

UCLA

UCLA Electronic Theses and Dissertations

Title

Visual Commonsense Reasoning: Functionality, Physics, Causality, and Utility

Permalink

<https://escholarship.org/uc/item/7sm0389z>

Author

Zhu, Yixin

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Visual Commonsense Reasoning:
Functionality, Physics, Causality, and Utility

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

by

Yixin Zhu

2018

© Copyright by

Yixin Zhu

2018

ABSTRACT OF THE DISSERTATION

Visual Commonsense Reasoning:
Functionality, Physics, Causality, and Utility

by

Yixin Zhu

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2018

Professor Song-Chun Zhu, Chair

Reasoning about commonsense from visual input remains an important and challenging problem in the field of computer vision. It is important because the ability to reason about commonsense, plan and act accordingly, represents the most distinct competence that tells human apart from other animals—the ability of analogy. It is challenging partially due to the absence of the observations of all the typical examples in a given category, in which the objects often present enormous intra-class variations, leading to a long-tail distribution in the dimensions of appearance and geometry. This dissertation focuses on four largely orthogonal dimensions—functionality, physics, causality, and utility—in computer vision, robotics, and cognitive science, and it makes six major contributions:

1. We rethink object recognition from the perspective of an agent: how objects are used as “tools” or “containers” in actions to accomplish a “task”. Here a task is defined as changing the physical states of a target object by actions, such as, cracking a nut or painting a wall. A tool is a physical object used in the human action to achieve the task, such as a hammer or a brush, and it can be any daily objects which are not restricted to conventional hardware tools. This leads us to a new framework—task-oriented object modeling, learning and recognition, which aims at understanding the underlying functions, physics and causality in using objects as tools in various task categories.
2. We propose to go beyond visible *geometric compatibility* to infer, through physics-based

simulation, the forces/pressures on various body parts as people interact with objects. By observing people’s choices in videos, we can learn the *comfort intervals* of the pressures on body parts as well as human preferences in distributing these pressures among body parts. Thus, our system is able to “feel”, in numerical terms, discomfort when the forces/pressures on body parts exceed comfort intervals. We argue that this is an important step in representing *human utilities*—the pleasure and satisfaction defined in economics and ethics (*e.g.*, by the philosopher Jeremy Bentham) that drives human activities at all levels.

3. We propose to go beyond modeling the *direct* and *short-term* human interaction with individual objects. Through accurately simulating thermodynamics and air fluid dynamics, our method can infer indoor room temperature distribution and air flow dynamics at arbitrary time and locations, thus establishing a form of *indirect* and *long-term* affordance. Unlike chairs in a sitting scenario, the objects (heating/cooling sources) that provide affordance do not directly interact with a person. Instead, the air in a room serves as an *invisible* medium to pass the affordance from an object to a person. We coin this new form of affordance as *intangible affordance*.
4. By fusing functionality and affordance into indoor scene generation, we propose a systematic learning-based approach to the generation of massive quantities of synthetic 3D scenes and numerous photorealistic 2D images thereof, with associated ground truth information, for the purposes of training, benchmarking, and diagnosing learning-based computer vision and robotics algorithms.
5. We present four case studies on integrating forces and functionality in object manipulations in the field of robotics, showcasing the significance and benefits of explicit modeling of the functionality in task executions.
6. We introduce an intuitive substance engine (ISE) model employing probabilistic simulation, which supports the hypothesis that humans infer future states of perceived physical situations by propagating noisy representations forward in time using approximated rational physics.

The dissertation of Yixin Zhu is approved.

Hongjing Lu

Demetri Terzopoulos

Ying Nian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2018

To my parents and Xingwen

TABLE OF CONTENTS

1	Introduction	1
I	Understanding Object and Scene by Reasoning Functionality and Physics	4
2	Understanding Tools: Task-Oriented Objects	5
2.1	Task-oriented Object Representation	8
2.1.1	Tool in 3D Space	9
2.1.2	Tool-use in Time	10
2.1.3	Physical Concept and Causality	10
2.2	Problem Definition	12
2.2.1	Learning Physical Concept	12
2.2.2	Recognizing Tools by Imagining Tool-uses	14
2.2.3	Parsing Human Demonstration	15
2.3	Experiments	17
2.3.1	Dataset	17
2.3.2	Learning Physical Concept	17
2.3.3	Inferring Tools and Tool-uses	21
2.4	Discussions	23
2.4.1	Related Work	24
2.4.2	Limitation and Future Work	26
3	Inferring Forces and Learning Human Utilities	27
3.1	Related Work	28

3.2	Representation	32
3.2.1	Spatial Entities and Relations in 3D Spaces	32
3.2.2	Physical Quantities of Human Utilities	33
3.2.3	Human Utilities in Time	33
3.3	Estimating the Forces in 3D Scenes	34
3.3.1	Dataset of 3D Scenes and Human Models	34
3.3.2	Reconstructing Watertight Scenes	34
3.3.3	Modeling Volumetric Human Pose	35
3.3.4	Simulating Human Interactions With Scenes	36
3.4	Learning and Inferring Human Utilities	39
3.4.1	Extracting Features	39
3.4.2	Learning Human Utilities	40
3.4.3	Inferring the Optimal Affordance	43
3.4.4	Sampling the Solution Space	44
3.5	Experiments	45
3.5.1	Learning Human Utilities From Demos	45
3.5.2	Inferring Optimal Affordance in Static Scenes	45
3.6	Discussion and Future Work	47
4	Learning Intangible Affordance	49
4.1	Related Work	50
4.2	Representation	52
4.3	Estimating Temperature and Velocity Field	53
4.3.1	Reconstruction of Volumetric Scene	53
4.3.2	Simulating Thermodynamic Air Flow	55

4.4	Learning Intangible Affordance	57
4.4.1	Learning Grammar of Human Activity	57
4.4.2	Extracting Features	61
4.4.3	Learning and Inference	63
4.4.4	Sampling Solution Spaces	64
4.5	Experiment	65
4.5.1	Learning Intangible Affordance	65
4.5.2	Inferring Optimal Affordance in Static Scenes	66
4.5.3	Evaluations	67
4.6	Discussion and Future Work	68
5	Inferring Containers and Containment Relations	69
5.1	What is Where: Inferring Containment Relations from Videos	69
5.1.1	Related Work	72
5.1.2	Problem Definition	74
5.1.3	Problem Formulation	75
5.1.4	Inference by Dynamic Programming	79
5.1.5	Experiments	80
5.1.6	Conclusion	85
5.2	Tracking Occluded Objects and Recovering Incomplete Trajectories by Reasoning about Containment Relations and Human Actions	85
5.2.1	Related Work	88
5.2.2	Probabilistic Formulation	89
5.2.3	Experiments	95
5.2.4	Conclusions and Discussions	103

6	Scene Synthesis by Integrating Functionality and Affordance	105
6.1	Related Work	107
6.2	Representation and Formulation	110
6.2.1	Representation: Attributed Spatial And-Or Graph	110
6.2.2	Probabilistic Formulation	113
6.3	Learning, Sampling and Synthesis	118
6.3.1	Learning the S-AOG	119
6.3.2	Sampling Scene Geometry Configurations	122
6.3.3	Scene Instantiation using 3D Object Datasets	126
6.3.4	Scene Attribute Configurations	126
6.4	Photorealistic Scene Rendering	128
6.5	Experiments	131
6.5.1	Normal Prediction	132
6.5.2	Depth Estimation	133
6.5.3	Benchmark and Diagnosis	135
6.6	Discussion	139
6.7	Appendix: Additional Results	141

II Integrating Forces and Functionality in Object Manipulations 146

7	Building Tactile Glove	149
7.1	A Glove-based System for Studying Hand-Object Manipulation via Joint Pose and Force Sensing	149
7.2	Unsupervised Learning of Hierarchical Models for Hand-Object Interactions .	150

8	Learning Object Manipulations by Integrating Forces and Functionality	154
8.1	Feeling the Force: Integrating Force and Pose for Fluent Discovery through Imitation Learning	154
8.2	Interactive Robot Knowledge Patching using Augmented Reality	157
III	Cognitive Studies	160
9	Intuitive Physics	161
9.1	Evaluating Human Cognition of Containing Relations with Physical Simulation	163
9.2	Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics	165
9.3	Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem	167
9.4	The Martian: Examining Human Physical Judgments Across Virtual Gravity Fields	169
9.5	Visuomotor Adaptation and Sensory Recalibration in Reversed Hand Movement Task	171
10	Causal Reasoning	172
10.1	Spatially Perturbed Collision Sounds Attenuate Perceived Causality in 3D Launching Events	173
10.2	Deep Reinforcement Learning Fails to Account for Human Causal Transfer .	176
11	Conclusion	179
	References	181

LIST OF FIGURES

2.1	An example of task-oriented object recognition problem	6
2.2	Inference of task-oriented objects	7
2.3	The task-oriented representation of tool-uses	9
2.4	Physical concepts involved in reasoning task-oriented objects	11
2.5	Pipeline of learning and inference of task-oriented object recognition	12
2.6	Spatial-temporal parsing of a human demonstration	16
2.7	Examples of tool instances in the dataset	18
2.8	Experiment: learning essential physical concepts of tool-use	19
2.9	Experiment: learning physical concept from a single human demonstration for cracking a nut	20
2.10	Experiment: recognizing tools for chopping wood	21
2.11	Experiment: comparison between human and algorithm prediction of tool- use for shoveling dirt	23
3.1	Examples of sitting activities in an office and a meeting room	28
3.2	An example of human utility: sitting behavior	31
3.3	Convert a stick-man model to a tetrahedralized human model	33
3.4	Reconstruction of watertight 3D scenes	35
3.5	Simulating human body with proper damping	37
3.6	Data pre-processing	39
3.7	Learning human utilities	41
3.8	Utilities of human sitting behavior	42
3.9	Experiment: ablative analysis on each component	44
3.10	Experiment: algorithm performance compared to human ratings	46

3.11	Experiment: top 3 poses in canonical, cluttered and novel scenarios	48
4.1	Examples of inferred intangible affordance	50
4.2	Reconstruction of volumetric scene	53
4.3	Thermodynamic air flow simulation	55
4.4	Clustered human skeletons	58
4.5	The grammar of human activity	60
4.6	Spatial segmentation of a 3D scene	62
4.7	Integrating additional planning cost	63
4.8	Sampling human activities and configurations of the heater/cooler	65
4.9	Human preferences learned from demonstrations	66
4.10	Experiment: inference results of intangible affordance	67
4.11	Experiment: inference by the proposed model compared with the human judgments	68
5.1	An example of inferring containment relations from video	70
5.2	An overview of the proposed approach	71
5.3	Temporal assumption	75
5.4	Containment relations in 3D space	76
5.5	Transition matrix of containment relation changes for an object	78
5.6	Experiment: probability of three different containment relation changes over time	80
5.7	Experiment: confusion matrix of relation change recognition	81
5.8	Experiment: inference of containment relations	82
5.9	Experiment: qualitative results	84
5.10	A scenario for tracking occluded objects in an indoor scene	86

5.11	The framework of the proposed method	87
5.12	Two causes of occlusions	90
5.13	Human action features at time t	92
5.14	Examples of occluded object tracking dataset	95
5.15	Experiment: transition probability of the object location	96
5.16	Confusion matrix of HOI	98
5.17	Experiment result	99
5.18	Experiment: different overlap ratios evaluated on different subsets	101
5.19	Experiment: more qualitative results	102
6.1	An example of automatically-generated 3D bedroom	106
6.2	Scene grammar as an attributed S-AOG	110
6.3	Attributed AOG	114
6.4	The learning-based pipeline for synthesizing images	118
6.5	Qualitative results in different types of scenes	124
6.6	Synthesis for different values of β	125
6.7	Various configurations of the synthesized scenes	127
6.8	Experiment: examples of normal estimation results predicted by the model trained with our synthetic data	132
6.9	Experiment: examples of depth estimation results predicted by the model trained with our synthetic data	134
6.10	Render an indoor scene as a video for SLAM	137
6.11	Experiment: benchmark on object detection	138
6.12	Additional results	142
6.13	Additional results (cont.)	143
6.14	Additional results (cont.)	144

6.15	Additional results (cont.)	145
7.1	Prototype of the tactile glove	149
7.2	An exmple of segmenting and parsing the noisy input of force and pose data in an unsupervised fashion	151
7.3	Unsupervised learning pipeline of hand-object motion recognition	151
8.1	The problem of opening a medicine bottle	155
8.2	System architecture	158
9.1	Two typical cases when a container fails to contain its contents	164
9.2	Sensor and interface of the 3D reconstruction	165
9.3	Illustration of experiment designs	170
10.1	Illustration of the experiment design	175
10.2	Common cause (CC) and common effect (CE) structures used in the present study	177

LIST OF TABLES

2.1	Experiment: accuracy of tool recognition	22
2.2	Experiment: errors of imagining tool-use for affordance / functional bases .	24
3.1	Physical simulation parameters	38
5.1	Experiment: accuracy of containment relation estimation	85
5.2	Experiment: racking accuracy of full model compared with three baselines .	100
5.3	Experiment: tracking accuracy on other datasets	103
6.1	Comparisons of rendering time vs quality	130
6.2	Experiment: Performance of normal estimation	131
6.3	Experiment: depth estimation performance	133
6.4	Experiment: benchmark on depth estimation	135
6.5	Experiment: benchmark on surface normal estimation	136

ACKNOWLEDGMENTS

I am fortunate enough to meet and collaborate with a collection of remarkable people:

My advisor Dr. Song-Chun Zhu, for his visionary guidance and mentorship, for showing me a broad picture of AI, computer vision and robotics research, for teaching me to do ground-breaking research, and for providing me many valuable opportunities in my career.

Dr. Demetri Terzopoulos, for his continuous support in various projects and papers we collaborated together, especially the days when he stayed overnight to revise our papers.

Dr. Hongjing Lu, for showing me another door of research—cognitive science, in particular for teaching me how to design impeccable cognitive experiments.

Dr. Ying Nian Wu, for the freedom to go where my heart and mind led me, for the patience to teach me any knowledge from scratch, and for clearing my puzzles when I lost in research.

Dr. Yibiao Zhao, certainly my peer advisor, for showing me how to actually do research and science, down to each line of code, as well as for the patience that I hope I had someday in the future.

Dr. Chenfanfu Jiang, my best friend and collaborator in the field of computer graphics, for many lessons in both the style and content of scientific research from the perspective of a different field.

Siyuan Qi, a genius of math, for showing me how to derive challenging formulations for unsolvable problems. It is always a pleasure to watch him make an impossible derivation possible.

Dr. Brandon Rothrock, a long-term collaborator and my “older brother” in engineering, for his meticulous planning and preparations during our collaborations, for his pursuit of perfection, and for all the hiking we had together when we both had some time in the weekend back then.

James Kubricht, the best scientific writer I ever know so far, for showing me how to address each important detail in cognitive science, and for teaching me how to write good

papers.

Feng Gao, Mark Edmonds, Hangxin Liu, Xu Xie, Chi Zhang, and Zhenliang Zhang, my peer “roommates” in the VCLA lab, for endless fun deadlines of papers and demos together. A special thank goes to Feng, who helps to ease my burden of maintaining the lab hardware and equipment.

Dr. Wei Liang, for all the container experiments and papers together, and for all the beer parties afterward.

Siyuan Huang, for your fearless in solving the most challenging scene understanding tasks.

Tianmin Shu, for all his brilliant ideas on social modeling, and for the fun moments when we traveled together and shared the same hotel room for various conferences.

Dr. Tao Gao, for teaching me intriguing and counter-intuitive facts in various fields, for inspiring Yibiao and I for our tool paper, and for his excellent ability of debate on any topics.

Dr. Craig Yu, Dr. Ping Wei, Jiajun Wu, and Dr. Peter Battaglia, for organizing and hosting FPIC workshops together with Yibiao, Tao and Fanfu in CVPR and CogSci conferences.

Zhi Han, my very first teacher in the field of computer vision, for bringing me into this fascinating field.

My appreciation further extends to the entire VCLA lab at UCLA Statistics Department.

Finally, those to whom this dissertation is dedicated—my parents and my girlfriend, Xingwen Guo.

Portions of this work were supported by DARPA XAI N66001-17-2-4029, ONR MURI N00014-16-1-2007, DARPA SIMPLEX N66001-15-C-4035, DARPA MSEE FA 8650-11-1-7149, ONR MURI N00014-10-1-0933, and NSF IIS-1423305.

VITA

2013–2018	Graduate Research Assistant, UCLA VCLA, Dr. Song-Chun Zhu
2017 Summer	Visiting Student, UPenn Computer Graphics, Dr. Chenfanfu Jiang
2017	Outstanding Reviewer, CVPR 2017
2015	Ph.D. Candidate in Statistics, UCLA
2013	M.S. in Computer Science, UCLA
2012 Summer	Research Intern, Harvard Medical School, Dr. Gil Alterovitz
2012	B.E. in Software Engineer, Xi'an Jiaotong University, China
2011 Summer	UCLA-CSST Research Program, Dr. Todd Millstein
2011	Google Scholarship
2011	Samsung Scholarship

PUBLICATIONS

★ denotes joint first authors

Human-centric Indoor Scene Synthesis using Stochastic Grammar. S. Qi, **Y. Zhu**, S. Huang, C. Jiang, S.-C. Zhu. *CVPR*, 2018.

Interactive Robot Knowledge Patching using Augmented Reality. H. Liu★, Y. Zhang★, W. Si, X. Xie, **Y. Zhu**, S.-C. Zhu. *ICRA*, 2018.

Unsupervised Learning of Hierarchical Models for Hand-Object Interactions using Tactile Glove. X. Xie★, H. Liu★, M. Edmonds, F. Gao, S. Qi, **Y. Zhu**, B. Rothrock, S.-C. Zhu. *ICRA*, 2018.

Spatially Perturbed Collision Sounds Attenuate Perceived Causality in 3D Launching Events. D. Wang★, J. Kubricht★, **Y. Zhu**★, W. Liang, S.-C. Zhu, C. Jiang, H. Lu. *IEEE VR*, 2018.

Tracking Occluded Objects and Recovering Incomplete Trajectories by Reasoning about Containment Relations and Human Actions. W. Liang, **Y. Zhu**, S.-C. Zhu. *AAAI*, 2018.

Learning Complex Functional Manipulations by Human Demonstration and Fluent Discovery. M. Edmonds*, F. Gao*, X. Xie, H. Liu, **Y. Zhu**, B. Rothrock, S.-C. Zhu. *IROS*, 2017.

A Glove-based System for Studying Hand-Object Manipulation via Pose and Force Sensing. H. Liu*, X. Xie*, M. Millar*, M. Edmonds, F. Gao, **Y. Zhu**, V. J. Santos, B. Rothrock, S.-C. Zhu. *IROS*, 2017.

Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics. J. Kubricht*, **Y. Zhu***, C. Jiang*, D. Terzopoulos, S.-C. Zhu, H. Lu. *CogSci*, 2017.

Visuomotor Adaptation and Sensory Recalibration in Reversed Hand Movement Task. J Lin*, **Y. Zhu***, J. Kubricht*, S.-C. Zhu, H. Lu. *CogSci*, 2017.

The Martian: Examining Human Physical Judgments Across Virtual Gravity Fields. T. Ye*, S. Qi*, J. Kubricht, **Y. Zhu**, H. Lu, S.-C. Zhu. *TVCG*, 2017.

What is Where: Inferring Containment Relations from Videos. W. Liang, Y. Zhao, **Y. Zhu**, S.-C. Zhu. *IJCAI*, 2016.

Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem. J. Kubricht*, C. Jiang*, **Y. Zhu***, S.-C. Zhu, D. Terzopoulos, H. Lu. *CogSci*, 2016.

Inferring Forces and Learning Human Utilities From Videos. **Y. Zhu***, C. Jiang*, Y. Zhao, D. Terzopoulos, S.-C. Zhu. *CVPR*, 2016.

Evaluating Human Cognition of Containing Relations with Physical Simulation. W. Liang, Y. Zhao, **Y. Zhu**, S.-C. Zhu. *CogSci*, 2015

Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition. **Y. Zhu***, Y. Zhao*, S.-C. Zhu. *CVPR*, 2015.

CHAPTER 1

Introduction

The goal of computer vision, as coined by Marr [Mar82], is to compute what is where by looking. This paradigm has guided the geometry-based approaches in the 1980s-1990s and the appearance-based methods in the past two decades. Although in certain tasks, *e.g.*, object recognition and detection, computer vision algorithms have achieved human-level performance, yet the current dominant methods have difficulties in understanding human needs, predicting human attentions and intentions, and further assisting human through a humanoid robot. We argue that such difficulties are mainly due to the lack of *visual commonsense reasoning*: an effective explicit representation of the knowledge as well as a suitable computational approach to reason about the unobservable factors. We believe that the current dominating end-to-end training with large dataset is not the ultimate solution for building an intelligent machine with the aforementioned capabilities; these challenges cannot be solved by the visible appearance or geometry alone, but require deeper understanding of the scenes, human actions, and interactions between humans and scenes.

To empower a machine with such capabilities, we must look for the *missing dimensions* that go beyond visual spectrum, which often requires to reason about the imperceptible entities. In images and videos, many entities (functional objects, liquid, granular material) and relations (causal-effects, physical supports, containment relations) are infeasible to detect purely by their appearances, yet they are pervasive and govern the visible entities that are visible for detection. The imperceptible entities not only include the invisible quantities, *e.g.*, forces exerted during interactions, liquid, and sand, but also include the relations among objects, *e.g.*, containment relations, heat transfer, and tool-uses.

In this dissertation, we explicitly study four dimensions: functionality, physics, causal-

ity and utility. By jointly and explicitly model these four dimensions, we hope to design reasoning systems, capable of jointly modeling the direct/short-term and indirect/long-term interactions. Such systems are demonstrated in the following three aspects:

- **Computer Vision:** reasoning about the hidden factors—functionality, physics, causality and utility—to go beyond the current geometry- and appearance-based paradigm. The majority of the computer vision community is working on deep learning to replace the feature engineering. Although it provides better performances, it fails to address challenges beyond observed data. In contrast, functionality, physics, causality and utility occurred during interactions are generally applicable to all categories of objects, scenes, actions and events, *i.e.*, transferable across datasets. These entities and relations are deeper and more invariant, than geometry and appearance—the dominating features used in visual recognition. In this way, we could go beyond example-based methods; instead, we can a) recognize an object by its essential purposes of use, b) recognize a scene by their functions to serve human activities, and c) recognize the containment relations by its relations and causal effects. We argue that it is the functionalities, physics, causality and utility of objects and scenes that decide their designs of geometry and appearance, and decide the planning of human actions and events. The goal of this dissertation is to develop methods that “understand” objects, scenes and actions, not merely classify them by memorizing typical examples. This is crucial for generalizing to novel examples in tests. Details are presented in Part I, and the published papers [ZZZ15, ZJZ16, LZZ16, JZQ17, LZZ18, QZH18].
- **Robotics:** integrating forces and functionality in object manipulations. Conventional learning from demonstration methods often only observe a teacher’s demonstration by watching, thus difficult to recover the important hidden and unobservable factors. By building a tactile glove, we are able to recover the visually latent force during human demonstrations, thereby enabling a Baxter robot to learn more complex manipulations. Details are summarized in Part II, and the published papers [LGS16, LXM17, EGX17, XLE18, LZS18].

- Cognitive Studies: building computational models that account for a) human vision in terms of intuitive physics, and b) causal reasoning mechanism. In this direction, we try to reverse engineer the human vision and reasoning system through psychological experiments. To study human cognition of complex phenomenon, by building intuitive physics models [BHT13] that achieves human-level cognition, we hope to leverage the state-of-the-art physics-based simulation to model the internal cognitive and reasoning mechanisms of human perception systems for complex phenomenon such as sand, liquid, heat, and gravity field in which the objects do not have a perceivable shapes. To study human causal reasoning capability in various tasks, by integrating virtual reality and robotics planning engine, we hopes to discover the differences between the current state-of-the-art methods and the human performance. Details are summarized in Part III, and the published papers [LZZ15, KJZ16, YQK17, LZK17, KZJ17, WKZ18].

This dissertation is intended to raise the awareness, in the computer vision community, of the missing dimensions and the potential benefits of integrating these dimensions to reason about the invisible entities and relations in scene understanding tasks, robotics tasks, and cognitive studies. Studying such invisible entities and relations in computer vision are crucial for filling the performance gaps in the recognition of objects, scenes, actions and events.

Part I

Understanding Object and Scene by Reasoning Functionality and Physics

CHAPTER 2

Understanding Tools: Task-Oriented Objects

In this chapter, we rethink object recognition from the perspective of an agent: how objects are used as “tools” in actions to accomplish a “task”. Here a task is defined as changing the physical states of a target object by actions, such as, cracking a nut or painting a wall. A tool is a physical object used in the human action to achieve the task, such as a hammer or a brush, and it can be any daily objects which are not restricted to conventional hardware tools. This leads us to a new framework—task-oriented object modeling, learning and recognition, which aims at understanding the underlying functions, physics and causality in using objects as tools in various task categories.

Figure 2.1 illustrates the two phases of this new framework. In the learning phase, our algorithm observes only one RGB-D video as an example, in which a rational human picks up one object, the hammer, among a number of candidates to accomplish the task. From this example, our algorithm reasons about the essential physical concepts in the task (*e.g.*, forces produced at the far end of the hammer), and thus learns the task-oriented model. In the inference phase, our algorithm is given a new set of daily objects (on the desk in Figure 2.1(b)), and makes the best choice available (the wooden leg) to accomplish the task.

From this new perspective, any objects can be viewed as a hammer or a shovel, and this generative representation allows computer vision algorithms to generalize object recognition to novel functions and situations by reasoning the physical mechanisms in various tasks, which goes beyond memorizing typical examples for each object category as the prevailing appearance-based recognition methods do in the literature.

Figure 2.2 shows some typical results in our experiments to illustrate this new task-oriented object recognition framework. Given three tasks: chop wood, shovel dirt, and paint

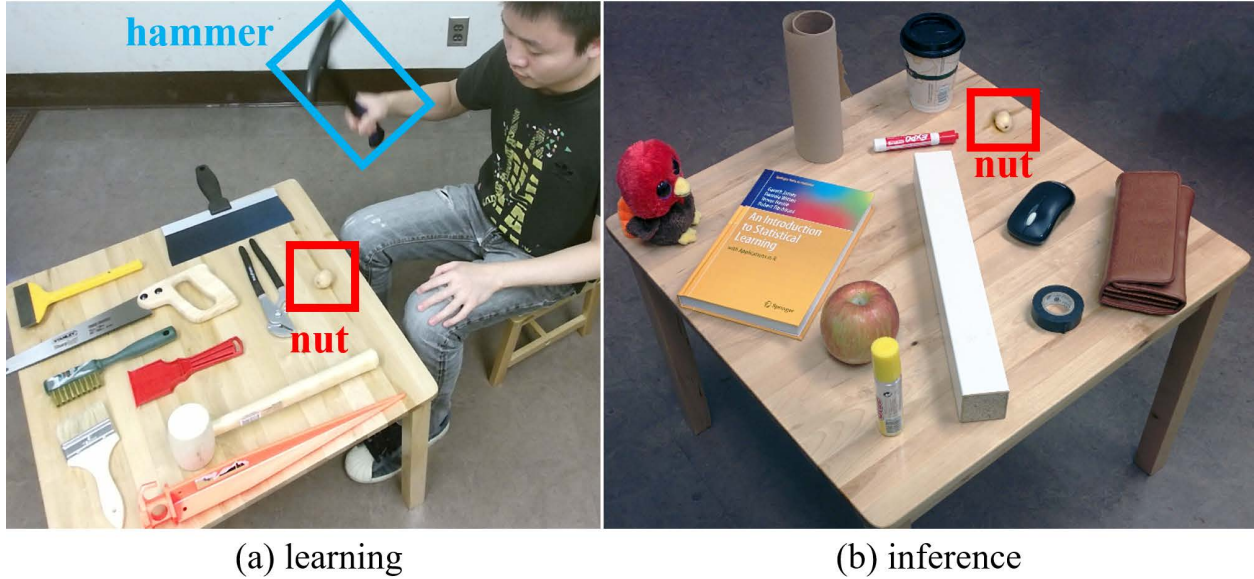


Figure 2.1: Task-oriented object recognition. (a) In the learning phase, a rational human is observed picking up a hammer among other tools to crack a nut. (b) In the inference phase, the algorithm is asked to choose the best object (*i.e.*, the wooden leg) on the table for the same task. This generalization entails physical reasoning.

wall, and three groups of objects: conventional tools, household objects, and stones, our algorithm ranks the objects in each group for a given task. Figure 2.2 shows the top two choices together with imagined actions using such objects for the tasks.

Our task-oriented object representation is a generative model consisting of four components in a hierarchical spatial-temporal parse graph:

- An *affordance basis* to be grasped by hand;
- A *functional basis* to act on the target object;
- An *imagined action* with pose sequence and velocity;
- The *physical concepts* produced, *e.g.*, force, pressure.

In the learning phase, our algorithm parses the input RGB-D video by simultaneously reconstructing the 3D meshes of tools and tracking human actions. We assume that the observed human makes rational decisions in demonstration: picks the best object, grasps

	Group 1: canonical tools	Group 2: household objects	Group 3: stones
tool candidates			
Task 1 chop wood			
Task 2 shovel dirt			
Task 3 paint wall			

Figure 2.2: Given three tasks: chop wood, shovel dirt, and paint wall. Our algorithm picks up and ranks objects for each task among objects in three groups: 1) conventional tools, 2) household objects, and 3) stones, and outputs the imagined tool-use: an affordance basis (the green spot to grasp with hand), a functional basis (the red area applied to the target object), and the imagined action pose sequence.

the right place, takes the right action (poses, trajectory and velocity), and lands on the target object on the right spots. These decisions are nearly optimal against a large number of compositional alternative choices. Using a ranking-SVM approach, our algorithm will discover the best underlying physical concepts in the human demonstration, and thus the essence of the task.

The proposed method makes four major contributions:

1. We propose a novel problem of task-oriented object recognition, which is more general than defining object categories by typical examples, and is of great importance for object manipulation in robotics applications.
2. We propose a task-oriented representation, which includes both the visible object and the imagined use (action and physics). The latter is the “dark matter” [XTZ13] in computer vision.
3. Given an input object, our method can imagine the plausible tool-use, thus allows vision algorithms to reason about innovative uses of daily objects—a crucial aspect of human and machine intelligence.
4. Our algorithm can learn the physical concepts from a single RGB-D video and reason about the essence of physics for a given task.

2.1 Task-oriented Object Representation

Tools and tool-uses are traditionally studied in the field of cognitive science [OJL10, Bec80, SWB11, Bab03] with verbal definitions and case studies, and an explicit formal representation is missing in the literature.

In our task-oriented modeling and learning framework, an object used for a task is represented in a joint spatial, temporal, and causal parse graph $\mathcal{G} = (\mathcal{G}_s, \mathcal{G}_t, \mathcal{G}_c)$ including three aspects shown in Figure 2.3:

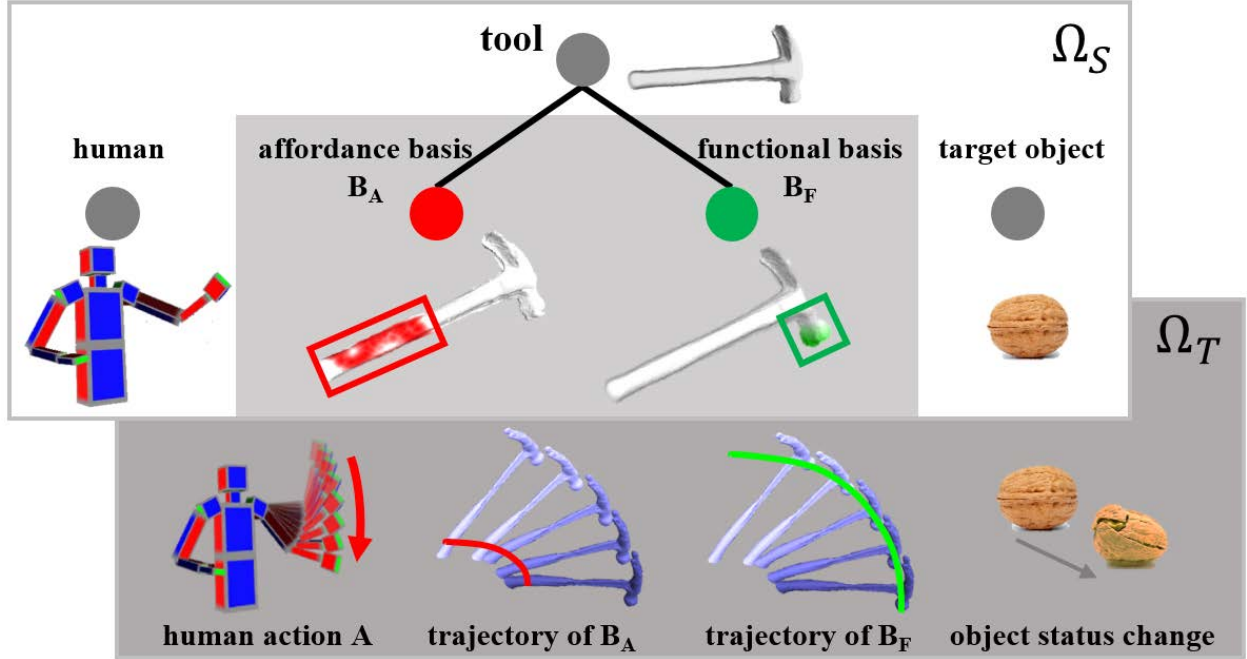


Figure 2.3: The task-oriented representation of a hammer and its use in a task (crack a nut), represented by a joint spatial, temporal, and causal space. The components in the gray area are imagined during inference phase.

- A *spatial parse graph* \mathcal{G}_s represents object decomposition and 3D relations with the imagined pose;
- A *temporal parse graph* \mathcal{G}_t represents the pose sequence in actions; and
- A *causal parse graph* \mathcal{G}_c represents the physical quantities produced by the action on the target object.

In this representation, only the object is visible as input; all other components are imagined.

2.1.1 Tool in 3D Space

An object (or tool) is observed in a RGB-D image in the inference stage, which is then segmented from the background and filled-in to become a 3D solid object denoted by X . The 3D object is then decomposed into two key parts in the spatial parse graph \mathcal{G}_s :

1. Affordance basis B_A , where the imagined human hand grasps the object with a certain

pose. Through offline training, we have collected a small set of hand poses for grasping. The parse graph \mathcal{G}_s encodes the 3D positions and 3D orientations between the hand poses and the affordance basis during the tool-use, using 3D geometric relations between the hand pose and the affordance basis, as it is done in [WZZ13]. The parse graph \mathcal{G}_s will have lower energy or high probability when the hand holds the object comfortably (see the trajectory of affordance basis B_A in Figure 2.3).

2. Functional basis B_F , where the object (or tool) is applied to a target object (the nut) to change its physical state (*i.e.*, fluent). The spatial parse graph \mathcal{G}_s also encodes the 3D relations between the functional basis B_F and the 3D shape of the target object during the action. We consider three types of the functional basis: (a) a single contact spot (*e.g.*, hammer), (b) a sharp contacting line segment or edge (*e.g.*, axe and saw), and (c) flat contacting area (*e.g.*, shovel).

2.1.2 Tool-use in Time

A tool-use is a specific action sequence that engages the tool in a task, and is represented by a temporal parse graph \mathcal{G}_t . \mathcal{G}_t represents the human action A as a sequence of 3D poses. In this work, since we only consider hand-hold objects, we collect some typical action sequences for the arm and hand movements using tools by RGB-D sensors, such as hammering, shoveling, *etc.*. These actions are then clustered into average pose sequences. For each of the sequence, we record the trajectories of the hand pose (or affordance basis) and the functional basis.

We define a space $\Omega_T = \{\mathcal{G}_t\}$ as the set of possible pose sequences and their associated trajectories of the affordance basis B_A and functional basis B_F .

2.1.3 Physical Concept and Causality

We consider thirteen basic physical concepts involved in the tool-use, which can be extracted or derived from the spatial and temporal parse graphs as Figure 2.4 illustrates.

- Blue dots and lines. We reconstruct the 3D mesh from the input 3D object, calculate

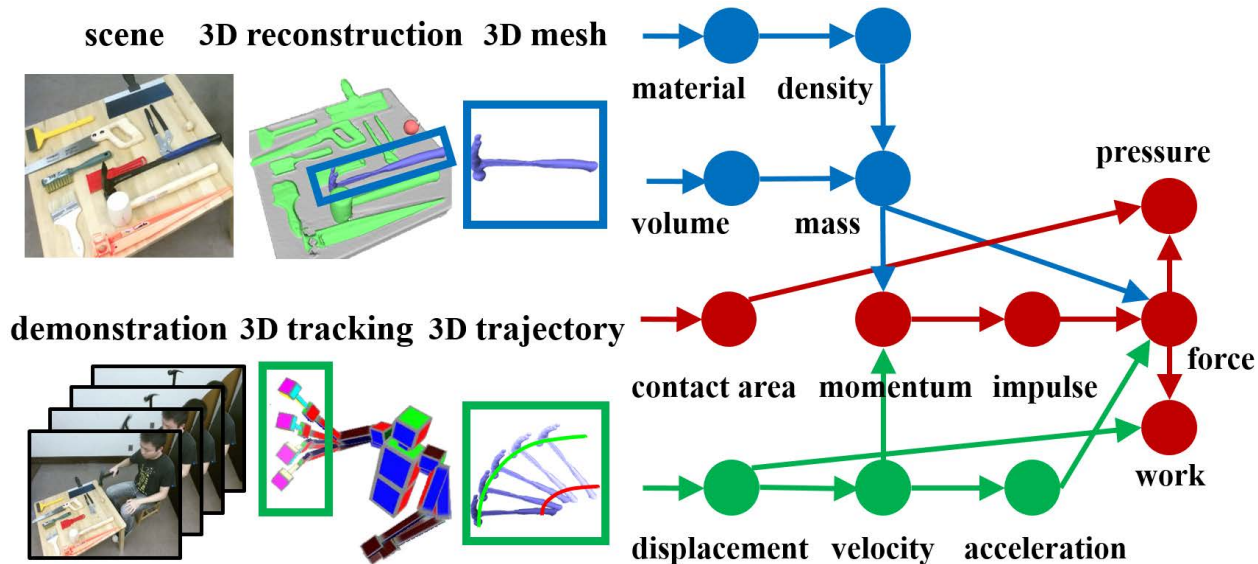


Figure 2.4: Thirteen physical concepts involved in tool-use and their compositional relations. By parsing human demonstration, the physical concepts of material, volume, concept area, and displacement are estimated from 3D meshes of the tool (blue), trajectories of tool-use (green) or jointly (red). The higher-level physical concepts can be further derived recursively.

its volume, and by estimating its material category, we obtain its density. From the volume and density, we further calculate the mass of the objects and its parts (when different materials are used).

- Green dots and lines. We can derive the displacement from the 3D trajectory of affordance basis and functional basis, and then calculate the velocity and acceleration of the two bases.
- Red dots and lines. We can estimate the contact spot, line and area from the functional basis and target object, and further compute the momentum, and impulse. We can also compute basic physical concepts, such as forces, pressure, work, *etc.*

Physical Concept Operators ∇ : We define a set of operators, including addition $\nabla_+(\cdot, \cdot)$, subtraction $\nabla_-(\cdot, \cdot)$, multiplication $\nabla_\times(\cdot, \cdot)$, division $\nabla_/\!(\cdot, \cdot)$, negation $\nabla_{\text{neg}}(\cdot)$, space integration $\nabla_{f_S}(\cdot)$, time integration $\nabla_{f_T}(\cdot)$, space derivation $\nabla_{\partial_S}(\cdot)$ and time derivation $\nabla_{\partial_T}(\cdot)$.

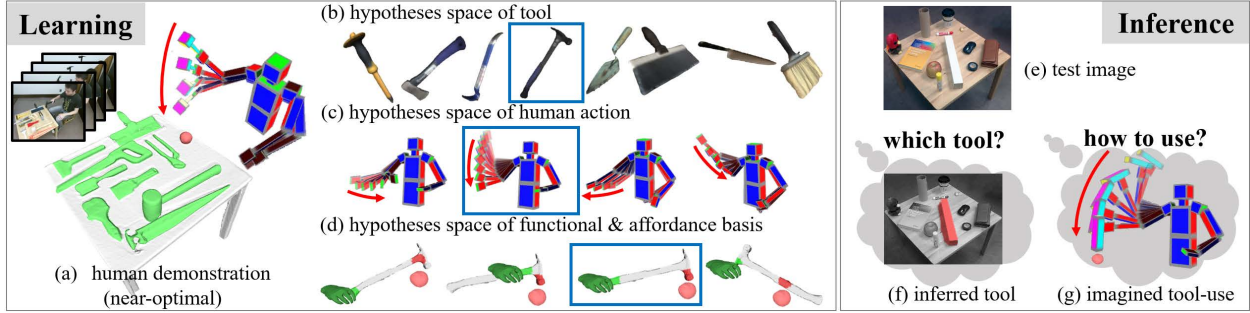


Figure 2.5: Illustration of the learning and the inference. (a)-(d) We assume the human choice (shown in blue bounding box) of tool and tool-use (action and affordance / functional bases) is near-optimal, thus most of the other combinations of tool and tool-use (action, affordance / functional bases) in the hypotheses spaces should not outperform human demonstration. Based on this assumption, we treat the human demonstration as positive example, and random sample other tools and tool-uses in the hypothesis spaces as negative examples. (e) During the inference, given an image of the static scene in a novel situation, (f) the algorithm infers the best tool and imagines the optimal tool-use.

For example, the concept of the force and acceleration are defined as:

$$\text{force} = \nabla_{\mathbf{x}}(\text{mass}, \text{acceleration}), \quad (2.1)$$

$$\text{acceleration} = \nabla_{\partial_t}(\text{velocity}). \quad (2.2)$$

The causal parse graph \mathcal{G}_c includes the specific physical concepts used in a tool-use which is often an instantiated sub-graph of the concept graph in Figure 2.4.

Since the law of physics is universally applicable, the major advantage of using physical concepts is the ability to generalize to novel situations.

2.2 Problem Definition

2.2.1 Learning Physical Concept

Given a task, the goal of the learning algorithm is to find the essential physical concept that best explains why a selected tool and tool-use is optimal.

Rational Choice Assumption states that human choices are rational and near-optimal. As shown in Figure 2.5(a-d), we assume that human chooses the optimal tool and tool-use \mathcal{G}^* (in blue box) based on the essential physical concept, so that most of other tools and tool-uses in the hypothesis spaces should not outperform the demonstration.

For instance, let us assume the essential physical concept to explain the choice of a tool is to maximize “mass”, then other tools should not offer more “mass” than the selected one. If there is a heavier tool not picked by human, it implies that “mass” is not the essential physical concept.

During the learning stage, we consider the selected tool and tool-use as the only positive training example, and we randomly sample n different combinations of tools and tool-uses \mathcal{G}_i , $i = 1 \cdots n$ in the hypothesis spaces as negative training samples.

Ranking Function: Based on the rational choice assumption, we pose the tool recognition as a ranking problem [Joa02], so that the human demonstration should be better than other tools and tool-uses with respect to the learned ranking function.

The goal of the learning is to find a ranking function indicating the essential purposes of tool-use in a given task

$$R(\mathcal{G}) = \boldsymbol{\omega} \cdot \boldsymbol{\phi}(\mathcal{G}), \quad (2.3)$$

where $\boldsymbol{\omega}$ are the weighting coefficients of the physical concepts. Intuitively, each coefficient reflects the importance of its corresponding physical concept for the task.

Learning ranking function is equivalent to find the weight coefficients so that the maximum number of pairwise constraints is fulfilled.

$$\forall i \in \{1, \dots, n\} : \boldsymbol{\omega} \cdot \boldsymbol{\phi}(\mathcal{G}^*) > \boldsymbol{\omega} \cdot \boldsymbol{\phi}(\mathcal{G}_i). \quad (2.4)$$

In this way, these constraints enforce the human demonstration \mathcal{G}^* has the highest ranking score compared with the other negative samples \mathcal{G}_i under the essential physical concept.

We approximate the solution by introducing nonnegative slack variables, similar to SVM

classification [Joa02]. This leads to the following optimization problem

$$\min \quad \frac{1}{2} \boldsymbol{\omega} \cdot \boldsymbol{\omega} + \lambda \sum_i^n \xi_i^2 \quad (2.5)$$

$$\text{s.t.} \quad \forall i \in \{1, \dots, n\} : \boldsymbol{\omega} \cdot \boldsymbol{\phi}(\mathcal{G}^*) - \boldsymbol{\omega} \cdot \boldsymbol{\phi}(\mathcal{G}_i) > 1 - \xi_i^2 \quad (2.6)$$

$$\xi_i \geq 0, \quad (2.7)$$

where ξ_i is a slack variable for each constraint, and λ is the trade-off parameter between maximizing the margin and satisfying the rational choice constraints.

This is a general formulation for the task-oriented modeling and learning problem, where the parse graph \mathcal{G} includes objects X , human action A and affordance / functional basis B_A / B_F . In this way, this framework subsumes following special cases: i) object recognition based on appearance and geometry $\boldsymbol{\phi}(X)$, ii) action recognition $\boldsymbol{\phi}(A)$, iii) detecting furniture by their affordance $\boldsymbol{\phi}(B_A)$, and iv) physical concept $\boldsymbol{\phi}(\mathcal{G}_c)$. In this work, we only focus on learning physical concepts.

In our experiment, we only consider the scenario that the learner only observes one demonstration of the teacher choosing one tool from a few candidates. Instead of feeding a large dataset for training, we are more interested in how much the algorithm can learn from such a small sample learning problem. Therefore, we only infer a single physical concept for functional and affordance basis simultaneously by iterating over the concept space, while this formulation can be naturally generalized to more sophisticated scenarios for future study.

2.2.2 Recognizing Tools by Imagining Tool-uses

Traditional object recognition methods assume that visual patterns of the objects in both training and testing sets share the same distribution. However, such assumption does not hold in tool recognition problem. The visual appearances of tools at different situations have fundamental differences. For instance, a hammer and a stone can be used to crack a nut, despite the fact that their appearances are quite different.

In order to address this challenge, we propose this algorithm to recognize tools by essential physical concepts and imagined tool-uses during the inference.

Recognize Tools by Essential Physical Concepts: Fortunately, as domain general mechanisms, the essential physical concepts in a given task are invariant across different situations. For instance, a hammer and a stone can be categorized as the same tool to crack a nut due to the similar ability to provide enough “force”. In the inference, we use the learned ranking function to recognize the best tool:

$$\mathcal{G}^* = \arg \max \omega \cdot \phi(\mathcal{G}). \quad (2.8)$$

Imagine Tool-use beyond Observations: Given an observed image of tool without actually seeing the tool-use, our algorithm first imagines different tool-uses (human action and affordance / functional bases), and then combines the imagined tool-uses with observed tools to recognize the best tool by evaluating the ranking function.

The imagined tool-uses are generated by sampling human action and affordance/functional bases from the hypothesis spaces as shown in Figure 2.5(c-d). We first assign the trajectories of imaged human hand movement to the affordance basis, then compute the trajectory of functional basis by applying the relative 3D transformation between the two bases. Lastly, we calculate the physical concepts recursively as discussed in subsection 2.1.3, and evaluate the ranking function accordingly.

The ability of imagining tool-use is particularly important for an agent to predict how they can use a tool, and physically interact with their environment.

Moreover, such ability of imagining tool-use enables the agent to actively explore different kinds of tool-uses instead of to simply mimic the observed tool-use in human demonstration. Although the tool-use in human demonstration is assumed to be optimal, other tool-uses may be better in different situations. For example, the way you use a stone to crack a nut may be quite different from the way you use a hammer.

2.2.3 Parsing Human Demonstration

In this section we show how we use the off-the-shelf computer vision algorithms to parse the input RGB-D video of human demonstration.

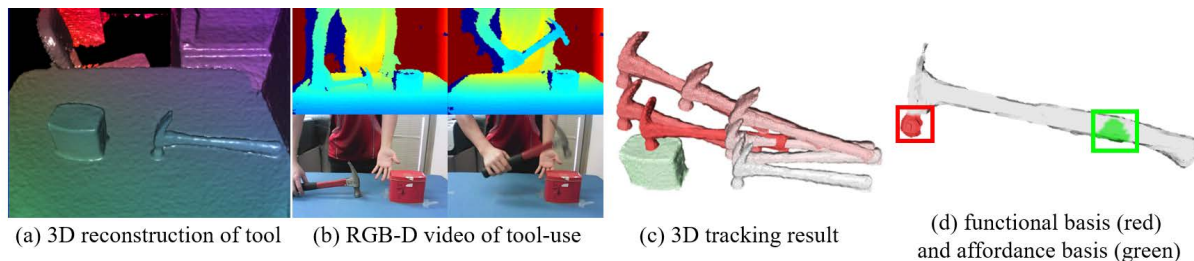


Figure 2.6: Spatial-temporal parsing of a human demonstration. (a) Using KinectFusion, we reconstruct 3D scene, including the tool and the target object. (b) Given a RGB-D video of tool-use by human demonstration, (d) affordance / functional bases can be detected by (c) 3D tracking.

3D Reconstruction: We apply the KinectFusion algorithm [NDI11] to generate a 3D reconstruction of the static scene, including a tool and an object. KinectFusion is GPU optimized such that it can run at interactive rates. Each frame of the depth image captured by RGB-D sensors has a lot of missing data. By moving the sensor around, the KinectFusion algorithm fills in these holes by combining temporal frames into a smooth 3D point cloud / mesh (Figure 2.6(a)). In this work, we only focus on medium sized tool that can be held in one hand, and can be well reconstructed by a consumer-level RGB-D sensor. By fitting the plane of the table, the tool and the target object then can be extracted from background.

3D Tracking of Tool and Target Object: Tracking the 3D mesh of tool and target object allows the algorithm to perceive the interactions and detect status changes. In this work, we use an off-the-shelf 3D tracking algorithm based on Point Cloud Library [RC11]. The algorithm first performs object segmentation using the first depth frame of the RGB-D video, and then invokes particle filtering [NUH07] to track each object segment as well as estimating the 3D orientation frame by frame (Figure 2.6(c)).

3D Hand Tracking: 3D tracking of hand positions and orientations are achieved by 3D skeleton tracking [SSK13]. The skeleton tracking outputs a full body skeleton, including 3D position and orientation of each joint. Without loss of generality, we assume the interacting

hand to be the right hand.

Contact Detection: Given the tracked 3D hand pose / tool / target object, we perform touch detection (Figure 2.6(d)) by measuring the euclidean distance among them. The touch detection between the human hand and the tool localizes the 3D location of the affordance basis, while the touch detection between the tool and the target object yields the 3D location of the functional basis.

2.3 Experiments

In this section, we first introduce our dataset, and evaluate our algorithm in three aspects: (i) learning physical concepts, (ii) recognizing tools, and (iii) imagining tool-uses.

2.3.1 Dataset

We designed a new Tool & Tool-Use (TTU) dataset for evaluating the recognition of tools and task-oriented objects. The dataset contains a collection of static 3D object instances, together with a set of human demonstrations of tool-use.

The 3D object instances include 452 static 3D meshes, ranging from typical tools, household objects and stones. Some of these object instances are shown in Figure 2.7. Some typical actions are illustrated in Figure 2.5. Each action contains a sequence (3-4 seconds) of full body skeletons. Since some action fragments only last for 0.5-1 second, which contain only around 10 frames, the interpolated actions are needed to generate smoother trajectories. Both 3D meshes and human actions are captured by consumer-level RGB-D sensors.

2.3.2 Learning Physical Concept

We first evaluate our learning algorithm by comparing with human judgments. Forty human subjects annotated the essential physical concepts for four different tasks. The distribution of annotated the essential physical concepts is shown as the blue bars in Figure 2.8. Interest-



Figure 2.7: Examples of tool instances in the dataset: (a) typical tools, (b) household objects, and (c) natural stones.

ingly, human subjects have relative consistent common knowledge that force and momentum are useful for cracking nuts, and pressure is important for chopping wood. Our algorithm learned very similar physical concepts as the red bars shown in Figure 2.8. For the other

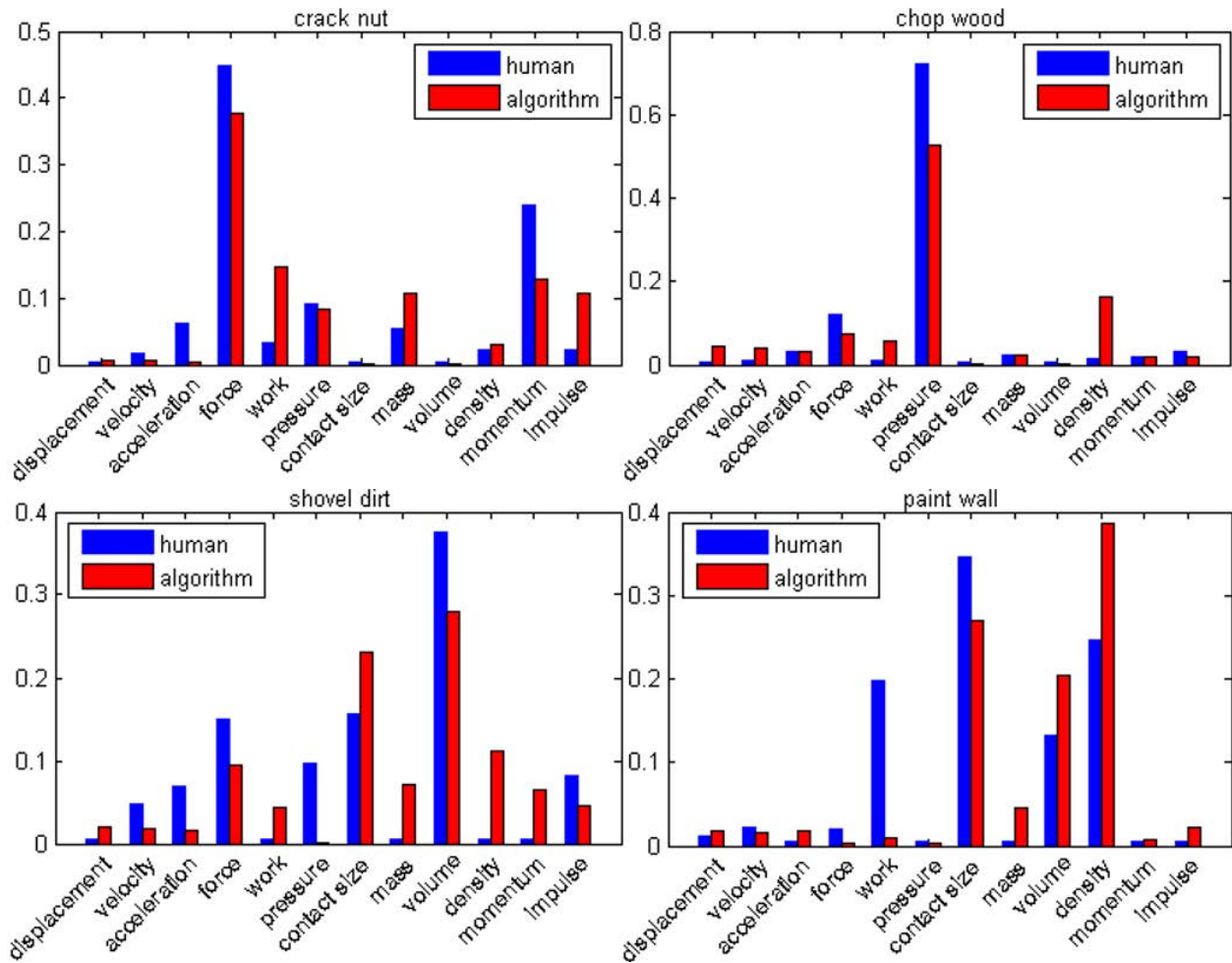


Figure 2.8: Learning essential physical concepts of tool-use. The red bars represent human judgments about what the essential physical concepts are for each task. The blue bars represent weight coefficients of different physical concepts learned by our algorithm.

two tasks, *i.e.*, shovel dirt and paint wall, although the human judgments are relatively ambiguous, our algorithm still produces relative similar results of learned physical concepts.

Figure 2.9 shows an example of learning physical concept for cracking a nut. Given a set of RGB-D images of ten tool candidates in Figure 2.9(a) and a human demonstration of tool-use in Figure 2.9(b), our algorithm imagines different kinds of tool-use as shown in Figure 2.9(c), and ranks them with respect to different physical concepts. By assuming human demonstration is rational and near-optimal, our learning algorithm selects physical concepts by minimizing the number of violations as the red area on the left of Figure 2.9(c).

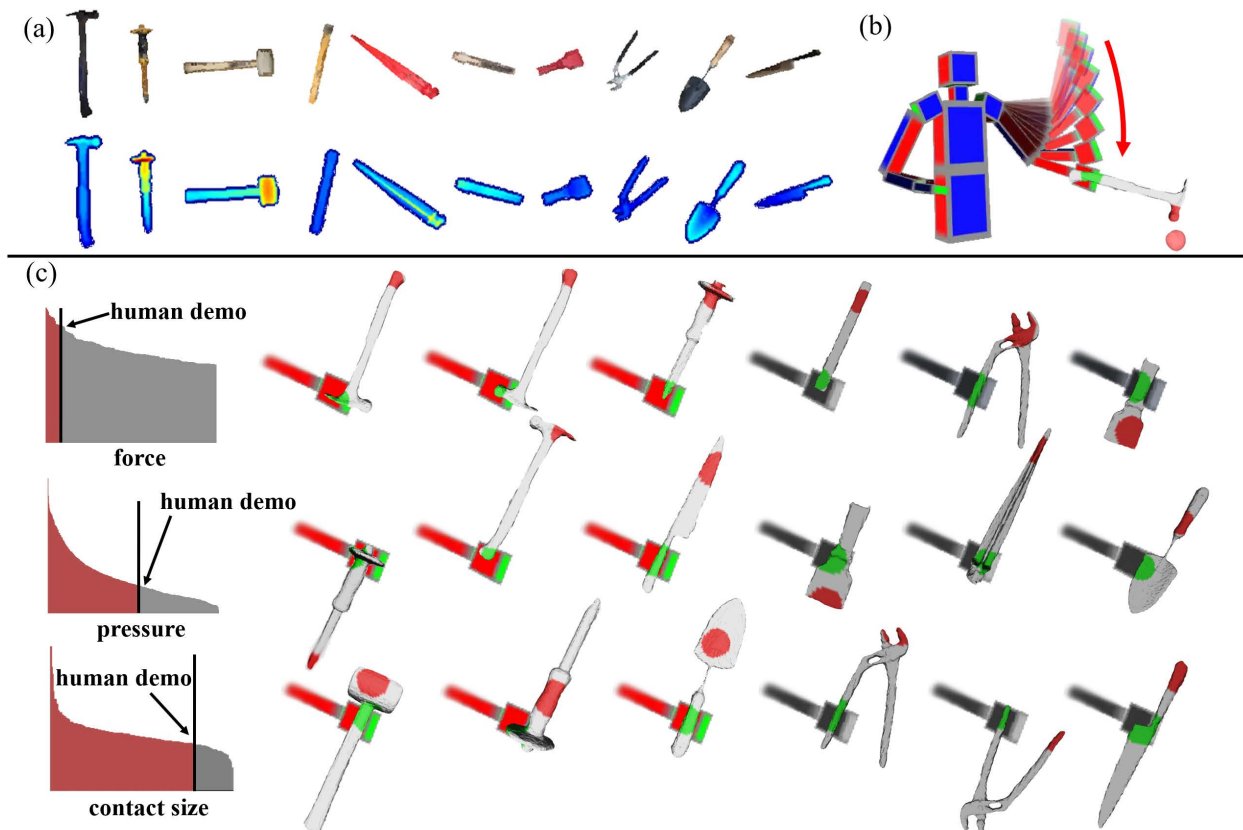


Figure 2.9: Learning physical concept from a single human demonstration for cracking a nut. (a) A set of tool candidates are given by RGB-D images. (b) The human demonstration of tool-use is assumed to be near-optimal. (c) The algorithm sorts all the samples of tool-uses with respect to different physical concepts. The black vertical bar represents the human demonstration of tool-use, while the red area and gray area represent samples that outperform and underperform human demonstration, respectively. We showed six sampled tools and tool-uses, three of which outperform human demonstration, and the others underperform human demonstration. In this cracking nut example, the “forces” is selected as the essential physical concept because there are minimum number of samples that violate the “rational choice assumption”.

For instance, the plot of “force” shows ranked pairs of tool and tool-use with respect to the forces applied on the functional basis. The force produced by human demonstration (the black vertical line) is larger than most of the generated tool-uses, thus it is near-optimal.

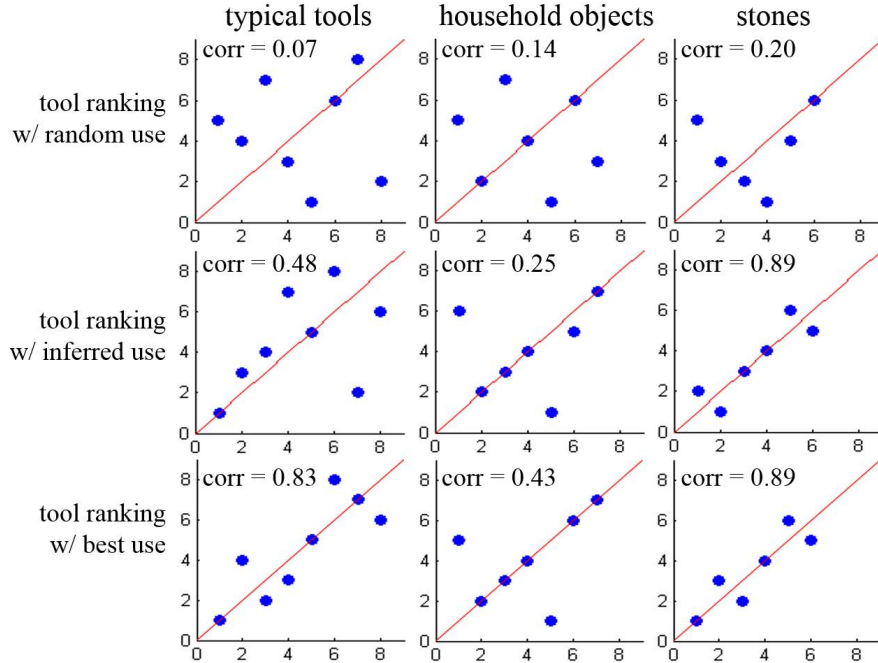


Figure 2.10: Recognizing tools for chopping wood. The scatters show tool candidates ranked by our algorithm (y-axis) with respect to the average ranking by human subjects (x-axis). The three columns show different testing scenarios, while the three rows represent different levels of tool-use imagined by inference algorithm.

The instances on the right of Figure 2.9(c) are sampled tools and tool-uses. The red ones are the cases outperform human demonstration, while the gray ones are the cases underperform human demonstration.

2.3.3 Inferring Tools and Tool-uses

In the Figure 2.2, we illustrate qualitative results of inferred tool and tool-use for three tasks, *i.e.*, chop wood, shovel dirt, and paint wall. By evaluating in three scenarios: (a) typical tools, (b) household objects, and (c) natural stones, we are interested in the generalization ability of the learned model.

Table 2.1: Accuracy of tool recognition. This table shows the correlation between the ranking generated by our algorithm and the average ranking annotated by human subjects. The three rows represent different levels of tool-use imagined by our inference algorithm. The qualitative and quantitative ranking results of tool candidates are illustrated in Figure 2.2 an Figure 2.10, respectively.

correlation of ranking algorithm vs. human	chop wood			shovel dirt			paint wall		
	tool	object	stone	tool	object	stone	tool	object	stone
tool + random use	0.07	0.14	0.20	0.52	0.32	0.09	0.12	0.11	0.31
tool + inferred use	0.48	0.25	0.89	0.64	0.89	0.14	0.10	0.64	0.20
tool + best use	0.83	0.43	0.89	0.64	0.89	0.14	0.10	0.64	0.20

2.3.3.1 Recognizing Tools

We asked four human subjects to rank tool candidates shown in Figure 2.2. For the task of chopping wood in Figure 2.10, we plot tool candidates in terms of their average ranking by human subjects (x-axis) and their ranking generated by our algorithm (y-axis).

The three columns show different testing scenarios. We can see that our model learned from canonical cases of tool-use can be easily generalized to recognize tools in novel situation, *i.e.*, household objects and natural stones. The correlation between algorithm ranking and human ranking is consistent across these three scenarios. Sometimes, the algorithm works even better on the stone scenarios.

The three rows represent different levels of tool-use: (a) the “tool-ranking with random use” evaluates the ranking of tools by calculating the expected scores of random tool-use; (b) the “tool-ranking with inferred use” evaluates the ranking of tools by calculating their optimal tool-use inferred by our algorithm; (c) the “tool-ranking with best use” evaluates the ranking of tools by their best uses given by human subjects. The Table 2.1 summarizes the correlations between human rankings and algorithm rankings on three tasks.

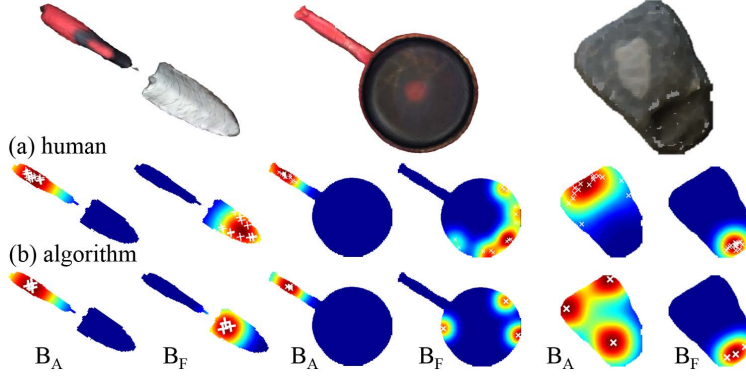


Figure 2.11: Comparison of (a) human predicted tool-use and (b) algorithm imagined tool-use for shoveling dirt.

2.3.3.2 Imagining Tool-uses

We also evaluate the imagined tool-uses in three aspects: human action A , affordance basis B_A , functional basis B_F .

The evaluation of human action is based on the classification of action directions, which are “up”, “down”, “forward”, “backward”, “left” and “right”. The classification accuracy for this problem over all the experiments is 89.3%. The algorithm can reliably classify the action of cracking a nut as “down”. But there are some ambiguities in classifying the action of shoveling dirt, because “left” and “right” are physically similar.

Figure 2.11 illustrates three example of imagined affordance basis B_A and functional basis B_F . Comparing to human annotations, the algorithm finds very similar positions of affordance basis B_A and functional basis B_F . In Table 2.2, we show the 3D distances between the positions imagined by our algorithm and the positions annotated by human subjects in centimeter.

2.4 Discussions

In this chapter, we present a new framework for task-oriented object modeling, learning and recognition. An object for a task is represented in a spatial, temporal, and causal parse graph including:

Table 2.2: Errors of imagining tool-use for affordance / functional bases (B_A and B_F) . The table shows the 3D distances between their positions imagined by our algorithm and the positions annotated by human subjects. The specific positions for sample tool candidates are shown in Figure 2.11.

3D distance (cm) algorithm vs. human	chop wood			shovel dirt			paint wall		
	tool	object	stone	tool	object	stone	tool	object	stone
B_A - top 1	1.75	3.02	3.19	1.17	2.03	3.28	0.43	2.48	2.86
B_A - top 3	1.04	2.17	2.81	0.97	0.52	2.21	0.31	2.32	2.67
B_F - top 1	0.48	5.97	3.91	6.98	6.38	0.23	2.35	2.74	2.65
B_F - top 3	0.27	5.92	3.95	2.85	3.29	0.31	1.43	2.64	2.71

- Spatial decomposition of the object and 3D relations with the imagined human pose;
- Temporal pose sequences of human actions; and
- Causal effects (physical quantities on the target object) produced by the object and action. In this inferred representation, only the object is visible, while all other components are imagined ‘dark’ matters. This framework subsumes other traditional problems, such as: (a) object recognition based on appearance and geometry; (b) action recognition based on poses; (c) object manipulation and affordance in robotics.

We argue that objects, especially man-made objects, are designed for various tasks in a broad sense [OJL10, Bec80, SWB11, Bab03], and therefore it is natural to study them in a task-oriented framework.

In the following, we briefly review related work in the literature of cognitive science, neuroscience, vision and robotics.

2.4.1 Related Work

Cognitive Science and Psychology: The perception of tools and tool-uses has been extensively studied in cognitive science and psychology. Our work is motivated by the

astonishing ability of animal tool-uses [Goo86, BW89, WGM99, Bec80, SWB11, SMT14]. For example, Santos *et al.* [SPS06] trained two species of monkeys on a task to choose one of the two canes to reach food under various conditions that involve physical concepts. Weir *et al.* [WCK02] reported that New Caledonian crows can bend a piece of straight wire into a hook and successfully use it to lift a bucket containing food from a vertical pipe. These discoveries suggest that animals can reason about the functional properties, physical forces and causal relations of tools using domain general mechanisms. Meanwhile, the history of human tool designing reflects the history of human intelligence development [McG92, Fre07, GGI94, Vae12]. One argument in cognitive science is that an intuitive physics simulation engine may have been wired in the brain through evolution [BHT13, TKG11, USG14], which is crucial for our capabilities of understanding objects and scenes.

Neuroscience: Studies in neuroscience [Lew06, FH05, CL05] found in the fMRI experiments that cortical areas in the dorsal pathway are selectively activated by tools in contrast to faces, indicating a very different pathway and mechanism for object manipulation from that of object recognition. Therefore, studying this mechanism will lead us to new directions for computer vision research.

Robotics and AI: There is also a large body of work studying tool manipulation in robotics and AI. Some related work focuses on learning affordance parts or functional object detectors, *e.g.*, [SLZ08, VV12, MLB08, Sto05, PEK13, JI11, VV11, MKF14, MTF15]. They, however, are still learning high-level appearance features, either selected by affordance / functional cues, or through human demonstrations [ACV09], not to reason about the underlying physical concepts.

Computer Vision: The most related work in computer vision is a recent stream that recognizes functional objects (*e.g.*, chairs) [Ho87, SB91, GGV11, KRK11, WZZ13, ZFF14, LFU13, KCG14] and functional scene (*e.g.*, bedroom) [ZZ13, GSE11, CCP13, JKS13] by fitting imagined human poses. The idea of integrating physical-based models has been used

for object tracking [TA15, OKA11] and scene understanding [ZZY13, ZZY14] in computer vision. But our work goes beyond affordance.

2.4.2 Limitation and Future Work

In this chapter, we only consider handhold physical objects as tools. We do not consider other tools, such as, electrical, digital, virtual or mental tools. Our current object model is also limited to rigid bodies, and cannot handle deformable or articulated objects, like scissors, which requires fine-grained hand pose and motion. All these tasks request richer and finer representations which we will study in the future work.

CHAPTER 3

Inferring Forces and Learning Human Utilities

In recent years, there has been growing interest in studying object affordance in computer vision and graphics. As many object classes, especially man-made objects and scene layouts, are designed primarily to serve human purposes, the latest studies on object affordance include reasoning about geometry and function, thereby achieving better generalizations to unseen instances than conventional appearance-based machine learning approaches. In particular, Grabner *et al.* [GGV11] designed an “affordance detector” for chairs by fitting typical human sitting poses to 3D objects.

In this chapter, we propose to go beyond visible *geometric compatibility* to infer, through physics-based simulation, the forces/pressures on various body parts (hip, back, head, neck, arm, leg, *etc.*) as people interact with objects. By observing people’s choices in videos—for example, in selecting a specific chair in which to sit among the many chairs available in a scene (Figure 3.1)—we can learn the *comfort intervals* of the pressures on body parts as well as human preferences in distributing these pressures among body parts. Thus, our system is able to “feel”, in numerical terms, discomfort when the forces/pressures on body parts exceed comfort intervals. We argue that this is an important step in representing *human utilities*—the pleasure and satisfaction defined in economics and ethics (*e.g.*, by the philosopher Jeremy Bentham) that drives human activities at all levels. In our work, human utilities explain why people choose one chair over others in a scene and how they adjust their poses to sit more comfortably, providing a deeper and finer-grained account not only of object affordance but also of people’s behaviors observed in videos.

In addition to comfort intervals for body pressures, our notion of human utilities also takes into consideration: (i) the tasks observed in a scene—for example, students conversing

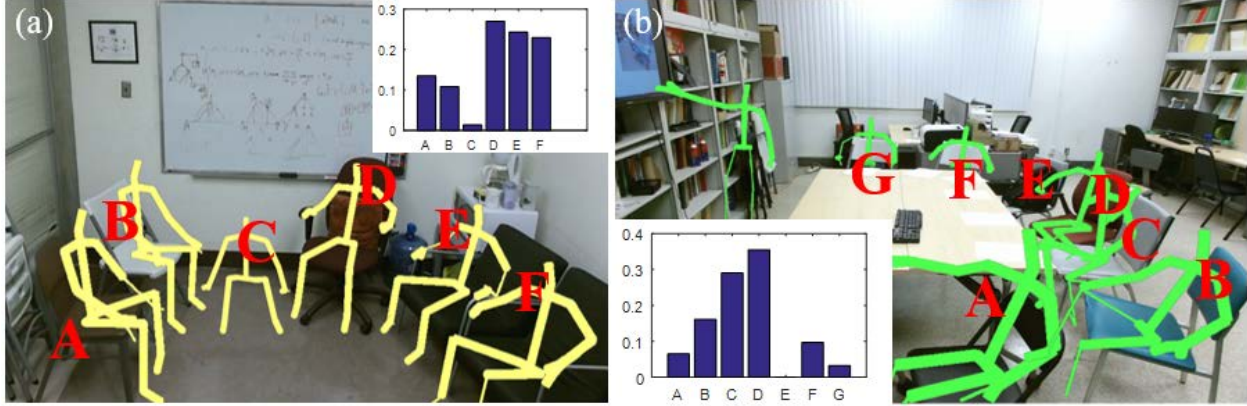


Figure 3.1: Examples of sitting activities in (a) an office and (b) a meeting room. In addition to geometry and appearance, people also consider other important factors including comfortability, reaching cost, and social goals when choosing a chair. The histograms indicate human preferences for different candidate chairs.

with a professor in an office (Figure 3.1(a)) or participating in a teleconference in a lab (Figure 3.1(b))—where people must attend to other objects and humans, and (ii) the space constraints in a planned motion—*e.g.*, the cost to reach a chair at a distance. In a full-blown application, we demonstrate that human utilities can be used to analyze human activities, such as in the context of robot task planning.

3.1 Related Work

Modeling Affordance: The concept of affordance was first introduced by Gibson [Gib77]. Hermans *et al.* [HRB11] and Fritz *et al.* [FPB06] predicted action maps for autonomous robots. Later, researchers incorporated affordance cues in shape recognition by observing people interacting with 3D scenes [DFL12, FDG14, WZZ13]. Adding geometric constraints, several researchers computed alignments of a small set of discrete poses [GGV11, GSE11, JS13b]. By searching a continuous pose parameter space of shapes, Kim *et al.* [KCG14] obtained accurate alignments between shapes and human skeletons. More recently, Savva *et al.* [SCH14] predicted regions in 3D scenes where actions may take place. Applications that use affordance in scene labeling and object placement are reported in [JS13a, JLS12,

JKS13]. A closely related topic is to infer the stability and the supporting relations in a scene [JGS13, ZZY14, LZZ15].

Inferring Forces from Videos: For pose tracking, Brubaker *et al.* [BF08, BSF09, BFH10] estimate contact forces and internal joint torques using a mass-spring system. More recently, Zhu *et al.* and Pham *et al.* [ZZZ15, PKQ15] use numerical differentiation methods to estimate hand manipulation forces. These methods are either limited to rigid body problems or employ oversimplified volumetric human models inadequate in simulating detailed human interactions with arbitrary 3D objects in scenes. In computer graphics, soft body simulation has been used to jointly track human hands and calculate contact forces from videos [ZZM13, WMZ13].

Task Planning in Robotics: Robotics has a rich history in seeking to understand human motion through synthesized trajectories. Hierarchical task planning through 2D human motion synthesis is explored in [ZRG09], but these models are constrained to 2D motion plans and relatively simplistic location-oriented goals. More complex models such as [KDD09] seek to understand task-oriented human motion on a musculoskeletal level, but they do not take into account the context of an entire 3D environment. To synthesize logical trajectories, we rely on robust planning algorithms developed for robotics control applications (*e.g.* [GO04]) and we apply these forward planning engines to scene understanding by synthesizing rational human trajectories, a well-studied robotics problem [LaV06].

Physics-based Human Simulation in Graphics: Physics-based techniques for simulating deformable objects have been widely employed in computer graphics after the pioneering work on the topic [TPB87, TF88]. Popular methods for simulating elastoplastic material include mass-spring-damper systems [Mil88, TT94], the Finite Element Method (FEM) [TBH03, ITF04, MZS11, HJS13], and the Material Point Method (MPM) [SSC13, SSJ14]. We adopt the FEM as it is physically accurate, robust, and computationally efficient. Among various deformable solids, the human body has received much attention due to its impor-

tance in character animation for movies and games. Significant prior work models human anatomical structure as a biomechanical musculoskeletal system including adipose tissues [LST09, LPK14, SLS14, SZK15]. For efficiency, our human body model is simply a single isotropic elastic body. This enables us to run a large number of simulations in a reasonable time limit and still achieve useful results.

Contributions

This work makes five major contributions:

1. We incorporate physics-based, soft body simulations to infer the *invisible* physical quantities—*e.g.*, forces and pressures—during human-object interactions. To our knowledge, this is the first work to adopt state-of-the-art, physically accurate simulations to scene understanding. A major advantage of our method is its robustness in inferring both the forces and pressures acting on the entire human body as our model, which is comprised of more than 2,000 vertices, deforms in a realistic manner.
2. Given a static scene acquired by RGB-D sensors, our proposed framework reasons about the relevant physics in order to synthesize creative, *physically stable* ways of sitting on objects.
3. By incorporating a conventional robotics path planner, our proposed framework can generalize a static sitting pose to extend over a *dynamic* moving sequence.
4. From human demonstrations, our system learns to generate the force histograms of each human body part, which essentially defines human utilities, such as comfortability, in terms of the force acting on each body part.
5. We propose a method to robustly generate *volumetric* human models from the widely-used stick-man models acquired using Kinect sensors [OKA11], and introduce a pipeline to reconstruct *watertight* 3D scenes with well-defined interior and exterior regions, which are critical to the success of physics-based scene understanding using advanced simulations.

Overview

The remainder of this chapter is organized as follows: In section 3.2, we introduce our representation, which incorporates physical quantities into the spatiotemporal spaces of interest.

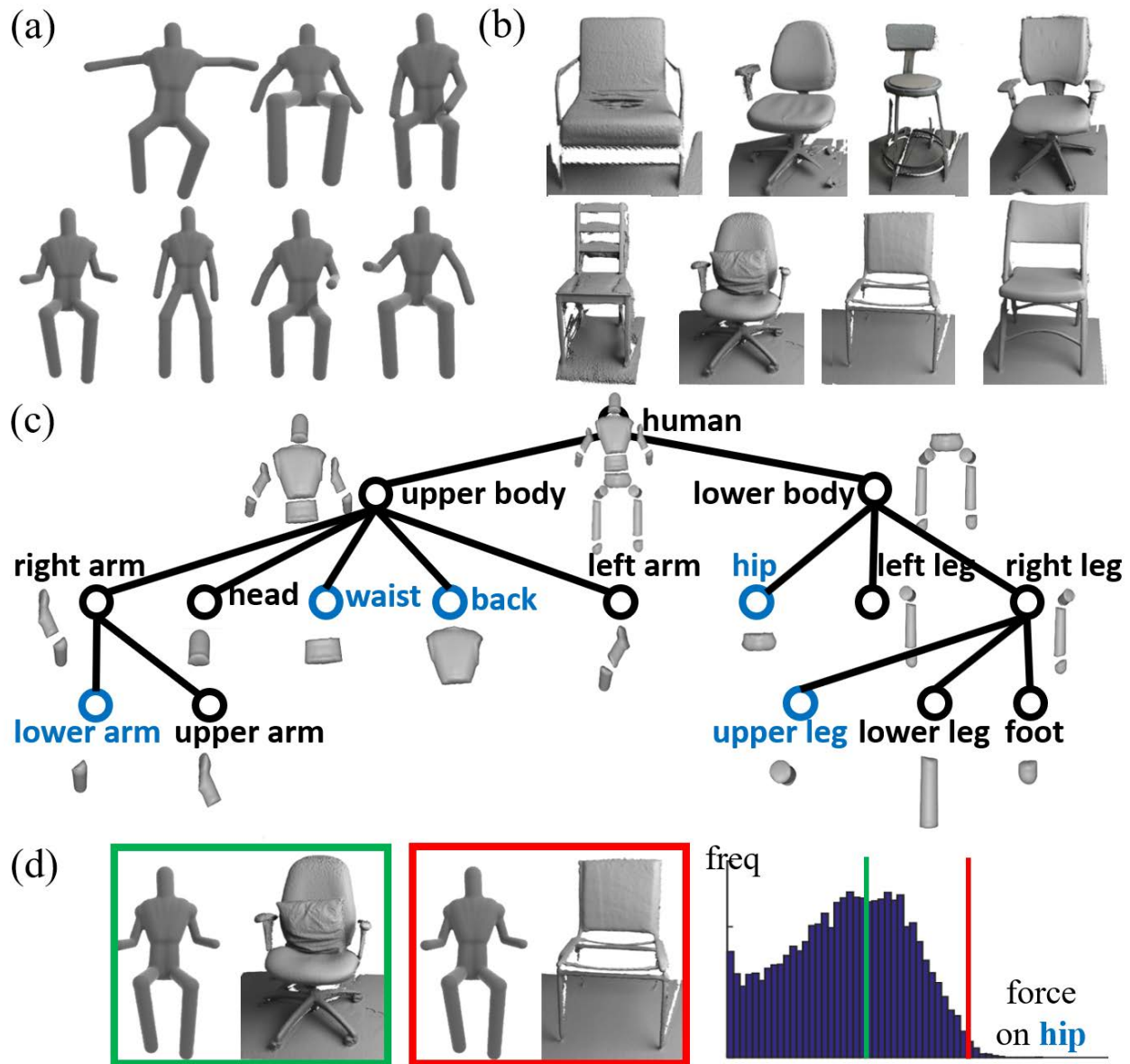


Figure 3.2: (a) We collect a set of human poses and cluster them into 7 average poses. (b) Various chairs extracted from scanned scenes. (c) Each human pose is decomposed into 14 body parts. When a human interacts with a chair, we infer the forces on each body part using FEM simulations. (d) Examples illustrating human preferences; green indicates a comfortable sitting activity, red an uncomfortable one.

In section 3.3, we describe the pipeline for calculating the relevant physical quantities, which makes use of the Finite Element Method (FEM). In section 3.4, we formulate the problem as a ranking task, and introduce a learning and inference algorithm under the assumption of rational choice. Section 3.5 demonstrates that our proposed framework can be easily generalized to challenging new situations. Section 3.6 concludes the chapter by discussing limitations and future work.

3.2 Representation

3.2.1 Spatial Entities and Relations in 3D Spaces

We represent sitting behaviors and associated relations in a parse graph \mathcal{G} , which includes (i) spatial entities—objects and human poses extracted from 3D scenes—and (ii) spatial relations—object-object and human-object relations.

Spatial Entities: For each frame of the input video, the parse graph \mathcal{G} is first decomposed into a static scene and a human pose. The static scene is further decomposed into a set of 3D objects, including chairs (Figure 3.2(b)). In this work, we consider only human poses related to sitting. We collect typical sitting poses using a Kinect sensor, and align and cluster them into 7 average poses (Figure 3.2(a)). For each average pose, we first convert the Kinect stick-man models (Figure 3.3(a)) into tetrahedralized human models (Figure 3.3(b)). These are then discretized into 14 pre-defined human body parts (Figure 3.3(c)) for simulations, as shown in Figure 3.3(d).

Spatial Relations: Pairs of objects extracted from 3D scenes form object-object relations, and each object and human pose pair forms a human-object relation. Figure 3.6(d)(e) show an example of spatial relations. For the purposes of this work, we define these two spatial relations as spatial features $\phi_s(\mathcal{G})$ that encode the relative spatial distances and orientations. At a higher level, human-object relations also encode visual attention and social goals.

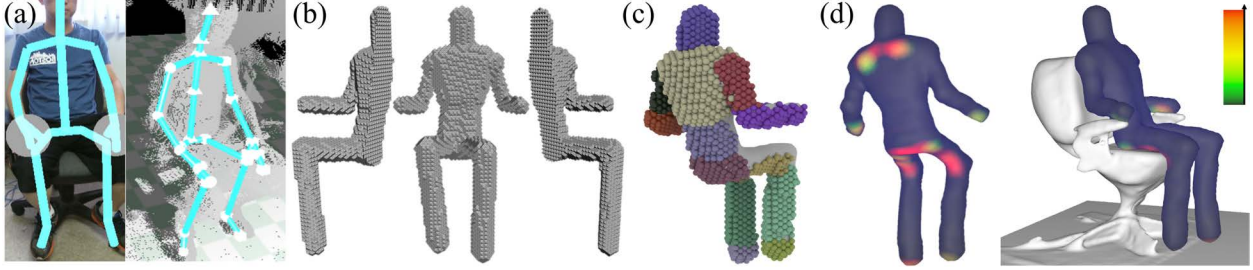


Figure 3.3: The stick-man model (a) captured using a Kinect is converted into a tetrahedralized human model (b) and then segmented into 14 body parts (c). Using FEM simulation the physical quantities $\phi_p(\mathcal{G})$ are estimated at each vertex of the FEM mesh; the forces at each vertex are visualized in (d).

3.2.2 Physical Quantities of Human Utilities

To date, researchers have mostly generated affordance maps by evaluating the geometric compatibility between people and objects [KCG14, JLS12, FDG14, JKS13, SCH14, WZZ13]. We employ a more meaningful and quantifiable metric—forces (including pressures) as physical quantities $\phi_p(\mathcal{G})$ produced during human-object interactions. The forces acting on each body part essentially determines the *comfortability* of a person interacting with the scene. People tend to choose more comfortable chairs that will apparently provide better distributions of supporting forces at each body part (Figure 3.2(d)).

Deploying our physically simulated volumetric human models in the reconstructed scenes, we can estimate fine-grained external forces at each vertex of the human model, as shown in Figure 3.3(d). In this work, we use the FEM to compute forces. The force acting on each body part can be estimated by summing up vertex-wise force contributions. A major advantage of using physical concepts is their ability to generalize to new situations.

3.2.3 Human Utilities in Time

To model the human utility, a plan cost $\phi_t(\mathcal{G})$ is incorporated into our proposed framework. This is defined as a body pose sequence from a given initial state to a goal state, which encodes people’s intentions and task planning through time. Compared to prior work, adding

plan cost extends the solution space from a static human pose to *dynamic* pose sequences.

To simplify the problem, we use the Probabilistic Roadmap (PRM) planner [KSL96] to calculate the plan cost. Viewed from above, we project the 3D scene to create a planar map, and use a 2D PRM to calculate the plan cost. However, our proposed framework does not preclude the use of more sophisticated planning methods in 3D space.

3.3 Estimating the Forces in 3D Scenes

3.3.1 Dataset of 3D Scenes and Human Models

Our dataset includes reconstructed *watertight* 3D scenes, 3D objects (including chairs) extracted from the scenes, tracked human skeletons and *volumetric* human poses. The skeletons and volumetric human poses are registered in the reconstructed scenes.

The most distinguishing feature of our dataset relative to previous ones (*e.g.*, [CZK15, HWM14, XOT13, SCH14]) is the watertight property of our reconstructed scenes. This is crucial for physics-based simulation methods such as the FEM. Furthermore, our dataset includes much larger variations of chair-shaped objects and human poses, as shown in Figure 3.2(a)(b), as well as more challenging and cluttered scenes.

3.3.2 Reconstructing Watertight Scenes

Reconstructing Closed-loop Scenes: Reconstruction methods that use purely geometric registration [NDI11, KPR15, NZI13, WKF12] suffer from aliasing of fine geometric details and an inability to disambiguate different locations based on local geometry. Such problems are compounded when attempting to register loop closure fragments with low overlap. In our work, we reconstruct 3D scenes with global optimization based on line processes [CZK15], resulting in detailed reconstructions with loop closures, as shown in Figure 3.4(a).

Converting to Watertight Scenes: Collision detection and resolution in the simulation requires a watertight scene mesh. We first use Poisson disk sampling [Bri07] to generate

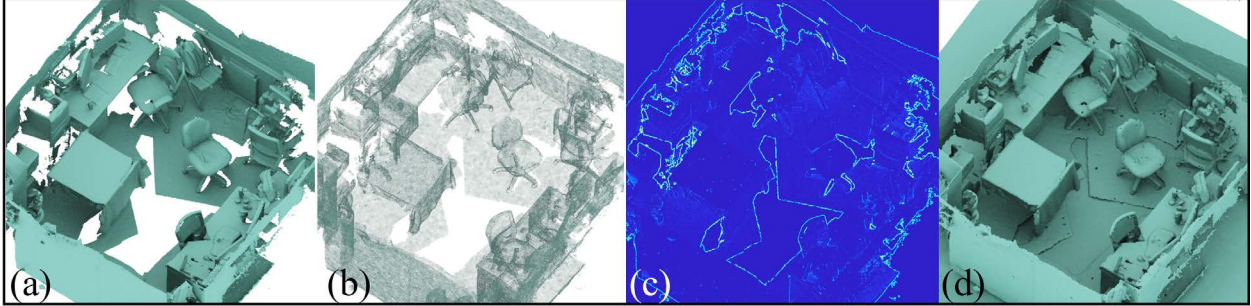


Figure 3.4: (a) From a reconstructed 3D indoor scene [CZK15, SCH14], (b) we uniformly sample vertices in the input mesh with Poisson disk sampling [Bri07], then convert them into a watertight mesh [LC87, MLJ13] with well-defined interior and exterior regions. Differences (c) between the input mesh and the converted watertight mesh. By adding a ground geometry, we obtain a detailed, watertight reconstruction (d) of the 3D scene, which is inputted to the simulation.

uniformly distributed vertices from the input triangle mesh, as illustrated in Figure 3.4(b). Each vertex is then replaced with a fixed-radius sphere level set [MLJ13]. Subsequently, the Constructive Solid Geometry (CSG) union operation is applied to this level set and a ground level set to produce a complete scene with a filled-in floor. Finally, the Marching Cubes algorithm [LC87] is applied to the level set in order to generate the watertight surface, as shown in Figure 3.4(d). The resulting scene has the well-defined interior and exterior regions required by the simulation.

3.3.3 Modeling Volumetric Human Pose

Skeleton Alignment and Clustering: The resting poses of human skeletons acquired using the Kinect are aligned by solving the absolute orientation problem using Horn’s quaternion-based method [Hor87]; *i.e.*, finding the optimal rotation and translation that maps one collection of vertices to another in a least squares sense:

$$\min \sum_i \|\mathbf{R}\mathbf{A}(:, i) + \mathbf{t} - \mathbf{B}(:, i)\|^2, \quad (3.1)$$

where \mathbf{A} and \mathbf{B} are a $3 \times N$ matrices whose columns comprise the coordinates of the N source vertices and N target vertices, respectively. Presently, we have $N = 3$ (left shoulder,

right shoulder, and spine base) for skeleton alignment. The K-means clustering algorithm [CGB07, Syl09, DLR77] is then applied to cluster the resting poses into 7 categories, as shown in Figure 3.2(a).

Skeleton Skinning: Human skeleton data comprise joints, segments, and their orientations. For simplicity, an analytic geometric primitive is assigned to each body part. The primitives include ellipsoids (including spheres), hexahedra, and cylinders. The parameters of the primitives are chosen such that they best fit the body parts. A high-resolution level set is then applied to wrap around the union of all the primitives [MLJ13]; its zero isocontour approximates the skin [LC87].

Volumetric Discretization: Although the Marching Cubes algorithm suffices to extract a triangulated skin mesh from the level set, our simulation requires a full discretization of the volume bounded by the skin. To achieve this, we embed the skin level set into a body-centered cubic tetrahedral lattice as in [MBT03]. This results in a tetrahedralized human shape geometry as shown in Figure 3.3(b).

3.3.4 Simulating Human Interactions With Scenes

As stated earlier, we chose the FEM to simulate human tissue dynamics. Our simulation requires only reconstructed watertight scenes and volumetric human poses as inputs. The outputs of the simulation are the relevant physical quantities $\phi_p(\mathcal{G})$; *e.g.*, forces and pressures.

Elasticity: The human body is modeled as an elastic material. The total elastic potential energy is defined as

$$\Phi^E(\mathbf{x}) = \int_{\Omega} \Psi^E(\mathbf{x}) d\mathbf{x} \approx \sum_e V_e^0 \Psi^E(\mathbf{F}(\mathbf{x})), \quad (3.2)$$

where Ω is the simulation domain defined by the tetrahedral body mesh, \mathbf{x} denotes the deformed vertex positions, and V_e^0 is the initial undeformed volume of tetrahedral element e . The hyperelastic energy density function Ψ^E is defined in terms of the deformation gradient

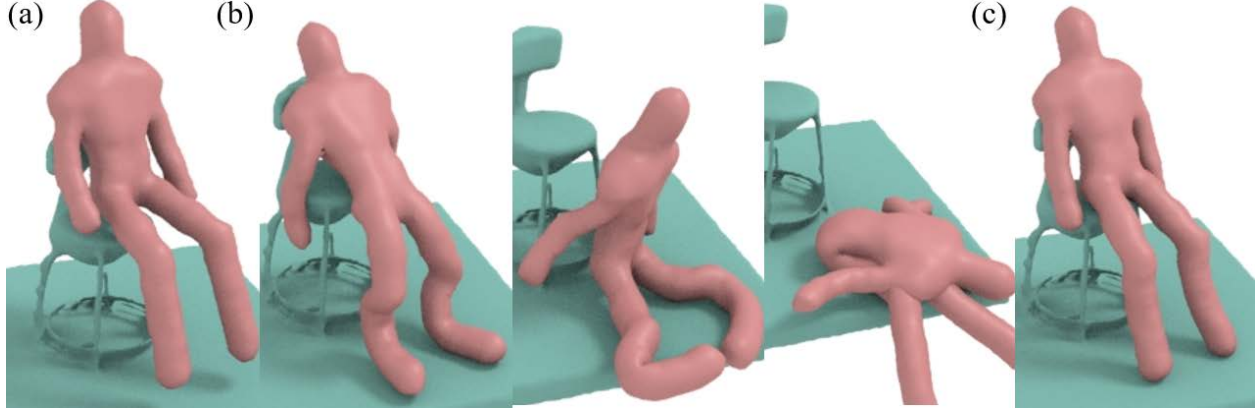


Figure 3.5: (a) Given an initial human pose in a 3D scene subject to gravity, (b) without adequate damping, the human body is too energetic and produces unnaturally bouncy motion. (c) With proper damping, the simulation converges to a physically stable rest pose in a small number of timesteps.

$\mathbf{F} = \frac{\partial \mathbf{x}}{\partial \mathbf{X}}$, where \mathbf{X} denotes the undeformed vertex positions. We use the fixed corotated elasticity model [SHS12] for Ψ^E due to its robustness in handling large deformations.

Contact Forces: To model contact forces, we need to penalize penetrations of the human body mesh into the scene mesh. This requires a differentiable volumetric description of the scene geometry. With watertight scenes, the level set reconstruction is performed by directly computing signed distances from level set vertices to the mesh surface. In each simulation timestep, all human mesh vertices are checked against the scene level set. If a penetration is detected for vertex i , a collision energy $\Phi^C(\mathbf{x}_i)$ that penalizes the penetration distance in the normal direction is assigned to the corresponding vertex

$$\Phi^C(\mathbf{x}_i) = \frac{1}{2}k_c(\mathbf{x}_i - \mathcal{P}(\mathbf{x}_i))^2, \quad (3.3)$$

where k_c is a penalty stiffness constant and $\mathcal{P}(\mathbf{x}_i)$ projects \mathbf{x}_i onto the closest point on the level set zero isocontour along its normal direction. To prevent free sliding along the collision geometry, we further introduce a friction force that slightly damps the tangential velocity for vertices in collision.

Table 3.1: Physical simulation parameters

Timestep:	Density:	Young's modulus:	Poisson's ratio:
$1 \times 10^{-3}s$	$1000kg/m^3$	$0.15kPa$	0.3
Collision stiffness:	Friction coeff:	Damping coeff:	Gravity:
$1 \times 10^4kg/s^2$	1×10^{-3}	$50kg/s$	$9.81m/s^2$

Dynamics Integration: Backward Euler time integration is used to solve the momentum equation. From time n to $n + 1$, the nonlinear system to solve is

$$\mathbf{M} \frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} = \mathbf{f}(\mathbf{x}^{n+1}, \mathbf{v}^{n+1}) + \mathbf{M}g, \quad (3.4)$$

$$\mathbf{f}(\mathbf{x}^{n+1}, \mathbf{v}^{n+1}) = \mathbf{f}^E(\mathbf{x}^{n+1}) + \mathbf{f}^C(\mathbf{x}^{n+1}) + \mathbf{f}^D(\mathbf{v}^{n+1}), \quad (3.5)$$

$$\mathbf{x}^{n+1} - \mathbf{x}^n = \mathbf{v}^{n+1} \Delta t. \quad (3.6)$$

Here \mathbf{M} is the mass matrix, \mathbf{x} denotes position, \mathbf{v} denotes velocity, $\mathbf{f}^E = -\frac{\partial \Phi^E}{\partial \mathbf{x}}$ is the elastic force, $\mathbf{f}^C = -\frac{\partial \Phi^C}{\partial \mathbf{x}}$ is the contact force, $g = 9.8m/s$ is gravity, and $\mathbf{f}^D = -\nu \mathbf{v}$ is an additional force to dampen the velocities, where ν is the damping coefficient. Figure 3.5(b) shows that without the damping force, the deformable human body model is too energetic and may produce unnaturally bouncy motion. While there exist more accurate viscoelastic material models of human tissue, our simple damping force is easy to implement and achieves similar behaviors for the simulation results. We solve the above nonlinear system for positions \mathbf{x}^{n+1} and velocities \mathbf{v}^{n+1} using Newton's method [GSS15].

Simulation Outputs: When the simulation comes to rest, $\mathbf{v} = \mathbf{0}$ and the damping forces vanish. The elastic, contact, and gravity forces sum to zero everywhere over the mesh. As the output of the simulation, we export the computed contact forces acting on the skin surface.

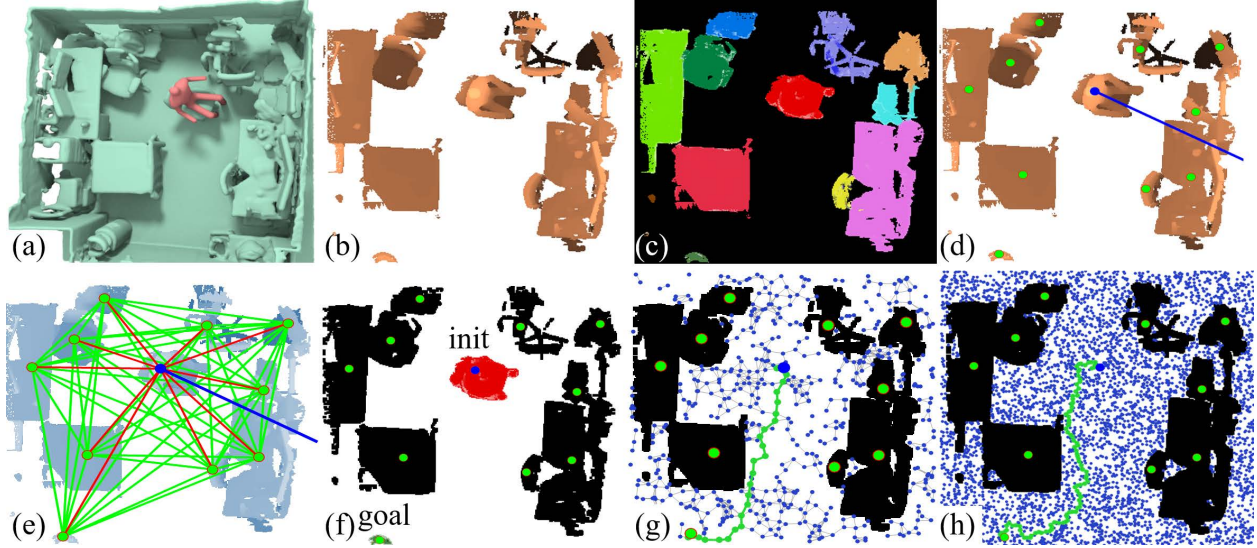


Figure 3.6: **Data pre-processing.** (a) Given a reconstructed 3D scene, (b) we project it down onto a planar map, and (c) segment 3D objects from the scene. (d) visualizes 3D object positions (green dots), human head position (blue dot), and orientation (blue line). (e) **Spatial features** $\phi_s(\mathcal{G})$ are defined as human-object (red lines) and object-object (green lines) relative distances and orientations. (f) **Temporal features** $\phi_t(\mathcal{G})$ are defined as the plan cost from a given initial position to a goal position. (g)(h) Two solutions generated by the PRM planner using graphs with different numbers of nodes (more nodes yield finer-grained plans at higher cost).

3.4 Learning and Inferring Human Utilities

3.4.1 Extracting Features

We craft features $\phi(\mathcal{G})$ of three types: (i) spatial features $\phi_s(\mathcal{G})$ encoding spatial relations, (ii) temporal features $\phi_t(\mathcal{G})$ associated with plan cost, and (iii) physical quantities $\phi_p(\mathcal{G})$ produced during human interactions with scenes.

Data pre-processing is illustrated in Figure 3.6(a)-(c). Given a reconstructed watertight scene, we remove the ground plane by setting a 0.05m depth threshold and projecting it down onto a planar map. 3D objects in the scene are first segmented into primitives [AFS06]

and then grouped into object segments as in [ZZY15, ZZY13]. Some manual labeling and processing is needed for certain cluttered scenes. Finally, a semantic label is manually assigned to each object; *e.g.*, a desk with a monitor, a door, *etc.*.

Spatial features $\phi_s(\mathcal{G})$ are defined as human-object / object-object relative distances and orientations as shown in Figure 3.6(d)(e). For each object, the geometric center is obtained by averaging over all the vertices. The human head position and orientation is acquired with the Kinect.

Temporal features $\phi_t(\mathcal{G})$ are defined as the plan cost from a given initial position to a goal position. To simplify the problem, we project the 3D scene down onto a planar map. We build a binary obstacle map where the free spaces devoid of objects have unit costs, whereas the spaces occupied by objects have infinite costs. We use a 2D PRM planner to calculate the costs using 2D human positions and head orientations. Thus the planner constructs a probabilistic roadmap to approximate the possible motions. Finally, the optimal path is obtained using Dijkstra’s shortest path algorithm [Dij59]. Figure 3.6(f)–(h) show two solutions using different numbers of nodes in the planner graph.

Physical quantities $\phi_p(\mathcal{G})$ produced by people interacting with scenes are computed using the FEM. Currently, we consider only the forces and pressures acting on 14 body parts of the tetrahedralized human model, as shown in Figure 3.2(c). The net force on each body part is obtained by summing up the forces at all its vertices. The net force divided by the number of contributing vertices yields the local pressure. Figure 3.3(d) illustrates a force heatmap for sitting.

3.4.2 Learning Human Utilities

The goal in the learning phase is to find the proper coefficient vector ω of the feature space $\phi(\mathcal{G})$ that best separates the positive examples of people interacting with the scenes from the negative examples.

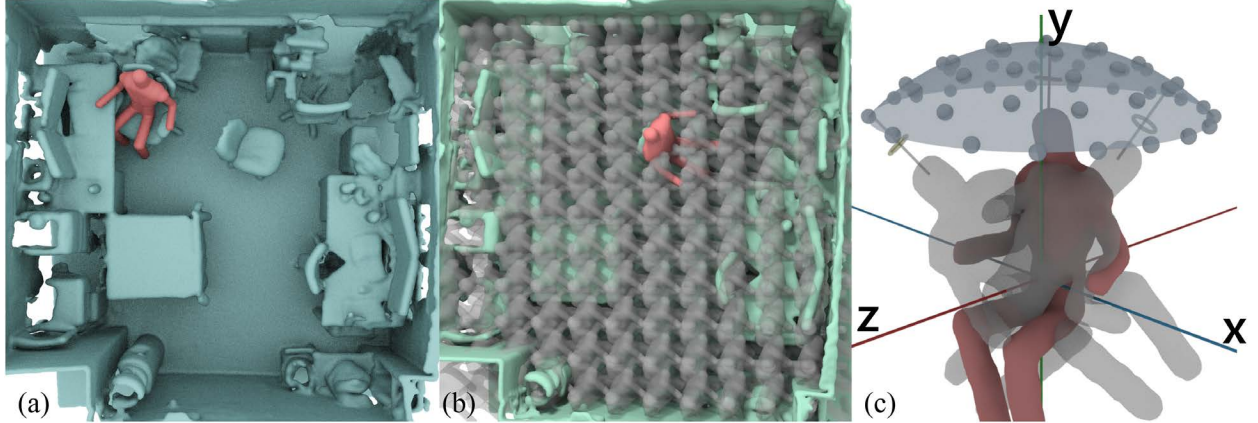


Figure 3.7: In the learning phase, based on rational choice theory, we assume that the observed demonstration is optimal, and therefore regard it a positive example. (a) In this example, a person is sitting on an armchair facing a desk with a monitor. The learning algorithm then imagines different configurations $\{\mathcal{G}_i\}$ in the solution space by initializing with different human poses P_a , (b) translations T_b , and (c) orientations O_c . The imagined randomly generated configurations $\{\mathcal{G}_i\}$ are regarded negative examples. In the inference phase, the inference algorithm performs the same sampling process (b)(c), and finds the optimal configuration \mathcal{G}^* with the highest score.

Rational Choice Assumption: We assume that in interacting with a 3D scene, the *observed* person makes near-optimal choices to minimize the cost of certain tasks. This is known as rational choice theory [Bec74, BE08, HS08, Loh08]. More concretely, the person tries to optimize one or more of the following factors: (i) the human-object and object-object orientations and distances defined as $\phi_s(\mathcal{G})$, (ii) the plan cost from the current position to a goal position $\phi_t(\mathcal{G})$, and (iii) the physical quantities $\phi_p(\mathcal{G})$ that quantify the comfortability of interactions with the scenes.

In accordance with rational choice theory, for an observed person choosing an object (*e.g.*, an armchair) on which to sit, their choice \mathcal{G}^* is assumed to be optimal; hence, this is regarded a positive example. If we *imagine* the same person making random choices $\{\mathcal{G}_i\}$ by randomly sitting on other objects (*e.g.*, the ground), the rational choice assumption implies that the costs of the imagined configurations $\{\mathcal{G}_i\}$ should be higher; hence, these should be

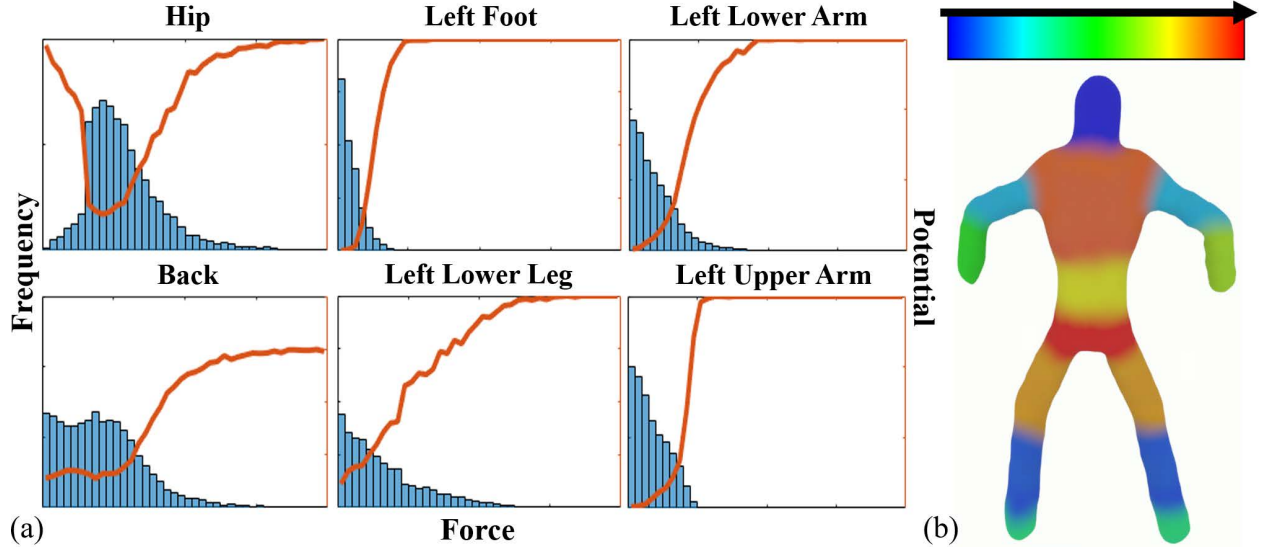


Figure 3.8: (a) The final force histograms of 6 (out of 14) body parts. The x axis indicates the magnitudes of the forces, the y axis their frequencies and potential energy. Histogram areas reflect the number of cases with non-zero forces. (b) The average forces of each body part normalized and remapped to a T pose.

regarded negative examples.

Let us consider a simplified scenario as an example: Suppose the ground-truth factors that best explain the observed demonstration are that the object is comfortable to sit on and that it faces the blackboard. Then, other objects in the imagined configurations should fall into one of the following three categories: they (i) may be more comfortable, but have less desirable orientations relative to the blackboard, or (ii) may have better orientations with the blackboard, but be less comfortable, or (iii) may be less comfortable and have worse orientations.

To summarize, under the rational choice assumption, we consider the *observed* rational person interacting with the scenes \mathcal{G}^* a positive example, and the *imagined* random configurations $\{\mathcal{G}_i\}$ as negative examples. However, the random generated configurations $\{\mathcal{G}_i\}$ may be similar or even identical to the observed optimal configuration \mathcal{G}^* . To avoid this problem, we remove random configurations that are too similar to observed configurations before applying the learning algorithm.

Ranking function: Based on the rational choice assumption, it is natural to formulate the learning phase as a ranking problem [Joa02]—the *observed* rational person interaction \mathcal{G}^* should have lower cost than any *imagined* random configurations $\{\mathcal{G}_i\}$ with respect to the correct coefficient vector $\boldsymbol{\omega}$ of $\boldsymbol{\phi}(\mathcal{G})$, which includes spatial relations $\phi_s(\mathcal{G})$, plan cost $\phi_t(\mathcal{G})$, and physical quantities $\phi_p(\mathcal{G})$. Each coefficient ω_i reflects the importance of its corresponding feature. The ranking function is defined as

$$R(\mathcal{G}) = \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}) \rangle. \quad (3.7)$$

Learning the ranking function is equivalent to finding the coefficient vector $\boldsymbol{\omega}$ such that the maximum number of the following inequalities are satisfied:

$$\langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^*) \rangle > \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle, \quad \forall i \in \{1, 2, \dots, n\}, \quad (3.8)$$

which corresponds to the rational choice assumption that the observed person’s choice is near-optimal.

To approximate the solution to the above NP-hard problem [HSV95], we introduce non-negative slack variables ξ_i [CV95]:

$$\min \frac{1}{2} \langle \boldsymbol{\omega}, \boldsymbol{\omega} \rangle + \lambda \sum_i^n \xi_i^2, \quad \forall i \in \{1, \dots, n\} \quad (3.9)$$

$$\text{s.t. } \xi_i \geq 0, \quad \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^*) \rangle - \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle > 1 - \xi_i^2, \quad (3.10)$$

where λ is the trade-off parameter between maximizing the margin and satisfying the pairwise relative constraints.

3.4.3 Inferring the Optimal Affordance

Given a static scene, the goal in the inference phase is to find, among all the *imagined* configurations $\{\mathcal{G}_i\}$ in the solution space, the best configuration \mathcal{G}^* that receives the highest score:

$$\mathcal{G}^* = \arg \max_{\mathcal{G}_i} \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle. \quad (3.11)$$

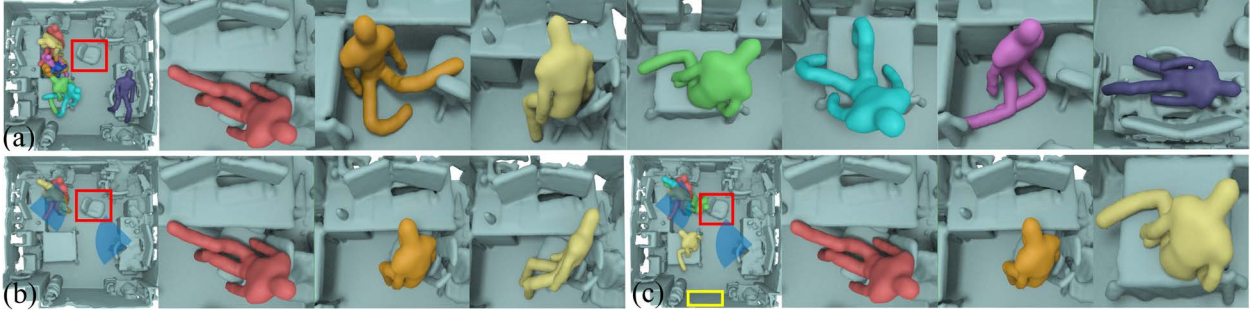


Figure 3.9: (a) The top 7 human poses using physical quantities $\phi_p(\mathcal{G})$. The algorithm seeks physically comfortable sitting poses, resulting in casual sitting styles; *e.g.*, lying on the desk. (b) Improved results after adding spatial features $\phi_s(\mathcal{G})$ to restrict the human-object relative orientations and distances. Further including temporal features $\phi_t(\mathcal{G})$ yields the most natural poses (c). The yellow bounding box indicates the door, the initial position for the path planner. Samples generated near the 3D chair labeled with a red bounding box do not produce high scores as forces apply on the arms of the person in the observed demonstration (Figure 3.7(a)). The lack of chair arms leads to low scores.

3.4.4 Sampling the Solution Space

Without observing a human interacting with the scenes, the inference algorithm must sample the solution space by imagining different configurations $\{\mathcal{G}_i\}$. The same sampling process is also required in the learning phase to generate negative examples.

We first quantize the human poses into the 7 categories shown in Figure 3.2(a). The imagined configurations of the human model are initialized with different poses P_a , translations T_b , and orientations O_c , as shown in Figure 3.7(b)(c). The tuple (P_a, T_b, O_c) specifies a unique human configuration. Given such a tuple, the simulation will impose gravity and the simulated human model will reach its rest state. The methods described in subsection 3.4.1 are then used to extract the features $\phi(\mathcal{G}_i)$.

In the learning phase, the $\phi(\mathcal{G}_i)$ are then used to learn the ranking function (Equation 3.7). In the inference phase, the extracted features are then evaluated by Equation 3.11. The configuration with the highest score is taken as the optimal configuration \mathcal{G}^* .

3.5 Experiments

3.5.1 Learning Human Utilities From Demos

A set of demonstrations of people sitting in the scene were collected using RGB-D sensors, as shown in Figure 3.7(a). The observed demonstrations were then used as positive training examples. For each 3D scene, we further generated over 4,000 different configurations \mathcal{G}_i by enumerating all poses and randomly sampling different initial human translations and rotations in the solution space, as shown in Figure 3.7(b)(c). The synthesized configurations that are similar to the human demonstrations were pruned. The remaining configurations were used as negative examples. The learning algorithm (Equation 3.7) learned the coefficient vector ω of the ranking function under three different settings: (i) physical quantities $\phi_p(\mathcal{G})$, (ii) with additional spatial relations $\phi_s(\mathcal{G})$, and (iii) with all features $\phi_p(\mathcal{G})$, $\phi_s(\mathcal{G})$, and $\phi_t(\mathcal{G})$.

Figure 3.8(a) shows the final force histograms of 6 (out of 14) body parts. Unsurprisingly when sitting, forces act on the hip in almost all cases, upper legs and lower arms also tend to be subject to relatively large magnitude forces, upper arms and heads are much less likely to interact with the scene, and the feet contact the scene in many cases, but with overall small force magnitudes. The heat map of the average forces acting on each human body part over all the collected human sitting activities is shown in Figure 3.8(b).

3.5.2 Inferring Optimal Affordance in Static Scenes

Next, we tested the learned models on our dataset as well as on prior 3D datasets [SCH14, CZK15] in three different scenarios: (i) canonical scenarios with chair-shaped objects, (ii) cluttered scenarios with severe object overlaps, and (iii) novel scenarios extremely different from the training data.

The first testing was done in the same scene as the training. Figure 3.9 shows examples of the top ranked human poses. Although using physical quantities $\phi_p(\mathcal{G})$ produced physically plausible sitting poses (Figure 3.9(a)), some of the results do not look like sitting poses (*e.g.*, lying poses and upside-down poses). Such diverse results are caused by the lack of spatial

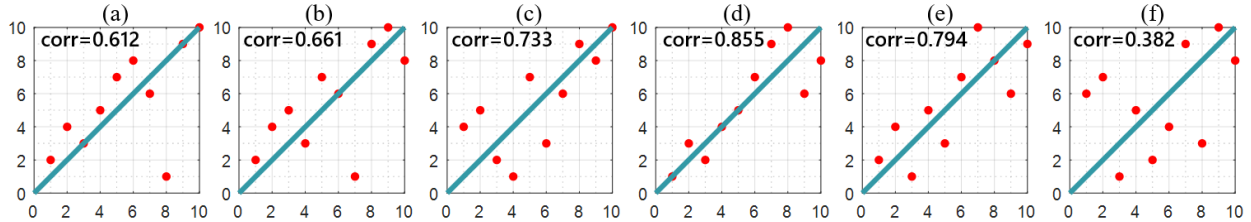


Figure 3.10: Correlations of the ranking by human subjects (x -axis) and our system’s output (y -axis). The closer the plotted points fall to the diagonal lines the better our proposed method matches the performance of the human subjects. Plots (a)–(e) correspond to Figure 3.11(a)–(e). Plot (f) corresponds to Figure 3.9(c).

and temporal constraints.

Including the spatial features $\phi_s(\mathcal{G})$, the relative orientations and distances between the human model and objects in the scene, improved the results, as shown in Figure 3.9(b). Intuitively, the top poses become more natural because they share similar human attentions and social goals to those in the observed demonstrations. For the case shown in Figure 3.9, the relative orientation between the human model and the desk with monitor prunes the configurations for which the human poses are not facing towards the monitor. The laying poses and upside-down poses are also pruned.

Integrating the temporal features $\phi_t(\mathcal{G})$ also takes into consideration the plan cost, which prunes the poses with large plan cost differences compared to the observed person demonstrations. Note that the plan cost used in temporal features enables our system to output a dynamic moving sequence, which extends the static sitting poses in previous work.

Additional results including canonical, cluttered, and novel scenarios from our dataset and other datasets [SCH14, XOT13, CZK15, ZMK13] are shown in Figure 3.11.

Evaluations: We asked 4 subjects to rank the highest-scored sitting poses. Figure 3.10 plots the correlations between their rankings and our system’s output.

3.6 Discussion and Future Work

The current stream of studies on object affordance [DFL12, FDG14, WZZ13, GGV11, GSE11, JS13b, KCG14, SCH14, ZZZ15] have attracted increasing interest on geometry-based methods, which offer more generalization power than the prevailing appearance-based machine learning approach. We have taken a step further by inferring the invisible physical quantities and learning human utilities based on rational human behaviors and choices observed in videos. Physics-based simulation is more general than geometric compatibility, as suggested by the various “lazy/casual seated poses” that are typically not observed in public videos. We argue that human utilities provide a deeper and finer-grained account for object affordance as well as for human behaviors. Incorporating spatial context features, temporal plan costs, and physical quantities computed during simulated human-object interactions, we demonstrated that our framework is general enough to handle novel cases using models trained from canonical cases.

Our current work has several limitations that we will address in future research: First, we have assumed a rigid scene. We shall consider various material properties of objects and allow two-way causal interactions between the objects and human models. This promises to enable deeper scene understanding with the help of more sophisticated hierarchical task planners. Second, currently we model the anatomically complex human body simply as a homogeneous elastodynamic material. We believe that a more realistic biomechanical human model with articulated bones actuated by muscles surrounded by other soft tissues (see, *e.g.*, [LST09, LPK14]) could enable our framework to yield more refined solutions. Optimal motor controllers could also be employed within the human simulation to support fine-grained motor planning, thus going beyond task planning, although this will increase computational complexity.

By solving these problems, we will be a step closer to consolidating several different research streams and associated methods in vision, graphics, cognition, and robotics.

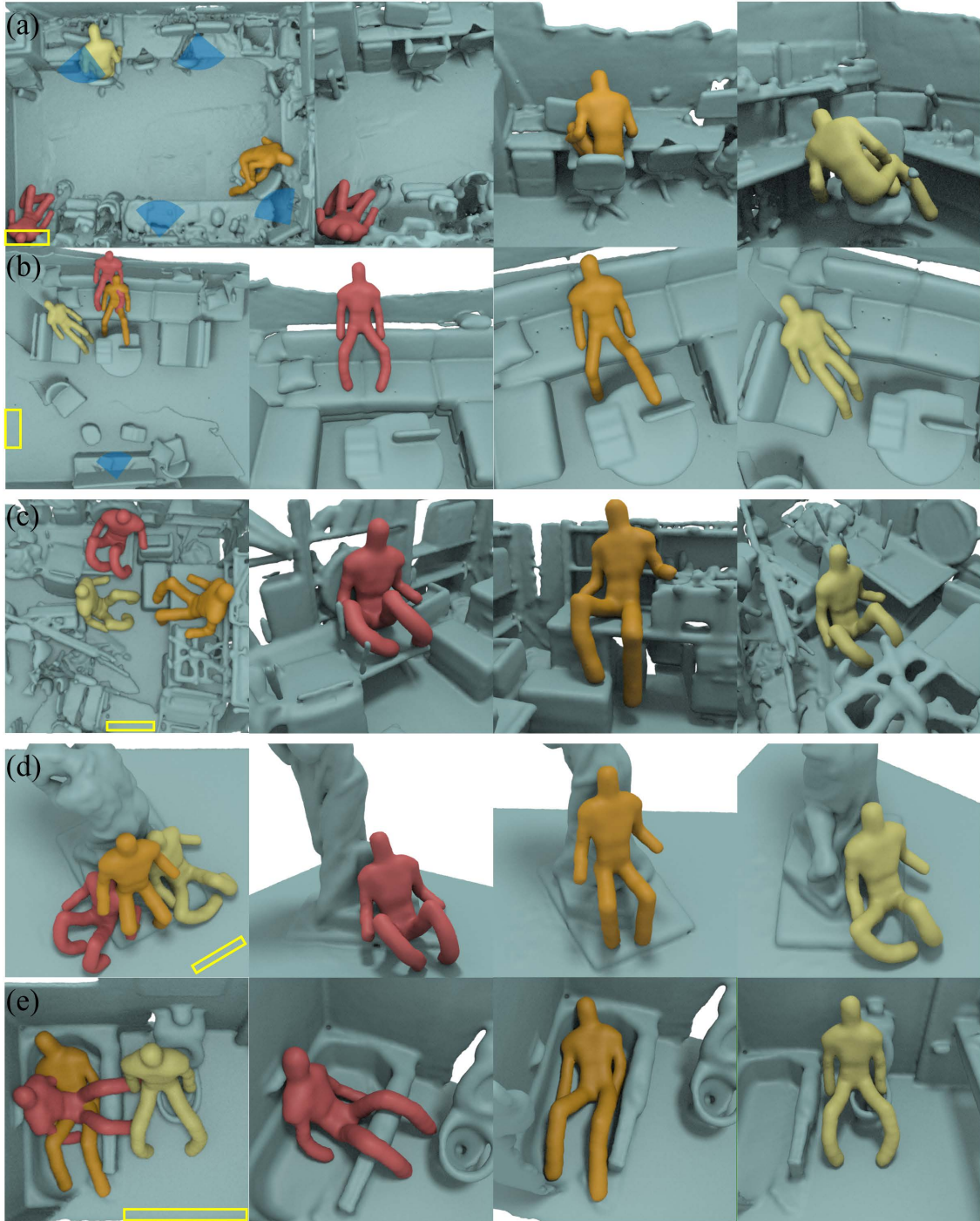


Figure 3.11: Top 3 poses in (a)(b) canonical scenarios, (c) cluttered scenarios, and (d)(e) novel scenarios. All the features $\phi(\mathcal{G})$ are used in (a) and (b). Both physical quantities $\phi_p(\mathcal{G})$ and plan costs $\phi_t(\mathcal{G})$ are used in (c)–(e). The initial position for the path planner is indicated by the yellow bounding box.

CHAPTER 4

Learning Intangible Affordance

The notion of object affordance has become an important topic of modeling human context and human-object interactions in scene understanding. As man-made objects in human living spaces are primarily designed to serve functions to fulfill daily human activities, some researchers [Gib77, Gib82, Min88, Gib14, Nel74, CCL99] argue that the recognition of affordance and function of an object assists the categorization of objects, events and places. Instead of using conventional appearance-based approaches, the latest studies on affordance reason about geometry compatibility and forces exerted during the interactions. In particular, an “affordance detector” was introduced by Grabner *et al.* [GGV11] to recognize chairs by fitting typical human poses to object candidates in a 3D scene. More recently, Zhu and Jiang [ZJZ16] devised an algorithm, through physics-based simulation using the finite element method, to infer the invisible forces/pressures on various human body parts, thereby achieving better generalization to learning human utilities in unseen objects and scenes.

In this chapter, we propose to go beyond modeling the *direct* and *short-term* human interaction with individual objects. Through accurately simulating thermodynamics and air fluid dynamics, our method can infer indoor room temperature distribution and air flow dynamics at arbitrary time and locations, thus establishing a form of *indirect* and *long-term* affordance. Unlike chairs in a sitting scenario, the objects (heating/cooling sources) that provide affordance do not directly interact with a person. Instead, the air in a room serves as an *invisible* medium to pass the affordance from an object to a person. We coin this new form of affordance as *intangible affordance*.

By observing people daily activities in indoor scenes with the labelled ground truth of the heater/cooler, we can unveil the human preferences of temperature and velocity through

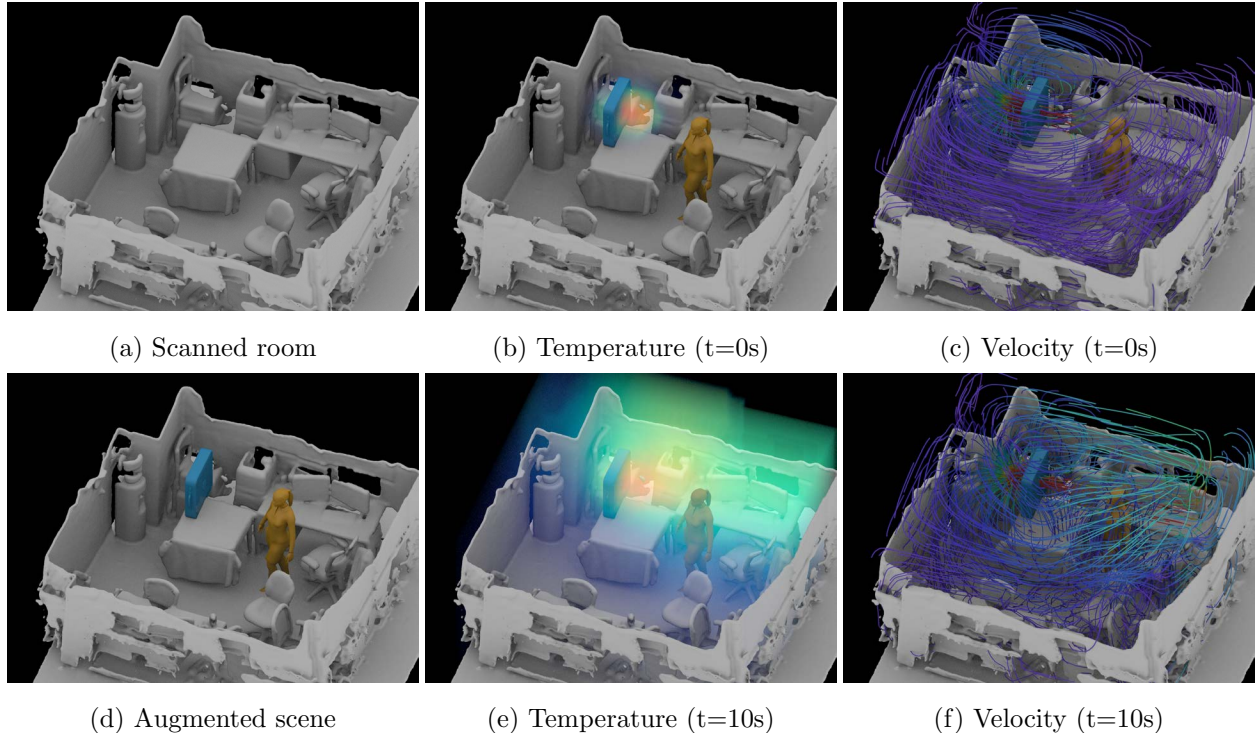


Figure 4.1: Given (a) a room scanned and reconstructed by a Kinect sensor, the proposed framework is capable of (d) imagining human activities sampled from a learned grammar, as well as sampling a plausible location and orientation of a heater/cooler. In such settings, the heater/cooler adjusts the (b)(e) temperature and (c)(f) velocity field based on human preferences without directly contacting a person, forming an *intangible affordance*.

simulating the thermodynamics and air fluid dynamics. Thus, our system can discover the indirect and long-term interactions between human and the heater/cooler. In inference, given only a static 3D scene, our system is capable of unfolding long-range human activities from a learned grammar, which in turn helps to infer proper configurations of the room layout.

4.1 Related Work

Affordance: The concept of affordance was first introduced by Gibson [Gib77], and later brought into computer vision by Ho [Ho87]. By observing people’s interactions with 3D scenes [DFL12, FDG14], researchers started to model affordance using geometry cues by fitting discretized poses [GGV11, GSE11, JX13] or poses in continuous spaces [KCG14].

Later, more attention was given to predicting the action map [SCH14] and affordance [HRB11, FPB06, RT16] given a 2D or 3D image. Object-level affordance in the context of tool-uses has been studied in [ZFF14, ZZC15, MKF14]. However, these types of affordance are all short-time with direct contact. Another closely related topic is inferring stability [JGS13, ZZY14] and containment relations [LZZ15, YDY15].

Fluid and Heat Simulation: Fluid dynamics such as liquid splashing and smoke motion is one of the major topics in computer graphics. We refer to Bridson [Bri15] for a recent survey. Many existing methods use a regular Cartesian grid due to its natural Eulerian advantages in solving the governing equations [Sta99, FF01, FSJ01, NFJ02]. Hybrid approaches are also popular, as they leverage the strengths of both particle and grid for reducing dissipation [ZB05, JSS15, ZBG15]. More recently, Chern *et al.* solved an incompressible Schrödinger's equation for simulating smoke [CKP16]. A lot of effort has been spent on preserving vorticity, reducing numerical dissipation, and improving boundary treatment for smoke simulation [SRF05, HKL15, BBB07, ABO16]. Beyond visual simulation of smoke, predicting thermodynamic air flow with computational fluid dynamics is also used for urban modeling [GAB17] and biosafety [Kol06]. In this work, we simulate indoor air flows with heat and mass transport by adopting the grid-based Eulerian approach with semi-Lagrangian advection [Sta99] for its unconditional stability and ease of implementation.

Contributions

This work makes three contributions:

1. We propose the notion of intangible affordance, which is a form of indirect and long-term interactions between an agent and objects. Inferring intangible affordance is a challenging problem as the algorithm needs to sample/imagine unobservable dynamic human interactions within scenes.
2. By simulating thermodynamic air flow given a scene geometry, our system uncovers the hidden motion of air dynamics and heat transport, thus unveiling the indirect human

interactions with scenes. To our knowledge, this is the first work to adopt numerical simulation of airflow and heat transport for the human utility of temperature in indoor scenes, enabling a deeper understanding of human activities for future human-robot interactions.

3. We address the challenge of long-term interactions by incorporating a stochastic grammar of human activity as well as a conventional robotics planning engine, which is capable of sampling a detailed action sequence in addition to motion trajectories.
4. We propose a volumetric scene completion pipeline, resulting in a voxelized 3D scene free from hollow regions or holes. Voxelized structures generated with our method enable robust geometric properties for the numerical discretization of the governing fluid equations.

4.2 Representation

We represent human activities in a video by a spatiotemporal parse graph \mathcal{G} . The spatiotemporal components are fully observable during the training, in which the intangible affordance is manually labelled, requiring forward simulation to uncover the human preferences.

Spatial Entities and Relations: At each frame, a static scene includes i) objects segmented from 3D scenes and human poses extracted from the current frame, and ii) the relations among entities. Spatial entities consist of i) human skeletons, and ii) objects segmented from 3D scenes, *e.g.*, tables, monitors, chairs, *etc.*. The human skeleton is extracted using the Kinect sensor [SSK13], and further aligned and clustered into 10 actions (Figure 4.4). 3D objects are first segmented by IAMs [BLC00] using a region growing scheme [PVB08]. Over-segmented fragments are then manually merged as shown in Figure 4.6b. Spatial relations include human-objects relations and object-object relations. The relative orientations and distances between each pair form the spatial relations.



(a) input 3D scene (b) voxelized scene (low/high resolution) (c) segmentation (d) scanline-fill diff

Figure 4.2: Given (a) an input 3D scene, we (b) voxelize the scene, and (c) reconstruct volumetric scene by jointly reasoning about geometry and physics. A scanline fill method is further applied to ensure there is no hollow regions. (d) The difference after applying scanline fill.

Human Activity and Cost in Time: To properly model the human interactions with objects developed in time, a plan cost is introduced into the proposed framework. We use the *additional* cost by comparing it before and after adding a potential heater/cooler. In addition, a human activity grammar is adopted to encode the activities frequency.

Intangible Affordance: To model intangible affordance, the temperature and velocity fields in the scene are simulated and extracted to account for the invisible, long-term, indirect, non-contact human interactions with the scene.

4.3 Estimating Temperature and Velocity Field

4.3.1 Reconstruction of Volumetric Scene

To estimate the temperature and velocity fields in a 3D scene, a volumetric representation of the scene and a proper volumetric completion pipeline are needed.

Reconstruction of 3D Point Cloud: 3D scenes are reconstructed using a real-time SLAM system [LM14] with appearance-based methods for loop closure [LM13, LM11]. A short-range RGB-D Primesense Sensor is used for its detailed and less cleaner stream compared to the Kinect sensor.

Mesh Constructions and Refinements: A 3D point cloud is scattered and refined with a fast Poisson-disk sampler using dart-throwing [DH06]. Meshes are constructed with Delaunay triangulation [Del34], resulting in a set of nearly equilateral triangles. Walls and floor are fitted to a set of planes with PCA [Jol02]. The ceiling is added at the proper height to complete a sealed mesh.

Volumetric Representation: To enable the simulation, we further convert the 2.5D surface into a 3D volumetric representation. A grid-based method is adopted (Figure 4.2b). We divide the 3D scene into a grid of voxels. A voxel is automatically labeled *occupied* if any point from the scanned point cloud falls into the boundary of the voxel; otherwise *empty*.

Volumetric Completion: Hollow regions in the volumetric representation leads to singular systems in the simulation and causes numerical difficulty. In the literature, the methods for volumetric scene completion can be categorized into two streams. i) object-level shape completion and fitting, and ii) scene-level geometric and physics reasoning.

- Object-level volumetric completion methods usually exploit geometry symmetry [KMY12, MPM14, THT16], directly fit similar 3D instances by their shapes [LF10, NXS12, SX14, GAG15, RGT15], indirectly approximate 3D shapes by primitives [LFU13, JX13], or train forest-based models [FMJ16] and deep neural network models [WSK15, VDR16, SYZ17a] based on geometric information. However, these methods require additional segmentation before they are applicable to the entire 3D scene.
- Scene-level volumetric completion methods utilize low level geometric cues by plane fitting [MMB15], jointly reason geometry and segmentation for outdoor [HZC13, BVR16] and indoor [ZZY13], or simultaneously recover semantic labels and volumetric structures from labelled datasets [SYZ17a].

In this work, we adopt the method described by Zhang *et al.* [ZZY13] and Jia *et al.* [JGS13] because it provides a segmented physically stable volumetric scene. A Manhattan structure [FCS09] is assumed, and a cluster sampling method Swendsen-Wang cut [BZ05] is applied

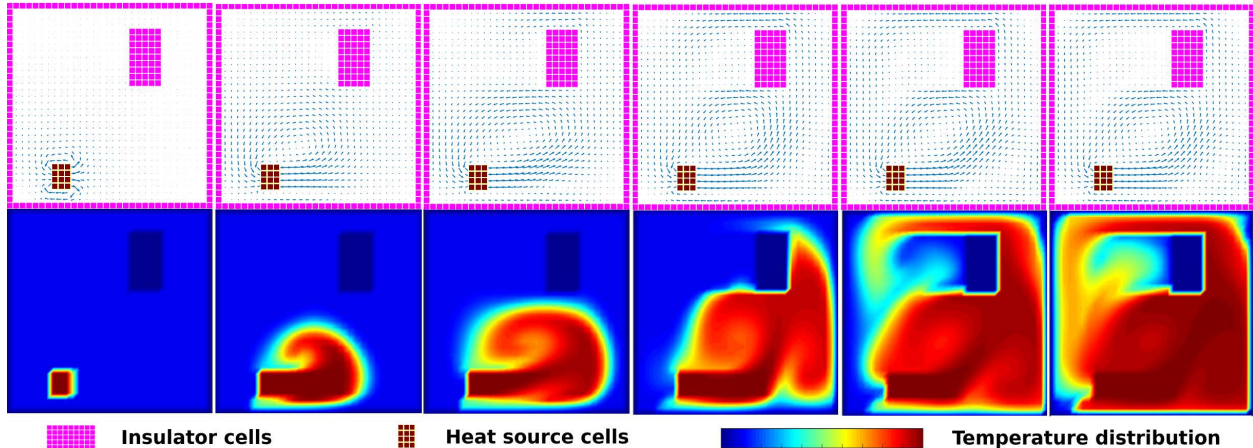


Figure 4.3: Thermodynamic air flow simulation in a 2D room. Top: air velocity field. Bottom: temperature evolution.

to segment objects in 3D. To ensure there are no hollow regions left, a 3D non-recursive scanline fill method [Hec90] is further applied. Figure 4.2 depicts the full pipeline for the reconstruction of hole-free volumetric scenes.

4.3.2 Simulating Thermodynamic Air Flow

Simulating the dynamics of gaseous phenomena has attracted a lot of attention in Computational Fluid Dynamics (CFD). While computer graphics applications often focus on the simulation of smoke due to its ubiquitousness in visual effects, we target at simulating the heat-driven air motion instead of the density-driven smoke concentration evolution.

Velocity: We model air as an inviscid and incompressible fluid. Such an assumption reduces Navier-Stokes equation to the incompressible Euler equations [Bri15]:

$$\rho \frac{D\mathbf{v}}{Dt} = -\nabla p + \mathbf{f}, \quad (4.1)$$

$$\nabla \cdot \mathbf{v} = 0, \quad (4.2)$$

where ρ is density, \mathbf{v} fluid velocity, p pressure, and \mathbf{f} external force. $\frac{D\mathbf{v}}{Dt}$ is the material derivative of the velocity; written as $\frac{D\mathbf{v}}{Dt} = \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v}$. Equation 4.2 accounts for the divergence free constraint on velocity field and implies incompressibility.

We discretize the velocity governing equations on a staggered Eulerian grid. Velocity degrees of freedom are sampled on voxel faces and pressure is discretized on cell centers. As in Zhu *et al.* [ZB05], an intermediate velocity field $\hat{\mathbf{v}}$ is assumed to split the solution procedure to an advection step and a projection step. Assuming the previous time step velocity field is \mathbf{v}^n , the advection step solves

$$\frac{\hat{\mathbf{v}} - \mathbf{v}^n}{\Delta t} = -(\mathbf{v}^n \cdot \nabla)\mathbf{v}^n + \mathbf{f}, \quad (4.3)$$

where we adopt the semi-Lagrangian advection scheme of Stam [Sta99] for its unconditional stability. The projection step comes from enforcing the incompressibility constraint on \mathbf{v}^{n+1} , the new velocity field. This results in solving a Poisson equation for pressure unknowns:

$$\nabla^2 p = \frac{\rho}{\Delta t} \nabla \cdot \hat{\mathbf{v}}. \quad (4.4)$$

Discretely, this can be written as $\mathbf{G}^T \mathbf{G} \{p\} = \frac{\rho}{\Delta t} \mathbf{G}^T \{\hat{\mathbf{v}}\}$ [BBB07], where $\{\hat{\mathbf{v}}\}$ is the vector containing all $\hat{\mathbf{v}}$ degrees of freedom, $\{p\}$ contains all pressures, \mathbf{G} is the discrete gradient operator. On the Marker And Cell (MAC) grid, this is equivalent to discretizing the Laplacian operator with the standard seven-point stencil. This system is typically symmetric positive definite and can be solved efficiently with Conjugate Gradient with incomplete Cholesky or Multigrid preconditioner [MST10]. We notice that, however, since the linear system and boundary conditions never change, it is much faster to prefactor the matrix with sparse Cholesky and perform back substitution. Furniture, heaters/coolers, walls, ground, and ceiling are all treated as Neumann boundaries. Additionally, we perform volumetric completion for filling in hollow regions to prevent empty rows in the system. After solving the pressures, velocities are updated with

$$\mathbf{v}^{n+1} = \hat{\mathbf{v}} - \frac{\Delta t}{\rho} \nabla p. \quad (4.5)$$

Temperature: Heat transfers in a room in multiple ways. For an air blowing heater/cooler, the effect mostly comes from heat convection, where temperature change is caused by the fluid motion of air particles. This corresponds to the advection of temperature:

$$\frac{\hat{T} - T^n}{\Delta t} = -(\mathbf{v}^n \cdot \nabla)T^n, \quad (4.6)$$

where \hat{T} is the updated temperature. We discretize T on cell centers and update them with semi-Lagrangian advection. Besides convection, we take into account slight heat diffusion in the air through an explicit temperature update:

$$\frac{T^{n+1} - \hat{T}}{\Delta t} = \kappa \nabla^2 T^{n+1}, \quad (4.7)$$

where κ is the diffusion coefficient. The heat equation also leads to a symmetric positive definite linear system, which we solve with sparse Cholesky factorization.

Finally, to model the temperature influence on the velocity, we take the Boussinesq approximation [FSJ01] and apply an external force to the velocity field as $\mathbf{f} = -\beta(T - T_a)\mathbf{z}$, where β is the buoyancy coefficient, T_a is the ambient temperature, $\mathbf{z} = (0, -1, 0)^T$ points to the gravity direction. This force is added to the advected velocity field $\hat{\mathbf{v}}$ to model the effect of rising hot air.

4.4 Learning Intangible Affordance

In this section, we first introduce a pipeline to learn the grammar of human activity from the training data, which serves as the input for extracting spatiotemporal features and affordance features. Next, we introduce a set of features crafted for learning intangible affordance in 3D indoor scenes. Using these features, we present our learning and inference algorithm based on rational agent theory and ranking functions.

4.4.1 Learning Grammar of Human Activity

Using the skeletons collected from Kinect sensor, we present a pipeline to build a stochastic grammar upon these raw noisy data, during which very limited manual labeling is involved. The pipeline includes four steps: i) skeleton alignment, ii) skeleton clustering as action sentences, iii) unsupervised structure learning of stochastic grammar, and iv) adding attributes.

Skeleton Alignment: The alignment of the skeleton is solved using Horns method [Hor87] to align absolute orientation based on quaternion: finding the optimal scale vector \mathbf{s} , rotation

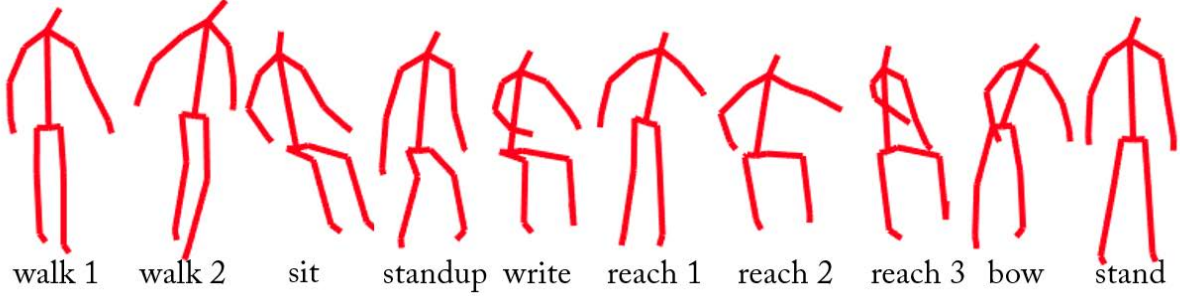


Figure 4.4: Human skeletons in the training data are clustered into 10 atomic actions.

matrix \mathbf{R} and translation vector \mathbf{t} that best map a target 3D pose A to a source 3D pose B in a least square fashion:

$$\min \sum_i w(i) \|\mathbf{sRA}(:, i) + \mathbf{t} - \mathbf{B}(:, i)\|^2, \quad (4.8)$$

where $w(i)$ is a user-specified vector, indicating significance of different joints of the 3D pose. In this work, the weights of 5 joints in human pose (left/right shoulder, left/right wrist, spine base) are set to 1, the others are set to 0.

Skeleton Clustering and Atomic Actions: Atomic actions [PSY13] or movement primitives [SPN05, PDP13] refer to elemental actions that cannot be further decomposed, such as sitting and standing. Instead of labeling all the actions in the training data, we perform unsupervised clustering on the skeleton data, resulting in a set of atomic actions. By only labeling very limited number of atomic actions, the action sequence in each video is automatically labeled at each frame.

A hierarchical agglomerative clustering scheme is applied: each skeleton starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy [War63]. Ward’s minimum variance method [War63] is chosen to join the two clusters A and B that minimizes the increase in the sum of squared errors:

$$I_{AB} = \frac{n_A n_B}{n_A + n_B} (\bar{\mathbf{a}} - \bar{\mathbf{b}})' (\bar{\mathbf{a}} - \bar{\mathbf{b}}), \quad (4.9)$$

where $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ represent the centroids of the cluster A and B , respectively. Results are shown in Figure 4.4.

By comparing the distances between the centroids of each voxelized objects (see subsection 4.4.2) and the centroids of human wrists, a pair of action-object label $\psi(A, O)$ is automatically generated at each frame by choosing the object with the minimal distance. Since the data is noisy, the number of different action-object pairs generated was originally over 60. We manually pruned some unreasonable pairs, *e.g.*, standing on a keyboard, making the final number of action-object pairs 26. The sequence of labeled action-object pairs in each video is treated as a sentence for grammar learning. To simplify the grammar structure, any two consecutive actions will be merged if they share the same action label.

Stochastic And-Or Grammar: An Attributed And-Or Graph [ZM07] is chosen as our representation for stochastic grammar. As a compact representation, it can easily represent repeated patterns of action and event data. Formally, an And-Or Graph is defined as a 5-element tuple: $\mathcal{G} = \langle S, T, N, R, P \rangle$, where S is the root node of the grammar, T the set of the terminal nodes, N the set of non-terminal nodes, R the set of the production rules, P the set of probabilities assigned to the grammar rules. In such representation, an And rule decomposes a pattern into a configuration of non-overlapped sub-patterns, whereas an Or rule indicates alternative configurations of a composite pattern.

Unsupervised Structure Learning: We adopt an unsupervised structure learning method for stochastic And-Or Grammar introduced by Tu *et al.* [TPZ13]. Starting from a flat grammar where each Or fragment denotes an individual training sample, the method iteratively induces compositions and reconfiguration, making the initial grammar more compact by mining and reducing the repeated patterns. The problem can be formulated as optimizing a posterior

$$P(\mathcal{G}|\Omega) \propto p(\mathcal{G})P(\Omega|\mathcal{G}) = \frac{1}{Z} e^{-\alpha\|\mathcal{G}\|} \prod_{a_i \in \Omega} P(a_i|\mathcal{G}), \quad (4.10)$$

where i is the i -th vertex of human pose, \mathcal{G} the grammar, $\Omega = \{a_i\} = \{\psi_i(A_i, O_i)\}$ the set of training action-object labels, Z the normalization factor of the prior, α a constant, and $\|\mathcal{G}\|$ the size of the grammar.

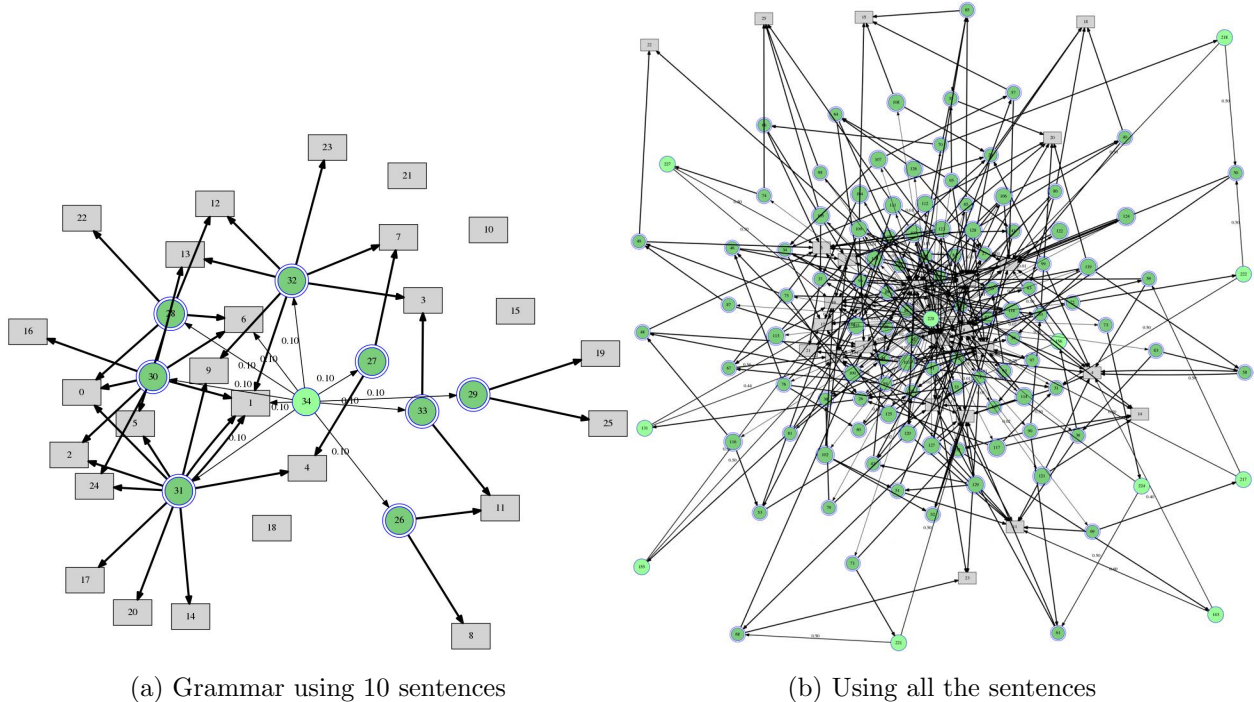


Figure 4.5: The grammar of human activity. Using labeled action-objects pairs from 94 short videos, the And-Or structures of the grammar with 228 nodes is learned in an unsupervised fashion, which represents all the actions from the training video in a compact way. The graph visualized here does not show the structure in a temporal order, but only shows the decomposition relations. Multiple edges may exist between any two nodes, indicating the decomposition happens more than once. Grey rectangle represents the 26 action-object pairs as terminal nodes, light and dark green circle refers to the And and Or nodes, respectively.

A matching pursuit [MZ93] scheme is applied, where we jointly learn And-Or fragments instead of And fragment and Or fragment individually. In each iteration, we find the And-Or fragments that lead to the highest gain in the posterior probability of the grammar.

Finding the highest posterior is equivalent to maximizing the production of the likelihood gain and the prior gain. The likelihood gain can be factorized as the product of two coherence measures [MO04]: i) n-gram matrix of the And-Or fragment, and ii) coherence of the context matrix:

$$\frac{P(D|\mathcal{G}_{t+1})}{P(D|\mathcal{G}_t)} = \frac{\prod_{i=1}^n \|r_i(a_{ij})\|^{\|r_i(a_{ij})\|}}{\|r\|^{n\|r\|}} \times \frac{\prod_c (\sum_e M(e, c)^{\sum_e M(e, c)})}{\prod_{e, c} M(e, c)^{M(e, c)}}, \quad (4.11)$$

where G_t and G_{t+1} are the grammars before and after learning from an And-Or fragment,

$r_i(a_{ij})$ the subset of reductions where a_{ij} of i -th node was reduced, e the sum or product over all possible configurations generated by And-Or fragments, and c the product over all the contexts represented by the context matrix M . Since the prior probability of the grammar is determined by the grammar size, the prior gain is defined as:

$$\frac{P(D|\mathcal{G}_{t+1})}{P(D|\mathcal{G}_t)} = e^{-\alpha(\|\mathcal{G}_{t+1}\| - \|\mathcal{G}_t\|)}. \quad (4.12)$$

The learned grammar structure is visualized in Figure 4.5. In total 228 nodes are used to encode all 94 videos of training data.

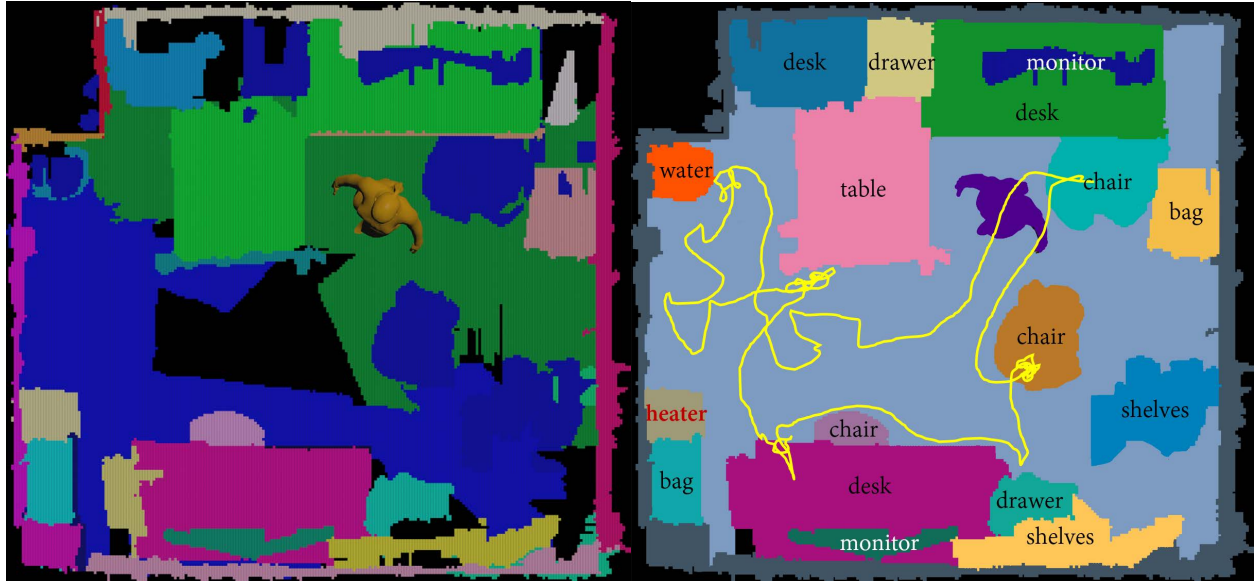
Attributed And-Or Grammar: On top of the And-Or structures, attributes [PZ15] are further integrated into terminal nodes in order to account for the motion speed and duration of human activities. To simplify the complexity of the model, we learn the distribution of the motion speed and duration independently from the structure learning. Gaussian Mixture Models are applied to model the distributions.

4.4.2 Extracting Features

We crafted three types of features to account for different perspectives of the intangible affordance: i) spatial features $\phi_S(\mathcal{G})$ characterizing the spatial relations among entities, ii) temporal features $\phi_T(\mathcal{G})$ indicating the additional planning cost introduced by the heater/cooler in the scene, and iii) affordance features $\phi_A(\mathcal{G})$ describing the human preferences in terms of air velocity and room temperature.

Spatial feature $\phi_S(\mathcal{G})$ characterize the spatial relations among humans, heaters/coolers and other objects.

- *Spatial Entities.* The over-segmented voxels are fit into 3D primitives [AFS06] (Figure 4.6a), and merged together or branched into sub-parts. Objects semantic label is manually assigned. Locations of the objects are computed by averaging over all the vertices. The location and orientation of a person is obtained using the head location and orientation acquired by Kinect. Result is shown in Figure 4.6b.



(a) Over-segmented voxels

(b) Spatial entities and relations

Figure 4.6: (a) After fitting a set of over-segmented voxels into 3D primitives, (b) the segmented voxels are merged together or branched into sub-parts using the scene graph structure described in Open Inventor [Wer94]. The head position is connected frame by frame, illustrated in yellow curve.

- *Spatial Relations.* Pair-wise spatial features $\phi_S(\mathcal{G})$ are extracted by computing the relative distances and orientation between human and any objects, as well as between the heater/cooler and any other objects.

Temporal feature $\phi_T(\mathcal{G})$ indicates the *additional* planning cost by comparing before and after placing the heater/cooler into a 3D scene. In this work, we project 3D voxels onto a 2D plane. All the objects above the floor is treated as obstacle, forming a binary obstacle map (Figure 4.7). A probabilistic roadmap [KSL96] is built by sampling a set of random placement of nodes and connecting k-nearest neighbors [Alt92] within a given distance range. Finally, the optimal path is obtained using Dijkstra’s shortest path algorithm [Dij59]. We assign unit costs to the free spaces devoid of objects in the obstacle map, and infinite costs to the occupied spaces. The final cost of a plan is calculated by accumulating the costs along the path.



(a) Before adding a heater/cooler

(b) After adding a heater/cooler

Figure 4.7: Additional planning cost introduced by adding a potential location of the heater/cooler. (a) Before adding the heater/cooler, the path between the start location and the end location is shorter than (b) after adding the heater/cooler as shown in the red rectangle.

Affordance feature $\phi_A(\mathcal{G})$ describes the human preferences in terms of air velocity and room temperature. In our staggered discretization of the fluid simulation, different velocity components are stored on cell faces. We average them to cell centers to get a net velocity. Temperature values are already stored at cell centers. We intersect the human geometry with the simulation grid, and extract all velocity and temperature values in the resulting cells. Velocity directions, velocity magnitudes and temperature values are added to feature vector.

4.4.3 Learning and Inference

We formulate the learning of intangible affordance as a ranking problem [Joa02], and assume the demonstrations in the videos are positive examples. This is known as rational choice theory [Loh08, HS08, Bec74, BE08] in economics and social science. This assumption is consistent with psychology studies [GBK02], indicating humans are intend to teach their children through successful demonstrations, unlike other animals.

In accordance to rational choice theory, the goal of the learning is to determine the best weights ω of the feature vector $\phi(\mathcal{G}) = (\phi_S(\mathcal{G}), \phi_T(\mathcal{G}), \phi_A(\mathcal{G}))$ that separates the observed positive examples from the the unobserved negative examples. Formally, the *observed* examples \mathcal{G}^* should have lower cost than any *imagined* random configurations $\{\mathcal{G}\}$ with respect to the proper coefficient vector ω of $\phi(\mathcal{G})$. The negative examples are randomly generated/imagined examples that are different from the observed human demonstration. Details are provided in subsection 4.4.4.

The ranking function is defined as:

$$R(\mathcal{G}) = \langle \omega, \phi(\mathcal{G}) \rangle. \quad (4.13)$$

Learning the above ranking function is equivalent to finding the coefficient vector ω such that the maximum number of the following inequalities hold:

$$\langle \omega, \phi(\mathcal{G}^*) \rangle > \langle \omega, \phi(\mathcal{G}_i) \rangle, \quad \forall i \in \{1, 2, \dots, n\}. \quad (4.14)$$

To approximate the solution to the above NP-hard problem [HSV95], non-negative slack variables ξ_i were introduced [CV95]:

$$\min \frac{1}{2} \langle \omega, \omega \rangle + \lambda \sum_i^n \xi_i^2, \quad \forall i \in \{1, \dots, n\} \quad (4.15)$$

$$\text{s.t. } \xi_i \geq 0, \quad \langle \omega, \phi(\mathcal{G}^*) \rangle - \langle \omega, \phi(\mathcal{G}_i) \rangle > 1 - \xi_i^2, \quad (4.16)$$

where λ is the trade-off parameter between maximizing the margin and satisfying the pairwise relative constraints. Given a static scene, the goal in the inference is to find the best configuration \mathcal{G}^* that receives the highest score among all the imagined configurations $\{\mathcal{G}_i\}$:

$$\mathcal{G}^* = \arg \max_{\mathcal{G}_i} \langle \omega, \phi(\mathcal{G}_i) \rangle. \quad (4.17)$$

4.4.4 Sampling Solution Spaces

Since we only observe the positive examples during the learning, negative examples are needed to sample from the solution spaces. During the inference, since the only input is a static scene, the same sampling process is used to generate both human activities and candidates of the heaters/coolers.

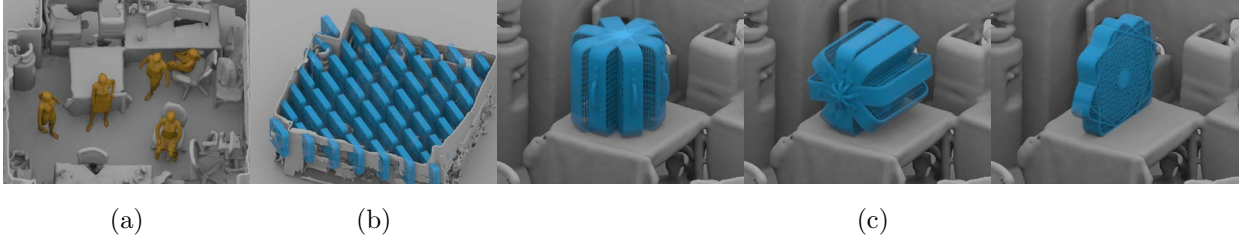


Figure 4.8: Sampling (a) human activities from temporal solution spaces, as well as the (b) location and (c) orientations of the heater/cooler from the spatial solution spaces.

Human Activities: An Earley parser [Ear70] is adopted to sample a sequence of action-object pairs. If there is no object in the current scene matched the sampled object, we will randomly choose an object in the current scene to replace the mismatched one. The same planning engine based on probabilistic roadmap is applied to generate a continuous motion trajectory. The motion speed and the duration of human activities are further sampled from the Attributed And-Or Graph to provide more realistic sampling of human activities.

Heater/Cooler: The imagined configurations of the heater/cooler include different locations and orientations. We quantize the entire 3D spaces of the scene into a set of discretized grids. Grids without any objects in its neighbors in any one of the six directions are pruned, as it cannot provide physical supports for heater/cooler to attach to.

4.5 Experiment

4.5.1 Learning Intangible Affordance

We collect Human demonstrations with ground truth labels (Figure 4.6b). These observed locations and orientations of the heaters/coolers are treated as positive examples. Additional planning cost is computed by randomly sampling other locations and orientations of the heaters/coolers. For each scene, 1000 samples with different configurations are generated and treated as negative examples. Samples that are closed to the positive examples are automatically pruned. Figure 4.9 illustrates the histogram of human preferences on temperature and net velocity. We can see that human prefers the temperature between 20 and 25 °C,

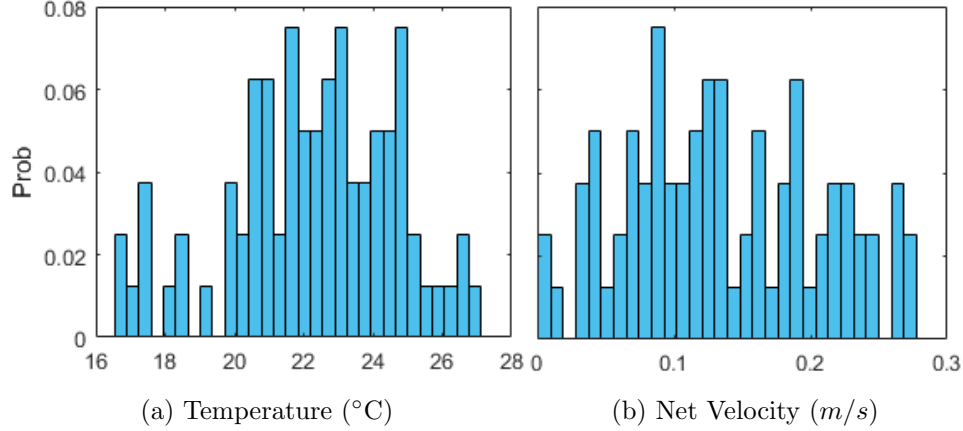


Figure 4.9: Human preferences learned from demonstrations. Human prefers (a) temperature between 20 and 25 °C, and (b) net velocity around 0.1 m/s in indoor environment.

and net velocity around 0.1 m/s .

4.5.2 Inferring Optimal Affordance in Static Scenes

Next, we test our learned model using three different setups: i) without a person, ii) with a static person, and iii) with a moving agent performing a set of human activities.

If we remove the factor of humans, the overall distribution of air flow velocity and temperature tends to be relatively uniform throughout the room in highly ranked results. For those with low scores, the fields exhibit strong anisotropy—temperature distribution has a bias in certain regions, and tends to concentrate in some corners or small parts inside the room.

When a static human pose is given, highly ranked configurations share similar velocity and temperature distributions. The learned preferences focus on spatial locations near the human. Other regions, even if reachable by different human activities, do not strongly influence the ranking results. The low score ones contain extremely high or low heat or velocity magnitudes on the human skin, therefore are less capable of providing affordance.

Finally, we consider a sequence of human activities sampled from the learned grammar in the room. In this case, top ranked configurations tend to match both velocity and temperature with human preferences in most regions, especially those reachable by human.

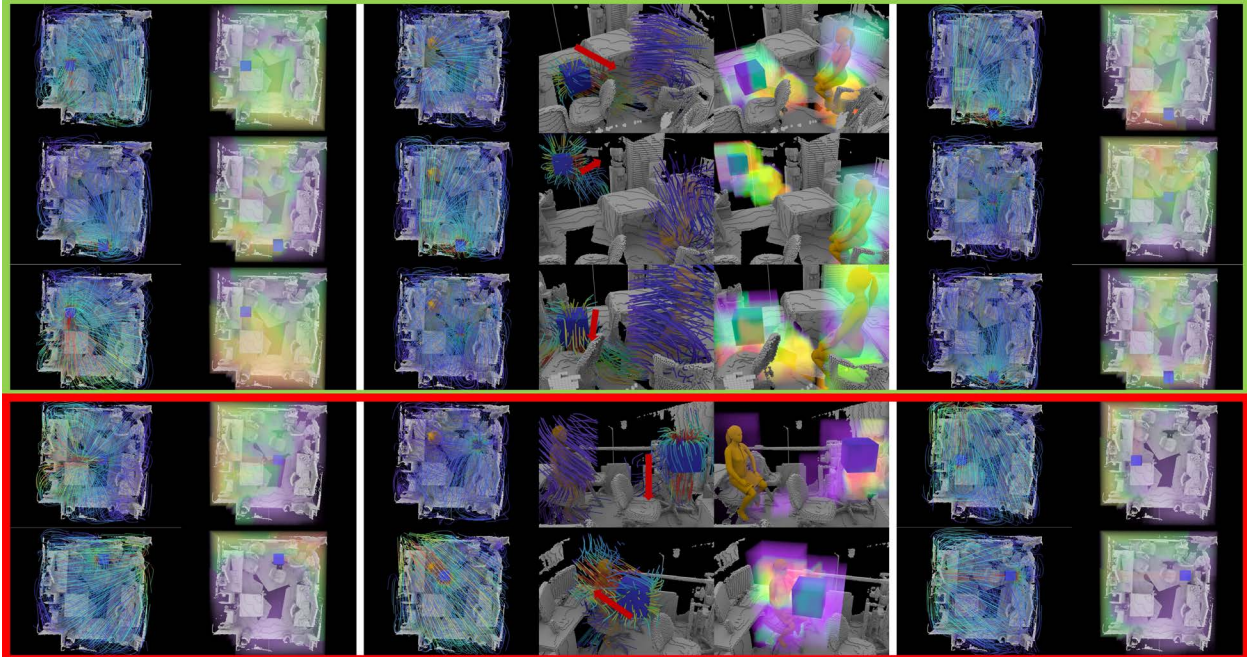


Figure 4.10: Inference results: top three (in green bounding box) and the worst two (in red bounding box) configurations. Left: If we do not consider human in a scene, the configurations receive higher ranks tend to have uniform distributed (left) velocity and (right) temperature across the entire scene. Middle: If we consider a static human, a person sitting on a chair in this case, both (left and middle) velocity and (right) temperature is tend to match the human preferences only near the location where the person sits. Right: When we consider human activities sampled from the learned grammar, the top ranked configurations tend to match both (left) velocity and (right) temperature with human preferences on most of the regions, excluding the regions where are not reachable by humans, *e.g.*, the regions near walls.

4.5.3 Evaluations

We asked 10 subjects to rank the high rank room configurations. Figure 4.11 plots the correlation between subject choices and our system’s output.

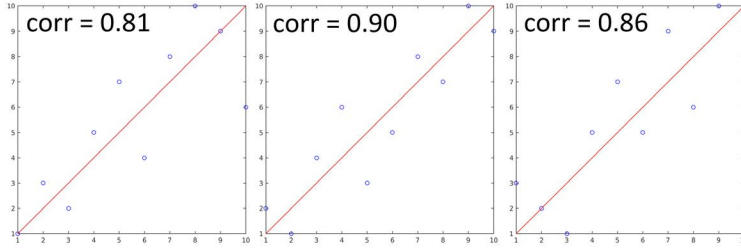


Figure 4.11: Correlation between our ranking (x-axis) vs human judgments (y-axis). Red lines denotes the perfect match.

4.6 Discussion and Future Work

By utilizing accurate physics-based simulation methods from computational fluid dynamics and computer graphics, we have taken a step further on object affordance by proposing and inferring a new form of affordance—intangible affordance. Exemplified with heater and cooler configurations, our system accurately predicts the air flow velocity field and temperature evolution in a room, thus providing a more in-depth and finer-grained account for object affordance and human behaviors that is missing in conventional scene understanding with pairwise human-object relationships.

Our system has several limitations that we will address in future research. First of all, we shall incorporate finer level task planning with more complex human activities and motion planning. Secondly, as an uncomplicated demonstration of the intangible affordance, we have excluded the change of airspeed, the heat loss due to radiation (such as those between human bodies and walls) and the influence on the air flow due to the motion of humans and objects. These are promising future directions that would potentially allow more realistic and useful learning results. We believe that using deeper features with the help of neural networks will improve the performance. A more detailed human body with different comfortability on various body parts could also be employed. By augmenting our system with these additional features, we will be a step closer to successfully applying our learning framework of intangible affordance to support many useful applications such as wind/temperature sensitive room layout design, robot/human task planning and human behavior reasoning.

CHAPTER 5

Inferring Containers and Containment Relations

In this chapter, we will present a framework to study and reason about the containment relations, following by a tracking algorithm integrating the reasoning of the containment relations to solve the challenging occlusion issues in the field of object tracking.

5.1 What is Where: Inferring Containment Relations from Videos

For many AI tasks, such as scene understanding in visual perception, task planning in robot autonomy, and symbol grounding in natural language understanding, a key problem is to infer “what is where over time”. A person may say “the pizza is in a pizza box, and the pizza box is in a fridge”. In such a description, the object locations are described in a qualitative and hierarchical way, in which *containers* play an important role in quantizing human perceptual space via *containment relations*. By *containers*, we refer to any general objects in a scene that can contain other objects, for example, fridge, mug, box, lunch bag, envelop and so on. The *containment relations* between containers and contained objects, *i.e.*, *containees*, may change over time by agents.

In this work, we propose a probabilistic approach to infer containment relations from RGB-D videos. Consider the example shown in Figure 5.1. The containers and containees are tracked in a 3D scene and highlighted in colored bounding boxes in the top panel. The inferred containment relations are constructed in the bottom panel, pointing from containees to the corresponding containers, and the numbers on edges denote the frames when the containment relations occur. It is worth noting that the containment relations are time varying, and can be changed by human actions. The presented containment relation inference

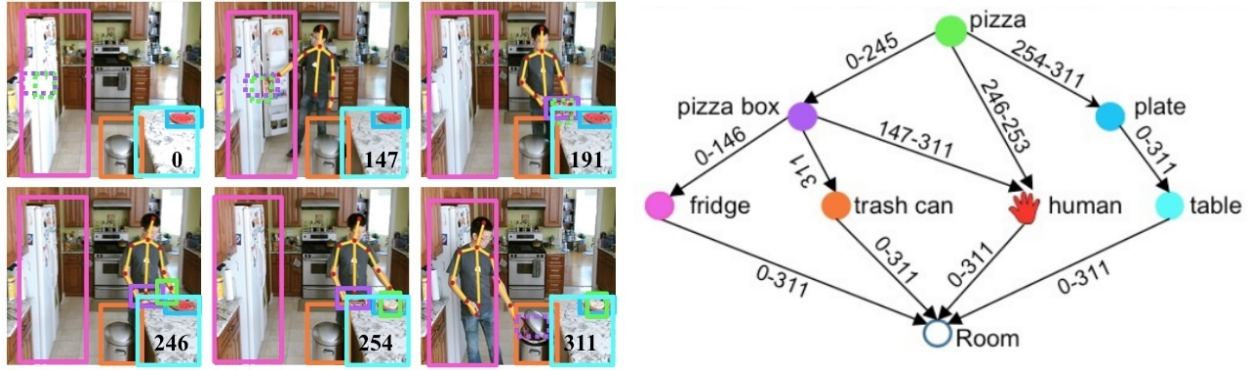


Figure 5.1: (Left) Structured, qualitative and abstract interpretation of containment relations over time in a scene. The goal is to answer “what is where over time”. (Right) The inferred containment relations. The numbers on edges denote the frames when the containment relations occur.

method is aim to address the following two tasks.

Recovering Hidden Objects with Severe Occlusions: Severe occlusions frequently happen in daily scene. Reasoning about containment relations helps to recover objects from tracking failures when objects are partially occluded or even completely unobservable. For instance, as shown in Figure 5.1, although we only observe a person taking a pizza from a pizza box at frame 246, we are able to infer that the pizza was contained by the pizza box from frame 1 to 245, during which period the pizza was unobservable. By simple commonsense that a containee shares the same position as its container when the containment relation between them holds, we are able to recover the hidden containee with severe occlusions or even without actually seeing it. This capability provides a potential solution to build a system (*e.g.*, an assistive robot) to answer “what is where over time”. For example, a robot can help to localize an object in a room if a person forgets where he or she left it.

Inferring Subtle Human Actions: Because there are self-occlusions or occlusions by other objects in a scene, subtle human actions that involve small and local movements, such as placing a phone in a bag, are difficult to detect. If a change of objects’ status (*i.e.*, a

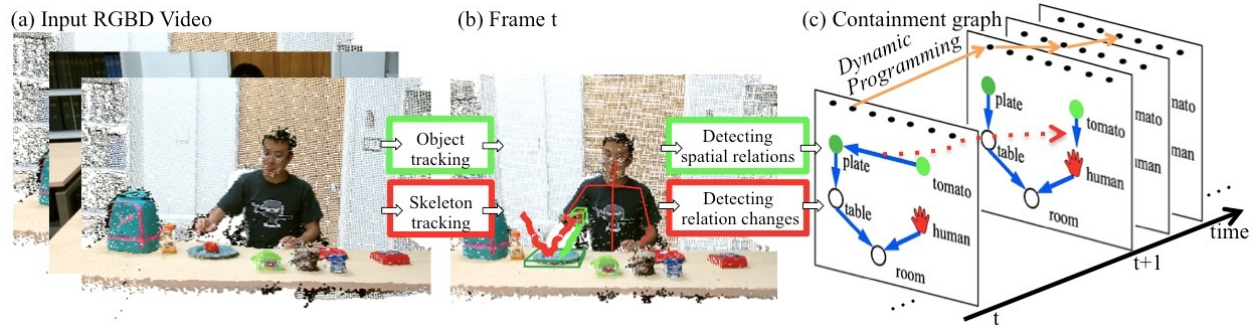


Figure 5.2: An overview of the proposed approach. (a) Given a RGB-D video, we first track objects and human skeletons in 3D space. (b) At each frame, the tracked 3D bounding boxes are used to construct containment relations, whereas tracked human skeletons are used to detect containment relation changes. (c) Across the video, a joint spatial-temporal inference method is used to find the optimal sequence of containment graphs. The containment graph sequence defines both spatial containment relations at each frame (blue edges in the graph) and temporal containment relation changes over time (changes of the blue edges, highlighted by red dashed arrows) caused by human actions.

containment relation change) is observable, it is natural to reason about that some human actions occurred. The ability of inferring human subtle actions by goals instead of observing and matching detailed action trajectories provides possibilities for a robot to understand the intentions of agents.

Figure 5.2 illustrates the framework of the proposed method. Given a RGB-D video captured by a consumer depth camera (Figure 5.2(a)), our method first tracks the objects of interest and the human skeletons (Figure 5.2(b)). Then at each frame t , the containment relations are represented by a containment graph; in time, containment relation changes are proposed based on human actions. To find the optimal interpretation across the full sequence, a dynamic programming algorithm is adopted to globally optimize both spatial and temporal space, resulting in the optimal sequence of containment graphs (Figure 5.2(c)).

This work makes three major contributions:

1. We propose a probabilistic approach to infer containment relations from videos over

time. A dynamic programming algorithm is applied to solve the ambiguities on both containment relations in space and containment relation changes in time, providing a globally optimized solution.

2. We propose a dynamic graph representation for containment relations over time (Figure 5.1). The dynamic graph quantizes 3D scene space and provides a qualitative way for tracking objects with heavy occlusions.
3. We model the containment relation changes in time by assuming that human actions are the only cause to change the containment relations. This constraint in return helps to recover hidden objects and infer time varying containment relations at each frame.

5.1.1 Related Work

Cognitive studies [HS07] has shown that infants can understand containers and containment relations as early as 3.5 months old. Strickland and Scholl [SS15] suggest that infants can detect containment before understanding occlusion. Liang *et al.* [LZZ15] evaluates human cognition of containing relations for adults through a series of experiments using physical-based simulations. In computer science, the problem of containers has been studied from various perspectives in the fields of AI, robotics and computer vision.

AI: Qualitative Spatial Representation / Reasoning (QSR) has been extensively studied in the AI community. Cohn and Hazarika [CH01] provided a survey of key ideas and results in QSR literature. Some typical methods include using ontology [HAB08, GS04], topology [GR02, Li07], metric spatial representation [Fra92, PS94], and other approaches [HHF10, SGR13, Ren12]. Since 1980s, the AI community began to study containers as a typical example for qualitative reasoning using symbolic input [BF03, Fra96]. In particular, Collins and Forbus [CF87] used containers to reason about liquid by introducing a new technique, namely molecular collection ontology. A knowledge base for qualitative reasoning about containers was developed by Davis *et al.* [DMC13], which was expressed in a first-order language of time, geometry, objects, histories, and events. However, existing methods for

spatial and temporal reasoning in the AI literature are mostly based on logic formulas, which are difficult to apply on processing real sensory inputs. The ability to handle noisy visual signal as inputs by introducing probability to model qualitative spatial relations makes our method different from previous work.

Robotics: Localizing objects using spatial relations has received considerable interests in the robotics community, *e.g.*, [ASF11, WKL13, FDP97, AAW10, NDD14]. Alper *et al.* [ASF11] utilized spatial relations to describe topological relationships between objects. Wong *et al.* [WKL13] studied occluded objects inside containers, and presented a novel generative model for representing container contents by using object co-occurrence information and spatial constraints. Feddema *et al.* [FDP97] applied two methods to control the surface of liquid in an open container which was moved by a robot. Most of the existing methods can only reason about containment relations in a known structured environment. In contrast, the proposed method aims to address the problem in arbitrary environments, where the number of objects are not fixed and relation changes occur more frequently.

Computer Vision: Two streams of studies are closely related to the present work: object affordance and tracking objects using context information. In recent literature, there is growing interest in understanding scenes and objects by their their affordances and functionalities [GGV11, GSE11, ZZ13, ZZZ15, ZJZ16] and their possible interactions with human poses [SLH12, KGS13, ZFF14, WNX14, WZZ13]. Using context information has also been extensively explored in human-object interaction and multi-object tracking. For instance, Yang *et al.* [YWH09] proposed to track multiple interacting objects by mining auxiliary objects, and [WNX14] formulated the interacting objects tracking as a network-flow Mixed Integer Program problem. More recently, a multiple objects tracking algorithm [YYL15] was proposed by maintaining spatial relations between objects using a Relative Motion Network. In comparison, our method utilizes the interactions between human and objects as temporal constraints to infer the explicitly modeled containment relations and their changes, resulting a probabilistic approach to recover objects with heavy occlusions.

5.1.2 Problem Definition

We use $\Omega = \{O^i | i = 1 \dots N\}$ to denote all the objects of interest in a scene, where O^i denotes the i th object. At each frame, we define the following variables:

- A containment indicator function is denoted by $C_t(\cdot) \in \Omega$. If O^j contains O^i directly, then $O^j = C_t(O^i)$, where O^i is the containee and O^j is the container.
- A containment relation $\mathcal{R}_t^i = \langle O^i, C_t(O^i) \rangle$ is an ordered pair representing the containment relation between O^i and $C_t(O^i)$. The set of all containment relations at time t is denoted as $\Lambda_t = \{\mathcal{R}_t^i | i = 1, \dots, N\}$.
- A containment graph is denoted as $\mathcal{G}_t = (\Omega, \Lambda_t)$, where Ω is the set of vertices and Λ_t is the set of directed edges.

To make the inference process tractable, we make the following assumptions about the properties of \mathcal{G}_t .

Spatial Assumptions: i) Each object must be contained by one and only one container, except the root node of \mathcal{G}_t , *i.e.*, the “room”, which does not have its container. ii) There is no loop in \mathcal{G}_t , that is, object cannot contain itself. The nested containment relations do not form loops. iii) A person becomes a container when he or she holds an object.

Temporal Assumption: We assume that all containment relation changes ought to be caused by a person, which means a scene does not have external disturbance other than human. In other words, if we know there is no human actions, the containment relations should not change. This temporal assumption couples containment relations in space with containment relation changes over time. Figure 5.3 gives an example: a person takes an apple out of a bowl, during which the person breaks the containment relation between the apple and the bowl, and establishes a new containment relation between the apple and the person.

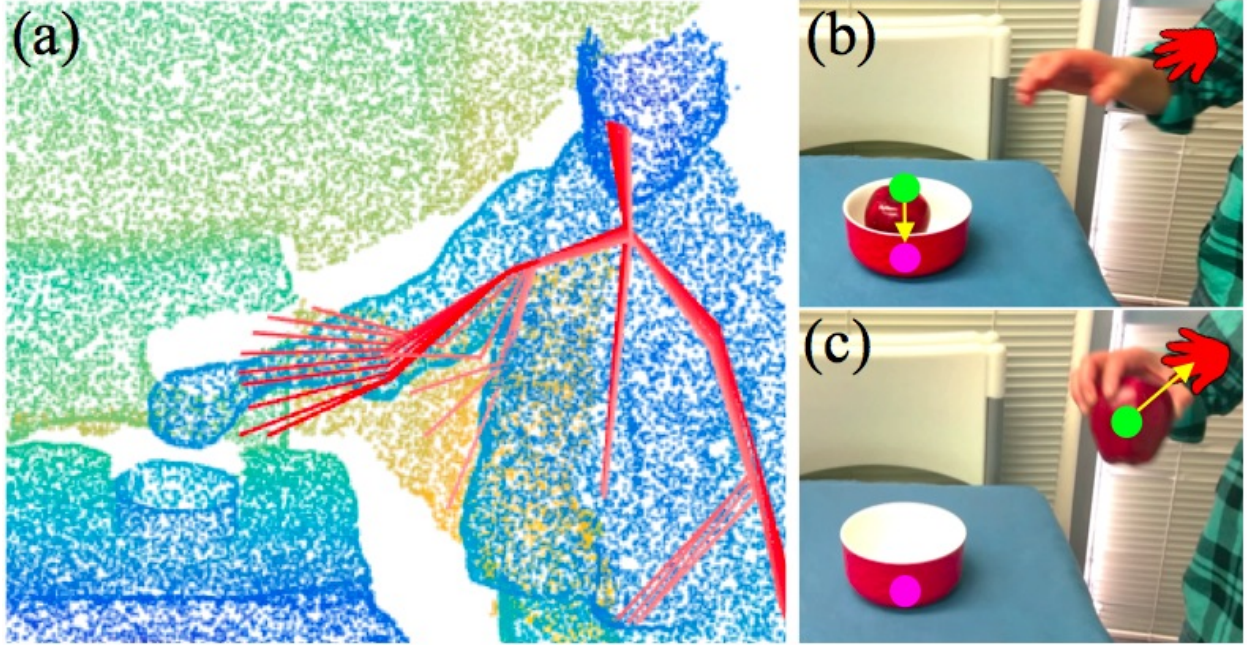


Figure 5.3: Temporal assumption. An apple is taken out of a bowl by a person, resulting a containment relation change. (a) Tracked skeletons during the time interval $[t, t + m]$. (b) At time t , the red bowl was the container of the apple. (c) At time $t + m$, the person became the container of the apple.

5.1.3 Problem Formulation

The objective of our work is to interpret the observed video as the optimal sequence of containment graphs $\{\mathcal{G}_t\}^*$ from a given RGB-D video $\mathcal{V}_{[1,T]} = \{V_t | t = 1, \dots, T\}$.

5.1.3.1 Containment Relations in 3D Space

At each frame t , a containee O^i is contained by a container $C_t(O^i)$ if and only if it satisfies all following relations defined in terms of an energy function Φ with the three components:

- IN relation defined by the energy term ϕ^{IN} ;
- ON relation defined by the energy term ϕ^{ON} ;
- AFFORD relation defined by the energy term ϕ^{AFF} .

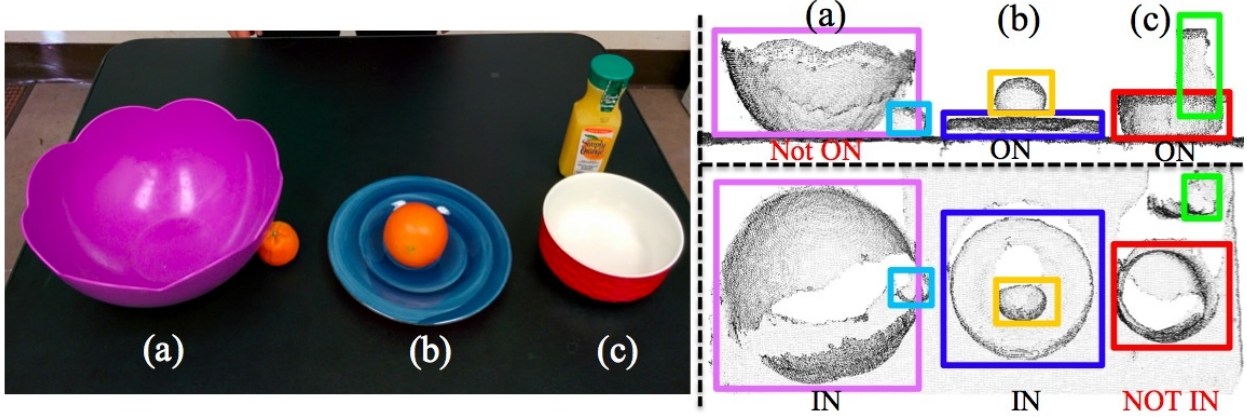


Figure 5.4: Containment relations in 3D space. (Left) A RGB image of a desktop scene. (Bottom right) Depth images from the top view. (Top right) Depth images from the front view. (a) and (c) violate ON relation and IN relation, respectively. Only (b) is considered to satisfy both IN and ON relations.

IN Relation describes containment relations from the top view:

$$\phi^{\text{IN}}(\mathcal{R}_t^i, V_t) = \Gamma(O^i) / [\Gamma(O^i) \cap \Gamma(C_t(O^i))], \quad (5.1)$$

where $\Gamma(O^i)$ is the projected area of containee O^i along the gravity axis, and $\Gamma(O^i) \cap \Gamma(C_t(O^i))$ is the overlapped area between containee O^i and its container $C_t(O^i)$ projected in 2D from the top view.

The bottom right of Figure 5.4 shows three examples of IN relation. If a containee is contained by its container, the boundary of the containee should be inside the contour of the container from the top view.

ON Relation describes containment relations from the front view:

$$\phi^{\text{ON}}(\mathcal{R}_t^i, V_t) = D(Z_b(O^i), Z_g(C_t(O^i))), \quad (5.2)$$

where $Z_b(O^i)$ is the bottom coordinates of the containee projected to 2D, $Z_g(C_t(O^i)) = [Z_t(C_t(O^i)), Z_b(C_t(O^i))]$ is the interval of the container's top and bottom coordinates projected to 2D, and D is a distance function which measures how well the bottom of the

containee $Z_b(O^i)$ falls into the intervals between the top and the bottom of the container $Z_g(C_t(O^i))$. If a containee is contained by its container, the bottom of the containee has to contact the container, and the containee should be above the container. Three examples of ON relation are illustrated in the top right of Figure 5.4.

AFF Relation $\phi^{\text{AFF}}(\mathcal{R}_t^i, V_t)$ measures the ability of a container to afford a containment relation with containee at frame V_t , which is a pair-wise term. The containment relation is subject to a set of physical and geometric constraints. For example, a porous basket can neither contain a containee bigger than itself, nor smaller containees like beads. In this work, we only consider the relative volume between the container and the containee.

Energy of Containment Relations is defined as:

$$\phi(\mathcal{G}_t, V_t) = \lambda_1 \cdot \phi^{\text{IN}} + \lambda_2 \cdot \phi^{\text{ON}} + \phi^{\text{AFF}}, \quad (5.3)$$

where λ_1 and λ_2 are the weights of the energy terms, obtained through cross-validation during the training phrase.

5.1.3.2 Containment Relation Changes in Time

The containment relation change between frame t and $t + 1$ is denoted as $\Delta\mathcal{R}_t^i$. We classify the changes based on human actions into the following four categories, shown in Figure 5.5.

Move-in is defined as $\Delta\mathcal{R}_t^i : \langle O^i, H_t \rangle \rightarrow \langle O^i, C_{t+1}(O^i) \rangle$, which describing a containee O^i moves from a person H_t at frame t to another container $C_{t+1}(O^i)$ at frame $t + 1$.

Move-out is the opposite change to move-in, defined as $\Delta\mathcal{R}_t^i : \langle O^i, C_t(O^i) \rangle \rightarrow \langle O^i, H_{t+1} \rangle$.

No-change implies there is no containment relation change between frame t and $t + 1$, defined as $C_t(O^i) = C_{t+1}(O^i)$.

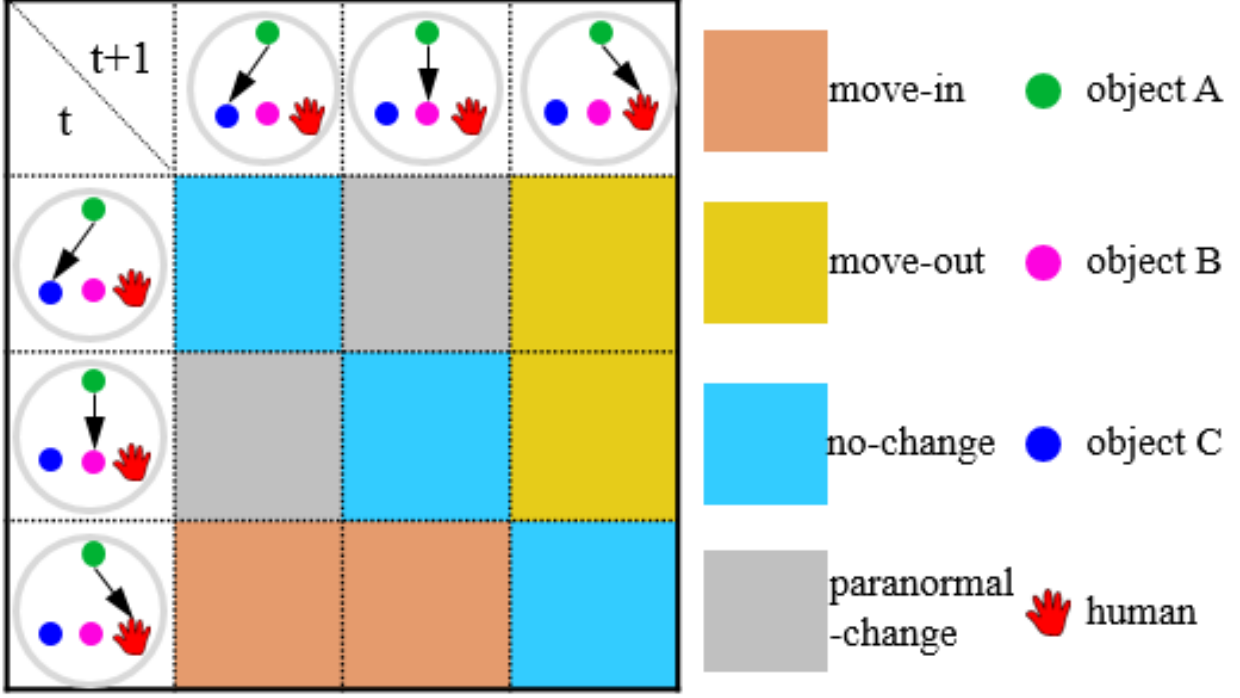


Figure 5.5: Transition matrix of containment relation changes for object A from frame t to $t + 1$. Move-in: the container of A changes from a person to an object. Move-out: the container of A changes from an object to a person. No-change: the container of A does not change. Paranormal-change: the containment relation changes without human intervention violate the *temporal assumption*, thus are ruled out.

Paranormal-change refers to the change that is in conflict with the *temporal assumption* that a person is the only cause that leads to containment relation changes, and thus is ruled out through reasoning. Formally, if the containment relation \mathcal{R}_t^i changes, *i.e.*, $C_t(O^i) \neq C_{t+1}(O^i)$, but no person H is involved, *i.e.*, $C_t(O^i) \neq H_t \& C_{t+1}(O^i) \neq H_{t+1}$, such changes are defined as paranormal-change.

Energy of containment relation changes is defined as

$$\psi(\mathcal{G}_t, \mathcal{G}_{t+1}, V_{[t-\epsilon, t+\epsilon]}) = \langle \omega_{\mathcal{L}_j}, \theta \rangle, \quad (5.4)$$

where $\omega_{\mathcal{L}_j}$ is the template parameter for four types of containment relation changes $\mathcal{L}_j, j \in \{1, 2, 3, 4\}$, $[t - \epsilon, t + \epsilon]$ is the time interval of the sliding windows for temporal feature

extractions, and θ is the extracted feature vector.

5.1.3.3 Joint Spatial-Temporal Energy

By combining both the energy of containment relations in space at each frame (Equation 5.3), and the energy of containment relation changes between adjacent frames (Equation 5.4), the optimal containment graph $\{\mathcal{G}_t\}^*$ is defined as:

$$\{\mathcal{G}_t\}^* = \underset{\{\mathcal{G}_t\}}{\operatorname{argmin}} E(\{\mathcal{G}_t\}, \{V_t\}) \quad (5.5)$$

$$= \underset{\{\mathcal{G}_t\}}{\operatorname{argmin}} \left[\mu \sum_{t=1}^T \phi(\mathcal{G}_t, V_t) + \sum_{t=1}^{T-1} \psi(\mathcal{G}_t, \mathcal{G}_{t+1}, V_{[t-\epsilon, t+\epsilon]}) \right], \quad (5.6)$$

where ϕ is the data term which models the energy of containment relations in space, ψ is the smooth term that models the containment relation changes in time, and μ is the trade-off parameter between the spatial-temporal cues.

5.1.4 Inference by Dynamic Programming

The goal of the inference process is to find the optimal sequence of containment graphs $\{\mathcal{G}_t\}^*$ for the input RGB-D video by optimizing the energy function defined in Equation 5.5.

The time complexity of searching the entire solution space is $O(N^{(N-1) \cdot T})$, where N is the number of objects and T is the video length. It is impractical to brute-force search the entire space.

Fortunately, at each frame, the container of the containee O^i is independent of the container of the containee O^j . By assuming such property that the container of each containee is independent, we can optimize the solution for each object separately. Hence, Equation 5.5 can be rewritten in terms of containment relations with respect to each object:

$$\{\mathcal{R}_t^i\}^* = \underset{\{\mathcal{R}_t^i | t=1, \dots, T\}}{\operatorname{argmin}} \left[\mu \sum_{t=1}^T \phi(\mathcal{R}_t^i, V_t) + \sum_{t=1}^{T-1} \psi(\mathcal{R}_t^i, \mathcal{R}_{t+1}^i, V_{[t-\epsilon, t+\epsilon]}) \right]. \quad (5.7)$$

By aggregating $\{\mathcal{R}_t^i\}^*$ from each object, we can recover the full sequence of containment graphs $\{\mathcal{G}_t\}^*$ of the scene. A dynamic programming is adopted to find the optimal solution of Equation 5.7 with the time complexity $O(N^2 \cdot T)$.

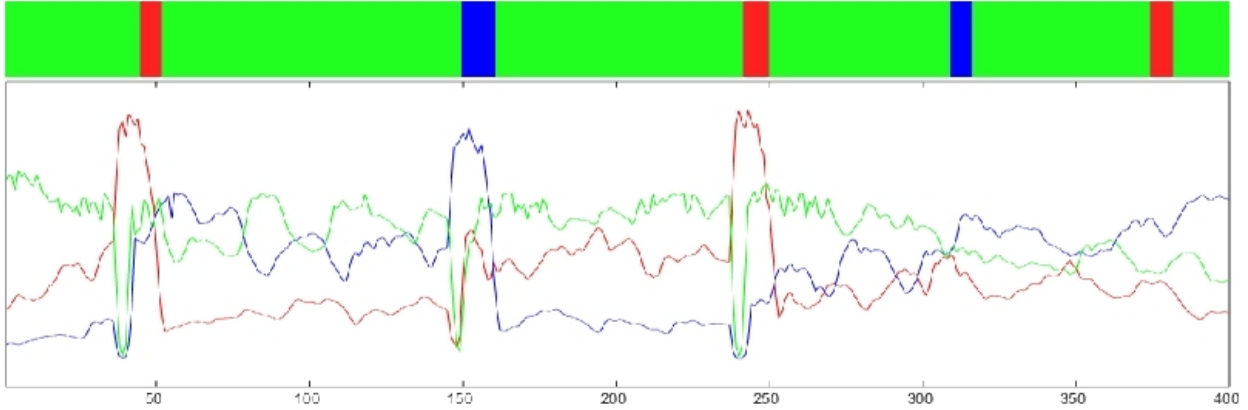


Figure 5.6: Probability of three different containment relation changes over time between two objects (in red and cyan bounding boxes). The ground truth is shown in the bottom.

5.1.5 Experiments

5.1.5.1 Dataset

We collected a RGB-D video dataset with diverse actions to evaluate the proposed method. Our dataset consists of 1326 video clips in 9 scenes captured by a Kinect sensor, in which 800 clips are used to train our model and the remaining clips are for testing. For each video clip, RGB and depth images, 3D human skeletons as well as point cloud information are used as the input of the proposed method.

Our dataset is unique, compared with traditional ones in the following aspects: i) it focuses on containers and containment relations; ii) it includes partially and completely occluded objects, such as an apple in a bowl (partial occlusion) and a laptop in a backpack (complete occlusion); iii) it includes diverse containment relation changes in different scenarios, such as throwing, picking up, opening lid, zipping zipper, *etc.*

5.1.5.2 Detection of Containment Relation Changes

Consider the case shown in Figure 5.6, where a person moves a containee from one container to another, during this process the containee changes directions, scales, and views. Severe

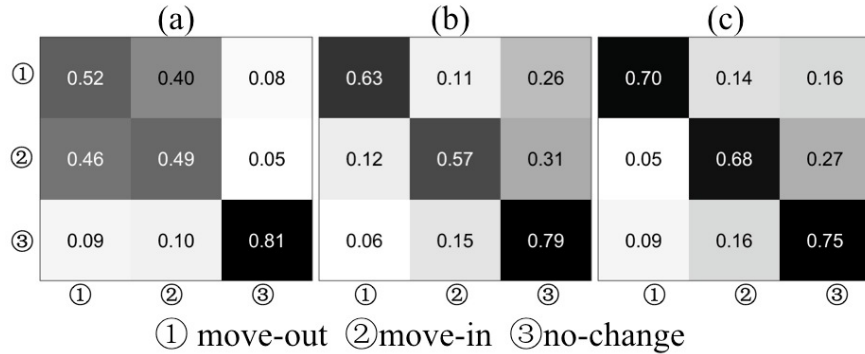


Figure 5.7: Confusion matrix of relation change recognition. (a) Human actions only. (b) Human actions with object context. (c) Joint inference using the proposed method.

occlusions by hands and other objects also occurred. We show the probabilities of three relation changes by detection: move-in, move-out and no-change between two objects highlighted with bounding boxes. At each frame, we compare the category which has the highest probability among three candidate categories with the ground truth (bars at the bottom). The detection of relation changes works well in most cases, but it fails in certain situations: i) when skeletons, the containee and the container are occluded at the same time (frame 54-70), the algorithm cannot distinguish the relation change of move-in or move-out from no-change; ii) some skeletons or the containee are occluded partly (frame 380-390), which causes difficulties in distinguishing move-in from move-out.

We quantitatively compare the results of containment relation changes between our method with two baseline methods, as shown in Figure 5.7: (a) recognition by human actions only, and (b) recognition by both human actions and object context. Both of them are trained by multi-class SVM on the same training data. There are obvious confusions between move-in and move-out in (a). By introducing object context, the proposed method improve the accuracy as shown in (b). The proposed method achieves the best performance in (c). The reason is that our method is able to correct some temporal detection errors.

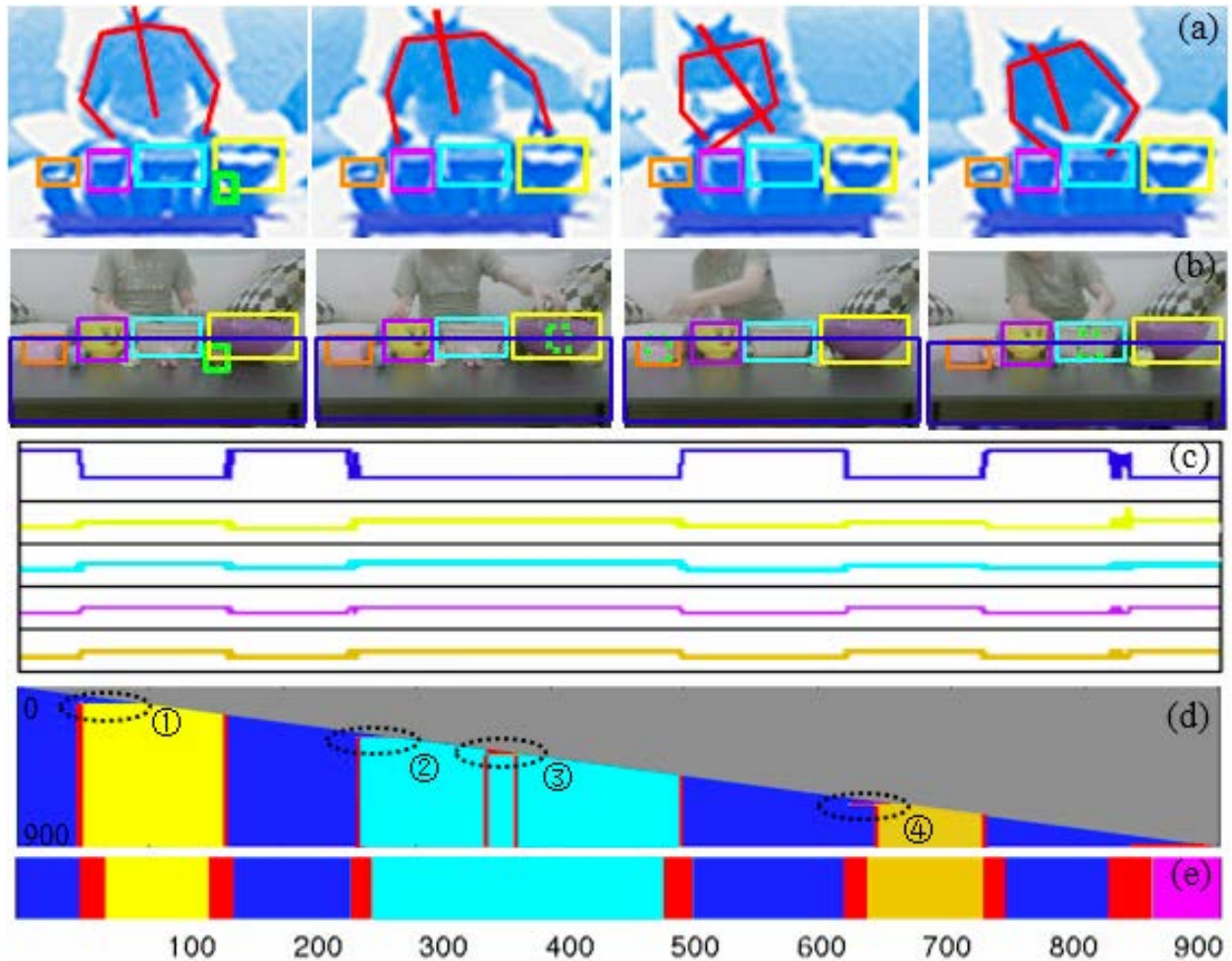


Figure 5.8: Inference of containment relations for the object in green bounding box. Each color denotes a different object. (a) The tracked objects and the human skeletons in 3D scene. (b) Refined tracking results. We recover the positions of hidden containee using the positions of its container. (c) The probability of the containee contained by each possible container in space. (d) The inference result matrix given different length of the same video. The results are corrected as more information provided ①-④. (e) Ground truth.

5.1.5.3 Inferring Containment Relations

Figure 5.8 demonstrates the inference process. Firstly, we track all objects of interest by state-of-the-art RGB-D trackers [SX13]. The objects are bounded by boxes with different colors in Figure 5.8(a). Take the object in the green box as an example, we infer the

containment relations it involves in.

Figure 5.8(c) shows the probability of containment relations between the object in the green box and its potential container, using the spatial cues only. Each row represents one possible container. The height of the line represents how likely this container contains the object in the green box. Due to severe occlusions, the spatial cues of objects are missing constantly. When the object is occluded, the probability for each possible container is evenly distributed.

Figure 5.8(d) shows the inference result matrix, in which the n^{th} row represents the DP result of each frame given the first n frames of the video. The grey area denotes the states that are not observed up to the n^{th} frame. As more information is given, the DP algorithm gradually corrects results and gets closer to the ground truth. Take ① for example, the inferred container from frame 53 to 90 does not change to the human hand until the 91st frame. The initial inferred container remains to be the table due to heavy occlusion. But as time goes, the temporal information accumulates and wins against the spatial cues. our method achieves good performance in comparison with the ground truth in Figure 5.8(e).

We also perform quantitative evaluations. For comparison, we transform the tracking results into containment relations as a baseline. Specially, we apply non-maximum suppression at each frame for all candidates, and the bounding box with the highest score is considered as the present position of the object. For each object O^i , its container is the object which is nearest to O^i , and satisfies both ON relation and IN relation.

We divide our dataset into three parts according to the visibility of objects: no occlusion, partial occlusion and complete occlusion. Partial occlusion is the situation that the containee or its container is observed partially, whereas complete occlusion means that the container or its container is occluded completely, such as a laptop is put into a backpack. We quantitatively evaluate the accuracy of containment relations on these situations. The results are shown in Table 5.1. In the completely occluded situation, both methods perform worse than in the other situations. The proposed method performs better by recovering some relations from complete occlusions. In the situation of no occlusion, there are some

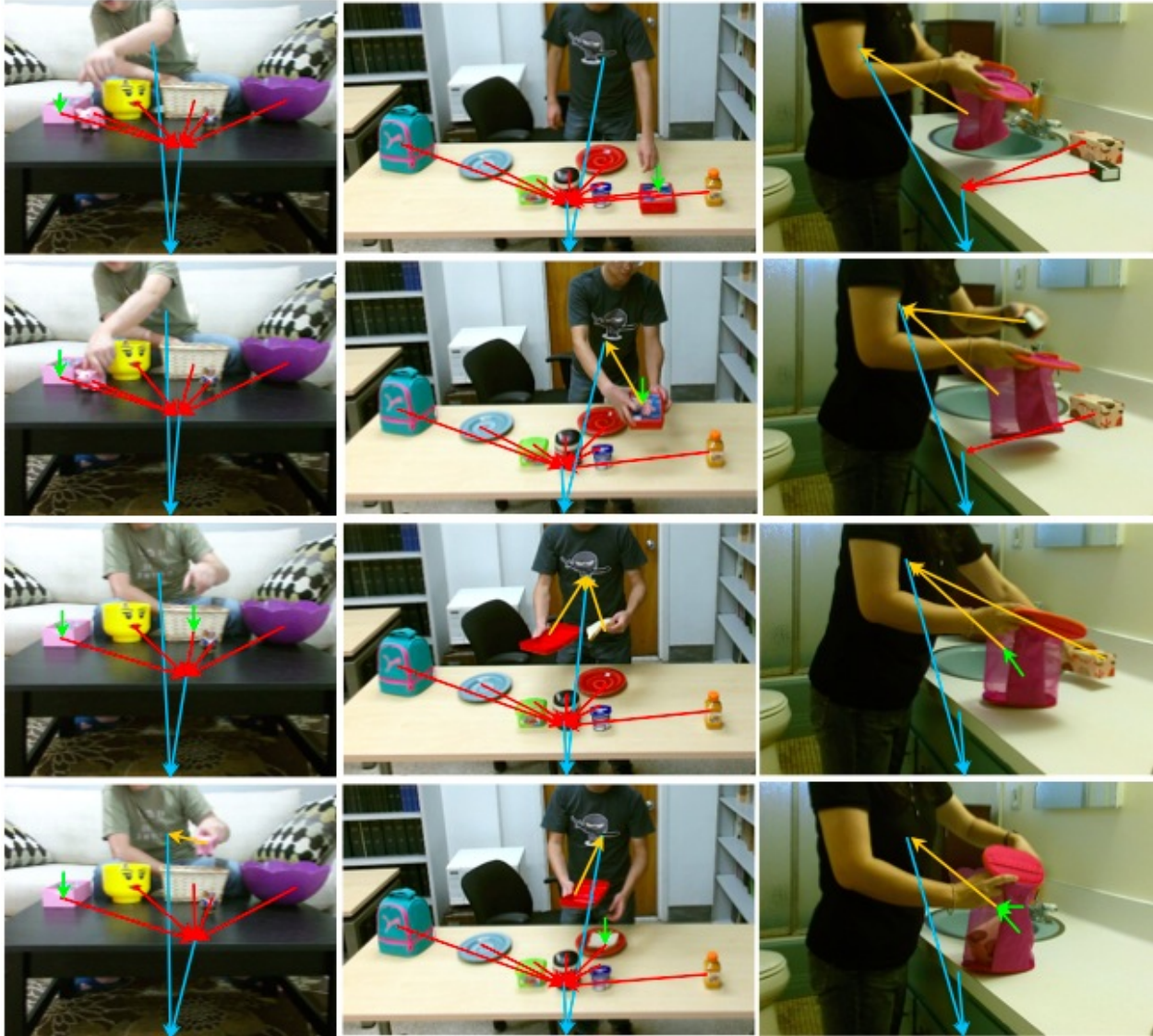


Figure 5.9: Some qualitative results. Each row shows the result of a specific scene. The arrows represent the containment relations between objects. Each arrow points from one containee to its container. The green arrows denote containment relations between the hidden containees and their containers.

false positives because of the observation noise in the detection process. Our method is efficient to eliminate some false positives.

Table 5.1: Accuracy of containment relation estimation in %.

Method	no occlusion	partial occlusion	complete occlusion	overall
Baseline	0.75	0.21	0.08	0.37
Ours	0.86	0.64	0.43	0.59

5.1.6 Conclusion

Containers and containment relations are ubiquitous in daily scenes and activities. They are useful not only to answer “what is where over time” for various AI tasks, but also for quantizing the perception of functional space, detecting and tracking hidden objects and heavily occluded objects, and reasoning about human subtle actions. The presented method achieves good performance in some challenging scenarios. However, it is still limited in the following aspects: i) IN and ON relation do not describe all containment relations, such as liquid or gas; ii) the objects with large deformation, such as plastic bags, are still difficult to solve.

5.2 Tracking Occluded Objects and Recovering Incomplete Trajectories by Reasoning about Containment Relations and Human Actions

We study the problem of tracking occluded objects during human daily activities in cluttered scenes, such as packing, playing, working, *etc.*. Figure 5.10 shows an example of a daily indoor scenario captured by a RGB-D sensor: an agent ① enters a room; ② puts down her backpack; ③ takes a laptop out of the backpack and puts it on the table; ④ grabs a cup, fetches some water from a water dispenser; ⑤ sits back and puts down the cup next to her. During the course of this event, objects disappear and then re-appear frequently.

Tracking objects in such scenarios is a challenging problem due to severe occlusions caused by two conditions:

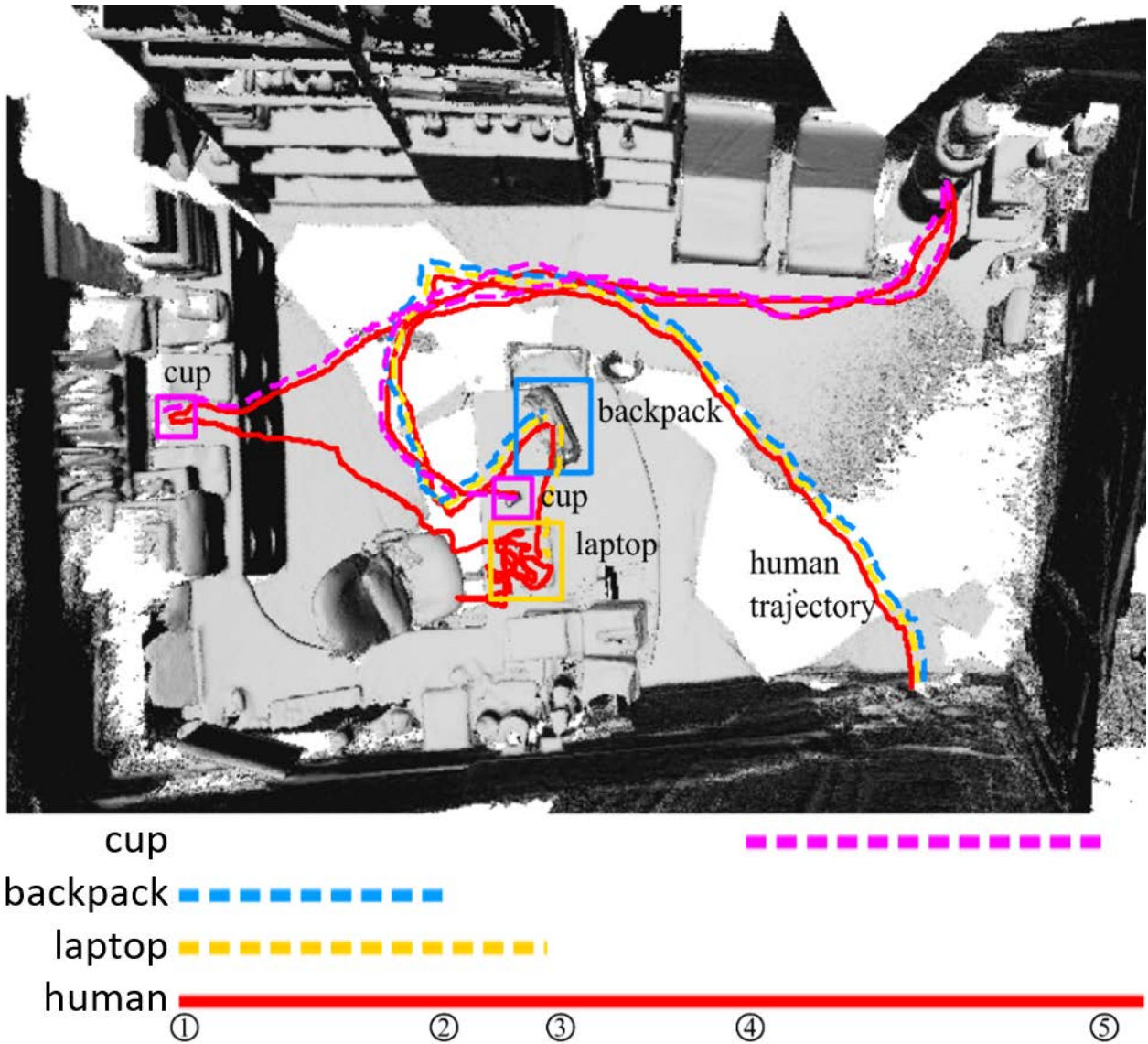


Figure 5.10: A scenario for tracking occluded objects in an indoor scene. The dashed lines represent the inferred trajectories and different colors indicate different objects in the scene. By explicitly reasoning about containment relations, the proposed algorithm is capable of recovering full trajectories of objects even they are contained or occluded by other objects in the video.

- **Contained.** The occlusion is caused by a *new containment relation formed* between two objects, *e.g.*, a person puts a laptop into a bag, which is view-independent;
- **Blocked.** The occlusion is caused by other objects observed from certain camera views,

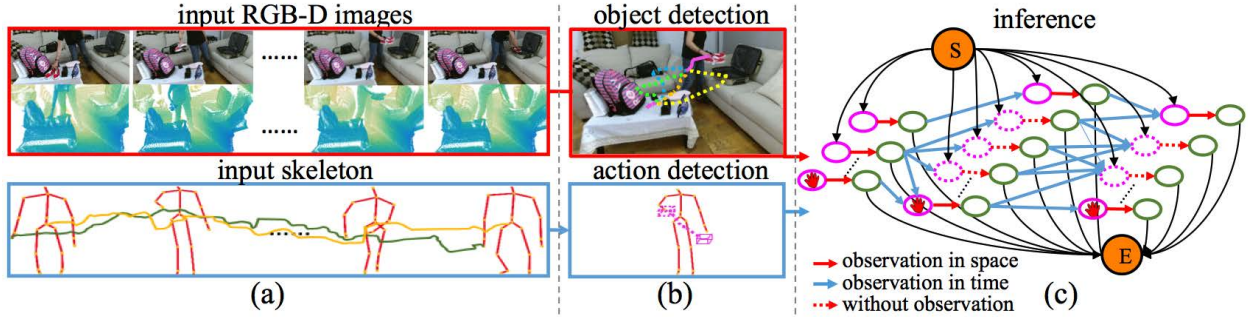


Figure 5.11: The framework of the proposed method. (a) Sensor input: a sequence of RGB-D images and human skeleton captured by a Kinect sensor. (b) Off-the-shelf state-of-the-art object detection and human action detection algorithms were applied to extract the object location and human actions per frame. (c) Inference on a network flow representation. The solid red lines denote the observations in space. The dashed red lines denote that the present state of the object is hidden and there is no observation. The blue lines denote the observations in time. S and E are the start and the end of the trajectories, respectively. A dynamic programming scheme is applied to search, optimize and recover the complete trajectory of each object.

in which the *containment relations unchanged*, e.g., a laptop is sitting in front of a cup, which blocks the view of the cup from the current camera view and is view-dependent.

We argue such problem is not merely a vision task compared to traditional visual tracking tasks, which primarily focuses on reliable object detectors and data association methods. Instead, significant reasoning processes are involved. To address this problem, we believe an explicit model of relations among objects as well as relations between objects and agents are needed.

The proposed framework is shown in Figure 5.11. Given a RGB-D video with extracted human skeleton sequence in a scene, state-of-the-art detection algorithms are applied to detect regions of interest of each object and human actions over time; the detection results serve as the initial proposals and the input to our algorithm. We pose the problem of recovering object trajectories as Maximizing a Posteriori (MAP) problem using a network flow repre-

sentation, in which a trajectory of an object is jointly interpreted and constrained by both object detection and containment relations in space as well as human actions over time. A dynamic programming scheme is applied to search, optimize and recover the complete trajectory of each object.

This work makes three contributions:

1. We propose a method to recover incomplete trajectories of objects by taking account of containment relations and two causes of occlusions: contained and blocked.
2. We assume that human action is the only cause that leads to occlusions and object status changes, and use it as a constraint to interpret trajectories of objects over time.
3. We introduce a new dataset including a set of RGB-D videos of human interacting with occluded objects.

5.2.1 Related Work

Spatial Reasoning: Spatial reasoning plays an essential role in human daily life. Although quantitative approaches can provide the most precise information, numerical information is often unnecessary or unavailable at human level. In computer vision, quantitative approaches usually study objects tracking problem, of which the literature is too expansive to survey here; we refer readers to recent survey and benchmark [WLY15, SCC14]. Here, we focus on spatial reasoning methods related to the presented work.

As a typical example, container has been used to study spatial reasoning problem [BF03, Fra96]. Using physical-based simulations, [LZZ15] evaluated human cognition of containing relations through human studies. [DMC13] developed a knowledge base for qualitative reasoning about containers, expressed in a first-order language of time, geometry, objects, histories, and events. Some exemplary tasks include i) computational approaches for reasoning about liquid transfer [KJZ16, YDY15, MSF17], ii) reason about containability and containment relations [LZZ16, YDY15, WLY17], and iii) occlusion modeling and reasoning [EF10, EES10, WWR11]. Compared to prior work, we integrate qualitative and quantita-

tive approach and explicitly model the occlusions by containment relations when an object is contained or blocked by others.

Detection using Context: Context has been widely explored in human-object interactions (HOI) and multi-object tracking. Some typical approaches and setups include: i) Learning deformable action templates [YZ09], actionlet [WLW14] and animated pose templates [YNL14]. ii) Combining spatial and functional constraints between human and objects [GKD09, YWH09, WZZ13]. iii) Task-oriented action recognition [ZZC15] and utility learning [ZJZ16, SHC17], including complex cooking tasks [RAA12, AAD11].

Different from the literature in context modeling, we explicitly model human action as a context cue. Specifically, we assume that human action is the only cause that leads to the changes of containment relations, and use the human action as a constraint to improve the tracking.

Although some recent work adopted deep neural networks to extract contexts for object detection and tracking, these data-driven feedforward methods have well-known problems: i) They are black-box models that cannot be explained and only applicable with supervised training by fitting the typical context of the object, thus difficult to generalize to new tasks. ii) Lacking explicit representation to handle occlusions, low resolution, and lighting variations—there are millions of ways to occlude an object in a given image [WZX17], making it impossible to have enough data for training and testing such black box models. In this work, we go beyond passive recognition by reasoning about time-varying containment relations.

5.2.2 Probabilistic Formulation

A key concept in the present work is “containment relation”. An object which contains or holds another object can serve as a container, forming containment relation with the object it contains. For instance, when a laptop is inside a backpack, a containment relation is formed between the laptop and the backpack, where the backpack plays the role of container.

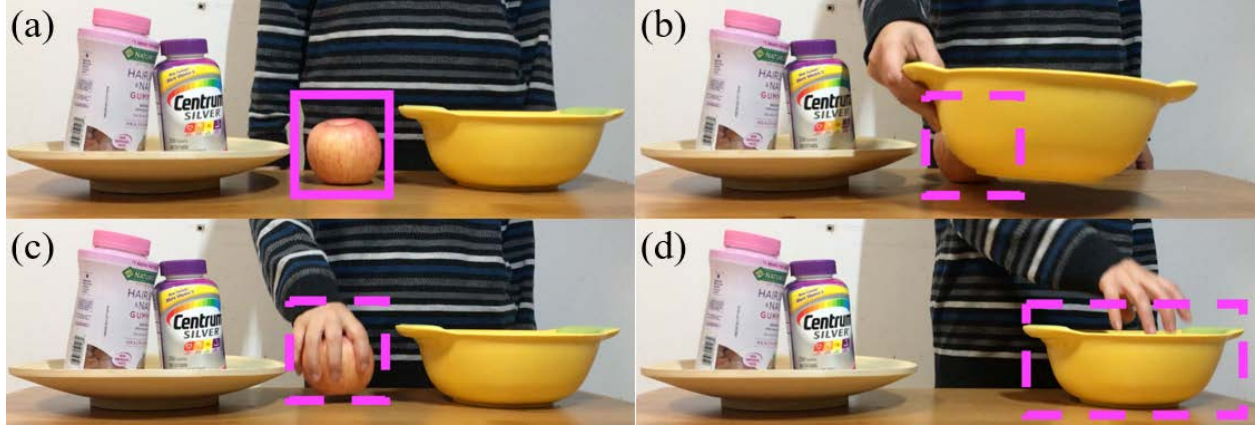


Figure 5.12: Two causes of occlusions. i) Blocked: An apple (a) can be detected at the beginning, but later (b) becomes occluded by a bowl. ii) Contained: an apple is contained by a person (c) and a bowl (d), respectively.

We make the following assumptions for containers and containment relations:

- When an object is contained by a container, the object will inherit the same trajectory from its container. For example, consider a case that a laptop is inside a backpack. If a person carries the backpack around, the laptop will move together with the backpack, sharing the same trajectory.
- The containment relation is a partially ordered relation constrained by the volume of the object and its container, *i.e.*, if a container's volume is smaller than an object's, the object cannot be contained by this container.
- An object can only be contained directly by one container.

Suppose there are K objects in a scene. Our goal is to recover trajectories of all K objects $T = \{T^1, T^2, \dots, T^K\}$ from a RGB-D image sequence $I = \{I_1, I_2, \dots, I_\tau\}$, where τ is the length of the image sequence. T^k is defined as an ordered set of object states $T^k = \{x_1^k, x_2^k, \dots, x_\tau^k\}$, where x_t^k is the state of the k th object in space at time t : $x_t^k = (l_t^k, c_t^k)$, where l_t^k is the location of the k th object at time t and $c_t^k \in \{1, 2, \dots, K\}$ is an object index, representing the inferred container of k th object at time t .

5.2.2.1 Spatial Hypotheses by Containment Relations

At each frame t , instead of purely relying on detection results, our algorithm further proposes two types of hypotheses generated based on the possible causes of occlusion. These two hypotheses provides additional cues, essentially competing with the detection results. As a result, such extra info recovered by the containment relations could later help overcome the miss or wrong detection described in the next section. The two types of hypotheses are:

Contained: In these situations, occlusion happens due to forming new containment relations as shown in Figure 5.12 (c)(d), where hypotheses are shown in dashed box. Formally, suppose such occlusion happens to the k th object at time t , the algorithm proposes that the location of the k th object the same as its container while keeping it's container the same as in the previous frame: $x_t^k = (l_t^{c_t^k}, c_t^k), c_t^k \neq k$.

Blocked: In such cases, an object is occluded due to another object sitting in between the object and the camera from certain camera views, as shown in Figure 5.12(a)(b), where an apple is occluded by a bowl and a person. The dashed box is the proposal for the apple's present location, which is the same as the location in the last frame before occlusion happened. Formally, suppose such occlusion happens to the k th object at time t , the algorithm proposes the object state as the same in previous state: $x_t^k = x_{t-1}^k = (l_{t-1}^k, c_{t-1}^k)$, where the location and containment relation remain the same.

5.2.2.2 Temporal Hypotheses by Human Actions

Across different frames, we consider human as the cause and the only cause of object state changes, assuming no other external disturbance in a scene. In other words, if there is no human action occurring, the objects should remain the same location and the containment relations will not change. As a result, human actions impose a hard constraint to rule out the implausible sudden jumps from the object detection, resulting in a smooth and plausible trajectory.

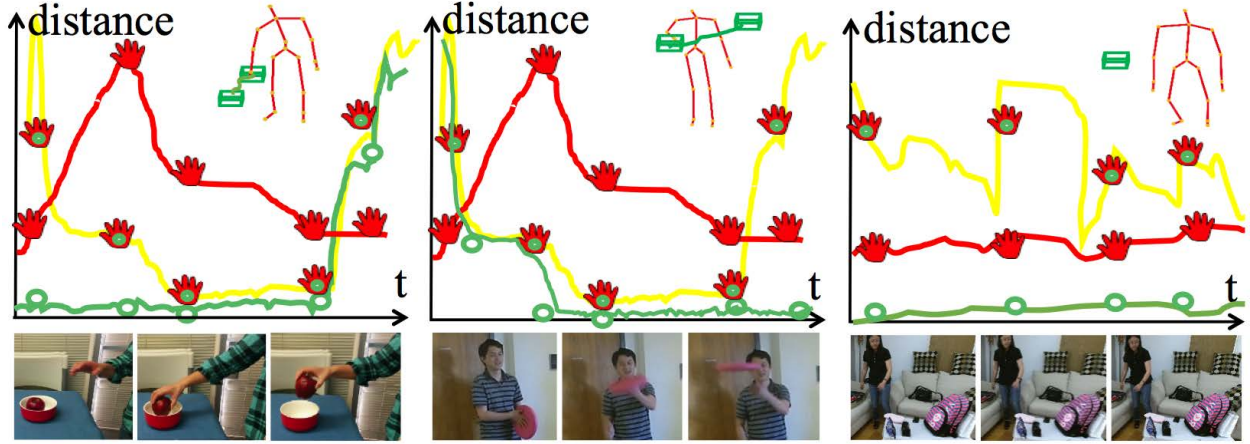


Figure 5.13: Human action features at time t in a sliding window with the length of 2ϵ . The red line represents the distance between the hand and the spine of the person. The yellow line represents the distance change between the hand and the object. The green line represents the location change of the object in the current sliding window.

In this work, we represent the human action as a skeleton sequence $H = \{H_1, H_2, \dots, H_\tau\}$, where τ is the length of the sequence. At time t , 25 joints of human skeleton captured by a Kinect sensor were used: $H_t = (h_t^1, h_t^2, \dots, h_t^{25})$.

5.2.2.3 Recovering Incomplete Trajectories

We recover incomplete trajectories using MAP by reasoning about containment relations and human actions:

$$T^* = \arg \max_T P(T|I) = \arg \max_T P(T|\mathcal{X}, H) \quad (5.8)$$

$$\propto \arg \max_T P(\mathcal{X}|T)P(T|H) \quad (5.9)$$

$$= \arg \max_T \prod_k P(\mathcal{X}|T^k)P(T^k|H), \quad (5.10)$$

where \mathcal{X} and H are the object detection in space and human action in time, respectively. $P(\mathcal{X}|T^k) = \prod_{t=1}^\tau P(\mathcal{X}_t|x_t^k)$ models the likelihood for object detector response $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_\tau\}$. $P(T^k|H)$ is a dynamic model which is a smoothness term for trajectory,

and can be decomposed as

$$P(T^k|H) = P(\{x_1^k, x_2^k, \dots, x_\tau^k\}|H) \quad (5.11)$$

$$= P_S(x_0^k) \prod_{t=1}^{\tau} P(x_t^k|x_{t-1}^k, H_{t-1}) P_E(x_\tau^k), \quad (5.12)$$

where $P_S(x_0^k)$ and $P_E(x_\tau^k)$ are the probability for initialization and termination, respectively, and $P(x_t^k|x_{t-1}^k, H_{t-1}^k)$ is the transition probability of two consecutive frames, which models the probability that the object status changes from time $t - 1$ to t based on the observation of human action H_{t-1} . Intuitively, this probability evaluates the consistency between the location of an object and human actions. As we discuss in previous section, the location changes of an object can be interpreted by the occurrence of human actions. Figure 5.13 illustrates three examples of the object location changes and the corresponding human actions, including (a) a person taking an apple from a bowl, (b) a person throwing a frisbee, and (c) the object keeps the same location without human action.

The transition probability of two consecutive states is

$$-\log P(x_t^k|x_{t-1}^k, H_{t-1}^k) = \langle \omega, \theta_{[\epsilon]} \rangle, \quad (5.13)$$

where ω is the template parameter. $\theta_{[\epsilon]}$ is the extracted human action feature in a time interval $[t - 1 - \epsilon, t - 1 + \epsilon]$.

For θ , we consider three types of features in a sliding window on the time axis: human pose, relative movements between the human and the object, and the object movements. Suppose that the sliding window size is 2ϵ , the feature vector sequence at time t is $\mathcal{F}_m = (\mathcal{F}_m^h, \mathcal{F}_m^r, \mathcal{F}_m^o)$, $m \in [t - 1 - \epsilon, t - 1 + \epsilon]$. Specifically,

1. \mathcal{F}_m^h is the relative distance of all the skeletons to three base points (two shoulders and one spine point), which encodes human action. In Figure 5.13, we show one component of \mathcal{F}_m^h in red lines: the distance between the hand and the spine point.
2. \mathcal{F}_m^r is the distance between human hand and the location of the object, which is denoted in yellow lines in Figure 5.13.

3. \mathcal{F}_m^o is the distance between the locations of the object at time m and t , depicted by green lines in Figure 5.13.

A sequence clip is first interpolated to a certain length. The wavelet transform is then applied to \mathcal{F}_m . The coefficients at the low frequency are kept as the action feature. The window sizes and sliding steps are both in multiple scales.

Substituting Equation 5.11 into Equation 5.8, we then have

$$T^* = \arg \max_T \prod_{k=1}^K \prod_{t=1}^{\tau} [P(\mathcal{X}_t | x_t^k) \cdot P_S(x_0^k) \cdot P(x_t^k | x_{t-1}^k, H_{t-1}) \cdot P_E(x_\tau^k)]. \quad (5.14)$$

We can reformulate Equation 5.14 as an Integer Linear Programming problem:

$$f^* = \arg \min_f C(f), \quad (5.15)$$

where

$$C(f) = \sum_i c_i^s f_i^s + \sum_{i,j} c_{ij} f_{ij} + \sum_i c_i f_i + \sum_i c_i^e f_i^e \quad (5.16)$$

$$c_{ij} = -\log P(x_j | x_i, H_i) \quad (5.17)$$

$$c_i = -\log P(x_i | T^k) \quad (5.18)$$

$$c_i^e = -\log P_E(x_i) \quad (5.19)$$

$$c_i^s = -\log P_S(x_i) \quad (5.20)$$

$$\text{s.t. } f_{ij}, f_i, f_i^s, f_i^e \in \{0, 1\}. \quad (5.21)$$

This is equivalent to finding a min-cost path in network flow with source S and sink E as shown in Figure 5.11: the red arrows denotes the detection on input RGB-D images with cost on the edge c_i , the dashed red arrows indicates that the object is hidden at the present state and there is no observation from current frame, and each transition between successive frames is denoted by blue lines with cost c_{ij} given by human actions, serving as a smoothness term.

Dynamic programming is applied to optimize Equation 5.16. By assuming objects will not affect each other's trajectory, we optimize the trajectory for each object individually. Firstly,

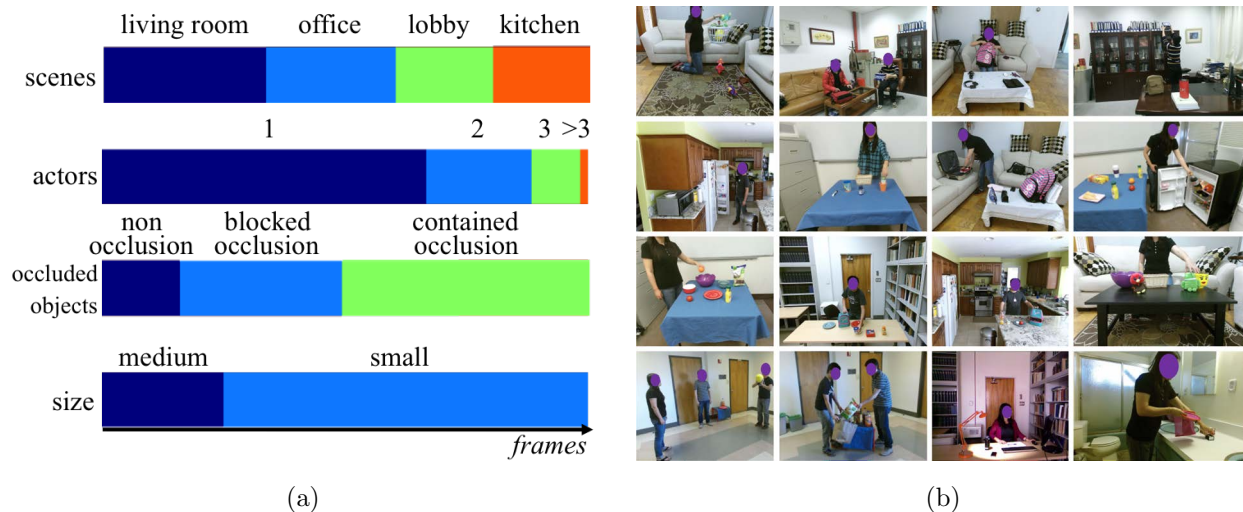


Figure 5.14: Our occluded object tracking dataset. (a) Statistic of the dataset. (b) Some examples of the activities.

we run K-Shortest Paths Algorithm [BFT11], which generates a set of tracklets. Then we use the Viterbi algorithm to connect these tracklets, which yields continuous trajectories for each object.

5.2.3 Experiments

5.2.3.1 Dataset

We collected a 3D dataset with diverse scenes, multiple actions and various objects to evaluate the proposed method (Figure 5.14). 1346 video clips in 10 scene categories were captured by Kinect sensors. RGB and depth images, 3D human skeletons as well as point cloud data were recorded in each video clip. Compared with existing dataset, the proposed dataset focuses on occluded objects for visual tracking, which consists of a large variety of human actions causing object location changes in different scenarios, such as throwing, catching, picking up, putting down, fetching, lifting, *etc.*

Each frame in the dataset was manually annotated with ground truth by drawing bounding boxes for each object. When an object is occluded, we annotate the ground truth based on two types of causes for the occlusions. i) Contained. The object shares the location with

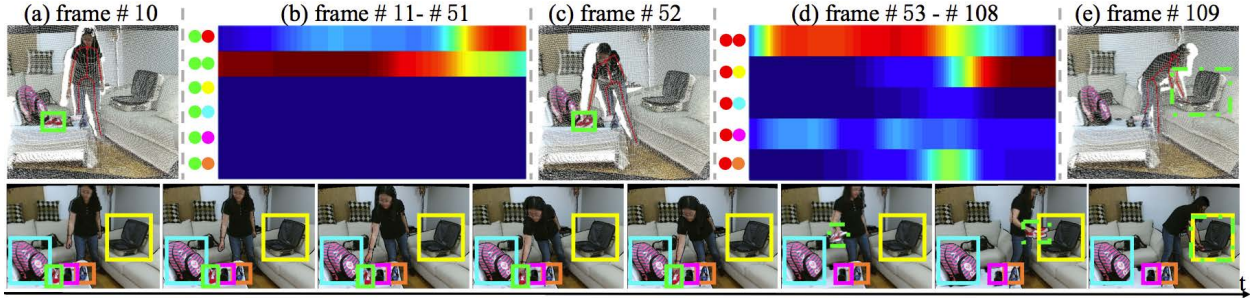


Figure 5.15: Transition probability of the object location in the green bounding box. The solid boxes depict that the object is tracked by object detectors. The dashed boxes depict that the object is recovered by inference. (a), (c) and (e) show detected bounding boxes and human skeletons on point cloud. (b) and (d) are the transition probabilities between two possible locations. In (b), the bottom four bars with low probability keep the same since we constrain the impossible object moving that are not caused by human actions.

its container, forming a new containment relation. ii) Blocked. The object is stationary, and the containment relation remains the same. For the situation that a person serves as a container, we draw a bounding box on the person’s hand.

5.2.3.2 Transitions in DP: an In-depth Example

Figure 5.15 shows an example of the trajectory inference process of an object bounded by a green box. The tracking results are visualized in the bottom panel, where the solid boxes denote the detected location, and the dashed boxes denote the inferred results. Specifically, we employed the state-of-the-art RGB-D based detectors [SX13] on a RGB-D image sequence. The detected objects are bounded by boxes with different colors shown in Figure 5.15(a), (c) and (e). The human skeletons from Kinect are in red color.

Figure 5.15(b) and (d) illustrate the partial transition probabilities changes between two consecutive states in an interval (frame 11 to frame 51, frame 53 to frame 108), equivalent to the probability of human actions and calculated by Equation 5.13. The left panel of (b) and (d) are some possible transitions. Take the first bar in (b) as an example. The green

and red dot represent the location of the object bounded by green bounding box and the person, respectively. The bar depicts the probabilities of the transition from the green box location to the human hand location over time. We can see that the probability increases from frame 11 to frame 51. At frame 51, the person picked up the object. From frame 59 to frame 108, the object was held by the person. The first bar of Figure 5.15 (d) shows the probability of the object being carried by this person.

It is worth noting that the bottom four bars in Figure 5.15(b) have low transition probabilities which are close to zero. Take the last bar in Figure 5.15(b) as an example. It shows the probability of the object bounded by the green box moving to the location of the object bounded by the orange box. This movement was not caused by human action and violated our assumption, which was ruled out during the inference.

From frame 51 to 109, the object was contained and thus cannot be visually detected. Human action provided a strong cue for the object location: a person picked up this object and moved it to a container bounded by a yellow bounding box.

5.2.3.3 Ablative Analysis: Roles of Interactions in HOI

In this section, we evaluate the roles and importance of HOI quantitatively by turning on and off certain components in the proposed method.

We consider the HOI as a binary classification problem: if the object movement is consistent with human action, it should be classified as positive; otherwise it is negative. We define whether the object movement is consistent with human action using two criteria: i) if no human action, the object should remain stationary, and vice versa; ii) if there is an object location change, the object should follow the trajectory of human action.

We first consider the simplest method using human pose only, *i.e.*, Equation 5.13 with feature vector $\mathcal{F}_m = (\mathcal{F}_m^h)$. As showed in Figure 5.16a, using human pose only is not sufficient to achieve reasonable performance. This was mainly caused by the lack of object context, disallowing a good classification between certain actions, *e.g.*, putting down and picking up.

Next, we consider the method using both human pose and object context, *i.e.*, Equa-

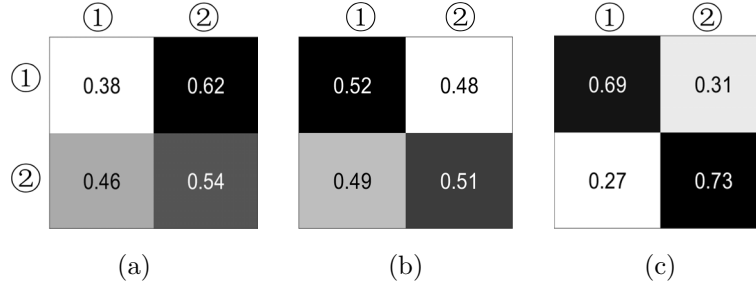


Figure 5.16: Confusion matrix of HOI. ① denotes that the object movement is consistent with HOI, whereas ② denotes that the object movement is not consistent with HOI. (a) Human pose sequence only. (b) Human pose sequence with objects context. (c) Joint inference in our method.

tion 5.13 with feature vector $\mathcal{F}_m = (\mathcal{F}_m^h, \mathcal{F}_m^r, \mathcal{F}_m^o)$. Although achieving reasonable results as shown in Figure 5.16b, this method only looks at local window $m \in [t-1-\epsilon, t-1+\epsilon]$, thus lacking of global optimization.

With back propagation using DP as described in Equation 5.14 and Equation 5.16, the proposed method globally adjust the inference, resulting in the best performance among three methods as shown in Figure 5.16c.

All the results report here were trained by SVM on the same training data. To address the problem of different scales of interaction, different step sizes and different sliding window sizes along time axis were used.

5.2.3.4 Ablative Analysis: Spatial/Temporal Suppression

In this section, we design two experiments (baseline1 and baseline2) to evaluate how spatial and temporal information influence the tracking. We compare the results of these two experiments with the approach of tracking with occlusion model (baseline3) and the proposed method (full model).

As an example, we show comparisons of results from different methods using a video of 530 frames (Figure 5.17). In this video, three actors threw and caught a ball highlighted by a yellow bounding box. The ball traveled fairly fast, appearing and then disappearing frequently. Directions, scales, and views of the ball also varied. Severe occlusions by hands

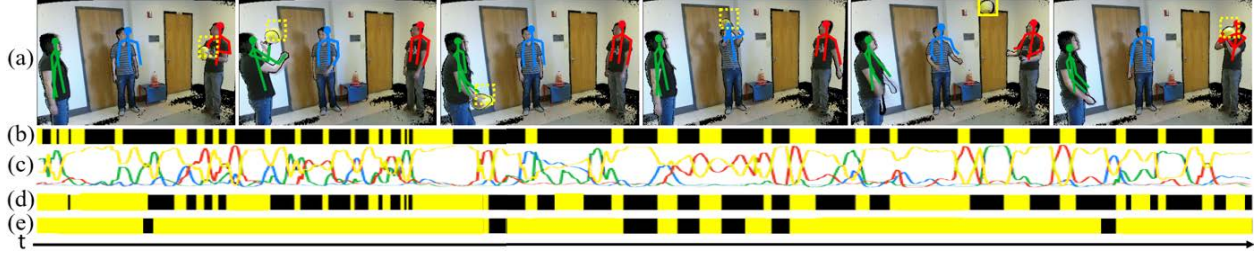


Figure 5.17: An example of the experiment results. The goal is to track the yellow ball. In each bar, the yellow represents the correct results, and the black represents the wrong results. The overlap ratio of bounding boxes were set to 0.5. Different colors denote different objects: actor 1 (green), actor 2 (blue), actor 3 (red) and ball (yellow). (a) Examples of tracking results. The dashed boxes depict the object is occluded. (b) Temporal-suppression results. (c) The scores of consistency between object movement and human action. (d) Spatial-suppression results. (e) Full model results.

or other body parts occurred.

Temporal Suppression (baseline1): In this setting, we do not consider the human actions, *i.e.*, set Equation 5.17 to a constant. As a result, it is equivalent to an online tracking problem: the trajectory of an object is determined only by the response of detectors. Non-maximum suppression was applied on all detection candidates per frame. Figure 5.17 (b) shows the results.

Spatial Suppression (baseline2): In this setting, we set Equation 5.18 to a constant, *i.e.*, not considering the detection score, but inferring object location only by human actions in time. In other words, the trajectory of an object is determined only by the transition probabilities modeled by human actions.

Results were shown in Figure 5.17 (d). Failure cases mostly fall into two categories: i) when human skeleton, the object and the container are occluded at the same time, and ii) when human skeleton or the object are partially occluded, it is difficult to distinguish the throwing action from the catching action as the lack of action cues or spatial context.

Table 5.2: Tracking accuracy of full model compared with three baselines on different subsets of the proposed dataset.

	baseline1	baseline2	baseline3	full
all	0.57	0.32	0.59	0.69
blocked	0.21	0.08	0.25	0.47
contained	0.15	0.02	0.16	0.42

Tracking with Occlusion Model (baseline3): A related topic in computer vision is multi-object tracking. Some recent efforts were trying to infer and recover both short-term and long-term occluded objects by occlusion assumption [ZLN08, AS11]. In this work, we use [ZLN08] as the baseline representing the state-of-the-art multi-object tracking algorithm with occlusion assumptions, which adopted an Explicit Occlusion Model (EOM) to track with long-term inter-object occlusions, adding occluded object hypothesis to model occlusions.

Full Model: The results of full model are shown in Figure 5.17 (e). Benefit from both spatial and temporal terms with back propagation, most of the occlusions were successfully recovered. The failure cases happened when the object was transferred continuously between containers without any valid object detection in space. For example, from frame 265-320, the ball was passed from actor 1 to actor 2 and then passed to actor 3. Later, at frame 320, the ball was passed back to actor 1. In this case, the ball was not detected during the entire process. As the result, our method believed the ball was in the hand of actor 1 all the time.

Results: To evaluate our method quantitatively, we extract two subsets of the video clips from the proposed dataset based on two causes of occlusions: contained by another object and the blocked camera views. We evaluate the accuracy on these two subsets as well as on the entire dataset.

Success rate was adopted for quantitative analysis, defined as the ratio between the number of frames with correct object localization and the number of all frames. Given an estimated bounding box of an object b_e and the ground truth bounding box b_g , the overlap

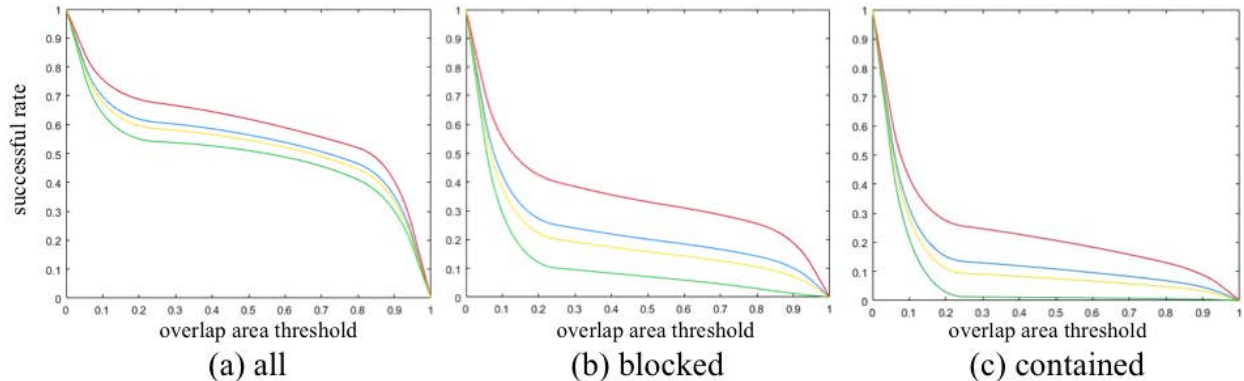


Figure 5.18: Different overlap ratios evaluated on different subsets. The red, yellow, green, and blue line represent the results of full model, baseline1, baseline2, and baseline3, respectively. The horizontal axis is the threshold axis, ranging from 0 to 1. The vertical axis is the success rate.

score is defined as $r = \frac{b_e \cap b_g}{b_e \cup b_g}$, where \cap and \cup are the intersection and union of two regions. An object bounding box is considered correct if $r \geq r_0$. The accuracy of the tracking results are shown in Table 5.2 with $r_0 = 0.5$. We further evaluate success rate when varying different overlap ratios r_0 . Results are shown in in Figure 5.18.

5.2.3.5 Evaluations on Existing Datasets

In addition to our proposed dataset designed for tracking severe objects which are “contained” or “blocked”, we further test our method on some existing datasets for modeling HOI: CAD-120 [SPS12], CMU interaction dataset [GKD09], MSR action recognition dataset [YLW09], and NW-UCLA Multiview Action 3D dataset [WNX14]. The major differences between these four datasets and other public available datasets (*e.g.*, the multiple objects tracking datasets) is: these four datasets focus on rich HOI, severe occlusions between human and objects, and large appearance variations of object, which is the main focus of this work.

To evaluate our method on these datasets, we apply the RGB-D detectors [SX13] for RGB-D datasets [SPS12, WNX14, YLW09], and RGB detectors [KMM12] for RGB-only

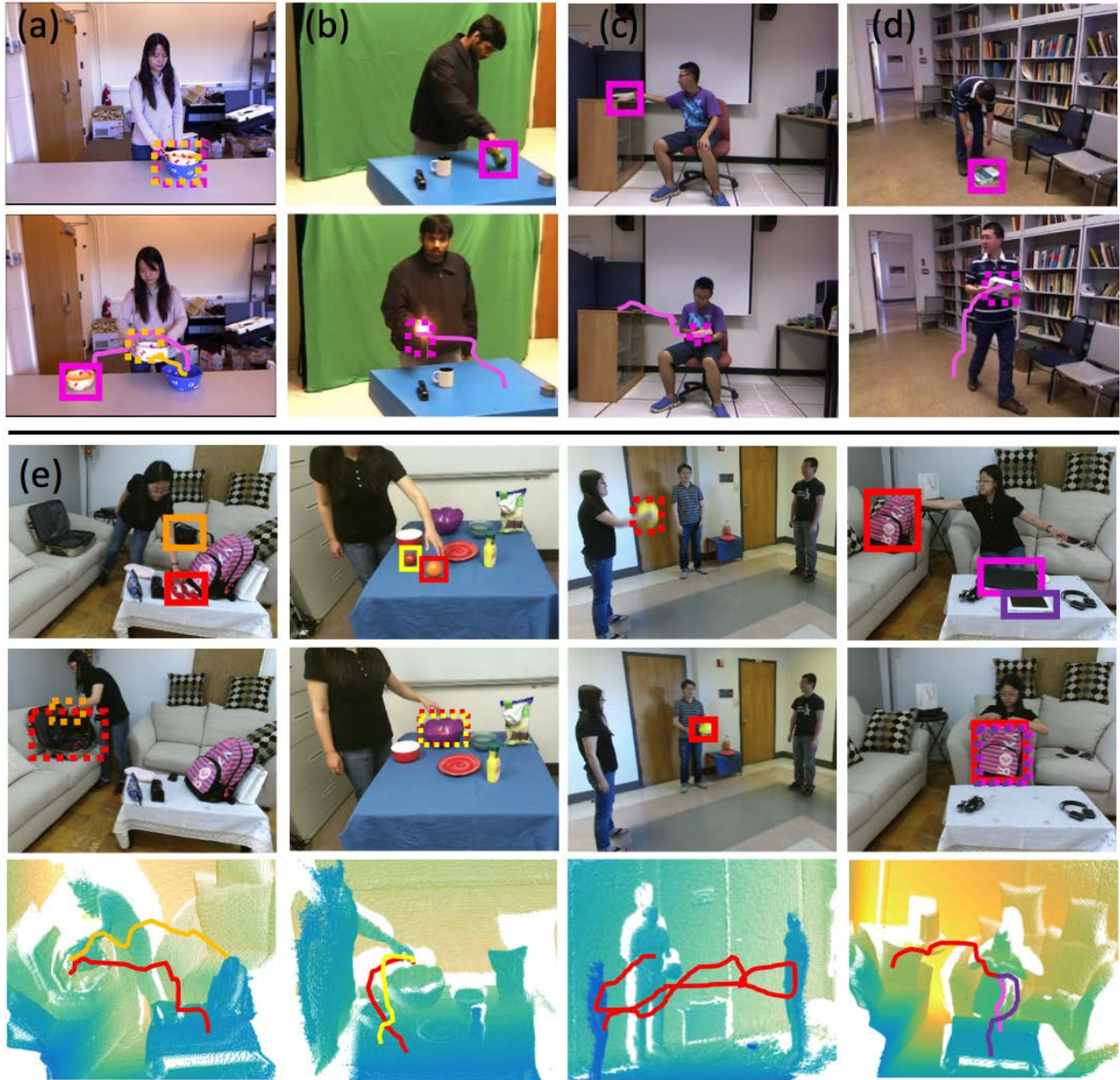


Figure 5.19: More qualitative results. Solid boxes are detected by tracking algorithm and the dashed boxes are inferred. Top two rows: (a) CAD-120, (b) CMU Dataset, (c) MSR Dataset, and (d) NW-UCLA Dataset. The bottom three rows (e) are the results on our proposed occluded objects dataset.

dataset [GKD09]. For action detection, we train a classifier on 2D data for CAD 120 and CMU interaction datasets which have no skeleton data. Examples of qualitative results are shown in Figure 5.19.

Table 5.3: Tracking accuracy on other datasets.

	baseline1	baeline2	baseline3	full
CAD-120	0.30	0.13	0.33	0.47
CMU	0.28	0.12	0.25	0.43
MSR	0.43	0.21	0.44	0.60
NW-UCLA	0.56	0.25	0.56	0.72

The quantitative tracking accuracy is shown in Table 5.3. The performance of our method on MSR action recognition and Northwestern-UCLA dataset is better than the results on CAD-120 and CMU dataset. We believe two reasons contributed to the performance differences: i) Some errors were caused by the unreliable action detections in 2D space compared to 3D space. ii) Small object detections are more challenging in 2D cases, such as pouring from a cup, lighting a flash light in the CMU dataset.

5.2.4 Conclusions and Discussions

We propose an algorithm to infer occluded objects and recover the incomplete trajectories for objects in a cluttered indoor scene by reasoning about containment relations and human actions. We assume that the movements of objects are only caused by human actions, and explicitly model occlusions from two causes: contained by others, or blocked camera views. A network flow representation is adopted to globally optimize trajectories based on two occlusion causes. In the experiment, we test our method on the collected occluded objects dataset and other four existing datasets, demonstrating the proposed method can provide better performance in challenging scenarios.

The current work is limited in the following aspects:

i) When the object detection is noisy, the performance of our method is likely to degenerate, especially when continuous transitions between occluded objects happen. High level knowledge may help to improve the results, *e.g.*, integrating an inference algorithm for the intention of the agent.

ii) We currently limit the scenarios where human is the only cause that leads to the object status changes, thus are unable to handle situations where objects move only by invisible force field, *e.g.*, gravity. Such challenging situations would require a much deeper understanding of the 3D scenes, particular the “dark matter” that is invisible [SXR15], *e.g.*, functionality [ZZY13, ZZC15] and causality [FZ13a].

iii) The majority of computer vision community is focusing on rigid body. However, properly modeling fluid (*e.g.*, water [BBY15, KJZ16]) and granular material (*e.g.*, sand [KZJ17]) is important for inferring containment relations.

CHAPTER 6

Scene Synthesis by Integrating Functionality and Affordance

Recent advances in visual recognition and classification through machine-learning-based vision algorithms have yielded similar or even better than human performance (*e.g.*, [HZR15, EEV15]) by leveraging large-scale, ground-truth-labeled RGB datasets [DDS09, LMB14]. However, indoor scene understanding remains a largely unsolved challenge due in part to the limitations of appropriate RGB-D datasets available for training purposes. To date, the most commonly used RGB-D dataset for scene understanding is the NYU-Depth V2 dataset [SHK12], which comprises only 464 scenes with only 1449 labeled RGB-D pairs provided while the remaining 407,024 pairs are unlabeled. This is clearly insufficient for the supervised training of modern computer vision methods, especially those based on deep learning. Furthermore, the manual labeling of per-pixel ground truth information, including the (crowd-sourced) labeling of RGB-D images captured by Kinect-like sensors, is tedious and error-prone, limiting both its quantity and accuracy.

To address this deficiency, recent years have seen the increased use of synthetic image datasets as training data, but little effort has been devoted to the learning-based systematic generation of massive quantities of sufficiently complex synthetic indoor scenes for the purposes of training scene understanding algorithms. This is partially due to the difficulties of (i) devising sampling processes capable of generating diverse scene configurations, and (ii) the intensive computational costs of photorealistically rendering large-scale scenes. Aside from a few efforts, reviewed in section 6.1, in generating small-scale synthetic scenes, the most notable work was recently reported by Song *et al.* [SYZ17b], in which a large scene layout dataset was downloaded from the Planner5D website.



Figure 6.1: (Top Left) An example automatically-generated 3D bedroom scene, rendered as a photorealistic RGB image, along with per-pixel ground truth of surface normal, depth, and object identity. (Top Right) Another synthesized bedroom scene. Synthesized scenes include fine details—objects (*e.g.*, the duvet and pillows on beds) and their textures are changeable, by sampling physical parameters of materials (reflectance, roughness, glossiness, *etc.*), and illumination parameters are sampled from continuous spaces of possible positions, intensities, and colors. (Bottom) Rendered images of 4 example synthetic indoor scenes.

By comparison, our work is unique in that we devise a complete learning-based pipeline for synthesizing large scale *learning-based configurable* scene layouts via intelligent sampling, as well as the photorealistic physics-based rendering of these scenes with associated per-pixel ground truth to serve as training data. Our pipeline has the following characteristics:

- By utilizing a stochastic grammar model, one represented by an attributed Spatial And-Or Graph (S-AOG), our sampling algorithm combines hierarchical compositions and contextual constraints to enable the systematic generation of 3D scenes with high variability, not only at the scene level (*e.g.*, control of size of the room and the number of objects within), but also at the object level (*e.g.*, control of the material properties of individual object parts).

- As Figure 6.1 shows, we employ state-of-the-art physics-based rendering, yielding photorealistic synthetic images. Our advanced rendering enables the systematic sampling of an infinite variety of environmental conditions and attributes, including illumination conditions (positions, intensities, colors, *etc.*, of the light sources), camera parameters (Kinect, fisheye, panorama, camera models and depth of field, *etc.*), and object properties (color, texture, reflectance, roughness, glossiness, *etc.*).

Since our synthetic data is generated in a forward manner—by rendering 2D images from 3D scenes of detailed geometric object models—ground truth information is naturally available without the need for any manual labeling. Hence, not only are our rendered images highly realistic, but they are also accompanied by accurate, per-pixel ground truth color, depth, surface normals, and object labels.

In our experimental study, we demonstrate the usefulness of our dataset by improving the performance in certain scene understanding tasks, showcasing the prediction of surface normals from RGB images, as well as the depth prediction from RGB images. Furthermore, by modifying object attributes and scene properties in a controllable manner, we provide benchmarks and diagnostics of trained models for common scene understanding tasks; *e.g.*, depth and surface normal prediction, semantic segmentation, reconstruction, *etc.*

6.1 Related Work

Synthetic image datasets have recently been a source of training data for object detection and correspondence matching [SGS10, SS14, SX14, FKI14, DFI15, PSA15, ZKA16, GWC16, MKS16, QSN16], single-view reconstruction [HWK15], view-point estimation [MSB14, SQL15], 2D human pose estimation [PJA12, RLB15, Qiu16], 3D human pose estimation [SSK13, SVD03, YIK16, DWL16, GKS16, RS16, ZZL16, CWL16, VRM17], depth prediction [SHM14], pedestrian detection [MVG10, PJW11, VLM14, HNK15], action recognition [RM15, RM16, SGC17], semantic segmentation [RVR16], scene understanding [HPS16, KIX16, QY16, HPB16], and in benchmark data sets [HWM14]. Previously, synthetic imagery, generated on the fly, online, had been used in visual surveillance [QT08] and active

vision / sensorimotor control [TR95]. Although prior work demonstrates the potential of synthetic imagery to advance computer vision research, to our knowledge no large synthetic RGB-D dataset of *learning-based configurable* indoor scenes has yet been released.

3D layout synthesis algorithms [YYT11, HPS16] have been developed to optimize furniture arrangements based on pre-defined constraints, where the number and categories of objects are pre-specified and remain the same. By contrast, we sample individual objects and create entire indoor scenes from scratch. Some work has studied fine-grained object arrangement to address specific problems; *e.g.*, utilizing user-provided examples to arrange small objects [FRS12, YYT16], and optimizing the number of objects in scenes using LARJ-MCMC [YYW12]. To enhance realism, Merrell *et al.* [MSL11] developed an interactive system that provides suggestions according to interior design guidelines.

Image synthesis has been attempted using various deep neural network architectures, including recurrent neural networks (RNN) [GDG15], generative adversarial networks (GAN) [WG16, RMC15], inverse graphics networks [KWK15], and generative convolutional networks [LZW16, XLZ16b, XLZ16a]. However, images of indoor scenes synthesized by these models often suffer from glaring artifacts, such as blurred patches. More recently, some applications of general purpose inverse graphics solutions using probabilistic programming languages have been reported [MKP13, LB14, KKT15]. However, the problem space is enormous, and the quality of inverse graphics “renderings” is disappointingly low and slow.

Stochastic scene grammar models have been used in computer vision to recover 3D structures from single-view images for both indoor [ZZ13, LZZ14] and outdoor [LZZ14] scene parsing. In the present work, instead of solving visual inverse problems, we sample from the grammar model to synthesize, in a forward manner, large varieties of 3D indoor scenes.

Domain adaptation Although the presented work does not directly involve domain adaptation, this plays an important role in learning from synthetic data, as the goal of using synthetic data is to transfer the learned knowledge and apply it to real-world scenarios. A

review of existing work in this area is beyond the scope of this work; we refer the reader to a recent comprehensive survey [Csu17]. Traditionally, the widely used techniques for domain adaptation can be divided into four categories: i) covariate shift with shared support [Hec77, GSH09, CMR08, BBS09], ii) learning shared representations [BMP06, BBC07, MMR09], iii) feature-based learning [EP04, Dau07, WDL09], and iv) parameter-based learning [CH05b, YTS05, XLC07, Dau09]. With the recent boost of deep learning, researchers have started to apply deep features to domain adaptation (*e.g.*, [GL15, THD15]).

Contributions

The present work makes five major contributions:

1. To our knowledge, ours is the first work that, for the purposes of indoor scene understanding, introduces a *learning-based configurable* pipeline for generating massive quantities of photorealistic images of indoor scenes with perfect per-pixel ground truth, including color, surface depth, surface normal, and object identity. The parameters and constraints are automatically learned from the SUNCG [SYZ17b] and ShapeNet [CFG15b] datasets.
2. For scene generation, we propose the use of a stochastic grammar model in the form of an attributed Spatial And-Or graph (S-AOG). Our model supports the arbitrary addition and deletion of objects and modification of their categories, yielding significant variation in the resulting collection of synthetic scenes.
3. By precisely customizing and controlling important attributes of the generated scenes, we provide a set of diagnostic benchmarks of previous work on several common computer vision tasks. To our knowledge, this is the first work to provide comprehensive diagnostics with respect to algorithm stability and sensitivity to certain scene attributes.
4. We have released a large dataset, AOGIndoor, which consists of rendered 2D images with depth, surface normals, and segmentation, as well as 3D layouts and models.

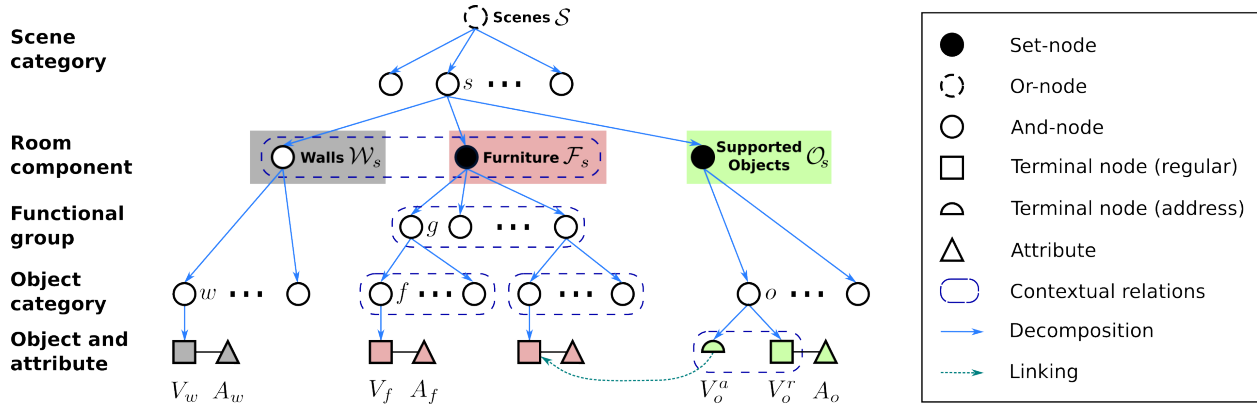


Figure 6.2: Scene grammar as an attributed S-AOG. The terminal nodes of the S-AOG are attributed with internal attributes (sizes) and external attributes (positions and orientations). A supported object node is combined by an address terminal node and a regular terminal node, indicating that the object is supported by the furniture pointed to by the address node. If the value of the address node is null, the object is situated on the floor. Contextual relations are defined between walls and furniture, among different furniture, between supported objects and supporting furniture, and for functional groups.

5. We demonstrate the effectiveness of the proposed synthesized scene dataset by advancing the state-of-the-art in the prediction of surface normals and depth.

6.2 Representation and Formulation

6.2.1 Representation: Attributed Spatial And-Or Graph

A scene model should be capable of: (i) representing the compositional/hierarchical structure of indoor scenes, and (ii) capturing the rich contextual relationships between different components of the scene. Specifically,

- *Compositional hierarchy* of the indoor scene structure is embedded in a graph-based model to model the decomposition into sub-components and the switch among multiple alternative sub-configurations. In general, an indoor scene can first be categorized into

different indoor settings (*i.e.*, bedrooms, bathrooms, *etc.*), each of which has a set of walls, furniture, and supported objects. Furniture can be decomposed into functional groups that are composed of multiple pieces of furniture; *e.g.*, a “work” functional group consists of a desk and a chair.

- *Contextual relations* between pieces of furniture are helpful in distinguishing the functionality of each furniture item and furniture pairs, providing a strong constraint for representing natural indoor scenes. In this work, we consider four types of contextual relations: (i) relations between furniture and walls; (ii) relations among furniture; (iii) relations between supported objects and their supporting objects (*e.g.*, monitor and desk); and (iv) relations between objects of a functional pair (*e.g.*, sofa and TV).

Representation: We represent the hierarchical structure of indoor scenes by an attributed Spatial And-Or Graph (S-AOG), which is a Stochastic Context-Sensitive Grammar (SCSG) with attributes on the terminal nodes. An example is shown in Figure 6.2. This representation combines (i) a stochastic context-free grammar (SCFG) and (ii) contextual relations defined on a Markov random field (MRF); *i.e.*, the horizontal links among the terminal nodes. The S-AOG represents the hierarchical decompositions from scenes (top level) to objects (bottom level), whereas contextual relations encode the spatial and functional relations through horizontal links between nodes.

Definitions: Formally, an S-AOG is denoted by a 5-tuple: $\mathcal{G} = \langle S, V, R, P, E \rangle$, where S is the root node of the grammar, $V = V_{\text{NT}} \cup V_{\text{T}}$ is the vertex set including non-terminal nodes V_{NT} and terminal nodes V_{T} , R stands for the production rules, P represents the probability model defined on the attributed S-AOG, and E denotes the contextual relations represented as horizontal links between nodes in the same layer.

Non-terminal Nodes: The set of non-terminal nodes $V_{\text{NT}} = V^{\text{And}} \cup V^{\text{Or}} \cup V^{\text{Set}}$ is composed of three set of nodes: *And-nodes* V^{And} denoting a decomposition of a large entity, *Or-nodes* V^{Or} representing alternative decompositions, and *Set-nodes* V^{Set} of which each child

branch represents an Or-node on the number of the child object. The Set-nodes are compact representations of nested And-Or relations

Production Rules: Corresponding to three different types of non-terminal nodes, three types of production rules are defined:

- And rules for an And-node $v \in V^{\text{And}}$, are defined as a deterministic decomposition

$$v \rightarrow u_1 \cdot u_2 \cdots u_{n(v)}. \quad (6.1)$$

- Or rules for an Or-node $v \in V^{\text{Or}}$, are defined as a switch

$$v \rightarrow u_1 | u_2 \cdots | u_{n(v)}, \quad (6.2)$$

with $\rho_1 | \rho_2 \cdots | \rho_{n(v)}$.

- Set rules for a Set-node $v \in V^{\text{Set}}$ are defined as

$$v \rightarrow (\text{nil} | u_1^1 | u_1^2 | \cdots) \cdots (\text{nil} | u_{n(v)}^1 | u_{n(v)}^2 | \cdots), \quad (6.3)$$

with $(\rho_{1,0} | \rho_{1,1} | \rho_{1,2} | \cdots) \cdots (\rho_{n(v),0} | \rho_{n(v),1} | \rho_{n(v),2} | \cdots)$, where u_i^k denotes the case that object u_i appears k times, and the probability is $\rho_{i,k}$.

Terminal Nodes: The set of terminal nodes can be divided into two types: (i) regular terminal nodes $v \in V_T^r$ representing spatial entities in a scene, with attributes A divided into internal A_{in} (size) and external A_{ex} (position and orientation) attributes, and (ii) address terminal nodes $v \in V_T^a$ that point to regular terminal nodes and take values in the set $V_T^r \cup \{\text{nil}\}$. These latter nodes avoid excessively dense graphs by encoding interactions that occur only in a certain context [Fri03].

Contextual Relations: The contextual relations $E = E_w \cup E_f \cup E_o \cup E_g$ among nodes are represented by horizontal links in the AOG. The relations are divided into four subsets:

- relations between furniture and walls E_w ;

- relations among furniture E_f ;
- relations between supported objects and their supporting objects E_o (e.g., monitor and desk); and
- relations between objects of a functional pair E_g (e.g., sofa and TV).

Accordingly, the cliques formed in the terminal layer may also be divided into four subsets:

$$C = C_w \cup C_f \cup C_o \cup C_g.$$

Note that the contextual relations of nodes will be inherited from their parents; hence, the relations at a higher level will eventually collapse into cliques C among the terminal nodes. These contextual relations also form an MRF on the terminal nodes. To encode the contextual relations, we define different types of potential functions for different kinds of cliques.

Parse Tree: A hierarchical parse tree pt instantiates the S-AOG by selecting a child node for the Or-nodes as well as determining the state of each child node for the Set-nodes. A parse graph pg consists of a parse tree pt and a number of contextual relations E on the parse tree: $pg = (pt, E_{pt})$. Figure 6.3 illustrates a simple example of a parse graph and four types of cliques formed in the terminal layer.

6.2.2 Probabilistic Formulation

The purpose of representing indoor scenes using an S-AOG is to bring the advantages of compositional hierarchy and contextual relations to the generation of realistic and diverse novel/unseen scene configurations from a learned S-AOG. In this section, we introduce the related probabilistic formulation.

Prior: We define the prior probability of a scene configuration generated by an S-AOG with the parameter set Θ . A scene configuration is represented by pg , including objects in the scene and their attributes. The prior probability of pg generated by an S-AOG parameterized

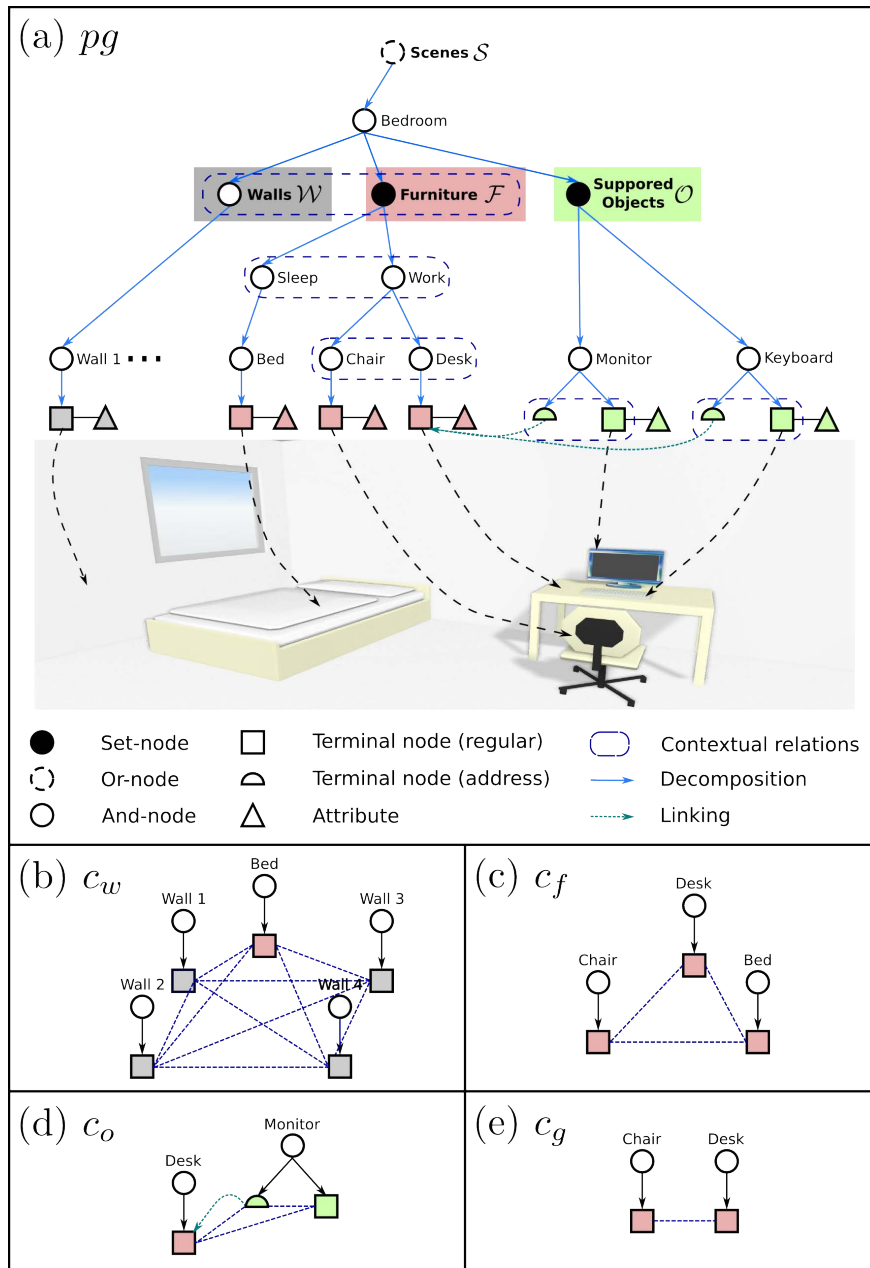


Figure 6.3: (a) A simplified example of a parse graph of a bedroom. The terminal nodes of the parse graph form an MRF in the bottom layer. Cliques are formed by the contextual relations projected to the bottom layer. (b)–(e) give an example of the four types of cliques, which represent different contextual relations.

by Θ is formulated as a Gibbs distribution

$$p(pg|\Theta) = \frac{1}{Z} \exp\{-\mathcal{E}(pg|\Theta)\} \quad (6.4)$$

$$= \frac{1}{Z} \exp\{-\mathcal{E}(pt|\Theta) - \mathcal{E}(E_{pt}|\Theta)\}, \quad (6.5)$$

where $\mathcal{E}(pg|\Theta)$ is the energy function of the parse graph, $\mathcal{E}(pt|\Theta)$ is the energy function of a parse tree, and $\mathcal{E}(E_{pt}|\Theta)$ is the energy function of the contextual relations. Here, $\mathcal{E}(pt|\Theta)$ is defined as combinations of probability distributions with closed-form expressions, and $\mathcal{E}(E_{pt}|\Theta)$ is defined as potential functions relating to the external attributes of the terminal nodes.

Energy of Parse Tree: Energy $\mathcal{E}(pt|\Theta)$ is further decomposed into energy functions of different types of non-terminal nodes, and energy functions of internal attributes of both regular and address terminal nodes:

$$\mathcal{E}(pt|\Theta) = \underbrace{\sum_{v \in V^{\text{Or}}} \mathcal{E}_{\Theta}^{\text{Or}}(v) + \sum_{v \in V^{\text{Set}}} \mathcal{E}_{\Theta}^{\text{Set}}(v)}_{\text{non-terminal nodes}} + \underbrace{\sum_{v \in V_T^r} \mathcal{E}_{\Theta}^{A_{\text{in}}}(v)}_{\text{terminal nodes}}, \quad (6.6)$$

where the choice of child node of an Or-node $v \in V^{\text{Or}}$ follows a multinomial distribution, and each child branch of a Set-Node $v \in V^{\text{Set}}$ follows a Bernoulli distribution. Note that the And-nodes are deterministically expanded; hence, Equation 6.6 lacks an energy term for the And-nodes. The internal attributes A_{in} (size) of terminal nodes follows a non-parametric probability distribution learned via kernel density estimation.

Energy of Contextual Relations: The energy $\mathcal{E}(E_{pt}|\Theta)$ is described by the probability distribution

$$p(E_{pt}|\Theta) = \frac{1}{Z} \exp\{-\mathcal{E}(E_{pt}|\Theta)\} \quad (6.7)$$

$$= \prod_{c \in C_w} \phi_w(c) \prod_{c \in C_f} \phi_f(c) \prod_{c \in C_o} \phi_o(c) \prod_{c \in C_g} \phi_g(c), \quad (6.8)$$

which combines the potentials of the four types of cliques formed in the terminal layer. The potentials of these cliques are computed based on the external attributes of regular terminal nodes:

- Potential function $\phi_w(c)$ is defined on relations between walls and furniture (Figure 6.3(b)). A clique $c \in C_w$ includes a terminal node representing a piece of furniture f and the terminal nodes representing walls $\{w_i\}$: $c = \{f, \{w_i\}\}$. Assuming pairwise object relations, we have

$$\phi_w(c) = \frac{1}{Z} \exp\left\{-\lambda_w \cdot \underbrace{\sum_{w_i \neq w_j} l_{\text{con}}(w_i, w_j)}_{\text{constraint between walls}} \right. \\ \left. \underbrace{\sum_{w_i} [l_{\text{dis}}(f, w_i) + l_{\text{ori}}(f, w_i)]}_{\text{constraint between walls and furniture}} >\right\}, \quad (6.9)$$

where λ_w is a weight vector, and l_{con} , l_{dis} , l_{ori} are three different cost functions:

- The cost function $l_{\text{con}}(w_i, w_j)$ defines the consistency between the walls; *i.e.*, adjacent walls should be connected, whereas opposite walls should have the same size. Although this term is repeatedly computed in different cliques, it is usually zero as the walls are enforced to be consistent in practice.
- The cost function $l_{\text{dis}}(x_i, x_j)$ defines the geometric distance compatibility between two objects

$$l_{\text{dis}}(x_i, x_j) = |d(x_i, x_j) - \bar{d}(x_i, x_j)|, \quad (6.10)$$

where $d(x_i, x_j)$ is the distance between object x_i and x_j , and $\bar{d}(x_i, x_j)$ is the mean distance learned from all the examples.

- Similarly, the cost function $l_{\text{ori}}(x_i, x_j)$ is defined as

$$l_{\text{ori}}(x_i, x_j) = |\theta(x_i, x_j) - \bar{\theta}(x_i, x_j)|, \quad (6.11)$$

where $\theta(x_i, x_j)$ is the distance between object x_i and x_j , and $\bar{\theta}(x_i, x_j)$ is the mean distance learned from all the examples. This terms represents the compatibility between two objects in terms of the relative orientations.

- Potential function $\phi_f(c)$ is defined on relations between pieces of furniture (Figure 6.3(c)). A clique $c \in C_f$ includes all the terminal nodes representing a piece of furniture:

$c = \{f_i\}$. Hence,

$$\phi_f(c) = \frac{1}{Z} \exp\{-\lambda_c \sum_{f_i \neq f_j} l_{\text{occ}}(f_i, f_j)\}, \quad (6.12)$$

where the cost function $l_{\text{occ}}(f_i, f_j)$ defines the compatibility of two pieces of furniture in terms of occluding accessible space

$$l_{\text{occ}}(f_i, f_j) = \max(0, 1 - d(f_i, f_j)/d_{\text{acc}}). \quad (6.13)$$

- Potential function $\phi_o(c)$ is defined on relations between a supported object and the furniture that supports it (Figure 6.3(d)). A clique $c \in C_o$ consists of a supported object terminal node o , the address node a connected to the object, and the furniture terminal node f pointed to by the address node: $c = \{f, a, o\}$

$$\phi_o(c) = \frac{1}{Z} \exp\{-\lambda_o \cdot \langle l_{\text{pos}}(f, o), l_{\text{ori}}(f, o), l_{\text{add}}(a) \rangle\}, \quad (6.14)$$

which incorporates three different cost functions. The cost function $l_{\text{ori}}(f, o)$ has been defined with potential function $\phi_w(c)$, and two new cost functions are:

- The cost function $l_{\text{pos}}(f, o)$ defines the relative position of the supported object o to the four boundaries of the bounding box of the supporting furniture f

$$l_{\text{pos}}(f, o) = \sum_i l_{\text{dis}}(f_{\text{face}_i}, o). \quad (6.15)$$

- The cost term $l_{\text{add}}(a)$ is the negative log probability of an address node $v \in V_T^a$, which is regarded as a certain regular terminal node and follows a multinomial distribution.

- Potential function $\phi_g(c)$ is defined for furniture in the same functional group (Figure 6.3(d)). A clique $c \in C_g$ consists of terminal nodes representing furniture in a functional group g : $c = \{f_i^g\}$

$$\phi_g(c) = \frac{1}{Z} \exp\{-\sum_{f_i^g \neq f_j^g} (\lambda_g \cdot \langle l_{\text{dis}}(f_i^g, f_j^g), l_{\text{ori}}(f_i^g, f_j^g) \rangle)\}. \quad (6.16)$$

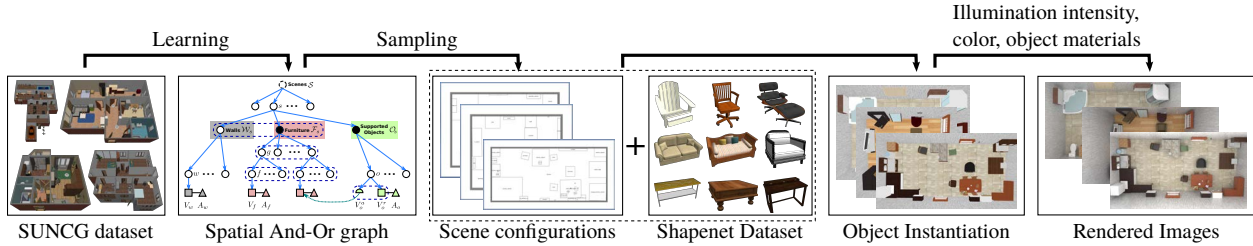


Figure 6.4: The learning-based pipeline for synthesizing images of indoor scenes.

6.3 Learning, Sampling and Synthesis

Before introducing the algorithm for learning all the parameters associated with an S-AOG, in subsection 6.3.1, note that our configurable scene synthesis pipeline includes the following components:

- A sampling algorithm based on the learned S-AOG for synthesizing realistic scene geometric configurations. This sampling algorithm controls the size of the individual objects as well as their pair-wise relations. More complex relations are recursively formed using pair-wised relations. The details are found in subsection 6.3.2.
- An attribute assignment process, which sets different material attributes to each object part, as well as various camera parameters and illuminations of the environment. The details are found in subsection 6.3.4.

The above two components are the essence of *configurable* scene synthesis; the first generates the structure of the scene while the second controls its detailed attributes. In between these two components, a scene instantiation process is applied to generate a 3D mesh of the scene based on the sampled scene layout. This step is described in subsection 6.3.3. Figure 6.4 illustrates the pipeline. At the end of this section, we showcase several examples of synthesized scenes with different configurable attributes.

6.3.1 Learning the S-AOG

The parameters Θ of a probability model can be learned in a supervised way from a set of N observed parse trees $\{pt_n, n = 1, 2, \dots, N\}$ by maximum likelihood estimation (MLE)

$$\Theta^* = \arg \max_{\Theta} \prod_{n=1}^N p(pt_n | \Theta). \quad (6.17)$$

We now describe how to learn all the parameters Θ , with the focus on learning the weights of the loss functions.

Weights of the Loss Functions: Recall that the probability distribution of cliques formed in the terminal layer is given by Equation 6.8

$$p(E_{pt} | \Theta) = \frac{1}{Z} \exp\{-\mathcal{E}(E_{pt} | \Theta)\} \quad (6.18)$$

$$= \frac{1}{Z} \exp\{-\langle \lambda, l(E_{pt}) \rangle\}, \quad (6.19)$$

where λ is the weight vector and $l(E_{pt})$ is the loss vector given by four different types of potential functions. To learn the weight vector, the traditional MLE maximizes the average log-likelihood:

$$\mathcal{L}(E_{pt} | \Theta) = \frac{1}{N} \sum_{n=1}^N \log p(E_{pt_n} | \Theta) \quad (6.20)$$

$$= -\frac{1}{N} \sum_{n=1}^N \langle \lambda, l(E_{pt_n}) \rangle - \log Z. \quad (6.21)$$

The average log-likelihood is usually maximized by following the gradient:

$$\begin{aligned} & \frac{\partial \mathcal{L}(E_{pt} | \Theta)}{\partial \lambda} \\ &= -\frac{1}{N} \sum_{n=1}^N l(E_{pt_n}) - \frac{\partial \log Z}{\partial \lambda} \end{aligned} \quad (6.22)$$

$$= -\frac{1}{N} \sum_{n=1}^N l(E_{pt_n}) - \frac{\partial \log \sum_{pt} \exp\{-\langle \lambda, l(E_{pt}) \rangle\}}{\partial \lambda} \quad (6.23)$$

$$= -\frac{1}{N} \sum_{n=1}^N l(E_{pt_n}) + \sum_{pt} \frac{1}{Z} \exp\{-\langle \lambda, l(E_{pt}) \rangle\} l(E_{pt}) \quad (6.24)$$

$$= -\frac{1}{N} \sum_{n=1}^N l(E_{pt_n}) + \frac{1}{\tilde{N}} \sum_{\tilde{n}=1}^{\tilde{N}} l(E_{pt_{\tilde{n}}}), \quad (6.25)$$

where $\{E_{pt_{\tilde{n}}}\}_{\tilde{n}=1, \dots, \tilde{N}}$ is the set of synthesized examples from the current model.

Unfortunately, it is computationally infeasible to sample a Markov chain that burns into an *equilibrium distribution* at every iteration of gradient ascent. Hence, instead of waiting for the Markov chain to converge, we adopt the contrastive divergence (CD) learning that follows the gradient of the difference of two divergences [Hin02]:

$$\text{CD}_{\tilde{N}} = \text{KL}(p_0||p_\infty) - \text{KL}(p_{\tilde{n}}||p_\infty), \quad (6.26)$$

where $\text{KL}(p_0||p_\infty)$ is the Kullback-Leibler divergence between the data distribution p_0 and the model distribution p_∞ , and $p_{\tilde{n}}$ is the distribution obtained by a Markov chain started at the data distribution and run for a small number \tilde{n} of steps (*e.g.*, $\tilde{n} = 1$).

Contrastive divergence learning has been applied effectively in addressing various problems, most notably in the context of Restricted Boltzmann Machines [HS06]. Both theoretical and empirical evidence shows its efficiency while maintaining a very small bias [CH05a]. The gradient of the contrastive divergence is given by:

$$\begin{aligned} \frac{\partial \text{CD}_{\tilde{N}}}{\partial \lambda} &= \frac{1}{N} \sum_{n=1}^N l(E_{pt_n}) - \frac{1}{\tilde{N}} \sum_{\tilde{n}=1}^{\tilde{N}} l(E_{pt_{\tilde{n}}}) \\ &\quad - \frac{\partial p_{\tilde{n}}}{\partial \lambda} \frac{\partial \text{KL}(p_{\tilde{n}}||p_\infty)}{\partial p_{\tilde{n}}}. \end{aligned} \quad (6.27)$$

Extensive simulations [Hin02] showed that the third term can be safely ignored since it is small and seldom opposes the resultant of the other two terms.

Finally, the weight vector is learned by gradient descent computed by generating a small number \tilde{n} of examples from the Markov chain

$$\lambda_{t+1} = \lambda_t - \eta_t \frac{\partial \text{CD}_{\tilde{N}}}{\partial \lambda} \quad (6.28)$$

$$= \lambda_t + \eta_t \left(\frac{1}{\tilde{N}} \sum_{\tilde{n}=1}^{\tilde{N}} l(E_{pt_{\tilde{n}}}) - \frac{1}{N} \sum_{n=1}^N l(E_{pt_n}) \right). \quad (6.29)$$

Or-nodes and Address-nodes: The MLE of the branching probabilities of Or-nodes and address terminal nodes is simply the frequency of each alternative choice [ZM07]

$$\rho_i = \frac{\#(v \rightarrow u_i)}{\sum_{j=1}^{n(v)} \#(v \rightarrow u_j)}. \quad (6.30)$$

However, the samples we draw from the distributions will rarely cover all possible terminal nodes to which an address node is pointing, since there are many unseen but plausible configurations. For instance, an apple can be put on a chair, which is physically and semantically plausible, but the training examples are highly unlikely to include such a case. Inspired by the Dirichlet process, we address this issue by altering the MLE to include a small probability α for all branches

$$\rho_i = \frac{\#(v \rightarrow u_i) + \alpha}{\sum_{j=1}^{n(v)} (\#(v \rightarrow u_j) + \alpha)}. \quad (6.31)$$

Set-nodes: Similarly, for each child branch of the Set-nodes, we use the frequency of samples as the probability if it is non-zero, otherwise we set the probability to be turned on as α . Based on the common practice—*e.g.*, choosing the probability of joining a new table in the Chinese restaurant process [Ald85]—we set α to have probability 1.

Parameters: To learn the S-AOG for sampling purposes, we use the SUNCG dataset [SYZ17b] to collect statistics, which contains over 45K different scenes with manually created realistic room and furniture layouts. We collect the statistics of room types, room sizes, furniture occurrences, furniture sizes, relative distances and orientations between furniture and walls, furniture affordance, grouping occurrences, and supporting relations.

The parameters of the loss functions are learned from the constructed scenes by computing the statistics of relative distances and relative orientations between different objects.

The grouping relations are manually defined (*e.g.*, nightstands are associated with beds, chairs are associated with desks and tables). We examine each pair of furniture pieces in the scene, and a pair is regarded as a group if the distance of the pieces is smaller than a

threshold (*e.g.*, 1m). The probability of occurrence is learned as a multinomial distribution. The supporting relations are automatically discovered from the dataset by computing the vertical distance between pairs of objects and checking if one bounding polygon contains another.

The distribution of object size among all the furniture and supported objects is learned from the 3D models provided by the ShapeNet dataset [CFG15a] and the SUNCG dataset [SYZ17b]. We first extracted the size information from the 3D models, and then fitted a non-parametric distribution using kernel density estimation (KDE). Not only is this more accurate than simply fitting a trivariate normal distribution, but it is also easier to sample from.

6.3.2 Sampling Scene Geometry Configurations

Based on the learned S-AOG, we sample scene configurations (parse graphs) based on the prior probability $p(pg|\Theta)$ using a Markov Chain Monte Carlo (MCMC) sampler. The sampling process comprises two major steps:

1. Top-down sampling of the parse tree structure pt and internal attributes of objects. This step selects a branch for each Or-node as well as chooses a child branch for every Set-node. In addition, internal attributes (sizes) of each regular terminal node are also sampled. Note that this can be easily done by sampling from closed-form distributions.
2. MCMC sampling of the external attributes (positions and orientations) of objects as well as the values of the address nodes. Samples are proposed by Markov chain dynamics, and are taken after the Markov chain converges to the prior probability. These attributes are constrained by multiple potential functions, hence it is difficult to directly sample from the true underlying probability distribution.

algorithm 1 overviews the sampling process. Some qualitative results are shown in Figure 6.5.

Algorithm 1: Sampling Scene Configurations

Input : Attributed S-AOG \mathcal{G}

Landscape parameter β

sample number n

Output: Synthesized room layouts $\{pg_i\}_{i=1,\dots,n}$

```
1 for  $i = 1$  to  $n$  do
2   Sample the child nodes of the Set nodes and Or nodes from  $\mathcal{G}$  directly to get the
   structure of  $pg_i$ .
3   Sample the sizes of room, furniture  $f$  and objects  $o$  in  $pg_i$  directly.
4   Sample the address nodes  $V^a$ .
5   Randomly initialize positions and orientations of furniture  $f$  and objects  $o$  in  $pg_i$ .
6    $iter = 0$ 
7   while  $iter < iter_{\max}$  do
8     Propose a new move and get proposal  $pg'_i$ .
9     Sample  $u \sim \text{unif}(0, 1)$ .
10    if  $u < \min(1, \exp(\beta(\mathcal{E}(pg_i|\Theta) - \mathcal{E}(pg'_i|\Theta))))$  then
11      |  $pg_i = pg'_i$ 
12    end
13     $iter += 1$ 
14  end
15 end
```

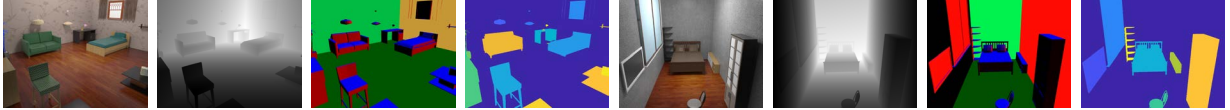
Markov Chain Dynamics: Four types of Markov chain dynamics $q_i, i = 1, 2, 3, 4$ are designed to be chosen randomly with probabilities to propose moves. Specifically, the dynamics q_1 and q_2 are diffusion, while q_3 and q_4 are reversible jumps:

1. *Translation of Objects.* Dynamic q_1 chooses a regular terminal node and samples a new position based on the current position of the object

$$\text{pos} \rightarrow \text{pos} + \delta\text{pos}, \quad (6.32)$$



(a) Different categories of the scenes using default attributes of object material, the lighting conditions, and camera parameters. Top row: top view. Bottom row: a random view.



(b) Additional examples of two bedrooms, with corresponding depth map, surface normal, and semantic segmentation.

Figure 6.5: Qualitative results in different types of scenes.

where δpos follows a bivariate normal distribution.

2. *Rotation of Objects.* Dynamic q_2 chooses a regular terminal node and samples a new orientation based on the current orientation of the object

$$\theta \rightarrow \theta + \delta\theta, \quad (6.33)$$

where $\delta\theta$ follows a normal distribution.

3. *Swapping of Objects.* Dynamic q_3 chooses two regular terminal nodes and swaps the positions and orientations of the objects.
4. *Swapping of Supporting Objects.* Dynamic q_4 chooses an address terminal node and samples a new regular furniture terminal node pointed to. We sample a new 3D location (x, y, z) for the supported object:

- Randomly sample $x = u_x * w_p$, where $u_x \sim \text{unif}(0, 1)$, and w_p is the width of the supporting object.
- Randomly sample $y = u_y * l_p$, where $u_y \sim \text{unif}(0, 1)$, and l_p is the length of the supporting object.
- The height z is simply the height of the supporting object.

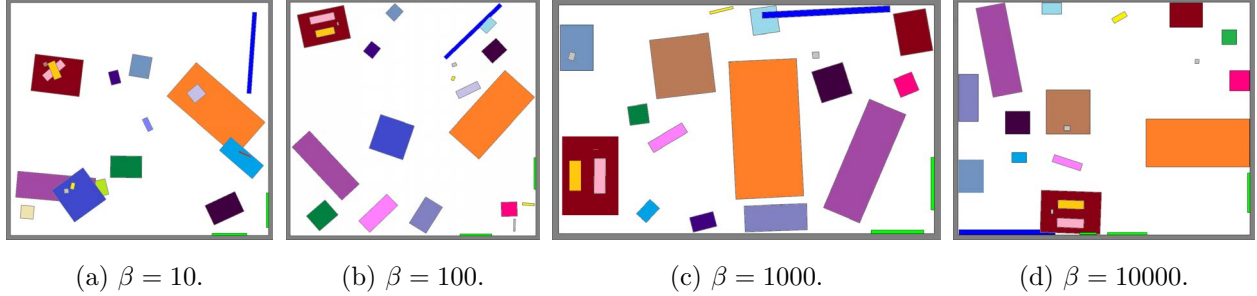


Figure 6.6: Synthesis for different values of β . Each image shows a typical configuration sampled from a Markov chain.

Adopting the Metropolis-Hastings algorithm, a newly proposed parse graph pg' is accepted according to the following acceptance probability:

$$\alpha(pg'|pg, \Theta) = \min\left(1, \frac{p(pg'|\Theta)p(pg|pg')}{p(pg|\Theta)p(pg'|pg)}\right) \quad (6.34)$$

$$= \min\left(1, \frac{p(pg'|\Theta)}{p(pg|\Theta)}\right) \quad (6.35)$$

$$= \min(1, \exp(\mathcal{E}(pg|\Theta) - \mathcal{E}(pg'|\Theta))). \quad (6.36)$$

The proposal probabilities are canceled since the proposed moves are symmetric in probability.

Convergence: To test if the Markov chain has converged to the prior probability, we keep a histogram of the energy of the last w samples. When the difference between two histograms at a distance of s sampling steps is smaller than a threshold ϵ , the Markov chain is considered to have converged.

Tidiness of Scenes: During the sampling process, a typical state is drawn from the distribution. We can easily control the level of tidiness of the sampled scenes by adding an extra parameter β to control the landscape of the prior distribution:

$$p(pg|\Theta) = \frac{1}{Z} \exp\{-\beta\mathcal{E}(pg|\Theta)\}. \quad (6.37)$$

Some examples are shown in Figure 6.6.

Note that the parameter β is analogous albeit differs from the temperature in simulated annealing optimization—the temperature in simulated annealing is time-variant; *i.e.*, it changes during the simulated annealing process. In our model, we simulate a Markov chain under one specific β to get typical samples at a certain level of tidiness. When β is small, the distribution is “smooth”; *i.e.*, the differences between local minima and local maxima are small.

6.3.3 Scene Instantiation using 3D Object Datasets

Given a generated 3D scene layout, the 3D scene is instantiated by assembling objects into it using 3D object datasets. In this work, we incorporate both ShapeNet dataset [CFG15b] and SUNCG dataset [SYZ17b] as our 3D model dataset. Scene instantiation includes five steps:

1. For each object in the scene layout, find the model has the closest the length/width ratio to the dimension specified in the scene layout.
2. Align the orientations of selected models according to the orientation specified in the scene layout.
3. Transform the models to the specified positions, and scales the models according to the generated scene layout.
4. Since we only fit the length and width in Step 1, an extra step to adjust object position along the gravity direction is needed, eliminating all the floating models and the models that penetrated into each other.
5. Add the floor, walls, and ceiling to complete the instantiated scene.

6.3.4 Scene Attribute Configurations

As we generate scenes in a forward fashion, our pipeline enables the precise customization and control of important attributes of the generated scenes. Some configurations are shown



(a) Lighting intensity: half and double

(b) Lighting color: purple and blue



(c) Different object materials: metal, gold, chocolate, and clay



(d) Different materials in each object part

(e) Using multiple light sources



(f) Fish eye lens

(g) Depth of field

(h) Panorama image



(i) Different background materials affect the rendering results

Figure 6.7: We can configure the scene with different (a) illumination intensities, (b) illumination colors, (c) materials, and (d) even on each object part. We can also control (e) the number of light source and their positions, (f) camera lenses (*e.g.*, fish eye), (g) depths of field, or (h) render the scene as a panorama for virtual reality and other virtual environments. (i) 7 different background wall textures. Note how the background affects the overall illumination.

in Figure 6.7. The rendered images are determined by combinations of the following four factors:

- Illuminations, including light source positions, intensities, colors, and the number of light sources.
- Material and texture of the environment; *i.e.*, the walls, floor and ceiling.
- Cameras, such as fisheye, panorama, and Kinect cameras, have different focal lengths and apertures, yielding dramatically different rendered images. By virtue of physics-based rendering, our pipeline can even control the F-stop and focal distance, resulting in different depths of field.
- Different object materials and textures will have various properties, represented by roughness, metallicness, and reflectivity.

6.4 Photorealistic Scene Rendering

We adopt Physics-Based Rendering (PBR) [PH04] to generate the photorealistic 2D images. PBR has become the industry standard in computer graphics applications in recent years, and has been widely adopted for both offline and real-time rendering. Unlike traditional rendering techniques where heuristic shaders are used to control how light is scattered by a surface, PBR simulates the physics of real-world light by computing the bidirectional scattering distribution function (BSDF) [BDW81] of the surface.

Formulation: Following the law of conservation of energy, PBR solves the rendering equation for the total spectral radiance of outgoing light in direction \mathbf{w} from point \mathbf{x} on a surface

$$L_o(\mathbf{x}, \mathbf{w}) = L_e(\mathbf{x}, \mathbf{w}) + \int_{\Omega} f_r(\mathbf{x}, \mathbf{w}', \mathbf{w}) L_i(\mathbf{x}, \mathbf{w}') (-\mathbf{w}' \cdot \mathbf{n}) d\mathbf{w}', \quad (6.38)$$

where L_o is the outgoing light, L_e is the emitted light (from a light source), Ω is the unit hemisphere uniquely determined by \mathbf{x} and its normal, f_r is the bidirectional reflectance

distribution function (BRDF), L_i is the incoming light from direction \mathbf{w}' , and $\mathbf{w}' \cdot \mathbf{n}$ accounts for the attenuation of the incoming light.










Advantages: In path tracing, the rendering equation is often solved with Monte Carlo methods. Contrasting what happens in the real world, the paths of photons in a scene are traced backwards from the camera (screen pixels) to the source lights. Objects in the scene receive lighting contributions as they interact with the photon paths. By computing both the reflected and transmitted components of rays in a physically accurate way while conserving energies and obeying the refraction equations, PBR photorealistically renders shadows, reflections, and refractions, thereby capturing unprecedented levels of detail compared to other existing shading techniques. Note PBR describes a shading process and does not dictate how images are rasterized in screen space. In this work we use the *Mantra*[®] PBR engine to render synthetic image data with raytracing for its accurate calculation of lighting and shading as well as its physically intuitive parameter configuration.

Indoor scenes are typically closed rooms. Various reflective and diffusive surfaces may exist throughout the space. Therefore the effect of secondary rays is particularly important in achieving realistic lighting. PBR robustly samples both direct lighting contributions on surfaces from light sources and indirect lighting from rays reflected and diffused by other surfaces. The BSDF shader on a surface manages and modifies its color contribution when hit by a secondary ray. Doing so results in more secondary rays being sent out from the surface in evaluation. The reflection limit (the number of times a ray can be reflected) and the diffuse limit (the number of times diffuse rays bounce on surfaces) need to be chosen wisely to balance the final image quality and the rendering time. Decreasing indirect lighting samples will likely yield a nice rendering time reduction, but at the cost of significantly diminished visual realism.

Rendering Time vs Rendering Quality: In summary, we use the following control parameters to adjust the render quality and speed:

- *Baseline pixel samples.* This is the minimum number of rays sent per pixel. Each pixel

Table 6.1: Comparisons of rendering time vs quality. The first column tabulates the reference number and rendering results used in this work, the second column lists all the criteria, and the remaining columns present comparative results. The color differences between the reference image and images rendered with various parameters are measured by LAB Delta E standard [SB02] tracing back to Helmholtz and Hering [BKW98, Val07].

Ref.	Criteria	Comparisons							
3×3	Baseline pixel samples	2×2	1×1	3×3	3×3	3×3	3×3	3×3	3×3
0.001	Noise level	0.001	0.001	0.01	0.1	0.001	0.001	0.001	0.001
22	Maximum additional rays	22	22	22	22	10	3	22	22
6	Bounce limit	6	6	6	6	6	6	3	1
203	Time (second)	131	45	196	30	97	36	198	178
	LAB Delta E difference								

is typically divided evenly along both directions. Common values for this parameter are 3×3 and 5×5 . The higher pixel sample counts are usually required to produce motion blur and depth of field effects.

- *Noise level.* Different rays sent from each pixel will not yield identical paths. This parameter determines the maximum allowed variance among the different results. If necessary, additional rays (in addition to baseline pixel sample count) will be generated to decrease the noise.
- *Maximum additional rays.* This parameter is the upper limit of the additional rays sent for satisfying the noise level.
- *Bounce limit.* The maximum number of secondary ray bounces. We use this parameter to restrict both diffuse and reflected rays. Note that in PBR the diffuse ray is one of the most significant contributors to realistic global illumination, while the other parameters are more important in controlling the Monte Carlo sampling noise.

Table 6.1 summarizes our analysis of how these parameters affect the render time and image

Table 6.2: Performance of normal estimation for the NYU-Depth V2 dataset with different training protocols.

pre-train	fine-tune	mean↓	median↓	11.25° ↑	22.5° ↑	30° ↑
	NYUv2	27.30	21.12	27.21	52.61	64.72
	Eigen	22.2	15.3	38.6	64.0	73.9
[ZSY17]	NYUv2	21.74	14.75	39.37	66.25	76.06
ours+[ZSY17]	NYUv2	21.47	14.45	39.84	67.05	76.72

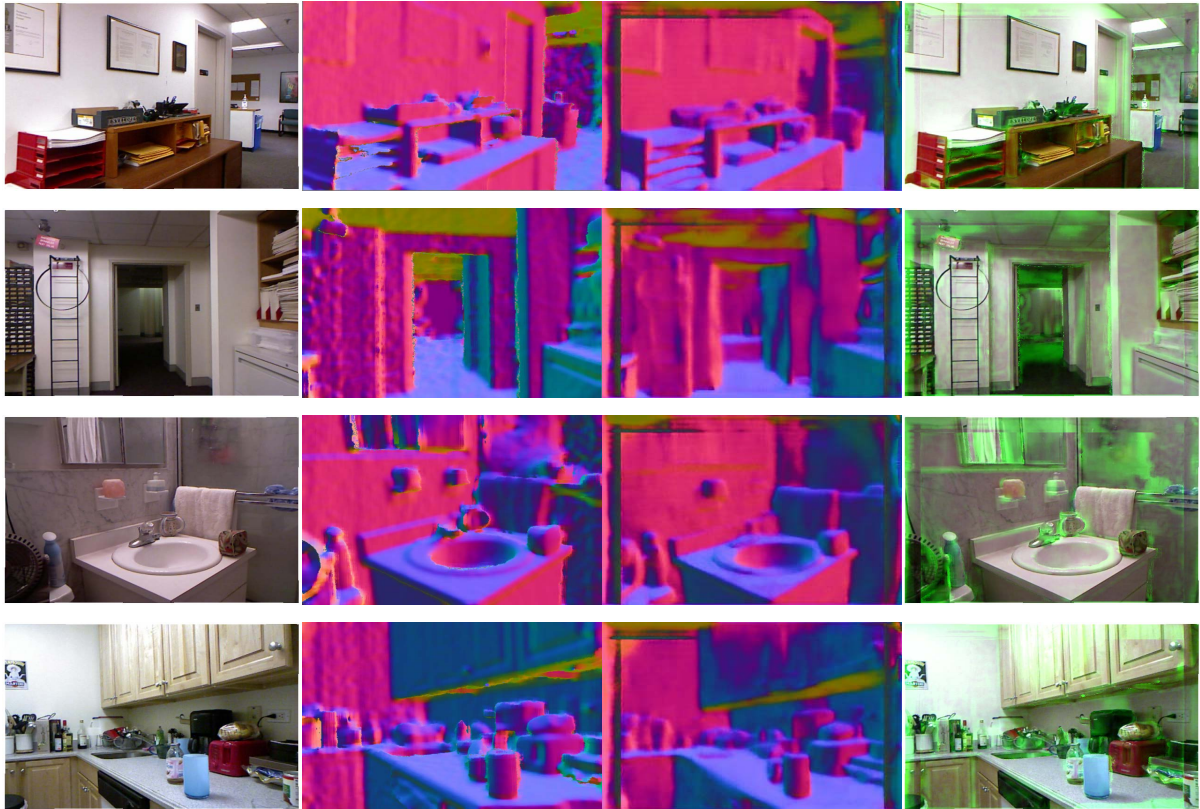
quality.

6.5 Experiments

In this section, we demonstrate the usefulness of the generated synthetic indoor scenes from two perspectives:

1. Improving state-of-the-art computer vision models by training with our synthetic data. We showcase our results on the task of normal prediction and depth prediction from a single RGB image, demonstrating the potential of using the proposed dataset.
2. Benchmarking common scene understanding tasks with configurable object attributes and various environments, which evaluates the stabilities and sensitivities of the algorithms, providing directions and guidelines for their further improvement in various vision tasks.

The reported results use the reference parameters indicated in Table 6.1. We found that choosing parameters for lower-quality rendering via the Mantra renderer does not provide training images that suffice to outperform state-of-the-art methods using the experimental setup described below.



(a) RGB (b) ground truth (c) estimation (d) error

Figure 6.8: Examples of normal estimation results predicted by the model trained with our synthetic data.

6.5.1 Normal Prediction

Predicting surface normals from a single RGB image is an essential task in scene understanding since it provides important information in recovering the 3D structure of the scenes. We train a neural network with our synthetic data to demonstrate that the perfect per-pixel ground truth generated using our pipeline could be utilized to improve upon the state-of-the-art performance on a specific scene understanding task. Using the fully convolutional network model described by Zhang *et al.* [ZSY17], we compare the normal estimation results given by models trained under two different protocols: (i) the network is directly trained and tested on the NYU-Depth V2 dataset, and (ii) the network is first pre-trained using our synthetic data, then fine-tuned and tested on NYU-Depth V2.

Table 6.3: Depth estimation performance on the NYU-Depth V2 dataset with different training protocols.

		Error					Accuracy		
Pre-Train	Fine-Tune	Abs Rel	Sqr Rel	Log10	RMSE(linear)	RMSE(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
NYUv2	-	0.233	0.158	0.098	0.831	0.117	0.605	0.879	0.965
Ours	-	0.241	0.173	0.108	0.842	0.125	0.612	0.882	0.966
Ours	NYUv2	0.226	0.152	0.090	0.820	0.108	0.616	0.887	0.972

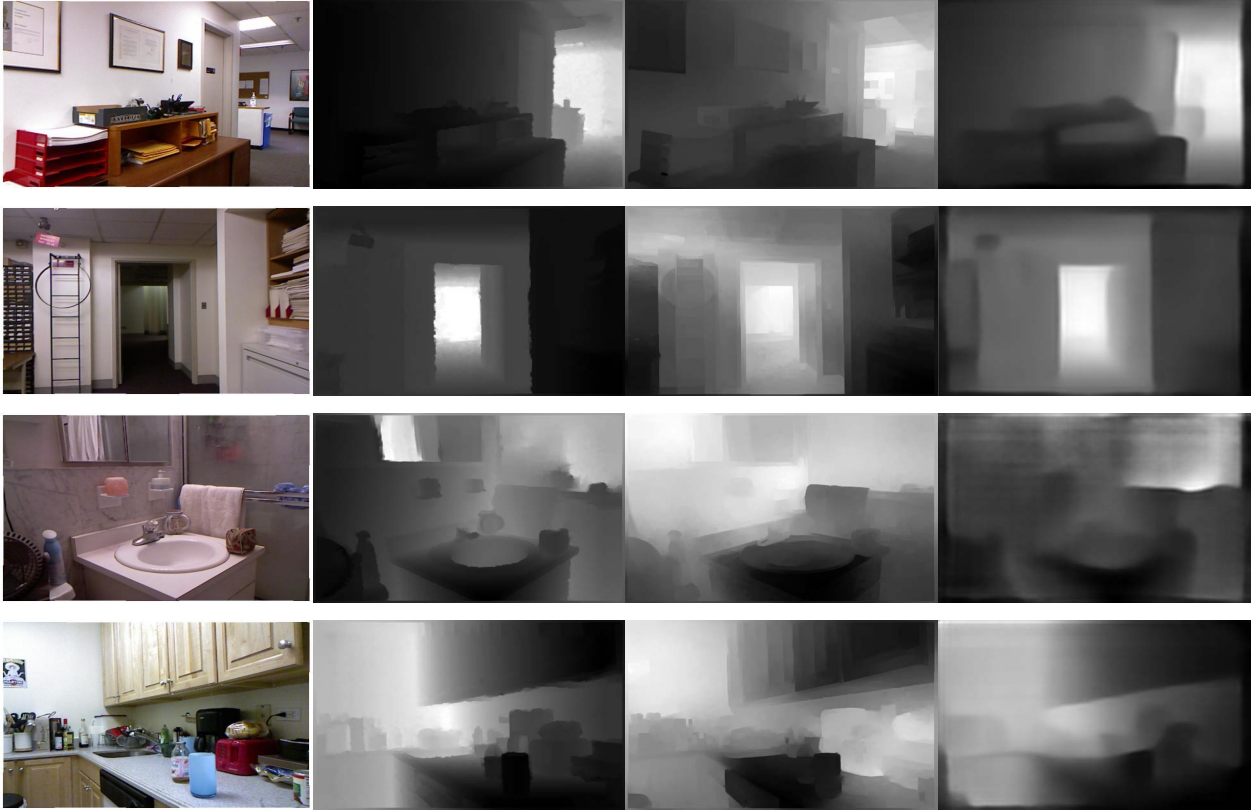
Following the standard evaluation protocol [FGH13, BRG16], we evaluate a per-pixel error over the entire dataset. To evaluate the prediction error, we computed the mean, median, and RMSE of angular error between the predicted normals and ground truth normals. Prediction accuracy is given by calculating the fraction of pixels that are correct within a threshold t , where $t = 11.25^\circ, 22.5^\circ, 30^\circ$. Our experimental results are summarized in Table 6.2. By utilizing our synthetic data, the model achieves better performance. From the visualized results in Figure 6.8, we can see that the error mainly accrues in the area where the ground truth normal map is noisy. We argue that part of the reason is due to the sensor’s noise or sensing distance limit. Such results in turn imply the importance to have perfect per-pixel ground truth for training and evaluation.

6.5.2 Depth Estimation

Single-image depth estimation is a fundamental problem in computer vision, which has found broad applications in scene understanding, 3D modeling, and robotics. The problem is challenging since no reliable depth cues are available. In this task, the algorithms output a depth image based on a single RGB input image.

To demonstrate the efficacy of our synthetic data, we compare the depth estimation results provided by models trained following protocols similar to those we used in normal prediction with the network in [LSL15]. To perform a quantitative evaluation, we used the metrics applied in previous work [EPF14]:

- Abs relative error: $\frac{1}{N} \sum_p \frac{|d_p - d_p^{gt}|}{d_p^{gt}}$,



(a) RGB (b) ground truth (c) NYUv2 (d) Ours+NYUv2

Figure 6.9: Examples of depth estimation results predicted by the model trained with our synthetic data.

- Square relative difference: $\frac{1}{N} \sum_p \frac{|d_p - d_p^{gt}|^2}{d_p^{gt}}$,
- Average \log_{10} error: $\frac{1}{N} \sum_x |\log_{10}(d_p) - \log_{10}(d_p^{gt})|$,
- RMSE : $\sqrt{\frac{1}{N} \sum_x |d_p - d_p^{gt}|^2}$,
- Log RMSE: $\sqrt{\frac{1}{N} \sum_x |\log(d_p) - \log(d_p^{gt})|^2}$,
- Threshold: % of d_p s.t. $\max(\frac{d_p}{d_p^{gt}}, \frac{d_p^{gt}}{d_p}) < \text{threshold}$,

where d_p and d_p^{gt} are the predicted depths and the ground truth depths at the pixel indexed by p , respectively, and N is the number of pixels in all the evaluated images. The first five metrics capture the error calculated over all the pixels; lower values are better. The threshold criteria capture the estimation accuracy; higher values are better.

Table 6.4: Depth estimation. Intensity, color, and material represent the scene with different illumination intensities, colors, and object material properties, respectively.

Setting	Method	Error					Accuracy		
		Abs Rel	Sqr Rel	Log10	RMSE(linear)	RMSE(log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Original	[LSL15]	0.225	0.146	0.089	0.585	0.117	0.642	0.914	0.987
	[EPF14]	0.373	0.358	0.147	0.802	0.191	0.367	0.745	0.924
	[EF15]	0.366	0.347	0.171	0.910	0.206	0.287	0.617	0.863
Intensity	[LSL15]	0.216	0.165	0.085	0.561	0.118	0.683	0.915	0.971
	[EPF14]	0.483	0.511	0.183	0.930	0.24	0.205	0.551	0.802
	[EF15]	0.457	0.469	0.201	1.01	0.217	0.284	0.607	0.851
Color	[LSL15]	0.332	0.304	0.113	0.643	0.166	0.582	0.852	0.928
	[EPF14]	0.509	0.540	0.190	0.923	0.239	0.263	0.592	0.851
	[EF15]	0.491	0.508	0.203	0.961	0.247	0.241	0.531	0.806
Material	[LSL15]	0.192	0.130	0.08	0.534	0.106	0.693	0.930	0.985
	[EPF14]	0.395	0.389	0.155	0.823	0.199	0.345	0.709	0.908
	[EF15]	0.393	0.395	0.169	0.882	0.209	0.291	0.631	0.889

Table 6.3 summarizes the results. We can see that the model pretrained on our dataset and fine-tuned on the NYU-Depth V2 dataset achieves the best performance, both in error and accuracy. Figure 6.9 shows qualitative results. This demonstrates the usefulness of our dataset in improving algorithm performance in scene understanding tasks.

6.5.3 Benchmark and Diagnosis

In this section, we show benchmark results and provide a diagnosis of various common computer vision tasks using our synthetic dataset.

Depth Estimation: In the presented benchmark, we evaluated three state-of-the-art single-image depth estimation algorithms due to Eigens *et al.* [EPF14, EF15] and Liu *et al.* [LSL15]. We evaluated those three algorithms with data generated from different settings including illumination intensities, colors, and object material properties. Table 6.4 shows

Table 6.5: Surface Normal Estimation. Intensity, color, and material represent the setting with different illumination intensities, illumination colors, and object material properties.

Setting	Method	Error			Accuracy		
		Mean	Median	RMSE	11.25°	22.5°	30°
Original	[EF15]	22.74	13.82	32.48	43.34	67.64	75.51
	[BRG16]	24.45	16.49	33.07	35.18	61.69	70.85
Intensity	[EF15]	24.15	14.92	33.53	39.23	66.04	73.86
	[BRG16]	24.20	16.70	32.29	32.00	62.56	72.22
Color	[EF15]	26.53	17.18	36.36	34.20	60.33	70.46
	[BRG16]	27.11	18.65	35.67	28.19	58.23	68.31
Material	[EF15]	22.86	15.33	32.62	36.99	65.21	73.31
	[BRG16]	24.15	16.76	32.24	33.52	62.50	72.17

a quantitative comparison. We see that both [EPF14] and [EF15] are very sensitive to illumination conditions, whereas [LSL15] is robust to illumination intensity, but sensitive to illumination color. All three algorithms are robust to different object materials. The reason may be that material changes do not alter the continuity of the surfaces. Note that [LSL15] exhibits nearly the same performance on both our dataset and the NYU-Depth V2 dataset, supporting the assertion that our synthetic scenes are suitable for algorithm evaluation and diagnosis.

Normal Estimation: Next, we evaluated two surface normal estimation algorithms due to Eigens *et al.* [EF15] and Bansal *et al.* [BRG16]. Table 6.5 summarizes our quantitative results. Compared with depth estimation, the surface normal estimation algorithms are stable to different illumination conditions as well as to different material properties. As in depth estimation, these two algorithms achieve comparable results on both our dataset and the NYU dataset.

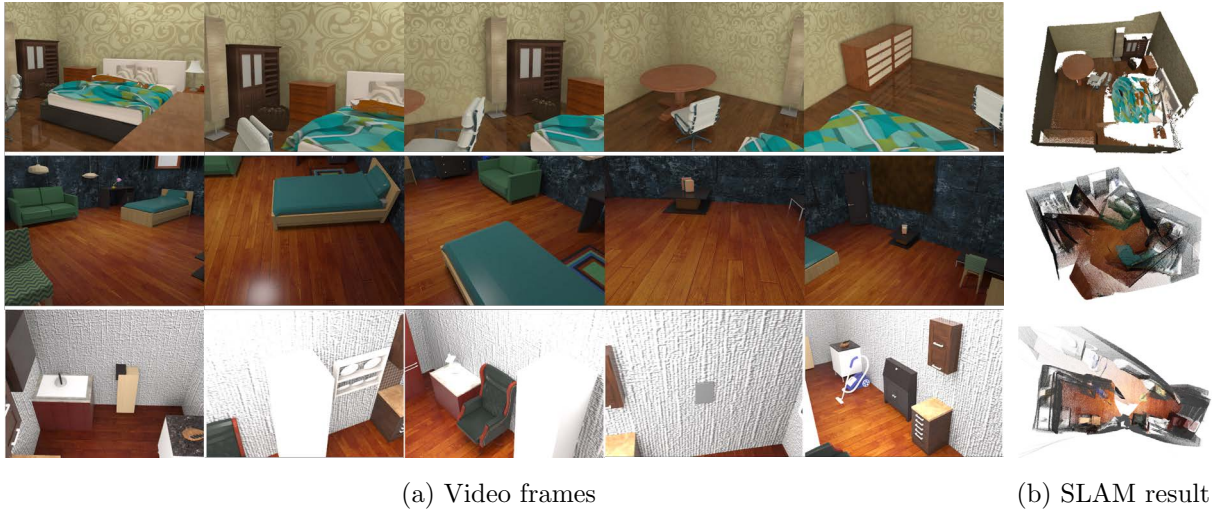


Figure 6.10: We can render the scenes as (a) a sequence of video frames after setting a camera trajectory, (b) which can be used to evaluate SLAM reconstruction [WLS15] results. The top row shows a successful reconstruction case, while the middle and bottom rows show two failure cases due to a fast moving camera and a plain, untextured surface, respectively.

Semantic Segmentation: Semantic segmentation has become one of the most popular tasks in scene understanding since the development and success of fully convolutional networks (FCNs). Given a single RGB image, the algorithm outputs a semantic label for every image pixel. We applied the semantic segmentation model described by Eigen *et al.* [EF15]. Since we have 129 classes of indoor objects whereas the model only has a maximum of 40 classes, we rearranged and reduced the number of classes to fit the prediction of the model. The algorithm achieves 60.5 pixel accuracy and 50.4 mIoU on our dataset.

3D Reconstructions and SLAM: We can evaluate 3D reconstruction and SLAM algorithms using images rendered from a sequence of different camera views. We generated different sets of images on diverse synthesized scenes with various camera motion paths and backgrounds to evaluate the effectiveness of the open-source SLAM algorithm ElasticFusion [WLS15]. A qualitative result is shown in Figure 6.10b. Some scenes can be robustly reconstructed when we rotate the camera evenly and smoothly, as well as when both the background and foreground objects have rich textures. However, other reconstructed 3D meshes are badly fragmented due to the failure to register the current frame with previous

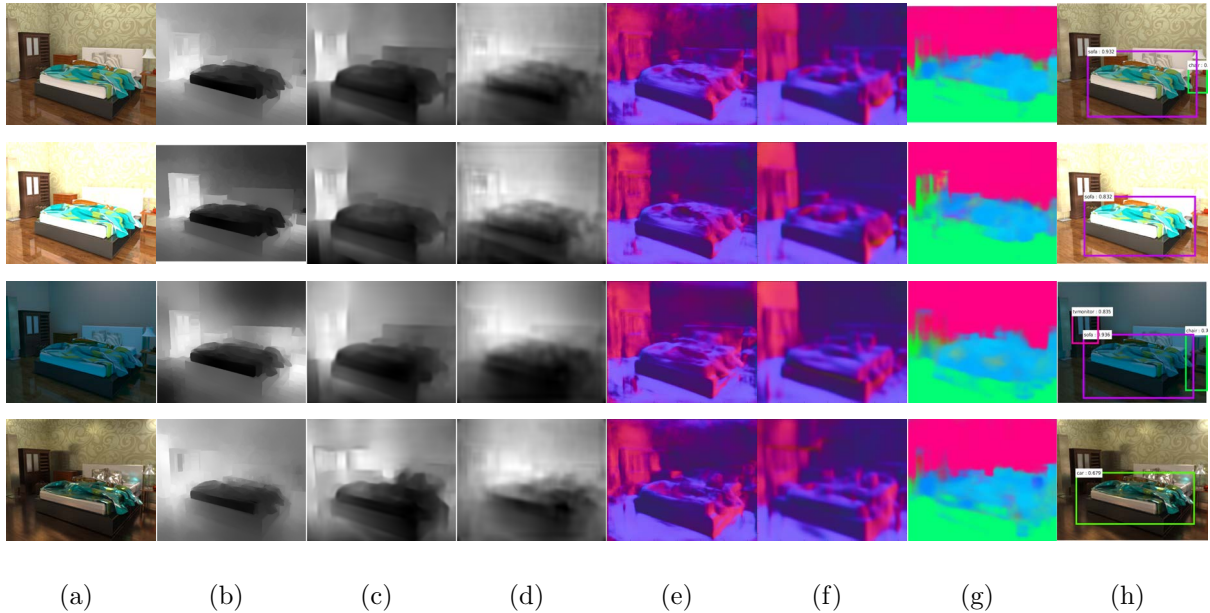


Figure 6.11: Benchmark results. (a) Given a set of generated RGB images rendered with different illuminations and object material properties (top to bottom: original settings, with high illumination, with blue illumination, and with metallic material properties), we evaluate (b)–(d) three depth prediction algorithms, (e)–(f) two surface normal estimation algorithms, (g) a semantic segmentation algorithm, and (h) an object detection algorithm.

frames due to fast moving cameras or the lack of rich textures. Experiments indicate that our synthetic scenes with configurable attributes and background can be utilized to diagnose the SLAM algorithm since we have full control of both the scenes and the camera trajectories.

Object Detection. The performance of object detection algorithms have greatly improved in recent years with the appearance and development of region-based convolutional neural networks. We apply the Faster R-CNN Model [RHG15] to detect objects. We again need to rearrange and reduce the number of classes for evaluation. Figure 6.11 summarizes our qualitative results with a bedroom scene. Note that a change of material can adversely affect the output of the model—when the material of objects is changed to metal, the bed is detected as a “car”.

6.6 Discussion

We now discuss in greater depth four specific topics related to the presented work.

Configurable scene synthesis: The most significant distinction between our work and prior work reported in the literature is our ability to generate large-scale *configurable* 3D scenes. But why is configurable generation desirable, given the fact that SUNCG [SYZ17b] already provided a large dataset of manually created 3D scenes?

A direct and obvious benefit is the potential to generate *unlimited* training data. As shown in a recent report by Sun *et al.* [SSS17], after introducing a dataset with 300 times of the size of ImageNet [DDS09], the performance of supervised learning appears to continue to increase linearly in proportion to the increased volume of labeled data. Such results indicate the usefulness of labeled datasets on a scale even larger than SUNCG. Although the SUNCG dataset is large by today’s standards, it is still a dataset limited by the manual specification of scene layouts.

A benefit of using configurable scene synthesis is to diagnose AI systems. Some preliminary results were reported in this work. In the future, we hope such methods can assist in building explainable AI. For instance, in the field of causal reasoning [Pea09], causal induction usually requires turning on and off specific conditions in order to draw a conclusion regarding whether or not a causal relation exists. Generating a scene in a controllable manner could provide a useful tool for studying these problems.

Furthermore, a configurable pipeline could be used to generate various virtual environment in a controllable manner in order to train virtual agents situated in virtual environments in order to learn task planning [LGS16, ZMK17] and control policy [HSL17, WMR17].

The importance of the different energy terms: In our experiments, the learned weights of the different energy terms indicate the importance of the terms. Based on the ranking from the largest weight to the smallest, the energy terms are 1) distances between furniture and the nearest wall, 2) relative orientations of furniture and the nearest wall, 3)

supporting relations, 4) functional group relations, and 5) occlusions of the accessible space of furniture by other furniture. We can regard such rankings learned from training data as human preferences of various factors in indoor layout designs, which is important for sampling and generating realistic scenes. For example, one can imagine that it is more important to have a desk aligned with a wall (relative distance and orientation), than it is to have a chair close to a desk (functional group relations).

Balancing rendering time and quality: The advantage of physically accurate representation of shadows, colors, and reflections comes at the cost of computation. High quality rendering (*e.g.*, rendering for movies) requires tremendous amounts of CPU time and computer memory that is practical only with distributed render farms. Low quality settings are prone to granular render noise due to stochastic sampling. Our comparisons between rendering time and rendering quality serve as a basic guideline for choosing the values of the rendering parameters. In practice, depending on the complexity of the scene (such as the number of light sources and reflective objects), manual adjustment is often needed in large-scale rendering (*e.g.*, an overview of a city) in order to achieve the best trade-off between rendering time and quality. Switching to GPU-based ray tracing engines is a promising alternative. This direction is especially useful for scenes with a small number of polygons and textures, which can fit into a modern GPU memory.

The speed of the sampling process: It takes roughly 3–5 minutes to render a 640×480 -pixel image, depending on settings related to illumination, environments, and the size of the scene. By comparison, the sampling process consumes approximately 3 minutes with the current setup. Although the convergence speed of the Monte Carlo Markov chain is fast enough relative to photorealistic rendering, it is still desirable to accelerate the sampling process. In practice, to speed up the sampling and improve the synthesis quality, we split the sampling process into five stages: (i) Sample the objects on the wall, *e.g.*, windows, switches, paints and lights, (ii) sample the core functional objects in *functional groups* (*e.g.*, desks and beds), (iii) sample the objects that are associated with the core functional objects

(*e.g.*, chairs and nightstands), (iv) sample the objects that are not paired with other objects (*e.g.*, wardrobes and bookshelves), and (v) Sample small objects that are supported by furniture (*e.g.*, laptops and books). By splitting the sampling process using functional groups, we effectively reduce the computational complexity, and different types of objects quickly converge to their final positions.

6.7 Appendix: Additional Results



Figure 6.12: Additional results of generated scenes.

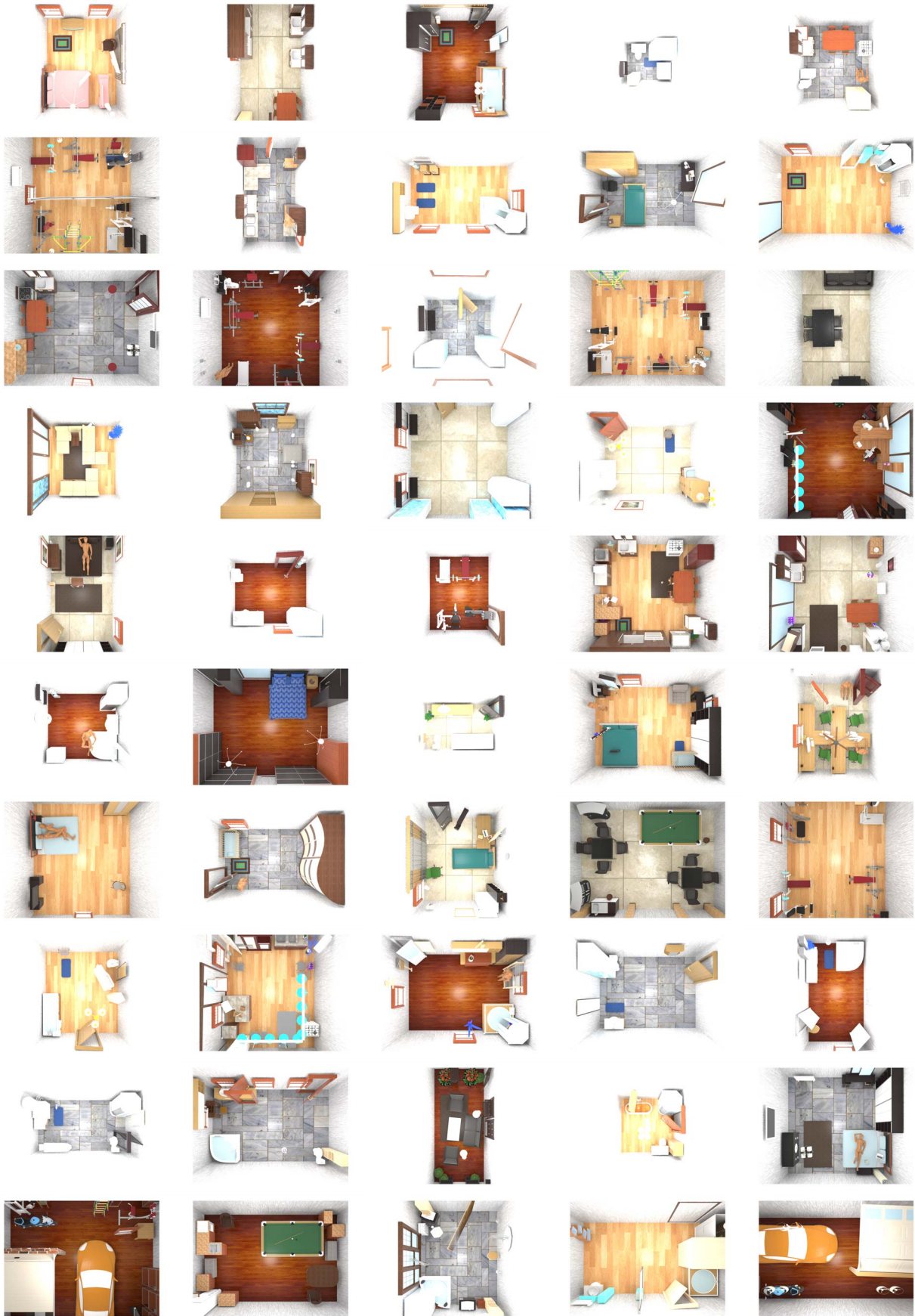


Figure 6.13: Additional results of generated scenes.

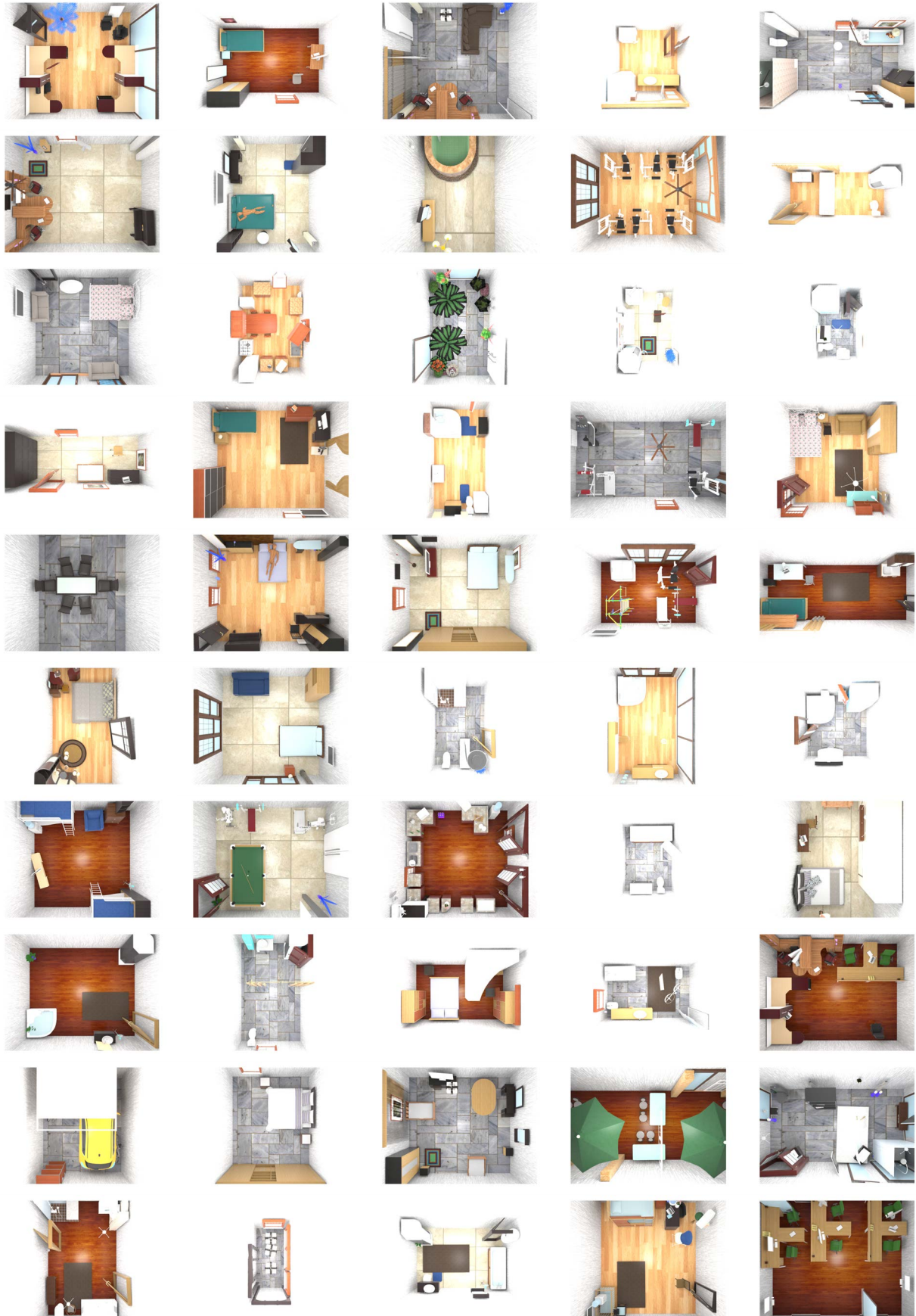


Figure 6.14: Additional results of generated scenes.

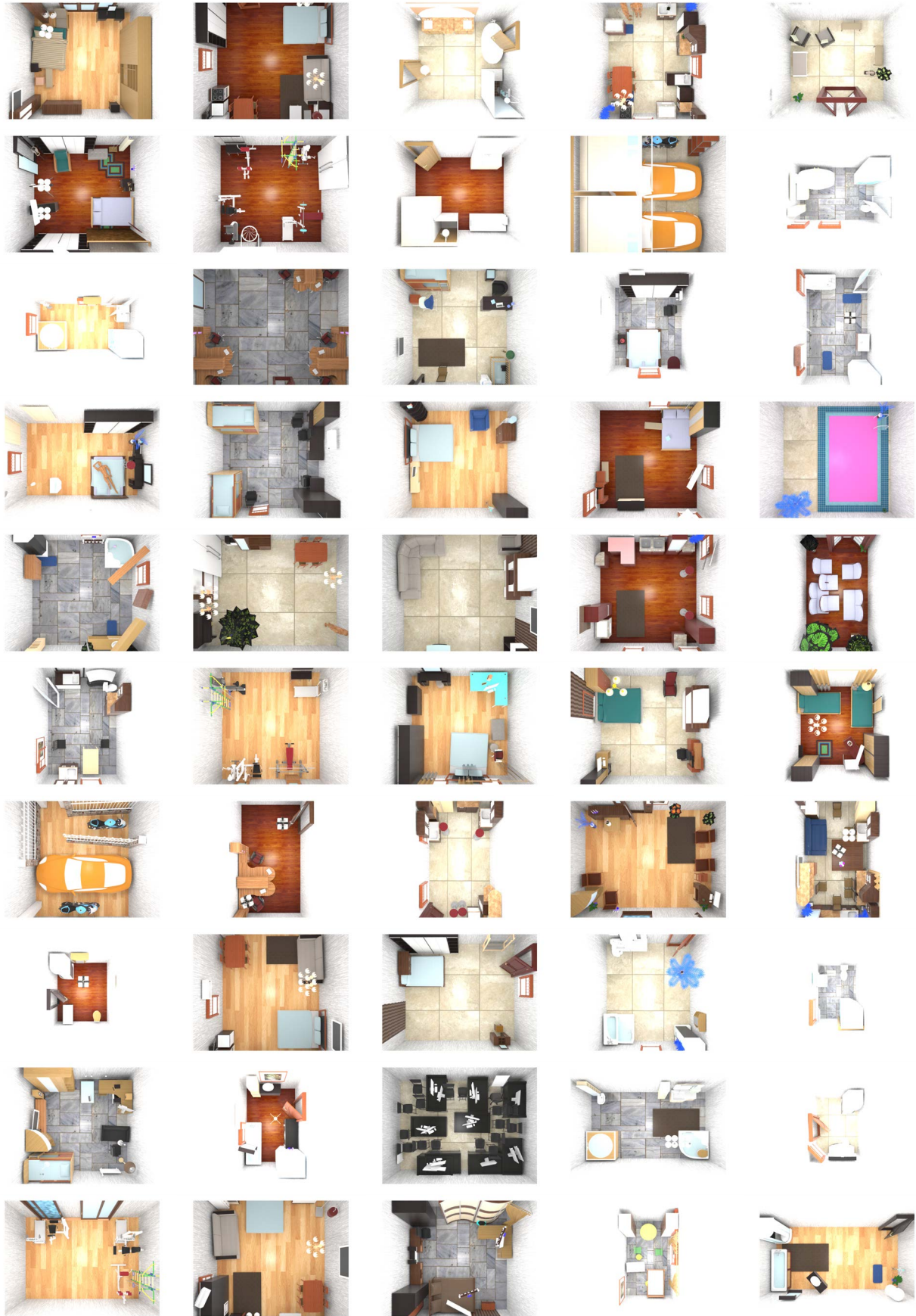


Figure 6.15: Additional results of generated scenes.

Part II

Integrating Forces and Functionality in Object Manipulations

We present 4 case studies on integrating forces and functionality in object manipulations in the field of robotics.

In section 7.1, we first present a design of an easy-to-replicate glove-based system that can reliably perform simultaneous hand pose and force sensing in real time, for the purpose of collecting human hand data during fine manipulative actions. The design consists of a sensory glove that is capable of jointly collecting data of finger poses, hand poses, as well as forces on palm and each phalanx. Specifically, the sensory glove employs a network of 15 IMUs to measure the rotations between individual phalanxes. Hand pose is then reconstructed using forward kinematics. Contact forces on the palm and each phalanx are measured by 6 customized force sensors made from Velostat, a piezoresistive material whose force-voltage relation is investigated. We further develop an open-source software pipeline consisting of drivers and processing code and a system for visualizing hand actions that is compatible with the popular Raspberry Pi architecture. In our experiment, we conduct a series of evaluations that quantitatively characterize both individual sensors and the overall system, proving the effectiveness of the proposed design.

In section 7.2, we propose an unsupervised learning approach for manipulation event segmentation and manipulation event parsing based on the data acquired by the tactile glove mentioned in section 7.1. The proposed framework incorporates hand pose kinematics and contact forces using a low-cost easy-to-replicate tactile glove. We use a temporal grammar model to capture the hierarchical structure of events, integrating extracted force vectors from the raw sensory input of poses and forces. The temporal grammar is represented as a temporal And-Or graph (T-AOG), which can be induced in an unsupervised manner. We obtain the event labeling sequences by measuring the similarity between segments using the Dynamic Time Alignment Kernel (DTAK). Experimental results show that our method achieves high accuracy in manipulation event segmentation, recognition and parsing by utilizing both pose and force data.

In section 8.1, we learn a manipulation model to execute tasks with multiple stages and variable structure, which typically are not suitable for most robot manipulation approaches. The model is learned from human demonstration using a tactile glove that mea-

sures both hand pose and contact forces. The tactile glove enables observation of visually latent changes in the scene, specifically the forces imposed to unlock the child-safety mechanisms of medicine bottles. From these observations, we learn an action planner through both a top-down stochastic grammar model (And-Or graph) to represent the compositional nature of the task sequence and a bottom-up discriminative model from the observed poses and forces. These two terms are combined during planning to select the next optimal action. We present a method for transferring this human-specific knowledge onto a robot platform and demonstrate that the robot can perform successful manipulations of unseen objects with similar task structure.

In section 8.2, we present a novel Augmented Reality (AR) approach, through Microsoft HoloLens, to address the challenging problems of diagnosing, teaching, and patching interpretable knowledge of a robot. A Temporal And-Or graph (T-AOG) of opening bottles is learned from human demonstration and programmed to the robot. This representation yields a hierarchical structure that captures the compositional nature of the given task, which is highly interpretable for the users. By visualizing the knowledge structure represented by a T-AOG and the decision making process by parsing the T-AOG, the user can intuitively understand what the robot knows, supervise the robot’s action planner, and monitor visually latent robot states (*e.g.*, the force exerted during interactions). Given a new task, through such comprehensive visualizations of robot’s inner functioning, users can quickly identify the reasons of failures, interactively teach the robot with a new action, and patch it to the current knowledge structure. In this way, the robot is capable of solving similar but new tasks only through minor modifications provided by the users interactively. This process demonstrates the interpretability of our knowledge representation and the effectiveness of the AR interface.

CHAPTER 7

Building Tactile Glove

7.1 A Glove-based System for Studying Hand-Object Manipulation via Joint Pose and Force Sensing

Robots that imitate the behaviors of humans may enable more natural and friendly interactions with humans in man-made environments, as with robotic handshaking [THG16]. Just as whole body sensing [KB13] is critical for the study of human movement, hand pose and force information is crucial to the investigation of manipulative tasks. While researchers can track hand pose based on perception [RA15], force estimation from vision using numerical differentiation methods [ZZC15, PKQ15], or sophisticated physics-based soft-body

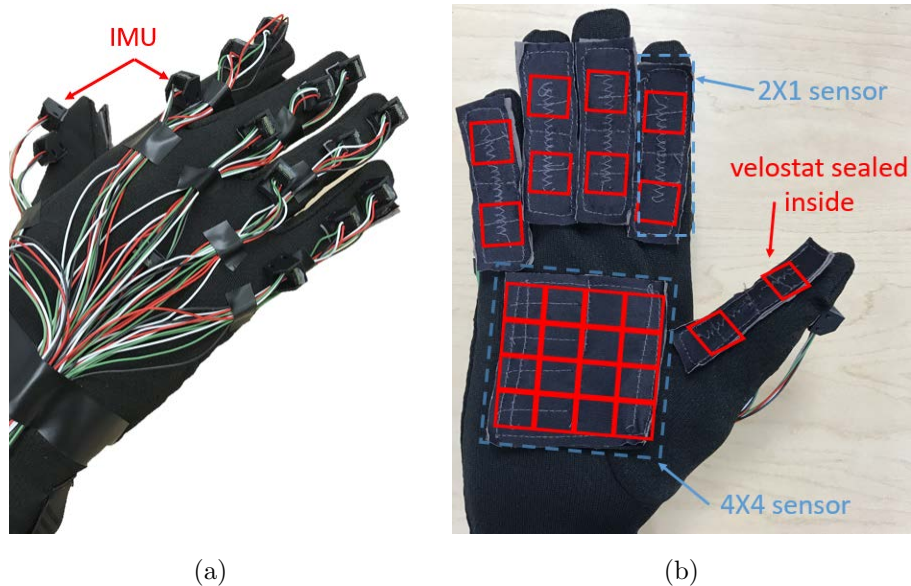


Figure 7.1: Prototype consisting of (a) 15 IMUs on the dorsum of the hand and (b) 6 integrated Velostat force sensor with 26 taxels on the palmar aspects of the hand.

simulation [WMZ13, ZZM13], glove-based devices still have their own advantages, presenting convenient, integrated solutions that can be natural and essential for collecting ground truth hand data during manipulations and interactions.

Designs of tactile gloves have long been proposed for a wide range of applications, and they remain an active research area. Dipietro *et al.* provided a comprehensive survey of glove-based system designs and their application from 1970s to 2008 [DSD08]. Since then, a number of novel designs have emerged to address existing limitations, including portability, reliability, and cost. As the main motivations of developing data/tactile gloves or other glove-based systems are obtaining the pose and force information during manipulative actions, we divide some notable recent designs since 2008 into two categories based on the types of data they can collect: gloves with i) only pose-sensing, and ii) joint pose- and force-sensing.

7.2 Unsupervised Learning of Hierarchical Models for Hand-Object Interactions

Consider a complex manipulation event of a person opening a medicine bottle with safety lock (Figure 7.2). During this process, a number of movement primitives were performed: *grasp*, *push-and-twist*, *push-and-twist*, *twist*, and finally *pull* the lid off the bottle. Even with the most state-of-the-art action understanding and recognition algorithms (see survey [Pop10, WRB11]), it is still challenging to segment such action sequence and parse the manipulation event. This is due to three major difficulties: i) severe occlusions happen during fine manipulation, especially self-occlusions, ii) in subtle manipulation tasks, visual data may not be able to reveal adequate knowledge to capture the quintessence. Certain actions are hard to detect using skeleton data alone but need additional force readings *e.g.*, whether an action of pushing was performed during twisting the lid, and iii) ground truth data is difficult to obtain using vision sensor alone, oftentimes impossible to obtain the needed information (*e.g.*, the force readings, and accurate finger poses during occlusions).

In this work, we present an unsupervised learning method for manipulation event segmentation, recognition and parsing. The method not only accounts for the aforementioned chal-

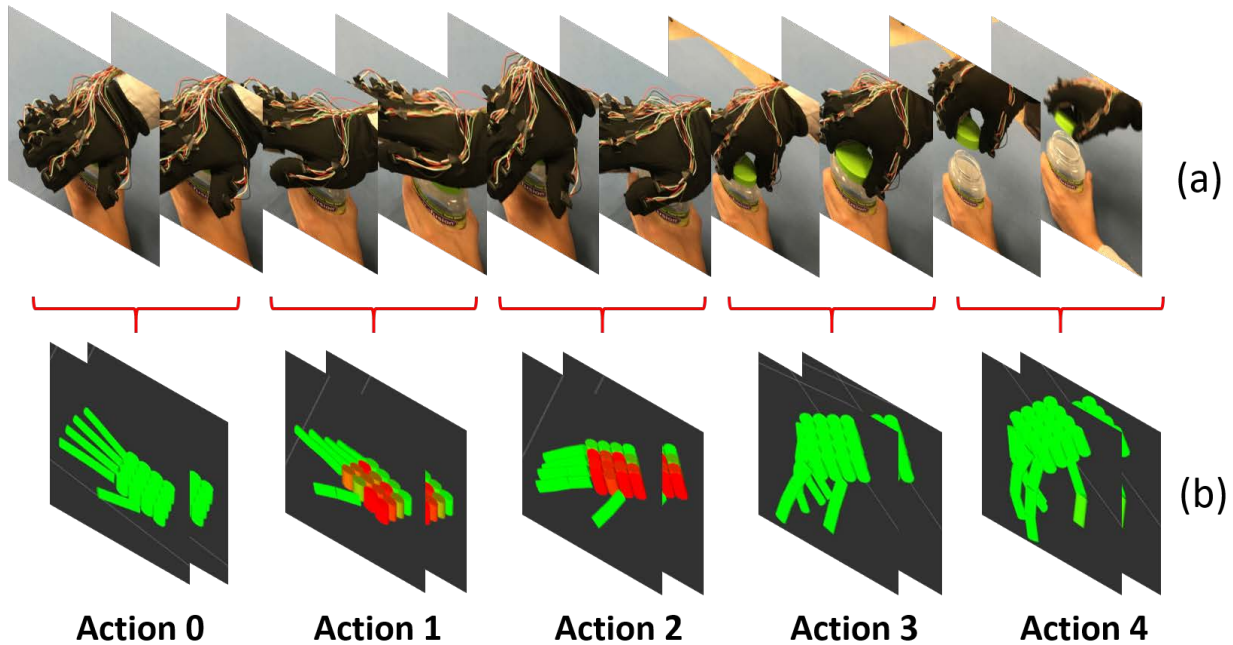


Figure 7.2: (a) A sequence of movement primitive demonstrated by an agent for a manipulation task—opening a medicine bottle captured by a tactile glove. (b) Reconstructed force and pose data using the tactile glove. Our proposed method segments and parses the noisy inputs of force and pose in an unsupervised fashion.

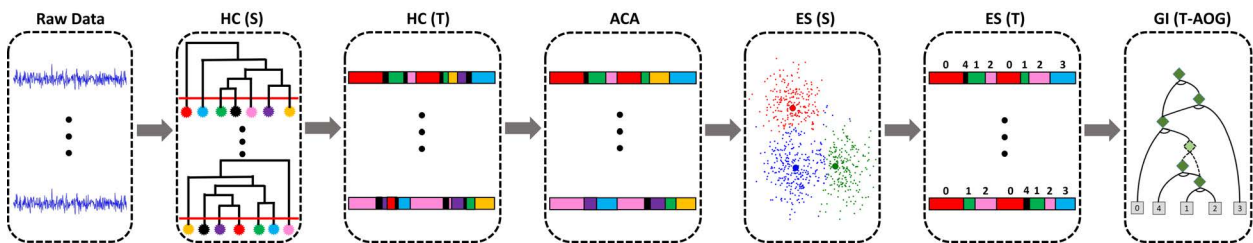


Figure 7.3: Unsupervised learning pipeline of hand-object motion recognition. After collecting the raw data using a tactile glove, a spatial (HC (S)) and temporal (HC (T)) hierarchical clustering is performed on both force and pose data. An aligned cluster analysis (ACA) is adopted to further reduce the noise. Event segmentation (ES (S) and ES (T)) is achieved by merging motion primitives based on the distance measured by DTAK. Finally, a grammar is induced (GI) based on the segmented events, forming a T-AOG.

lenges, but also captures the temporal hierarchical structure of the manipulation sequence using a grammar model—a temporal And-Or graph (T-AOG). Specifically, we investigate the manipulation actions of opening different types of medicine bottles. *Bottle 1* has no safety lock and can be opened by simply twisting the lid. *Bottle 2* requires pressing the lid while twisting. Pinching the safety lock is needed to open *Bottle 3*. Importantly, some actions (*e.g.*, pressing, pinching) are difficult to observe visually, thus require additional sensing for action recognition.

To obtain the force readings during manipulations, we propose to study hand-object interactions with additional force information through a low-cost, easy-to-replicate tactile glove [LXM17]. By observing the data collected using the tactile glove, such as the force exerted on the palm, we can learn that a push-down action is performed as well as a set of motion primitives that can best describe the action sequences. Thus, our system is able to “see”, in numerical terms, the forces during hand-object interactions. We argue that this is an important step in recognizing manipulation actions with visually latent force information.

Still, it is nearly impossible to understand and transfer the raw data (poses and forces) retrieved from the tactile glove *directly* to a robot due to different embodiments. Therefore, we need to reconstruct the semantic meanings of manipulation events from the human demonstration, allowing the transfer of abstract knowledge to a robot.

To recover the semantic meaning and model the temporal structure of actions in a hand-object interaction, we represent the manipulation sequence using a T-AOG, a temporal grammar model that captures the hierarchical structure of the action sequences. Its terminal nodes are motion primitives, *e.g.*, twisting and pressing, which is learned by unsupervised clustering over extracted features of the pose and force sensory inputs. To evaluate the effectiveness of our model, we compare the segmentation and labeling results of different sensory data with several baseline methods.

This work makes three contributions:

1. We incorporate *invisible* force in addition to the conventional pose-based methods for event segmentation and parsing during fine-grained manipulation tasks. We show in

the experiment that a better performance of motion recognition is achieved by jointly considering hand pose and force data.

2. We propose an unsupervised learning framework to learn a temporal grammar model (T-AOG) for hand-object interactions. The framework incorporates automatic clustering, segmentation, labeling, and high-level grammar induction. The grammar structure is shown to significantly improve the action recognition results compared to using clustering method alone.
3. We introduce a general method for modeling noisy and heterogeneous sensory data of hand-object manipulation.

CHAPTER 8

Learning Object Manipulations by Integrating Forces and Functionality

8.1 Feeling the Force: Integrating Force and Pose for Fluent Discovery through Imitation Learning

Consider the task of opening medicine bottles that have child-safety locking mechanisms (Figure 8.1a). These bottles require the user to push or squeeze in various places to unlock the cap. By design, attempts to open these bottles using a standard procedure will result in failure. Even if the agent visually observes a successful demonstration, imitation of this procedure will likely omit critical steps in the procedure. The visual procedure for opening both medicine and traditional bottles are typically identical. The agent lacks understanding of the tactile interaction required to unlock the safety mechanism of the bottle. Only direct observation of forces or instruction can elucidate the correct procedure (Figure 8.1e). Even with knowledge of the correct procedure, opening medicine bottles poses several manipulation challenges that involve feeling and reacting to the internal mechanisms of the bottle cap. Although the presented study takes opening medicine bottles as an example, many other tasks share similar properties and require non-trivial reasoning such as opening locked doors [SGG08].

In this work, we learn a manipulation model from human demonstration that captures observed motion and kinematics, as well as visually latent changes such as forces and internal state (Figure 8.1e). We learn this manipulation model for objects that have similar functional properties, but exhibit different geometries and internal configurations that affect how the

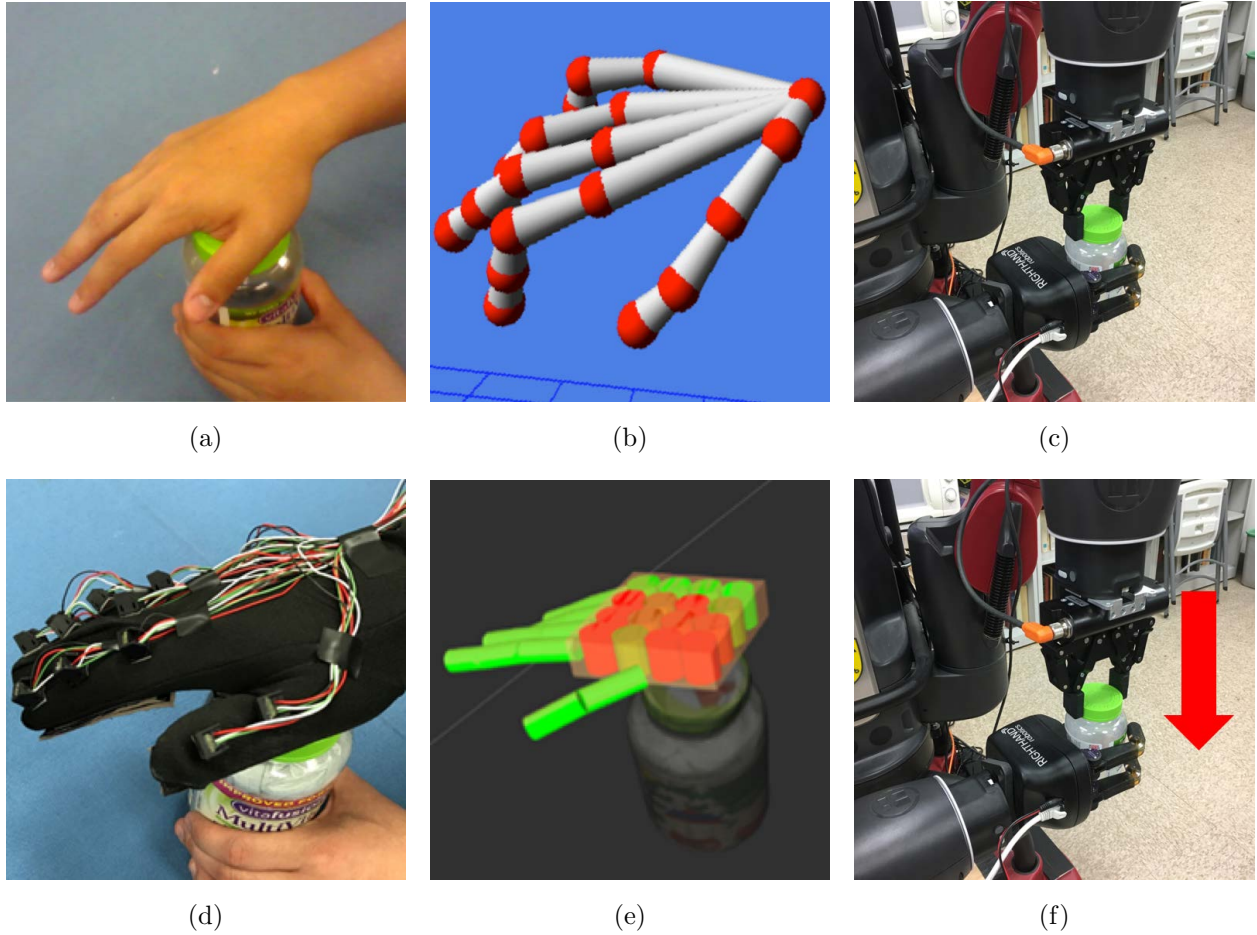


Figure 8.1: Given a RGB-D-based image sequence (a), although we can infer the skeleton of hand using vision-based methods (b), such knowledge cannot be easily transferred to a robot to open a medicine bottle (c), due to the lack of force sensing during human demonstrations. In this work, we utilize a tactile glove (d) and reconstruct both forces and poses from human demonstrations (e), enabling robot to directly observe forces used in demonstrations so that the robot can successfully open a medicine bottle (f).

object must be manipulated.

Two key problems are discussed in this work:

1. how to naturally recover the visually latent force data from the human demonstrations;
2. how to represent such knowledge and successfully transfer it to a robot?

For the first problem, although some initial results have been reported to reconstruct

poses and/or forces exerted by the demonstrator using vision-based methods [ZZM13, WMZ13, ZZC15, PKQ15, ZJZ16], these methods still have difficulty providing pose and force data precise enough for robot learning. Instead, we utilize an open-source tactile glove [LXM17] designed to measure both hand pose and contact forces across the surface of the hand. These demonstrations are performed naturally, and within a motion capture setup to obtain ground-truth tracking of the objects and human wrist.

For the second problem, our system takes into consideration: i) an And-Or-Graph (AOG) [ZM07] learned from human demonstrations as top-down knowledge for manipulations of an unseen medicine bottle, in which the AOG model uses *fluents* [NC36] to model the changes between pre- and post-conditions of demonstrations in a low-dimensional subspace; and ii) A bottom-up process learned from raw signal data when robot executes to encode transition between pre- and post-conditions. Together, these two processes learn a manipulation model to open medicine bottles.

This work makes four contributions:

1. Using a tactile glove during demonstrations that enable the robot to utilize both the poses and forces exerted by the demonstrator. In contrast with previous work, our method focuses on integrating visual measurements with physical measurements not observable from vision (*e.g.*, forces), capturing latent relationships that are imperceptible from vision alone.
2. Learning a stochastic grammar model that represents the compositional task hierarchy comprising of atomic actions for manipulation tasks, compactly capturing the admissible sequence of actions for all the bottles demonstrated.
3. Learning a bottom-up process that encodes raw haptic signals to account for the transition from a previous state to a new state. Together with the stochastic grammar model as a top-down process, these two processes jointly form the manipulation model.
4. Transferring the learned model from human demonstrations onto a Baxter robot by solving a correspondence problem [DH02]. This embodiment mapping function directly

relates hand pose and contact force from the human to the force-torque sensing and gripper state of the robot; enabling the robot to reason about its haptic measurements using the relations learned from human demonstration.

8.2 Interactive Robot Knowledge Patching using Augmented Reality

The ever-growing vast amount of data and rapid-increasing computing power have enabled a data-driven machine learning paradigm in the past decade. Using Deep Neural Networks (DNNs) [HOT06], the performance of machine learning methods has reached a remarkable level in some specific tasks, even arguably better than human, *e.g.*, control [DCH16, MKS15], grasping [MLN17, LLS15], object recognition [HZR15, IS15], learning from demonstration [ACV09], and playing the game of go [SHM16] and poker [MSB17, BS17]. However, despite these recent encouraging progress, DNN-based methods have well-known limitations; one of these limitations is the lack of interpretability of the knowledge representation, especially about how and why a decision is made, which plays a vital role in the scenarios where robots work alongside humans.

Meanwhile, contextual adaptation models using And-Or-Graph (AOG) and Probabilistic Programming start to demonstrate the interpretability using small amount of training data in robot learning [XSX16, ZZC15], recognition [ZM07, GLK17], reconstruction [LZZ17], social interactions [SGR17], causal reasoning [Pea09, FZ13b], playing video games [KSM17], and human-level concept learning [LST15]. Although these types of models have been identified by DARPA as the representative models in the third wave of artificial intelligence [Lau17], a natural and convenient way to teach and interact with a robot to acquire and accumulate such interpretable knowledge is still missing.

In this work, we propose an augmented reality (AR) interface, through Microsoft HoloLens, to interact with a Rethink Baxter robot for teaching and patching its interpretable knowledge represented by the AOGs. In the experiments, we demonstrate the proposed AR interface develops interpretations at three different levels:

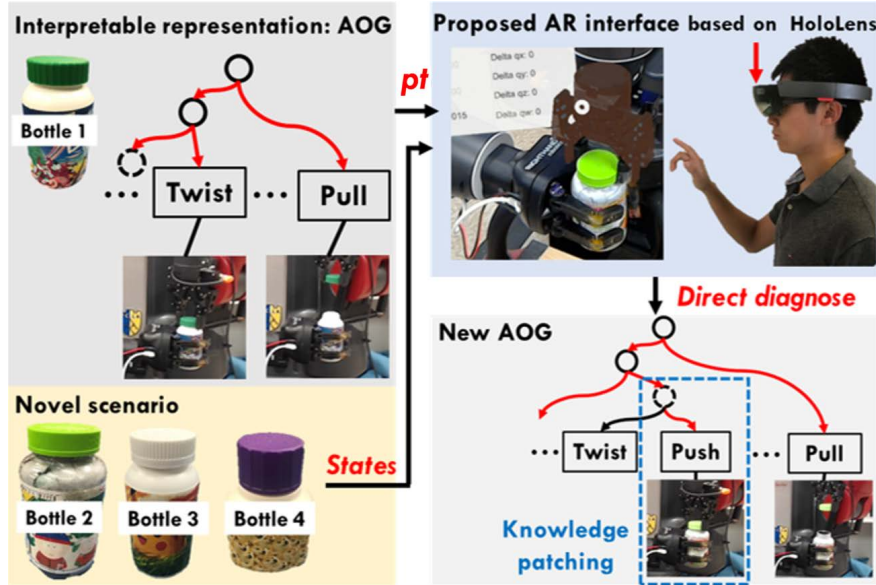


Figure 8.2: System architecture. Given a knowledge represented by a T-AOG of opening conventional bottles, the robot tries to open unseen medicine bottles with safety lock. The proposed AR interface can visualize the inner functioning of the robot during action executions. Thus, the user can understand the knowledge structure inside the robot’s mind, directly oversee the entire decision making process through HoloLens in real-time, and finally interactively correct the missing action (push) to open a medicine bottle successfully.

1. **Knowledge structure by compositional models.** We take an example of a robot opening various medicine bottles, and represent the robot’s knowledge structure using a Temporal And-Or Graph (T-AOG) [ZM07]. The T-AOG encodes a repertoire of a successful action sequence for a robot to open medicine bottles. Visualizing through the holographical interface, the state of robot represented by a T-AOG can be naturally inquired through gesture control.
2. **Interpretable decision making.** Unlike a teacher can usually query students to verify whether they obtain the knowledge structure correctly, it is nontrivial for users to check and understand robots’ inner functioning, making it difficult for users to diagnose the robot decision-making process. By visualizing the decision-making process on top of T-AOG through the holographical interface, information of interests can be better

associate to the actual robot and the actual scene, thus help to gain insight about how a robot behaves and why it behaves in a certain way.

3. **Interactive knowledge structure patching.** Once the users find out the reason why a certain action sequence leads to a failure, the users can interactively patch the knowledge structure represented by the T-AOG: adding a missing node, deleting a redundant node, and updating a node representing a wrong action.

This work makes the following three contributions:

1. We introduce a new AR interface based on the state-of-the-art head-mounted display, Microsoft HoloLens, providing users a much more natural way to interact with a robot. In addition to visualizing robot's states, intentions, or controlling robots, we further visualize robot's knowledge representation so that users can understand why and how a robot will behave.
2. In contrast to using additional force sensing [LXM17] to perceive the visually hidden force [EGX17], the present study provides an intuitive way for users to augment the visually imperceptible knowledge on top of the learned action sequence represented by a T-AOG. In this way, the AR interface affords a much more effectively diagnose and knowledge patching process. Furthermore, it often has a much lower cost, as users do not need to build any additional sensors or apparatus to demonstrate the tasks.
3. We build a communication interface between the HoloLens platform and ROS, and are publicly available online. It allows a variety of interchangeable messages, which we hope would ease the development difficulties across commonly used platforms.

Part III

Cognitive Studies

CHAPTER 9

Intuitive Physics

In this chapter, we present five case studies on intuitive physics.

In section 9.1, we first study the human cognition of containing relation. Containers are ubiquitous in daily life. By container, we consider any physical object that can contain other objects, such as bowls, bottles, baskets, trash cans, refrigerators, *etc.*. In this work, we are interested in following questions: What is a container? Will an object contain another object? How many objects will a container hold? We study those problems by evaluating human cognition of containers and containing relations with physical simulation. In the experiments, we analyze human judgments with respect to results of physical simulation under different scenarios. We conclude that the physical simulation is a good approximation to the human cognition of container and containing relations.

In section 9.2, we study the perception of fluid viscosity. The physical behavior of moving fluids is highly complex, yet people are able to interact with them in their everyday lives with relative ease. To investigate how humans achieve this remarkable ability, the present study extended the classical water-pouring problem [SB99] to examine how humans take into consideration physical properties of fluids (*e.g.*, viscosity) and perceptual variables (*e.g.*, volume) in a reasoning task. We found that humans do not rely on simple qualitative heuristics to reason about fluid dynamics. Instead, they rely on the perceived viscosity and fluid volume to make quantitative judgments. Computational results from a probabilistic simulation model can account for human sensitivity to hidden attributes, such as viscosity, and their performance on the water-pouring task. In contrast, non-simulation models based on statistical learning fail to fit human performance. The results in the present work provide converging evidence supporting mental simulation in physical reasoning, in addition

to developing a set of experimental conditions that rectify the dissociation between explicit prediction and tacit judgment through the use of mental simulation strategies.

In section 9.3, we compare the human perception of three types of substances: liquid, sand and rigid balls. A growing body of evidence supports the hypothesis that humans infer future states of perceived physical situations by propagating noisy representations forward in time using rational (approximate) physics. In the present study, we examine whether humans are able to predict (1) the resting geometry of sand pouring from a funnel and (2) the dynamics of three substances—liquid, sand, and rigid balls—flowing past obstacles into two basins. Participants’ judgments in each experiment are consistent with simulation results from the intuitive substance engine (ISE) model, which employs a Material Point Method (MPM) simulator with noisy inputs. The ISE outperforms ground-truth physical models in each situation, as well as two data-driven models. The results reported herein expand on previous work proposing human use of mental simulation in physical reasoning and demonstrate human proficiency in predicting the dynamics of sand, a substance that is less common in daily life than liquid or rigid objects.

In section 9.4, we examine how humans adapt to novel physical situations with unknown gravitational acceleration in immersive virtual environments. We designed four virtual reality experiments with different tasks for participants to complete: strike a ball to hit a target, trigger a ball to hit a target, predict the landing location of a projectile, and estimate the flight duration of a projectile. The first two experiments compared human behavior in the virtual environment with real-world performance reported in the literature. The last two experiments aimed to test the human ability to adapt to novel gravity fields by measuring their performance in trajectory prediction and time estimation tasks. The experiment results show that: 1) based on brief observation of a projectile’s initial trajectory, humans are accurate at predicting the landing location even under novel gravity fields, and 2) humans’ time estimation in a familiar earth environment fluctuates around the ground truth flight duration, although the time estimation in unknown gravity fields indicates a bias toward earth’s gravity.

In section 9.5, we study visuomotor adaptation. Visuomotor adaptation plays an impor-

tant role in motor planning and execution. However, it remains unclear how sensorimotor transformations are recalibrated when visual and proprioceptive feedback are decoupled. To address this question, the present study asked participants to reach toward targets in a virtual reality (VR) environment. They were given visual feedback of their arm movements in VR that was either consistent (normal motion) with the virtual world or reflected (reversed motion) with respect to the left-right and vertical axes. Participants completed two normal motion experimental sessions, with a reversed motion session in between. While reaction time in the reversed motion session was longer than in the normal motion session, participants showed the learning improvement by completing trials in the second normal motion session faster than in the first. The reduction in reaction time was found to correlate with greater use of linear reaching trajectory strategies (measured using dynamic time warping) in the reversed and second normal motion sessions. This result appears consistent with linear motor movement planning guided by increased attention to visual feedback. Such strategical bias persisted into the second normal motion session. Participants in the reversed session were grouped into two clusters depending on their preference for proximal/distal and awkward/smooth motor movements. We found that participants who preferred distal-smooth movements produced more linear trajectories than those who preferred proximal-awkward movements.

9.1 Evaluating Human Cognition of Containing Relations with Physical Simulation

Containers are ubiquitous objects in daily life, such as bowls, bottles, baskets, trash cans, refrigerators, *etc.*. Containing relation is a general and fundamental relation in the scene. Containers offer containing relations for carrying, hiding, or ensuring the objects remain in a safe place. The contained objects are called contents. The containing relation characterizes the “affordance” that how likely a container can hold its content.

Different from visual object recognition problems, recognition of containers involves the cognitive process of commonsense reasoning, such as analysis of physical properties, geomet-

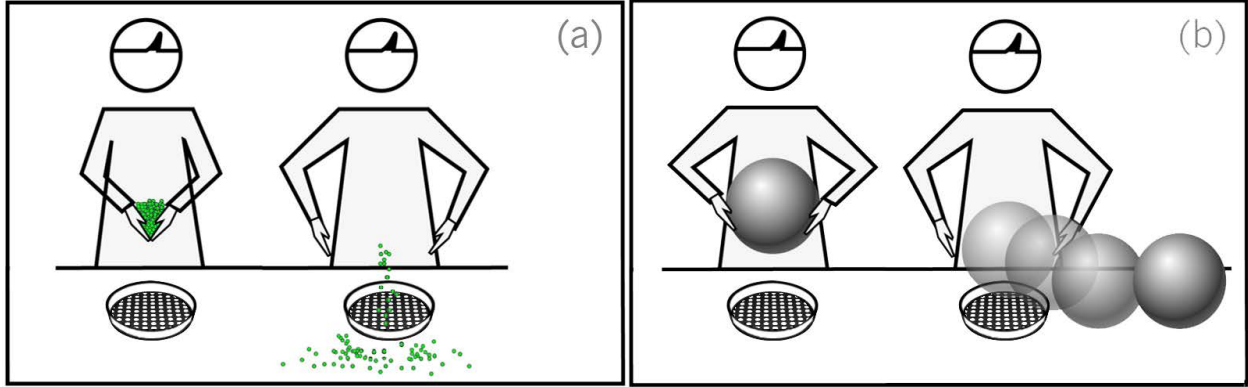


Figure 9.1: Two typical cases when a container fails to contain its contents: (a) the container with holes can not contain tiny objects; (b) the container with a low wall fails to contain a big ball. The left figures of these two panels illustrate a stimuli of our experiments, and the right figures illustrate simulation results with physical engine or in human mind.

ric shapes, and material properties, *etc.*. Figure 9.1 shows two examples when a container fails to contain its content: (a) the container with holes cannot contain tiny objects or staffs, like beads, sand or water; (b) the container with a low wall fails to contain a big ball.

Containers quantize and organize our perceptual scene space. For example, when people are asked “where the chilled beer is”, the answer will usually be that “it is in the refrigerator” without mentioning the exact 3D coordinates. By containers, the perceptual space of 3D scene is discretized and quantized, and objects are often organized in a hierarchy with respect to their containing relations [ZZ13]. This quantization largely simplifies many tasks, such as planning, detection and tracking.

Inspired by [BHT13] and [ZZY13], human perceive physical scenes by making approximate and probabilistic inference, and the physical engine helps us to reason about commonsense in complex scenes. When we ask about whether a container will hold another object, human may do similar mental simulations. The definition of containers are related to physical properties of containers and contents. In Figure 9.1, the container and its contents are not compatible in these two cases. In this work, we model and infer the containing relations between two objects by imagining what would happen when one puts an object



Figure 9.2: A 3D Structure Sensor attached to a tablet (left) and a physical simulation interface (right) are used in this work. The interface simulates a few balls falling onto a bag into a container.

In order to study containers and the factors which affect containing relations, we collected a 3D container dataset and carry out our experiments based on it. In the experiment, we presented some random sampled 3D objects from our dataset to the subjects. The subjects answered questions about container and containing relations according to these pictures. We also built an online physical simulation system with Unity 3D engine on a tablet platform as shown in Figure 9.2. The system is used for evaluating containing relations between objects and comparing with human judgments.

9.2 Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics

Imagine that you are preparing to pour pancake batter onto a griddle. To pour the correct amount, you must decide where to hold the container, at what angle, and for how long. We encounter similar situations frequently in our daily lives when interacting with viscous fluids ranging from water to honey, and with different volumes, contained in receptacles of various shapes and sizes.

In the majority of these situations, we are able to reason about fluid-related physical processes so as to implicitly predict how far a filled container can be tilted before the fluid

inside begins to spill over its rim. However, people perform significantly worse when asked to make explicit reasoning judgments in similar situations [MP91, SB99]. In the well-known Piagetian water level task (WLT; [How78]), participants receive instructions to draw the water level at indicated locations on the inside of tilted containers. Surprisingly, about 40% of adults predict water levels that deviate from the horizontal by 5 degrees or more (*e.g.*, [MP91]). [SB99] modified the WLT to include two containers, one wider than the other. The investigators asked participants to judge which container would need to be tilted farther before the water inside begins to pour out. Only 34% of the participants correctly reported that the thinner container would need to be tilted farther than the wider one. However, when instructed to complete the task by closing their eyes and imagining the same situation, nearly all (95% of) the participants rotated a thinner, imaginary container (or a real, empty one) farther. These findings suggest that people are able to reason successfully about relative pour angles by mentally simulating the tilting event. An apparent contrast in human performance between an explicit reasoning task and a simulated-doing task has also been found in people’s inferences about the trajectories of falling objects [KFW93, SBV13]. Thus, empirical findings in the literature of physical reasoning suggest that people employ both explicit knowledge about physical rules *and* mental simulation when making inferences [Heg04].

The *noisy Newton* framework for physical reasoning hypothesizes that inferences about dynamical systems can be generated by combining noisy perceptual inputs with the principles of classical (*i.e.*, Newtonian) mechanics, given prior beliefs about represented variables [BBY15, BHT13, GGL15, San14, SMG13, SV13]. In this framework, the locations, motions and physical attributes of objects are sampled from noisy distributions and propagated forward in time using an *intuitive physics engine*. The resulting predictions are queried and averaged across simulations to determine the probability of the associated human judgment. [BBY15] extended the framework from physical scene understanding (*e.g.*, [BHT13]) to fluid dynamics using an *intuitive fluid engine* (IFE), where future fluid states are approximated by probabilistic simulation via a Smoothed Particle Hydrodynamics (SPH) method [Mon92]. The particle-based IFE model matched human judgments about final fluid states and provided a better quantitative fit than alternative models that did not employ simulation or

account for physical uncertainty.

The present study aims to determine whether a particle-based IFE model coupled with noisy input variables can account for human judgments about the relative pour angle of two containers filled with fluids differing in their volume and viscosity. The experiment reported here was inspired by previous empirical findings in water-pouring tasks; *e.g.*, participants tilt containers filled with imagined molasses farther than those with an equal volume of water, suggesting that people are able to take physical attributes such as viscosity into account when making fluid-related judgments [SB99]. [BBY15] also found that their participants' judgments were sensitive to latent attributes of the fluid (*e.g.*, stickiness and viscosity).

To quantify the extent that humans employ their perceived viscosity of fluids in subsequent reasoning tasks, we utilized a recent development in graphical fluid simulation [Bri08, JSS15] to simulate the dynamic behavior of fluids in vivid animations. Previous work has shown that realistic animations can facilitate representation of *dynamic* physical situations [TMB02]. Furthermore, recent research on human visual recognition indicates that latent attributes of fluids (*e.g.*, viscosity) are primarily perceived from visual motion cues [KMF15], so displaying realistic fluid movement is needed to provide the input of key physical properties that enable mental fluid simulations. The present study, which uses a modification of [SB99]'s water-pouring problem coupled with advanced techniques in computer graphics, aims to test the hypothesis that animated demonstrations of flow behavior facilitate inference of latent fluid attributes, which inform mental simulations and enhance performance in subsequent reasoning tasks.

9.3 Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem

Consider *KerPlunk*, a children's game in which marbles are suspended in the air by a lattice of straws within a cylindrical tube. The goal of the game is for each player to take turns removing straws while minimizing the number of marbles that fall through the lattice. The task requires players to reason about the interaction between rigid bodies and obstacles in

3D space. But what if the marbles were replaced by balls of liquid or sand? Could humans predict how those substances would move? Would those predictions agree with a generative model based on ground-truth, Newtonian physics?

Recent computational evidence has demonstrated that human predictions *do* agree with Newtonian physics, given noisy perception and prior beliefs about spatially represented variables: *i.e.*, the *noisy Newton* hypothesis [BBY15, BHT13, GGL15, HBG16, KJZ16, San14, SMG13, SBV13]. The hypothesis suggests that humans rationally infer the values of physical variables and utilize normative conservation principles (approximately) to make predictions about future scene states. Computationally, this is achieved by sampling the initial locations, motions from noisy sensory input, and sampling physical attributes in a physical scene, propagating these variables forward in time according to approximated physical principles, and aggregating queries on the final scene states to form predicted response distributions.

[BBY15] extended the noisy Newton framework from block tower judgments [BHT13] to liquid dynamics using an intuitive fluid engine (IFE). In their IFE, ground-truth physics was approximated using smoothed particle hydrodynamics (SPH [Mon92]), a particle-based computational method for simulating non-solid dynamics. Their model predictions matched human judgments about future fluid states and outperformed alternative models that did not employ probabilistic simulation or account for physical uncertainty. Furthermore, the authors found that their participants' predictions were sensitive to latent fluid attributes (stickiness and viscosity), suggesting that humans have rich knowledge about the intrinsic properties of liquid.

The present study argues for the same general class of model as [BBY15] IFE and extends their work by examining (1) whether human predictions about future states of multiple substances (*i.e.*, rigid balls, liquid, and sand) differ, and (2) whether those differences can be consistently modeled using approximate, probabilistic simulation based on a hybrid particle/grid simulator adapted from previous work [KJZ16]. Although granular materials (*e.g.*, sand) are encountered in everyday life, they are far less common than liquid; can humans accurately predict how sand will interact with obstacles and support surfaces? We present two experiments exploring the human capacity to predict the dynamics of substances varying in

familiarity and physical properties, examining how human judgments and model predictions vary for different substances. Experiment 1 examines human predictions about the resting composition of sand after pouring from a funnel. In Experiment 2, participants make predictions about the flow of liquid, sand, and rigid balls past obstacles using a design similar to [BBY15]’s study.

9.4 The Martian: Examining Human Physical Judgments Across Virtual Gravity Fields

Sending a manned spacecraft to Mars would be a fantastic adventure, yet living on another planet could lead to significant challenges to human perceptual and cognitive systems. The change in gravity itself could alter daily activities (*e.g.*, throwing an object toward a desired location or pouring water into a container) that require adjustment of prediction and action in light of changing physical properties on the new planet. Imagine that you are living in an environment with a different gravity field than earth. Would you be able to adapt to it quickly? And how accurate would your predictions about the physical world be compared to when you were on earth?

Consumer-level virtual reality (VR) devices, with rapidly increasing popularity, provide a useful means for researchers to conduct experiments that were traditionally too costly or impossible to carry out in the real world. VR allows users to experience an artificial world in a manner similar to how they experience the real world: *i.e.*, head-mounted displays give the impression of three-dimensional observation, and remote controllers afford interactions with the virtual world from an embodied egocentric perspective. In particular, VR technology allows for both the control of many underlying factors of the virtual world (*e.g.*, time [SBS16] and gravity) and direct measurement of behavioral changes in novel environments.

In this work, we conducted four experiments to measure human performance in different tasks under novel and familiar gravity fields. In the first two experiments, participants were asked to strike a ball off of a track onto a target location and to trigger a ball to hit a target given a speed rating input. In Experiments 3 and 4, participants were asked to make predic-

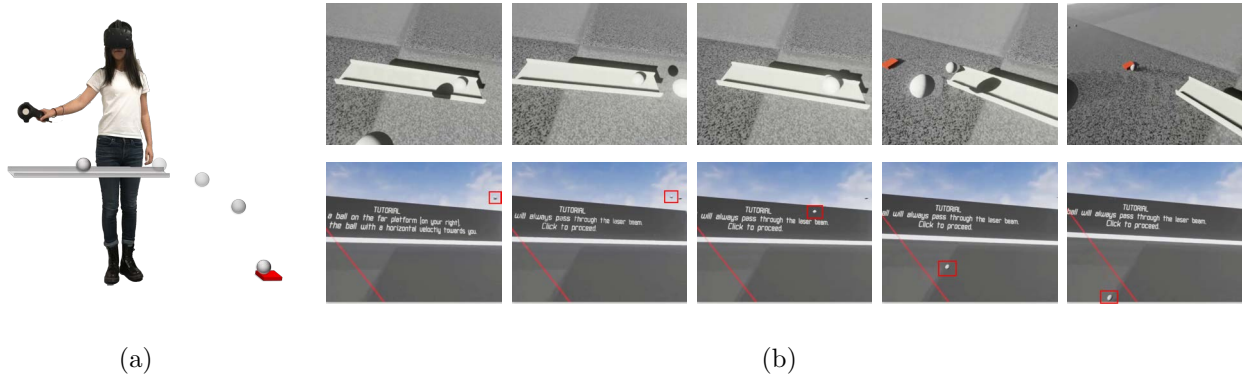


Figure 9.3: Illustration of experiment designs. (b) Two examples of the experiments: (upper) speed production and (lower) trajectory prediction.

tions about the location and flight duration of a projectile given the initial 0.2 seconds of its trajectory. The purpose of the experiments was to examine how humans learn and reason about object motion in novel gravity fields: are humans able to spontaneously habituate to new gravity fields? Do humans implicitly use prior knowledge about earth’s gravity to reason about new environments? Are humans implicitly simulating physical motion or predicting the movements using low-level visual features exclusively?

The first pair of experiments in the present work compare human performance in the VR setting with findings in similar real-world situations [KFW93]. The second pair of experiments compare two types of intuitive physical judgments (location predictions and time estimates) under different gravity fields. In summary, this work made the following contributions: 1) replicated a previous study on speed production and rating in novel virtual environments to demonstrate that VR is a feasible and reliable tool for studying human perception and cognition, 2) carried out a novel experimental design and method which precludes real-world replication, and 3) measured the effect of gravity field on human behavior in tasks varying in their cognitive demands.

9.5 Visuomotor Adaptation and Sensory Recalibration in Reversed Hand Movement Task

In the present study, we examined whether humans can adapt to environments where visual estimates of objects' positions are inconsistent with proprioceptive input. Participants interacted with virtual targets using two motion controllers in a VR application, where the movement of the virtual controller either matched the motion of the physical controller or was flipped on certain axes (both vertical and left-right). Participants were instructed to touch a series of virtual targets with the virtual controllers and then return to a neutral pose in between targets. Response time and arm movement trajectories were recorded and analyzed.

CHAPTER 10

Causal Reasoning

In this chapter, we present two studies on causal reasoning.

In section 10.1, we study the relation between the spatially perturbed collision sound and the perceived causality. When a moving object collides with an object at rest, people immediately perceive a causal event: *i.e.*, the first object has *launched* the second object forwards. However, when the second object's motion is delayed, or is accompanied by a collision sound, causal impressions attenuate and strengthen. Despite a rich literature on causal perception, researchers have exclusively utilized 2D visual displays to examine the launching effect. It remains unclear whether people are equally sensitive to the spatiotemporal properties of observed collisions in the real world. The present study first examined whether previous findings in causal perception with audiovisual inputs can be extended to immersive 3D virtual environments. We then investigated whether perceived causality is influenced by variations in the spatial position of an auditory collision indicator. We found that people are able to localize sound positions based on auditory inputs in VR environments, and spatial discrepancy between the estimated position of the collision sound and the visually observed impact location attenuates perceived causality.

In section 10.2, we study the capability of causal discovery in transfer cases. Discovery and application of causal knowledge in novel problem contexts is a prime example of human intelligence. As new information is obtained from the environment, people develop and refine causal schemas to establish a parsimonious explanation of underlying problem constraints. The aim of the current study is to systematically examine the human ability to discover causal schemas by exploring the environment and transferring knowledge to new situations with greater structural complexity. We developed a novel OpenLock task, in which participants

explored a virtual “escape room” environment by moving levers that served as “locks” to open a door. In each situation, the sequential movements of the levers that opened the door formed a branching causal sequence that began with either a common-cause (CC) or a common-effect (CE) structure. Participants in a baseline condition completed five trials with high structural complexity (*i.e.*, four active levers). Those in the transfer conditions completed six training trials with low structural complexity (*i.e.*, three active levers) before completing a high-complexity transfer trial. The causal schema acquired in the transfer condition was either congruent or incongruent with that in the transfer condition. Baseline performance under the CC schema was superior to performance under the CE schema, and schema congruency facilitated transfer performance when the congruent schema was the less difficult CC schema. We compared human performance to a deep reinforcement learning model and found that our deep reinforcement learning model is unable to capture the causal abstraction presented between trials with the same causal schema and trials with a transfer of causal schema.

10.1 Spatially Perturbed Collision Sounds Attenuate Perceived Causality in 3D Launching Events

Consider the following visual display: a red circle moves in a straight line towards a blue circle until the edges of the two circles touch. After the two circles make contact, the blue circle moves away from the red circle along the same straight line. Although there is no directly observable information in the display signaling a causal connection between the motions of the two circles, you will most likely perceive the red circle as having *launched* the blue circle forward [Hum78]. This is an example of causal perception: *i.e.*, the immediate, automatic, and irresistible impression of causality and animacy from low-level perceptual inputs [ST00]. Such impressions lie in contrast with high-level causal inference, which describes how real-world interpretations are made using logical rules and conceptual knowledge [RFD05].

A key characteristic of causal perception is that impressions are constructed at the perceptual level and do not rely on explicit background knowledge or experience [Mic63, RFD05,

SS92, ST00]. However, causal impressions are incredibly sensitive to the spatiotemporal properties of dynamic events [Boy60, Mic63, Nat61]. For example, in the aforementioned visual display, (1) if there is a temporal delay between when the red circle stops and the blue circle begins moving, (2) if the edges of the circles are separated or overlapped at the time of impact, or (3) if the blue ball moves perpendicular to the red ball’s motion prior to impact, the impression that the red ball launched the blue ball will diminish [MW14, SMG13, SN02]. Impressions of launching are also influenced by briefly observed motions of nearby shapes [SN02], indicating that the human perceptual system rapidly processes spatiotemporal information and visual context information to form immediate causal impressions from perceptual inputs [FRC05, RFD05].

To date, researchers have exclusively utilized 2D visual displays to examine causal perception [MW14, SMG13, ST00, SN02]. This is largely due to the difficulty in varying the spatiotemporal characteristics of moving objects in the real world. Specifically, it is prohibitively difficult to “pause” a real-world collision at the moment of impact without some costly external apparatus: *e.g.*, using magnets and a digital controller to move metallic objects at a given speed across an opaque track. However, virtual reality (VR) provides the means to manipulate such characteristics in immersive 3D environments. A secondary manipulation that VR technology affords is the perturbation of a sound’s location in 3D space. Previous work has shown that when a collision event is accompanied by an auditory cue indicating contact between two objects (*e.g.*, a *clack* sound), observers report a greater causal impression than when the sound is absent [GT03]. It remains unclear, however, whether the human perceptual system encodes the location of the sound when forming such impressions, as it does when it infers that a ventriloquist’s voice emanates from a nearby dummy [AB04]. Thus, the present study sought to answer the following questions: (1) do classical findings in causal perception extend to 3D virtual environments, and (2) does the spatial position of an auditory collision indicator influence perceived causality?

Three experiments were conducted to address these questions. In each experiment, an initially moving red object collides with an initially stationary blue object, and after a 0 to 400 msec delay, the blue object begins moving forwards (see Figure 10.1 (b)(d)). In Experiment 1,

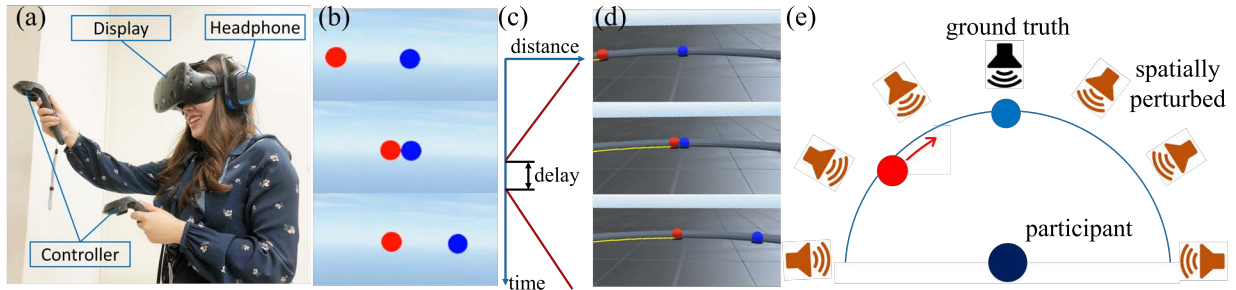


Figure 10.1: Illustration of (a) VR materials utilized in the present study as well as the virtual environment for (b) 2D and (d) 3D collision events. (c) In each environment, the red ball moves to the impact location, stops, and after a delay the blue ball moves forwards. (e) The collision was accompanied by an auditory collision indicator which was spatially perturbed across trials in the 3D environment.

participants reported causal impressions in 2D launching events in the presence and absence of an auditory collision indicator. The first experiment was a direct replication of Guski and Troje’s [GT03] previous study and was designed to determine whether their findings extend to tasks presented via a VR apparatus. In Experiment 2, participants completed an identical task but in a 3D virtual environment. The collision sound in the second experiment was always located at the ground-truth position: *i.e.*, at the location of impact. The purpose of Experiment 2 was to ensure that previous findings in audiovisual causal perception extend to 3D collision situations. Experiment 3 was identical to Experiment 2, except that the spatial position of the auditory collision indicator was perturbed $\pm 90^\circ$ around the observer in increments of 30° (see Figure 10.1 (e)).

In summary, the present study made the following contributions: (1) replicated previous work of causal perception in a virtual environment to demonstrate the viability of VR technology in examinations of human perception and cognition, (2) examined the effect of spatially perturbed auditory collision indicators on impressions of causality in delayed launching events, and (3) measured how well humans can estimate sound location in a VR setup.

10.2 Deep Reinforcement Learning Fails to Account for Human Causal Transfer

Causality has been dubbed the “cement of the university” [Mac74]. The key research question in the field of causal learning is how various intelligent systems, ranging from rats to humans and machines, can acquire knowledge about cause-effect relations in novel situations. Decades ago, a number of researchers (*e.g.*, [SD88, Sha91]) suggested that causal knowledge can be acquired by a basic learning mechanism, associative learning, that non-human animals commonly employ in classical conditioning paradigms to learn the relationship between stimuli and responses. A major theoretical account of associative learning is the Rescorla-Wagner model, guided by prediction error in updated associative weights on cue-effect links [RW72].

However, subsequent research has produced extensive evidence that human causal learning depends on more sophisticated processes than associative learning of cue-effect links [HC11]. Human learning and reasoning involves the acquisition of abstract causal structure [WH92] and strength values for cause-effect relations [Che97]. Causal graphical models [Pea00] have been integrated with Bayesian statistical inference [GT05, GT09, LYL08] to provide a general representational framework for human causal learning [HC11].

However, most models of human causal learning assume that the hypothesis space of causal variables and causal structures is given, and that inference focuses on selecting the best causal structure to explain the observed contingency information relating causal cues to effects. It is unclear how an agent could actively explore a completely novel situation and narrow down the set of potential causal structures to enable efficient inference.

In situations in which outcomes depend on the learner’s actions rather than simply observations, reinforcement learning (RL) is a widely-used modeling tool. It is useful for designing autonomous, dynamic agents capable of exploration in complex environments. RL focuses on learning what to do by mapping situations to actions, so as to maximize a reward signal [SB98]. RL has historically been closely linked with associative learning theory and conceives of learning as essentially a process of trial and error. The connection between classical

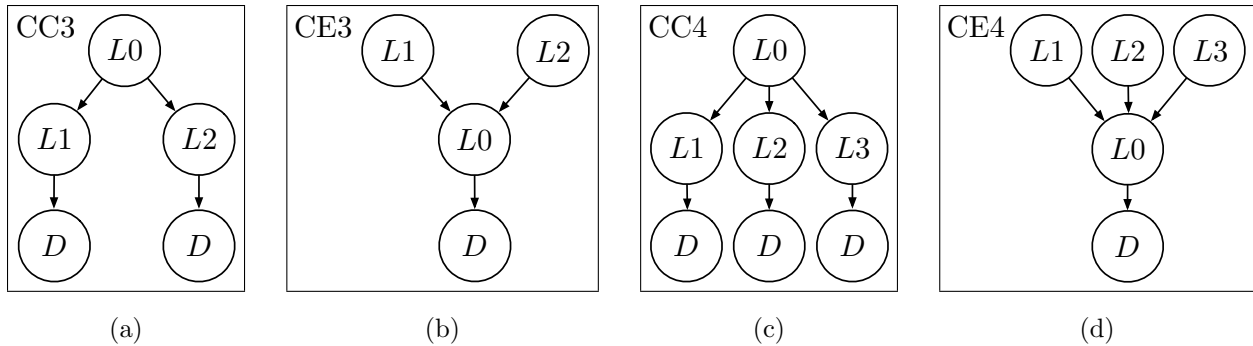


Figure 10.2: Common cause (CC) and common effect (CE) structures used in the present study. D indicates the effect of opening the door. (a) CC3 condition, three lock cues; (b) CE3 condition, three lock cues. (c) CC4 condition, four lock cues; (d) CE4 condition, four lock cues.

conditioning and temporal-difference learning, a central element of RL, is widely acknowledged [SB90]. Hence, RL could be considered as a modern version of associative learning, where learning is not only guided by prediction error but also by other learning mechanisms, notably the estimation of the reward function. Recent advances in RL, especially deep RL, have demonstrated impressive success in applications involving the design of autonomous, dynamic agents for exploration, including playing Atari and Go [MKS15, VGS16, SHM16] and learning complex robot control policies [LFD16].

With these significant developments in RL, is it possible for modern learning models to acquire human-like causal knowledge? To address this question, we designed a novel task to examine learning of action sequences governed by different causal structures, allowing us to determine in what situations humans can transfer their learned causal knowledge. Our design involves two types of basic causal structures (common cause and common effect; see Figure 10.2). When multiple causal chains are consolidated into a single structure, they can form either common cause or common effect schemas. Previous studies using an observational paradigm have found an asymmetry in human learning for common-cause and common-effect structures [WH92].

To design a novel environment for humans, we developed a virtual “escape room”. Imagine that you find yourself trapped in an empty room where the only means of escape is

through a door that will not open. Although there is no visible keyhole on the door—nor do you see any keys lying around—there are some conspicuous levers sticking out of the walls. Your first instinct might be to pull the levers at random to see what happens, and given the outcome, you might revise your theory about how lever interactions relate to the opening of the door. We refer to this underlying theory as a causal schema: *i.e.*, a conceptual organization of events identified as cause and effect [Hei58]. These schemas are discovered with experience and can potentially be transferred to novel target problems to infer their characteristics [KLH17].

In the escape room example, one method of unlocking the door is to induce the causal schema connecting lever interactions to the door’s locking mechanism. However, it remains unclear whether people are equally proficient in uncovering common cause and common effect schemas in novel situations. In the current study, we first assessed whether human causal learning can be impacted by the underlying structure, comparing learning of a common cause structure with learning of a common effect structure. We then examined whether learning one type of causal structure can facilitate subsequent learning of a more complex version of the same schema involving a greater number of causal variables. We compared human performance in a range of learning situations with that of a deep RL model to determine whether behavioral trends can be captured by an algorithm which learns solely by reward optimization, with no prior knowledge about causal structure.

CHAPTER 11

Conclusion

In this dissertation, we show that by integrating four hidden dimensions—functionality, physics, causality and utility—into the current computer vision and robotics systems, we are able to advance the state-of-the-art methods, in particular in the context of learning from limited training examples and transfer learning.

In Part I, we design methods that go beyond example-based methods: a) recognize an object by its essential purposes of use, b) recognize a scene by their functions to serve human activities, c) recognize the containment relations by its causal effects, d) reason about long-term indirect intangible affordance in daily scenes, and e) synthesis realistic scene configuration by integrating functionality and affordance. We argue that it is the missing dimensions of objects and scenes—functionality, physics, causality and utility—that decide their designs of geometry and appearance, as well as the planning of human actions and events. Our objective is to develop methods that “understand” objects, scenes and actions, not merely classify them by memorizing typical examples. This is crucial for generalizing to novel examples in tests.

In Part II, we reason about the hidden and time-varying facts of objects and scenes, and the actions that cause their changes, going beyond passive recognition. As each task is executed by actions to change object status until a desired composite state is reached, by understanding the unfolding tasks, we can reason about the scene history, through past experiences or physics-based simulation—what has resulted in the observed scene, what are the current invisible factors, and predict what will happen next.

In Part III, a set of cognitive studies have shown that human perception of liquid, sand, and motion indeed can be explained by the intuitive physics theories, and have shown that

human demonstrated a superior ability of causal reasoning, especially on the abstract knowledge, which is extremely challenging for the current state-of-the-art machine learning methods.

Going forward, two other critical missing dimensions that are not included in this dissertations are human intention [QZ18, QHW17] and social interactions [SPF18, SXR15, SRZ16, STZ17, SGR17], which worth further explorations in the future.

REFERENCES

- [AAD11] Eren Erdal Aksoy, Alexey Abramov, Johannes Dörr, Kejun Ning, Babette Dellen, and Florentin Wörgötter. “Learning the semantics of object–action relations by observation.” *International Journal of Robotics Research (IJRR)*, **30**(10):1229–1249, 2011.
- [AAW10] Eren Erdal Aksoy, Alexey Abramov, Florentin Wörgötter, and Babette Dellen. “Categorizing object-action relations from semantic scene graphs.” In *International Conference on Robotics and Automation (ICRA)*, 2010.
- [AB04] David Alais and David Burr. “The ventriloquist effect results from near-optimal bimodal integration.” *Current Biology*, **14**(3):257–262, 2004.
- [ABO16] Vinicius C Azevedo, Christopher Batty, and Manuel M Oliveira. “Preserving geometry and topology for fluid flows with thin obstacles and narrow gaps.” *ACM Transactions on Graphics (TOG)*, **35**(4):97, 2016.
- [ACV09] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. “A survey of robot learning from demonstration.” *Robotics and Autonomous Systems (RAS)*, **57**(5):469–483, 2009.
- [AFS06] Marco Attene, Bianca Falcidieno, and Michela Spagnuolo. “Hierarchical mesh segmentation based on fitting primitives.” *The Visual Computer*, **22**(3):181–193, 2006.
- [Ald85] David J Aldous. “Exchangeability and related topics.” In *École d’Été de Probabilités de Saint-Flour XIII1983*. Springer, 1985.
- [Alt92] Naomi S Altman. “An introduction to kernel and nearest-neighbor nonparametric regression.” *The American Statistician*, **46**(3):175–185, 1992.
- [AS11] Anton Andriyenko and Konrad Schindler. “Multi-target tracking by continuous energy minimization.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [ASF11] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. “Search in the real world: Active visual object search based on spatial relations.” In *International Conference on Robotics and Automation (ICRA)*, 2011.
- [Bab03] Christopher Baber. *Cognition and tool use: Forms of engagement in human and animal use of tools*. CRC Press, 2003.
- [BBB07] Christopher Batty, Florence Bertails, and Robert Bridson. “A fast variational framework for accurate solid-fluid coupling.” *ACM Transactions on Graphics (TOG)*, **26**(3):100, 2007.

- [BBC07] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. “Analysis of representations for domain adaptation.” In *Advances in Neural Information Processing Systems (NIPS)*, 2007.
- [BBS09] Steffen Bickel, Michael Brückner, and Tobias Scheffer. “Discriminative learning under covariate shift.” *Journal of Machine Learning Research*, **10**(Sep):2137–2155, 2009.
- [BBY15] Christopher Bates, Peter Battaglia, Ilker Yildirim, and Joshua B Tenenbaum. “Humans predict liquid dynamics using probabilistic simulation.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2015.
- [BDW81] FO Bartell, EL Dereniak, and WL Wolfe. “The theory and measurement of bidirectional reflectance distribution function (BRDF) and bidirectional transmittance distribution function (BTDF).” In *Radiation scattering in optical systems*, 1981.
- [BE08] Lawrence E Blume and David Easley. “Rationality.” In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.
- [Bec74] Gary S Becker. “Crime and punishment: An economic approach.” In *Essays in the Economics of Crime and Punishment*. NBER, 1974.
- [Bec80] Benjamin B Beck. *Animal tool behavior: the use and manufacture of tools by animals*. Garland STPM Publisher, 1980.
- [BF03] Bert Bredeweg and Kenneth D Forbus. “Qualitative modeling in education.” *AI magazine*, **24**(4):35, 2003.
- [BF08] Marcus A Brubaker and David J Fleet. “The kneed walker for human pose tracking.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [BFH10] Marcus A Brubaker, David J Fleet, and Aaron Hertzmann. “Physics-based person tracking using the anthropomorphic walker.” *International Journal of Computer Vision (IJCV)*, **87**(1-2):140–155, 2010.
- [BFT11] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. “Multiple object tracking using k-shortest paths optimization.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **33**(9):1806–1819, 2011.
- [BHT13] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. “Simulation as an engine of physical scene understanding.” *Proceedings of the National Academy of Sciences (PNAS)*, **110**(45):18327–18332, 2013.
- [BKW98] Werner GK Backhaus, Reinhold Kliegl, and John S Werner. *Color vision: Perspectives from different disciplines*. Walter de Gruyter, 1998.

- [BLC00] Michael M Blane, Zhibin Lei, Hakan Çivi, and David B Cooper. “The 3L algorithm for fitting implicit polynomial curves and surfaces to data.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **22**(3):298–313, 2000.
- [BMP06] John Blitzer, Ryan McDonald, and Fernando Pereira. “Domain adaptation with structural correspondence learning.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 2006.
- [Boy60] D. G. Boyle. “A contribution to the study of phenomenal causation.” *Quarterly Journal of Experimental Psychology*, **12**(3):171–179, 1960.
- [BRG16] Aayush Bansal, Bryan Russell, and Abhinav Gupta. “Marr revisited: 2d-3d alignment via surface normal prediction.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [Bri07] Robert Bridson. “Fast Poisson disk sampling in arbitrary dimensions.” In *SIGGRAPH sketches*, 2007.
- [Bri08] Robert Bridson. *Fluid simulation for computer graphics*. CRC Press, 2008.
- [Bri15] Robert Bridson. *Fluid simulation for computer graphics*. CRC Press, 2015.
- [BS17] Noam Brown and Tuomas Sandholm. “Superhuman AI for heads-up no-limit poker: Libratus beats top professionals.” *Science*, p. eaao1733, 2017.
- [BSF09] Marcus A Brubaker, Leonid Sigal, and David J Fleet. “Estimating contact dynamics.” In *International Conference on Computer Vision (ICCV)*, 2009.
- [BVR16] Maros Blaha, Christoph Vogel, Audrey Richard, Jan D Wegner, Thomas Pock, and Konrad Schindler. “Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [BW89] Richard W Byrne and Andrew Whiten. *Machiavellian Intelligence: Social Expertise And The Evolution Of Intellect In Monkeys, Apes, And Humans (Oxford Science)*. Oxford University Press, 1989.
- [BZ05] Adrian Barbu and Song-Chun Zhu. “Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **27**(8):1239–1253, 2005.
- [CCL99] Laura A Carlson-Radvansky, Eric S Covey, and Kathleen M Lattanzi. ““What” Effects on “Where”: Functional Influences on Spatial Relations.” *Psychological Science*, **10**(6):516–521, 1999.
- [CCP13] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. “Understanding indoor scenes using 3d geometric phrases.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

- [CF87] John W Collins and Kenneth D Forbus. “Reasoning about fluids via molecular collections.” In *AAAI Conference on Artificial Intelligence (AAAI)*, 1987.
- [CFG15a] 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.
- [CFG15b] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. “ShapeNet: An Information-Rich 3D Model Repository.” *arXiv preprint arXiv:1512.03012*, 2015.
- [CGB07] Sylvain Calinon, Florent Guenter, and Aude Billard. “On learning, representing, and generalizing a task in a humanoid robot.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **37**(2):286–298, 2007.
- [CH01] Anthony G. Cohn and Shyamanta M. Hazarika. “Qualitative spatial representation and reasoning: An overview.” *Fundamenta informaticae*, **46**(1-2):1–29, 2001.
- [CH05a] Miguel A Carreira-Perpinan and Geoffrey E Hinton. “On contrastive divergence learning.” In *AI Stats*, volume 10, pp. 33–40, 2005.
- [CH05b] Olivier Chapelle and Zaid Harchaoui. “A machine learning approach to conjoint analysis.” In *Advances in Neural Information Processing Systems (NIPS)*, 2005.
- [Che97] Patricia W Cheng. “From covariation to causation: a causal power theory.” *Psychological Review*, **104**(2):367–405, 1997.
- [CKP16] Albert Chern, Felix Knöppel, Ulrich Pinkall, Peter Schröder, and Steffen Weissmann. “Schrödinger’s smoke.” *ACM Transactions on Graphics (TOG)*, **35**(4):77, 2016.
- [CL05] Sarah H Creem-Regehr and James N Lee. “Neural representations of graspable objects: are tools special?” *Cognitive Brain Research*, **22**(3):457–469, 2005.
- [CMR08] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. “Sample selection bias correction theory.” In *International Conference on Algorithmic Learning Theory*, 2008.
- [Csu17] Gabriela Csurka. “Domain adaptation for visual applications: A comprehensive survey.” *arXiv preprint arXiv:1702.05374*, 2017.
- [CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks.” *Machine Learning*, **20**(3):273–297, 1995.
- [CWL16] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Dani Lischinsk, Daniel Cohen-Or, Baoquan Chen, et al. “Synthesizing Training Images for Boosting Human 3D Pose Estimation.” In *International Conference on 3D Vision (3DV)*, 2016.

- [CZK15] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. “Robust Reconstruction of Indoor Scenes.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [Dau07] Hal Daumé III. “Frustratingly Easy Domain Adaptation.” In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.
- [Dau09] Hal Daumé III. “Bayesian multitask learning with latent hierarchies.” In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.
- [DCH16] Yan Duan, Xi Chen, Rein Houthoofd, John Schulman, and Pieter Abbeel. “Benchmarking deep reinforcement learning for continuous control.” In *International Conference on Machine Learning (ICML)*, 2016.
- [DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [Del34] Boris Delaunay. “Sur la sphere vide.” *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, **7**(793-800):1–2, 1934.
- [DFI15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. “Flownet: Learning optical flow with convolutional networks.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [DFL12] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. “Scene semantics from long-term observation of people.” In *European Conference on Computer Vision (ECCV)*, 2012.
- [DH02] Y Derimis and G Hayes. “Imitations as a dual-route process featuring predictive and learning components: a biologically plausible computational model.” *Imitation in animals and artifacts*, pp. 327–361, 2002.
- [DH06] Daniel Dunbar and Greg Humphreys. “A spatial data structure for fast Poisson-disk sample generation.” *ACM Transactions on Graphics (TOG)*, **25**(3):503–508, 2006.
- [Dij59] Edsger W Dijkstra. “A note on two problems in connexion with graphs.” *Numerische Mathematik*, **1**(1):269–271, 1959.
- [DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm.” *Journal of the Royal Statistical Society, Series B*, pp. 1–38, 1977.
- [DMC13] Ernest Davis, Gary Marcus, and Angelica Chen. “Reasoning from radically incomplete information: The case of containers.” In *Annual Conference on Advances in Cognitive Systems*, 2013.

- [DSD08] Laura Dipietro, Angelo M Sabatini, and Paolo Dario. “A survey of glove-based systems and their applications.” *Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **38**(4):461–482, 2008.
- [DWL16] Yu Du, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan Kankanhalli, and Weidong Geng. “Marker-less 3D human motion capture with monocular image sequence and height-maps.” In *European Conference on Computer Vision (ECCV)*, 2016.
- [Ear70] Jay Earley. “An efficient context-free parsing algorithm.” *Communications of the ACM*, **13**(2):94–102, 1970.
- [EES10] Markus Enzweiler, Angela Eigenstetter, Bernt Schiele, and Dariu M Gavrilă. “Multi-cue pedestrian classification with partial occlusion handling.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [EEV15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. “The pascal visual object classes challenge: A retrospective.” *International Journal of Computer Vision (IJCV)*, **111**(1), 2015.
- [EF10] Marcin Eichner and Vittorio Ferrari. “We are family: Joint pose estimation of multiple persons.” In *European Conference on Computer Vision (ECCV)*, 2010.
- [EF15] David Eigen and Rob Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [EGX17] Mark Edmonds, Feng Gao, Xu Xie, Hangxin Liu, Siyuan Qi, Yixin Zhu, Brandon Rothrock, and Song-Chun Zhu. “Feeling the Force: Integrating Force and Pose for Fluent Discovery through Imitation Learning to Open Medicine Bottles.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [EP04] Theodoros Evgeniou and Massimiliano Pontil. “Regularized multi-task learning.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.
- [EPF14] David Eigen, Christian Puhrsch, and Rob Fergus. “Depth map prediction from a single image using a multi-scale deep network.” In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [FCS09] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. “Manhattan-world stereo.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [FDG14] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. “People watching: Human actions as a cue for single view geometry.” *International Journal of Computer Vision (IJCV)*, **110**(3):259–274, 2014.

- [FDP97] John T Feddema, Clark R Dohrmann, Gordon G Parker, Rush D Robinett, Vicente J Romero, and Dan J Schmitt. “Control for slosh-free motion of an open container.” *IEEE Control Systems*, **17**(1):29–36, 1997.
- [FF01] Nick Foster and Ronald Fedkiw. “Practical animation of liquids.” In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.
- [FGH13] David F Fouhey, Abhinav Gupta, and Martial Hebert. “Data-driven 3D primitives for single image understanding.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [FH05] Fang Fang and Sheng He. “Cortical responses to invisible objects in the human dorsal and ventral pathways.” *Nature Neuroscience*, **8**(10):1380–1385, 2005.
- [FKI14] Sean Ryan Fanello, Cem Keskin, Shahram Izadi, Pushmeet Kohli, David Kim, David Sweeney, Antonio Criminisi, Jamie Shotton, Sing Bing Kang, and Tim Paek. “Learning to be a depth camera for close-range human capture and interaction.” *ACM Transactions on Graphics (TOG)*, **33**(4):86, 2014.
- [FMJ16] Michael Firman, Oisín Mac Aodha, Simon Julier, and Gabriel J Brostow. “Structured prediction of unobserved voxels from a single depth image.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [FPB06] Gerald Fritz, Lucas Paletta, Ralph Breithaupt, Erich Rome, and Georg Dorffner. “Learning Predictive Features in Affordance based Robotic Perception Systems.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2006.
- [Fra92] Andrew U Frank. “Qualitative spatial reasoning about distances and directions in geographic space.” *Journal of Visual Languages & Computing*, **3**(4):343–371, 1992.
- [Fra96] Andrew U Frank. “Qualitative spatial reasoning: Cardinal directions as an example.” *International Journal of Geographical Information Science*, **10**(3):269–290, 1996.
- [FRC05] J. A. Fugelsang, M. E. Roser, P. M. Corballis, M. S. Gazzaniga, and K. N. Dunbar. “Brain mechanisms underlying perceptual causality.” *Cognitive Brain Research*, **24**(1):41–47, 2005.
- [Fre07] Scott H Frey. “What puts the how in where? Tool use and the divided visual streams hypothesis.” *Cortex*, **43**(3):368–375, 2007.
- [Fri03] Arthur Fridman. “Mixed markov models.” *Proceedings of the National Academy of Sciences (PNAS)*, **100**(14), 2003.
- [FRS12] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. “Example-based synthesis of 3D object arrangements.” *ACM Transactions on Graphics (TOG)*, **31**(6), 2012.

- [FSJ01] Ronald Fedkiw, Jos Stam, and Henrik Wann Jensen. “Visual simulation of smoke.” In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.
- [FZ13a] Amy Sue Fire and Song-Chun Zhu. “Learning perceptual causality from video.” In *Learning Rich Representations from Low-Level Sensors, AAAI Workshop*, 2013.
- [FZ13b] Amy Sue Fire and Song-Chun Zhu. “Using Causal Induction in Humans to Learn and Infer Causality from Video.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2013.
- [GAB17] Ignacio Garcia-Dorado, Daniel G Aliaga, Saiprasanth Bhalachandran, Paul Schmid, and Dev Niyogi. “Fast weather simulation for inverse procedural design of 3d urban models.” *ACM Transactions on Graphics (TOG)*, **36**(2):21, 2017.
- [GAG15] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. “Aligning 3D models to RGB-D images of cluttered scenes.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [GBK02] György Gergely, Harold Bekkering, and Ildikó Király. “Developmental psychology: Rational imitation in preverbal infants.” *Nature*, **415**(6873):755–755, 2002.
- [GDG15] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. “DRAW: A recurrent neural network for image generation.” *arXiv preprint arXiv:1502.04623*, 2015.
- [GGI94] Kathleen R Gibson, Kathleen Rita Gibson, and Tim Ingold. *Tools, language and cognition in human evolution*. Cambridge University Press, 1994.
- [GGL15] Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. “How, whether, why: Causal judgments as counterfactual contrasts.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2015.
- [GGV11] Helmut Grabner, Juergen Gall, and Luc Van Gool. “What makes a chair a chair?” In *International Conference on Computer Vision (ICCV)*, 2011.
- [Gib77] James J Gibson. *The theory of affordances*. Hilldale, USA, 1977.
- [Gib82] Eleanor J Gibson. “The Concept of Affordances in Development: The Renaissance of Functionalism.” In *The concept of development: The Minnesota symposia on child psychology*, volume 15, pp. 55–81. Lawrence Erlbaum Hillsdale, NJ, 1982.
- [Gib14] James J Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, 2014.

- [GKD09] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. “Observing human-object interactions: Using spatial and functional compatibility for recognition.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **31**(10):1775–1789, 2009.
- [GKS16] Mona Fathollahi Ghezalghieh, Rangachar Kasturi, and Sudeep Sarkar. “Learning camera viewpoint using CNN to improve 3D body pose estimation.” In *International Conference on 3D Vision (3DV)*, 2016.
- [GL15] Yaroslav Ganin and Victor Lempitsky. “Unsupervised domain adaptation by backpropagation.” In *International Conference on Machine Learning (ICML)*, 2015.
- [GLK17] Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. “A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs.” *Science*, **358**(6368):eaag2612, 2017.
- [GO04] RJ Geraerts and MH Overmars. “A Comparative Study of Probabilistic Roadmap Planners.” *Tracts in Advanced Robotics*, **7**:43, 2004.
- [Goo86] Jane Goodall. *The chimpanzees of Gombe: Patterns of behavior*. Belknap Press of Harvard University Press, Cambridge, 1986.
- [GR02] Alfonso Gerevini and Jochen Renz. “Combining topological and size information for spatial reasoning.” *Artificial Intelligence*, **137**(1):1–42, 2002.
- [GS04] Pierre Grenon and Barry Smith. “SNAP and SPAN: Towards dynamic spatial ontology.” *Spatial cognition and computation*, **4**(1):69–104, 2004.
- [GSE11] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. “From 3d scene geometry to human workspace.” In *International Conference on Computer Vision (ICCV)*, 2011.
- [GSH09] Arthur Gretton, Alex J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. “Covariate Shift by Kernel Mean Matching.” In *Dataset shift in machine learning*. MIT Press, 2009.
- [GSS15] Theodore F Gast, Craig Schroeder, Alexey Stomakhin, Chenfanfu Jiang, and Joseph M Teran. “Optimization integrator for large time steps.” *IEEE Transactions on Visualization and Computer Graph (TVCG)*, **21**(10):1103–1115, 2015.
- [GT03] R. Guski and N. F. Troje. “Audiovisual phenomenal causality.” *Attention, Perception, & Psychophysics*, **65**(5):789–800, 2003.
- [GT05] Thomas L Griffiths and Joshua B Tenenbaum. “Structure and strength in causal induction.” *Cognitive Psychology*, **51**(4):334–384, 2005.
- [GT09] Thomas L Griffiths and Joshua B Tenenbaum. “Theory-based causal induction.” *Psychological Review*, **116**(4):661–716, 2009.

- [GWC16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. “Virtual worlds as proxy for multi-object tracking analysis.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [HAB08] Céline Hudelot, Jamal Atif, and Isabelle Bloch. “Fuzzy spatial relation ontology for image interpretation.” *Fuzzy Sets and Systems*, pp. 1929–1951, 2008.
- [HBG16] J. B. Hamrick, P. W. Battaglia, T. L. Griffiths, and J. B. Tenenbaum. “Inferring mass in complex scenes by mental simulation.” *Cognition*, **157**:61–76, 2016.
- [HC11] Keith J Holyoak and Patricia W Cheng. “Causal learning and inference as a rational process: The new synthesis.” *Annual Review of Psychology*, **62**:135–163, 2011.
- [Hec77] James J Heckman. “Sample selection bias as a specification error (with an application to the estimation of labor supply functions).”, 1977.
- [Hec90] Paul S Heckbert. “A seed fill algorithm.” In *Graphics Gems*, pp. 275–277. Academic Press Professional, Inc., 1990.
- [Heg04] Mary Hegarty. “Mechanical reasoning by mental simulation.” *Trends in Cognitive Sciences*, **8**(6):280–285, 2004.
- [Hei58] Fritz Heider. *The Psychology of Interpersonal Relations*. Psychology Press, 1958.
- [HHF10] Varsha Hedau, Derek Hoiem, and David Forsyth. “Thinking inside the box: Using appearance models and context based on room geometry.” In *European Conference on Computer Vision (ECCV)*, 2010.
- [Hin02] Geoffrey E Hinton. “Training products of experts by minimizing contrastive divergence.” *Neural Computation*, **14**(8):1771–1800, 2002.
- [HJS13] Jan Hegemann, Chenfanfu Jiang, Craig Schroeder, and Joseph M Teran. “A level set method for ductile fracture.” In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2013.
- [HKL15] Zhanpeng Huang, Ladislav Kavan, Weikai Li, Pan Hui, and Guanghong Gong. “Reducing numerical dissipation in smoke simulation.” *Graphical Models*, **78**:10–25, 2015.
- [HNK15] Hironori Hattori, Vishnu Naresh Boddeti, Kris M Kitani, and Takeo Kanade. “Learning scene-specific pedestrian detectors without real data.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [Ho87] Seng-Beng Ho. *Representing and using functional definitions for visual recognition*. PhD thesis, The University of Wisconsin-Madison, 1987.
- [Hor87] Berthold KP Horn. “Closed-form solution of absolute orientation using unit quaternions.” *Journal of the Optical Society of America A*, **4**(4):629–642, 1987.

- [HOT06] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets.” *Neural Computation*, **18**(7):1527–1554, 2006.
- [How78] Ian P Howard. “Recognition and knowledge of the water-level principle.” *Perception*, **7**(2):151, 1978.
- [HPB16] Ankur Handa, Viorica Pătrăucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. “Understanding real world indoor scenes with synthetic data.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [HPS16] Ankur Handa, Viorica Patraucean, Simon Stent, and Roberto Cipolla. “SceneNet: an Annotated Model Generator for Indoor Scene Understanding.” In *International Conference on Robotics and Automation (ICRA)*, 2016.
- [HRB11] Tucker Hermans, James M Rehg, and Aaron Bobick. “Affordance prediction via learned object attributes.” In *Workshop on Semantic Perception, Mapping, and Exploration, ICRA*, 2011.
- [HS06] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks.” *Science*, **313**(5786):504–507, 2006.
- [HS07] Susan J Hespos and ES Spelke. “Precursors to spatial language: The case of containment.” *The categorization of spatial entities in language and cognition*, pp. 233–245, 2007.
- [HS08] Peter Hedström and Charlotta Stern. “Rational choice and sociology.” In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.
- [HSL17] Nicolas Heess, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, Ali Eslami, Martin Riedmiller, et al. “Emergence of Locomotion Behaviours in Rich Environments.” *arXiv preprint arXiv:1707.02286*, 2017.
- [HSV95] Klaus-Uwe Hoffgen, Hans-Ulrich Simon, and Kevin S Vanhorn. “Robust trainability of single neurons.” *Journal of Computer and System Sciences*, **50**(1):114–125, 1995.
- [Hum78] D. Hume. *A treatise of human nature*. Oxford University Press, Oxford, England, 1738/1978.
- [HWK15] Qixing Huang, Hai Wang, and Vladlen Koltun. “Single-view reconstruction via joint analysis of image and shape collections.” *ACM Transactions on Graphics (TOG)*, **34**(4), 2015.
- [HWM14] Ankish Handa, Thomas Whelan, John McDonald, and Andrew J Davison. “A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM.” In *International Conference on Robotics and Automation (ICRA)*, 2014.

- [HZC13] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. “Joint 3D scene reconstruction and class segmentation.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [IS15] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In *International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.
- [ITF04] Geoffrey Irving, Joseph Teran, and Ronald Fedkiw. “Invertible finite elements for robust simulation of large deformation.” In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2004.
- [JGS13] Zhaoyin Jia, Andrew Gallagher, Ashutosh Saxena, and Tsuhan Chen. “3D-based reasoning with blocks, support, and stability.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [JI11] R Jain and T Inamura. “Learning of Tool Affordances for autonomous tool manipulation.” In *IEEE/SICE International Symposium on System Integration (SII)*, 2011.
- [JKS13] Yun Jiang, Hema Koppula, and Ashutosh Saxena. “Hallucinated humans as the hidden context for labeling 3d scenes.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [JLS12] Yun Jiang, Marcus Lim, and Ashutosh Saxena. “Learning object arrangements in 3D scenes using human context.” In *International Conference on Machine Learning (ICML)*, 2012.
- [Joa02] Thorsten Joachims. “Optimizing search engines using clickthrough data.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.
- [Jol02] Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.
- [JS13a] Yun Jiang and Ashutosh Saxena. “Hallucinating humans for learning robotic placement of objects.” In *Experimental Robotics*, pp. 921–937. Springer, 2013.
- [JS13b] Yun Jiang and Ashutosh Saxena. “Infinite Latent Conditional Random Fields for Modeling Environments through Humans.” In *Robotics: Science and Systems (RSS)*, 2013.
- [JSS15] Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. “The affine particle-in-cell method.” *ACM Transactions on Graphics (TOG)*, **34**(4):51, 2015.

- [JX13] Hao Jiang and Jianxiong Xiao. “A linear approach to matching cuboids in RGBD images.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [JZQ17] Chenfanfu Jiang, Yixin Zhu, Siyuan Qi, Siyuan Huang, Jenny Lin, Xiongwen Guo, Lap-Fai Yu, Demetri Terzopoulos, and Song-Chun Zhu. “Configurable, Photorealistic Image Rendering and Ground Truth Synthesis by Sampling Stochastic Grammars Representing Indoor Scenes.” *arXiv preprint arXiv:1704.00112*, 2017.
- [KB13] Andrea Kleinsmith and Nadia Bianchi-Berthouze. “Affective body expression perception and recognition: A survey.” *Transactions on Affective Computing*, **4**(1):15–33, 2013.
- [KCG14] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. “Shape2pose: Human-centric shape analysis.” *ACM Transactions on Graphics (TOG)*, **33**(4):120, 2014.
- [KDD09] Oussama Khatib, Emel Demircan, Vincent De Sapio, Luis Sentis, Thor Besier, and Scott Delp. “Robotics-based synthesis of human motion.” *Journal of Physiology Paris*, **103**(3):211–219, 2009.
- [KFW93] Horst Krist, Edgar L Fieberg, and Friedrich Wilkening. “Intuitive physics in action and judgment: The development of knowledge about projectile motion.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**(4):952, 1993.
- [KGS13] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. “Learning human activities and object affordances from rgb-d videos.” *International Journal of Robotics Research (IJRR)*, **32**(8):951–970, 2013.
- [KIX16] Yinda Zhang Mingru Bai Pushmeet Kohli, Shahram Izadi, and Jianxiong Xiao. “DeepContext: Context-Encoding Neural Pathways for 3D Holistic Scene Understanding.” *arXiv preprint arXiv:1603.04922*, 2016.
- [KJZ16] J.R. Kubricht, C. Jiang, Y. Zhu, S. Zhu, D. Terzopoulos, and H. Lu. “Probabilistic simulation predicts human performance on viscous fluid-pouring task.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2016.
- [KKT15] Tejas D Kulkarni, Pushmeet Kohli, Joshua B Tenenbaum, and Vikash Mansinghka. “Picture: A probabilistic programming language for scene perception.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [KLH17] James R Kubricht, Hongjing Lu, and Keith J Holyoak. “Individual Differences in Spontaneous Analogical Transfer.” *Memory and Cognition*, **45**(4):576–588, 2017.
- [KMF15] Takahiro Kawabe, Kazushi Maruya, Roland W Fleming, and Shinya Nishida. “Seeing liquids from visual motion.” *Vision Research*, **109**:125–138, 2015.

- [KMM12] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. “Tracking-learning-detection.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **34**(7):1409–1422, 2012.
- [KMY12] Young Min Kim, Niloy J Mitra, Dong-Ming Yan, and Leonidas Guibas. “Acquiring 3d indoor environments with variability and repetition.” *ACM Transactions on Graphics (TOG)*, **31**(6):138, 2012.
- [Kol06] Alexy Kolesnikov. “Use of computational fluid dynamics to predict airflow and contamination concentration profiles within laboratory floor plan environment.” *Applied Biosafety*, **11**(4):197–214, 2006.
- [KPR15] Olaf Kahler, Victor Prisacariu, Carl Ren, Xin Sun, Philip Torr, and David Murray. “Very High Frame Rate Volumetric Integration of Depth Images on Mobile Devices.” *IEEE Transactions on Visualization and Computer Graph (TVCG)*, **21**:1241–1250, 2015.
- [KRK11] Hedvig Kjellström, Javier Romero, and Danica Kragić. “Visual object-action recognition: Inferring object affordances from human demonstration.” *Computer Vision and Image Understanding (CVIU)*, **115**(1):81–90, 2011.
- [KSL96] Lydia E Kavraki, Petr Švestka, Jean-Claude Latombe, and Mark H Overmars. “Probabilistic roadmaps for path planning in high-dimensional configuration spaces.” *IEEE Transactions on Robotics and Automation*, **12**(4):566–580, 1996.
- [KSM17] Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. “Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics.” In *International Conference on Machine Learning (ICML)*, 2017.
- [KWK15] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. “Deep convolutional inverse graphics network.” In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [KZJ17] J.R. Kubricht, Y. Zhu, C. Jiang, D. Terzopoulos, Song-Chun Zhu, and H. Lu. “Consistent Probabilistic Simulation Underlying Human Judgment in Substance Dynamics.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2017.
- [Lau17] J Launchbury. “A DARPA Perspective on Artificial Intelligence.”, 2017.
- [LaV06] Steven M LaValle. *Planning algorithms*. Cambridge University Press, 2006.
- [LB14] Matthew M Loper and Michael J Black. “OpenDR: An approximate differentiable renderer.” In *European Conference on Computer Vision (ECCV)*, 2014.
- [LC87] William E Lorensen and Harvey E Cline. “Marching cubes: A high resolution 3D surface construction algorithm.” *ACM Transactions on Graphics (TOG)*, **21**(4):163–169, 1987.

- [Lew06] James W Lewis. “Cortical networks related to human use of tools.” *The Neuroscientist*, **12**(3):211–231, 2006.
- [LF10] Kevin Lai and Dieter Fox. “Object recognition in 3D point clouds using web data and domain adaptation.” *International Journal of Robotics Research (IJRR)*, **29**(8):1019–1037, 2010.
- [LFD16] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. “End-to-end training of deep visuomotor policies.” *The Journal of Machine Learning Research*, **17**(1):1334–1373, 2016.
- [LFU13] Dahua Lin, Sanja Fidler, and Raquel Urtasun. “Holistic scene understanding for 3d object detection with rgb-d cameras.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [LGS16] Jenny Lin, Xingwen Guo, Jingyu Shao, Chenfanfu Jiang, Yixin Zhu, and Song-Chun Zhu. “A virtual reality platform for dynamic human-scene interaction.” In *Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments, SIGGRAPH ASIA 2016 Workshop*, 2016.
- [Li07] Sanjiang Li. “Combining Topological and Directional Information for Spatial Reasoning.” In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 435–440, 2007.
- [LLS15] Ian Lenz, Honglak Lee, and Ashutosh Saxena. “Deep learning for detecting robotic grasps.” *International Journal of Robotics Research (IJRR)*, **34**(4-5):705–724, 2015.
- [LM11] Mathieu Labbé and François Michaud. “Memory management for real-time appearance-based loop closure detection.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2011.
- [LM13] Mathieu Labbe and Francois Michaud. “Appearance-based loop closure detection for online large-scale and long-term operation.” *Transactions on Robotics (TRO)*, **29**(3):734–745, 2013.
- [LM14] Mathieu Labbe and François Michaud. “Online global loop closure detection for large-scale multi-session graph-based slam.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2014.
- [LMB14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context.” In *European Conference on Computer Vision (ECCV)*, 2014.
- [Loh08] Susanne Lohmann. “Rational choice and political science.” In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Basingstoke, UK, 2008.
- [LPK14] Yoonsang Lee, Moon Seok Park, Taesoo Kwon, and Jehee Lee. “Locomotion control for many-muscle humanoids.” *ACM Transactions on Graphics (TOG)*, **33**(6):218, 2014.

- [LSL15] Fayao Liu, Chunhua Shen, and Guosheng Lin. “Deep convolutional neural fields for depth estimation from a single image.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [LST09] Sung-Hee Lee, Eftychios Sifakis, and Demetri Terzopoulos. “Comprehensive biomechanical modeling and simulation of the upper body.” *ACM Transactions on Graphics (TOG)*, **28**(4):99, 2009.
- [LST15] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. “Human-level concept learning through probabilistic program induction.” *Science*, **350**(6266):1332–1338, 2015.
- [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 *International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [LYL08] Hongjing Lu, Alan L Yuille, Mimi Liljeholm, Patricia W Cheng, and Keith J Holyoak. “Bayesian generic priors for causal learning.” *Psychological Review*, **115**(4):955–984, 2008.
- [LZK17] J.H. Lin, Y. Zhu, J.R. Kubricht, S.-C. Zhu, and H. Lu. “Visuomotor Adaptation and Sensory Recalibration in Reversed Hand Movement Task.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2017.
- [LZS18] Hangxin Liu, Yaofang Zhang, Wenwen Si, Xu Xie, Yixin Zhu, and Song-Chun Zhu. “Interactive Robot Knowledge Patching using Augmented Reality.” In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [LZW16] Yang Lu, Song-Chun Zhu, and Ying Nian Wu. “Learning FRAME Models Using CNN Filters.” In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.
- [LZZ14] Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. “Single-view 3d scene parsing by attributed grammar.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [LZZ15] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. “Evaluating Human Cognition of Containing Relations with Physical Simulation.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2015.
- [LZZ16] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. “What Is Where: Inferring Containment Relations from Videos.” In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.
- [LZZ17] Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. “Single-View 3D Scene Reconstruction and Parsing by Attribute Grammar.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

- [LZZ18] Wei Liang, Yixin Zhu, and Song-Chun Zhu. “Tracking Occluded Objects and Recovering Incomplete Trajectories by Reasoning about Containment Relations and Human Actions.” In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [Mac74] John L Mackie. *The Cement of the Universe: A Study of Causation*. Oxford, 1974.
- [Mar82] David Marr. *Vision: A computational approach*. Freeman.[aAC], 1982.
- [MBT03] Neil Molino, Robert Bridson, Joseph Teran, and Ronald Fedkiw. “A Crystalline, Red Green Strategy for Meshing Highly Deformable Objects with Tetrahedra.” In *International Meshing Roundtable (IMR)*, 2003.
- [McG92] William Clement McGrew. *Chimpanzee material culture: implications for human evolution*. Cambridge University Press, 1992.
- [Mic63] A. Michotte. *The perception of causality*. Basic Books, New York, NY, 1963.
- [Mil88] Gavin SP Miller. “The motion dynamics of snakes and worms.” *ACM Transactions on Graphics (TOG)*, **22**(4):169–173, 1988.
- [Min88] Marvin Minsky. *Society of Mind*. Simon and Schuster, 1988.
- [MKF14] Austin Myers, Angjoo Kanazawa, Cornelia Fermuller, and Yiannis Aloimonos. “Affordance of Object Parts from Geometric Features.” In *Workshop on Vision meets Cognition, CVPR*, 2014.
- [MKP13] Vikash Mansinghka, Tejas D Kulkarni, Yura N Perov, and Josh Tenenbaum. “Approximate bayesian image interpretation using generative probabilistic graphics programs.” In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [MKS15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. “Human-level control through deep reinforcement learning.” *Nature*, **518**(7540):529–533, 2015.
- [MKS16] Yair Movshovitz-Attias, Takeo Kanade, and Yaser Sheikh. “How useful is photo-realistic rendering for visual learning?” In *European Conference on Computer Vision (ECCV)*, 2016.
- [MLB08] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. “Learning Object Affordances: From Sensory–Motor Coordination to Imitation.” *Transactions on Robotics (T-RO)*, **24**(1):15–26, 2008.
- [MLJ13] Ken Museth, Jeff Lait, John Johanson, Jeff Budsberg, Ron Henderson, Mihai Alden, Peter Cucka, David Hill, and Andrew Pearce. “OpenVDB: An open-source data structure and toolkit for high-resolution volumes.” In *ACM SIGGRAPH 2013 Courses*, 2013.

- [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.
- [MMB15] Aron Monszpart, Nicolas Mellado, Gabriel J Brostow, and Niloy J Mitra. “RAPter: rebuilding man-made scenes with regular arrangements of planes.” *ACM Transactions on Graphics (TOG)*, **34**(4):103, 2015.
- [MMR09] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. “Domain adaptation: Learning bounds and algorithms.” In *Annual Conference on Learning Theory (COLT)*, 2009.
- [MO04] Sara C Madeira and Arlindo L Oliveira. “Biclustering algorithms for biological data analysis: a survey.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**(1):24–45, 2004.
- [Mon92] Joe J Monaghan. “Smoothed particle hydrodynamics.” *Annual Review of Astronomy and Astrophysics*, **30**:543–574, 1992.
- [MP91] Ellen A McAfee and Dennis R Proffitt. “Understanding the surface orientation of liquids.” *Cognitive Psychology*, **23**(3):483–514, 1991.
- [MPM14] Oliver Mattausch, Daniele Panozzo, Claudio Mura, Olga Sorkine-Hornung, and Renato Pajarola. “Object detection and classification from large-scale cluttered indoor scans.” In *Computer Graphics Forum*, 2014.
- [MSB14] Yair Movshovitz-Attias, Yaser Sheikh, Vishnu Naresh Boddeti, and Zijun Wei. “3D Pose-by-Detection of Vehicles via Discriminatively Reduced Ensembles of Correlation Filters.” In *British Machine Vision Conference (BMVC)*, 2014.
- [MSB17] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. “Deepstack: Expert-level artificial intelligence in heads-up no-limit poker.” *Science*, **356**(6337):508–513, 2017.
- [MSF17] Roozbeh Mottaghi, Connor Schenck, Dieter Fox, and Ali Farhadi. “See the Glass Half Full: Reasoning about Liquid Containers, their Volume and Content.” In *International Conference on Computer Vision (ICCV)*, 2017.
- [MSL11] Paul Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. “Interactive furniture layout using interior design guidelines.” *ACM Transactions on Graphics (TOG)*, **30**(4), 2011.
- [MST10] Aleka McAdams, Eftychios Sifakis, and Joseph Teran. “A parallel multigrid Poisson solver for fluids simulation on large grids.” In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2010.

- [MTF15] Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. “Affordance Detection of Tool Parts from Geometric Features.” In *International Conference on Robotics and Automation (ICRA)*, 2015.
- [MVG10] Javier Marin, David Vázquez, David Gerónimo, and Antonio M López. “Learning appearance in virtual scenarios for pedestrian detection.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [MW14] Ralf Mayrhofer and Michael R. Waldmann. “Indicators of causal agency in physical interactions: The role of the prior context.” *Cognition*, **132**(3):485–490, 2014.
- [MZ93] Stéphane G Mallat and Zhifeng Zhang. “Matching pursuits with time-frequency dictionaries.” *IEEE Transactions on Signal Processing*, **41**(12):3397–3415, 1993.
- [MZS11] Aleka McAdams, Yongning Zhu, Andrew Selle, Mark Empey, Rasmus Tamstorf, Joseph Teran, and Eftychios Sifakis. “Efficient elasticity for character skinning with contact and collisions.” *ACM Transactions on Graphics (TOG)*, **30**(4):37, 2011.
- [Nat61] T. Natsoulas. “Principles of momentum and kinetic energy in the perception of causality.” *The American Journal of Psychology*, **74**(3):394–402, 1961.
- [NC36] Isaac Newton and John Colson. *The Method of Fluxions and Infinite Series; with Its Application to the Geometry of Curve-lines*. Henry Woodfall; and sold by John Nourse, 1736.
- [NDD14] Davide Nitti, Tinne De Laet, and Luc De Raedt. “Relational object tracking and learning.” In *International Conference on Robotics and Automation (ICRA)*, 2014.
- [NDI11] Richard A Newcombe, Andrew J Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. “KinectFusion: Real-time dense surface mapping and tracking.” In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.
- [Nel74] Katherine Nelson. “Concept, word, and sentence: Interrelations in acquisition and development.” *Psychological Review*, **81**(4):267, 1974.
- [NFJ02] Duc Quang Nguyen, Ronald Fedkiw, and Henrik Wann Jensen. “Physically based modeling and animation of fire.” *ACM Transactions on Graphics (TOG)*, **21**(3):721–728, 2002.
- [NUH07] S Nakano, G Ueno, and T Higuchi. “Merging particle filter for sequential data assimilation.” *Nonlinear Processes in Geophysics*, **14**(4):395–408, 2007.
- [NXS12] Liangliang Nan, Ke Xie, and Andrei Sharf. “A search-classify approach for cluttered indoor scene understanding.” *ACM Transactions on Graphics (TOG)*, **31**(6):137, 2012.

- [NZI13] Matthias Niesner, Michael Zollhofer, Shahram Izadi, and Marc Stamminger. “Real-time 3D reconstruction at scale using voxel hashing.” *ACM Transactions on Graphics (TOG)*, **32**(6):169, 2013.
- [OJL10] François Osiurak, Christophe Jarry, and Didier Le Gall. “Grasping the affordances, understanding the reasoning: toward a dialectical theory of human tool use.” *Psychological Review*, **117**(2):517, 2010.
- [OKA11] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. “Efficient model-based 3D tracking of hand articulations using Kinect.” In *British Machine Vision Conference (BMVC)*, 2011.
- [PDP13] Alexandros Paraschos, Christian Daniel, Jan R Peters, and Gerhard Neumann. “Probabilistic movement primitives.” In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [Pea00] J Pearl. *Causality: Models, reasoning and inference*. Cambridge University Press, 2000.
- [Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [PEK13] Alessandro Pieropan, Carl Henrik Ek, and Hedvig Kjellstrom. “Functional object descriptors for human activity modeling.” In *International Conference on Robotics and Automation (ICRA)*, 2013.
- [PH04] Matt Pharr and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2004.
- [PJA12] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. “Articulated people detection and pose estimation: Reshaping the future.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [PJW11] Leonid Pishchulin, Arjun Jain, Christian Wojek, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. “Learning people detection models from few training samples.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [PKQ15] Tu-Hoa Pham, Abderrahmane Kheddar, Ammar Qammar, and Antonis A Argyros. “Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [Pop10] Ronald Poppe. “A survey on vision-based human action recognition.” *Image and Vision Computing*, **28**(6):976–990, 2010.
- [PS94] Dimitris Papadias and Timos Sellis. “Qualitative representation of spatial knowledge in two-dimensional space.” *The VLDB Journal*, **3**(4):479–516, 1994.

- [PSA15] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. “Learning deep object detectors from 3D models.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [PSY13] Mingtao Pei, Zhangzhang Si, Benjamin Z Yao, and Song-Chun Zhu. “Learning and parsing video events with goal and intent prediction.” *Computer Vision and Image Understanding (CVIU)*, **117**(10):1369–1383, 2013.
- [PVB08] Jann Poppinga, Narunas Vaskevicius, Andreas Birk, and Kaustubh Pathak. “Fast plane detection and polygonalization in noisy 3D range images.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2008.
- [PZ15] Seyoung Park and Song-Chun Zhu. “Attributed grammars for joint estimation of human attributes, part and pose.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [QHW17] Siyuan Qi, Siyuan Huang, Ping Wei, and Song-Chun Zhu. “Predicting Human Activities Using Stochastic Grammar.” In *International Conference on Computer Vision (ICCV)*, 2017.
- [Qiu16] Weichao Qiu. *Generating Human Images and Ground Truth using Computer Graphics*. PhD thesis, UNIVERSITY OF CALIFORNIA, LOS ANGELES, 2016.
- [QSN16] Charles R Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. “Volumetric and Multi-View CNNs for Object Classification on 3D Data.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [QT08] Faisal Qureshi and Demetri Terzopoulos. “Smart camera networks in virtual reality.” *Proceedings of the IEEE*, **96**(10):1640–1656, 2008.
- [QY16] Weichao Qiu and Alan Yuille. “UnrealCV: Connecting Computer Vision to Unreal Engine.” *arXiv preprint arXiv:1609.01326*, 2016.
- [QZ18] Siyuan Qi and Song-Chun Zhu. “Intent-aware multi-agent reinforcement learning.” In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [QZH18] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. “Human-centric Indoor Scene Synthesis Using Stochastic Grammar.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [RA15] Siddharth S Rautaray and Anupam Agrawal. “Vision based hand gesture recognition for human computer interaction: a survey.” *Artificial Intelligence Review*, **43**(1):1–54, 2015.
- [RAA12] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. “A database for fine grained activity detection of cooking activities.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

- [RC11] Radu Bogdan Rusu and Steve Cousins. “3d is here: Point cloud library (pcl).” In *International Conference on Robotics and Automation (ICRA)*, 2011.
- [Ren12] Jochen Renz. “Implicit Constraints for Qualitative Spatial and Temporal Reasoning.” In *International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.
- [RFD05] M. E. Roser, J. A. Fugelsang, K. N. Dunbar, P. M. Corballis, and M. S. Gazzaniga. “Dissociating processes supporting causal perception and causal inference in the brain.” *Neuropsychology*, **19**(5):591–602, 2005.
- [RGT15] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. “Completing 3D object shape from one depth image.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [RHG15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks.” In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [RLB15] Javier Romero, Matthew Loper, and Michael J Black. “FlowCap: 2D human pose from optical flow.” In *German Conference on Pattern Recognition*, 2015.
- [RM15] Hossein Rahmani and Ajmal Mian. “Learning a non-linear knowledge transfer model for cross-view action recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [RM16] Hossein Rahmani and Ajmal Mian. “3d action recognition from novel viewpoints.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [RMC15] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” *arXiv preprint arXiv:1511.06434*, 2015.
- [RS16] Grégory Rogez and Cordelia Schmid. “MoCap-guided data augmentation for 3D pose estimation in the wild.” In *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [RT16] Anirban Roy and Sinisa Todorovic. “A Multi-scale CNN for Affordance Segmentation in RGB Images.” In *European Conference on Computer Vision (ECCV)*, 2016.
- [RVR16] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. “Playing for data: Ground truth from computer games.” In *European Conference on Computer Vision (ECCV)*, 2016.
- [RW72] Robert A Rescorla, Allan R Wagner, et al. “A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.” *Classical conditioning II: Current research and theory*, **2**:64–99, 1972.

- [San14] Adam N Sanborn. “Testing Bayesian and heuristic predictions of mass judgments of colliding objects.” *Frontiers in Psychology*, **5**, 2014.
- [SB90] Richard S Sutton and Andrew G Barto. “Time-derivative models of Pavlovian reinforcement.” In *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, 1990.
- [SB91] Louise Stark and Kevin Bowyer. “Achieving generalized object recognition through reasoning about association of function to structure.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **13**(10):1097–1104, 1991.
- [SB98] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.
- [SB99] Daniel L Schwartz and Tamara Black. “Inferences through imagined actions: knowing by simulated doing.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25**(1):116, 1999.
- [SB02] Gaurav Sharma and Raja Bala. *Digital color imaging handbook*. CRC press, 2002.
- [SBS16] Christian Schatzschneider, Gerd Bruder, and Frank Steinicke. “Who turned the clock? Effects of Manipulated Zeitgebers, Cognitive Load and Immersion on Time Estimation.” *IEEE Transactions on Visualization and Computer Graph (TVCG)*, **22**(4):1387–1395, 2016.
- [SBV13] K Smith, Peter Battaglia, and Edward Vul. “Consistent physics underlying ballistic motion prediction.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2013.
- [SCC14] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. “Visual tracking: An experimental survey.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **36**(7):1442–1468, 2014.
- [SCH14] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Niesner. “Scenegrok: Inferring action maps in 3D environments.” *ACM Transactions on Graphics (TOG)*, **33**(6):212, 2014.
- [SD88] David R Shanks and Anthony Dickinson. “Associative accounts of causality judgment.” In *Psychology of learning and motivation*, volume 21, pp. 229–261. Elsevier, 1988.
- [SGC17] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez. “Procedural Generation of Videos to Train Deep Action Recognition Networks.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [SGG08] Andreas J Schmid, Nicolas Gorges, Dirk Goger, and Heinz Worn. “Opening a door with a humanoid robot using multi-sensory tactile feedback.” In *International Conference on Robotics and Automation (ICRA)*, 2008.
- [SGR13] Hajar Sadeghi Sokeh, Stephen Gould, and Jochen Renz. “Efficient extraction and representation of spatial information from video data.” In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1076–1082, 2013.
- [SGR17] Tianmin Shu, Xiaofeng Gao, Michael S Ryoo, and Song-Chun Zhu. “Learning Social Affordance Grammar from Videos: Transferring Human Interactions to Human-Robot Interactions.” In *International Conference on Robotics and Automation (ICRA)*, 2017.
- [SGS10] Michael Stark, Michael Goesele, and Bernt Schiele. “Back to the Future: Learning Shape Models from 3D CAD Data.” In *British Machine Vision Conference (BMVC)*, 2010.
- [Sha91] David R Shanks. “Categorization by a connectionist network.” *Journal of Experimental Psychology*, **17**(3):433–443, 1991.
- [SHC17] Nishant Shukla, Yunzhong He, Frank Chen, and Song-Chun Zhu. “Learning Human Utility from Video Demonstrations for Deductive Planning in Robotics.” In *Conference on Robot Learning*, 2017.
- [SHK12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. “Indoor segmentation and support inference from RGBD images.” In *European Conference on Computer Vision (ECCV)*, 2012.
- [SHM14] Hao Su, Qixing Huang, Niloy J Mitra, Yangyan Li, and Leonidas Guibas. “Estimating image depth using shape collections.” *ACM Transactions on Graphics (TOG)*, **33**(4):37, 2014.
- [SHM16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. “Mastering the game of Go with deep neural networks and tree search.” *Nature*, **529**(7587):484–489, 2016.
- [SHS12] Alexey Stomakhin, Russell Howes, Craig Schroeder, and Joseph M Teran. “Energetically consistent invertible elasticity.” In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2012.
- [SLH12] Scott Satkin, Jason Lin, and Martial Hebert. “Data-driven scene understanding from 3D models.” In *British Machine Vision Conference (BMVC)*, 2012.
- [SLS14] Weiguang Si, Sung-Hee Lee, Eftychios Sifakis, and Demetri Terzopoulos. “Realistic biomechanical simulation and control of human swimming.” *ACM Transactions on Graphics (TOG)*, **34**(1):10, 2014.

- [SLZ08] Michael Stark, Philipp Lies, Michael Zillich, Jeremy Wyatt, and Bernt Schiele. “Functional object class detection based on learned affordance cues.” In *International Conference on Computer Vision Systems*, 2008.
- [SMG13] Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. “Reconciling intuitive physics and Newtonian mechanics for colliding objects.” *Psychological Review*, **120**(2):411, 2013.
- [SMT14] Gloria Sabbatini, Héctor Marín Manrique, Cinzia Trapanese, Aurora De Bortoli Vizioli, Josep Call, and Elisabetta Visalberghi. “Sequential use of rigid and pliable tools in tufted capuchin monkeys (*Sapajus* spp.).” *Animal Behaviour*, **87**:213–220, 2014.
- [SN02] Brian J. Scholl and Ken Nakayama. “Causal capture: Contextual effects on the perception of collision events.” *Psychological Science*, **13**(6):493–498, 2002.
- [SPF18] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, and Song-Chun Zhu. “Perception of Human Interaction Based on Motion Trajectories: From Aerial Videos to Decontextualized Animations.” *Topics in cognitive science*, **10**(1):225–241, 2018.
- [SPN05] Stefan Schaal, Jan Peters, Jun Nakanishi, and Auke Ijspeert. “Learning movement primitives.” In *The Eleventh International Symposium on Robotics Research*, 2005.
- [SPS06] Laurie R Santos, Heather M Pearson, Geertrui M Spaepen, Fritz Tsao, and Marc D Hauser. “Probing the limits of tool competence: experiments with two non-tool-using species (*Cercopithecus aethiops* and *Saguinus oedipus*).” *Animal Cognition*, **9**(2):94–109, 2006.
- [SPS12] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. “Unstructured human activity detection from rgbd images.” In *International Conference on Robotics and Automation (ICRA)*, 2012.
- [SQL15] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. “Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [SRF05] Andrew Selle, Nick Rasmussen, and Ronald Fedkiw. “A vortex particle method for smoke, water and explosions.” *ACM Transactions on Graphics (TOG)*, **24**(3):910–914, 2005.
- [SRZ16] Tianmin Shu, Michael S Ryoo, and Song-Chun Zhu. “Learning social affordance for human-robot interaction.” In *International Conference on Robotics and Automation (ICRA)*, 2016.
- [SS92] A. Schlottmann and D. R. Shanks. “Evidence for a distinction between judged and perceived causality.” *The Quarterly Journal of Experimental Psychology*, **44**(2):321–342, 1992.

- [SS14] Baochen Sun and Kate Saenko. “From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains.” In *British Machine Vision Conference (BMVC)*, 2014.
- [SS15] Brent Strickland and Brian J Scholl. “Visual perception involves event-type representations: The case of containment versus occlusion.” *Journal of Experimental Psychology: General*, **144**(3):570, 2015.
- [SSC13] Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. “A material point method for snow simulation.” *ACM Transactions on Graphics (TOG)*, **32**(4):102, 2013.
- [SSJ14] Alexey Stomakhin, Craig Schroeder, Chenfanfu Jiang, Lawrence Chai, Joseph Teran, and Andrew Selle. “Augmented MPM for phase-change and varied materials.” *ACM Transactions on Graphics (TOG)*, **33**(4):138, 2014.
- [SSK13] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. “Real-time human pose recognition in parts from single depth images.” *Communications of the ACM*, **56**(1):116–124, 2013.
- [SSS17] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. “Revisiting unreasonable effectiveness of data in deep learning era.” In *International Conference on Computer Vision (ICCV)*, 2017.
- [ST00] Brian J Scholl and Patrice D Tremoulet. “Perceptual causality and animacy.” *Trends in cognitive sciences*, **4**(8):299–309, 2000.
- [Sta99] Jos Stam. “Stable fluids.” In *Proceedings of the 26th annual conference on computer graphics and interactive techniques*, 1999.
- [Sto05] Alexander Stoytchev. “Behavior-grounded representation of tool affordances.” In *International Conference on Robotics and Automation (ICRA)*, 2005.
- [STZ17] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. “CERN: confidence-energy recurrent network for group activity recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [SV13] Kevin A Smith and Edward Vul. “Sources of uncertainty in intuitive physics.” *Topics in Cognitive Science*, **5**(1):185–199, 2013.
- [SVD03] Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. “Fast pose estimation with parameter-sensitive hashing.” In *International Conference on Computer Vision (ICCV)*, 2003.
- [SWB11] Robert W Shumaker, Kristina R Walkup, and Benjamin B Beck. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

- [SX13] Shuran Song and Jianxiong Xiao. “Tracking revisited using RGBD camera: Unified benchmark and baselines.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [SX14] Shuran Song and Jianxiong Xiao. “Sliding shapes for 3d object detection in depth images.” In *European Conference on Computer Vision (ECCV)*, 2014.
- [SXR15] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song-Chun Zhu. “Joint inference of groups, events and human roles in aerial videos.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4584, 2015.
- [Syl09] C Sylvain. *Robot programming by demonstration: A probabilistic approach*. EPFL Press, 2009.
- [SYZ17a] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. “Semantic Scene Completion From a Single Depth Image.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [SYZ17b] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. “Semantic Scene Completion from a Single Depth Image.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [SZK15] Shunsuke Saito, Zi-Ye Zhou, and Ladislav Kavan. “Computational bodybuilding: Anatomically-based modeling of human bodies.” *ACM Transactions on Graphics (TOG)*, **34**(4):41, 2015.
- [TA15] A Qammaz T.H.Pharm, A Kheddar and A.A Argyros. “Towards Force Sensing from Vision: Observing Hand-Object Interactions to Infer Manipulation Forces.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [TBH03] Joseph Teran, Sylvia Blemker, V Hing, and Ronald Fedkiw. “Finite volume methods for the simulation of skeletal muscle.” In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2003.
- [TF88] Demetri Terzopoulos and Kurt Fleischer. “Deformable models.” *The Visual Computer*, **4**(6):306–331, 1988.
- [THD15] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. “Simultaneous deep transfer across domains and tasks.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [THG16] Gilles Tagne, Patrick Hénaff, and Nicolas Gregori. “Measurement and analysis of physical parameters of the handshake between two persons according to simple social contexts.” In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016.
- [THT16] Duc Thanh Nguyen, Binh-Son Hua, Khoi Tran, Quang-Hieu Pham, and Sai-Kit Yeung. “A field model for repairing 3d shapes.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

- [TKG11] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. “How to grow a mind: Statistics, structure, and abstraction.” *Science*, **331**(6022):1279–1285, 2011.
- [TMB02] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. “Animation: can it facilitate?” *International Journal of Human-computer Studies*, **57**(4):247–262, 2002.
- [TPB87] Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. “Elastically deformable models.” *ACM Transactions on Graphics (TOG)*, **21**(4):205–214, 1987.
- [TPZ13] Kewei Tu, Maria Pavlovskaja, and Song-Chun Zhu. “Unsupervised structure learning of stochastic and-or grammars.” In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [TR95] Demetri Terzopoulos and Tamer F Rabe. “Animat vision: Active vision in artificial animals.” In *International Conference on Computer Vision (ICCV)*, 1995.
- [TT94] Xiaoyuan Tu and Demetri Terzopoulos. “Artificial fishes: Physics, locomotion, perception, behavior.” In *Computer Graphics (Proceedings of ACM SIGGRAPH)*, 1994.
- [USG14] Tomer Ullman, Andreas Stuhlmüller, Noah Goodman, and Josh Tenenbaum. “Learning physics from dynamical scenes.” In *Annual Conference of the Cognitive Science Society (CogSci)*, 2014.
- [Vae12] Krist Vaesen. “The cognitive bases of human tool use.” *Behavioral and Brain Sciences*, **35**(04):203–218, 2012.
- [Val07] Arne Valberg. *Light vision color*. John Wiley & Sons, 2007.
- [VDR16] Jacob Varley, Chad DeChant, Adam Richardson, Avinash Nair, Joaquín Ruales, and Peter Allen. “Shape Completion Enabled Robotic Grasping.” *arXiv preprint arXiv:1609.08546*, 2016.
- [VGS16] Hado Van Hasselt, Arthur Guez, and David Silver. “Deep Reinforcement Learning with Double Q-Learning.” In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.
- [VLM14] David Vázquez, Antonio M Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. “Virtual and real world adaptation for pedestrian detection.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **36**(4):797–809, 2014.
- [VRM17] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael Black, Ivan Laptev, and Cordelia Schmid. “Learning from Synthetic Humans.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [VV11] Karthik Mahesh Varadarajan and Markus Vincze. “Affordance based part recognition for grasping and manipulation.” In *Workshop on Autonomous Grasping, ICRA*, 2011.
- [VV12] Karthik Mahesh Varadarajan and Markus Vincze. “AfRob: The affordance network ontology for robots.” In *International Conference on Robotics and Automation (ICRA)*, 2012.
- [War63] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function.” *Journal of the American Statistical Association*, **58**(301):236–244, 1963.
- [WCK02] Alex AS Weir, Jackie Chappell, and Alex Kacelnik. “Shaping of hooks in New Caledonian crows.” *Science*, **297**(5583):981–981, 2002.
- [WDL09] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. “Feature hashing for large scale multitask learning.” In *International Conference on Machine Learning (ICML)*, 2009.
- [Wer94] Josie Wernecke et al. *The Inventor mentor: programming object-oriented 3D graphics with Open Inventor, release 2*, volume 1. Citeseer, 1994.
- [WG16] Xiaolong Wang and Abhinav Gupta. “Generative Image Modeling using Style and Structure Adversarial Networks.” *arXiv preprint arXiv:1603.05631*, 2016.
- [WGM99] Andrew Whiten, Jane Goodall, William C McGrew, Toshisada Nishida, Vernon Reynolds, Yukimaru Sugiyama, Caroline EG Tutin, Richard W Wrangham, and Christophe Boesch. “Cultures in chimpanzees.” *Nature*, **399**(6737):682–685, 1999.
- [WH92] Michael R Waldmann and Keith J Holyoak. “Predictive and diagnostic learning within causal models: asymmetries in cue competition.” *Journal of Experimental Psychology: General*, **121**(2):222–236, 1992.
- [WKF12] T Whelan, M Kaess, and MF Fallon. “Kintinuous: Spatially Extended KinectFusion.” In *Workshop on RGB-D: Advanced Reasoning with Depth Cameras, RSS*, 2012.
- [WKL13] Lawson LS Wong, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. “Manipulation-based active search for occluded objects.” In *International Conference on Robotics and Automation (ICRA)*, 2013.
- [WKZ18] Duotun Wang, James Kubricht, Yixin Zhu, Wei Liang, Song-Chun Zhu, Chenfanfu Jiang, and Hongjing Lu. “Spatially Perturbed Collision Sounds Attenuate Perceived Causality in 3D Launching Events.” In *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2018.
- [WLS15] Thomas Whelan, Stefan Leutenegger, Renato F Salas-Moreno, Ben Glocker, and Andrew J Davison. “ElasticFusion: Dense SLAM without a pose graph.” In *Robotics: Science and Systems (RSS)*, 2015.

- [WLW14] Jiang Wang, Zicheng Liu, and Ying Wu. “Learning actionlet ensemble for 3D human action recognition.” In *Human Action Recognition with Depth Cameras*, 2014.
- [WLY15] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. “Object tracking benchmark.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **37**(9):1834–1848, 2015.
- [WLY17] Hanqing Wang, Wei Liang, and Lap-Fai Yu. “Transferring Objects: Joint Inference of Container and Human Pose.” In *International Conference on Computer Vision (ICCV)*, 2017.
- [WMR17] Ziyu Wang, Josh S Merel, Scott E Reed, Nando de Freitas, Gregory Wayne, and Nicolas Heess. “Robust imitation of diverse behaviors.” In *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- [WMZ13] Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, and Jinxiang Chai. “Video-based hand manipulation capture through composite motion control.” *ACM Transactions on Graphics (TOG)*, **32**(4):43, 2013.
- [WNX14] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. “Cross-view action modeling, learning and recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [WRB11] Daniel Weinland, Remi Ronfard, and Edmond Boyer. “A survey of vision-based methods for action representation, segmentation and recognition.” *Computer Vision and Image Understanding (CVIU)*, **115**(2):224–241, 2011.
- [WSK15] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. “3d shapenets: A deep representation for volumetric shapes.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [WWR11] Christian Wojek, Stefan Walk, Stefan Roth, and Bernt Schiele. “Monocular 3D scene understanding with explicit occlusion reasoning.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [WZX17] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. “Visual Concepts and Compositional Voting.” *arXiv preprint arXiv:1711.04451*, 2017.
- [WZZ13] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. “Modeling 4d human-object interactions for event and object recognition.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [XLC07] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. “Multi-task learning for classification with dirichlet process priors.” *Journal of Machine Learning Research (JMLR)*, **8**(Jan):35–63, 2007.

- [XLE18] Xu Xie, Hangxin Liu, Mark Edmonds, Feng Gao, Siyuan Qi, Yixin Zhu, Brandon Rothrock, and Song-Chun Zhu. “Unsupervised Learning of Hierarchical Models for Hand-Object Interactions.” In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [XLZ16a] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. “Cooperative Training of Descriptor and Generator Networks.” *arXiv preprint arXiv:1609.09408*, 2016.
- [XLZ16b] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. “A theory of generative convnet.” In *International Conference on Machine Learning (ICML)*, 2016.
- [XOT13] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. “SUN3D: A database of big spaces reconstructed using SfM and object labels.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [XSX16] Caiming Xiong, Nishant Shukla, Wenlong Xiong, and Song-Chun Zhu. “Robot learning with a spatial, temporal, and causal and-or graph.” In *International Conference on Robotics and Automation (ICRA)*, 2016.
- [XTZ13] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. “Inferring ”Dark Matter” and ”Dark Energy” from Videos.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [YDY15] Lap-Fai Yu, Noah Duncan, and Sai-Kit Yeung. “Fill and transfer: a simple physics-based approach for containability reasoning.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [YIK16] Hashim Yasin, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall. “A Dual-Source Approach for 3D Pose Estimation from a Single Image.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [YLW09] Junsong Yuan, Zicheng Liu, and Ying Wu. “Discriminative subvolume search for efficient action detection.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [YNL14] Benjamin Z Yao, Bruce X Nie, Zicheng Liu, and Song-Chun Zhu. “Animated pose templates for modeling and detecting human actions.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **36**(3):436–452, 2014.
- [YQK17] Tian Ye, Siyuan Qi, James Kubricht, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. “The martian: Examining human physical judgments across virtual gravity fields.” *IEEE Transactions on Visualization and Computer Graph (TVCG)*, **23**(4):1399–1408, 2017.
- [YTS05] Kai Yu, Volker Tresp, and Anton Schwaighofer. “Learning Gaussian processes from multiple tasks.” In *International Conference on Machine Learning (ICML)*, 2005.

- [YWH09] Ming Yang, Ying Wu, and Gang Hua. “Context-aware visual tracking.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **31**(7):1195–1209, 2009.
- [YYL15] Ju Hong Yoon, Ming-Hsuan Yang, Jongwoo Lim, and Kuk-Jin Yoon. “Bayesian multi-object tracking using motion context from multiple objects.” In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.
- [YYT11] Lap Fai Yu, Sai Kit Yeung, Chi Keung Tang, Demetri Terzopoulos, Tony F Chan, and Stanley J Osher. “Make it home: automatic optimization of furniture arrangement.” *ACM Transactions on Graphics (TOG)*, **30**(4), 2011.
- [YYT16] Lap-Fai Yu, Sai-Kit Yeung, and Demetri Terzopoulos. “The clutterpalette: An interactive tool for detailing indoor scenes.” *IEEE Transactions on Visualization and Computer Graph (TVCG)*, **22**(2):1138–1148, 2016.
- [YYW12] Yi-Ting Yeh, Lingfeng Yang, Matthew Watson, Noah D Goodman, and Pat Hanrahan. “Synthesizing open worlds with constraints using locally annealed reversible jump mcmc.” *ACM Transactions on Graphics (TOG)*, **31**(4), 2012.
- [YZ09] Benjamin Yao and Song-Chun Zhu. “Learning deformable action templates from cluttered videos.” In *International Conference on Computer Vision (ICCV)*, 2009.
- [ZB05] Yongning Zhu and Robert Bridson. “Animating sand as a fluid.” *ACM Transactions on Graphics (TOG)*, **24**(3):965–972, 2005.
- [ZBG15] Xinxin Zhang, Robert Bridson, and Chen Greif. “Restoring the missing vorticity in advection-projection fluid solvers.” *ACM Transactions on Graphics (TOG)*, **34**(4):52, 2015.
- [ZFF14] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. “Reasoning about Object Affordances in a Knowledge Base Representation.” In *European Conference on Computer Vision (ECCV)*, 2014.
- [ZJZ16] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. “Inferring Forces and Learning Human Utilities From Videos.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [ZKA16] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. “Learning Dense Correspondence via 3D-guided Cycle Consistency.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [ZLN08] Li Zhang, Yuan Li, and Ramakant Nevatia. “Global data association for multi-object tracking using network flows.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [ZM07] Song-Chun Zhu, David Mumford, et al. “A stochastic grammar of images.” *Foundations and Trends® in Computer Graphics and Vision*, **2**(4):259–362, 2007.

- [ZMK13] Qian-Yi Zhou, Steven Miller, and Vladlen Koltun. “Elastic fragments for dense scene reconstruction.” In *International Conference on Computer Vision (ICCV)*, 2013.
- [ZMK17] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. “Target-driven visual navigation in indoor scenes using deep reinforcement learning.” In *International Conference on Robotics and Automation (ICRA)*, 2017.
- [ZRG09] Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. “Planning-based prediction for pedestrians.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2009.
- [ZSY17] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. “Physically-Based Rendering for Indoor Scene Understanding Using Convolutional Neural Networks.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [ZZ13] Yibiao Zhao and Song-Chun Zhu. “Scene parsing by integrating function, geometry and appearance models.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [ZZC15] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. “Understanding tools: Task-oriented object modeling, learning and recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [ZZL16] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. “Sparseness meets deepness: 3D human pose estimation from monocular video.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [ZZM13] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. “Robust real-time physics-based motion control for human grasping.” *ACM Transactions on Graphics (TOG)*, **32**(6):207, 2013.
- [ZZY13] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. “Beyond point clouds: Scene understanding by reasoning geometry and physics.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [ZZY14] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. “Detecting potential falling objects by inferring human action and natural disturbance.” In *International Conference on Robotics and Automation (ICRA)*, 2014.
- [ZZY15] Bo Zheng, Yibiao Zhao, Joey Yu, Katsushi Ikeuchi, and Song-Chun Zhu. “Scene Understanding by Reasoning Stability and Safety.” *International Journal of Computer Vision (IJCV)*, **112**(2):221–238, 2015.

- [ZZZ15] Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. “Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.