Title

Hierarchical Modeling of Human-Object Interactions: from Concurrent Action Parsing to Physics-Based Grasping

Author

Liu, Tengyu

Publication Date

2021

UNIVERSITY OF CALIFORNIA

Los Angeles

Hierarchical Modeling of Human-Object Interactions:
from Concurrent Action Parsing to Physics-Based Grasping

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Tengyu Liu

2021

ABSTRACT OF THE DISSERTATION

Hierarchical Modeling of Human-Object Interactions: from Concurrent Action Parsing to Physics-Based Grasping

by

Tengyu Liu
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2021
Professor Song-Chun Zhu, Chair

The study of human-object interaction (HOI) aims at modeling the geometric relationship between a human and an object in an interaction. Understanding HOI is an essential step towards holistic scene understanding and generating realistic scenarios that involve humans. Conventionally, the study of HOI focuses on detecting and classifying instance-level HOI on 2D images. Given an image, an example output would be a triplet ⟨person, chair, sit⟩, or ⟨person, apple, eat⟩, where the person, chair, and apple are all represented by bounding boxes. This dissertation aims to understand HOI in 3D.

Extending HOI to 3D faces two significant challenges. The first challenge lies in the difficulty of obtaining high-fidelity 3D annotation of HOI data. Existing methods of collecting 3D datasets all suffer from high occlusion, poor resolution, and high annotation costs. Another critical challenge in modeling 3D HOI lies in the representation of the objects. Existing methods treat each object as a unity, usually represented as an axis-aligned bounding box. Such methods ignore the complexity of objects' shapes and therefore fail to model complex geometrical relationships in HOIs such as sitting. The root cause of this challenge traces back to the first challenge, where we do not have the high-fidelity data necessary to reflect the details in object shapes.

This dissertation addresses both challenges by collecting a large-scale high-fidelity 3D HOI dataset and by proposing hierarchical modeling of HOI. By using instance-level HOI annotation, our dataset improves scene reconstruction performance by a significant margin. This high-fidelity nature of the collected dataset enables part-level HOI modeling, which addresses the second challenge. This dissertation also addresses the second challenge by decomposing shape-level HOI into physics-level, which significantly improves the quality and robustness of grasp synthesis.

The dissertation of Tengyu Liu is approved.

Kai-Wei Chang

Demetri Terzopoulos

Ying Nian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2021

In memory of my late father, who showed me to always keep calm and never stop fighting.

To my mother, for her unconditional and unreserved love and support to me.

I am forever in debt.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Song-Chun Zhu, for his tremendous help in my Ph.D. program. I am in debt to his overwhelming support to my academic career. His vision and quest in computer vision, cognition, and theory of mind will be an invaluable guide for my future research. His dedication and undisturbed focus will be my lifelong inspiration.

I would also like to thank my committee members: Professor Ying Nian Wu, Professor Demetri Terzopoulos, Professor Kai-Wei Chang. This dissertation would not be possible without their guidance and support.

In addition, I would like to thank all my labmates at VCLA. They are all responsible for my growth as a researcher and as a person. My Ph.D. career would be miserable without them. Particularly, I would like to thank Dr. Yuanlu Xu, Dr. Jianwen Xie, and Dr. Yixin Zhu for their generous guidance and support. I would like to thank Dr. Siyuan Huang, Dr. Siyuan Qi, Dr. Zilong Zheng, Dr. Hangxin Liu, Dr. Ruiqi Gao, Yifei Xu, Hanlin Zhu, Yixin Chen, Xu Xie, Baoxiong Jia, Pan Lu, and many others for all the endless discussions and debates, as well as for all the fun that we have had together.

I would like to thank my best friends Dr. Zhengxiang Yi, Wenxuan Mao, Zeyu Li, and Lynn Zhang for helping me through the darkest times.

Finally, I would like to thank my family and my fiancée, Xinyi Wu, for their endless love and support, and for tolerating my emotions and stress.

Portions of this dissertation are supported by ONR N00014-19-1-2153, ONR MURI N00014-16-1-2007, and DARPA XAI N66001-17-2-4029.

VITA

2015	B.S. (Computer Science), UIUC.
2018	M.S. (Computer Science), UCLA.
2018–2019	Research Assistant, Computer Science Department, UCLA.
2019–2020	Teaching Assistant, Computer Science Department, UCLA.

PUBLICATIONS

Monocular 3D Pose Estimation via Pose Grammar and Data Augmentation, Yuanlu Xu, Wenguan Wang, Tengyu Liu, Xiaobai Liu, Jianwen Xie and Song-Chun Zhu, IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 2021.

Synthesizing Diverse and Physically Stable Grasps with Arbitrary Hand Structures by Differentiable Force Closure Estimation , Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu and Song-Chun Zhu, IEEE Robotics and Automation Letters (RA-L) , 2021.

CHAPTER 1

Introduction

The study of computer vision is to answer the question of “what” and “where” given 2D images. With the help of deep learning and large-scale image datasets, modern computer vision systems are getting comfortable answering the question of “what” but are struggling to place the detected human and objects in 3D spaces correctly.

Unlike modern computer vision algorithms, the ability to understand the 3D relationships between humans and objects from 2D signals is innate to human beings. One could argue that human vision relies on the stereo vision for depth information. However, we can still easily tell the 3D arrangement of a scene when given its 2D projection, *i.e.*, a picture or a video. If we compare the current computer vision system to our human vision system, both systems can detect, classify, and localize objects in 2D images. What is the difference, then? It appears that our human mind can leverage *commonsense* to determine the best 3D arrangement that satisfies both our commonsense and the visual cues. Those that do not align with our commonsense are known to create illusions.

Human visual commonsense includes various priors, including geometrical relationships between objects, object motion, object persistence, *etc.* This dissertation focuses on the 3D geometrical relationship between a human body and an object in a human-object interaction (HOI). An HOI can be anything involving a human and an object. Examples include sitting in a chair, drinking from a coffee mug, or working on a computer.

Most existing studies of HOI focus on HOI detection on 2D images. With the development of deep learning and large-scale datasets [CWH15, CLL18], 2D HOI detection becomes another task that can be learned in an end-to-end fashion [ML16, CLL18, GGD18, QWJ18]. These methods represent both the human and the object by their image patches and determine the interaction by joining image features of both patches in some way. By joining the image features of both patches, these methods capture both coarse level and fine-grained level information implicitly.

In 3D, on the other hand, the study of HOI has only gained a few tractions [WZZ13, WZZ17]. These methods model the human as a 3D skeleton and the object as a 3D axis-aligned bounding box. The interaction is modeled as a geometrical relationship between a critical body part and the position of the object bounding box. By representing an object using its axis-aligned bounding box, all the details of the object shape are removed. The models, therefore, are bounded to learn only a highly abstracted geometrical relationship for each HOI.

The main challenge that stands between researchers and understanding 3D HOI is two-fold. Firstly, it is extremely expensive to collect high-fidelity 3D HOI data with accurate shapes of both humans and objects. Secondly, human-object interaction is usually represented on the instance level, where an HOI is a triplet $\langle$human, object verb$\rangle$. While the instance-level representation has served 2D HOI detection well, it fails to model the complex geometrical relationship between 3D shapes, such as a person sitting in a chair. Instead, the geometrical relationship should be modeled between shapes or parts.

This dissertation addresses the first challenge by proposing SHADE, a high-fidelity dynamic 3D HOI dataset collected from well-established 3D video game assets. SHADE improves the SOTA performance of 3D human pose estimation and 3D scene reconstruction by a significant margin. We address the second challenge by modeling HOI in a hierarchical fashion. By decomposing human-object interactions into part-level and physics-level interactions, the models are capable of generative diverse, and realistic HOI snapshots that are otherwise difficult if not impossible to generate otherwise.

In this dissertation, an HOI node is defined as a component of human activity when

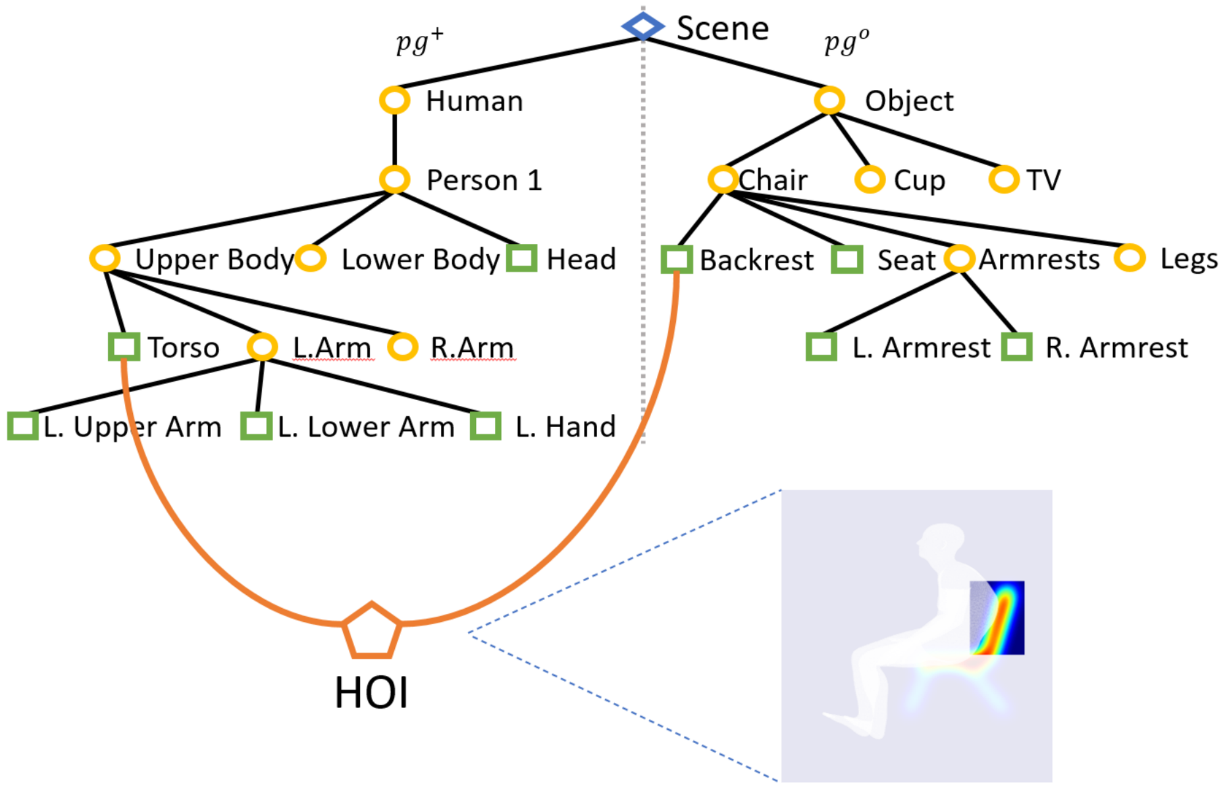


Figure 1.1: An illustrative example spatial parse graph of a sitting scene. We only show the HOI node between torso and backrest for clarity. The blue diamond is the root node of the scene pg. Each terminal node is represented by a green sphere, and each non-terminal node is represented by a yellow circle. The orange pentagon shows the HOI node. The geometric relationship described by the HOI is illustrated by the heat map on the lower-right.

human-object interaction is involved. It describes the geometric relationship between both parties of an interaction. When we represent a scene as a spatial parse graph (pg) where each entity is a node, the HOI nodes are connections between those nodes that provide spatial constraints to help localize each other. Figure 1.1 shows an illustration of a spatial parse graph of a sitting scene.

1.1 Definition

In order to formally define an HOI node, we first need to define the spatial parse graph pg^S and action parse graph pg^A. A spatial parse graph can be described as a pair

$$pg^S = (pg^o, pg^+) \tag{1.1}$$

, where the object pg $pg^o = \{V^o, E^o, \psi\}$ represents the attributed parse graph for objects, and the human pg $pg^+ = \{V^+, E^+, \phi\}$ represents the attributed parse graph for human. V^o and V^+ are the sets of object nodes and human nodes, respectively. E^o and E^+ are edges within object pg and human pg, respectively. $E^o = \{(v_1, v_2) : v_1 \in V^o, v_2 \in V^o\}$ is a set of edges where v_2 is a part of v_1. E^+ is defined similarly on the human pg. $\psi : V^o \mapsto X^o$ is the function of object node features, where X^o is the object feature space. Similarly, ϕ defines the function of human node features.

An action parse graph $pg^A = \{V^N, V^I, E^N\}$ describes the decomposition of human actions, where V^N is the set of human action nodes, and V^I is the set of HOI nodes. $E^N = \{(v_1, v_2), v_1 \in V^N, v_2 \in V^N \cup V^I\}$ describes the decomposition of action nodes.

An HOI node is therefore defined as $v^I = (v^o, v^+, f)$, where v^o is an object node, v^+ is a human node, and $f : X^o \times X^+ \mapsto \mathbb{R}$ is the energy function describing the geometrical relationship between the two nodes. An HOI node functions as a bridge between the human parse graph and the object parse graph, and a bridge between the spatial parse graph and the action parse graph.

1.2 Formulation

We can then formulate scene understanding as finding the optimal parse graph given an image

$$\begin{aligned} pg^* &= \arg\max_{pg} P(pg|I) && (1.2) \\ &= \arg\max_{pg} \frac{P(I|pg)P(pg)}{P(I)} && (1.3) \\ &= \arg\max_{pg} \frac{1}{Z} \exp\{-\mathcal{E}(I|pg) - \mathcal{E}^S(pg^S) - \mathcal{E}^I(pg^S, V^I)\} && (1.4) \end{aligned}$$

, where Z is the intractable normalizing constant. Here, $\mathcal{E}(I|pg)$ is the grounding energy that can be computed by off-the-shelf human and object detectors. $\mathcal{E}^S(pg^S)$ is the prior energy for spatial configuration. This term usually contains human pose priors, physics constraints, functionality constraints, *etc.* $\mathcal{E}^I(pg^S, V^I) = \sum_{(v_i^o, v_i^+, f_i) \in V^I} f_i(\psi(v_i^o), \phi(v_i^+))$ describes the energy of geometric relationships between the interacting parties.

We can then describe different tasks using this formulation. In addition to scene understanding which is finding $\arg\max_{pg} P(pg|I)$ where I is the input image, hallucinating human in a 3D scene can be solved by sampling from $\tilde{pg}^+ \sim P(pg^+|pg^o)$, and HOI detection can be formulated as finding $\arg\max_{V^I} P(pg|I)$.

CHAPTER 2

Video Game Exploit: Hacking Game Assets to Learn 3D Human-Object Interactions

A major challenge in studying 3D human-object interaction is the lack of high-quality data. Existing human action datasets are either in 2D [SZS12, SVW16] where each frame is an RGB image, or in 2.5D [KGS13, SLN16] where each frame is an RGB image associated with a depth map. These datasets fail to describe the accurate 3D shapes of both the human and the objects. With expensive motion capture systems, 3D human action datasets [IPO14] can be collected. However, these datasets fail to collect accurate 3D shapes of the objects interacting with a human. Without a high-fidelity 3D HOI dataset, it is challenging to model HOI in 3D, let alone break down HOI into smaller pieces. This chapter addresses this challenge by collecting a dynamic 3D HOI dataset from video game assets.

2.1 Introduction

Understanding the geometric relationships in human-object interactions (HOI) is beneficial to many real-life tasks such as robot grasping, surveillance, human activity analysis, and object detection. Although we have seen rapid growth in the analysis of 3D humans and 3D scenes over the past few years, very few works focus on modeling the interaction between human and dynamic objects in 3D.

Notations. In this chapter, we use the term *static object* for objects that do not move over time and *dynamic object* for objects whose position or orientation changes over time.

This chapter only considers objects that move due to HOI and neglect other factors such as gravity.

The difficulty lying behind the challenge of modeling dynamic human-object interactions is mainly two-fold:

Heavy occlusion. When a person interacts with an object, it is natural for the person to partially, if not wholly, occlude it in front of a camera. Smaller objects such as a cup or a pen are very likely to be completely occluded by the interacting person. This creates formidable challenges for detection-based object localization algorithms. [WZZ13] argues that we can predict dynamic object location from the 3D skeleton of the interacting human, whose estimation has been widely studied. However, detailed annotation of 3D object locations is exceptionally difficult to acquire in real life, which leads to the next difficulty.

Data scarcity. Existing datasets are most likely focused on two aspects separately: dynamic human analysis [IPO14] or static scene analysis [SYZ17]. Although some existing datasets [SCH14, WZZ17] do contain 3D human-object interactions, they lack either the annotation of dynamic object location or annotation of object location in general. In [KGS13] they provide annotation of dynamic object location but are limited by its data complexity as well as its annotation granularity since it only contains 3D positions of 15 joints and does not provide the 3D geometry ground truth for objects.

In this chapter, we present a large-scale dataset SHADE (Synthetic Human Activities with Dynamic Environment) to alleviate both difficulties by utilizing the graphics engine in a video game containing abundant human-object interactions. Our dataset tracks the 3D skeleton of every human accurate to three knuckle joints in each finger and contains real-time 3D position, orientation, and geometry of every object as small as a potato chip. Our experiment reveals three properties of our dataset: i) modeling of human-object interaction provides a significant edge to understanding human behavior; ii) the geometric relationship in human-object interactions can be generalized to real-world human activities; iii) in addition to having more detailed annotation, the human skeleton in our dataset is a complement to

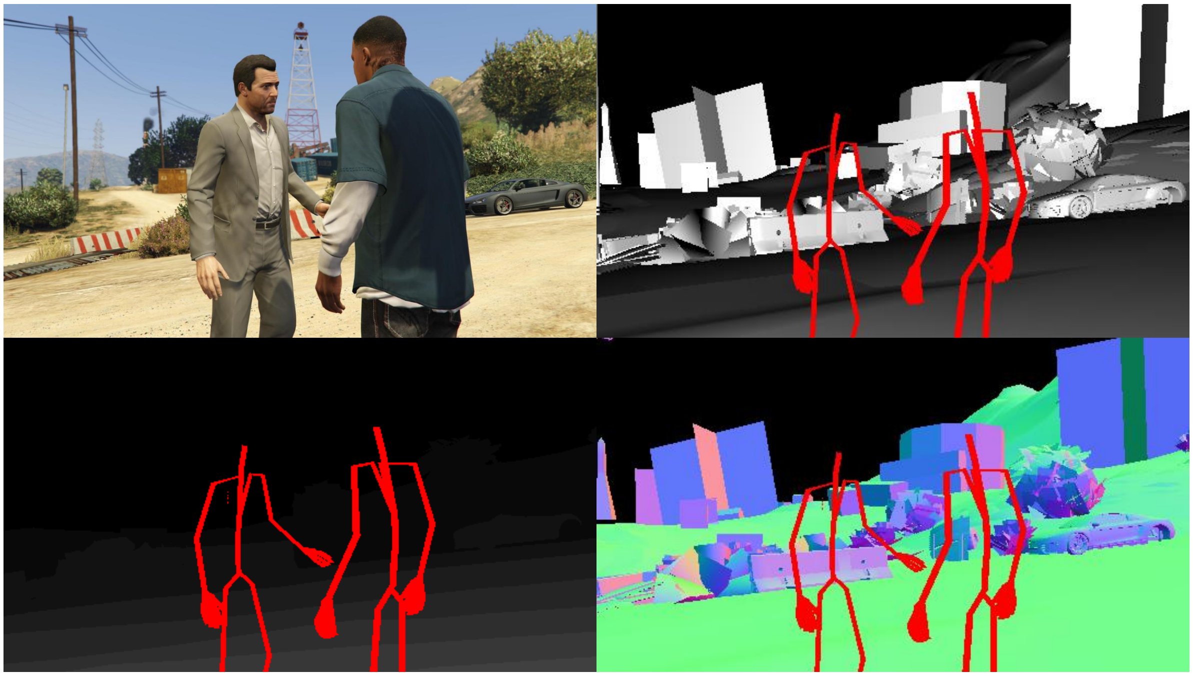


Figure 2.1: Illustration of photo-realistic synthetic data in our SHADE dataset.

other public human pose datasets.

We conduct experiments on three vision tasks: HOI recognition, object localization, and 3D pose estimation using our dataset, and show improved performance to existing methods when used as external training data.

2.2 Related Work

The work presented in this chapter is closely related to the following three research streams.

2D/3D human-object interaction. Rather than detecting objects or estimating articulated human pose individually, recognizing human-object interactions (HOIs) requires a deeper and more comprehensive understanding of the mutual spatial structure information and rich semantic relations between humans and objects. HOI recognition has gained increasing research interests over the past few years. With the popularity of deep learning techniques in computer vision, various network architectures [ML16, CLL18, GGD18, QWJ18]

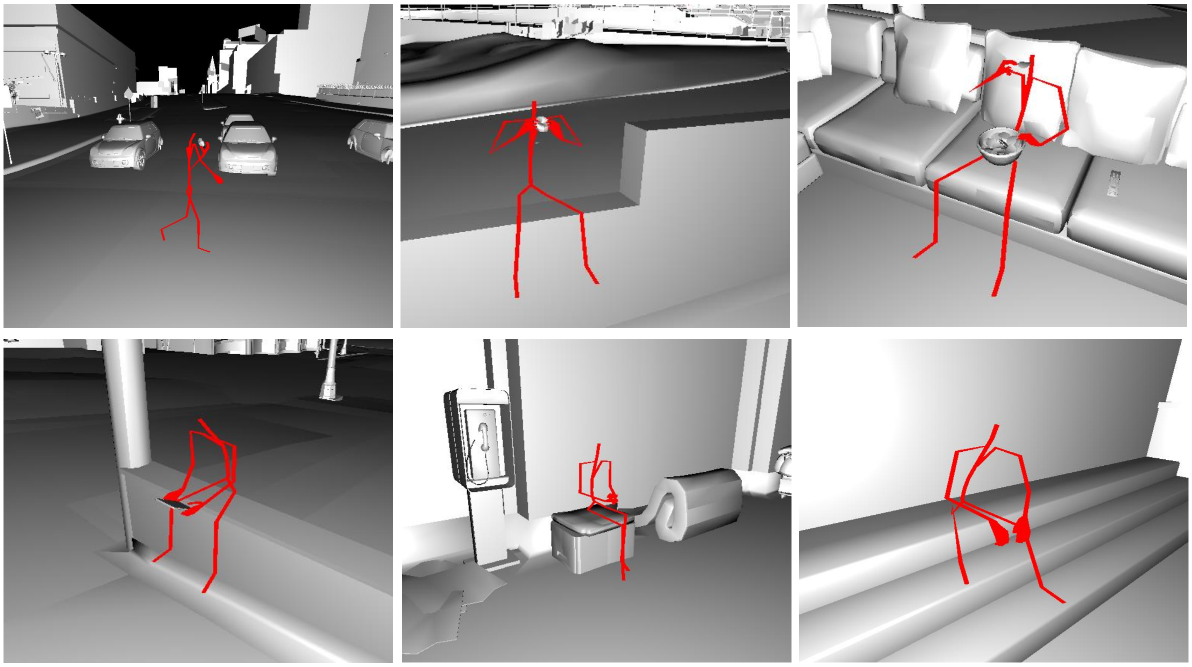


Figure 2.2: Illustration of variances of the same action. The first row belongs to eat and the second row belongs to sit.

were explored for tackling this task. Some large-scale 2D datasets [CWH15, CLL18] were also proposed to support the training of deep HOI models.

However, most of the previous attempts focused on HOI recognition in 2D images. Only a few methods [WZZ13, WZZ17] were proposed for modeling HOI in 3D scenes. Despite the difficulties brought from the extra dimension, the lack of a large-scale, well-annotated 3D HOI dataset severely restricted the development of 3D HOI recognition. This chapter proposes a large-scale, synthetic 3D dataset for HOI recognition, which is long-time urged in this field. We believe that this dataset would open up new possibilities for moving HOI recognition and analysis into 3D.

Action recognition. There are two main streams in current action recognition literature: appearance-based methods and skeleton-based methods. Similar to HOI recognition methods, researches in appearance-based action recognition have moved from hand-crafted features to learning deep features with neural networks [JXY13, SZ14, KTS14]. Recently, appearance-based action recognition methods [ZLS17, HBE17] have seen significant improvement in both classification accuracy and generalization capability by incorporating contextual information. Skeleton-based methods [WW17, YXL18], on the other hand, are more robust against appearance and lighting changes since they ignore image features altogether. However, the use of contextual information is very limited in skeleton-based action recognition, largely due to the lack of well-annotated data.

Object localization has long been a challenging task for computer vision. In 2D object localization, a common practice is to use a sliding window and run object detection algorithm on each window. This stream naturally extends to convolutional neural networks. Others [OBL15, TGJ15] regress heatmaps of object presence on images directly. The recent development of convolutional neural networks has yielded a huge leap [HGD17] in 2D object localization by extending and combining both ideas into the region of interest (ROI) operations. In 3D, however, object localization remains challenging due to the cubic growth of data size brought by the extra dimension. [SX16] extended the sliding window to 2.5D by applying a convolutional neural network on RGB-D images. However, such a method is sensitive to occlusion.

In addition, many works [ISS17, LBM17, HQZ18] have been done to estimate the static scene layout given a 2D image. However, small and dynamic object localization in 3D has yet to be addressed due to the lack of data.

2.3 Proposed Framework

In this section, we will describe the framework we use to collect data. As illustrated in Fig. 2.3, we first seek a video game environment that simulates people's daily activities and then develop a plugin to fetch the critical game assets from the graphics engine.

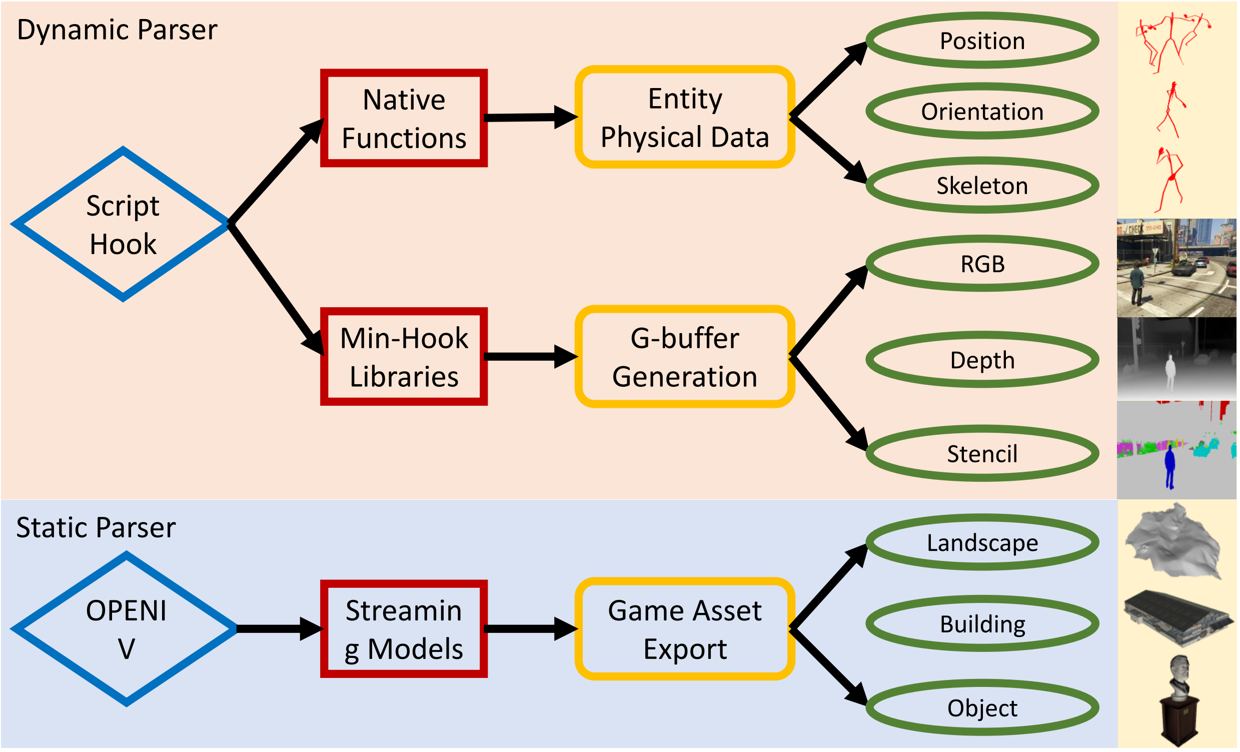


Figure 2.3: Pipeline of our data collection pipeline. Given the video game GTA V, our data acquisition plugin operates on accessible gaming interfaces, parsing and fetching both static (e.g., objects, buildings, landscapes) and dynamic (e.g., actions, interactions, cinematic videos) gaming resources.

2.3.1 Photo-Realistic Physics-Realistic Synthetic Game Environment

Although human activities involving objects are ubiquitous in daily life, the effort to record such fruitful interaction data to a fine-grained level remains challenging.

Some resort to optical motion capturing systems for target localization, *e.g.*, VICON cameras [IPO14, SBB10]. Others make tactile sensors to estimate hand pose during an interaction, such as tactile gloves [LXM17]. These approaches require an elaborated system set up to serve real-time data recording. In our approach, instead, we build our data acquisition pipeline based on a video game platform – Grand Theft Auto V (GTA V). Unlike other video games that simplify human-object interaction dynamics, GTA V is well-known for its richness in photo- and physics-realistic daily activities. In this video game, abundant human-object interaction events are incorporated. For instance, we can see a human agent walking in the street eating a sandwich and another human agent sitting on a low wall reading from a tablet. In order to obtain the interaction data of agents and objects, we develop a game plugin as the game data parser running parallel with the rendering process.

2.3.2 Game Plugin Design and Characteristics

The development of our data acquisition plugin is based upon the Script Hook library, which provides an accessible interface to the GTA V script native functions. The released plugin is portable to the GTA game running environment and can parse the game data in real-time. We characterize the main features of our plugin as follows:

Asset Exploit. By using the native functions in GTA, we can access the states of gaming agents with our plugin. We collect the data in two means.

First, we collect human-object interactions with dynamic objects (e.g., drinking, smoking) in real-time. Such interactions are marked in the graphics engine so that the interacting objects are attached to the corresponding agent. We develop a simple detection algorithm to handle such objects. If a certain object is within a threshold distance to the character, this object is considered an interacting object. We record the locations and motions of both the character and the object and couple them into an interacting relation.

Second, static objects (e.g., walls, trees, benches) are unmarked and thus untracked in the game interface. We manually mark such objects (i.e., categories, asset ids, locations) to track human-object interactions with such static objects, e.g., sitting on a low wall, climbing over a tree. Given marked assets, we can dump such environment data from the game asset library using OPENIV GTA static parser. Like the dynamic parser, we also use a simple action detection algorithm to record human interactions with static objects.

Data Scope. Once the plugin is hooked up inside the game, it runs silently in the back end for data collection. Though our plugin can retrieve data in the area of the whole game map, we limit the data collection range to a fixed radius w.r.t the main character's position for efficiency considerations. In order to collect the different body motion styles featured in different areas in the game map, we periodically teleport the tracking character to a predefined series of locations across the map, covering common environments (e.g., streets, parks, downtown areas, outskirts) in daily activities. In this sense, we guarantee the diversity of collected data.

Data Formation. Our plugin runs in the background to fetch gaming assets in every frame. The data collection rate is empirically set to 10Hz to not interfere with the rendering process. The raw data incorporates three types of entities in the GTA environment in each frame, including human agents, objects, and vehicles. The plugin captures the real-time physical quantities such as position, orientation, velocity, acceleration, and heading for each entity. Besides, for human agents, our plugin also records skeleton data which contains 98 key points, of which 55 are skeletal joints, 21 are facial bone joints, and the rest are control nodes. We also collected the 3D geometry of each object in the form of 3D meshes, which are dumped from the OPENIV GTA static parser mentioned above.

2.3.3 Dataset Collection

We adopt two modes in the data collection process: street mode and theater mode.

Street Mode. We uniformly create 595 grid coordinates across the game map. We observe and record all humans and objects that reside in the graphics engine at each coordinate, regardless of whether it is rendered on the screen. The humans include pedestrians, drivers, business people, construction workers, gangsters, police officers, *etc*. Although the action space of the observed agents is limited to a predefined collection of activities, each person adopts a different style of body motion according to their gender, age, occupation, and physique. Therefore, we observe a wide variety of body motion sequences. Fig. 2.2 illustrates the wide variance within two action categories.

Theater Mode. In addition to the constrained set of activities collected from the street mode, we also record human and object dynamics in cutscenes. Cutscenes are CG video clips between game events that are performed by real actors and are perfected by professional artists. The dynamics in cutscenes are more diverse and realistic than those collected in street mode.

Notice that there are multiple characters at each time step, referring to both a time step in the game engine and a snapshot of a human skeleton as a frame. To avoid miscommunication, we denote each time step in the game engine as a *world-frame* and denote a snapshot of a human character as a *person-frame*.

Action Annotation. We ask volunteers to label human actions to each frame and up to one associated object for each action. For example, if a person is sitting while drinking, our volunteer would label the current frame as (sit, chair), (drink, cup) where 'chair' and 'cup' each refer to a specific object instance in the scene. It is impossible to annotate every frame of our dataset since it contains 902,478 world-frames and, on average, 32 person-frames in each world-frame. We took our best effort to annotate 609,045 person-frames in the training set and 164,628 person-frames in the testing set. We made sure that we have annotated the

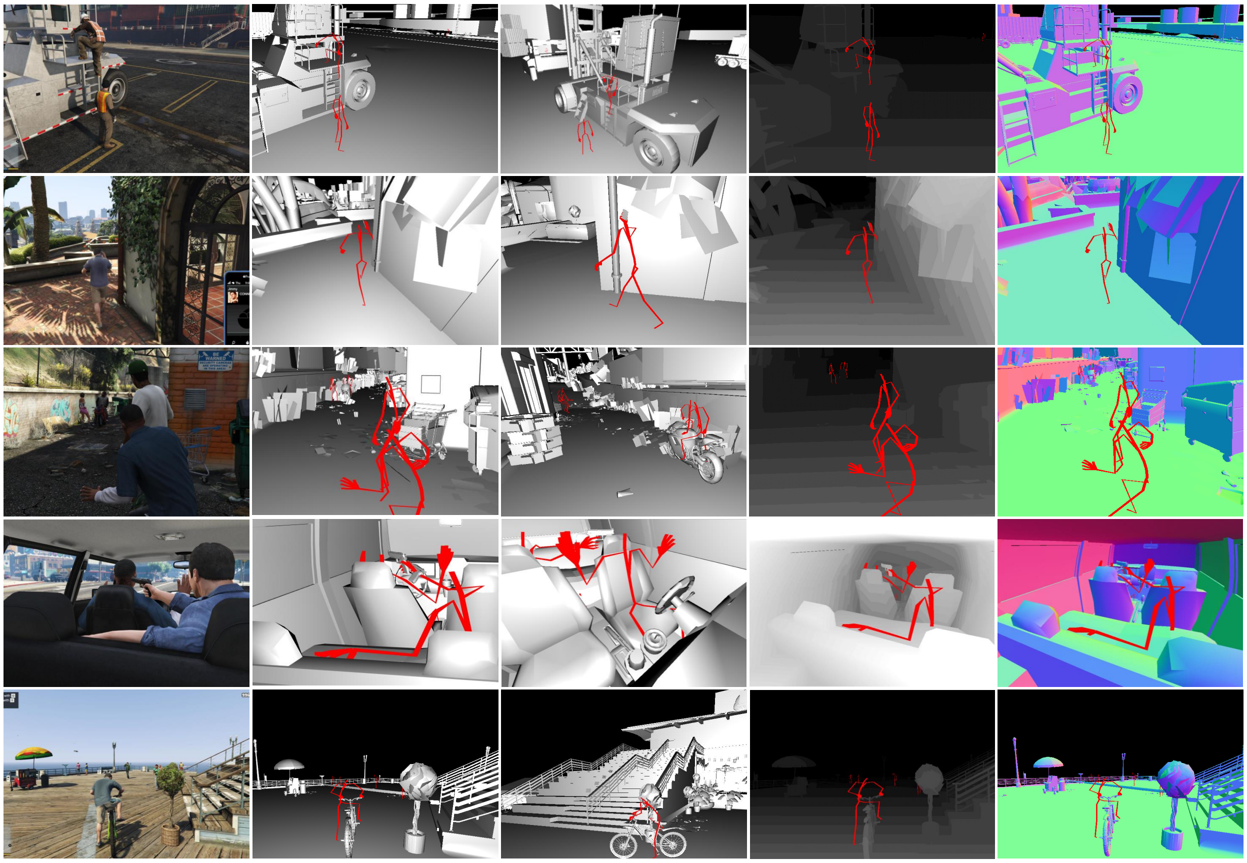


Figure 2.4: Overview of our SHADE dataset. The six columns are: RGB scene, 3D mesh model, 3D mesh model from novel viewpoint, depth map, surface normal map.

actions of every performed activity in our testing set.

2.3.4 Copyright Issues

Grand Theft Auto V allows non-commercial use of its content as long as certain conditions, such as no spoilers, are met [Roc17]. The content of this game has been used in [RVR16] for acquiring semantic segmentation annotations for self-driving cars.

2.4 Dataset Overview

In this section, we describe the design and composition of our dataset.

2.4.1 Detailed Statistics

Data Scale. We collected 902,478 world-frames and 29,164,913 person-frames, of which 772,229 person-frames are annotated. On average, each annotated person-frame contains 2.03 action labels and 0.89 interacting objects. Detailed action/interaction frequencies are reported in Fig. 2.6.

Human Action. We record the 3D positions of 55 human body keypoints for each person, including 25 major skeleton joints, 30 finger joints. Fig. 2.5 illustrates the human skeleton representation used in our dataset. In addition to skeletal joints, our dataset contains 21 key points on facial bones for expression and gaze analysis, although we do not provide annotations for expression and gaze.

Object Geometry. We represent the geometry of each object as a 3D mesh accompanied by its translation and rotation in each frame. We use the mesh representation instead of the more popular bounding box representation because it contains much richer information and can support more detailed analysis such as analyzing forces, modeling fine-grained geometric relationships, or modeling the relationship between shape and affordance. We express

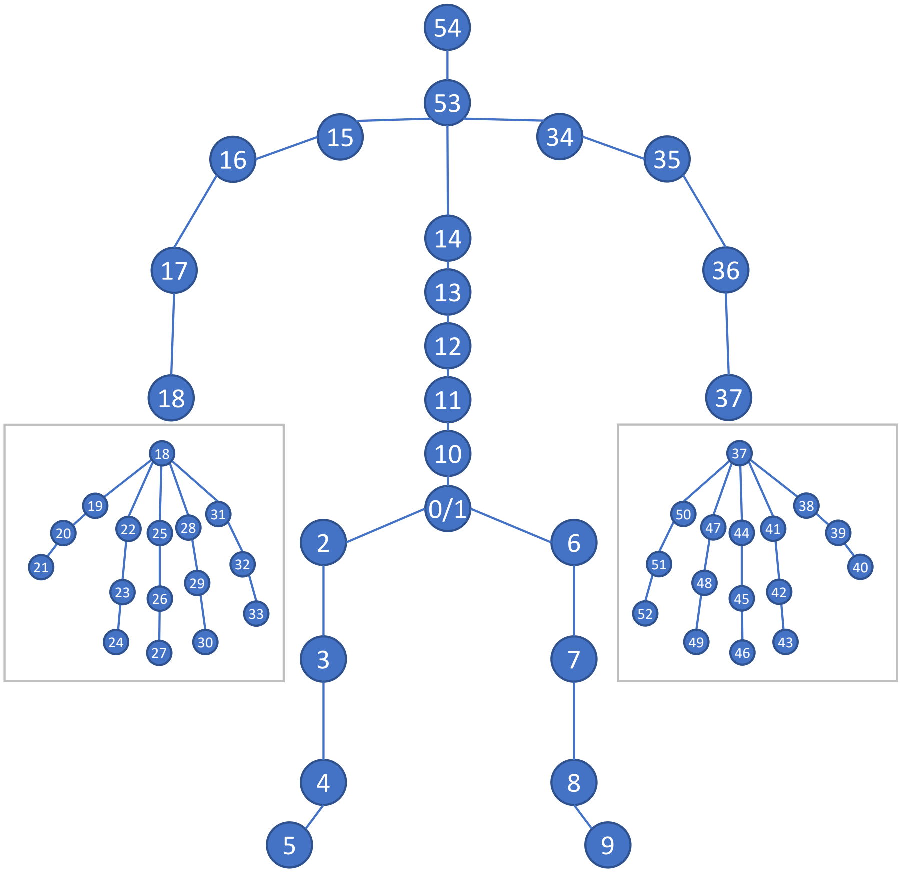


Figure 2.5: Illustration of human skeleton used in SHADE dataset.

Dataset	# Joint	# Action	Object	Sequence	Mesh GT	Bbox GT
HumanEva	16	6	No	No	No	No
Human3.6M	32	16	No	No	No	No
UCLA Multiview	20	8	No	No	No	No
MSRA DA3D	20	16	No	No	No	No
SYSU 3DHOI	20	12	No	No	No	No
SceneGrok	25	7	**Yes**	No	No	**Yes**
CAD-120	15	20	**Yes**	**Yes**	No	**Yes**
SunCG	N/A	N/A	**Yes**	No	**Yes**	**Yes**
SHADE	**55**	**161**	**Yes**	**Yes**	**Yes**	**Yes**

Table 2.1: Comparisons between our dataset and existing 3D datasets.

the rotation of an object in quaternions to avoid the singularity problem in the Euler angle expression.

Dataset Partition. We segment our dataset into training and testing set according to the way they are collected. Since the street mode produces varieties of repeating activities and the theater mode produces more diverse variations of the same set of activities, but with a smaller number of frames, it is natural to assign the street data to the training set and the theater data to the testing set.

2.4.2 Comparison with Other Datasets

Annotation Richness. Existing 3D datasets [SBB10, IPO14, HZL15, SYZ17, WZZ17] focus on either human or environment instead of both, with the only exceptions of CAD-120 [KGS13] and SceneGrok [SCH14]. Table 2.1 shows the qualitative comparison between our dataset and other 3D datasets. We show that our dataset has richer and more fine-grained annotations than other public datasets.

Pose & Action Diversity. We show that the human pose distribution in SHADE is more diverse than that in H36M by comparing the t-SNE embedding of the poses in both datasets. In order to make the two datasets comparable, we map the poses from SHADE

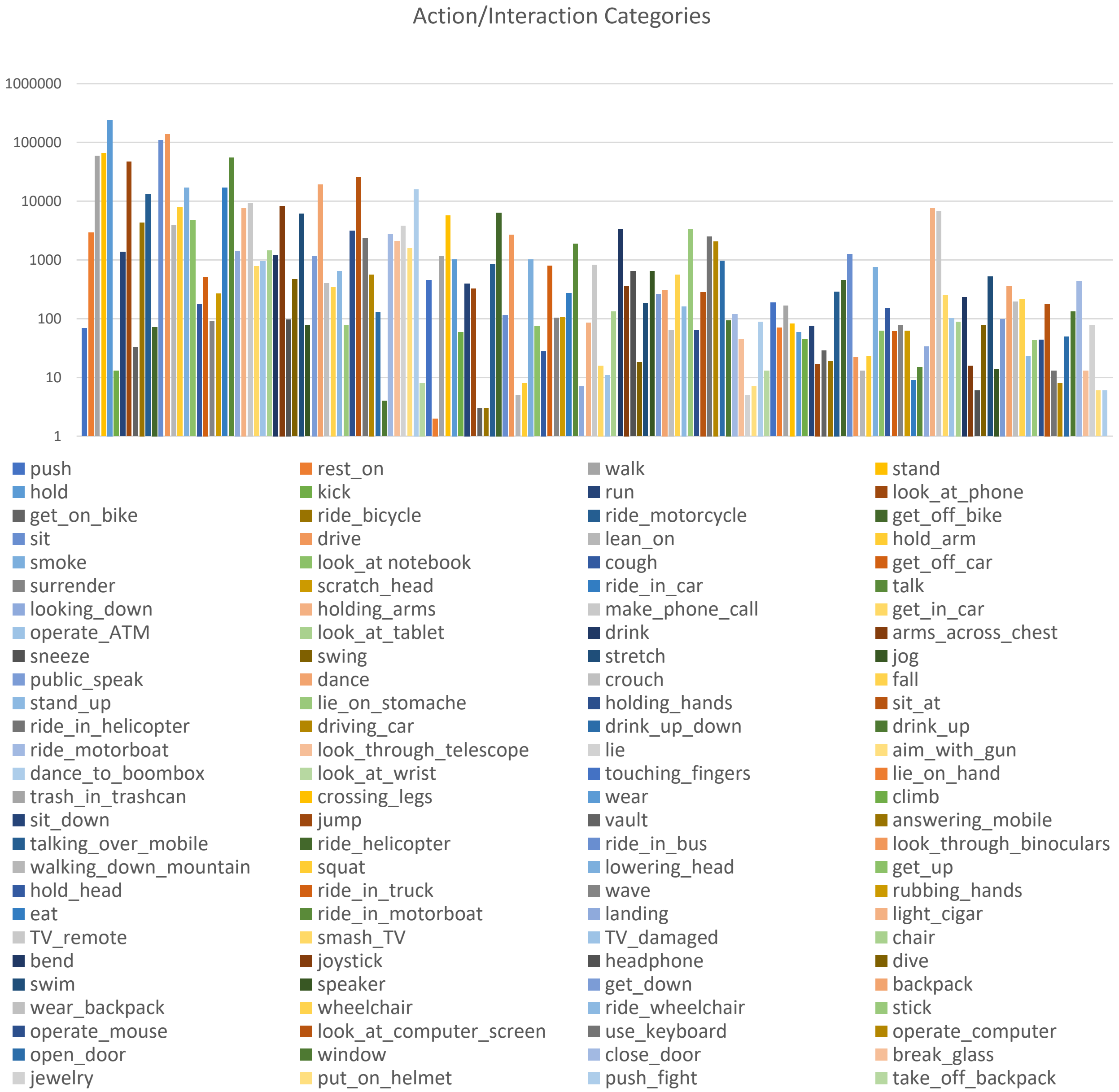


Figure 2.6: Action/interaction category frequencies in our SHADE dataset. Log scale is used for Y-axis (i.e., sample size).

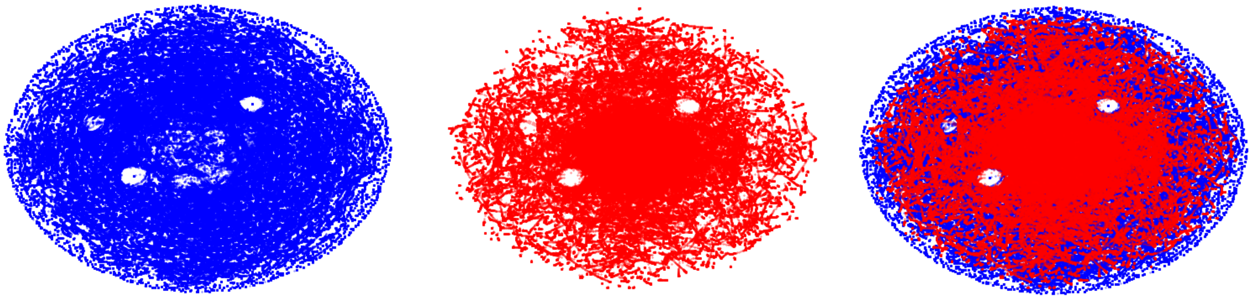


Figure 2.7: 2D t-SNE analysis of the human pose distribution. The left figure shows the human pose distribution in SHADE (in blue) and the middle figure shows the human pose distribution in H36M (in red). The right figure shows the overlap of the two.

to the same skeleton structure in H36M by removing joints on both hands. Notice that this modification reduces the number of joints by more than half. Figure 2.7 shows that SHADE has a more extensive coverage of human pose space than H36M, even with a reduced skeleton.

Pose & Action Quality. Since the human actions in SHADE are synthesized, we evaluate the quality by conducting two user studies on Amazon Mechanical Turk (AMT). In the first user study, we show each worker with 20 human motion sequences (with 3D skeleton only) and ask the worker to rate each sequence from 1 to 9 based on whether the sequence looks like natural human motion. In the second user study, we show each worker a mix of 10 sequences of eating action and 10 sequences of smoking action and ask workers to rate the sequences from 1 to 9 based on whether the sequences look like eat (1) or smoke (9). We select eating and smoking because they look similar in the H36M skeleton, and both actions exist in H36M and SHADE. We conduct both user studies with three input variants, i.e., H36M skeleton, SHADE with H36M-like reduced skeleton, and SHADE with a full skeleton. For each study, We sample 100 sequences (200 frames each) from each dataset and have around 200 participants. The average user score is reported in Figure 2.8. We can observe that human motions from the SHADE dataset are more natural and contain more information for action recognition than those in H36M and that the additional information from hand poses makes an additional contribution to motion quality.

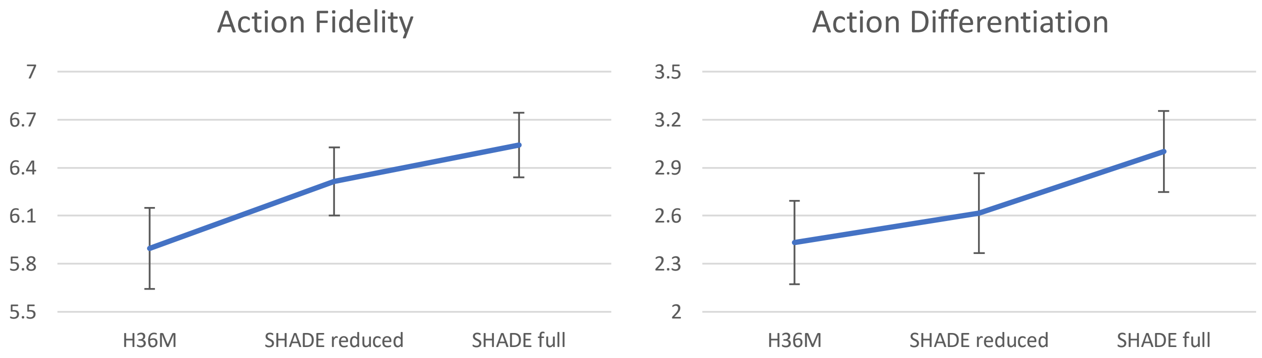


Figure 2.8: Results of user studies of human motion quality. The left figure shows the average user rating of how natural a sequence looks, where 1 means the least natural and 9 means the most natural. The right figure shows the difference in average scores between eating sequences and smoking sequences, where 1 means most like eating and 9 means most like smoking. Both values are better if larger.

2.5 Experiment

We evaluate our dataset with four tasks: HOI recognition, object localization, human pose estimation, and scene reconstruction.

2.5.1 HOI Recognition

We run state-of-the-art skeleton-based action recognition models [WW17, YXL18] on our data, and augmented the better one with an additional coarse contextual feature, richness-of-object, around each joint. Table 2.2 shows that this simple feature has already provided a significant edge for the state-of-the-art action recognition model.

Richness-of-object. We first uniformly sample point clouds on the surfaces of all objects. Then we compute the number of points within a fixed radius around each joint. We then divide the number by 1000 and clip the result to be between 0 and 1. We append the resulting number to each joint in the human skeleton to reflect the richness of contextual objects around each joint.

	mAP	Top-1 Acc.	Top-3 Acc.
ST-GCN [YXL18]	0.54	0.35	0.59
2stream [WW17]	0.76	0.61	**0.94**
ST-GCN + SHADE	0.61	0.42	0.75
2stream + SHADE	**0.84**	**0.78**	**0.94**

Table 2.2: Quantitative results and comparisons of the accuracy on skeleton-based action recognition. Higher values are better. The best score is marked in **bold**.

IoU	Smoke	Eat	Drink	Sit	Sit at
KNN	0.08	0.02	0.10	0.37	0.14
DNN-single	0.11	0.07	0.13	0.42	0.14
DNN-joint	**0.14**	**0.15**	**0.20**	**0.50**	**0.16**

Table 2.3: Quantitative results and comparisons of the accuracy (IoU) on interacting object localization. Higher values are better. The best score is marked in **bold**.

Protocol #1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
[TMS17]	85.0	108.7	84.3	98.9	119.3	95.6	98.4	93.7	73.7	170.4	85.0	116.9	113.7	62.0	94.8	100.0
[PZD17]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
[ZHS17]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
[MHR17]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
[FXW18]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
[MHR17] + SHADE	49.7	56.6	57.1	58.0	67.2	77.4	54.7	57.8	81.1	91.5	61.0	58.5	65.8	49.47	53.2	62.6
[FXW18] + SHADE	49.3	54.9	56.6	57.1	65.8	75.4	53.5	56.0	73.0	88.8	60.6	57.1	61.9	45.8	48.7	**60.3**
Protocol #3	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
[PZD17]	79.2	85.2	78.3	89.9	86.3	87.9	75.8	81.8	106.4	137.6	86.2	92.3	72.9	82.3	77.5	88.6
[ZHS17]	61.4	70.7	62.2	76.9	71.0	81.2	67.3	71.6	96.7	126.1	68.1	76.7	63.3	72.1	68.9	75.6
[MHR17]	65.7	68.8	92.6	79.9	84.5	100.4	72.3	88.2	109.5	130.8	76.9	81.4	85.5	69.1	68.2	84.9
[FXW18]	57.5	57.8	81.6	68.8	75.1	85.8	61.6	70.4	95.8	106.9	68.5	70.4	73.8	58.5	59.6	72.8
[MHR17] + SHADE	64.8	64.1	83.8	78.2	80.2	100.5	67.6	84.2	113.9	129.1	73.5	78.0	85.9	67.8	67.2	82.6
[FXW18] + SHADE	59.6	61.0	73.5	68.0	67.6	81.3	62.5	67.4	87.0	100.4	64.2	71.5	78.0	61.9	61.5	**71.0**

Table 2.4: Quantitative comparisons of Average Euclidean Distance (in mm) between the estimated pose and the ground-truth on *Human3.6M* under *Protocol #1* and *Protocol #3*. Lower values are better. The best score is marked in **bold**.

2.5.2 Object Localization

We establish three baselines on five common activities in our dataset for the reference of future research. The results are listed in Table 2.3.

Referred object. For the first four activities, the referred objects are cigarettes, food, drinks, and chairs, respectively. For the last activity, the referred object is the table in front of the sitting person if there exists one.

Baseline methods. We develop three simple baselines for the task of 3d object localization: i) KNN. We normalize each joint coordinate to zero-mean and unit variance and find the nearest neighbor given the query skeleton in training data. We return the associated bounding box of the nearest neighbor as our prediction result; ii) DNN-single. We design a neural network based on the structure proposed in [MHR17]. We consider the object 3D bounding box boundaries as keypoints and learn to regress their coordinates end-to-end; iii) DNN-joint. We further consider human poses and extend the architecture from ii) to jointly optimize the locations of human skeleton joints and interacting object bounding boxes. We evaluate the baseline models on intersection over union (IOU). Notice that the first two activities suffer from extremely low IOU since the referred objects are usually much smaller than other objects, and therefore it is harder for the predicted bounding boxes to intersect with the ground truth ones.

Generalizing to real humans. To show that the geometric relationship learned in our dataset can be generalized to real-world cases, we evaluate the KNN method on a pose chosen from Human3.6M [IPO14] and show four synthesized objects for eat, sit and sit_at in Fig. 2.10. The selected pose is sitting on a chair and is acting as if she is eating.

2.5.3 Human Pose Estimation

We demonstrate the diversity of our collected human pose in this subsection by training a state-of-the-art 3D pose estimation model [MHR17] on a combination of our dataset and

H36M and compare it with the same model trained solely on the H36M dataset. Table 2.4 shows that our dataset helps improve performance on state-of-the-arts in action recognition. We further test the two trained models on less common human poses in our testing set, *i.e.*, Yoga poses. Fig. 2.9 qualitatively illustrate that our dataset allows better generalization of the state-of-the-art model than H36M does. We make sure that no Yoga poses or any similar poses to the testing poses are present in the training data.

2.6 Conclusion

This chapter presents a large-scale synthetic dataset SHADE (Synthetic Human Activities with Dynamic Environment). Our dataset is the first that contains rich and fine-grained 3D annotations of human-object interactions. Our experiments show that the human pose in our dataset complements existing human pose datasets and that the geometrical relationship in our dataset can be applied to real-life human behaviors. We believe that this dataset would open up new possibilities in modeling 3D human-object interactions.

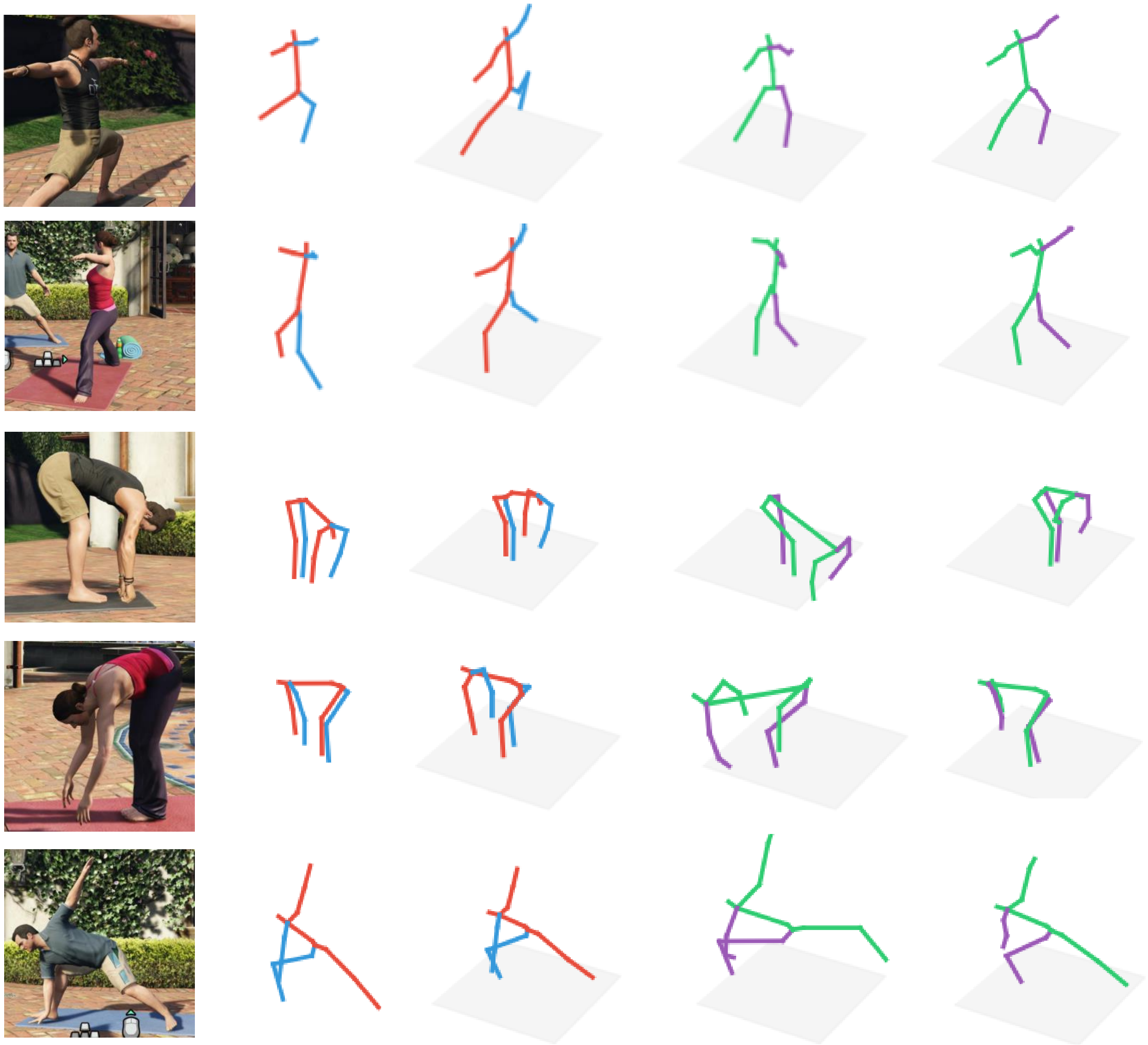


Figure 2.9: Qualitative results on 3D pose estimation. The first column are querying 2D poses, the second column are ground truth 3D poses, the third column are 3D poses predicted by model trained on H36M, and the fourth column are 3D poses predicted by model trained on both H36M and SHADE.

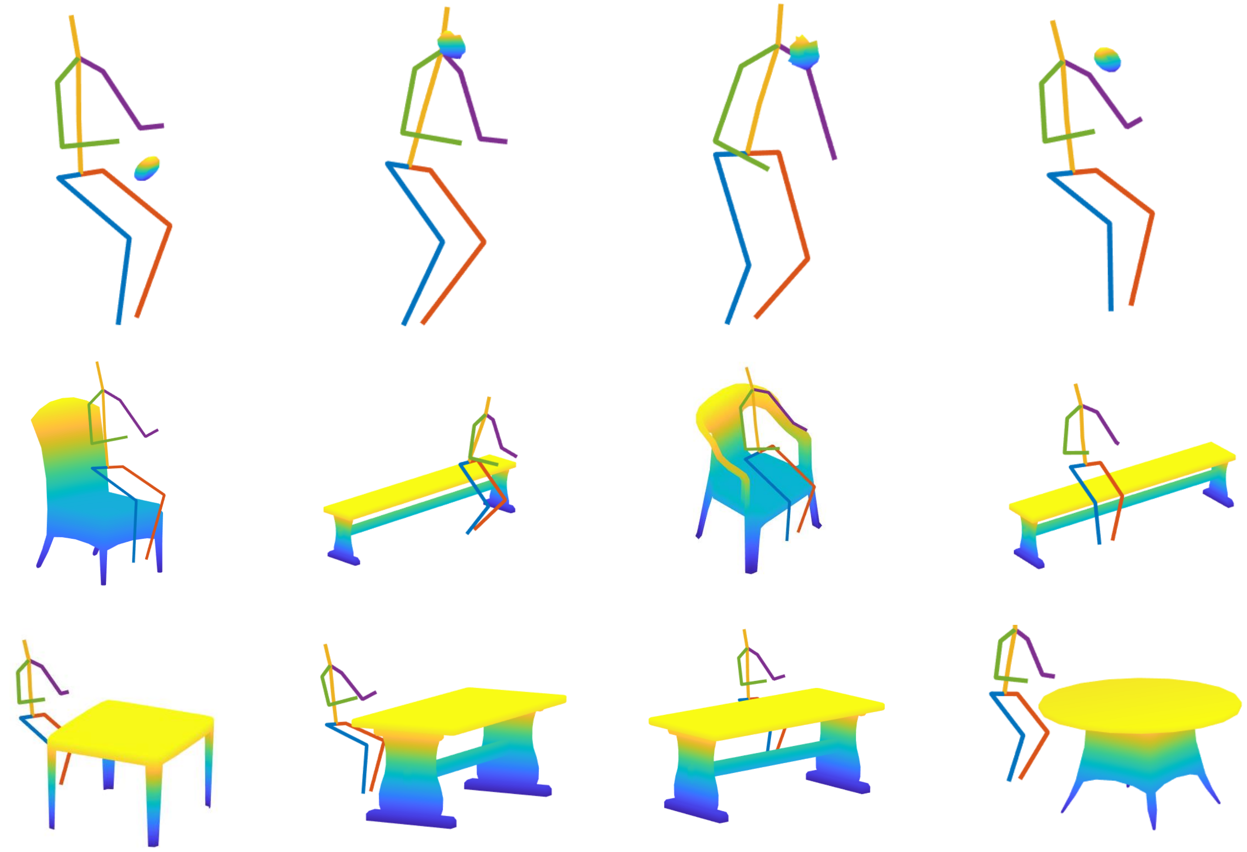


Figure 2.10: Qualitative results of model generalization to Human3.6M. We apply the learned human-object interaction model from the SHADE dataset and sample the possible interactions given poses from Human3.6M datasets. The first row synthesizes a bagel or a sandwich for eating, the second row synthesizes a chair or a bench for sitting, and the third row synthesizes a table or a desk for sitting-at. The four columns show four different samples.

CHAPTER 3

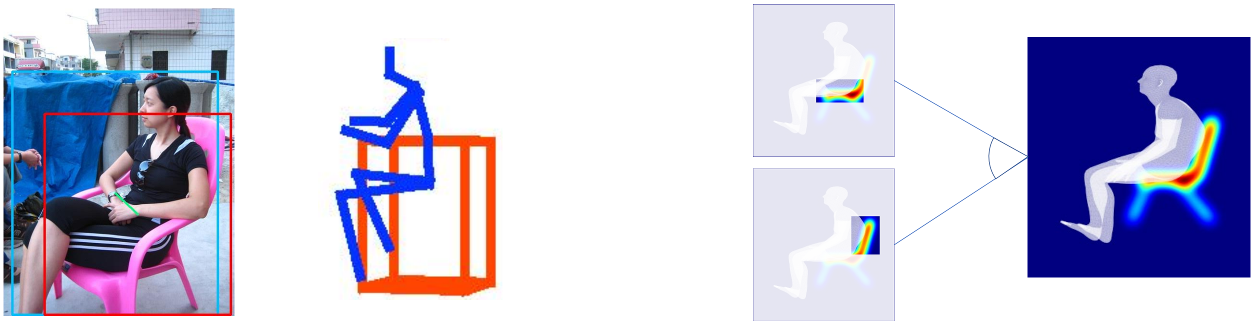
Hallucinating Sitting Human by Exploiting Part-Level Human-Object Interaction

3.1 Introduction

The previous chapter shows that understanding HOI does improve scene understanding. However, there is one assumption that is problematic when we look at it closely. Namely, we assumed that a single bounding box could represent each object, and its position follows a multivariate Gaussian distribution anchored from a specific joint of the human skeleton. This assumption seems to be working effectively and has been widely adopted in multiple publications involving HOI [HQZ18, WZZ13, WZZ17]. However, all the mentioned works evaluated their models' performance in a minimal environment, where the chairs are predominately office chairs and have a consistent shape distribution.

By representing each chair as a single bounding box, we assumed the seats and backs of a chair are always around certain positions of the chair's bounding box and that the geometric relationship between the human hip and the chair's geometrical center is consistent across the dataset. This assumption is most likely valid when we only consider one specific type of chair, but it would break as soon as we consider chairs with different shapes, such as sofas and stools. We show this dilemma in Fig. 3.1a, where the instance-level HOI does not accurately show the relationship between the human and the object.

In this chapter, we argue that instead of defining HOI between the object-level bounding box and a key joint, we should build HOI between lower-level parts of the object and body



(a)

(b)

Figure 3.1: Instance-level HOI vs. Part-level HOI.

parts, as shown in Fig. 3.1b. We propose a novel algorithm of hallucinating sitting humans given a decomposed chair. By leveraging part-level HOI relationships, our method is capable of achieving

- simpler geometric relationship in each HOI node,
- more diverse hallucinated human pose, and
- more physically plausible hallucinated human pose.

3.2 Related Works

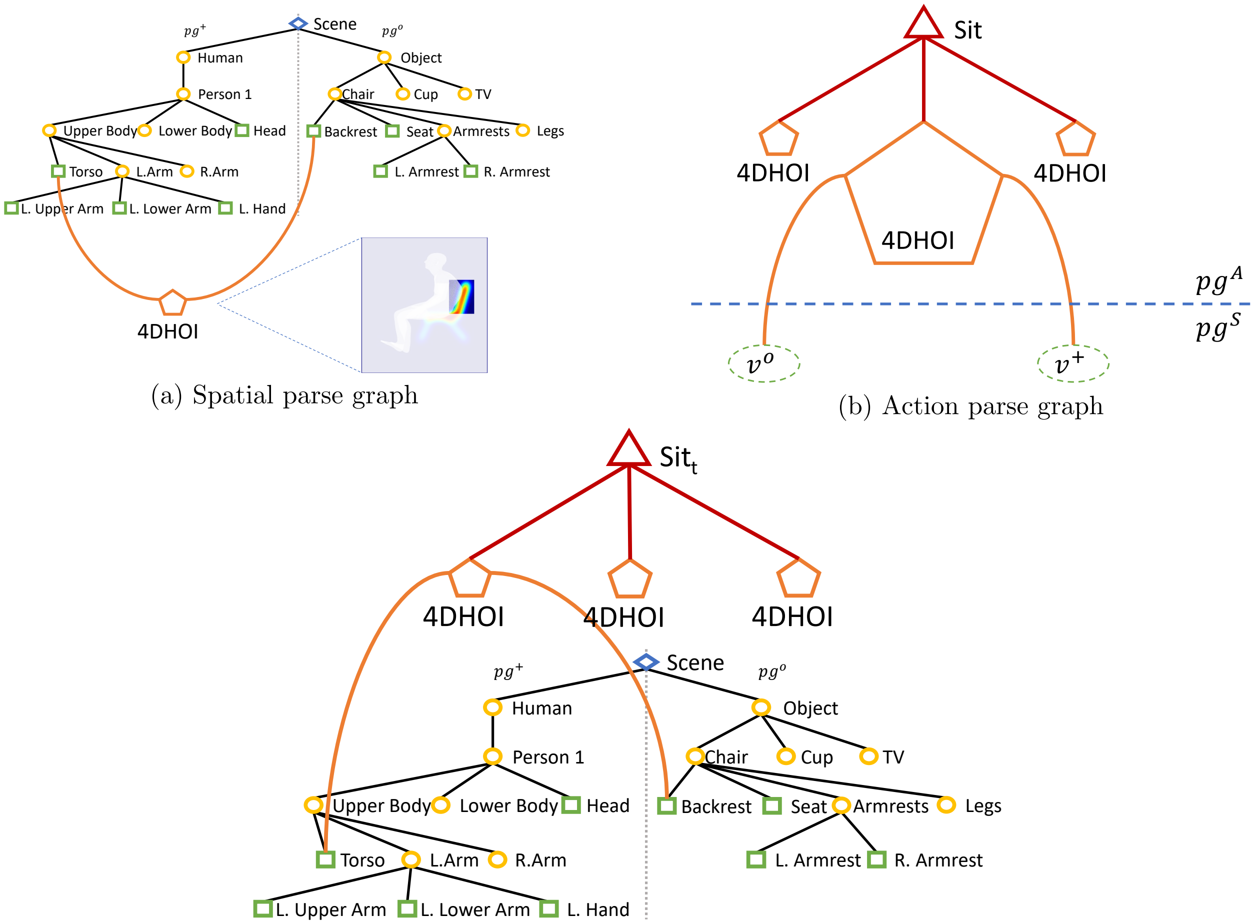
Human Object Interaction (HOI) has been widely studied as both a 2D detection problem and a 3D reconstruction problem. 2D HOI detection often involves identifying an interacting pair of human and object and classifying the verb of interaction given an interacting human and object image. Most of the current HOI detection algorithms [YF10, LZH19, WYD20, KLG18] project both the human and object patch into image feature space, and classify the interaction verb using both features. Some recent approaches use graph neural networks to predict human-object pairs that contain HOI relationship [QWJ18]. In addition to detecting instance-level HOI, several recent works [LXL20, LSL18] demonstrate improved generalization capability by identifying the interaction between body parts and objects.

3D HOI is often used as the prior knowledge of arranging estimated 3D human and objects in scene reconstruction tasks [WZZ13, WZZ17, HQZ18]. While these methods all improve existing 3D scene reconstruction methods by a significant margin, they all rely on consistent shape distribution for each object type. If, for example, the object category chair includes both high stools and office chairs, they would have very similar bounding boxes but significantly different positions for a person to sit on. This chapter aims to solve this issue by using a set of part-level oriented bounding boxes to represent an object accurately and efficiently.

Hallucinating Human is another trending research topic that utilizes 3D HOI to improve its hallucination quality [HGT21, ZHN20, ZZM20, HCT19]. Publications along this stream use the exact 3D shape, usually represented by triangular meshes, to regulate the placement of hallucinated human bodies. The use of mesh representation includes high-frequency details of the interacting objects. However, the overwhelming amount of detail limits the proposed models from learning significant features between humans and objects. As a result, the hallucinated human had to re-use poses from demonstration data and then be placed in the given scene to match some pre-defined features. Our proposed method can learn pairwise part-level interaction features that can produce highly diverse interacting human poses.

3.3 HOI Representation

Existing algorithms often represent objects as 2D and 3D bounding boxes. Although bounding boxes are excellent in representation efficiency, it does eliminate important details for complex shapes such as a chair. Another extreme of the spectrum is to use expensive representations such as voxels, point clouds, or meshes to include the high-frequency details of



(a) Spatial parse graph

(b) Action parse graph

(c) Full view

Figure 3.2: An illustration of our HOI representation as a bridge between spatial parse graph and action parse graph.

the object. We observe that most manufactured objects, especially furniture, can be decomposed into semantic parts, and each part can be approximated by a shape primitive. In this chapter, we use oriented bounding boxes to represent each part.

Since we can decompose both object and human into semantic parts, we propose to use a hierarchical structure to represent actions that involve human-object interactions. In this chapter, we use the action sitting as an example to illustrate our idea.

Consider a scene where a person sits in an office chair, with both arms on the armrest. The static scene includes a human H and an object O, which is the chair. Both H and O can be organized into a parse graph. We use pg^O to represent the object and pg^H to represent the human. $pg^O = \langle V^O, E^O \rangle$ is a graph where $V^O = \{v_i^o, i = 1, ..., n\}$ is the set of object part nodes that each describes a component of the object. $E^O = \{(v_1, v_2), v_1, v_2 \in V^O\}$ is the set of edges in pg^O where v_2 is a sub-part of v_1. The root node represents the entire object, and every level down in pg^O represents a more fine-grained decomposition of the object. pg^H is defined similarly following the hierarchical decomposition of a human skeleton. We represent action A as a hierarchical structure pg^A where the action is decomposed into several HOI nodes. Each HOI node represents a geometrical relationship between an object part and a human body part. We illustrate the representation in Fig. 3.2.

3.4 Formulation

We formulate the human hallucination problem as sampling from the conditional distribution $p(H|O; \Theta)$, where Θ is the set of learnable parameters. We model the distribution as a Gibbs distribution

$$p_A(H|O; \Theta) = \frac{1}{Z(\Theta)} \exp -\mathcal{E}_A(H|O; \Theta) p_0(H) \tag{3.1}$$

, where we can learn the parameters Θ by maximizing the log-likelihood

$$L_p(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \log p(H_i|O_i; \Theta) \tag{3.2}$$

. $Z(\Theta) = \int \exp -\mathcal{E}_A(H|O; \Theta) dH$ is the normalizing constant. To maximize L_p, we use standard gradient-ascent algorithm and compute the gradient

$$L'_p(\Theta) = \mathbb{E}_\Theta \left[\frac{\partial}{\partial \Theta} \mathcal{E}_A(H|O;\Theta) \right] - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta} \mathcal{E}(H_i|O_i;\Theta) \tag{3.3}$$

. The $\mathbb{E}_\Theta$ term is the expectation of energy with respect to $p(H|O;\Theta)$, which is intractable. We use Langevin sampling to sample $\{\tilde{H}_i\}$ from $p(H|O;\Theta)$ with given $\{\tilde{O}_i\}$, and use the sampled results to estimate $\mathbb{E}_\Theta$. The gradient then becomes

$$L'_p(\Theta) = \mathbb{E}_\Theta \left[\frac{\partial}{\partial \Theta} \mathcal{E}_A(H|O;\Theta) \right] - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta} \mathcal{E}_A(H_i|O_i;\Theta) \tag{3.4}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta} \mathcal{E}_A(\tilde{H}_i|\tilde{O}_i;\Theta) - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta} \mathcal{E}_A(H_i|O_i;\Theta) \tag{3.5}$$

In light of Eq. (3.5), we implement Algorithm 1 for hallucinating interacting human given an object: h

```
Algorithm 1: Hallucinating Interacting Human
   Input: HOI datasets $\{(H_i, O_i), i = 1, ..., N\}$, query objects $\{\tilde{O}_i, i = 1, ..., N\}$, initial
          parameters $\Theta^0$, total iterations $L$, langevin steps $l$, step size $\eta$, random noise
          size $\zeta$
   Output: Hallucinated human $\tilde{H}_i$, final parameters $\Theta^*$
 1 Initialize $\Theta^0$
 2 for $iter = 1 : L$ do
 3     for $step = 1 : l,\ i = 1 : N$ do
 4         $\tilde{H}_i \leftarrow \tilde{H}_i - \frac{\partial}{\partial H_i} \mathcal{E}_A(\tilde{H}_i|\tilde{O}_i;\Theta) + \epsilon, \epsilon \sim \mathcal{N}(0, \zeta)$;
 5     end
 6     $\Theta^{(iter)} = \Theta^{(iter-1)} + \eta \cdot L'_p(\Theta)$;
 7 end
 8 $\Theta^* \leftarrow \Theta^L$
```

Algorithm 1 can be interpreted as a minimax game, where Line 4 updates the synthesized examples $\tilde{H}$ to minimize its energy, and Line 6 updates the parameters Θ to maximize the energy of synthesized examples and minimize the energy of observed ones. It can further be interpreted as two competing forces attacking each other's weaknesses. In one turn, Line 4 tries to exploit the weakness of the energy function and fool the energy function into thinking the synthesized examples are real. In the next turn, the energy function tries to identify the weaknesses in the synthesized examples in order to tell the synthesized ones apart from the observed ones in Line 6.

3.5 Human Object Interaction Energy

We use a neural network $\mathcal{F}(x;\Theta) : \mathbb{X} \mapsto \mathbb{R}$ to model the energy function $\mathcal{E}_A$, where $x(H,O) \in \mathbb{X}$ is the collection of all part-level HOI features between human H and object O. For each object part o and human joint j, we compute two scalar part-level HOI features between o and j as

1. distance from j to o's bounding box center, and
2. distance from j to o's closest bounding box surface.

For each object part o and human bone k, we compute six scalar part-level HOI features between o and k as

1-3. angle between bone and bounding box axes, and

4-6. angle between bone and world axes.

In total, we have 24 joints and 23 bones in a human skeleton, and therefore 186-vector for each object part. We then aggregate the features with max-pooling by part name and concatenate the features of different part names into a long vector before feeding the features to the energy network $\mathcal{F}$.

Method	Physical Plausibility ↓	Human Pose Diversity ↑
cVAE (Instance-Level)	1.82×10^{-2}	0.032
cVAE (Part-Level)	$\mathbf{9.32 \times 10^{-3}}$	0.057
Ours (Instance-Level)	3.16×10^{-2}	0.081
Ours (Part-Level)	2.14×10^{-2}	**0.110**

Table 3.1: Comparisons between our dataset and existing 3D datasets. Physical plausibility is measured by the penetration between a human skeleton and object part bounding boxes, and human pose diversity is measured by the standard deviation of human joint positions.

3.6 Experiment

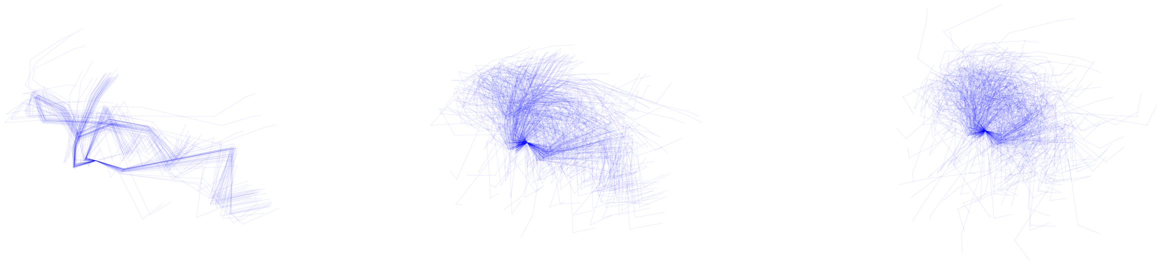
We collect 108 instances of sitting pose from the SHADE dataset described in the previous chapter as our training data. We evaluate our model with two quantitative metrics: physical plausibility and human pose diversity.

- **Physical Plausibility.** We compute the total penetration between the human pose skeleton and the oriented bounding boxes of the interacting object. This metric is better when lower.

- **Human Pose Diversity.** We compute the average standard deviation of per-joint rotation angles for the hallucinated human pose. This metric is better when larger.

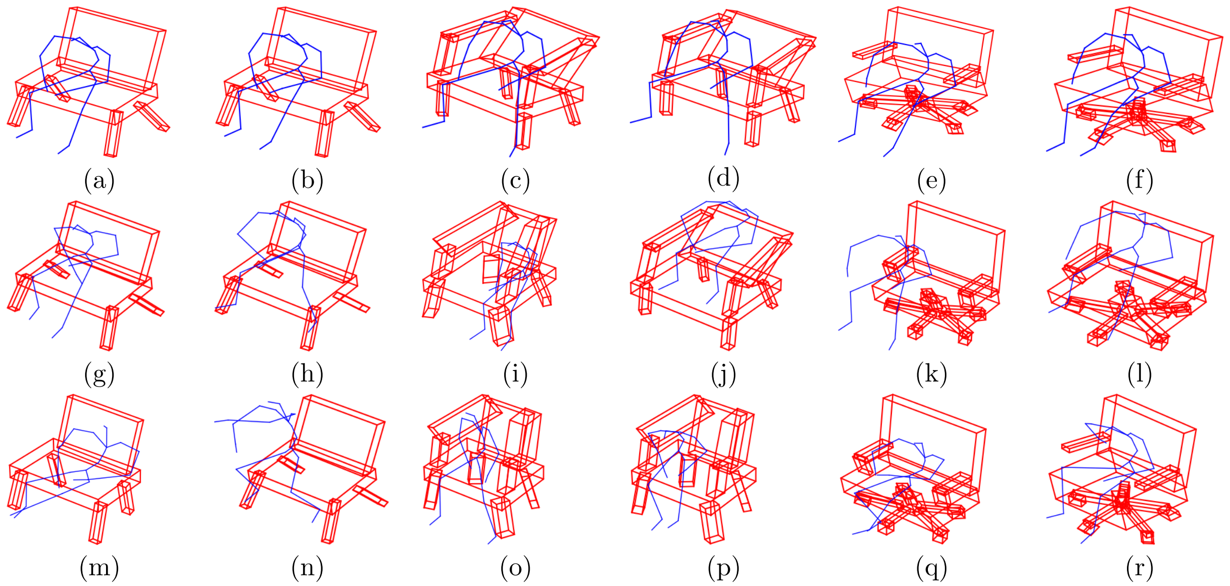
We compare our model with a baseline method of naive conditional VAE. We also conducted an ablation study where our model is used with the entire object as a single part, forcing the model to use instance-level HOI. We show in Table 3.1 that our model performs significantly better in both metrics, and in Fig. 3.3 to qualitatively show how our method improves synthesis diversity by a large margin at a slight cost in physical plausibility. We also show that by including part-level HOI, we can reduce physics violations.

We also show a qualitative result of the synthesized sitting poses in Fig. 3.4 to demonstrate that our method is capable of hallucinating diverse poses that are not included in the training set. We further show the 3D views of sitting poses synthesized by our full model in Fig. 3.5.



(a) cVAE (b) Ours (instance-level HOI) (c) Ours (part-level HOI)

Figure 3.3: **Synthesized pose distribution of different methods.** We show that our method is capable of synthesizing much more diverse human pose than the baseline method, and that part-level HOI provides even stronger diversity.



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

Figure 3.4: Synthesized sitting human pose conditioned on chairs. We show examples synthesized using (top) cVAE with object parts as condition, (middle) our method with instance-level HOI, and (bottom) our method with part-level HOI. We show that our method can create significantly more diverse sitting poses, and that incorporating part-level information creates more realistic examples.

3.7 Conclusion

This chapter proposed representing human-object interaction as a combination of part-level HOIs, where an oriented bounding box represents each object part. We introduced a novel algorithm of hallucinating interacting humans to prove the expressiveness of our representation. We show that we can express a richer set of HOIs using naive geometric relationships by decomposing the HOI relationships into lower levels.

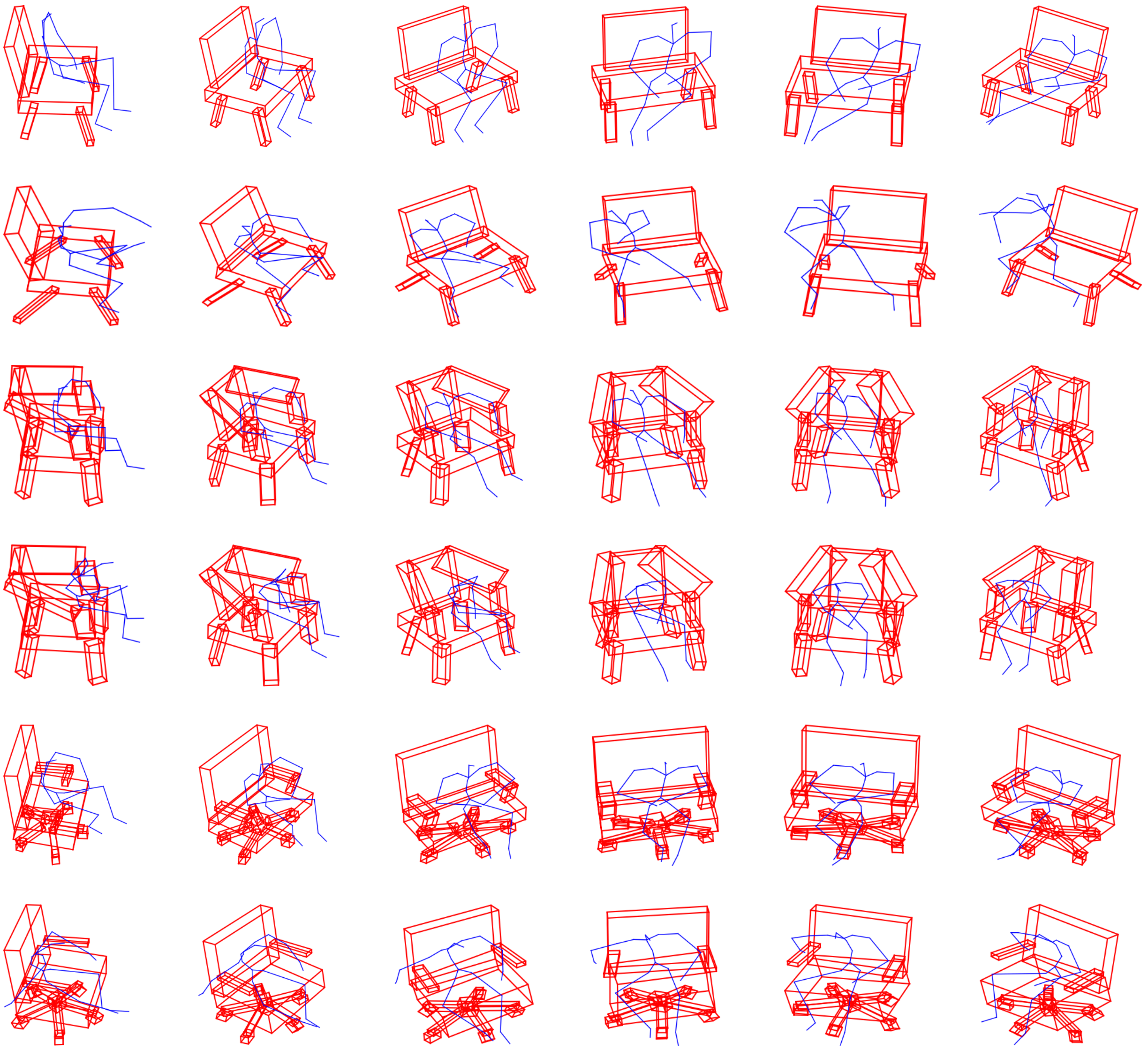


Figure 3.5: 3D view of sitting poses synthesized by our method. Each row contains 6 views of the same synthesized example. We observe that our method is capable of synthesizing diverse and realistic examples.

CHAPTER 4

InteractionField: Learning Intuitive Grasping from Human Demonstrations via Conditional Descriptor Networks

In addition to complex body-scale human-object interactions such as sitting, this dissertation also investigates hand-object interaction, which is an equally complex human-object interaction on a smaller scale. This chapter proposes to learn a geometry-aware grasping energy function that describes the shape distribution of humanoid grasping of daily objects.

4.1 Introduction

In recent years, the robotics community views grasping as a physics problem of finding a hand configuration that ensures force closure on an object. Although this is a widely accepted stream in the robotics community, it is different from the internal dynamics of human grasping. When a person performs grasping of an object, two stages are involved. First, the person positions his hand next to the object and places his fingers around it in anticipation of a grasp. Then, the person contracts the muscle on his fingers to hold the object steadily. The first stage is commonly referred to as the preshape stage, and the second is named the holding stage. A recent study found that the electroencephalographic (EEG) activity, which reflects the brain activity, in the preshape stage corresponds to the shape and size of the object, while the EEG activity in the holding stage reflects muscle activity [SM18]. This finding suggests that when the hand configuration is determined, the human brain is more likely to focus on geometries than physics. We refer to this phenomenon as "intuitive grasping" instead of the "physical grasping" that the majority of the robotics community is focusing on. Specifically, intuitive grasping refers to synthesizing a grasping snapshot (or a sequence) of a given object without reasoning on physics. We argue that we are converting a challenging optimization problem under complex physical constraints to a much simpler problem of learning geometric relationships by studying intuitive grasping.

In this chapter, we model intuitive grasping as a conditional distribution of the human hand configuration given an object shape. We propose to learn the conditional distribution with a conditional descriptor network, a conditional energy-based model that aims to describe the energy landscape of its input. We can sample from the learned energy landscape with Markov chain Monte Carlo (MCMC) methods such as Langevin dynamics. We develop a 3D shape-aware energy function for the proposed model, which can extract 3D hand-object interaction from hand and object configuration variables. The energy function consists of two parts: (1) some pre-trained interaction-field modules (IF modules) that facilitates differentiable mappings from geometric variables to 3D voxel-based representations of hand and object, and (2) a trainable 3D bottom-up ConvNet structure that takes the channel-wise concatenation of the voxel-based representations of hand and object as input and outputs the negative energy. We designed a platform to collect a dataset of human demonstrations of grasping for training our model. We demonstrate that the proposed model can generate meaningful grasping when a seen or unseen object is given. and it outperforms the baseline methods we develop based on GAN [GPM14, MO14] and VAE [KW13].

Our contribution is four-fold: i) Despite a naming collision with an earlier work [US00], we are the first to learn the intuitive grasping that is decoupled from physics; ii) We propose the conditional descriptor network, which is a deep conditional energy-based model, for modeling the geometric relationship in grasping; iii) We propose a 3D shape aware energy function for the proposed model; iv) We design a platform to collect a dataset of human demonstrations of grasping to train our model.

We will discuss our similarities and differences with related works in section 4.2. We will then introduce our framework for learning intuitive grasping from human demonstrations in section 4.3. Next, we will introduce the 3D shape aware energy function for the proposed model in section 4.4. In section 4.5, we will introduce how to collect a human demonstration dataset and evaluate our framework by conducting experiments. A conclusion is made in section 4.6.

4.2 Related Works

Hand-Object Grasping. Current research in grasping can be organized into three streams, analytic approach, data-driven approach, ([BMA13]) and contact-based approach. The analytic approach [RMF12, PMG12, RSG12, Mur17] attempts to generate grasps by ensuring force closure of rigid objects, assuming simplified contact and friction models. Although this stream of work established a good foundation in analyzing stable grasp quality, it largely relies on the precise knowledge of the 3D shape being grasped. Their performance drops significantly if the shape is obtained via estimation. The data-driven approach leverages recent advancements in machine learning and attempts to estimate grasping points from input images. Although many works have shown progress along this stream [SDN08, CA09, RKK09, LLS15, MLN17, LPK18], this approach relies on huge datasets to learn successful grasping. In addition, the mapping from grasping points to a grasping hand is non-trivial. Most work along the data-driven stream focuses on grippers with limited DOF. The contact-based approach [AKH12, LS15, BHH19] extends data-driven approach to more complex hand models. Given an object and a contact map, contact-based algorithms optimize hand poses to fit the contact map. This approach assumes that the contact point of at least one finger must be given in addition to the contact map. This chapter focuses on modeling intuitive grasping with probabilistic models and learning the distribution of hand pose given the object to be grasped, which is a fundamental problem in cognitive robotics.

Human Hand Grasp Datasets. Collecting a high-quality human grasping dataset has been challenging due to the musculoskeletal complexity of the human hand and occlusion between the interacting parties. Several attempts have been made to record the image and 6D poses of hands and objects. [MA04] generated grasping data by sampling random grasps and accepting valid ones according to a force closure criteria introduced in [FC92]. Although this approach reduces manual labor in labeling grasp poses, the collected grasping gestures are not necessarily natural human grasps. Some datasets ([KRK11, BFD15, SXL16]) collect 2D images from human demonstrations. Due to high redundancy in visual information, estimating human poses from images/videos may be expensive and inaccurate. [HAW07, LS14] introduced data gloves to collect the exact hand poses in 3D. However, aligning an object with the hand still requires additional effort. [GYB18] can record the 6D poses of hand and object with seven magnetic 6D pose sensors. However, recording the exact shapes of hand and object during a hand-object interaction remains a challenging task.

Descriptive Models. Our model is related to a stream of publications in the field of descriptive models. [ZWM98] proposed the FRAME (Filters, Random Fields and Maximum Entropy) model for modeling and synthesizing textures. The resulting model is energy-based in the form of Gibbs distribution. [XHZ15, XLZ16a] proposed the sparse FRAME models for object patterns. The above two models use linear filters to capture local image features. [LZW15] extends the FRAME model by using pre-trained ConvNet structure as non-linear filters. Instead of using filters from a pre-trained ConvNet, [XLZ16b] learns a deep convolutional energy-based model from the observed data. [XZN17, XZG18] further explore the possibilities of learning such deep energy-based models for representing videos and 3D shapes using voxels, respectively. Previous works show that the most difficult part in training a descriptive network is the sampling process from high dimensional spaces such as the image space. [XZG18, LZW15] use a warm-start technique to overcome the problem, where a very long sampling chain (probably more than 1000 Langevin steps) is required before good examples can be synthesized. [GLZ18] first samples from a low-dimensional space (low-resolution image space) and gradually move up in resolution. [XLG18] learns a separate network for sampling. This chapter proposes a conditional version of a descriptive model with a novel 3D shape aware energy function that extracts 3D hand-object geometric interaction from their configuration variables.

Deep Learning with 3D Shapes. Currently, most deep learning algorithms involving 3D shapes adopt one of three representations: occupancy grid (voxel), point cloud, and mesh. **voxel** is the most commonly used 3D representation for analysis on 3D shapes [XZG18, ROG17, CAL16] due to its innate similarity to pixels and compatibility to 3D ConvNets. However, the voxel representation suffers from many drawbacks. In addition to the commonly criticized problem of cubic complexity with respect to resolution, [LPS16] shown that the voxel representation also suffers from increasing sparsity with increasing resolution. Recently [ROG17] aims to solve both problems so that voxel of higher resolution can fit in the deep learning framework. The voxel representation also lacks surface information such as surface normal. **Point cloud** is a set of points sampled (often uniformly) from object surfaces. Although point cloud is a more compact representation than voxels, it does not fit well in most machine learning frameworks as there can be $N!$ permutation of the same set of points. Point clouds also do not host volumetric information and surface normal information. Although recent works [QSM17, QYS17] have demonstrated possibilities in working on point cloud data directly in a deep learning framework, it is still an open problem to process point cloud for generic tasks. **mesh** is more commonly used in computer graphics than in deep learning due to its compactness but is less favored in the machine learning community. Our model involves voxel-based representations of 3D shapes in the 3D shape-aware energy function.

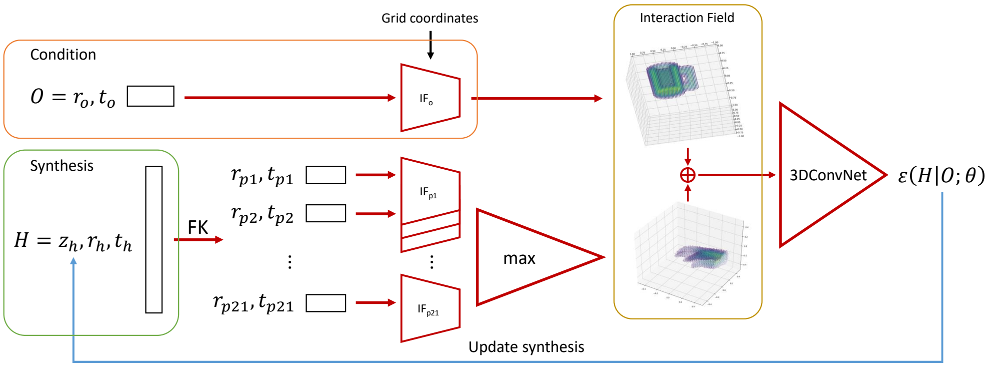


Figure 4.1: Pipeline of our model. Red shapes indicate differentiable steps. Blue arrow indicates the synthesis step using Langevin dynamics.

4.3 Conditional Descriptor Nets as Intuitive Grasping

Problem definition. This chapter aims to train a conditional descriptor network as intuitive grasping, a conditional distribution $p(H|O)$ for the geometric configuration H of a grasping hand conditioned on a given object with geometric configuration O to be grasped. We shall learn $p(H|O)$ from human demonstrations, which are represented by hand-and-object pairs $\{(H_i, O_i), i = 1, ..., n\}$. With the learned model $p(H|O)$, given an object with observed geometric configuration O_{obs}, we can generate meaningful geometric configurations of grasping hands by sampling from $p(H|O_{\text{obs}})$ via Markov Chain Monte Carlo (MCMC).

Model and learning. The model is based on an objective function $f(H, O; \theta)$ defined on (H, O), where θ collects all parameters. Serving as a negative energy function, $f(H, O; \theta)$ defines a joint energy-based model

$$P(H, O; \theta) = \frac{1}{Z(\theta)} \exp\left[f(H, O; \theta)\right], \tag{4.1}$$

where $Z(\theta) = \int \exp\left[f(H, O; \theta)\right] dH dO$ is the normalizing constant that is analytically intractable. We denote energy function $\mathcal{E}(H, O; \theta) = -f(H, O; \theta)$. Fixing an object with geometric configuration O, $f(H, O'; \theta)$ evaluates the value of the grasping H for the object represented by O, and $-f(H, O; \theta)$ plays a role of conditional energy function. The conditional probability for intuitive grasping is defined by

$$P(H|O;\theta) = \frac{P(H,O;\theta)}{P(O;\theta)} = \frac{P(H,O;\theta)}{\int P(H,O;\theta)dH} = \frac{1}{Z(O,\theta)} \exp\left[f(H,O;\theta)\right], \tag{4.2}$$

where $Z(O,\theta) = Z(\theta)p(O;\theta)$. We call this model the conditional descriptor nets, because $f(H,O;\theta)$ describes some statistical features about object grasping. Given the observed human demonstrations of grasping objects $\{(H_i,O_i), i = 1,...,n\}$, we learn the model by finding the optimal θ to maximize the log-likelihood function $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log P(H_i|O_i;\theta)$. The gradient of $L(\theta)$ is

$$L'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} f(H_i,O_i;\theta) - \mathrm{E}_{p(H|O;\theta)}\left[\frac{\partial}{\partial\theta} f(H,O;\theta)\right] \tag{4.3}$$

where the $\mathrm{E}_{p(H|O;\theta)}$ is the expectation with respect to $p(H,O;\theta)$. Since the expectation term is analytically intractable, we approximate it with a MCMC, such as Langevin dynamics, which samples from the current conditional distribution $P(H|O;\theta)$ by iterating the following steps:

$$H_{\tau+1} = H_\tau - \frac{\Delta}{2}\frac{\partial}{\partial H}\mathcal{E}(H_\tau,O;\theta) + \sqrt{\Delta}\epsilon_\tau = H_\tau + \frac{\Delta}{2}\left[\frac{\partial}{\partial H} f(H_\tau,O;\theta)\right] + \sqrt{\Delta}\epsilon_\tau \tag{4.4}$$

where τ is the current step and Δ is the step size. $\epsilon_\tau \sim N(0,I)$ is a random noise sampled from a Gaussian distribution. Each step of Langevin dynamics performs gradient descent with a random perturbation to escape local minima to minimize the energy function. Suppose we sample $\tilde{H}_j$ for each O_i from the distribution $p(H|O_i;\theta)$, Equation 4.3 can be approximated by

$$L'(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} f(H_i,O_i;\theta) - \frac{1}{n}\sum_{j=1}^{n} \frac{\partial}{\partial\theta} f(\tilde{H}_j,\tilde{O}_j;\theta) \tag{4.5}$$

We then update the model parameters by $\theta^{(t+1)} = \theta^{(t)} + \beta L'(\theta^{(t)})$.

Model understanding. The learning of our model follows what Grenander [GMM07] called "analysis by synthesis" scheme, which alternates the sampling step defined by equation (4.4) and the learning step involving computing the gradient of θ in equation (4.5). The keys of both steps are the computations of $\frac{\partial}{\partial H} f(H, O; \theta)$ and $\frac{\partial}{\partial \theta} f(H, O; \theta)$, each of which can be easily computed by back-propagation. The learning process corresponds to learning a value function $f(H, O; \theta)$ for intuitive grasping by shifting high value region of $f(H, O; \theta)$ from the currently generated grasping $\{(\tilde{H}_i, O_i), i = 1, ..., \tilde{n}\}$ toward human demonstration $\{(H_i, O_i), i = 1, ..., n\}$, while the MCMC sampling process corresponds to the exploration of the hand configuration space in order to generate hand poses given an object that maximize the value function.

4.4 3D Shape Aware Energy function

3D Shape Aware. The conditional descriptor model proposed in section 4.3 relies on a well-designed energy function $f(H, O)$ that captures the interaction between the hand poses H and the object O to be grasped. Given that the hand-object geometry relationship of grasping can be better captured in the 3D space than in the configuration parameter space, we propose a 3D shape-aware energy function that maps H and O into negative energy. The energy function captures the hand-object interaction in the 3D voxel space instead of directly building a bottom-up multilayer perceptron on the concatenated version of hand-object configuration (H, O).

Let $O = (r_o, t_o)$ denote the geometric configuration of object, where r_o and t_o are the rotation and translation of the object respectively, and $H = (z_h, r_h, t_h)$ denote the geometric configuration of hand, where z_h, r_h, and t_h are the joint angle, rotation and translation of the hand respectively. We define $\mathbf{I}$ as an operation that takes O and H as inputs and output a 3D voxel-based hand-object interaction V, which is represented by a channel-wise concatenation of the voxel-based 3D shapes of object V_o and hand V_h. The 3D shape aware energy function is given by $f(H, O; \theta) = g(\mathbf{I}(H, O); \theta)$, where g is a bottom-up 3D convolutional neural network that maps $\mathbf{I}(H, O)$ into negative energy with parameters θ. Figure 4.1 illustrates the structure of the proposed 3D shape aware energy function.

IF-module. Let X be a set of coordinates from a regular grid. Let $\mathcal{D}_o^O(X)$ represent the signed distances from coordinate X to the object under transformation O, and $\mathcal{D}_h^H(X)$ represent the signed distances from coordinate X to the hand under transformation H. $\mathcal{D}_S(X)$ is positive if X is inside shape S, is negative if X is outside of S, and is 0 if X is on the surface of S. Both $\mathcal{D}_o^O(X)$ and $\mathcal{D}_h^H(X)$ can be reshaped to voxel-based 3D representations V_o and V_h. Here we expect $\mathbf{I}$ to be a differentiable function such that the derivative of the energy function can be performed by back-propagation. We approximate $\mathcal{D}_o^O(X)$ with an IF-module and $\mathcal{D}_h^H(X)$ with a combination of IF-modules.

The IF-module for object is a differentiable function $\mathbf{IF}_o(r_o, t_o, X)$, which takes the rotation r_o, translation t_o and coordinate X as inputs and outputs the signed distance from X to the object with transformation of r_o and t_o. The DeepSDF model [PFS19] learned an MLP to approximate a transformation-invariant function $\bar{\mathbf{IF}}_o(X; \alpha^{\bar{\mathbf{IF}}_o})$, whose output is the signed distance from X to the object with no transformation. We build $\mathbf{IF}_o$ on the pre-trained $\bar{\mathbf{IF}}_o$ and fix $\alpha^{\bar{\mathbf{IF}}_o}$, i.e., $\mathbf{IF}_o(r_o, t_o, X) = \bar{\mathbf{IF}}_o(r_o^{-1}(X - t_o); \alpha^{\bar{\mathbf{IF}}_o})$.

The shape of a human hand is a combination of multiple rigid parts, and the joint angles control the overall hand shape. Due to the complexity of hand shape and possible topology changes of a hand shape across different gestures, it is challenging to represent the hand with a single IF module. To address this issue, we implemented a differentiable forward kinematics module to compute the rotation and translation of each rigid part from the joint angles. The signed distance to the hand is the max of signed distances to all parts.

Comparison to Interaction Bisector Surface. Interaction Bisector Surface (IBS) [ZWK14] is a general representation for two-party 3D interactions. For a pair of 3D shapes A, B, the IBS is defined as the equidistant surface between A and B. We argue that the IBS is a special case of Interaction Field. For IBS, we are interested in a surface on which all points satisfy the constraint $\mathcal{D}_A(\cdot) == \mathcal{D}_B(\cdot)$. Although IBS is a strong baseline for classifying and retrieving 3D interactions, it is a computationally intensive task. With pretrained IF modules, we can quickly sample a set of points on the IBS via MCMC.

Physics. Although we can not write all the physics laws in the term of interaction field, many important ones involved in two-party interactions can be encoded. **Penetration** of two objects A, B can be written as $\sum_{\mathbf{x}\in\mathbf{X}} \sqrt{\mathcal{D}_A(\mathbf{x})^2 + \mathcal{D}_B(\mathbf{x})^2}$, where $\mathbf{X} = \{x : \mathcal{D}_A(x) > 0 \text{ and } \mathcal{D}_B(x) > 0\}$. **Optimization for contact** can be written as $\arg\min_H \sum_{\mathbf{x}\in\mathbf{C}}$ $\sqrt{\mathcal{D}_A^H(\mathbf{x})^2 + \mathcal{D}_B^H(\mathbf{x})^2}$, where $\mathbf{C}$ is a collection of points on predefined contact regions on A and B. Recall that $\frac{\partial \mathcal{D}_O(X)}{\partial X}$ is the normal vector of object O at coordinate X. We can also enforce **surface alignment** with $\arg\min_H \sum_{\mathbf{x}\in\mathbf{C}} \| \frac{\partial}{\partial \mathbf{x}} \mathcal{D}_A^H(\mathbf{x}) - \frac{\partial}{\partial \mathbf{x}} \mathcal{D}_B^H(\mathbf{x}) \|^2$.

4.5 Experiment

Collecting Human Demonstrations of Grasping Objects. In order to learn intuitive grasping by the conditional descriptor networks proposed in section 4.3 from human demonstration, we design a framework to collect a dataset of grasping objects, which contains over 56k frames of natural human grasping of 10 household objects. Each frame is represented by a vector containing the hand pose, object pose, joint angles, and forces on each predefined region. Table 4.1 describes the composition of a frame vector. The 3D scene can be reconstructed from the frame vectors. Figure 4.2 shows a visualization of our dataset.

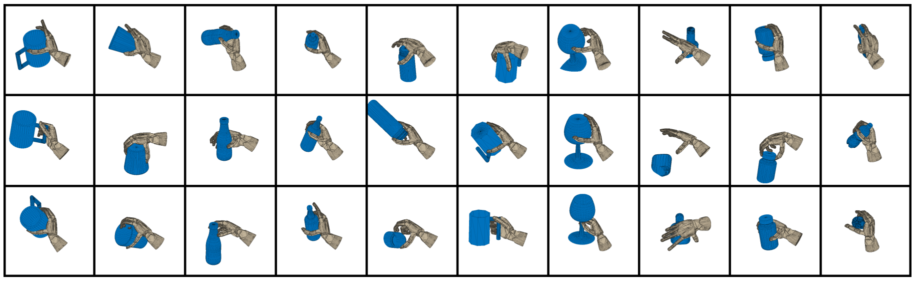


Figure 4.2: Snapshots in our dataset. Each column shows three different forms of grasping of an object.

time	obj. position	obj. orientation	joint angle	force	total
1	28 × 3 x-y-z	28 × 4 quaternion	36	19	252

Table 4.1: Frame vector composition

Our data collection environment consists of a set of sensors, a VR environment, and a communication process. Our sensor composition consists of a data glove developed in [LXM17] and an HTC Vive tracker. We use the glove to collect joint angles and the HTC Vive tracker to collect the 6D pose of the upper arm, which is the root node of the hand. The collected data is streamed to MuJoCo [KT15] in real-time to control an MPL hand model [JBB11]. MuJoCo renders the hand model according to the received data to provide visual feedback for participants to adjust their actions. The communication channel between our tracker and the VR environment is implemented in ROS. Figure 4.3 depicts the overall schematic diagram of our environment setup.

We ask ten right-handed subjects to grasp the given objects in VR steadily without external support except for visual feedback. Each participant is given 60 seconds to perform as many forms of natural stable grasping for each object. To increase data variance, we ask the participants to move their fingers slightly for each form of grasping. To avoid interfering with human intuitive grasping strategy, we asked subjects to grasp with whatever poses they think comfortable and natural, without telling the subjects to perform the specific, well-defined grasping poses as categorized in [FRS15]. As proved in [US00], people can grasp arbitrary virtual objects with realistic grasping poses intuitively without any haptic feedback. The collected dataset serves as human demonstrations, from which our model learns intuitive grasping.

Data Processing. For each frame, we align the data so that its center of gravity is at the origin, and the palm of the human hand is pointing along the x-axis. We use a 6-vector to represent a 3D rotation and a 2-vector to represent and 1D rotation for better performance with neural networks [ZBL18]. Therefore, our H is a 53-vector with the following correspon-

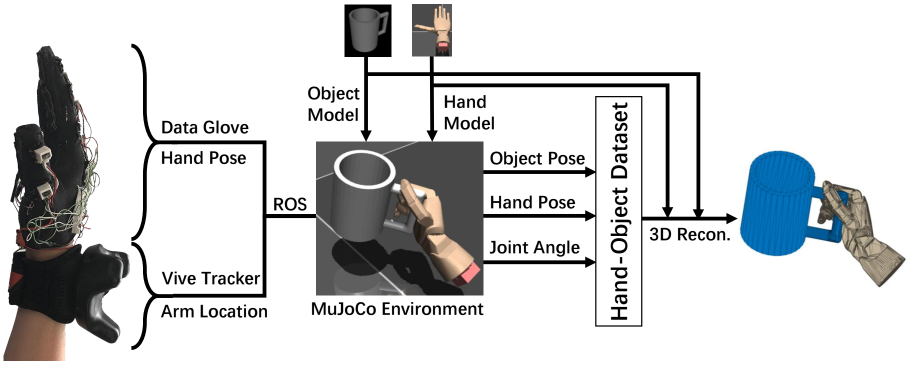


Figure 4.3: The schematic diagram of data collection process.

dence. 0-43: joint angles for 22 joints. 44-49: rotation of the hand. 50-52: translation of hand. Our Y is a 9-vector for rotation and translation of the object. We further perform a singular value decomposition on H and keep the top 24 principal components.

Training. We use a 3D convolutional neural network to estimate the energy as a function of the interaction field. The network consists of 4 3D convolutional layers, with the input size being $N \times 64 \times 64 \times 64 \times 2$, where N is the batch size. The layers have 128 $8 \times 8 \times 8$, 128 $8 \times 8 \times 8$, 64 $8 \times 8 \times 8$ and 64 $6 \times 6 \times 6$ filters respectively. The bottom layer is followed by a fully-connected layer with one output unit. There is no downsampling operation, and each layer is followed by ReLU activation except for the last layer.

We use 4 Titan RTX graphics cards in parallel to train our model. Due to memory limits, we randomly pick 4000 frames from our dataset to train our model. We use $l = 90$ Langevin steps in the sampling process with step size $\Delta\tau = 0.1$. Each epoch takes $\sim$ 20 hours to train. During training, we make two key observations and applied two tricks for easier training. **Soft start.** We observe that in the first epoch, when the learned energy landscape is yet to be meaningful, we do not need a long sequence of Langevin dynamics to sample from $P(H|O;\theta)$. Therefore, we adjust the number of Langevin steps $l_t = \min\left(90, \text{floor}\left(\frac{t}{10}\right)\right)$ at the t-th step. **Adaptive step size.** We also observe that the gradient on the control vector H suffers from unbalanced gradients. Specifically, more sensitive signals (such as translation and rotation of hand) receive much larger gradients than less sensitive ones (such as the angle of the terminal joint of each finger). In our experiment, the difference between gradients can be as large as 20X. We adopt the idea from Adam [KB14] to set the step size of each signal to be inversely proportional to a moving average of its gradient.

Results and Analysis. Figure 4.4 shows qualitative synthesis results from our model and two baseline models. For each example, we run 16 parallel syntheses and show the one with the lowest energy. We can observe that our model can synthesize different grasping patterns for the same object. We can also observe failure cases where the fingers either penetrate or do not touch the object. We expect such behavior because we only model the geometric features of human grasping that correspond to the brain activity in the preshape stage. Physical constraints can be met with additional optimization procedures and is out of the scope of this chapter. We also see some failure cases with complex shapes, such as wine glass. A possible explanation is that the energy landscape for grasping complex shapes is much more complex than simpler shapes. Additional effort in the sampling process may be required to improve our robustness over complex shapes. We also observe that our model significantly outperforms baseline methods [MO14, KW13] with the same architecture. A possible reason is that both baseline methods use two networks. The interplay between two networks makes efficient training exceptionally hard. Our method only contains one network and therefore is easier to train. Another explanation is that both the generator in conditional GAN and the encoder in VAE computes H as a function of the geometrical features of the object, without leveraging the geometrical relationship between both shapes.

Since our method does not model physical stability, we cannot evaluate our result with simulation. Instead, we ask volunteers to rate synthesis quality for our synthesized results. The results shown in Table 4.2 show that our method can synthesize much more realistic graspings compared to conditional GAN and VAE. Compared with GAN and VAE, our model does not rely on any extra network structures, such as discriminator in GAN and inference network in VAE.

To illustrate that intuitive grasping can be applied to actual grasping with physically-

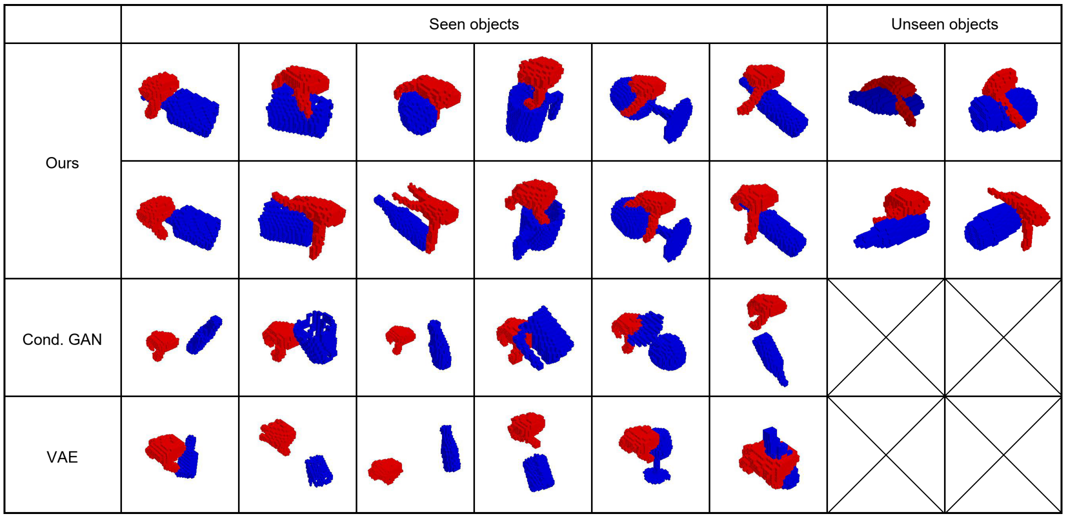
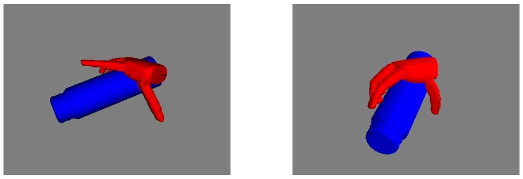


Figure 4.4: Qualitative evaluation of our synthesized examples. The blue voxels are the objects, and the red ones are the hands. The first 6 columns show synthesis results on objects that have been seen during the training stage. The last two columns show synthesis results on objects that have never been seen during training. The animated version of this figure can be found in supplementary materials.

Description	Our method	Our method (U)	Conditional GAN	VAE
(1)	16.25%	10.63%	0.63%	1.25%
(2)	30.63%	21.88%	1.25%	6.88%
(3)	35.00%	44.38%	4.38%	6.88%
(4)	18.13%	23.13%	93.75%	85.00%

Table 4.2: Human evaluation. We show ten unmarked animations of synthesized examples for each method to 16 participants and ask them to choose the best description from 4 options for each example. We report the proportions of chosen descriptions for each method. *Our method (U)* shows the result of our method on unseen objects. The options are: (1) A perfect grasping. (2) A good grasping with error in details. (3) Close to a good grasping. (4) Not a grasping at all.



(a) Before physics (b) After physics

Figure 4.5: An example of physically-based optimization. We observe that the grasping hand is qualitatively more natural than the proposed shape after optimization.

based optimization, we implement a naive optimization algorithm to reduce penetration and promote contact. Our algorithm alternatively performs penetration reduction and contact promotion. **Penetration reduction.** We define a penetration loss $\mathcal{L}^p(H) = \sum_{x \in P} \mathcal{D}_h(x)^2 + \mathcal{D}_o(x)^2$ where P is a set of points sampled from the surfaces of the object and the hand, where the signed distances from each point to both the hand and the object are non-negative. **Contact promotion.** We define a contact loss $\mathcal{L}^c(H) = \sum_{x \in P^c} \mathcal{D}_h(x)^2 + \mathcal{D}_o(x)^2$ where P^c is a set of sampled points between the object and the hand. P^c is sampled with MCMC sampling where the energy term is $\mathcal{E}(x) = \mathcal{D}_o(x) + \mathcal{D}_h(x)$. The MCMC sampling starts from points on the inner surface of the hand model. We update H in both steps by gradient descent on the corresponding loss. Figure 4.5 shows an example of our physically-based optimization.

4.6 Conclusion

In this chapter, we propose a conditional descriptor network to represent intuitive grasping, a deep conditional energy-based model with ConvNet structure defined on the configurations of both hand pose and object as negative energy. The model can be learned from human demonstrations and can synthesize possible intuitive grasping for given objects. We show that our method is superior to the baseline methods we developed based on GAN and VAE. We also introduce a 3D shape-aware energy function with Interaction Field as an intermediate representation for 3D shape interactions. In addition, we collect a grasping dataset serving as human demonstrations, from which the proposed model learns intuitive grasping. The dataset contains over 56k frames of natural and realistic human grasping that contains shape, joint angles, contact surfaces, and forces. We expect future works to perfect the generation of intuitive grasping and optimize generated intuitive grasping proposals based on physical constraints.

CHAPTER 5

Synthesizing Diverse and Physically Stable Grasps with Arbitrary Hand Structures by Differentiable Force Closure Estimation

The previous chapter describes an energy-based method of learning humanoid grasping. While the method can synthesize realistic grasps, the synthesis results are restricted to a single type of grasping. This chapter breaks down the shape-level interaction into physics-level and shows significant improvement of synthesis diversity and robustness of the new method.

5.1 Introduction

Grasp synthesis has been a challenging task due to the complexity of hand kinematics. Although force closure has been commonly accepted to evaluate the quality of the generated grasps, researchers usually avoid using it as an optimization objective: Computing force closure requires solving for contact forces, which is an optimization problem itself. As a result, using force closure as the optimization objective in grasp synthesis would produce a notoriously slow and nested optimization problem. Instead, researchers have turned to analytical or data-driven methods [BMA13].

Analytical methods use manually derived algorithms. Due to the intrinsic complexity of the grasp synthesis, these methods [PSB93, PSS97, LLC03] typically perform only in limited settings (usually on power grasps as defined in grasp taxonomy) and are only applicable to specific robotic hand structures. Modern approaches focus more on data-driven methods [TGB20, BKK20, KYZ20], which relies on large datasets of human demonstrations. Although these methods are able to reproduce (and even interpolate) similar but different grasps compared to human demonstrations, they are inherently difficult to generalize (especially to extrapolate) to arbitrary hand kinematics and unseen grasp types. Furthermore, these data-driven methods usually do not consider the physical stability in producing grasps, making them difficult to deploy on physical robots.

In this chapter, different from analytical or data-driven approaches, we derive a fast and differentiable estimation of force closure. It can be computed within **milliseconds** on modern desktops, significantly faster than classic algorithms. Such fast computation of force closure opens a new venue for grasp synthesis. Since it does not rely on training data or restrict to specific robotic hand structures, the proposed method can be applied to arbitrary hand structures to synthesize diverse types of grasps with physical stability.

Specifically, our method is based on two simple yet reasonable and effective assumptions: **zero friction** and **equal magnitude of contact forces**, which avoid solving the contact forces in the inner optimization problem. Intuitively, such assumptions indicate that the contact force on each contact point becomes simply the object's surface normal on that point. As such, the overall nested optimization problem is converted to minimizing the errors that violate the above assumptions; see an example in Fig. 5.1. In experiments, we demonstrate that our estimated error reflects the difference between surface normal vectors and force closure contact force vectors. We further devise a grasp energy function based on the estimated force closure and validate the force-closure grasp synthesis by minimizing the energy function.

This chapter makes two primary contributions:[1]

[1] See additional material on our website `https://sites.google.com/view/ral2021-grasp/`.

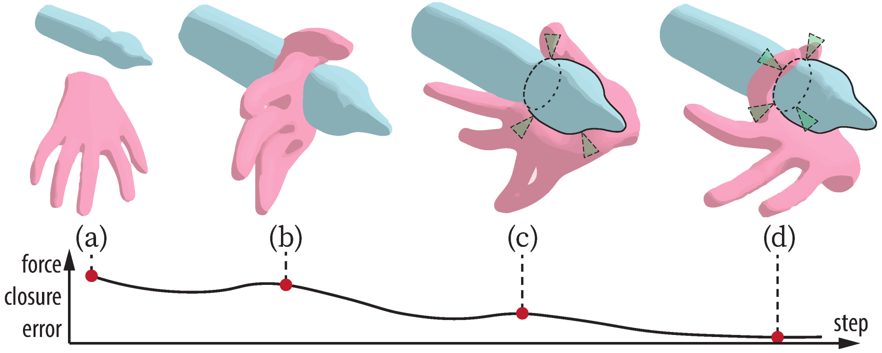


Figure 5.1: **Grasp synthesis process by minimizing force closure error.** The green trianglets in (c)(d) denote the friction cones at contact points used to calculate force closure.

1. We formulate a fast and differentiable estimation of force closure, computed within milliseconds.
2. We propose a grasp synthesis algorithm that can generate diverse types of grasps with arbitrary hand structures without any training data.

5.2 Related Work

Grasp synthesis literature can be roughly categorized into two schools of thought: analytic and data-driven approach.

The analytic approach generates grasps by considering kinematics and physics constraints [SEB12]. Although force closure has been commonly adopted as the physics constraint [RMF12, PMG12, RSG12, Mur17], primary efforts have been devoted to simplify the search space (*e.g.*, [PSB93, PSS97, LLC03]) as testing force closure is expensive. However, these methods are only effective in specific settings.

The data-driven approach leverages recent advancements in machine learning to estimate grasp points. Despite promising progress [SDN08, CA09, RKK09, LLS15, MLN17, LPK18], this approach relies heavily on large datasets to learn successful grasps, with a particular focus on grippers with limited DoF. Although recent literature [AKH12, LS15, BHH20, GTT21] extends this approach to more complex hand models, capable of generating more realistic grasps, they still rely on the expensive and tedious collection of human demonstration data. Fundamentally, it is non-trivial for a data-driven approach to generalize the learned model to other hand kinematics.

An example that does not fall into either of the above categories is the popular toolkit of GraspIt! [MA04]. It generates grasps by initializing hand pose randomly, squeezing the fingers as much as possible, and ranking them by a user-defined grasp metric (*e.g.*, a force closure metric). Although this method can generate valid grasps, it is highly inefficient and incapable of generating diverse grasps [CPA20].

A **force-closure** grasp is a grasp with contact points $\{x_i \in \mathbb{R}^3, i = 1, ..., n\}$ such that $\{x_i\}$ can resist arbitrary external wrenches with contact forces f_i, where f_i lies within the friction cones rooted from x_i. The angles of the friction cones are determined by the surface friction coefficient: The stronger the friction, the wider the cone. The force-closure metric is, therefore, irrelevant to the actual hand pose, but only relevant to the contact points and friction cones.

To test whether a set of contact points form a force-closure grasp, the first step is solving an optimization problem regarding contact forces rooted from the points [BW07, HTL00]. Although various methods have been devised, they all require iterations to jointly solve an auxiliary function, *e.g.*, a support function [ZC09], a bilinear matrix inequality [DMT18], or a ray shooting problem [Liu99]. As a result, solving force-closure grasps under the constraint of hand kinematics and force closure becomes a nested optimization problem.

Human grasps can be organized into a **grasp taxonomy** [FRS15]; humans perform grasps to provide different levels of power and precision. According to the taxonomy [FRS15], most existing grasp synthesis methods focus on synthesizing power grasp, including both analytical approaches [RMF12, PMG12, RSG12, Mur17] and data-driven approaches [KYZ20]. At a high cost of annotating object-centric grasp contact information, some data-driven approaches [BHH20, TGB20] demonstrate a certain level of capability to generate a broader range of grasp types.

The diversity of grasp synthesis can be evaluated by comparing the types of generated grasps against the ones in the grasp taxonomy. Corona *et al.* [CPA20] provide a dataset, YCB-Affordance, of 3D grasps with corresponding grasp types, which covers all 33 grasp types as defined in [FRS15].

5.3 Differentiable Force Closure

Formally, given a set of n contact points $\{x_i \in \mathbb{R}^3, i = 1, ..., n\}$ and their corresponding friction cones $\{(c_i, \mu)\}$, where c_i is the friction cone axis and μ is the friction coefficient, a grasp is in *force closure* if there exists contact forces $\{f_i\}$ at $\{x_i\}$ within $\{(c_i, \mu)\}$ such that $\{x_i\}$ can resist arbitrary external wrenches. We follow the notations in Dai *et al.* [DMT18] to define a set of contact forces to be force closure if it satisfies the following constraints:

$$GG' \geq \epsilon I_{6\times 6}, \tag{5.1a}$$

$$Gf = 0, \tag{5.1b}$$

$$f_i^T c_i > \frac{1}{\sqrt{\mu^2+1}} |f_i|, \tag{5.1c}$$

$$x_i \in S, \tag{5.1d}$$

where S is the object surface, and

$$G = \begin{bmatrix} I_{3\times 3} & I_{3\times 3} & ... & I_{3\times 3} \\ \lfloor x_1 \rfloor_\times & \lfloor x_2 \rfloor_\times & ... & \lfloor x_n \rfloor_\times \end{bmatrix}, \tag{5.2}$$

$$\lfloor x_i \rfloor_\times = \begin{bmatrix} 0 & -x_i^{(3)} & x_i^{(2)} \\ x_i^{(3)} & 0 & -x_i^{(1)} \\ -x_i^{(2)} & x_i^{(1)} & 0 \end{bmatrix}. \tag{5.3}$$

The form of $\lfloor x_i \rfloor_\times$ ensures the cross product $\lfloor x_i \rfloor_\times f_i = x_i \times f_i$; $f = [f_1^T f_2^T ... f_n^T]^T \in \mathbb{R}^{3n}$ is the unknown variable of contact forces. In Eq. (5.1a), ϵ is a small constant. $A \geq B$ means $A - B$ is positive semi-definite, *i.e.*, it is symmetric, and all its eigenvalues are non-negative. Eq. (5.1a) states that G is full rank. Eq. (5.1c) describes the constraint that f_i should not deviate from the friction cone $\{(c_i, \mu)\}$.

5.3.1 Relaxation

Of note, Eq. (5.1b) is bilinear on x_i and f_i. Given a set of contact points $\{x_i\}$, verification of force closure requires finding a solution of $\{f_i\}$. The time complexity for computing such a solution is linear w.r.t. the number of contact points [DMT18]. However, we observe that under the assumption of zero friction and the contact forces have equal magnitude, Eq. (5.1b) can be relaxed and rewritten to

$$GG' \geq \epsilon I_{6\times 6}, \tag{5.4a}$$

$$Gc < \delta, \tag{5.4b}$$

$$x_i \in S, \tag{5.4c}$$

where $c = [c_1^T c_2^T ... c_n^T]^T$ is the set of friction cone axes; c_i can be simply replaced by the surface normal of the object on x_i, which is easily accessible in many shape representations. By combining Eq. (5.4b) along with Eq. (5.1a) and Eq. (5.1c), we no longer need to solve the unknown variable f. The constraints of x_i becomes quadratic. Hence, the verification of force closure can now be computed extremely fast. The error in Gc reflects the difference between force closure contact forces and friction cone axes.

5.3.2 Implications of Assumptions

Enforcing zero friction and equal magnitude contact forces may *seem* to eliminate a large pool of force-closure contact-point compositions. In practice, however, this is not the case: A residual in $||Gc||_2$ indicates that the zero-friction and equal-magnitude contact forces do

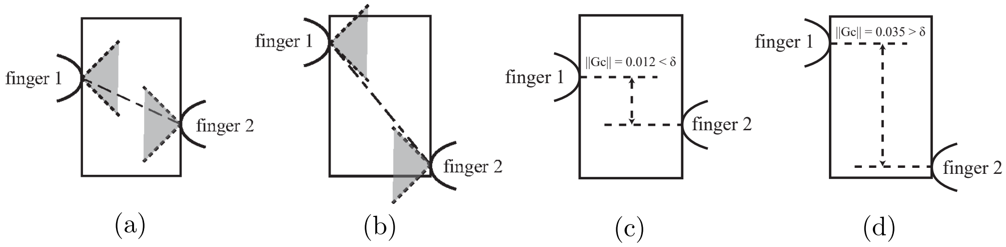


Figure 5.2: **A 2D illustration of the classical force closure test and our estimated force closure error.** (a)(b) Two scenarios passed and failed the classical force closure test. (c)(d) Our estimated force closure error on the same scenarios as in (a)(b).

not perfectly cancel out. Such residual could have been reduced to zero should friction and magnitude difference be allowed. Fig. 5.2 illustrates the implication of our assumptions in 2D. Specifically, for cases where our assumptions are violated, $||Gc||_2$ would have a non-zero error δ. Eq. (5.1b) can be rewritten as

$$Gf = G(f_n + f_t) = 0, \tag{5.5a}$$

$$G\frac{f_n}{||f_n||_2} = -\frac{Gf_t}{||f_n||_2}, \tag{5.5b}$$

$$Gc = -\frac{Gf_t}{||f_n||_2}, \tag{5.5c}$$

where f_n and f_t are the normal and tangential components of contact force f in the force closure model. Having an error in $||Gc||_2$ essentially implies that there is a friction components in the contact forces to form a force closure grasp, and the error δ indicates the magnitudes of the friction components.

To further verify our interpretation, we randomly sample 500,000 grasps, each containing three contact points on the surface of a unit sphere. For each grasp, we compute the *minimum* friction coefficient μ_0 required for the grasp to satisfy the traditional force closure constraints described in Eq. (5.1). We plot the error δ of our estimated force closure value against μ_0 in Fig. 5.3 to show that the relation between μ_0 and δ is almost linear.

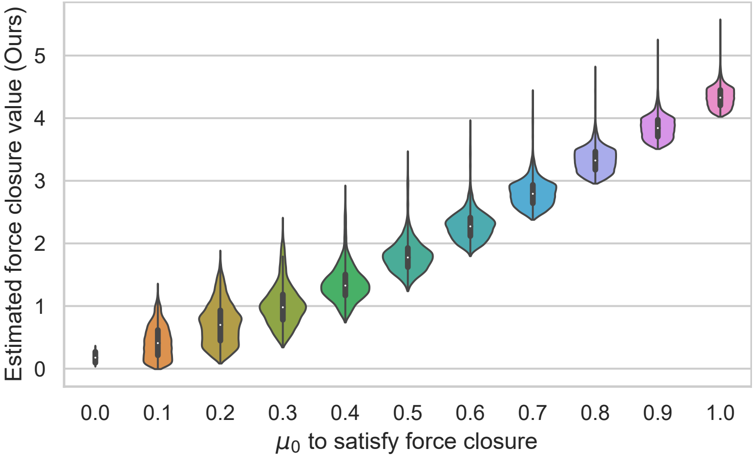


Figure 5.3: **Estimated force closure error δ (x-axis) against minimum friction coefficient μ_0 (y-axis).** The violinplots [HN98] show the distributions of all estimated force closure errors that require a minimum friction coefficient μ_0 to pass the classical force closure test. Overall, these two are linearly correlated.

5.4 Grasp Synthesis

5.4.1 Formulation

We formulate the grasp synthesis problem as sampling from a conditional Gibbs distribution:

$$P(H|O) = \frac{P(H|O)P(O)}{P(O)} = \frac{P(H,O)}{P(O)} \tag{5.6}$$

$$\propto P(H,O) = \frac{1}{Z}\exp^{-E(H,O)}, \tag{5.7}$$

where H denote the hand, O the object, Z the intractable normalizing constant, and $E(H,O)$ is the grasp energy. $E(H,O)$ can be further decomposed to

$$E(H,O) = \min_{x\subset S(H)} E_{grasp}(H,x,O) \tag{5.8}$$

$$= \min_{c\subset S(H)} FC(c,O) + E_{prior}(H) + E_{pen}(H,O), \tag{5.9}$$

where $S(H)$ is a set of points sampled from the hand surface determined by the hand pose H, $c \subset S(H)$ the set of contact points selected from hand surface, and $FC(c,O)$ the force closure formulation from Eq. (5.11).

$E_{prior}(H)$ is the *energy prior* of the hand kinematic tree; its exact form depends on the hand definition. $E_{pen}(H,O) = \sum_{x \in S(H)} \sigma(x,O)$ is the *penetration energy*, where $S(H)$ is a set of points sampled from hand surface, and $\sigma(x,O)$ is a modified distance function between a point x and an object O:

$$\sigma(x,O) = \begin{cases} 0 & \text{if } x \text{ outside } O \\ |d| & \text{otherwise} \end{cases}, \tag{5.10}$$

where d is the distance from x to surface of O.

5.4.2 Algorithm

Due to the complexity of human hand kinematics, the landscape of our grasp energy is highly non-convex. With a naive gradient-based optimization algorithm, it is very likely to get stuck at bad local minima. We use a modified Metropolis-adjusted Langevin algorithm (MALA) to overcome this issue; see the algorithm details in Algorithm 2. The random walk aspect of Langevin dynamics provides the chance of escaping bad local minima. Our algorithm starts with random initialization of hand configuration H and contact points $c \subset S(H)$. Next, we run our algorithm L iterations to update H, c and maximize $P(H,O)$. In each iteration, our algorithm randomly decides to update either the hand configuration by Langevin dynamics or one of the contact points to a point uniformly sampled from the hand surface.

To sample contact points from the hand surface, we start with random initialization of contact points, and randomly update one of the contact points to a point uniformly sampled from the hand surface each time. Notice that different compositions of contact points in fact correspond to different grasp types as they contribute to some of the classification basis of the grasp taxonomy, including virtual finger assignment and opposition type. Hence, this step is crucial for exploring different types of grasps. In practice, we also empirically find that this step is crucial for escaping bad local minima.

```
Algorithm 2: Modified MALA Algorithm
Input: Energy function $E_{grasp}$, object shape $O$, step size $\eta$, Langevin steps $L$,
       switch probability $\rho$
Output: grasp parameters $H, c$
Initialize $H, c$
for $step = 1 : L$ do
    if $rand() < \rho$ then
        Propose $H^*$ according to Langevin dynamics
            $H^* = H - \frac{\eta^2}{2}\frac{\partial}{\partial H}E_{grasp}(H, c, O) + \eta\epsilon,$
          where $\epsilon \sim N(0, 1)$ is a Gaussian noise
    else
        Propose $c^*$ by sampling from $S(H)$
    end
    Accept $H \leftarrow H^*, c \leftarrow c^*$ by Metropolis algorithm using energy function $E_{grasp}$
end
```

5.4.3 Refinement

While our modified MALA algorithm can produce realistic results, there may still be physical inconsistencies in the synthesized examples, such as penetrations and gaps between contact points and object surface. To resolve these issues, we further refine the synthesized results by minimizing E_{grasp} using gradient descent on H. We do not update the contact point selection c in this step since we hope to focus on optimizing the physical consistency in this step, rather than exploring the grasp landscape for diverse grasp types.

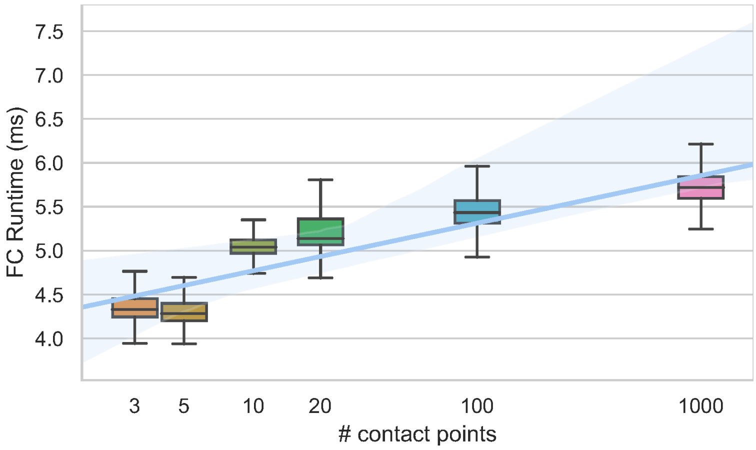


Figure 5.4: **Boxplot and log-linearly fitted curve of the runtime w.r.t. to the number of contact points.** We run a simulated test of force closure with 3, 5, 10, 20, 100, and 1000 contact points for 1,000 iterations. X-axis is the number of contact points in log scale. Y-axis is the runtime of our force closure error estimation. The shaded area denotes the 95% confidence interval.

5.5 Experiment

5.5.1 Experiment Setup

Hand Model We use MANO [RTB17] to model the humanoid hand. It is a parameterized 3D hand shape model that maps low-dimensional hand poses to 3D human hand shapes. We use the norm of the PCA weights of the hand pose as $E_{prior}(H)$. Since MANO vertices are distributed uniformly across the hand surface, we sample points from the hand surface by directly sampling from MANO vertices.

Object Model We use the DeepSDF model [PFS19] to model the objects to be grasped. DeepSDF is a densely connected neural network that implicitly represents the surface of a shape. The model estimates the signed distance from a position to an object surface. The signed distance is negative if the point is inside the object, and is positive if the point is outside the object. The set of points with zero distance compose the surface of the object. We can obtain the surface normal of the object by taking the derivative of the signed distance w.r.t. the input position.

5.5.2 Runtime Efficiency

Figure 5.4 shows that the time complexity of testing force closure with a fitted log-linear curve w.r.t. the number of contact points. Each test takes 4-6ms to run on an NVIDIA Titan RTX GPU, significantly faster than the exact solution [DMT18]. We also observe that roughly 80% of the total runtime is spent at the computation of surface normal; this operation is particularly slow because it takes a derivative of the DeepSDF model. Taken together, these empirical simulated results indicate that a further improvement in runtime efficiency would be achievable with a more computationally tractable object shape representation.

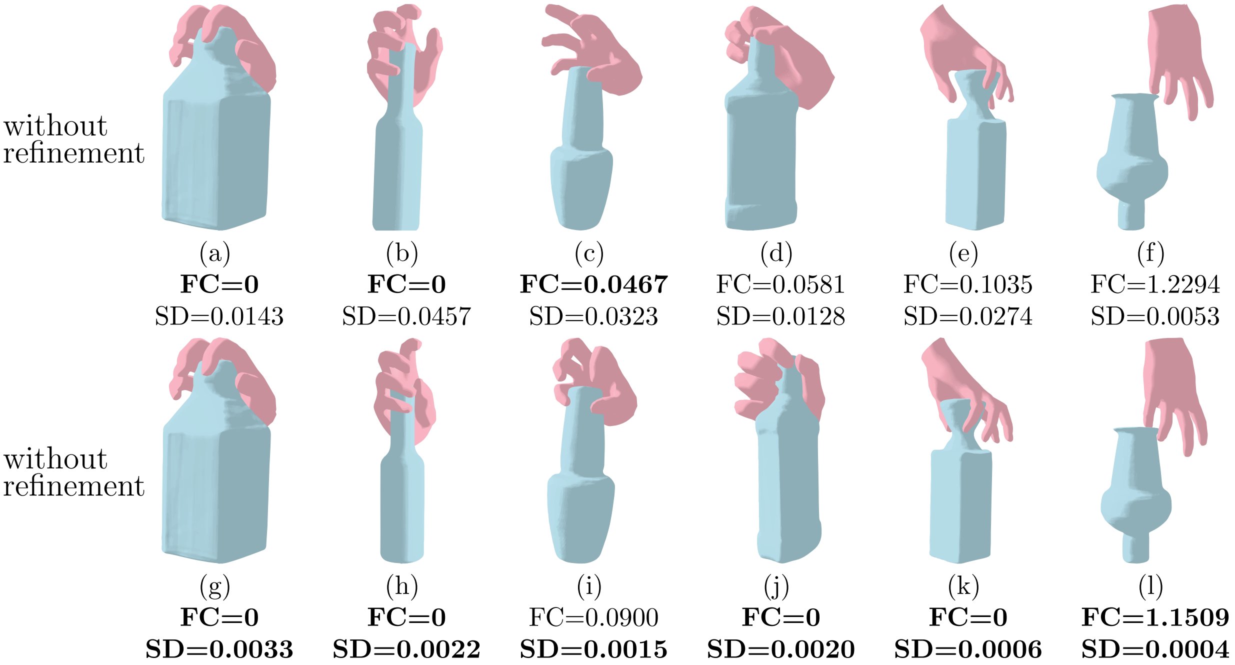
5.5.3 Force-closure Contact-point Generation

By directly minimizing the proposed force closure estimate, we can synthesize force closure contact points with *arbitrary* shapes. Specifically, we rewrite the solution of constraint in Eq. (5.4) as

$$\begin{aligned} x^* &= \arg\min_x FC(x, O), \\ FC(x, O) &= \lambda_0(GG' - \epsilon I_{6\times 6}) + ||Gc||_2 + w \sum_{x_i \in x} d(x_i, O), \end{aligned} \tag{5.11}$$

where G is defined in Eq. (5.2), and $c = \{c_i\}$, where c_i is the surface normal of object O at point x_i. $\lambda_0(\cdot)$ gives the smallest eigenvalue. $d(x, O)$ returns the distance from point x to the surface of object O. w is a scalar that controls the weight for the distance between contact points and object. By minimizing the three terms, we are looking for $\{x_i\}$ that satisfies the constraints in Eqs. (5.4a) to (5.4c), respectively.

We run gradient descent on contact point positions to minimize $FC(x, O)$; the computed contact points on a unit sphere and some daily objects are shown in Fig. 5.6. Despite our assumptions, minimizing our force closure estimate can properly produce force closure contact points.



(a) **FC=0** SD=0.0143

(b) **FC=0** SD=0.0457

(c) **FC=0.0467** SD=0.0323

(d) FC=0.0581 SD=0.0128

(e) FC=0.1035 SD=0.0274

(f) FC=1.2294 SD=0.0053

(g) **FC=0** **SD=0.0033**

(h) **FC=0** **SD=0.0022**

(i) FC=0.0900 **SD=0.0015**

(j) **FC=0** **SD=0.0020**

(k) **FC=0** **SD=0.0006**

(l) **FC=1.1509** **SD=0.0004**

Figure 5.5: **Examples of synthesized grasps.** Top: synthesized grasps before refinement. Bottom: the same set of synthesized grasps after refinement. FC: estimated force closure error. SD: mean distance from each contact point to the object surface. Left to right: examples with zero FC error, small FC error, and high FC error qualitatively illustrate how our estimation of force closure correlates to grasp quality.

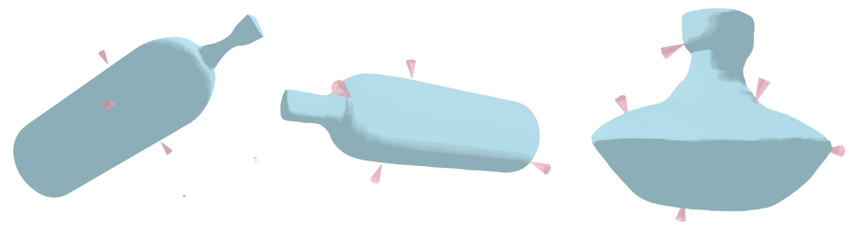


Figure 5.6: **Force-closure contact-point generations on unit spheres (top) and daily objects (bottom) by minimizing Eq. (5.11).** Objects in each columns have 3, 4, and 5 contact points, respectively.

5.5.4 Grasp Synthesis

We test our grasp synthesis algorithm on various bottle shapes retrieved from the ShapeNet dataset [CFG15]. Given the pre-trained DeepSDF model of an object, we randomly initialize a MANO hand and use Algorithm 2 to sample the hand configuration as well as contact points from $P(H|O)$. We set the step size $\eta = 0.1$, switch probability $\rho = 0.85$, distance weight $w = 1$, temperature $T = 0.1$, and Langevin steps $L = 10^6$. We filter out samples trapped in bad local minima by keeping samples that satisfy the constraint:

$$||Gc||_2 < 0.5 \tag{5.12a}$$

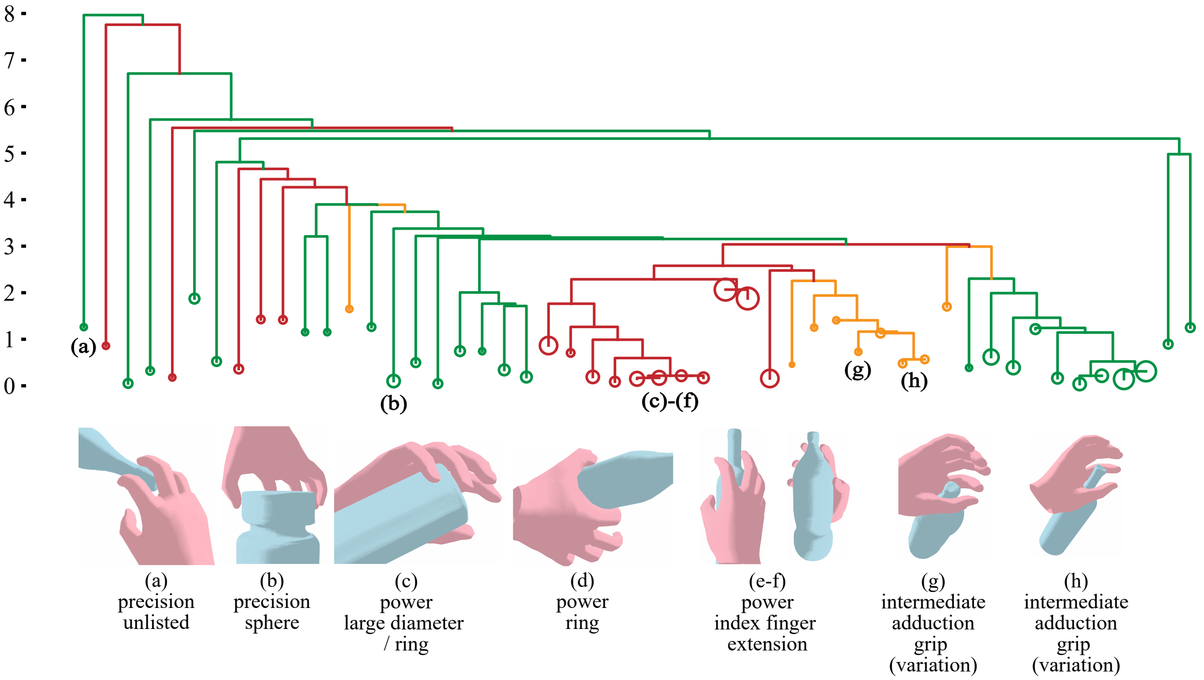
$$\sum_{x_i \in x} d(x_i, O)^2 < 0.02 \tag{5.12b}$$

$$E_{pen}(H, O) < 0.02 \tag{5.12c}$$

where x is the set of contact points on the hand surface, and c is the friction cone axes at contact points. Fig. 5.5 shows a collection of synthesis results with and without the refinement step: Higher values of our force closure estimation corresponds to non-grasps, whereas force closure estimation closed to zero is as good as the ones with force closure estimation equal to zero. **This observation confirms our previous analysis.** We also notice cases when the synthesis is trapped in bad local minima; these examples exhibit large values in our force closure estimation. We show two examples in the last column of Fig. 5.5. Such errors happened since the optimization problem is highly non-convex; one cannot avoid every bad minimum with gradient-based methods. Fortunately, we can identify these examples by their high force closure scores.

5.5.5 Physical Stability

We verify the physical stability of our synthesized examples by simulating the samples in PyBullet. Specifically, we set gravity to be $[0, 0, -10]$; an example is deemed to be a suc-



(a)
precision
unlisted

(b)
precision
sphere

(c)
power
large diameter
/ ring

(d)
power
ring

(e-f)
power
index finger
extension

(g)
intermediate
adduction
grip
(variation)

(h)
intermediate
adduction
grip
(variation)

Figure 5.7: **Energy landscape mapping generated by the ADELM algorithm [HNZ19] (best viewed in color).** Top: disconnectivity diagram of the energy landscape of our energy function $E(H, O)$. Green minima denote precision grasps, red power grasps, and yellow intermediate grasps. Bottom: examples from selected local minima; minima with lower energy barriers in between have similar grasps. We also label the grasp taxonomy of each example according to [FRS15]. Examples marked as *unlisted* do not belong to any manually classified type.

cessful grasp if the object's vertical drop is less than 0.3 after 1000 steps of simulation. Notice that a grasp's physical stability depends not only on the force closure score of the contact points, but also on whether the contact points are close enough to the object surface. We set two different thresholds on the contact point distance; Table 5.1 tabulates detailed comparisons of the success rate between our method against state-of-the-art algorithms [PAD10, ORG19]. To the best of our knowledge, [PAD10] is the state-of-the-art analytic approach, whereas [ORG19] is the state-of-the-art data-driven approach. Of note, although [ORG19] reported 95% success rate in the original paper, many of the objects being tested have simple shapes, such as a sphere or a box; the success rate would drop to 85% when we remove these simple objects. Additionally, neither of the two state-of-the-art methods has demonstrated the ability to synthesize diverse types of grasps. Although some other data-driven methods have demonstrated a certain level of diverse grasp synthesis, they fail to report their physical stability as it is not their primary focus.

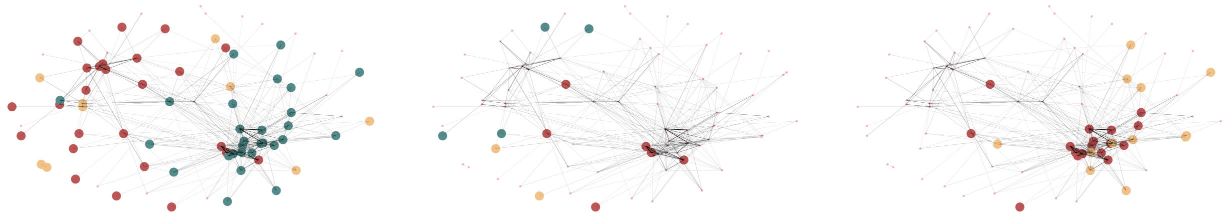
Table 5.1: Grasp success rates

method	**success rate**
Unions of Balls [PAD10]	72.53%
Visuo-Haptic [ORG19]	**85.00%**
Ours ($\sigma < 0.0015$)	76.98%
Ours (σ <0.0005)	**85.00%**

5.5.6 Diversity of the Grasp Types

To evaluate the grasp synthesis's diversity generated by the proposed method, we examine the energy landscape of our grasp energy function. Below, we show that the distribution of grasps defined by our energy function loosely aligns with the carefully organized grasps taxonomy [FRS15] when applied to humanoid hands. We use the ADELM algorithm [HNZ19] to build the energy landscape mapping of our grasps energy function $E(H, O)$.

Specifically, we collected 371 synthesized grasp examples and adopted the ADELM algorithm [HNZ19] to find minimum energy pathways (MEPs) between them. We project the MEPs between examples to a disconnectivity graph in Fig. 5.7. In the disconnectivity graph, each circle at the bottom represents a local minima group. The size of the circle indicates how many synthesized examples fall into this group. The height of the horizontal bar between two groups represent the maximum energy (or energy barrier) along the MEPs between two groups. The MEPs with lowest barriers connect smaller groups into larger groups, and this process is repeated until all examples are connected. The produced disconnectivity graph is an estimation of the true landscape of the energy function. Energy landscape mapping in Fig. 5.7 shows that the local minima with low energy barriers between them have similar grasps, and those with high energy barriers between them tend to have different grasps. We also observe that the energy landscape contains all three categories in the power/precision



(a) Red: power grasps. Yellow: intermediate grasps. Green: Precision grasps. Other: Unlisted.

(b) Red: power sphere grasps. Yellow: power disk grasp. Green: power cylinder grasps (large diameter, medium wrap, small diameter)

(c) Red: power sphere and precision sphere grasps. Yellow: tri- and quad- pod grasps.

Figure 5.8: Alignment between our energy landscape and existing grasp taxonomy [FRS15]. Best viewed in color.

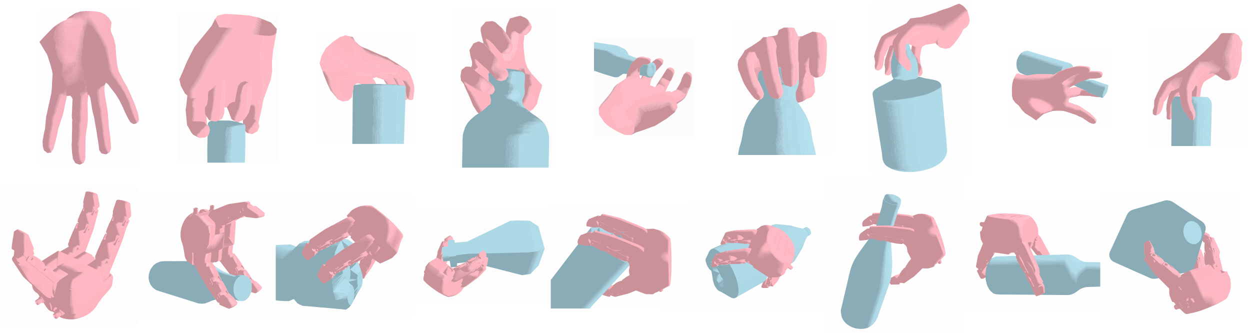


Figure 5.9: **Synthesized grasps of different hands using our formulation.** Top: A MANO hand with its thumb removed. Bottom: A Robotiq 3-finger gripper. The left-most figure shows the hand used in each row.

dimension as described in [FRS15].

To provide a more comprehensive understanding of the alignment between our energy landscape and the existing taxonomy, we further plot the local minima groups as a 2D graph in Fig. 5.8, which supplements the 1D energy landscape shown in Fig. 5.7. In Fig. 5.8, each node represents a local minima group. The edges between nodes denote the energy barriers between the minima groups they connect: Thicker edges indicate lower energy barriers and therefore closer minima groups, and no edge between two nodes means no pathway has been found between the two groups. Nodes with lower barriers between them are placed closer to each other.

Fig. 5.8a shows that the power grasps and precision grasps are mostly separated from

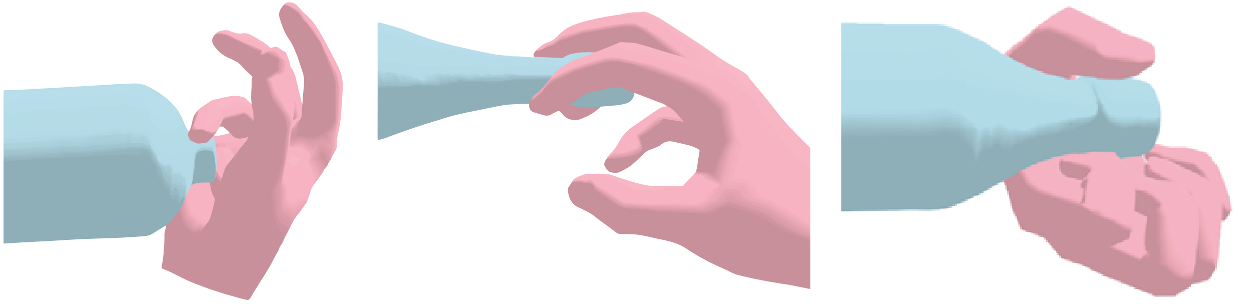


Figure 5.10: Examples of novel grasp poses that, to the best of our knowledge, are not included in any grasp taxonomy.

each other, indicating a high energy barrier between the two. One interpretation is that there is no smooth transition between a power grasp and a precision grasp without a non-force-closure grasp along with the transition. Intermediate grasps are scattered around. Nodes that are not colored are grasp types not listed in any existing grasp taxonomy, indicating the manually-defined grasp taxonomy, though carefully collected and designed, still falls short when facing a large variety of grasps.

In Fig. 5.8b, we draw different types of power grasps in different colors. Only the power grasps close to the precision grasps belong to the power sphere type. This observation matches our intuition as a power sphere grasp is similar to a precision sphere grasp, with a slight difference in the distance between the object and the palm. In other words, there exists a smooth transition between a precision sphere grasp and a power sphere grasp such that all snapshots along the transition are force-closure grasps. Please refer to [FRS15] for more details on power and precision sphere grasp.

In Fig. 5.8c, we observe that sphere grasps and tri- or quad-pod grasps are close to each other. This observation is also expected since many sphere grasps can be converted to tri- or quad-pod grasps by merely lifting one or two fingers.

We further demonstrate that our algorithm can find natural but novel and stable grasps in Fig. 5.10. These grasps are rarely collected in any of the modern 3D grasp datasets (*e.g.*, [TGB20, CPA20]), since they do not belong to any type as defined in the grasp taxonomy. However, these grasps are valid grasps and could well exist in physical interactions. For example, the left example in Fig. 5.10 is commonly used when one needs to twist-open a bottle when some of your fingers are occupied or injured. The second example would occur if one is already holding something in the palm while picking up another bottle. These grasp poses happen because the human hand is excellent in doing multiple tasks simultaneously; they have not been well recognized in the grasp literature as we always assumed otherwise. Such limitation would hinder a robotic hand's capacity from developing to its full potential. Our method provides possibilities to explore grasps in different types beyond the grasp taxonomy, which is a crucial step toward exploiting the total capacity of a complex hand structure such as human hands.

5.5.7 Grasp Synthesis for Arbitrary Hand Structures

Although the above experiments primarily rely on MANO for hand modeling and grasp taxonomy, our method in fact makes no assumption on the hand kinematics except for having a differentiable mapping between pose and shape. As a result, we can synthesize grasps for arbitrary hand so long as there exists such a mapping. In Fig. 5.9, our method, without modifications, can synthesize grasps of a MANO hand with its thumb removed and a Robotiq 3-finger gripper. Specifically, for the 3-finger gripper, we used a differentiable forward kinematics [SWL20] as the mapping from joint states to the hand shape. These examples demonstrate that our method can explore a wide range of grasps for arbitrary hand structure, which could provide valuable insights for understanding the task affordance of prosthetic or robotic hands, and hands with injuries or disabilities. Our method is also applicable to animations, wherein grasps of non-standard hands or claws are common.

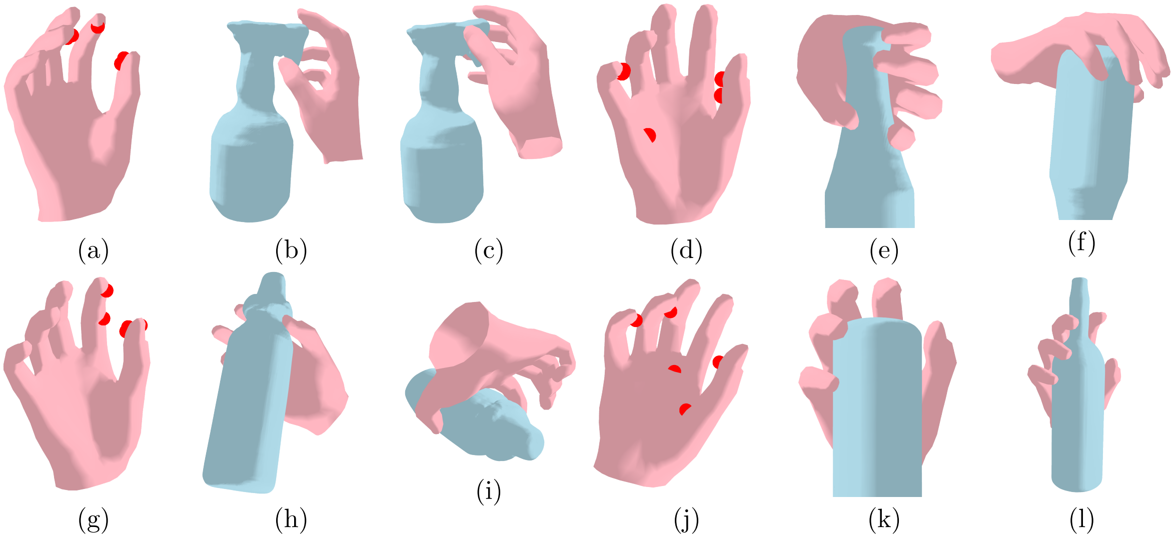


Figure 5.11: **Synthesizing specific types of grasping by enforcing contact points.** (a)(d)(g)(j) show the query contact points in red, each followed by two synthesized examples using the queried contact points. Grasp types can be determined by enforcing the choice of contact points on the hand surface.

5.5.8 Synthesizing Specific Grasp Type

As mentioned in Section 5.4.2, the choice of contact points on the hand surface primarily determines the grasp type. Hence, specific grasp types can be synthesized by mandating the choice of contact point; see examples in Fig. 5.11.

5.6 Conclusion

We formulated a fast and differentiable approximation of the force closure test computed within milliseconds, which enables a new grasp synthesis algorithm. In a series of experiments, we verified that our force closure estimation correctly reflects the quality of a grasp, and demonstrated the proposed grasp synthesis algorithm could generate diverse and physically stable grasps with arbitrary hand structures. The diversity of the generated grasps is validated by its alignment with widely accepted grasp taxonomy.

We believe that exploring different grasp types is crucial for future works of understanding the hand's total functional capacity, whether it is a prosthetic hand, a robotic hand, or an animated character's hand.

CHAPTER 6

Conclusion

This dissertation addresses two major missing parts in the study of 3D human-object interaction, namely, the lack of high-quality data and the lack of a hierarchical representation of 3D HOI.

We collected a large-scale synthetic dataset SHADE by exploiting the asset of a popular 3D video game. Our dataset contains rich and fine-grained 3D annotations of human-object interactions. In addition, the human pose in our dataset is a complement to existing human pose datasets, and the geometrical relationship in our dataset can be applied to real-life human behaviors. This dataset opens up new possibilities in modeling 3D human-object interactions. The SHADE dataset contains high-fidelity 3D shapes of the objects in HOI, which allowed us to address the second challenge by decomposing the object shape into functional parts.

We also propose a hierarchical modeling of human-object interaction and demonstrate that we can learn a more robust descriptive model by modeling HOI in a hierarchical fashion. We further demonstrated that we could obtain an explicitly derived descriptive model of grasping for arbitrary hand structure by decomposing shape-level geometrical relationships into physics-level.

We conclude this dissertation by observing that human activities are the results of fulfilling specific goals, where these goals can be written as physical or social constraints. In this dissertation, we explored the activity of sitting, where the goal is learned by an energy-based model, and the activity of grasping, where the goal is to resist arbitrary external wrench.

Some other examples include drinking, where the goal is to let water flow into the mouth, and group activity, where the goal is to maximize communication efficiency within a group. Once we understand these goals, understanding human behavior becomes trivial. However, it is a more challenging task to understand the goal behind observed human activities. It is an especially challenging task to have a unified representation to facilitate automatic goal discovery. Such a task may require joint understanding of object affordance and fluents.

REFERENCES

[AKH12] Heni Ben Amor, Oliver Kroemer, Ulrich Hillenbrand, Gerhard Neumann, and Jan Peters. "Generalization of human grasping for multi-fingered robot hands." In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2043–2050. IEEE, 2012.

[BFD15] Ian M Bullock, Thomas Feix, and Aaron M Dollar. "The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments." *The International Journal of Robotics Research*, **34**(3):251–255, 2015.

[BHH19] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. "ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact." *CoRR*, **abs/1904.03754**, 2019.

[BHH20] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. "ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact." In *ECCV*, 2020.

[BKK20] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. "Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects." In *CVPR*, 2020.

[BMA13] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. "Data-driven grasp synthesis—a survey." *IEEE Transactions on Robotics*, **30**(2):289–309, 2013.

[BW07] Stephen P Boyd and Ben Wegbreit. "Fast computation of optimal contact forces." *T-RO*, **23**(6):1117–1132, 2007.

[CA09] Matei T Ciocarlie and Peter K Allen. "Hand posture subspaces for dexterous robotic grasping." *The International Journal of Robotics Research*, **28**(7):851–867, 2009.

[CAL16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation." In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

[CFG15] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. "Shapenet: An information-rich 3d model repository." *arXiv preprint arXiv:1512.03012*, 2015.

[CLL18] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. "Learning to Detect Human-Object Interactions." In *WACV*, 2018.

[CPA20] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. “Ganhand: Predicting human grasp affordances in multi-object scenes.” In *CVPR*, 2020.

[CWH15] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. “HICO: A Benchmark for Recognizing Human-Object Interactions in Images.” In *ICCV*, 2015.

[DMT18] Hongkai Dai, Anirudha Majumdar, and Russ Tedrake. “Synthesis and optimization of force closure grasps via sequential semidefinite programming.” In *Robotics Research*, pp. 285–305. Springer, 2018.

[FC92] Carlo Ferrari and John F Canny. “Planning optimal grasps.” In *ICRA*, volume 3, pp. 2290–2295, 1992.

[FRS15] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. “The grasp taxonomy of human grasp types.” *IEEE Transactions on human-machine systems*, **46**(1):66–77, 2015.

[FXW18] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, and Song-Chun Zhu. “Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation.” In *AAAI*, 2018.

[GGD18] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. “Detecting and recognizing human-object interactions.” In *CVPR*, 2018.

[GLZ18] Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. “Learning Generative ConvNets via Multi-grid Modeling and Sampling.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2018.

[GMM07] Ulf Grenander, Michael I Miller, Michael Miller, et al. *Pattern theory: from representation to inference.* Oxford university press, 2007.

[GPM14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets.” In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[GTT21] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. “ContactOpt: Optimizing Contact to Improve Grasps.” In *CVPR*, 2021.

[GYB18] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. “First-person hand action benchmark with RGB-D videos and 3D hand pose annotations.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–419, 2018.

[HAW07] Guido Heumer, Heni Ben Amor, Matthias Weber, and Bernhard Jung. "Grasp recognition with uncalibrated data gloves-a comparison of classification methods." In *2007 IEEE Virtual Reality Conference*, pp. 19–26. IEEE, 2007.

[HBE17] Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. "SCC: Semantic Context Cascade for Efficient Action Detection." In *CVPR*, pp. 3175–3184, 2017.

[HCT19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. "Resolving 3D Human Pose Ambiguities with 3D Scene Constraints." In *Proceedings International Conference on Computer Vision*, pp. 2282–2292. IEEE, October 2019.

[HGD17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN." In *ICCV*, pp. 2980–2988, 2017.

[HGT21] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. "Populating 3D Scenes by Learning Human-Scene Interaction." In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[HN98] Jerry L Hintze and Ray D Nelson. "Violin plots: a box plot-density trace synergism." *The American Statistician*, **52**(2):181–184, 1998.

[HNZ19] Mitch Hill, Erik Nijkamp, and Song-Chun Zhu. "Building a telescope to look into high-dimensional image spaces." *Quarterly of Applied Mathematics*, **77**(2):269–321, 2019.

[HQZ18] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. "Holistic 3D Scene Parsing and Reconstruction from a Single RGB Image." In *CVPR*, 2018.

[HTL00] Li Han, Jeffrey C Trinkle, and Zexiang X Li. "Grasp analysis as linear matrix inequality problems." *IEEE Transactions on Robotics and Automation*, **16**(6):663–674, 2000.

[HZL15] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. "Jointly learning heterogeneous features for RGB-D activity recognition." In *CVPR*, pp. 5344–5352, 2015.

[IPO14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." *IEEE TPAMI*, **36**(7):1325–1339, 2014.

[ISS17] Hamid Izadinia, Qi Shan, and Steven M Seitz. "Im2cad." In *CVPR*, pp. 2422–2431, 2017.

[JBB11] Matthew S Johannes, John D Bigelow, James M Burck, Stuart D Harshbarger, Matthew V Kozlowski, and Thomas Van Doren. "An overview of the developmental process for the modular prosthetic limb." *Johns Hopkins APL Technical Digest*, **30**(3):207–216, 2011.

[JXY13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. "3D convolutional neural networks for human action recognition." *IEEE TPAMI*, **35**(1):221–231, 2013.

[KB14] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[KGS13] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. "Learning human activities and object affordances from RGB-D videos." *The International Journal of Robotics Research*, 2013.

[KLG18] Keizo Kato, Yin Li, and Abhinav Gupta. "Compositional learning for human object interaction." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 234–251, 2018.

[KRK11] Hedvig Kjellström, Javier Romero, and Danica Kragić. "Visual object-action recognition: Inferring object affordances from human demonstration." *Computer Vision and Image Understanding*, **115**(1):81–90, 2011.

[KT15] Vikash Kumar and Emanuel Todorov. "Mujoco haptix: A virtual reality system for hand manipulation." In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 657–663. IEEE, 2015.

[KTS14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks." In *CVPR*, pp. 1725–1732, 2014.

[KW13] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*, 2013.

[KYZ20] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Micheal Black, Siyu Tang, and Krikamol Muandet. "Grasping Field: Learning Implicit Representations for Human Grasps." In *3DV*, 2020.

[LBM17] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. "Roomnet: End-to-end room layout estimation." In *ICCV*, pp. 4875–4884, 2017.

[Liu99] Yun-Hui Liu. "Qualitative test and force optimization of 3-D frictional form-closure grasps using linear programming." *IEEE Transactions on Robotics and Automation*, **15**(1):163–173, 1999.

[LLC03] Jia-Wei Li, Hong Liu, and He-Gao Cai. "On computing three-finger force-closure grasps of 2-D and 3-D objects." *IEEE Transactions on Robotics and Automation*, **19**(1):155–161, 2003.

[LLS15] Ian Lenz, Honglak Lee, and Ashutosh Saxena. "Deep learning for detecting robotic grasps." *The International Journal of Robotics Research*, **34**(4-5):705–724, 2015.

[LPK18] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection." *The International Journal of Robotics Research*, **37**(4-5):421–436, 2018.

[LPS16] Yangyan Li, Soeren Pirk, Hao Su, Charles R Qi, and Leonidas J Guibas. "Fpnn: Field probing neural networks for 3d data." In *Advances in Neural Information Processing Systems*, pp. 307–315, 2016.

[LS14] Yun Lin and Yu Sun. "Grasp planning based on strategy extracted from demonstration." In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4458–4463. IEEE, 2014.

[LS15] Yun Lin and Yu Sun. "Robot grasp planning based on demonstrated grasp strategies." *The International Journal of Robotics Research*, **34**(1):26–42, 2015.

[LSL18] Cewu Lu, Hao Su, Yonglu Li, Yongyi Lu, Li Yi, Chi-Keung Tang, and Leonidas J Guibas. "Beyond holistic object recognition: Enriching image understanding with part states." In *CVPR*, 2018.

[LXL20] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. "PaStaNet: Toward Human Activity Knowledge Engine." In *CVPR*, 2020.

[LXM17] Hangxin Liu, Xu Xie, Matt Millar, Mark Edmonds, Feng Gao, Yixin Zhu, Veronica J Santos, Brandon Rothrock, and Song-Chun Zhu. "A glove-based system for studying hand-object manipulation via joint pose and force sensing." In *IROS*, pp. 6617–6624, 2017.

[LZH19] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. "Transferable interactiveness knowledge for human-object interaction detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3585–3594, 2019.

[LZW15] Yang Lu, Song-Chun Zhu, and Ying Nian Wu. "Learning frame models using cnn filters." *arXiv preprint arXiv:1509.08379*, 2015.

[MA04] Andrew T Miller and Peter K Allen. "Graspit! a versatile simulator for robotic grasping." 2004.

[MHR17] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. "A simple yet effective baseline for 3d human pose estimation." In *ICCV*, pp. 2659–2668, 2017.

[ML16] Arun Mallya and Svetlana Lazebnik. "Learning models for actions and person-object interactions with transfer to question answering." In *ECCV*, 2016.

[MLN17] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics." *arXiv preprint arXiv:1703.09312*, 2017.

[MO14] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784*, 2014.

[Mur17] Richard M Murray. *A mathematical introduction to robotic manipulation.* CRC press, 2017.

[OBL15] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." In *CVPR*, pp. 685–694, 2015.

[ORG19] Simon Ottenhaus, Daniel Renninghoff, Raphael Grimm, Fabio Ferreira, and Tamim Asfour. "Visuo-haptic grasping of unknown objects based on gaussian process implicit surfaces and deep learning." In *International Conference on Humanoid Robots (Humanoids)*, 2019.

[PAD10] Markus Przybylski, Tamim Asfour, and Rüdiger Dillmann. "Unions of balls for shape approximation in robot grasping." In *IROS*, 2010.

[PFS19] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation." *arXiv preprint arXiv:1901.05103*, 2019.

[PMG12] Domenico Prattichizzo, Monica Malvezzi, Marco Gabiccini, and Antonio Bicchi. "On the manipulability ellipsoids of underactuated robotic hands with compliance." *Robotics and Autonomous Systems*, **60**(3):337–346, 2012.

[PSB93] Jean Ponce, Steve Sullivan, J-D Boissonnat, and J-P Merlet. "On characterizing and computing three-and four-finger force-closure grasps of polyhedral objects." In *ICRA*, 1993.

[PSS97] Jean Ponce, Steve Sullivan, Attawith Sudsang, Jean-Daniel Boissonnat, and Jean-Pierre Merlet. "On computing four-finger equilibrium and force-closure grasps of polyhedral objects." *IJRR*, **16**(1):11–35, 1997.

[PZD17] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose." In *CVPR*, pp. 1263–1272, 2017.

[QSM17] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. "Pointnet: Deep learning on point sets for 3d classification and segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.

[QWJ18] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. "Learning Human-Object Interactions by Graph Parsing Neural Networks." In *ECCV*, 2018.

[QYS17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." In *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017.

[RKK09] Javier Romero, Hedvig Kjellstrom, and Danica Kragic. "Modeling and evaluation of human-to-robot mapping of grasps." In *2009 International Conference on Advanced Robotics*, pp. 1–6. IEEE, 2009.

[RMF12] Alberto Rodriguez, Matthew T Mason, and Steve Ferry. "From caging to grasping." *The International Journal of Robotics Research*, **31**(7):886–900, 2012.

[Roc17] RockStar-Games. "Policy on posting copyrighted Rockstar Games material." 2017.

[ROG17] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. "Octnet: Learning deep 3d representations at high resolutions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3577–3586, 2017.

[RSG12] Carlos Rosales, Raúl Suárez, Marco Gabiccini, and Antonio Bicchi. "On the synthesis of feasible and prehensile robotic grasps." In *2012 IEEE International Conference on Robotics and Automation*, pp. 550–556. IEEE, 2012.

[RTB17] Javier Romero, Dimitrios Tzionas, and Michael J Black. "Embodied hands: Modeling and capturing hands and bodies together." *TOG*, **36**(6):1–17, 2017.

[RVR16] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. "Playing for data: Ground truth from computer games." In *ECCV*, pp. 102–118, 2016.

[SBB10] Leonid Sigal, Alexandru O Balan, and Michael J Black. "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." *IJCV*, **87**(1):4–27, 2010.

[SCH14] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. "SceneGrok: Inferring action maps in 3D environments." *ACM TOG*, **33**(6):212, 2014.

[SDN08] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. "Robotic grasping of novel objects using vision." *The International Journal of Robotics Research*, **27**(2):157–173, 2008.

[SEB12] Anis Sahbani, Sahar El-Khoury, and Philippe Bidaud. "An overview of 3D object grasp synthesis algorithms." *Robotics and Autonomous Systems*, **60**(3):326–336, 2012.

[SLN16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. "NTU RGB+D: A large scale dataset for 3D human activity analysis." In *CVPR*, pp. 1010–1019, 2016.

[SM18] Andreea I Sburlea and Gernot R Müller-Putz. "Exploring representations of human grasping in neural, muscle and kinematic signals." *Scientific reports*, **8**(1):16669, 2018.

[SVW16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. "Hollywood in homes: Crowdsourcing data collection for activity understanding." In *European Conference on Computer Vision*, pp. 510–526. Springer, 2016.

[SWL20] Giovanni Sutanto, Austin Wang, Yixin Lin, Mustafa Mukadam, Gaurav Sukhatme, Akshara Rai, and Franziska Meier. "Encoding physical constraints in differentiable newton-euler algorithm." In *Learning for Dynamics and Control*, 2020.

[SX16] Shuran Song and Jianxiong Xiao. "Deep sliding shapes for amodal 3d object detection in rgb-d images." In *CVPR*, pp. 808–816, 2016.

[SXL16] Pierre Sermanet, Kelvin Xu, and Sergey Levine. "Unsupervised perceptual rewards for imitation learning." *arXiv preprint arXiv:1612.06699*, 2016.

[SYZ17] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. "Semantic Scene Completion from a Single Depth Image." *CVPR*, 2017.

[SZ14] Karen Simonyan and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." In *NIPS*, pp. 568–576, 2014.

[SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." *arXiv preprint arXiv:1212.0402*, 2012.

[TGB20] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. "GRAB: A dataset of whole-body human grasping of objects." In *ECCV*, 2020.

[TGJ15] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. "Efficient object localization using convolutional networks." In *CVPR*, pp. 648–656, 2015.

[TMS17] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. "Learning to fuse 2d and 3d image cues for monocular body pose estimation." In *ICCV*, pp. 3961–3970, 2017.

[US00] Thomas Ullmann and Joerg Sauer. "Intuitive virtual grasping for non haptic environments." In *Proceedings the Eighth Pacific Conference on Computer Graphics and Applications*, pp. 373–457. IEEE, 2000.

[WW17] Hongsong Wang and Liang Wang. "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks." In *CVPR*, 2017.

[WYD20] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. "Learning human-object interaction detection using interaction points." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4116–4125, 2020.

[WZZ13] 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*, pp. 3272–3279, 2013.

[WZZ17] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. "Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization." *IEEE TPAMI*, **39**(6):1165–1179, 2017.

[XHZ15] Jianwen Xie, Wenze Hu, Song-Chun Zhu, and Ying Nian Wu. "Learning sparse FRAME models for natural image patterns." *International Journal of Computer Vision*, **114**(2-3):91–112, 2015.

[XLG18] Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. "Cooperative learning of energy-based model and latent variable model via mcmc teaching." In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[XLZ16a] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. "Inducing wavelets into random fields via generative boosting." *Applied and Computational Harmonic Analysis*, **41**(1):4–25, 2016.

[XLZ16b] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. "A theory of generative convnet." In *International Conference on Machine Learning*, pp. 2635–2644, 2016.

[XZG18] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. "Learning descriptor networks for 3d shape synthesis and analysis." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8629–8638, 2018.

[XZN17] Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. "Synthesizing dynamic patterns by spatial-temporal generative convnet." In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 7093–7101, 2017.

[YF10] Bangpeng Yao and Li Fei-Fei. "Modeling mutual context of object and human pose in human-object interaction activities." In *CVPR*, 2010.

[YXL18] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition." In *AAAI*, 2018.

[ZBL18] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. "On the Continuity of Rotation Representations in Neural Networks." *arXiv preprint arXiv:1812.07035*, 2018.

[ZC09] Yu Zheng and Chee-Meng Chew. "Distance Between a Point and a Convex Cone in n-Dimensional Space: Computation and Applications." *T-RO*, **25**(6):1397–1412, 2009.

[ZHN20] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. "Generating 3D People in Scenes without People." In *Computer Vision and Pattern Recognition (CVPR)*, pp. 6194–6204, June 2020.

[ZHS17] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. "Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach." In *ICCV*, pp. 398–407, 2017.

[ZLS17] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. "Towards Context-Aware Interaction Recognition for Visual Relationship Detection." In *ICCV*, pp. 589–598, 2017.

[ZWK14] Xi Zhao, He Wang, and Taku Komura. "Indexing 3d scenes using the interaction bisector surface." *ACM Transactions on Graphics (TOG)*, **33**(3):22, 2014.

[ZWM98] Song Chun Zhu, Yingnian Wu, and David Mumford. “Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling.” *International Journal of Computer Vision*, **27**(2):107–126, 1998.

[ZZM20] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. “PLACE: Proximity learning of articulation and contact in 3D environments.” In *8th international conference on 3D Vision (3DV 2020)(virtual)*, 2020.