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

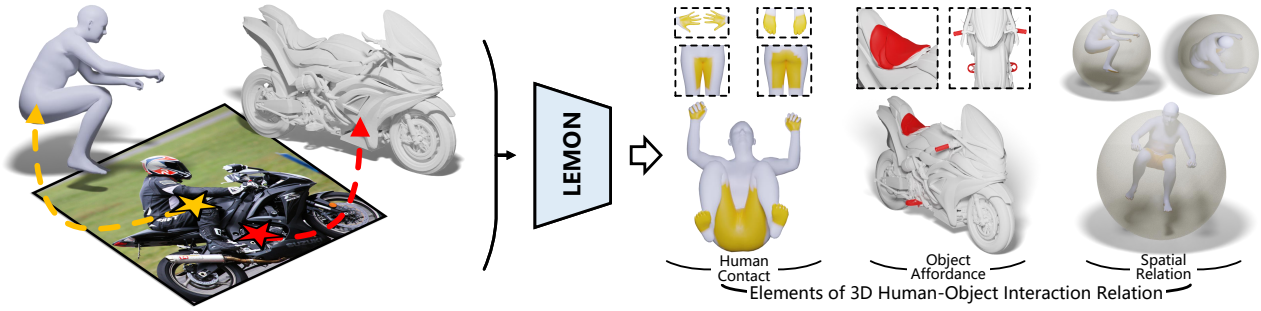

Figure 1. For an interaction image with paired geometries of the human and object, LEMON learns 3D human-object interaction relation by jointly anticipating the interaction elements, including human contact, object affordance, and human-object spatial relation. Vertices in yellow denote those in contact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

Abstract

*Our work, **LEMON**, aims to understand 3D human-object interactions (HOI) from a novel perspective. HOI understanding is a fundamental task in computer vision. It is pivotal for fields like embodied AI and graphics. How to enable machines to comprehend HOIs has perennially remained a pivotal crux. Some approaches take proxy tasks, such as HOI detection and image caption, to capture semantics pertaining to interactions. However, in real application environments, whether real or virtual world, interactions mostly occur within 3D space. Mere comprehension of semantic interaction aspects is inadequate for supporting downstream applications. Spatial priors related to the interaction should be additionally captured. Some studies anticipate interaction elements* e.g.*, human contact and object affordance in 3D space, but they isolate the human and object, which underexploits certain correlations between interaction counterparts and struggles to address the uncertainty inherent in interactions.*

In this work, 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). We focus on perceiving certain elements capable of revealing the interaction, thereby facilitating the machine to learn and understand HOIs, and providing spatial priors for downstream applications. LEMON addresses the interaction uncertainty by unearthing the semantic and geometric affinities in-between the interacting humans and objects. Actually, humans and objects are intertwined and possess affinities in the interaction. In specific, the design of objects typically adheres to certain human needs. Therefore, object affordances inherently hint at "what" interactions humans intend to make, revealing the intention affinity. Meanwhile, the interacting human and object exhibit matching geometries (either posture or configuration), which present "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 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. LEMON employs multibranch 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. Besides, LEMON integrates geometric curvatures to capture the geometry affinity and reveal the representation of human contact and object affordance. These representations then assist in anticipating the spatial relation constrained by a combined distance loss.

*To support the proposed task, we collect the **3DIR** dataset, which contains natural HOI images paired with 3D objects and SMPL-H pseudo-GTs. We also make multiple annotations for the data, including human contact, 3D object affordance, and relative human-object spatial relation. It serves as the test bed for the model training and evaluation. We believe this work offers fresh insights and paves a new way for 3D human-object interaction understanding.*

LEMON gets accepted to CVPR2024.