the donut-shaped distribution as in Fig. 10 when we change the prompt into "The commuter standing next to a bicycle". We argue that such distribution conveys accurate semantics of spatial arrangements as humans can be oriented in arbitrary directions even if standing next to a bicycle, where we can assume such arrangement leads to rotationally-invariant distribution (e.g., donut-shape).
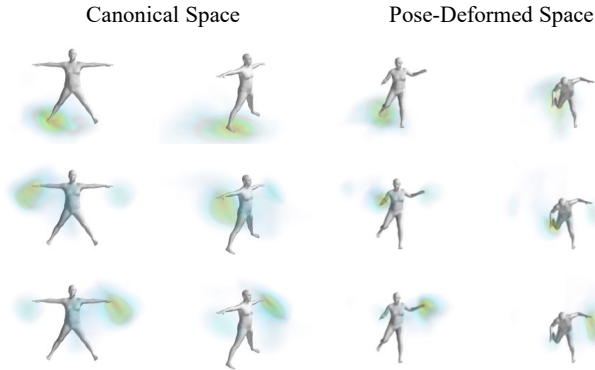
Figure 9. **Effects of body-part specification.** For category *sports ball* and prompt "A man playing with a sports ball", body parts play a significant role as a condition for further specifying the HOI type.
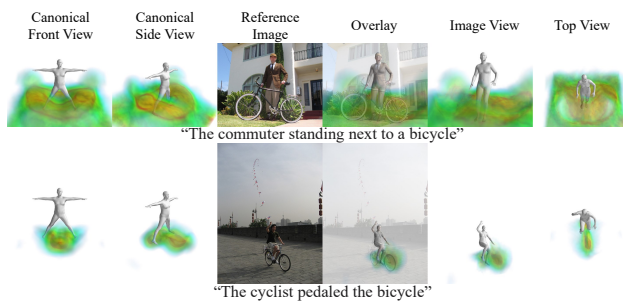

Figure 10. **Effects of prompt specification.** For category *bicycle*, object distribution changes significantly as prompts representing the HOI type differs.
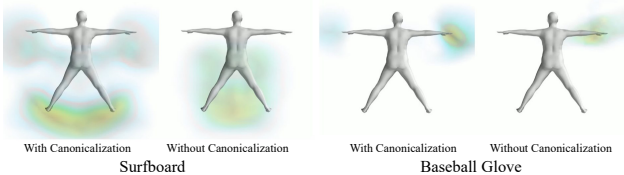

Figure 11. **Effects of canonicalization.** Removing canonicalization from our approach leads to scattered aggregation of 2D mask occupancies, resulting in blurry outcomes and diminished precision.

Effects of Canonicalization. We ablate the effects of canonicalization and report results in Fig. 11. From results, we can observe that precision is enhanced and variance is reduced when pose-canonicalization is applied, as the aggregation of 2D cues becomes more coherent compared to direct rays due to pose-variance.

4.5. Application

Our method returns generalizable 3D human-object spatial knowledge represented as probability occupancy distribution of the object, which can be easily applied to many downstream tasks. This section provides an example case of application to the following downstream task: *3D Human-Object Reconstruction from a Single-view Image*. Given a single-view image, we apply off-the-shelf monocular 3D
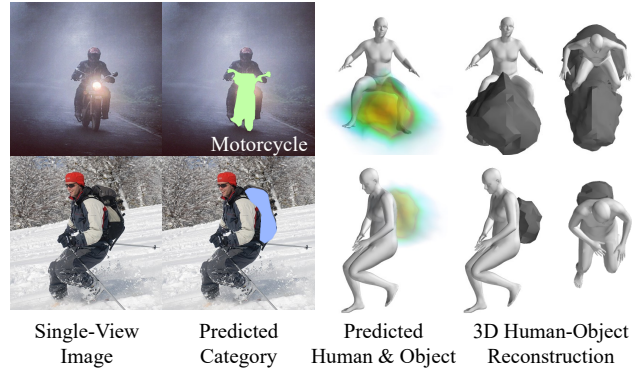

Figure 12. **3D Human-Object Reconstruction from a Single-view Image.** Using learned distribution as a prior, we can reconstruct 3D human and object from single image without any template mesh.

human pose estimator [53] to extract SMPL pose and off-the-shelf object detector [33] to find the category of interacting object, where we find the object of interaction by computing bounding box overlap between human and object and picking the object with maximum value. Next, we extract 3D object occupancy distribution by simply inputting the category of the found object as a keyword to our method. Finally, we apply the marching cube algorithm [38] to the object distribution to obtain 3D object occupancy, which is subsequently converted into a 3D object mesh. Our method is applicable to in-the-wild images, fully automatic and template agnostic, in contrast to previous works that use template mesh to represent 3D objects [70, 75]. See Fig. 12 for results.

5. Discussion

Our method introduces a novel approach, enabling machines to probabilistically learn 3D human-object spatial relationships based on pose and HOI types. Unlike prior methods, ours extracts this information through self-supervision, eliminating the need for laborious annotations. We present multiple strategies, including leveraging a generative model for prompt/image synthesis to enhance image control, and a canonicalization-based framework for computing 3D spatial HOI fields from synthesized 2D images. Furthermore, we introduce a novel metric that demonstrates the high quality of our learned HOI distribution. As the work presents a method to extract an intermediate representation, our approach offers numerous potential downstream applications; however, while possessing such potentials, the work is new and thus has limitations, such as low granularity and inaccurate modeling of distributions for small objects. We extensively discuss these limitations and potential future avenues in Supp. Mat. D.

References

[1] V. Besnier, H. Jain, A. Bursuc, M. Cord, and P. Pérez. This dataset does not exist: training models from generated images. In *ICASSP*, 2020. 2

[2] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 3

[3] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 3

[4] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, 2019. 3

[5] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020. 3

[6] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 3

[7] L. Chai, J.-Y. Zhu, E. Shechtman, P. Isola, and R. Zhang. Ensembling with deep generative views. In *CVPR*, 2021. 2

[8] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 1

[9] X. Chen, Y. Zheng, M. J. Black, O. Hilliges, and A. Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 3, 5

[10] X. Chen, T. Jiang, J. Song, J. Yang, M. J. Black, A. Geiger, and O. Hilliges. gdna: Towards generative detailed neural avatars. In *CVPR*, 2022. 3, 5

[11] Y. Chen, S. Huang, T. Yuan, S. Qi, Y. Zhu, and S.-C. Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *ICCV*, 2019. 3

[12] M. Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 4

[13] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 3

[14] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[15] K. Ehsani, S. Tulsiani, S. Gupta, A. Farhadi, and A. Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *CVPR*, 2020. 3

[16] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020. 3

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[18] A. Gupta, P. Dollar, and R. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6, 8

[19] Z. Hao, A. Mallya, S. Belongie, and M.-Y. Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *ICCV*, 2021. 2

[20] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, 2021. 3

[21] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 3

[22] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2

[23] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[24] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[25] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 3

[26] A. Jahanian, X. Puig, Y. Tian, and P. Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 2

[27] B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 3, 5

[28] Y. Jiang, S. Jiang, G. Sun, Z. Su, K. Guo, M. Wu, J. Yu, and L. Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *CVPR*, 2022. 3

[29] H. Joo, N. Neverova, and A. Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 6, 8

[30] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[31] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 2

[32] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020. 3

[33] A. Kirillov, Y. Wu, K. He, and R. Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 4, 9

[34] Z. Li, J. Sedlar, J. Carpentier, I. Laptev, N. Mansard, and J. Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *CVPR*, 2019. 3

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6, 8

[36] L. Liu, Y. Ren, Z. Lin, and Z. Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 2

[37] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 3, 5, 6

[38] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*. 1998. 9

[39] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick. Generative interventions for causal learning. In *CVPR*, 2021. 2

[40] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi.

Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021. 2

[41] A. Monszpart, P. Guerrero, D. Ceylan, E. Yumer, and N. J. Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM TOG*, 38(4):1–15, 2019. 3

[42] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 5

[43] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[44] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[45] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2023. 2, 4

[46] X. Pan, B. Dai, Z. Liu, C. C. Loy, and P. Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844*, 2020. 2

[47] W. Peebles, J.-Y. Zhu, R. Zhang, A. Torralba, A. A. Efros, and E. Shechtman. Gan-supervised dense visual alignment. In *CVPR*, 2022. 2

[48] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[50] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[51] D. Rempe, L. J. Guibas, A. Hertzmann, B. Russell, R. Villegas, and J. Yang. Contact and human dynamics from monocular video. In *ECCV*, 2020. 3

[52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[53] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021. 4, 9

[54] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[55] S. Saito, J. Yang, Q. Ma, and M. J. Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 3, 5

[56] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner. Pigraphs: learning interaction snapshots from observations. *ACM TOG*, 35(4):1–12, 2016. 3

[57] Y. Shi, D. Aggarwal, and A. K. Jain. Lifting 2d stylegan for 3d-aware face generation. In *CVPR*, 2021. 2

[58] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[59] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit

models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[60] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

[61] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[62] G. Sun, X. Chen, Y. Chen, A. Pang, P. Lin, Y. Jiang, L. Xu, J. Yu, and J. Wang. Neural free-viewpoint performance rendering under complex human-object interactions. In *ACMM*, 2021. 3

[63] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 4

[64] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3

[65] F. H. K. d. S. Tanaka and C. Aranha. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019. 2

[66] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2

[67] N. Tritrong, P. Rewatbowornwong, and S. Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *CVPR*, 2021. 2

[68] E. Wu, K. Wu, D. Cox, and W. Lotter. Conditional infilling gans for data augmentation in mammogram classification. In *MICCAI Workshop on Breast Image Analysis*, 2018. 2

[69] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 4

[70] X. Xie, B. L. Bhatnagar, and G. Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, 2022. 3, 9

[71] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, and C. Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 3

[72] H. Yi, C.-H. P. Huang, D. Tzionas, M. Kocabas, M. Hassan, S. Tang, J. Thies, and M. J. Black. Human-aware object placement for visual environment reconstruction. In *CVPR*, 2022. 3

[73] K. Yoongi. Autocrawler. https://github.com/YoongiKim/AutoCrawler, 2023. 6

[74] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020. 4

[75] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 3, 9

[76] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang. Place: Proximity learning of articulation and contact in 3d environments. In *3DV*, 2020. 3

[77] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll. Couch: Towards controllable human-chair interactions. In *ECCV*, 2022. 3

[78] Y. Zhang, W. Chen, H. Ling, J. Gao, Y. Zhang, A. Torralba, and S. Fidler. Image gans meet differentiable rendering for

inverse graphics and interpretable 3d neural rendering. *arXiv preprint arXiv:2010.09125*, 2020. 2

[79] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2

[80] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *ECCV*, 2022. 3