LEMON: Learning 3D Human-Object Interaction Relation from 2D Images

Yuhang Yang[1], Wei Zhai[1,†], Hongchen Luo[1], Yang Cao[1,2], Zheng-Jun Zha[1]
[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yyuhang@mail., wzhai056@, lhc12@mail., forrest@, zhazj@}ustc.edu.cn

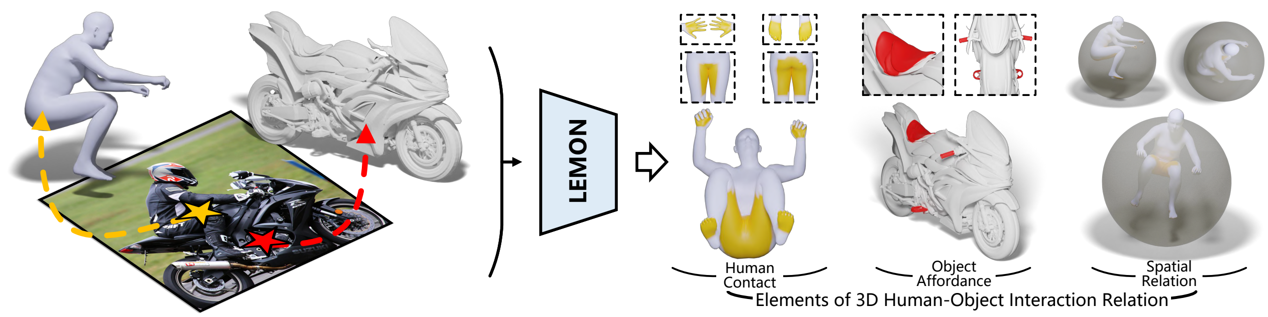

Figure 1. For an interaction image with paired geometries of the human and object, LEMON learns 3D human-object interaction relation by jointly anticipating the interaction elements, including human contact, object affordance, and human-object spatial relation. Vertices in yellow denote those in contact with the object, regions in red are object affordance regions, and the translucent sphere is the object proxy.

Abstract

Learning 3D human-object interaction relation is pivotal to embodied AI and interaction modeling. Most existing methods approach the goal by learning to predict isolated interaction elements, e.g., *human contact, object affordance, and human-object spatial relation, primarily from the perspective of either the human or the object. Which underexploit certain correlations between the interaction counterparts (human and object), and struggle to address the uncertainty in interactions. Actually, objects' functionalities potentially affect humans' interaction intentions, which reveals what the interaction is. Meanwhile, the interacting humans and objects exhibit matching geometric structures, which presents how to interact. In light of this, we propose harnessing these inherent correlations between interaction counterparts to mitigate the uncertainty and jointly anticipate the above interaction elements in 3D space. To achieve this, we present **LEMON** (**LE**arning 3D hu**M**an-**O**bject i**N**teraction relation), a unified model that mines interaction intentions of the counterparts and employs curvatures to guide the extraction of geometric correlations, combining them to anticipate the interaction elements. Besides, the **3D I**nteraction **R**elation dataset (**3DIR**) is collected to serve as the test bed for training and evaluation. Extensive experiments demonstrate the superiority of LEMON over methods estimating each element in isolation. The code and dataset are available at https://yyvhang.github.io/LEMON.*

†Corresponding Author.

1. Introduction

Learning 3D human-object interaction (HOI) relation seeks to capture semantic co-occurrence and geometric compatibility between humans and objects in 3D space [5, 68, 80]. *How can machines learn the interaction relation?* One possible solution involves perceiving certain elements capable of revealing the interaction. Contact [6, 21, 64, 70], affordance [10, 13, 46, 92], and spatial relation [19, 31, 52] that elucidate "where" the interaction manifests between the human and object garner great attention. Capturing representations of such elements is pivotal for applications like AR/VR [7], imitation learning [1, 24], embodied AI [17, 58, 60], and interaction modeling [20, 81].

Humans predominantly manipulate and interact with objects in 3D space. Thus, many methods devise task-specific models to anticipate certain interaction elements, thereby perceiving 3D HOI relation. Methods [21, 62, 64] estimate dense **human contact** based on the interaction semantics depicted in images. Some studies anticipate the **object affordance** according to objects' structures or 2D visuals [10, 44, 67, 78]. Several works delve into predicting the human-object **spatial relation** through synthetic images [19] or posed human geometries [52]. A prevailing trend in these methods involves taking certain attributes (*e.g.*, appearances, geometries) of either humans or objects to predict an isolated interaction element, capturing one aspect of the interaction relation. However, they overlook the impact of factors mutually determined by humans and objects on interactions, such as the interaction intention and geometric correlation, which leads to struggles in addressing the uncertainty within interactions. Specifically, the uncertainty could be attributed to two principal aspects. On the one hand, the diversity of HOIs introduces challenges in capturing explicit interaction intentions solely centering on humans or objects, raising the intention uncertainty. On the other hand, the limited view and mutual occlusions give rise to invisible interaction regions in images, complicating the constitution of linkages between these regions and target 3D geometries, causing the region uncertainty. These uncertainties may culminate in ambiguous anticipations.

In this paper, we propose leveraging both counterparts of the interaction to jointly anticipate human contact, object affordance, and human-object spatial relation in 3D space (Fig. 1), addressing the uncertainty by unearthing the correlation between the interacting humans and objects. Actually, humans and objects are intertwined and possess affinities in the interaction (Fig. 2). In specific, the design of objects typically adheres to certain human needs. Therefore, object affordances inherently hint at "what" interactions humans intend to make [39, 61, 63], revealing the intention affinity of interactions. Meanwhile, the interacting human and object exhibit matching geometries (either posture or configuration), which presents "how" to interact, arising the geometry affinity. The intention affinity clarifies the interaction type and implicates the interaction regions. Geometry affinity could serve as pivotal clues for excavating correlations between geometries corresponding to invisible regions in the image. These interaction-related regions *e.g.*, contact regions, further reflect the human-object spatial relation.

To achieve this, we present the **LEMON**, a novel framework that correlates the intention semantics and geometric correspondences to jointly anticipate human contact, object affordance, and human-object spatial relation, in 3D space. To capture the intention affinity, LEMON employs multi-branch attention to model the correlation between the interaction content in images and geometries of humans and objects, revealing intention representations of the interaction corresponding to geometries. The cosine similarity is utilized to further ensure their semantic consistency. Taking intention representations as conditions, LEMON integrates geometric curvatures to capture the geometry affinity and reveal the human contact and object affordance representations. These representations then assist in anticipating the spatial relation constrained by a combined distance loss.

In addition, we collect the **3DIR** dataset, which contains natural HOI images paired with object point clouds and SMPL-H [57] pseudo-GTs. Multiple annotations are made for these data, *e.g.*, dense human contact, object affordance, and human-object spatial relation. It serves as the test bed for the model training and evaluation.

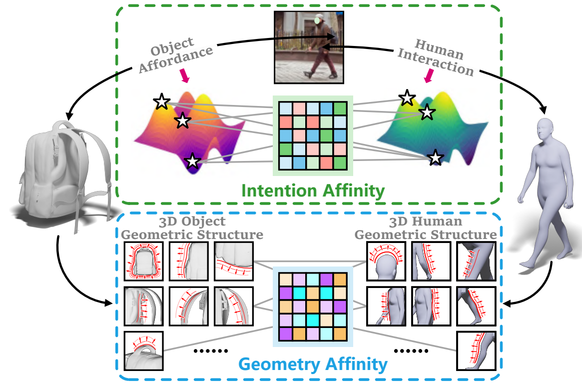

Figure 2. **Motivation.** Affinities within the HOI. The object affordance inherently reveals the human's interaction intention, arising the intention affinity. The interacting human and object possess matching structures, exhibiting the geometry affinity.

The contributions are summarized as follows:

1) We thoroughly exploit the correlation between the interaction counterparts to jointly anticipate human contact, object affordance, and human-object spatial relation in 3D space. It furnishes essential interaction elements to comprehend 3D HOIs.

2) We present the LEMON, a novel framework that correlates semantic interaction intention and geometries to capture the affinity between humans and objects, eliminating the impact of the interaction uncertainty and anticipating plausible 3D interaction elements.

3) We provide the 3DIR dataset that contains paired HOI data and multiple annotations, including dense human contact, object affordance, and human-object spatial relation, to support the anticipation. Extensive experiments demonstrate the superiority of LEMON.

2. Related Work

2.1. Human-Object Interaction Relation in 2D

Various approaches perceive elements pertinent to human-object interaction relation from multiple perspectives in the 2D domain. Methods [4, 16, 30, 72, 94] represent human-object interaction as triplets <human, action, object>. Grounding bounding boxes of interaction counterparts and actions corresponding to them. They output the instance-level human-object spatial relation in pixel space and action semantics. Delving into the object side, the concept of "interaction possibility" is encapsulated in the term affordance [14], and bountiful methods explore to anticipate the object affordance, whether at instance-level or part-level

[11, 12, 37, 38, 47, 85]. Turning to the human side, the contact explicitly indicates where the human interacts with objects. Some works detect the contact for specific parts (*e.g.*, hand and foot) at a box-region level [48, 55, 59]. HOT [6] provides 2D contact annotations to support human-object contact estimation in images. These efforts achieve substantial results in understanding the human-object interaction relation in 2D space. However, they still encounter challenges when extrapolating to 3D space for practical applications due to the lack of a dimension.

2.2. Human-Object Interaction Relation in 3D

To facilitate the incorporation of interaction comprehension into applications, extensive works explore to anticipate interaction elements in 3D space. For object affordance, methods [10, 44, 67, 69, 73, 86, 91] constitute the linkage between object shapes and affordances by a learning-based mapping, approaching the matter from the perspective of objects' functionality and structure. Some approaches learn object affordances from a distinct perspective, based on interactions. Whether making agents actively interact with 3D synthetic scenarios [46], or ground 3D object affordances through object-object [45] and 2D [78] interactions.

The dense human contact estimation is primarily based on the SMPL series [36, 50, 57]. Numerous methods model human contact for various tasks [8, 20, 22, 23, 71, 82, 83]. HULC [62], BSTRO [21], and DECO [64] are more similar to the estimation of contact in our method. They learn a mapping from the interaction semantics in images to the human vertices sequence, hardly leveraging geometries. Different from them, our method harnesses semantic and geometric correlations of interaction counterparts to address the uncertainty and infer the vertices that contact with objects. DECO also contributes the DAMON [64] dataset that possesses dense 3D contact annotation for in-the-wild images. In addition to human contact annotations, 3DIR includes other annotations related to the interaction, *e.g.*, 3D object affordance and human-object spatial relation.

Bounding boxes obtained by HOI detection methods give the human-object spatial relation in pixel space. To lift this relation to 3D space, DJ-RN [31] takes hollow spheres with defined radii to represent objects, and projects the spatial relation in images to 3D space based on the bounding boxes and defined radii. CHORUS [19] utilizes synthetic multi-view images generated by the diffusion model [56] to learn the spatial relation between objects and canonical humans. Given a posed human, Object Pop-Up [51] anticipates the objects and their spatial positions to match the human body for certain interactions.

The above methods commonly focus on one side of the interaction in isolation. In contrast, our method captures the inherent affinity between both sides of the interaction to jointly anticipate interaction elements in 3D space. Such anticipations are beneficial for tasks *e.g.*, robot manipulation [3, 40, 93], interaction generation [27, 29, 75, 81, 90] and reconstruction [20, 71, 76, 87].

3. Method

Given the inputs $\{H, O, I\}$, where $H \in \mathbb{R}^{N_h \times 3}$ indicates vertices of the SMPL-H [57], which represents the human as pose $\theta \in \mathbb{R}^{52 \times 3}$, shape $\beta \in \mathbb{R}^{10}$ parameters, and output a mesh $M(\theta, \beta) \in \mathbb{R}^{6890 \times 3}$. $O \in \mathbb{R}^{N_o \times 3}$ is an object point cloud, $I \in \mathbb{R}^{H \times W \times 3}$ is an image. N_h, N_o are the number of points, H, W are image's height and width. LEMON jointly anticipates the human contact $\bar{\phi}_c \in \mathbb{R}^{N_h \times 1}$, object affordance $\bar{\phi}_a \in \mathbb{R}^{N_o \times 1}$, and object center position $\bar{\phi}_p \in \mathbb{R}^3$, in 3D space. As shown in Fig. 3, initially, the inputs are sent to image and point cloud backbones, obtain respective features $\mathbf{F}_h, \mathbf{F}_o, \mathbf{F}_i$. Then, LEMON correlates $\mathbf{F}_i$ with $\mathbf{F}_o$, $\mathbf{F}_h$ to extract features of the interaction intention ($\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$) inherent in geometries through multi-branch attention, and employs cosine similarity to constrain their semantic consistency (Sec. 3.1). With $\bar{\mathbf{T}}_o$, $\bar{\mathbf{T}}_h$ as conditions, LEMON integrates curvatures to guide the modeling of geometric correlations, capturing the geometry affinity and revealing the contact ϕ_c and affordance ϕ_a features (Sec. 3.2). Next, geometric features and $\phi_c, \bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$ are utilized to model the object spatial feature ϕ_p (Sec. 3.3). Eventually, ϕ_c, ϕ_a, ϕ_p are projected to $\bar{\phi}_c$, $\bar{\phi}_a$, and $\bar{\phi}_p$ in the decoder, the whole process is optimized by a combined loss (Sec. 3.4).

3.1. Interaction Intention Excavation

We add human and object masks on I, and utilize the HRNet [65] and DGCNN [66] as backbones to extract the image feature $\mathbf{F}_i \in \mathbb{R}^{C \times h \times w}$, the geometric features of human $\mathbf{F}_h \in \mathbb{R}^{C \times N_h}$, and object $\mathbf{F}_o \in \mathbb{R}^{C \times N_o}$, $\mathbf{F}_i$ is flattened to $\mathbb{R}^{C \times hw}$. Images contain rich interaction semantics, which could serve as clues to unearth the interaction intention within the geometries. In detail, tokens $\mathbf{T}_o, \mathbf{T}_h \in \mathbb{R}^{C \times 1}$ are generated to represent intention features inherent in geometries. $\mathbf{T}_o, \mathbf{T}_h$ are concatenated with $\mathbf{F}_o$ and $\mathbf{F}_h$ respectively to get the feature sequences $\mathbf{F}_{to} \in \mathbb{R}^{C \times (N_o+1)}, \mathbf{F}_{th} \in \mathbb{R}^{C \times (N_h+1)}$. Taking $\mathbf{F}_i$ as the shared key and value, $\mathbf{F}_{to}$, $\mathbf{F}_{th}$ as queries (q_1, q_2) in two branches, the multi-branch attention is employed to model the intention features, expressed as $\bar{\mathbf{F}}_{to}, \bar{\mathbf{F}}_{th} = f_\delta([\mathbf{F}_{to}, \mathbf{F}_{th}], \mathbf{F}_i)$. Where f_δ indicates the multi-branch attention layer [28, 54, 84], $[\cdot]$ denotes two branches with different queries. Human and object geometries possess multiple interaction possibilities, which may introduce semantic ambiguity. To mitigate this, we further constrain the consistency of semantic tokens:

$$\varphi = \frac{\bar{\mathbf{T}}_o \cdot \bar{\mathbf{T}}_h}{||\bar{\mathbf{T}}_o||_2 \times ||\bar{\mathbf{T}}_h||_2}, \tag{1}$$

where $\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h \in \mathbb{R}^{C \times 1}$ are split from $\bar{\mathbf{F}}_{to}$, $\bar{\mathbf{F}}_{th}$, representing the semantic intention features of geometries. φ is the

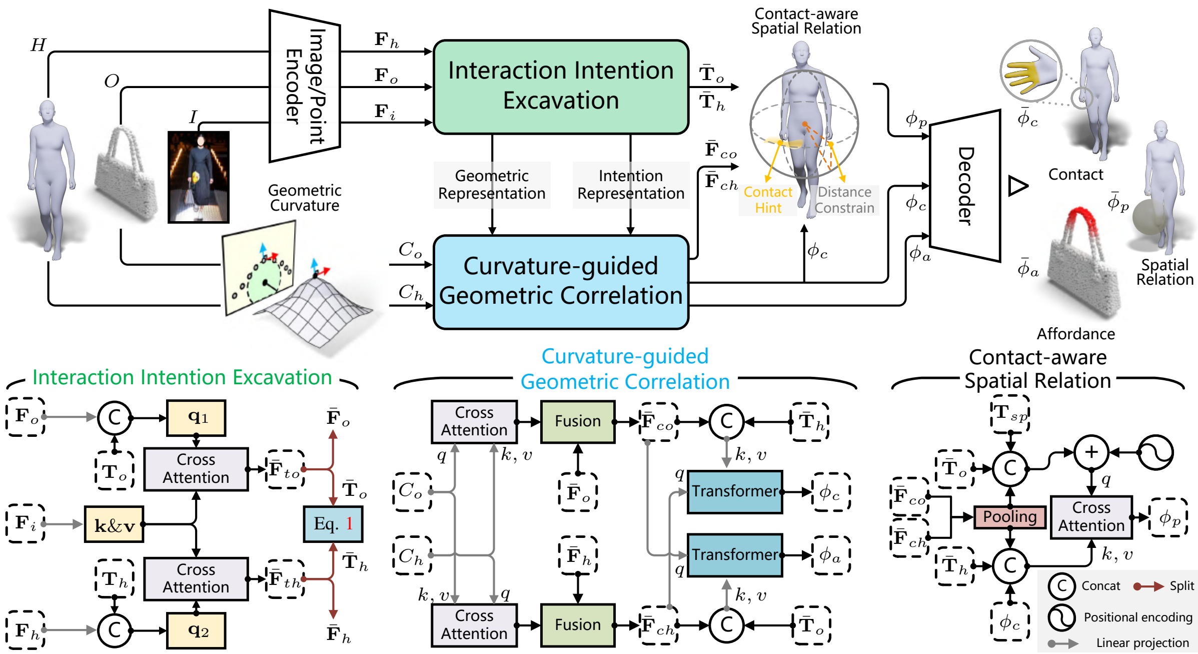

Figure 3. **LEMON pipeline.** Initially, it takes modality-specific backbones to extract respective features $\mathbf{F}_h, \mathbf{F}_o, \mathbf{F}_i$, which are then utilized to excavate intention features ($\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$) of the interaction (Sec. 3.1). With $\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$ as conditions, LEMON integrates curvatures (C_o, C_h) to model geometric correlations and reveal the contact ϕ_c, affordance ϕ_a features (Sec. 3.2). Following, the ϕ_c is injected into the calculation of the object spatial feature ϕ_p (Sec. 3.3). Eventually, the decoder projects ϕ_c, ϕ_a, ϕ_p to the final outputs $\bar{\phi}_c, \bar{\phi}_a, \bar{\phi}_p$.

semantic cosine similarity, a part of $\mathcal{L}_s$ in Sec. 3.4.

3.2. Curvature-guided Geometric Correlation

The interacting humans and objects exhibit certain geometry affinity [77], manifesting in matching geometric structures and correlative curvatures. For geometric curvatures, the normal curvature could better represent local structures like interaction regions [26, 49]. Thus, we encode the normal curvatures into geometric features, and regard the $\bar{\mathbf{T}}_o, \bar{\mathbf{T}}_h$ as conditions to capture the affinity among human and object geometries. Normal curvatures $C_o \in \mathbb{R}^{1\times N_o}$, $C_h \in \mathbb{R}^{1\times N_h}$ of object and human are obtained by local fitting method [74, 89]. To correlate the curvatures, C_o, C_h are encoded to high dimension $C_o^{'} \in \mathbb{R}^{C\times N_o}, C_h^{'} \in \mathbb{R}^{C\times N_h}$, and the cross-attention f_m is mutually performed on them, formulated as $\bar{C}_o = f_m(C_o^{'}, C_h^{'}), \bar{C}_h = f_m(C_h^{'}, C_o^{'})$. Following, to make the curvature guide the calculation of geometric correlation, $\bar{C}_o, \bar{C}_h$ are integrated and fused with geometric features:

$$\bar{\mathbf{F}}_{co} = f(\Gamma(\bar{C}_o, \bar{\mathbf{F}}_o)), \bar{\mathbf{F}}_{ch} = f(\Gamma(\bar{C}_h, \bar{\mathbf{F}}_h)), \tag{2}$$

where $\bar{\mathbf{F}}_{co}, \bar{\mathbf{F}}_o \in \mathbb{R}^{C\times N_o}$ and $\bar{\mathbf{F}}_{ch}, \bar{\mathbf{F}}_h \in \mathbb{R}^{C\times N_h}$. $\bar{\mathbf{F}}_o, \bar{\mathbf{F}}_h$ are split from $\bar{\mathbf{F}}_{to}$ and $\bar{\mathbf{F}}_{th}$, Γ denotes the concatenation, f indicates convolution layers with 1×1 kernel. Then, $\bar{\mathbf{T}}_o$ and $\bar{\mathbf{T}}_h$ are considered as conditions [41] to further screen candidate regions that match the interaction depicted in images and participate in the modeling of geometry affinity, revealing the affordance and contact features ϕ_a, ϕ_c. The process could be formulated as:

$$\phi_a = f_\theta(\bar{\mathbf{F}}_{co}, \Gamma(\bar{\mathbf{F}}_{ch}, \bar{\mathbf{T}}_o)), \phi_c = f_\theta(\bar{\mathbf{F}}_{ch}, \Gamma(\bar{\mathbf{F}}_{co}, \bar{\mathbf{T}}_h)), \tag{3}$$

where $\phi_a \in \mathbb{R}^{C\times N_o}$, to calculate it, $\bar{\mathbf{F}}_{co}$ serves as the query, and the concatenation of $\bar{\mathbf{F}}_{ch}$ and $\bar{\mathbf{T}}_o$ serves as the key and value, f_θ is the cross-transformer layer. $\phi_c \in \mathbb{R}^{C\times N_h}$ is obtained through the analogous way.

3.3. Contact-aware Spatial Relation

Human-object interactions are extremely diverse, rendering the reasoning of their 3D spatial relation very challenging. Nevertheless, human contact implicitly constrains the position of the object, thereby assisting in inferring plausible human-object spatial relation. Thus, the object's center is represented as a token sequence $\mathbf{T}_{sp} \in \mathbb{R}^{C\times 3}$, with introducing the contact feature as a constraint, LEMON takes the semantic intention and geometric features of the object to query the corresponding features of the human. Modeling $\mathbf{T}_{sp}$ as the object spatial feature, formulated as:

$$\phi_p = f_\rho(\Gamma(\Theta(\bar{\mathbf{F}}_{co}), \bar{\mathbf{T}}_o, \mathbf{T}_{sp}) + pe, \Gamma(\Theta(\bar{\mathbf{F}}_{ch}), \bar{\mathbf{T}}_h, \phi_c)), \tag{4}$$

where $\phi_p \in \mathbb{R}^{C\times 3}$, is the spatial feature of the object center, f_ρ is a cross-attention layer, Θ denotes the pooling layer, and pe indicates a learnable positional encoding.

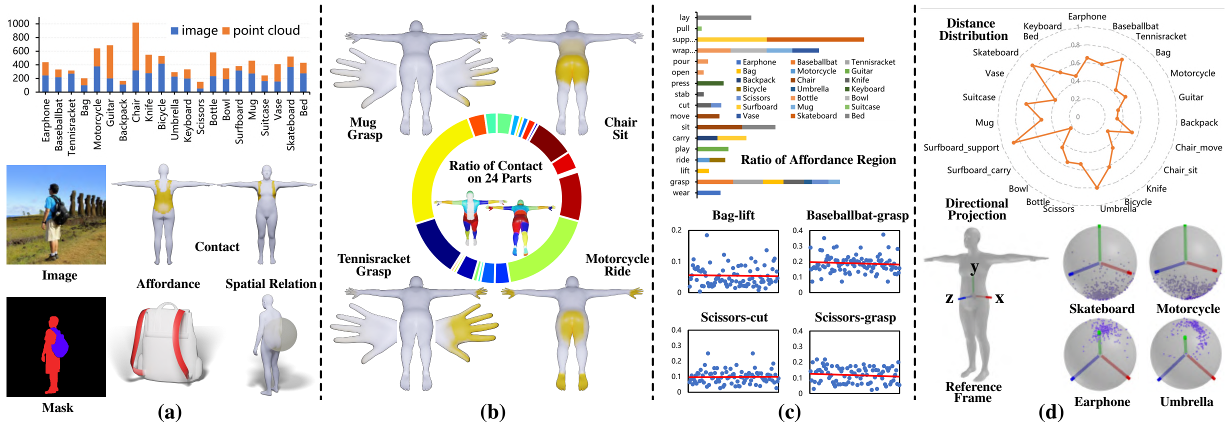

Figure 4. **3DIR Dataset.** **(a)** The quantity of images and point clouds for each object, and a data sample containing the image, mask, dense human contact annotation, 3D object with affordance annotation, and the fitted human mesh with the object proxy sphere. **(b)** The proportion of our contact annotations within 24 parts on SMPL [36], and distributions of contact vertices for certain HOIs. **(c)** The ratio of annotated affordance regions to the whole object geometries, and the distribution of this ratio for some categories. **(d)** Mean distances (unit: m) between annotated object centers and human pelvis joints, and directional projections of annotated centers for several objects.

3.4. Loss Functions

Eventually, ϕ_a, ϕ_c, ϕ_p are sent to the decoder that contains three projection heads with linear, normalization, and activation layers. Which output object affordance $\bar{\phi}_a \in \mathbb{R}^{N_o \times 1}$, human contact $\bar{\phi}_c \in \mathbb{R}^{N_h \times 1}$, and object's center position $\bar{\phi}_p \in \mathbb{R}^3$ respectively. With defined radii for specific objects, we treat a sphere as the object proxy [31], representing the spatial relation in the camera coordinates of fitted humans. The overall training loss is expressed as:

$$\mathcal{L} = \omega_1 \mathcal{L}_c + \omega_2 \mathcal{L}_a + \omega_3 \mathcal{L}_s + \omega_4 \mathcal{L}_p, \quad (5)$$

where ω_{1-4} are hyper-parameters to balance the losses. $\mathcal{L}_c$ and $\mathcal{L}_a$ possess the same formulation, a focal loss [34] combined with a dice loss [42]. $\mathcal{L}_c$ is calculated by $\bar{\phi}_c$ and the contact label $\hat{\phi}_c$, while $\mathcal{L}_a$ is calculated by $\bar{\phi}_a$ and the affordance annotation $\hat{\phi}_a$. $\mathcal{L}_s$ is employed to ensure $\bar{\mathbf{T}}_o$, $\bar{\mathbf{T}}_h$ align with interaction semantics of the image and maintain consistency in the semantic space. Specifically, $\mathbf{F}_i$ is mapped to a logit y, which is used to calculate a cross-entropy loss $\mathcal{L}_{ce}$ with interaction categorical label $\hat{y}$, $\mathcal{L}_s$ is the sum of $\mathcal{L}_{ce}$ and φ (Eq. 1). For 3D HOIs, object positions are diverse due to variations in human orientation, relying solely on absolute coordinates as supervision makes it hard for the model to learn a consistent mapping. However, the relative distances between humans and objects are similar for specific HOIs. This smoother distance distribution contributes to reducing the optimization space to an approximate sphere for specific HOIs. Consequently, we take the distance between the object center and the human pelvis joint as an additional constraint, and $\mathcal{L}_p$ is formulated as:

$$\mathcal{L}_p = \mathcal{L}_{pa} + \mathcal{L}_{pr}, \quad \mathcal{L}_{pa} = ||\bar{\phi}_p - \hat{\phi}_p||_2, \mathcal{L}_{pr} = ||\bar{\xi} - \hat{\xi}||_2, \quad (6)$$

where $\hat{\phi}_p$ is the annotation of the object center. $\bar{\xi}, \hat{\xi}$ denotes distances between the pelvis joint and $\bar{\phi}_p, \hat{\phi}_p$ respectively.

4. Dataset

We introduce the collection and annotation protocols of the 3DIR and give some statistical analysis of the collected data and annotations, shown in Fig. 4.

Collection. We collect images with explicit interaction contents, in which humans interact with specific objects. These images adhere to the condition that the human's upper body is present, ensuring the efficient recovery of the human mesh. In total, we collect 5k in-the-wild images from HAKE [32], V-COCO [18], PIAD [78], and websites with free licenses, spanning 21 object classes and 17 interaction categories. Additionally, we collect over 5k 3D object instances from several 3D datasets [9, 10, 35, 43], based on the category of objects in collected images.

Annotation. We make over 25k annotations with multiple types for the collected data. **1) Masks**: With the assistance of the SAM [25], we manually mask the interacting human and object in the image. **2) Human Mesh**: In natural images, some humans only exhibit the upper body. Thus, we utilize the pipeline of UBody [33] to fit SMPL-H pseudo-GTs. The 2D body and hand joints needed by the pipeline are obtained through the DWPose [79]. **3) Contact**: Similar to DAMON [64] and HOT [6], the human contact is annotated based on human knowledge. For each image, we clearly provide masks and a specific interaction type for the human-object pair, professional annotators are hired to "draw" vertices on the human template that contact with objects. The entire annotation process cycles three rounds, with each round incorporating subjective cross-checks and objective metric checks. **4) Affordance**: We refer to the 3D-AffordanceNet [10] to annotate the object affordance. The annotations of 11 objects included in 3D-AffordanceNet are directly utilized, and we

Table 1. **Comparison on the 3DIR.** Evaluation metrics of comparison methods on the benchmark, the best results are covered with the mask. ⋄ indicates the improvement relative to the baseline. P. means take PointNet++ [53] as the backbone, and D. means DGCNN [66].

Human Contact					Object Affordance				Spatial Relation	
Methods	**Precision** ↑	**Recall** ↑	**F1** ↑	**geo. (cm)** ↓	**Methods**	**AUC** ↑	**aIOU** ↑	**SIM** ↑	**Methods**	**MSE** ↓
Baseline	0.49	0.52	0.49	32.83	Baseline	82.36	32.63	0.50	Baseline	0.051
BSTRO [21]	0.57 ⋄16.3%	0.58 ⋄11.5%	0.55 ⋄12.2%	28.58 ⋄12.9%	3DA. [10]	85.49 ⋄3.8%	35.42 ⋄8.5%	0.56 ⋄12.0%	DJ-RN [31]	0.042 ⋄17.6%
DECO [64]	0.70 ⋄42.8%	0.72 ⋄38.4%	0.69 ⋄40.8%	15.25 ⋄53.5%	IAG [78]	86.63 ⋄5.1%	38.57 ⋄18.2%	0.59 ⋄18.0%	PopUp [51]	0.027 ⋄47.0%
Ours P.	0.76 ⋄55.1%	0.81 ⋄55.7%	0.77 ⋄57.1%	9.02 ⋄72.5%	Ours P.	87.91 ⋄6.7%	40.97 ⋄25.5%	0.63 ⋄26.0%	Ours P.	0.012 ⋄76.4%
Ours D.	0.78 ⋄59.1%	0.82 ⋄57.6%	0.78 ⋄59.1%	7.55 ⋄77.0%	Ours D.	88.51 ⋄7.4%	41.34 ⋄26.6%	0.64 ⋄28.0%	Ours D.	0.010 ⋄80.3%

annotate an additional 10 object categories excluded in 3D-AffordanceNet. **5) Spatial Relation**: For each sample, we color the fitted human mesh with annotated per-vertex human contact and treat a sphere as the object proxy. The radius of the sphere for each object category is pre-defined, referring to DJ-RN [31]. Given the posed human in color and the proxy sphere, annotators adjust the sphere relative to the human to align with the human-object spatial relation depicted in the image. Ultimately, we record the center coordinates of the adjusted sphere. Due to the page limitation, please refer to the Sup. Mat. for more annotation details.

Statistical Analysis. The quantity of images and 3D instances for each object category in the 3DIR is shown in Fig. 4 (a). For contact annotation, we count its proportion within 24 human parts defined on the SMPL [36]. Moreover, we visualize the distribution of contact annotation for several HOIs, where the deeper color indicates more contact annotations at the vertex, as shown in Fig. 4 (b). Fig. 4 (c) demonstrates the proportion of annotated affordance regions to the entire object geometry, encompassing the mean of each category and the distribution of all instances for several categories. As can be seen, there are differences in distinct objects with the same affordance, as well as variations in distinct affordances of the same object. For spatial relation annotation, we count the mean distance between annotated centers and human pelvis joints, and project the spatial direction of annotated centers onto a fixed-radius sphere with the pelvis joint as the center. Fig. 4 (d) shows several cases and the distribution of mean distances.

5. Experiment

5.1. Benchmark Setting

We refer to methods that anticipate each interaction element [21, 47, 64, 78] to thoroughly benchmark the 3DIR. For the training, in addition to the training data in 3DIR, we select another 5k data with low redundancy in BEHAVE [2]. Evaluation is conducted on the test set of 3DIR. The baseline model adopts a multitask-like framework, directly utilizing three branches to make anticipations. Besides, we compare LEMON with advanced methods that anticipate respective elements. Plus, we also conduct evaluation experiments on DAMON [64], BEHAVE [2], and PIAD [78], the results and implementation details are in the Sup. Mat.

5.2. Comparison Results

The comparison results of evaluation metrics are presented in Tab. 1. The baseline model, which does not capture correlations between interaction counterparts, indicates that directly anticipating these elements through multiple branches yields poor results. Our method outperforms comparative methods across all metrics for respective elements, demonstrating that leveraging correlations between humans and objects indeed benefits the comprehension of interaction relation, and the anticipation of interaction elements. It seems to be "the best of both worlds". Furthermore, we conduct a visual comparison of methods with higher evaluation metrics, as shown in Fig. 5. The results showcase multiple objects interacting with distinct human parts. As can be seen, our anticipations are more precise, for some uncertain regions that are not visible in images (*e.g.*, the first and third rows), LEMON could also anticipate plausible results. This is attributed to the modeling of correlations between geometries, which compensates for missing features in invisible regions of the image.

5.3. Ablation Study

We conduct a thorough ablation study to validate the effectiveness of the model design. Tab. 2 (a) reports the model performance without modeling the semantic intention and the geometric correlation, demonstrating their impacts on the model performance. Moreover, the absence of constraining semantic consistency hinders the modeling of intentions, while not introducing curvatures affects the extraction of geometric correlations, resulting in a decrease in model performance. Both results are detailed in Tab. 2 (a). Besides, we test the impact of ϕ_c and $\mathcal{L}_{pr}$ (Eq. 6) on spatial relation prediction, shown in Tab. 2 (b). Notably, removing the extraction of ϕ_a in f_θ (Eq. 3) also affects the performance. This is attributed to the interrelation between ϕ_c and ϕ_a, the absence of ϕ_a impacts the representation of ϕ_c, and subsequently influencing the final spatial prediction.

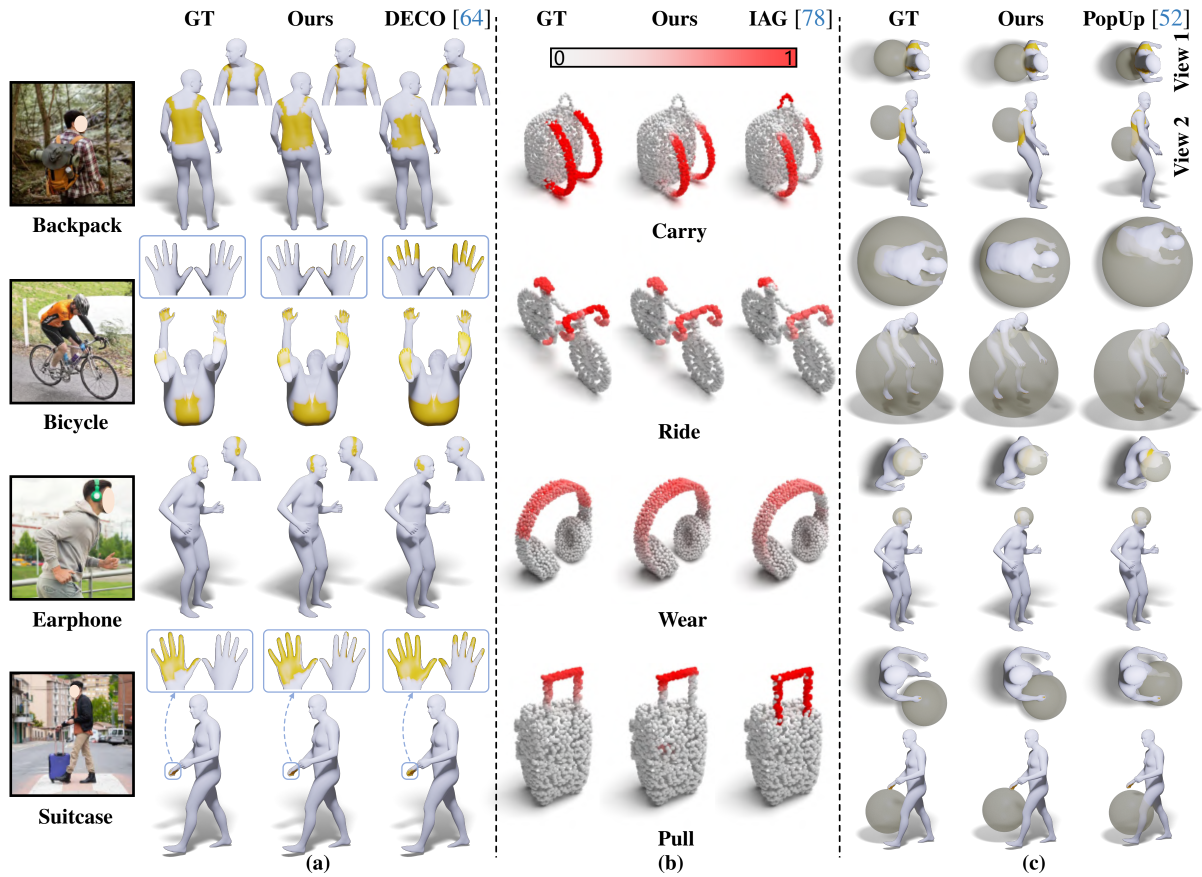

Figure 5. **Visualization Results.** **(a)** Results of the estimated human vertices in contact with objects, the estimated contact vertices are shown in yellow. **(b)** The anticipations of 3D object affordance, the depth of red represents the probability of anticipated affordance. **(c)** Two views of the predicted spatial relation, translucent spheres are object proxies. Please zoom in for a better visualization.

To further illustrate the effectiveness of capturing the intention and geometry affinity, we visualize attention maps on the human and object geometries when removing one of them, as shown in Fig. 6. The first row indicates without capturing intention affinity, solely relying on geometries leads the model to focus on multiple candidate regions, differing from attention regions in the image. Furthermore, the second row demonstrates that while semantic intention guides the model to identify approximate candidate regions, modeling geometry affinity further locates corresponding regions most related to the interaction of the geometries.

5.4. Performance Analysis

Multiple Interactions. To validate whether the model reasons interaction elements based on the understanding of interaction relation. We employ the model to infer different interactions between humans and the same object, as demonstrated in Fig. 7. The results showcase that, for the same object, when interactions are different, the model outputs distinct results, encompassing human contact, object affordance, and spatial relation, all of which align with the interaction relation depicted in the image. This indicates

Table 2. **Ablation studies.** **(a)** Performance when not modeling intention representations (intent.), geometric correlations (geom.), and not introducing φ (Eq. 1), curvatures (cur.). **(b)** The impact of several factors on spatial relation prediction. ✗ means without.

Metrics	Ours	✗ intent.	✗ φ	✗ geom.	✗ cur.
Precision	0.78	0.71	0.74	0.68	0.75
Recall	0.82	0.73	0.79	0.70	0.78
F1	0.78	0.70	0.74	0.67	0.75
geo. (cm)	7.55	11.87	10.26	14.87	9.13
AUC	88.51	85.87	86.66	83.23	86.98
aIOU	41.34	38.19	40.03	37.13	39.58
SIM	0.64	0.58	0.60	0.55	0.62
MSE	0.010	0.027	0.022	0.031	0.018

(a)

	MSE
✗ ϕ_c	0.037
✗ $\mathcal{L}_{pr}$	0.024
✗ ϕ_a	0.019

(b)

that the model infers interaction elements according to the interaction relation rather than relying on direct mapping solely driven by the categories of human-object pairs.

Multiple Objects. In scenarios where distinct objects are involved in concurrent interactions with humans, the model should indeed possess the ability to comprehend the interaction relation with different objects. We give an experi-

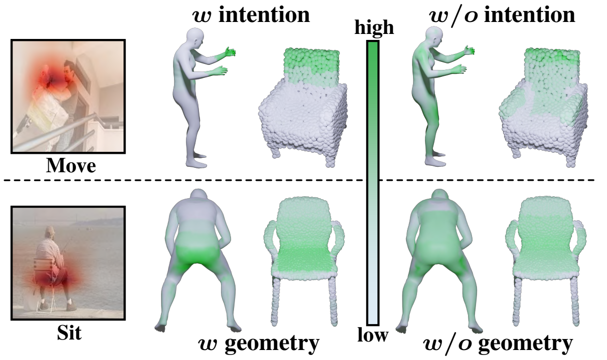

Figure 6. **Visual Attention.** Attention heatmaps on geometries, in cases with ($\boldsymbol{w}$) and without ($\boldsymbol{w/o}$) the extraction of intention and geometry affinity. Images are also shown with attention heatmaps.

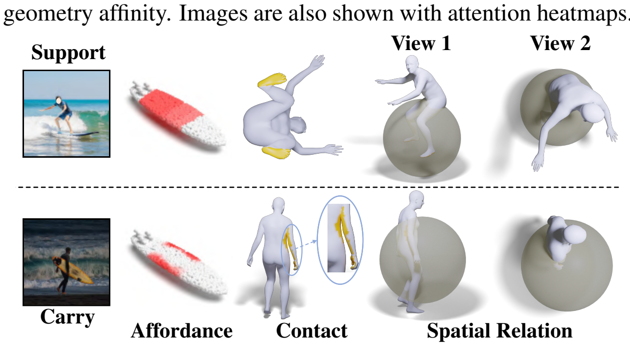

Figure 7. **Multiple Interactions.** Distinct anticipations when interactions are different with the same object.

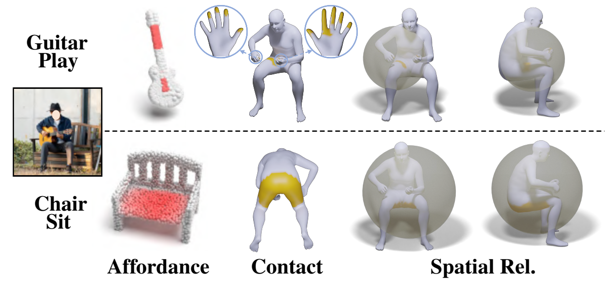

Figure 8. **Multiple Objects.** The anticipations for multiple objects with different human-object interactions.

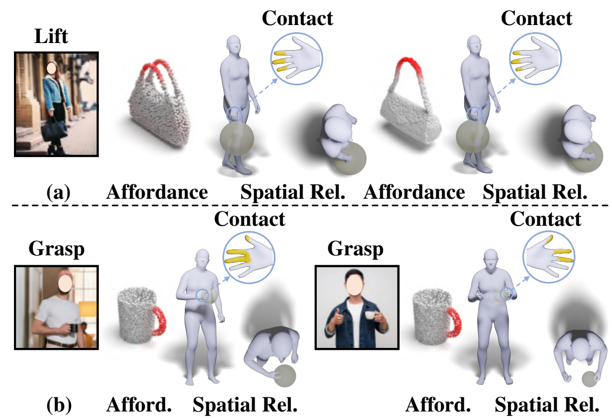

Figure 9. **Multiple Instances. (a)** Different object instances $w.r.t.$ the same image and human geometry. **(b)** Different images and human instances $w.r.t.$ the same object.

ment pertaining to this property. As shown in Fig. 8, when reasoning with different objects, the human contact, object affordance, and spatial relation are distinct, while all anticipations are plausible. This is attributed to the collaboration of the semantic interaction intention and geometric correlation, ensuring that the model could explicitly capture interaction intentions with different objects and reveal regions related to the interaction on corresponding geometries.

Multiple Instances. We conduct an experiment to validate the model's understanding of interaction relation across various instances, assessing its generalization and robustness. Specifically, distinct object instances are utilized to infer with the same image and human geometry, shown in Fig. 9 (a). As can be seen, even though the contact and spatial relation are slightly different, the results are consistent with the interaction shown in the image, and object affordances are correctly anticipated. Additionally, the same object instance is utilized to infer with different human geometries and images that contain similar human-object interactions, as demonstrated in Fig. 9 (b). These results indicate that the model could generalize the understanding of interaction relation to different human and object instances.

6. Conclusion

We propose leveraging both counterparts of the HOI to jointly anticipate human contact, object affordance, and human-object spatial relation in 3D space. It holds the potential to drive embodied intelligence to learn from interactions and assist the interaction modeling in graphics. With the provided dataset 3DIR containing paired HOI data and multiple annotations, we train LEMON, a novel model that harnesses the interaction intention and geometric correlation between humans and objects to capture the affinity, eliminating interaction uncertainties and anticipating plausible interaction elements. Extensive experiments demonstrate the effectiveness and superiority of LEMON. We believe this work offers fresh insights and paves a new way for 3D human-object interaction understanding.

Limitations and Future Work. In the current formulation, LEMON needs the pre-inferred human mesh as input. In the future, we plan to integrate HMR into the entire framework, freeing input constraints required for inference, and taking 3D interaction elements to improve the accuracy of HMR. Plus, another interesting direction is to leverage multi-modal methods [15, 88], boosting the interaction relation understanding from other sources *e.g.*, text, audio.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207 and 62306295.

References

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009. 1

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 6

[3] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, pages 348–353. IEEE, 2000. 3

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2

[5] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 1

[6] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 5

[7] Kun-Hung Cheng and Chin-Chung Tsai. Affordances of augmented reality in science learning: Suggestions for future research. *Journal of science education and technology*, 22:449–462, 2013. 1

[8] Henry M Clever, Zackory Erickson, Ariel Kapusta, Greg Turk, Karen Liu, and Charles C Kemp. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6215–6224, 2020. 3

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 5

[10] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021. 1, 3, 5, 6

[11] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 3

[12] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018. 3

[13] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 1

[14] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 2

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 8

[16] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 2

[17] Elena Corina Grigore, Kerstin Eder, Anthony G Pipe, Chris Melhuish, and Ute Leonards. Joint action understanding improves robot-to-human object handover. In *2013 IEEE/RSJ international conference on intelligent robots and systems*, pages 4622–4629. IEEE, 2013. 1

[18] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5

[19] Sookwan Han and Hanbyul Joo. Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 3

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 1, 3

[21] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 1, 3, 6

[22] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3

[23] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 3

[24] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017. 1

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5

[26] Jan J Koenderink and Andrea J Van Doorn. Surface shape and curvature scales. *Image and vision computing*, 10(8): 557–564, 1992. 4

[27] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis, 2023. 3

[28] Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12032–12042, 2023. 3

[29] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *arXiv preprint arXiv:2309.16237*, 2023. 3

[30] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[31] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 1, 3, 5, 6

[32] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 5

[33] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023. 5

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[35] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 5

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3, 5, 6

[37] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2252–2261, 2022. 3

[38] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Leverage interactive affinity for affordance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6809–6819, 2023. 3

[39] Bertram F Malle and Joshua Knobe. The folk concept of intentionality. *Journal of experimental social psychology*, 33 (2):101–121, 1997. 2

[40] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 6169–6176. IEEE, 2021. 3

[41] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 4

[42] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[43] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 5

[44] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 1, 3

[45] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on Robot Learning*, pages 1666–1677. PMLR, 2022. 3

[46] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. *Advances in Neural Information Processing Systems*, 33:2005–2015, 2020. 1, 3

[47] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 3, 6

[48] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 33:7841–7851, 2020. 3

[49] Abdul Nurunnabi, Geoff West, and David Belton. Outlier detection and robust normal-curvature estimation in mobile laser scanning 3d point cloud data. *Pattern Recognition*, 48 (4):1404–1419, 2015. 4

[50] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3

[51] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4726–4736, 2023. 3, 6

[52] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 7

[53] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 6

[54] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12802–12813, 2023. 3

[55] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 71–87. Springer, 2020. 3

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[57] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 3

[58] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 1

[59] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 3

[60] Thomas B Sheridan. Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532, 2016. 1

[61] Chris Shilling. The body in culture, technology and society. *The Body in Culture, Technology and Society*, pages 1–256, 2004. 2

[62] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision*, pages 516–533. Springer, 2022. 1, 3

[63] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005. 2

[64] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8001–8013, 2023. 1, 3, 5, 6, 7

[65] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 3

[66] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 3, 6

[67] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *European Conference on Computer Vision*, pages 90–107. Springer, 2022. 1, 3

[68] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3272–3279, 2013. 1

[69] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. *arXiv preprint arXiv:2303.11057*, 2023. 3

[70] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023. 1

[71] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3

[72] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[73] Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. 3

[74] Rui Xu, Zhiyang Dou, Ningna Wang, Shiqing Xin, Shuangmin Chen, Mingyan Jiang, Xiaohu Guo, Wenping Wang, and Changhe Tu. Globally consistent normal orientation for point clouds by regularizing the winding-number field. *ACM Transactions on Graphics (TOG)*, 42(4):1–15, 2023. 4

[75] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 3

[76] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3

[77] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 4

[78] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10905–10915, 2023. 1, 3, 5, 6, 7

[79] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. *arXiv preprint arXiv:2307.15880*, 2023. 5

[80] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2010. 1

[81] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1, 3

[82] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12976, 2023. 3

[83] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023. 3

[84] Guochen Yu, Andong Li, Chengshi Zheng, Yinuo Guo, Yutian Wang, and Hui Wang. Dual-branch attention-in-attention transformer for single-channel speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7847–7851. IEEE, 2022. 3

[85] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10): 2472–2500, 2022. 3

[86] Wei Zhai, Pingyu Wu, Kai Zhu, Yang Cao, Feng Wu, and Zheng-Jun Zha. Background activation suppression for weakly supervised object localization and semantic segmentation. *International Journal of Computer Vision*, pages 1–26, 2023. 3

[87] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 3

[88] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 8

[89] Xiaopeng Zhang, Hongjun Li, Zhanglin Cheng, and Yikuan Zhang. Robust curvature estimation and geometry analysis of 3d point cloud surfaces. *J. Inf. Comput. Sci*, 6(5):1983–1990, 2009. 4

[90] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023. 3

[91] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *arXiv preprint arXiv:2207.01971*, 2022. 3

[92] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper manipulation. In *International Conference on Learning Representations*, 2023. 1

[93] Xinghao Zhu, JingHan Ke, Zhixuan Xu, Zhixin Sun, Bizhe Bai, Jun Lv, Qingtao Liu, Yuwei Zeng, Qi Ye, Cewu Lu, et al. Diff-lfd: Contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering. In *7th Annual Conference on Robot Learning*, 2023. 3

[94] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834, 2021. 2