UNIVERSITY OF CALIFORNIA

Los Angeles

A Cognition Platform for Joint Inference of 3D Geometry, Object States, and Human Belief

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

by

Tao Yuan

2019

ABSTRACT OF THE DISSERTATION

A Cognition Platform for Joint Inference of
3D Geometry, Object States, and Human Belief

by

Tao Yuan
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2019
Professor Song-Chun Zhu, Chair

Humans can extract rich information from visual scenes, such as the 3D locations of objects and humans, the actions of humans, the states of objects, the belief of humans. Although various state-of-the-art algorithms can achieve good results for solving individual tasks, building a system to jointly infer these different tasks for scene understanding is still an underexplored area. Most of these tasks are not independent with each other, and humans can jointly infer hidden information with their commonsense knowledge among these tasks. In this dissertation, we propose a spatio-temporal framework to jointly infer and optimize multiple tasks across different times and views with a unified explicit probabilistic graphical representation.

This dissertation contains four main parts. 1) we describe the system overview, the data flow in the system, and engineering efforts to make the system scalable under different scenarios. 2) we propose an algorithm for holistic 3D scene parsing and human pose estimation with human-object interaction and physical commonsense. Human-object interaction can model the fine-grained relations between agents and objects, and physical commonsense can model the physical plausibility of the reconstructed scene. 3) we introduce a joint parsing framework that integrates view-centric proposals into scene-centric parse graphs that represent a coherent scene-centric understanding of cross-view scenes. 4) we present a joint inference algorithm to understanding object states, robot knowledge, and human beliefs under multi-view settings by maintaining three types of parse graphs. The algorithm can be applied to the cross-view small object tracking problem and some false-belief problems. Experiments show that our joint inference framework can achieve better results than individual algorithms.

The dissertation of Tao Yuan is approved.

Ying Nian Wu

Demetri Terzopoulos

Tao Gao

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2019

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Prof. Song-Chun Zhu, for inspiring me with his unique and deep insight in Computer Vision, Artificial Intelligence, and Statistical Learning during my doctoral research. He is an excellent explorer and always encourages us to challenge new problems and develop a unified framework. I would also like to thank other committee members: Prof. Tao Gao, Prof. Demetri Terzopoulos, and Prof. Ying Nian Wu for their guidance and support.

I always enjoy the time learning, discussing, working, and traveling with my creative labmates at VCLA. Specifically, I would like to thank Hang Qi, Yixin Zhu, Siyuan Huang, Yixin Chen, Hangxin Liu, Lifeng Fan, Zilong Zheng, Yuanlu Xu, Siyuan Qi, and Tianfu Wu for their collaboration of many topics related to this dissertation. I also feel very fortunate to work with my friends: Baoxiong Jia, Qing Li, Tengyu Liu, Zeyu Zhang, Chi Zhang, Weixin Li, Yang Liu, Tianmin Shu, Nishant Shukla, Mitchell Hill, Ruiqi Gao, Arjun Akula, Feng Shi, Mark Edmonds, Luyao Yuan, Xiaofeng Gao, Feng Gao, Xu Xie, Erik Nimp, and many others.

Most importantly, I would like to thank my parents for their love and support. None of this could have happened without them.

VITA

2015–2019	Graduate Student Researcher, Department of Statistics, UCLA
2011–2015	B.E. (Computer Science), Shanghai Jiao Tong University
2014	Algorithm Engineer Intern, Alibaba Group

PUBLICATIONS

Understanding False-Belief by Joint Inference of Object States, Robot Knowledge, and Human (False-)Beliefs, Tao Yuan, Hangxin Liu, Lifeng Fan, Zilong Zheng, Tao Gao, Yixin Zhu, and Song-Chun Zhu, *under review by ICRA*, 2019

3D Object Detection from a Single RGB Image via Perspective Points, Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu, *NeurlPS*, 2019

Holistic++ Scene Understanding: Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-Object Interaction and Physical Commonsense, Yixin Chen*, Siyuan Huang*, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu, *ICCV*, 2019

Scene-centric Joint Parsing of Cross-view Videos, Hang Qi*, Yuanlu Xu*, Tao Yuan*, Tianfu Wu, and Song-Chun Zhu, *AAAI*, 2018

CHAPTER 1

Introduction and System Overview

1.1 Motivation

During the past decades, remarkable progress has been made in many vision tasks, e.g. image classification, object detection, pose estimation. However, in most works, these vision tasks are studied independently without considering the information provided by each other. And more and more complex AI tasks, such as visual question answering, and social norm understanding, require the results from various vision tasks and multiple inputs from different fields of views. According to the observation that there are consistent correlations accross different cognition tasks and input sources, and contraints provided by commonsense. We propose a cognition platform for joint inference of 3D geometry, object states, and human belief.

Our humans, even babies, can understand and extract rich information from complex scenes. For example, given a picture of a girl is using a computer, we can still infer there is a mouse in her hand, even we cannot see it, by applying our knowledge of the human-object interaction. And we can also infer the 3D location of the girl is on the chair by using our physical commonsense that the chair can support the girl. Inspired by this, our system first focuses on a new 3D holistic++ problem includes two main tasks: 1) 3D scene parsing and reconstruction, 2) 3D human pose estimation. We jointly tackle these two tasks by exploiting the human-object interaction and physical commonsense constraints.

Multiple-camera systems, such as security systems, are widely applied in our daily lives. It is essential to have a framework to resolve ambiguity and establish cross-reference among information from different views. We tackle the problem with a spatio-temporal joint parsing framework that integrates information from each view into scene-centric parse graphs to present the coherent understanding of cross-view videos by maintaining some correlations and constraints in commonsense.

Once we have the unified explicit graphical model to track object states, our system can further infer human beliefs and discover false-beliefs from segments of the graphs. With the proposed joint inference algorithm, our system can achieve good results in recognizing some human belief tasks.

1.2 Related Work

Our work is related to the following areas in computer vision and artificial intelligence.

AI Cloud Platforms Various companies provide computer vision services on their cloud AI platforms, such as Google Cloud AI Platform, Microsoft Azure AI Platform, and Face++ AI Open Platform. These platforms can provide cognitive services such as object detection, optical character recognition, and face detection. However, most of these modules are run separately. In our system, different modules can be inferred and optimized with the joint inference algorithms. Inputs on most cloud AI platforms are single images. On our system, inputs could be video streams from single-view or multi-view, which can utilize both the spatial and temporal information.

Multi-view video analytics Typical multi-view visual analytics tasks include object detection [LS10a, UB11a], cross-view tracking [BFT11a, LPR12, XLL16, XLQ17], action recognition [WNX14], person re-identification [XLZ13, XMH14] and 3D reconstruction [HWR13a]. While heuristics such as appearances and motion consistency constraints have been used to regularize the solution space, these methods focus on a specific multi-view vision task whereas we aim to propose a general framework to jointly resolve a wide variety of tasks.

Visual Cognitive Reasoning Related work includes recovering incomplete trajectories [LZZ18], predicting human activities [QHW17], learning utility and affordance [ZJZ16, ZZZ15], inferring human intention and attention [FCW18, WLS18], *etc.* As to understanding (false-)belief, despite many psychological experiments and theoretical analysis [CT99, WAS18, BBP16], very few attempts have been made to solve (false-)belief in man-made scenes with visual input; handcrafted constraints are usually required for specific problems in prior work. In contrast, our work utilizes a unified representation across different domains with heterogeneous information to model human mental states.

1.3 System Architecture

The system includes three main components: the input module, the joint inference module, and the visualization module.

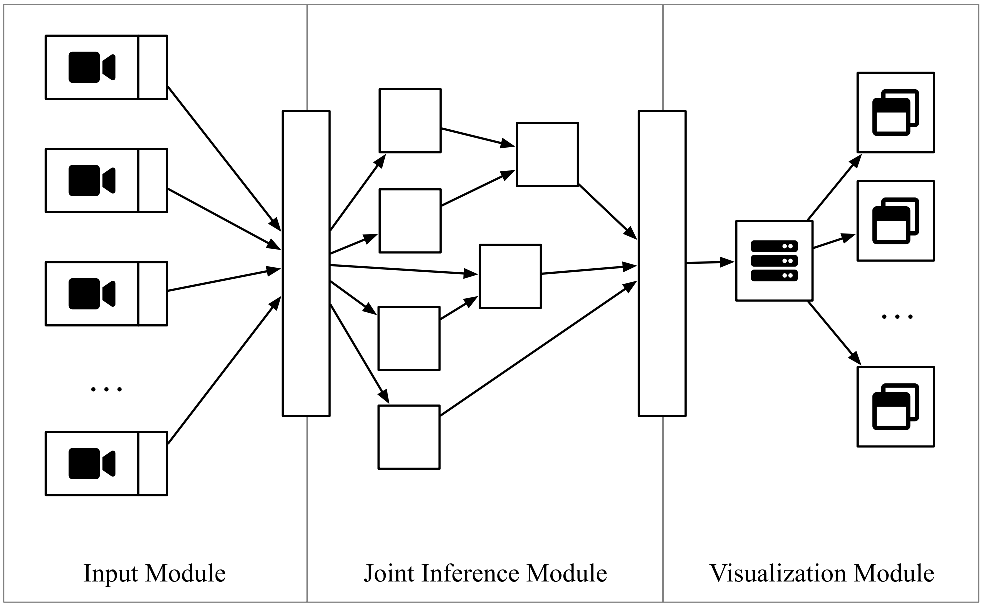


Figure 1.1: System Architecture

Input Module: The system can receive frame sequence input from different types of sources, such as webcams via USB interfaces, remote surveillance cameras, cameras from robots, and offline videos. Our system pre-processes frame sequence from these sources and encodes them into a unified binary format data stream. The frame size, frame rate, and frame quality can be set in the configuration files.

Joint Inference Module: the most important part of the system is the joint inference engine. It receives encoded image streams from input devices and feeds them into various computer vision modules. The computer vision modules are organized in a hierarchical structure, because many high-level modules may depend on other mid-level modules. We will introduce some modules in the following chapters. The module dependencies are handled with a topological sorting algorithm. All modules are run in separate containers with several benefits:

- Isolation: The environments of each container could be different. Developers and researchers can build and test their own modules without worry about breaking the environments of other modules.
- Compatibility: Once the environment is configured, the module can be run on every other device, whatever a personal computer or a cloud server. This compatibility enables rapid deployment and continuous deployment and testing.
- Scalability: The modules can be swarm with many copies. The system automatically does the load balance according to compute resources and requests.

Visualization Module: we employ a set of web-based tools to visualize results from the joint inference module in realtime. The web server receives serialized results and then send them to the client-side web application via WebSocket. The javascript based web application draws 2D results, such as 2D bounding boxes, on the original frames in SVG containers and renders 3D results, such as 3D poses, with WebGL. We also provide APIs to let modules render 2D results themselves on the server.

1.4 Outline

The remainder of the dissertation is organized as follows:

In Chapter 2, we focus on the 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. The HOI priors and physical commonsense constraints are helpful to infer the spatial relations between human and the scene. The indoor scenes can be represented by parse graphs, and the system can iteratively optimize these parse graphs by MCMC sampling.

In Chapter 3, we propose a scene-centric joint parsing algorithm for understanding cross-view videos. We find that overlapping fields of views embed rich appearance and geometry correlations, and individual vision tasks are governed by consistency constraints available in commonsense knowledge. Our system represents such correlations and constraints explicitly and generates semantic scene-centric parse graphs.

In Chapter 4, we present the joint inference algorithm of object states, robot knowledge, and human beliefs to understand false-belief. The system first infers single-view parse graphs from each individual view and then fuses these parse graphs into a joint parse graph. Belief parse graphs can be extracted from the joint parse graph and provide the capability of recognizing human false-belief.

CHAPTER 2

Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-Object Interaction and Physical Commonsense

2.1 Introduction

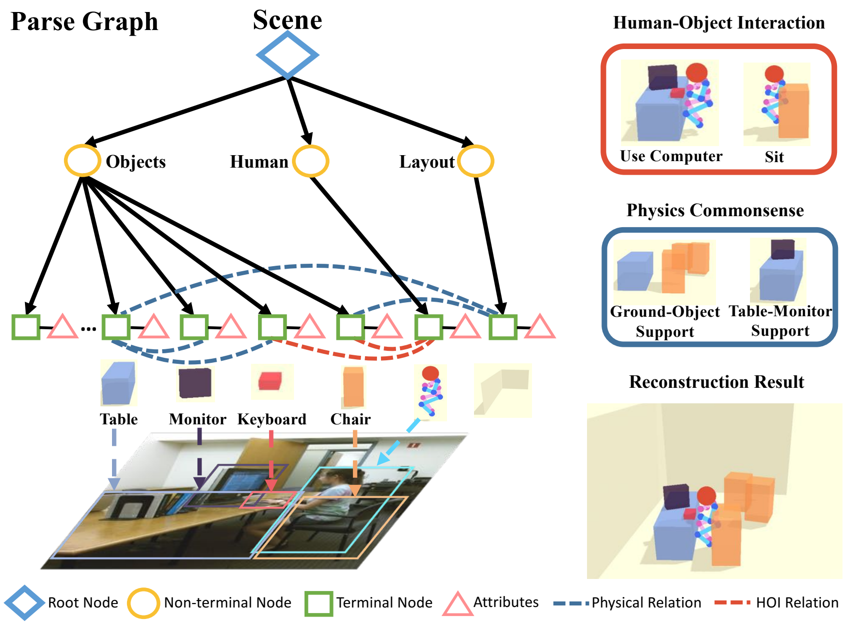


Figure 2.1: *holistic^{++} scene understanding* task requires to jointly recover a parse graph that represents the scene, including human poses, objects, camera pose, and room layout, all in 3D. Reasoning human-object interaction (HOI) helps reconstruct the detailed spatial relations between humans and objects. Physical commonsense (*e.g.*, physical property, plausibility, and stability) further refines relations and improves predictions.

Humans, even young infants, are adept at perceiving and understanding complex indoor scenes. Such an amazing vision system not only relies on the data-driven pattern recognition but also roots from the visual reasoning system, known as the core knowledge [SK07] that facilitates the 3D holistic scene understanding tasks. Consider a typical indoor scene shown in Figure 2.1 where a person sits in an office. We can effortlessly extract rich knowledge from the static scene, including 3D room layout, 3D position of all the objects and agents, and correct human-object interaction (HOI) relations in a physically plausible manner. In fact, psychology studies have established that even infants employ at least two constraints—HOI and physical commonsense—in perceiving occlusions [THK87, KS83], tracking small objects even if contained by other objects [FC03], realizing object permanence [BSW85], recognizing rational HOI [Woo99, SCS13], understanding intuitive physical [GBK02, Nee97, Bai04], and using exploratory play to understand the environment [SF15]. All these evidence in literature call for a treatment to integrate HOI and physical commonsense with a modern computer vision system.

In contrast, few attempts have been made to achieve this goal. This challenge is difficult partially due to the fact that the algorithm has to jointly accomplish both 3D holistic scene understanding task and the 3D human pose estimation task in a physically plausible fashion. Since this task is beyond the scope of holistic scene understanding in the literature, we define this comprehensive task as *holistic^{++} scene understanding*—to simultaneously estimate human pose, objects, room layout, and camera pose, all in 3D.

Existing work only individually solves the task of 3D holistic scene understanding [HQZ18, ZLH17, BRG16, SYZ17] or 3D human pose estimation [ZWM17, RKS12, FXW18] from a single RGB image, although one can achieve an impressive performance in a single task by training with an enormous amount of annotated data. We, however, argue that these two tasks are intertwined tightly since the indoor scenes are invented and constructed by human designs to support the daily activities, generating affordance for various tasks and human activities [Gib79].

To solve the *holistic^{++} scene understanding* task, we attempt to address four fundamental challenges:

1. How to utilize the coupled nature of human pose estimation and holistic scene under-

standing, and make them benefit each other? How to reconstruct the scene with complex human activities and interactions?

2. How to constrain the solution space of the 3D estimations from a single 2D image?
3. How to make a physically plausible and stable estimation for complex scenes with human agents and objects?
4. How to improve the generalization ability to achieve a more robust reconstruction across different datasets?

To address the first two challenges, we take a novel step to incorporate **HOI** as constraints for **joint parsing** of both 3D human pose and 3D scene. The integration of HOI is inspired by crucial observations of human 3D scene perception, which are challenging for existing systems. Take Figure 2.1 as an example, humans are able to not only infer the relative position and orientation between the girl and chair but also impose a constraint by recognizing the girl is sitting in the chair. Similarly, such constraint can help to recover the small objects (*e.g.*, recognizing keyboard by detecting the girl is using a computer in Figure 2.1). By learning HOI priors and using the inferred HOI as visual cues to adjust the fine-grained spatial relations between human and scene (objects and room layout), the geometric ambiguity (3D estimation solution space) in the single-view reconstruction would be largely eased, and the reconstruction performances of both tasks would be improved.

To address the third challenge, we incorporate **physical commonsense** into the proposed method. Specifically, the proposed method reasons about the physical relations (*e.g.*, support relation) and penalizes the physical violations to predict a physically plausible and stable 3D scene. The HOI and physical commonsense serve as **general prior** knowledge across different datasets, thus help address the fourth issue.

To jointly parse 3D human pose and 3D scene, we represent the configuration of an indoor scene by a parse graph shown in Figure 2.1, which consists of a parse tree with hierarchical structure and a Markov random field (MRF) over the terminal nodes, capturing the rich contextual relations among human, objects, and room layout. The optimal parse graph to reconstruct both the 3D scene and human poses is achieved by a MAP estimation, where the prior characterizes the prior distribution of the contextual HOI and physical relations among the nodes. The likelihood measures the similarity between (i) the detection results directly from 2D object and pose detector, and (ii) the 2D results projected from the 3D parsing results. The parse graph can be iteratively optimized by sampling an MCMC with simulated annealing based on posterior probability. The joint optimization relies less on a specific training dataset since it benefits from the prior of HOI and physical commonsense which is almost invariant across environments and datasets, and other knowledge learned from well-defined vision task (*e.g.*, 3D pose estimation, scene reconstruction), improving the generalization ability significantly across different datasets compared with purely data-driven methods.

Experimental results on PiGraphs [SCH16], Watch-n-Patch [WZS15], and SUN RGB-D [SLX15] demonstrate that the proposed method outperforms state-of-the-art methods for both 3D scene reconstruction and 3D pose estimation, demonstrating an improved generalization ability across different datasets comparing with pure data-driven methods. Moreover, the ablative analysis shows that the HOI prior improves the reconstruction and the physical common sense helps to make physically plausible predictions.

This work makes four major contributions:

1. We propose a new *holistic*$^{++}$ *scene understanding* task with a computational framework to jointly infer human poses, objects, room layout, and camera pose, all in 3D.
2. We integrate HOI into the proposed algorithm to bridge the human pose estimation and the scene reconstruction, reducing the geometric ambiguity (solution space) of the single-view reconstruction.
3. We incorporate physical commonsense, which helps to predict physically plausible scenes and improve the 3D localization of both humans and objects.
4. We demonstrate the joint inference improves the performance of each sub-module and achieves better generalization ability across various indoor scene datasets compared with purely data-driven methods.

2.1.1 Related Work

Single-view 3D Human Pose Estimation. Previous methods on 3D pose estimation can be divided into two streams: (i) directly learning 3D pose from a 2D image [SRA12, LC14], and (ii) cascaded frameworks that first perform 2D pose estimation and then reconstruct 3D pose from the estimated 2D joints [ZWM17, MSS17, RKS12, WXL16, CLO16, TRA17]. Although these researches have produced impressive results in scenarios with relatively clean background, the problem of estimating the 3D pose in a typical indoor scene with arbitrary cluttered objects has rarely been discussed. Recently, Zanfir *et al.* [ZMS18] adopts constraints of ground plane support and volume occupancy by multiple people, but the detailed relations between human and scene (objects and layout) are still missing. In contrast, the proposed model not only estimates the 3D poses of multiple people with an absolute scale but also models the physical relations between humans and 3D scenes.

Single-view 3D Scene Reconstruction. Single-view 3D scene reconstruction has three main approaches: (i) Predict room layouts by extracting geometric features to rank 3D cuboids proposals [ZLH17, SYZ17, ISS17]. (ii) Align object proposals to RGB or depth image by treating objects as geometric primitives or CAD models [BRG16, SX14, ZLX14]. (iii) Joint estimation of the room layout and 3D objects with contexts [SYZ17, ZZ13, CCP13, ZSY17, ZLH17]. A more recent work by Huang *et al.* [HQZ18] models the hierarchical structure, latent human context, physical constraints, and jointly optimizes in an analysis-by-synthesis fashion. Although human context and functionality were taken into account, indoor scene reconstruction with human poses and HOI remains untouched.

Human-Object Interaction. Reasoning fine-grained human interactions with objects are essential for a more holistic indoor scene understanding as it provides important cues for human activities and physical interactions. There have been a great deal of work in robotics and computer vision that exploits human-object relations in event, object and scene modeling, but most work focuses on human-object relation detection in image space [CLL18, QWJ18, ML16, KRK11], probabilistic modeling from multiple data sources [WZZ13, SCH14, GKD09], and snapshots generation or scene synthesis [SCH16, MLZ16, QZH18, JQZ18].

Different from all previous work, we use the learned 3D HOI priors to refine the relative spatial relations between human and scene, enabling a top-down prediction of interacted objects.

Physical Commonsense The ability to infer hidden physical properties is a well established human cognitive ability [KHL17]. By exploiting the underlying physical properties of scenes and objects, recent efforts have demonstrated the capability of estimating both current and future dynamics of static scenes [WYL15, MBR16] and objects [ZZZ15], understanding the support relationships and stability of objects [ZZY13], volumetric and occlusion reasoning [SHK12, ZZY15], inferring the hidden force [ZJZ16], and reconstructing the 3D scene [HQX18, DLB18] and 3D pose [ZMS18]. In addition to the physical properties and support relations among objects adopted in previous methods, we further model the physical relations (i) between human and objects, and (ii) between human and room layout, resulting in a physically plausible and stable scene.

2.2 Representation of the Scene

We represent the configuration of an indoor scene by a parse graph $pg = (pt, E)$ as shown in Figure 2.1. It combines a parse tree pt and contextual relations E among the leaf nodes. Here $pt = (V, R)$ and we denote $V = V_r \cup V_m \cup V_t$ the vertex set and R the decomposing rules. The tree has three levels. The first level is the root node V_r that represents the scene, and the second level V_m has three nodes (objects, human, and room layout). The third level (terminal nodes V_t) contains child nodes of the second level nodes, representing the detected instances of the parent node in this scene. $E \subset V_t \times V_t$ is the set of contextual relations among the terminal nodes, represented by horizontal links.

Terminal Nodes $\mathbf{V_t}$ in pg can be further decomposed as $V_t = V_{\text{layout}} \cup V_{\text{object}} \cup V_{\text{human}}$:

- The room layout $v \in V_{\text{layout}}$ is represented by a 3D bounding box $X^L \in \mathbb{R}^{3\times 8}$ in the world coordinate. The 3D bounding box is parametrized by the node's attributes, including its 3D size $S^L \in \mathbb{R}^3$, center $C^L \in \mathbb{R}^3$, and orientation $Rot(\theta^L) \in \mathbb{R}^{3\times 3}$. See the supplementary for the parametrization of the 3D bounding box.

- Each 3D object $v \in V_{\text{object}}$ is represented by a 3D bounding box with its semantic label. We keep the same parameterization of the 3D bounding box as the one for room layout.
- Each human $v \in V_{\text{human}}$ is represented by 17 3D joints $X^H \in \mathbb{R}^{3\times 17}$ with their action labels. These 3D joints are parametrized by the pose scale $S^H \in \mathbb{R}$, pose center (*i.e.*, hip) $C^H \in \mathbb{R}^3$, local joint position $Rel^H \in \mathbb{R}^{3\times 17}$, and pose orientation $Rot(\theta^H) \in \mathbb{R}^{3\times 3}$. Each person is also attributed by a concurrent action label a, which is a multi-hot vector representing the current actions of this person: one can "sit" and "drink", or "walk" and "make phone call" at the same time.

Contextual Relations E contains three types of relations in the scene $E = \{E_s, E_c, E_{hoi}\}$. Specifically:

- E_s and E_c denote support relation and physical collision, respectively. These two relations penalize the physical violations among objects, between objects and layout, and between human and layout, resulting in a physically plausible and stable prediction.
- E_{hoi} models HOI and gives us more constraints to reconstruct 3D from 2D. For instance, if a person is detected as sitting on the chair, we can constrain the relative 3D positions between this person and chair using a pre-learned spatial relation of "sitting".

2.3 Probabilistic Formulation

The parse graph pg is a comprehensive interpretation of the observed image I. The goal of the holistic^{++} scene understanding is to infer the optimal parse graph pg^* given I by a MAP estimation:

$$\begin{aligned} pg^* &= \arg\max_{pg} p(pg|I) = \arg\max_{pg} p(pg)\cdot p(I|pg) \\ &= \arg\max_{pg} \frac{1}{Z}\exp\{-\mathcal{E}_{phy}(pg) - \mathcal{E}_{hoi}(pg) - \mathcal{E}(I|pg)\}, \end{aligned} \tag{2.1}$$

We model the joint distribution by a Gibbs distribution, where the prior probability of parse graph can be decomposed into physical prior and HOI prior.

Physical Prior $\mathcal{E}_{phy}(pg)$ represents physical commonsense in a 3D scene. We consider two types of physical relations among the terminal nodes: support relation E_s and collision relation E_c. Therefore, the energy of physical prior is defined as $\mathcal{E}_{phy}(pg) = \lambda_s \mathcal{E}_s(pg) +$

$\lambda_c \mathcal{E}_c(pg)$, where λ_s and λ_c are balancing factors. Specifically:

- *Support Relation* $\mathcal{E}_s(pg)$ defines the energy between the supported object/human and the supporting object/layout:

$$\mathcal{E}_s(pg) = \sum_{(v_i,v_j)\in E_s} \mathcal{E}_o(v_i, v_j) + \mathcal{E}_{height}(v_i, v_j), \tag{2.2}$$

where $\mathcal{E}_o(v_i, v_j) = 1 - \text{area}(v_i \cap v_j)/\text{area}(v_i)$ is the overlapping ratio in the xy-plane, and $\mathcal{E}_{height}(v_i, v_j)$ is the absolute height difference between the lower surface of the supported object v_i and the upper surface of the supporting object v_j. We define

$$\mathcal{E}_o(v_i, v_j) = \mathcal{E}_{height}(v_i, v_j) = 0$$

if the supporting object is floor or wall.

- *Physical Collision* $\mathcal{E}_c(pg)$ denotes the physical violations. We penalize the intersection among human, objects, and room layout except the objects in HOI and objects that could be a container. The potential function is defined as:

$$\mathcal{E}_c(pg) = \sum_{v\in(V_{object}\cup V_{human})} \mathcal{C}(v, V_{layout}) + \sum_{\substack{v_i\in V_{object}\\ v_j\in V_{human}\\ (v_i,v_j)\notin E_{hoi}}} \mathcal{C}(v_i, v_j) + \sum_{\substack{v_i,v_j\in V_{object}\\ v_i,v_j\notin V_{container}}} \mathcal{C}(v_i, v_j), \tag{2.3}$$

where $\mathcal{C}()$ denotes the volume of intersection between entities. $V_{container}$ denotes the objects that can be a container, such as a cabinet, desk, and drawer.

Human-object Interaction Prior $\mathcal{E}_{hoi}(pg)$ is defined on the interactions between human and objects:

$$\mathcal{E}_{hoi}(pg) = \sum_{(v_i,v_j)\in E_{hoi}} \mathcal{K}(v_i, v_j, a_{v_j}), \tag{2.4}$$

where $v_i \in V_{object}, v_j \in V_{human}$, and $\mathcal{K}$ is an HOI function that evaluates the interaction between an object and a human given the action label a:

$$\mathcal{K}(v_i, v_j, a_{v_j}) = -\log l(v_i, v_j | a_{v_j}), \tag{2.5}$$

where $l(v_i, v_j | a_{v_j})$ is the likelihood of the relative position between node v_i and v_j given an action label a, and λ_a the balancing factor. We formulate the action detection as a *multi-label classification*; see subsection 2.5.3 for details. The likelihood $l(\cdot)$ models the distance between key joints and the center of the object; *e.g.*, for "sitting", it models the relative spatial relation between the hip and the center of a chair. The likelihood can be learned from 3D HOI datasets with a multivariate Gaussian distribution $(\Delta x, \Delta y, \Delta z) \sim \mathcal{N}_3(\mu, \Sigma)$, where $\Delta x, \Delta y$, and Δz are the relative distances in the directions of three axes.

Likelihood $\mathcal{E}(I|pg)$ characterizes the consistency between the observed 2D image and the inferred 3D result. The projected 2D object bounding boxes and human poses can be computed by projecting the inferred 3D objects and human poses onto a 2D image plane. The likelihood is obtained by comparing the directly detected 2D bounding boxes and human poses with projected ones from inferred 3D results:

$$\mathcal{E}(I|pg) = \sum_{v \in V_{object}} \lambda_o \cdot \mathcal{D}_o(B(v), B'(v)) + \sum_{v \in V_{human}} \lambda_h \cdot \mathcal{D}_h(Po(v), Po'(v)), \tag{2.6}$$

where $B()$ and $B'()$ are the bounding boxes of detected and projected 2D objects, $Po()$ and $Po'()$ the poses of detected and projected 2D humans, $\mathcal{D}_o(\cdot)$ the IoU between the detected 2D bounding box and the convex hull of the projected 3D bounding box, and $\mathcal{D}_h(\cdot)$ the average pixel-wise Euclidean distance between two 2D poses.

2.4 SHADE Dataset

We collect SHADE (Synthetic Human Activities with Dynamic Environment), a self annotated dataset that consists of dynamic 3D human skeletons and objects, to learn the prior model for each HOI. It is collected from a video game Grand Theft Auto V with various daily activities and HOIs. Currently, there are over 29 million frames of 3D human poses, where 772,229 frames are annotated. On average, each annotated frame is associated with 2.03 action labels and 0.89 HOIs. There are 19 different HOI relation categories in the dataset; we choose 6 that usually occur in indoor scenes. Figure 2.2 shows some typical examples and relations in the dataset.

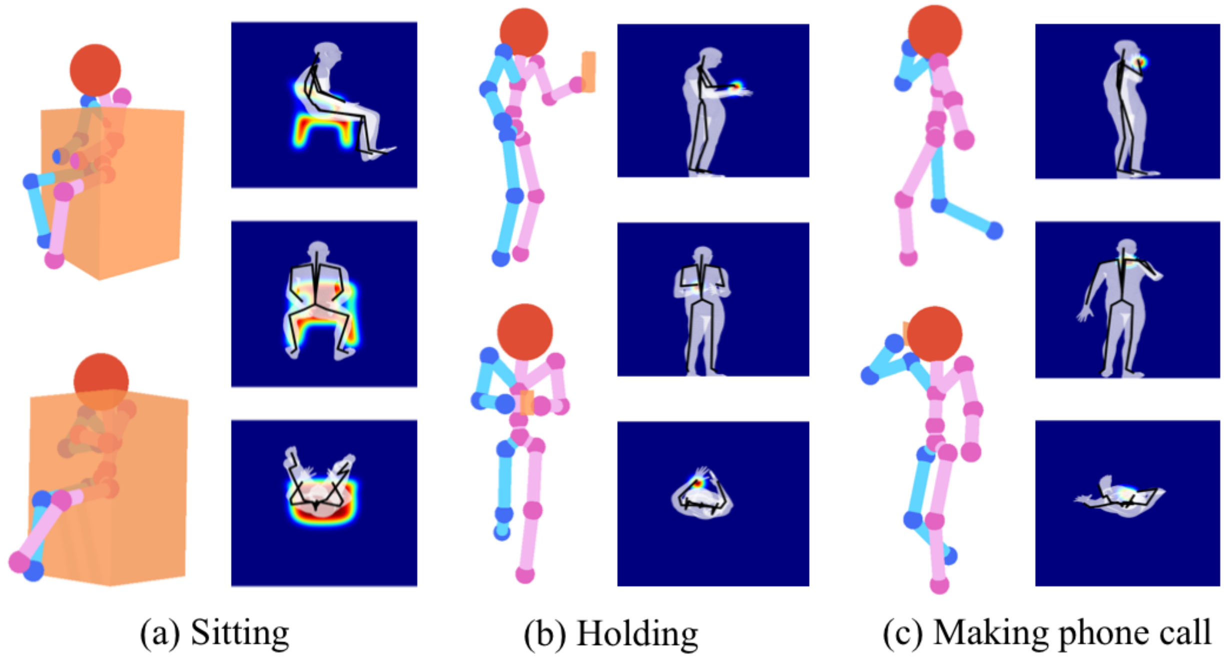


Figure 2.2: Examples of typical HOIs and examples from the SHADE dataset. The heatmap indicates the probable locations of HOI.

2.5 Joint Inference

Given a single RGB image as the input, the goal of joint inference is to find the optimal parse graph that maximizes the posterior probability $p(pg|I)$. The joint parsing is a four-step process: (i) 3D scene initialization of the camera pose, room layout, and 3D object bounding boxes, (ii) 3D human pose initialization that estimates rough 3D human poses in a 3D scene, (iii) concurrent action detection, and (iv) joint inference to optimize the objects, layout, and human poses in 3D scenes by maximizing the posterior probability.

2.5.1 3D Scene Initialization

Following [HQX18], we initialize the 3D objects, room layout, and camera pose cooperatively, where the room layout and objects are parametrized by 3D bounding boxes. For each object $v_i \in V_{object}$, we find its supporting object/layout by minimizing the supporting energy:

$$v_j^* = \arg\min_{v_j} \mathcal{E}_o(v_i, v_j) + \mathcal{E}_{height}(v_i, v_j) - \lambda_s \log p_{spt}(v_i, v_j), \tag{2.7}$$

where $v_j \in (V_{object}, V_{layout})$ and $p_{spt}(v_i, v_j)$ are the prior probabilities of the supporting relation modeled by multinoulli distributions, and λ_s a balancing constant.

2.5.2 3D Human Pose Initialization

We take 2D poses as the input and predict 3D poses in a local 3D coordinate following [TRA17], where the 2D poses are detected and estimated by [CSW17]. The local 3D coordinate is centered at the human hip joint, and the z-axis is aligned with the up direction of the world coordinate. To transform this local 3D pose into the world coordinate, we find the 3D world coordinate $\mathbf{v_{3D}} \in \mathbb{R}^3$ of one visible 2D joint $\mathbf{v_{2D}} \in \mathbb{R}^2$ (*e.g.*, head) by solving a linear equation with the camera intrinsic parameter K and estimated camera pose R. Per the pinhole camera projection model, we have

$$\alpha \begin{bmatrix} \mathbf{v_{2D}} \\ 1 \end{bmatrix} = K \cdot R \cdot \mathbf{v_{3D}}, \tag{2.8}$$

where α is a scaling factor in the homogeneous coordinate. To make the function solvable, we assume a pre-defined height h_0 for the joint position $\mathbf{v_{3D}}$ in the world coordinate. Lastly, the 3D pose initialization is obtained by aligning the local 3D pose and the corresponding joint position with $\mathbf{v_{3D}}$.

2.5.3 Concurrent Action Detection

We formulate the concurrent action detection as a multi-label classification problem to ease the ambiguity in describing the action. We define a portion of the action labels (*e.g.*, "eating", "making phone call") as the HOI labels, and the remaining action labels (*e.g.*, "standing", "bending") as general human poses without HOI. The mixture of HOI actions and non-HOI actions covers most of the daily human actions in indoor scenes. We manually map each of the HOI action labels to a 3D HOI relation learned from the SHADE dataset, and use the HOI actions as cues to improve the accuracy of 3D reconstruction by integrating it as prior knowledge in our model. The concurrent action detector takes 2D skeletons as the input and predicts multiple action labels with a three-layer multi-layer perceptron (MLP).

Algorithm 1 Joint Inference Algorithm

```
Given: Image $I$, initialized parse graph $pg_{init}$
procedure PHASE 1
    for Different temperatures do
        Inference with physical commonsense $\mathcal{E}_{phy}$ but without HOI $\mathcal{E}_{hoi}$: randomly select from room layout, objects, and human poses to optimize $pg$
procedure PHASE 2
    Match each agent with their interacting objects
procedure PHASE 3
    for Different temperatures do
        Inference with total energy $\mathcal{E}$, including physical commonsense and HOI: randomly select from layout, objects, and human poses to optimize $pg$
procedure PHASE 4
    Top-down sampling by HOIs
```

The dataset for training the concurrent action detectors consists of both synthetic data and real-world data. It is collected from: (i) The synthetic dataset described in section 2.4. We project the 3D human poses of different HOIs into 2D poses with random camera poses. (ii) The dataset proposed and collected by [JSL17], which also contains 3D poses of multiple persons in social interactions. We project 3D poses into 2D using the same method as (i). (iii) The 2D poses in an action recognition dataset [YJK11]. Our results show that the synthetic data can significantly expand the training set and help to avoid overfitting in concurrent action detection.

2.5.4 Inference

Given an initialized parse graph, we use MCMC with simulated annealing to jointly optimize the room layout, 3D objects, and 3D human poses through the non-differentiable energy space; see 1 as a summary. To improve the efficiency of the optimization process, we adopt a scheduling strategy that divides the optimization process into following four phases with different focuses: (i) Optimize objects, room layout, and human poses without HOIs. (ii) Assign HOI labels to each human in the scene, and search the interacting objects of each human. (iii) Optimize objects, room layout, and human poses jointly with HOIs. (iv) Generate possible miss-detected objects by top-down sampling.

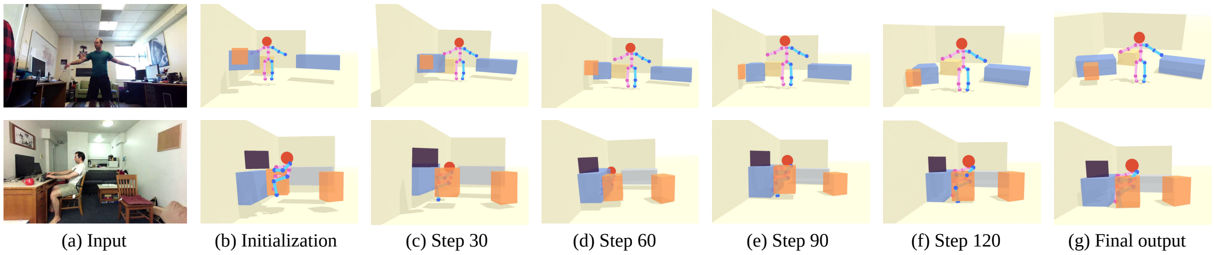


Figure 2.3: The optimization process of the scene configuration by simulated annealing MCMC. Each step is the number of accepted proposal.

Dynamics. In Phase (i) and (iii), we use distinct MCMC processes. To traverse non-differentiable energy spaces, we design Markov chain dynamics q_1^o, q_2^o, q_3^o for objects, q_1^l, q_2^l for room layout, and q_1^h, q_2^h, q_3^h for human poses.

• Object Dynamics: Dynamics q_1^o adjusts the position of an object, which translates the object center in one of the three Cartesian coordinate axes or along the depth direction. The depth direction starts from the camera position and points to the object center. Translation along depth is effective with proper camera pose initialization. Dynamics q_2^o proposes rotation of the object with a specified angle. Dynamics q_3^o changes the scale of the object by expanding or shrinking corner positions of the cuboid with respect to object center. Each dynamic can diffuse in two directions: each object can translate in the direction of '$+x$' and '$-x$,' or rotate in the direction of clockwise and counterclockwise. To better traverse in energy space, the dynamics may propose to move along the gradient descent direction with a probability of 0.95 or the gradient ascent direction with a probability of 0.05.

• Human Dynamics: Dynamics q_1^h proposes to translate 3D human joints along x, y, z, or depth direction. Dynamics q_2^h is designed to rotate the human pose with a certain angle. Dynamics q_3^h adjusts the scale of human poses by a scaling factor on the 3D joints with respect to the pose center.

• Layout Dynamics: Dynamics q_1^l translates the wall towards or away from the layout center. Dynamics q_2^l adjusts the floor height, equivalent to change the camera height.

In each sampling iteration, the algorithm proposes a new pg' from current pg under the proposal probability of $q(pg \to pg'|I)$ by applying one of the above dynamics. The generated proposal is accepted with respect to an acceptance rate $\alpha(\cdot)$ as in the Metropolis-Hastings algorithm [Has70]:

$$\alpha(pg \to pg') = \min(1, \frac{q(pg' \to pg) \cdot p(pg'|I)}{q(pg \to pg') \cdot p(pg|I)}), \tag{2.9}$$

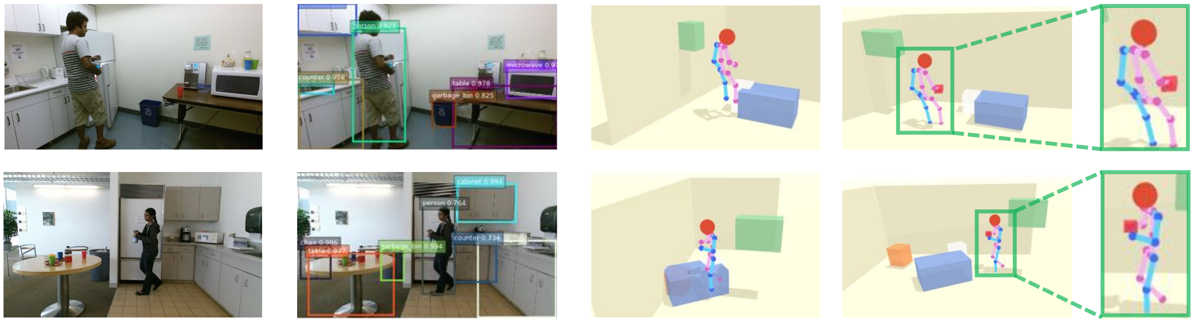
A simulated annealing scheme is adopted to obtain pg with high probability.

Top-down sampling. By top-down sampling objects from HOIs, the proposed method can recover the interacting 3D objects that are too small or novel to be detected by the state-of-the-art 2D object detector. In Phase (iv), we propose to sample an interacting object from the person if the confidence of HOI is higher than a threshold. Specifically, we minimize the HOI energy in Equation 2.4 to determine the category and location of the object; see examples in Figure 2.4.

Implementation Details. In Phase (ii), we search the interacting objects for each agent involved in HOI by minimizing the energy in Equation 2.4. In Phase (iii), after matching each agent with their interacting objects, we can jointly optimize objects, room layout, and human poses with the constraint imposed by HOI. Figure 2.3 shows examples of the simulated annealing optimization process.

2.6 Experiments

Since the proposed task is new and challenging, limited data and state-of-the-art methods are available for the proposed problem. For fair evaluations and comparisons, we evaluate



(a) Input (b) 2D Detection (c) Initialization (d) Model Output

Figure 2.4: Illustration of the top-down sampling process. The object detection module misses the detection of the bottle held by the person, but our model can still recover the bottle by reasoning HOI.

the proposed algorithm on three types of datasets: (i) Real data with full annotation on PiGraphs dataset [SCH16] with limited 3D scenes. (ii) Real data with partial annotation on daily activity dataset Watch-n-Patch [WZS15], which only contains ground-truth depth information and annotations of 3D human poses. (iii) Synthetic data with generated annotations to serve as the ground truth: we sample 3D human poses of various activities in SUN RGB-D dataset [SLX15] and project the sampled skeletons back onto the 2D image plane.

2.6.1 Comparative methods

To the best of our knowledge, no previous algorithm jointly optimizes the 3D scene and 3D human pose from a single image. Therefore, we compare our model against state-of-the-art methods for each task. Particularly, we compare with [HQX18] for single-image 3D scene reconstruction and VNect [MSS17] for 3D pose estimation in the world coordinate.

Since VNect can only estimate a single person during the estimation, we also design an additional baseline for multi-person 3D human pose estimation in the world coordinate. We first extract a 2048-D image feature vector using the Global Geometry Network (GGN) [HQX18] to capture the global geometry of the scene. The concatenated vector (GGN image feature, 2D pose, 3D pose in the local coordinate, and the camera intrinsic matrix) is then fed into a 5-layer fully connected network to predict the 3D pose. The fully-connected layers are trained using the mean squared error loss. We train the network on the training set of the synthetic SUN RGB-D dataset. Please refer to supplementary materials for more details of the baseline model.

2.6.2 Dataset

PiGraphs [SCH16] contains 30 scenes and 63 video recordings obtained by Kinect v2, designed to associate human poses with object arrangements. There are 298 actions available in approximately 2-hours of recordings. Each recording is about 2-minute long with an average 4.9 action annotation. We removed the frames without human appearance or annotations, resulting in 36,551 test images.

Watch-n-Patch (WnP) [WZS15] is an activity video dataset recorded by Kinect v2. It

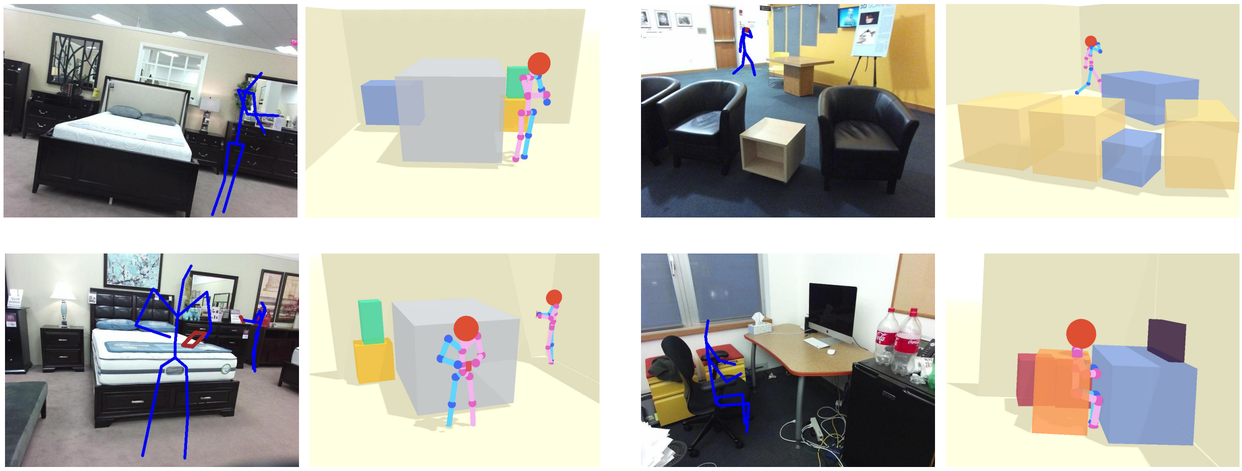


Figure 2.5: Augmenting SUN RGB-D with synthetic human poses.

contains several human daily activities as compositions of multiple actions interacting with various objects. The dataset comes with activity annotations, depth maps, and 3D human poses by Kinect. We test our algorithm on 1,210 randomly selected frames.

SUN RGB-D [SLX15] contains rich indoor scenes that are densely annotated with 3D bounding boxes, room layouts, and camera poses. The original dataset has 5,050 testing images, but we discarded images with no detected 2D objects, invalid 3D room layout annotation, limited space, or small field of view, resulting in 3,476 testing images.

Synthetic SUN RGB-D is an augmented SUN RGB-D dataset by sampling human poses in the scenes. Following methods of sampling imaginary human poses in [HQZ18], we extend the sampling to more generalized settings for various poses. The augmented human is represented by a 6-tuple $\langle a, \mu, t, r, s, \hat{\mu} \rangle$, where a is the action type, μ the pose template, t translation, r rotation, s scale, and $\hat{\mu} = \mu \cdot r \cdot s + t$ the imagined human skeleton. For each action label, we sample an imagined human pose inside a 3D scene: $\langle t^*, r^*, s^* \rangle = \arg\min_{t,r,s} \mathcal{E}_{phy} + \mathcal{E}_{hoi}$. If a is involved with any HOI unit, we further augment the 3D bounding box of the object. After sampling a human pose, we project the augmented 3D scenes back onto the 2D image plane using the ground truth camera matrix and camera pose; see examples in Figure 2.5. For a fair comparison of 3D human pose estimation on synthetic SUN RGB-D, all the algorithms are provided with the ground truth 2D skeletons as the input.

For 3D scene reconstruction, both [HQX18] and the proposed 3D scene initialization are learned using SUN RGB-D training data and tested on the above three datasets. For 3D pose estimation, both [MSS17] and the initialization of the proposed method are trained on public datasets, while the baseline is trained on synthetic SUN RGB-D. Note that we only use the SHADE dataset for learning a dictionary of HOIs.

2.6.3 Quantitative and Qualitative Results

We evaluate the proposed model on holistic^{++} scene understanding task by comparing the performances on both 3D scene reconstruction and 3D pose estimation.

Scene Reconstruction: We compute the 3D IoU and 2D IoU of object bounding boxes to evaluate the 3D scene reconstruction and the consistency between 3D world and 2D image. Following the metrics [HQX18], we compute the 3D IoU between the estimated 3D bounding boxes and the annotated 3D bounding boxes on PiGraphs and SUN RGB-D. For dataset without ground-truth 3D bounding boxes (*i.e.*, Watch-n-Patch), we evaluate the distance between the camera center and the 3D object center. To evaluate the 2D-3D consistency, the 2D IoU is computed between the projected 2D boxes of the 3D object bounding boxes and the ground-truth 2D boxes or detected 2D boxes (*i.e.*, Watch-n-Patch). As shown in Table 2.1, the proposed method improves the state-of-the-art 3D scene reconstruction results on all three datasets without specific training on each of them. More importantly, it significantly improves the results on PiGraphs and Watch-n-Patch compared with [HQX18]. The most likely reason is [HQX18] is trained on SUN RGB-D dataset in a purely data-driven fashion, therefore difficult to generalize across to other datasets (*i.e.*, PiGraphs, and Watch-n-Patch). In contrast, the proposed model incorporates more general prior knowledge of HOI and physical commonsense, and combined such knowledge with 2D-3D consistency (likelihood) for joint inference, avoiding the over-fitting caused by the direct 3D estimation from 2D. Figure 2.6 shows the qualitative results on all three datasets.

Pose Estimation: We evaluate the pose estimation in both 3D and 2D. For 3D evaluation, we compute the Euclidean distance between the estimated 3D joints and the 3D ground truth and average it over all the joints. For 2D evaluation, we project the estimated 3D pose

Table 2.1: Quantitative Results of 3D Scene Reconstruction

Methods	Huang *et al.* [HQX18]			Ours		
Metric	2D IoU (%)	3D IoU (%)	Depth (m)	2D IOU (%)	3D IoU (%)	Depth (m)
PiGraphs	68.6	21.4	-	**7**5.1	**2**4.9	-
SUN RGB-D	63.9	17.7	-	**7**2.9	**1**8.2	-
WnP	-	-	0.375	-	-	**0**.162

Table 2.2: Quantitative Results of Global 3D Pose Estimation

Methods	VNect[MSS17]		Baseline		Ours	
Metrics	2D (pix)	3D (m)	2D (pix)	3D (m)	2D (pix)	3D (m)
PiGraphs	63.9	0.732	284.5	2.67	**1**5.9	**0**.472
SUNRGBD	-	-	45.81	**0**.435	**1**4.03	0.517
WnP	50.51	0.646	325.2	2.14	**2**0.5	**0**.330

Table 2.3: Ablative results of HOI on 3D object IoU (%), 3D pose estimation error (m), and miss-detection rate (MR, %)

Methods	*w/o hoi*			*Full model*		
HOI Type	Object $\uparrow$	Pose $\downarrow$	MR $\downarrow$	Object $\uparrow$	Pose $\downarrow$	MR $\downarrow$
Sit	26.9	0.590	15.2	**2**7.8	**0**.521	**1**3.1
Hold	17.4	0.517	78.9	**1**7.6	**0**.490	**5**4.6
Use Laptop	14.1	0.544	58.8	**1**5.0	**0**.534	**4**3.3
Read	**1**4.5	0.466	65.3	14.3	**0**.453	**4**1.9

back to 2D image plane and compute the pixel distance against ground truth. Quantitative results are shown in Table 2.2. The proposed method outperforms two other methods in both 2D and 3D. On the synthetic SUN RGB-D dataset, all algorithms are given the ground truth 2D poses as the input for fair comparison. Although the baseline model achieves better performances since the 3D human poses are synthesized with limited templates and the baseline model fits it well, the 3D poses estimated by VNect and baseline model deviate a lot from the ground truth for datasets with real human poses (*i.e.*, PiGraph, and Watch-n-Patch). In contrast, the proposed algorithm still performs well, demonstrating an outstanding generalization ability across various datasets.

Ablative Analysis to analyze the contributions of HOI and physical commonsense by comparing two variants of the proposed full model: (i) model *w/o HOI*: without HOI $\mathcal{E}_{hoi}(pg)$, and (ii) model *w/o phy.*: without physical commonsense $\mathcal{E}_{phy}(pg)$.

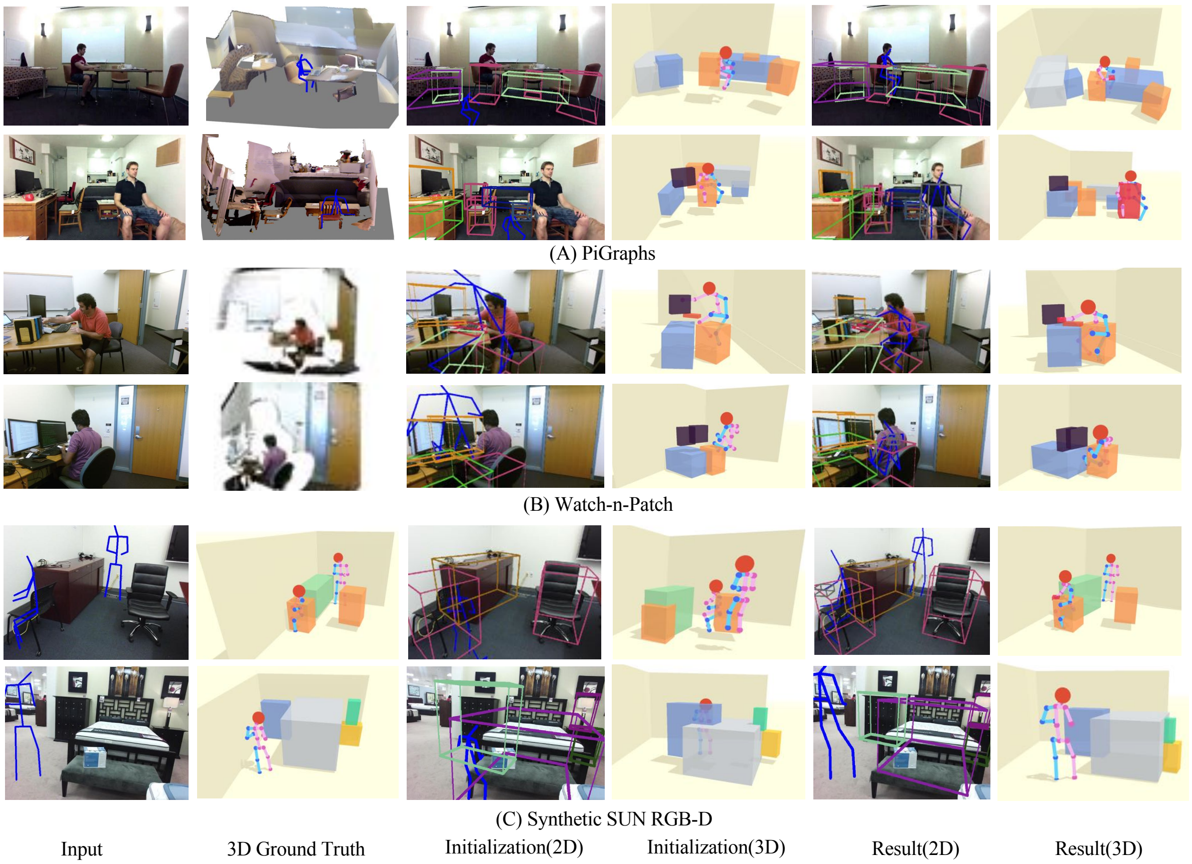


Figure 2.6: Qualitative results of the proposed method on three datasets. The proposed model improves the initialization with accurate spatial relations and physical plausibility and demonstrates an outstanding generalization across various datasets.

Human-Object Interaction. We compare our full model with model *w/o hoi* to evaluate the effects of each category of HOI. Evaluation metrics include 3D pose estimation error, 3D bounding box IoU, and miss-detection rate (MR) of the objects interacted with agents. The experiments are conducted on PiGraphs dataset and Synthetic SUN RGB-D dataset with the annotated HOI labels. As shown in Table 2.3, the performances of both scene reconstruction and human pose estimation are hindered without reasoning HOI, indicating HOI helps to infer the relative spatial relationship between humans and interacted objects to further improve the performance of both two tasks. Moreover, a marked performance gain of miss-detection rate implies the effectiveness of the top-down sampling process during the joint inference.

Physical Commonsense. Reasoning about physical commonsense drives the reconstructed

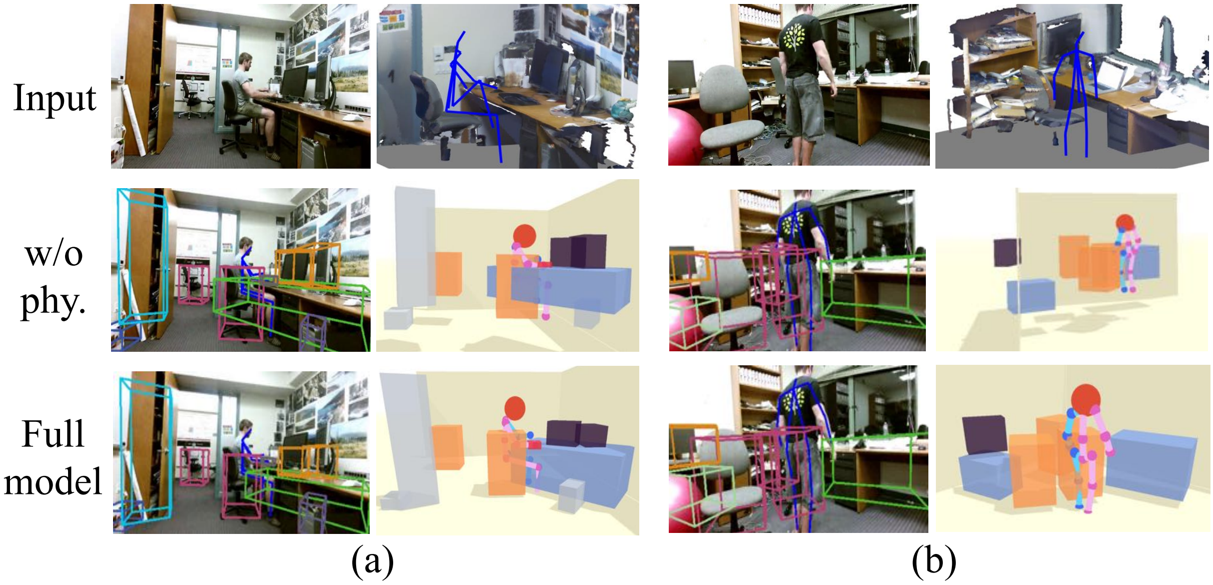


Figure 2.7: Qualitative comparison between (a) model *w/o phy.* and (b) the full model on PiGraphs dataset.

3D scene to be physically plausible and stable. We tested 3D estimation of object bounding boxes on the PiGraphs dataset using *w/o phy.* and the full model. The full model outperforms *w/o phy.* from two aspects: (i) 3D object detection IoU (from 23.5% to 24.9%), and (ii) physical violation (from 0.223m to 0.150m). The physical violation is computed as the distance between the lower surface of an object and the upper surface of its supporting object. The qualitative comparisons are shown in Figure 2.7. Objects detected by model *w/o phy.* may float in the air or penetrate each other, while the full model yields physically plausible results.

2.7 Conclusion

This work tackles a challenging holistic^{++} scene understanding problem to jointly solve 3D scene reconstruction and 3D human pose estimation from a single RGB image. By incorporating physical commonsense and reasoning about HOI, our approach leverages the coupled nature of these two tasks and goes beyond merely reconstructing the 3D scene or human pose by reasoning about the concurrent action of human in the scene. We design a joint inference algorithm which traverses the non-differentiable solution space with MCMC and optimizes the scene configuration. Experiments on PiGraphs, Watch-n-Patch, and Synthetic SUN RGB-D demonstrate the efficacy of the proposed algorithm, and the general prior knowledge of HOI and physical commonsense.

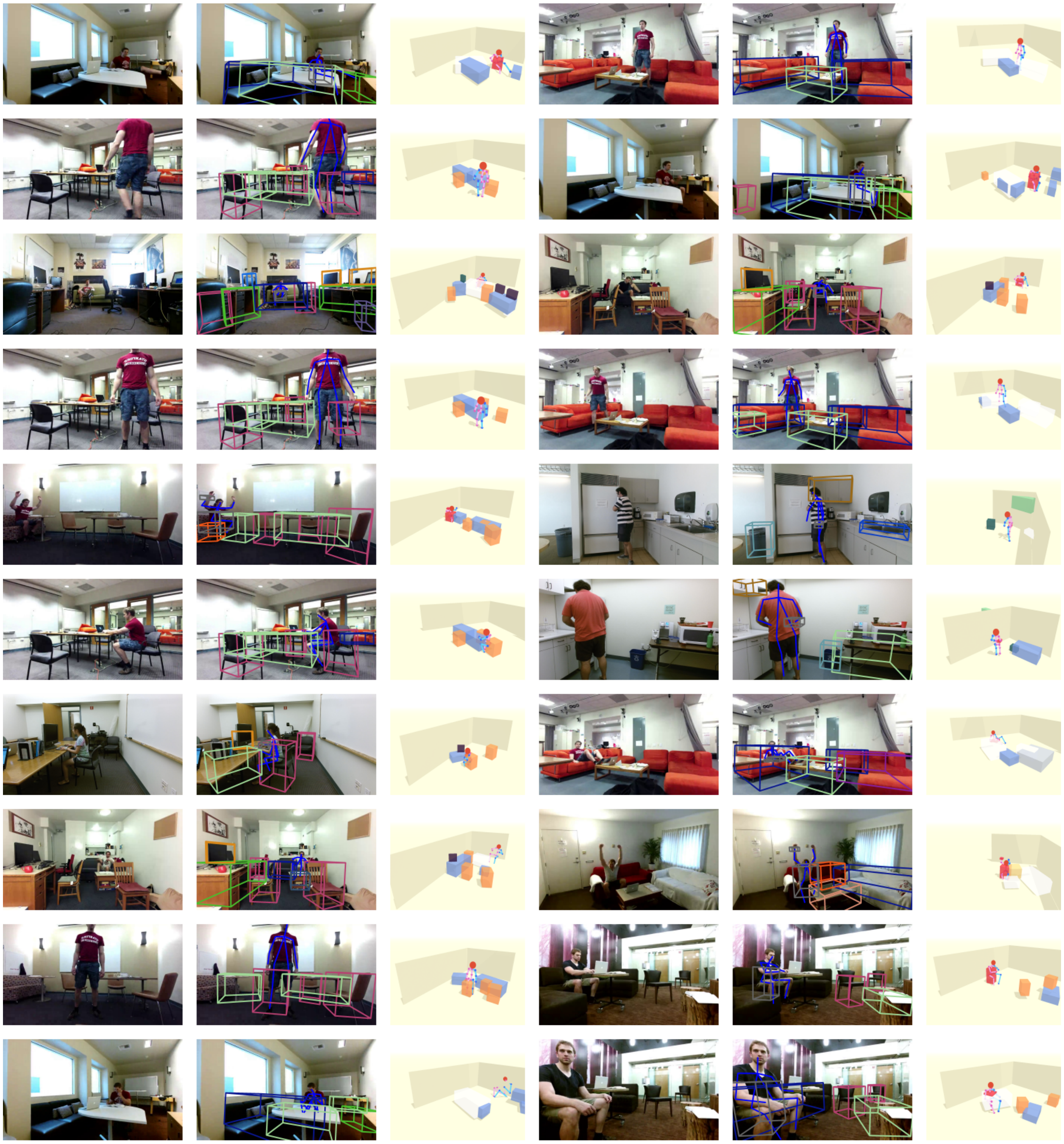


Figure 2.8: More results

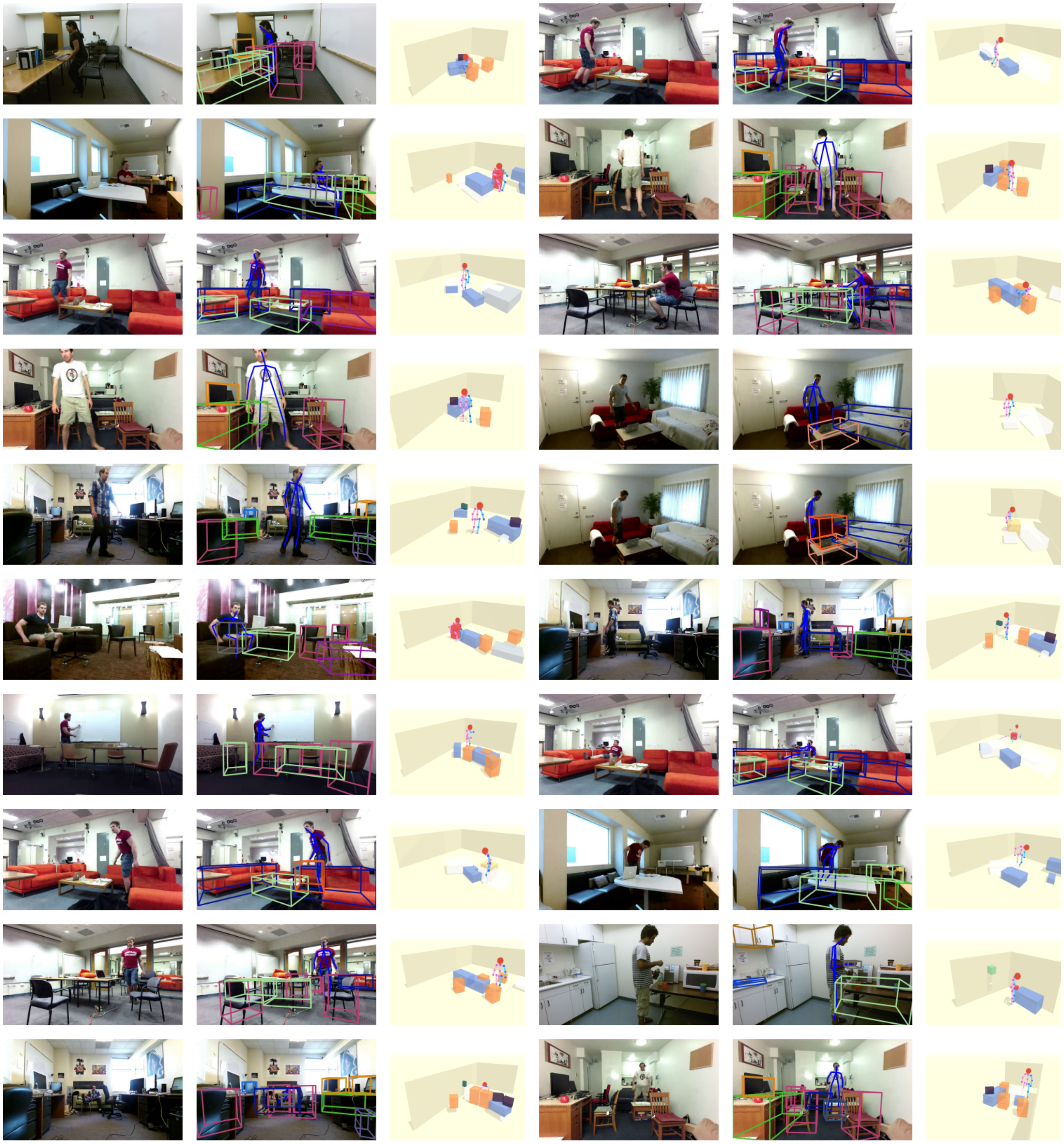


Figure 2.9: More results

CHAPTER 3

Scene-centric Joint Parsing of Cross-view Videos

3.1 Introduction

During the past decades, remarkable progress has been made in many vision tasks, e.g., image classification, object detection, pose estimation. Recently, more comprehensive visual tasks probe deeper understanding of visual scenes under interactive and multi-modality settings, such as visual Turing tests [GGH15, QWL15] and visual question answering [AAL15]. In addition to discriminative tasks focusing on binary or categorical predictions, emerging research involves representing fine-grained relationships in visual scenes [KZG17, ABY16] and unfolding semantic structures in contexts including caption or description generation [YYL10], and question answering [TML14, ZGB16].

In this work, we present a framework for uncovering the semantic structure of scenes in a cross-view camera network. The central requirement is to resolve ambiguity and establish cross-reference among information from multiple cameras. Unlike images and videos shot from single static point of view, cross-view settings embed rich physical and geometry constraints due to the overlap between fields of views. While multi-camera setups are common in real-word surveillance systems, large-scale cross-view activity dataset are not available due to privacy and security reasons. This makes data-demanding deep learning approaches infeasible.

Our joint parsing framework computes a hierarchy of spatio-temporal parse graphs by establishing cross-reference of entities among different views and inferring their semantic attributes from a scene-centric perspective. For example, Fig. 3.1 shows a parse graph hierarchy that describes a scene where two people are playing a ball. In the first view,

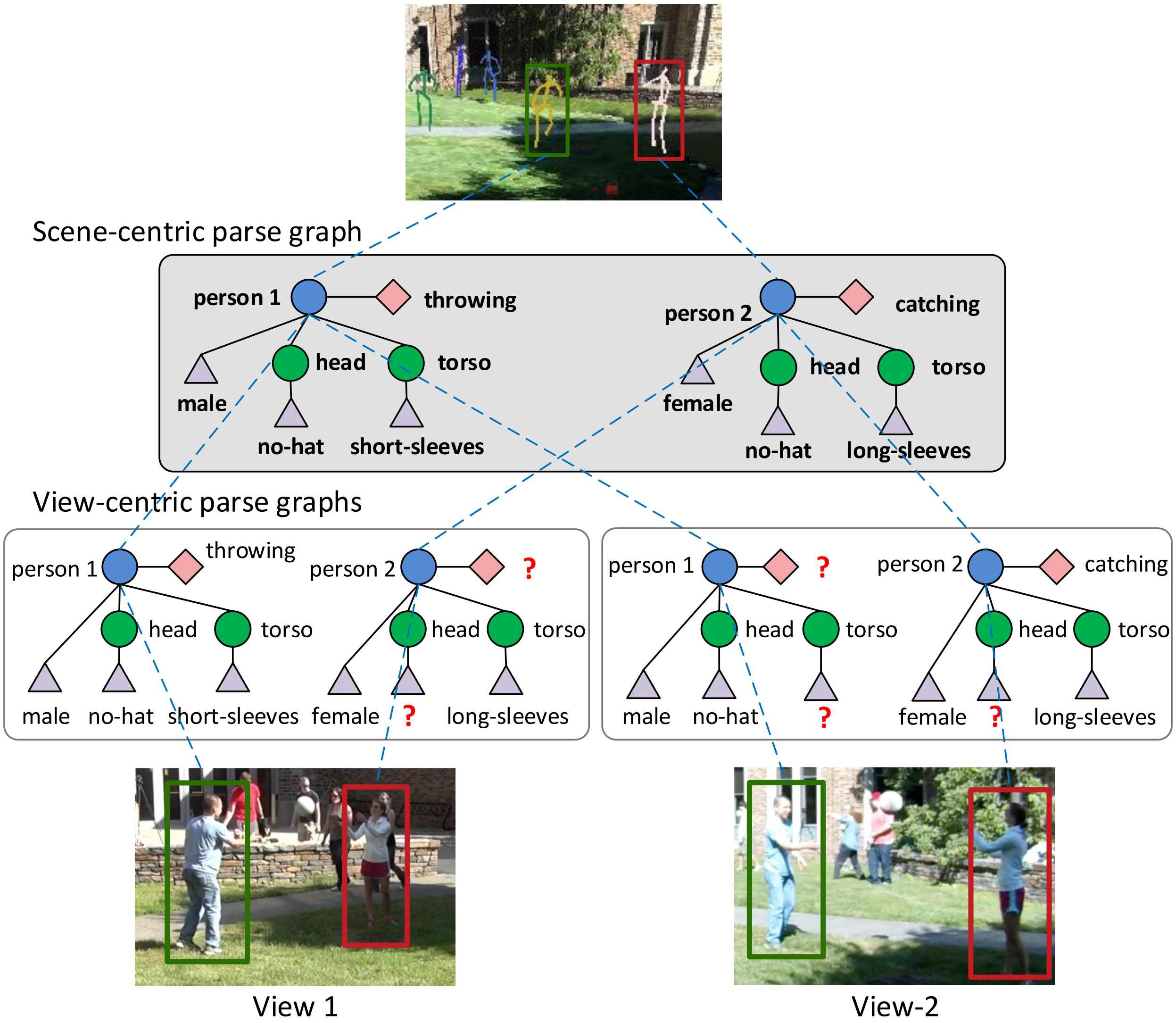


Figure 3.1: An example of the spatio-temporal semantic parse graph hierarchy in a visual scene captured by two cameras.

person 2's action is not grounded because of the cluttered background, while it is detected in the second view. Each view-centric parse graph contains local recognition decisions in an individual view, and the scene centric parse graph summaries a comprehensive understanding of the scene with coherent knowledge.

The structure of each individual parse graph fragment is induced by an ontology graph that regulates the domain of interests. A parse graph hierarchy is used to represent the correspondence of entities between the multiple views and the scene. We use a probabilistic model to incorporate various constraints on the parse graph hierarchy and formulate the joint parsing as a MAP inference problem. A MCMC sampling algorithm and a dynamic programming algorithm are used to explore the joint space of scene-centric and view-centric interpretations and to optimize for the optimal solutions. Quantitative experiments show that scene-centric parse graphs outperforms the initial view-centric proposals.

Contributions. The contributions of this work are three-fold: (i) a unified hierarchical parse graph representation for cross-view person, action, and attributes recognition; (ii) a stochastic inference algorithm that explores the joint space of scene-centric and view-centric interpretations efficiently starting with initial proposals; (iii) a joint parse graph hierarchy that is an interpretable representation for scene and events.

3.2 Related Work

Our work is closely related to three research areas in computer vision and artificial intelligence.

Multi-view video analytics. Typical multi-view visual analytics tasks include object detection [LS10a, UB11a], cross-view tracking [BFT11a, LPR12, XLL16, XLQ17], action recognition [WNX14], person re-identification [XLZ13, XMH14] and 3D reconstruction [HWR13a]. While heuristics such as appearances and motion consistency constraints have been used to regularize the solution space, these methods focus on a specific multi-view vision task whereas we aim to propose a general framework to jointly resolve a wide variety of tasks.

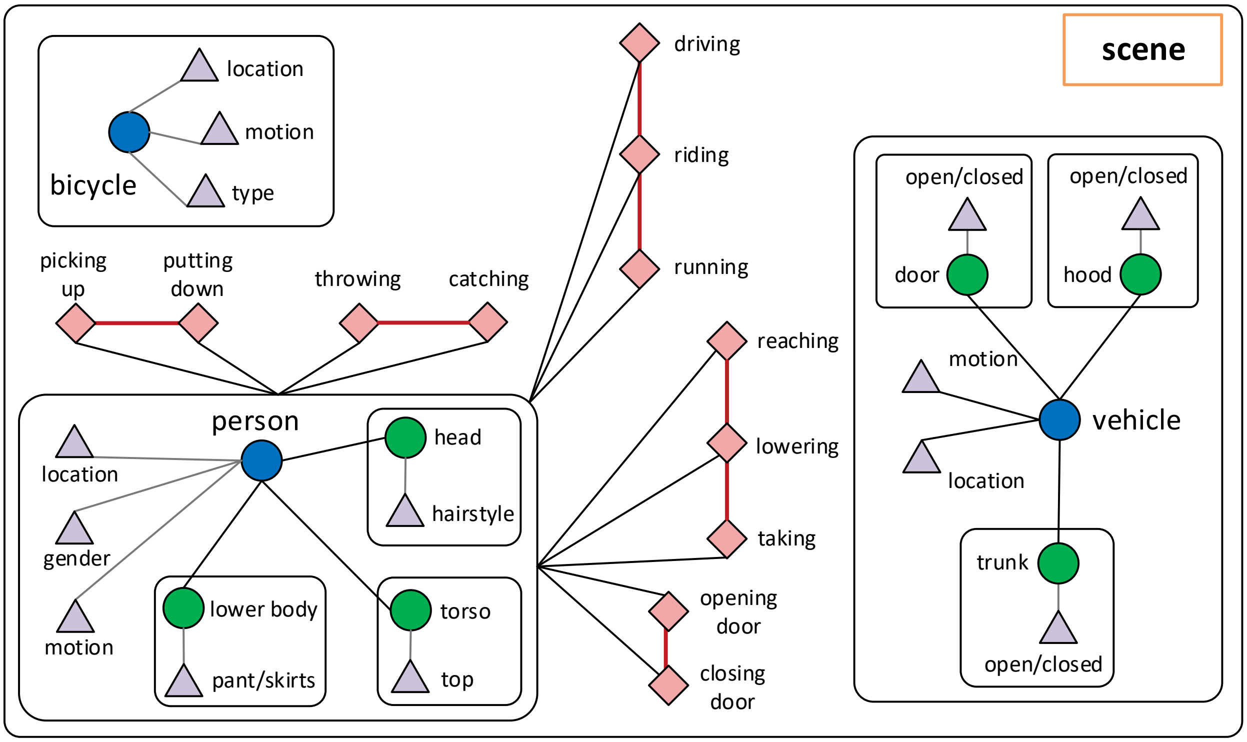


Figure 3.2: An illustration of the proposed ontology graph describing objects, parts, actions and attributes.

Semantic representations. Semantic and expressive representations have been developed for various vision tasks, e.g., image parsing [HZ09], 3D scene reconstruction [LZZ14, PBH13], human-object interaction [KS16], pose and attribute estimation [WZZ16]. In this work, our representation also falls into this category. The difference is that our model is defined upon cross-view spatio-temporal domain and is able to incorporate a variety of tasks.

Interpretability. Automated generation of explanations regarding predictions has a long and rich history in artificial intelligence. Explanation systems have been developed for a wide range of applications, including simulator actions [VFM04, LCV05, CLV06], robot movements [LCC12], and object recognition in images [BM14, HAR16]. Most of these approaches are rule-based and suffer from generalization across different domains. Recent methods including [RSG16] use proxy models or data to interpret black box models, while our scene-centric parse graphs are explicit representations of the knowledge by definition.

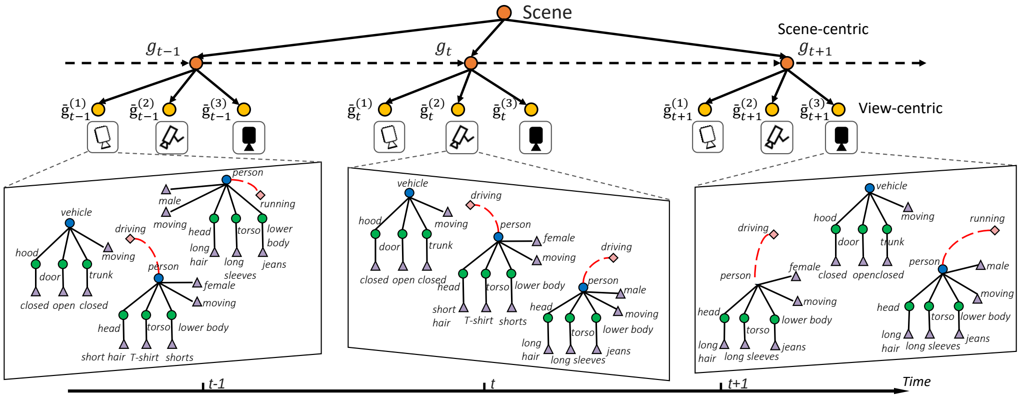


Figure 3.3: The proposed spatio-temporal parse graph hierarchy. (Better viewed electronically and zoomed).

3.3 Representation

A scene-centric spatio-temporal parse graph represents humans, their actions and attributes, interaction with other objects captured by a network of cameras. We will first introduce the concept of ontology graph as domain definitions, then we will describe parse graphs and parse graph hierarchy as view-centric and scene-centric representations respectively.

Ontology graph. To define the scope of our representation on scenes and events, an ontology is used to describe a set of plausible objects, actions and attributes. We define an ontology as a graph that contains nodes representing objects, parts, actions, attributes respectively and edges representing the relationships between nodes. Specifically, every object and part node is a concrete type of object that can be detected in videos. Edges between object and part nodes encodes "part-of" relationships. Action and attribute nodes connected to an object or part node represent plausible actions and appearance attributes the object can take. For example, Fig. 3.2 shows an ontology graph that describes a domain including people, vehicles, bicycles. An object can be decomposed into parts (i.e., green nodes), and enriched with actions (i.e., pink nodes) and attributes (i.e., purple diamonds). The red edges among action nodes denote their incompatibility. The ontology graph can be considered a compact AOG [LZZ14, WZZ16] without the compositional relationships and event hierarchy. In this work, we focus on a restricted domain inspired by [QWL15], while larger ontology graphs can be easily derived from large-scale visual relationship datasets such as [KZG17] and open-domain knowledge bases such as [LS04].

Parse graphs. While an ontology describes plausible elements, only a subset of these concepts can be true for a given instance at a given time. For example, a person cannot be both “standing” and “sitting” at the same time, while both are plausible actions that a person can take. To distinguish plausible facts and satisfied facts, we say a node is *grounded* when it is associated with data. Therefore, a subgraph of the ontology graph that only contains grounded nodes can be used to represent a specific *instance* (e.g. a specific person) at a specific time. In this work, we refer to such subgraphs as *parse graphs*.

Parse graph hierarchy. In cross-view setups, since each view only captures an incomplete set of facts in a scene, we use a spatio-temporal hierarchy of parse graphs to represent the collective knowledge of the scene and all the individual views. To be concrete, a view-centric parse graph $\tilde{g}$ contains nodes grounded to a video sequence captured by an individual camera, whereas a scene-centric parse graph g is an aggregation of view-centric parse graphs and therefore reflects a global understanding of the scene. As illustrated in Fig. 3.3, for each time step t, the scene-centric parse graph g_t is connected with the corresponding view-centric parse graphs $\tilde{g}_t^{(i)}$ indexed by the views, and the scene-centric graphs are regarded as a Markov chain in the temporal sequence. In terms of notations, in this work we use a tilde notation to represent the view-centric concepts $\tilde{x}$ corresponding to scene-centric concepts x.

3.4 Probabilistic Formulation

The task of joint parsing is to infer the spatio-temporal parse graph hierarchy

$$G = \langle \Phi, g, \tilde{g}^{(1)}, \tilde{g}^{(2)}, \ldots, \tilde{g}^{(M)} \rangle$$

from the input frames from video sequences $I = \{I_t^{(i)}\}$ captured by a network of M cameras , where Φ is an object identity mapping between scene-centric parse graph g and view-centric parse graphs $\tilde{g}^{(i)}$ from camera i. Φ defines the structure of parse graph hierarchy. In this section, we discuss the formulation assuming a fixed structure, while defer the discussion of how to traverse the solution space to section 3.5.

We formulate the inference of parse graph hierarchy as a MAP inference problem in a posterior distribution $p(G|I)$ as follows

$$G^* = \arg\max_G p(I|G) \cdot p(G). \tag{3.1}$$

Likelihood. The likelihood term models the grounding of nodes in view-centric parse graphs to the input video sequences. Specifically,

$$\begin{aligned} p(I|G) &= \prod_{i=1}^{M} \prod_{t=1}^{T} p(I_t^{(i)}|\tilde{g}_t^{(i)}) \\ &= \prod_{i=1}^{M} \prod_{t=1}^{T} \prod_{v \in V(\tilde{g}_i^{(t)})} p(I(v)|v), \end{aligned} \tag{3.2}$$

where $\tilde{g}_t^{(i)}$ is the view-centric parse graph of camera i at time t and $V(\tilde{g}_t^{(i)})$ is the set of nodes in the parse graph. $p(I(v)|v)$ is the node likelihood for the concept represented by node v being grounded on the data fragment $I(v)$. In practice, this probability can be approximated by normalized detection and classifications scores [PRF11].

Prior. The prior term models the compatibility of scene-centric and view-centric parse graphs across time. We factorize the prior as

$$p(G) = p(g_1) \prod_{t=1}^{T-1} p(g_{t+1}|g_t) \prod_{i=1}^{M} \prod_{t=1}^{T} p(\tilde{g}_t^{(i)}|g_t), \tag{3.3}$$

where $p(g_1)$ is a prior distribution on parse graphs that regulates the combination of nodes, and $p(g_t|g_{t-1})$ is a transitions probability of scene-centric parse graphs across time. Both probability distributions are estimated from training sequences. $p(\tilde{g}_t^{(i)}|g_t)$ is defined as a Gibbs distribution that models the compatibility of scene-centric and view-centric parse graphs in the hierarchy (we drop subscripts t and camera index i for brevity).

$$\begin{aligned} p(\tilde{g}|g) &= \frac{1}{Z}\exp\{-\mathcal{E}(g,\tilde{g})\} \\ &= \frac{1}{Z}\exp\{-w_1\mathcal{E}_S(g,\tilde{g}) - w_2\mathcal{E}_A(g,\tilde{g}) \\ &\qquad - w_3\mathcal{E}_{Act}(g,\tilde{g}) - w_4\mathcal{E}_{Attr}(g,\tilde{g})\}, \end{aligned} \tag{3.4}$$

where energy $\mathcal{E}(g,\tilde{g})$ is decomposed into four different terms described in detail in the subsection below. The weights are tuning parameters that can be learned via cross-validation. We consider view-centric parse graphs for videos from different cameras are independent conditioned on scene-centric parse graph under the assumption that all cameras have fixed and known locations.

3.4.1 Cross-view Compatibility

In this subsection, we describe the energy function $\mathcal{E}(g,\tilde{g})$ for regulating the compatibility between the occurrence of objects in the scene and an individual view from various aspects. Note that we use a tilde notation to represent the node correspondence in scene-centric and view-centric parse graphs (i.e., for a node $v \in g$ in a scene-centric parse graph, we refer to the corresponding node in a view-centric parse graph as $\tilde{v}$).

Appearance similarity. For each object node in the parse graph, we keep an appearance descriptor. The appearance energy regulates the appearance similarity of object o in the scene-centric parse graph and $\tilde{o}$ in the view-centric parse graphs.

$$\mathcal{E}_A(g,\tilde{g}) = \sum_{o\in g} ||(\phi(o) - \phi(\tilde{o})||_2, \tag{3.5}$$

where $\phi(\cdot)$ is the appearance feature vector of the object. At the view-level, this feature vector can be extracted by pre-trained convolutional neural networks; at the scene level, we use a mean pooling of view-centric features.

Spatial consistency. At each time point, every object in a scene has a fixed physical location in the world coordinate system while appears on the image plane of each camera according to the camera projection. For each object node in the parse graph hierarchy, we keep a scene-centric location $s(o)$ for each object o in scene-centric parse graphs and a view-centric location $s(\tilde{o})$ on the image plane in view-centric parse graphs. The following energy is defined to enforce the spatial consistency:

$$\mathcal{E}_S(g, \tilde{g}) = \sum_{o \in g} ||s(o) - h(s(\tilde{o}))||_2, \tag{3.6}$$

where $h(\cdot)$ is a perspective transform that maps a person's view-centric foot point coordinates to the world coordinates on the ground plane of the scene with the camera homography, which can be obtained via the intrinsic and extrinsic camera parameters.

Action compatibility. Among action and object part nodes, scene-centric human action predictions shall agree with the human pose observed in individual views from different viewing angles:

$$\mathcal{E}_{Act}(g, \tilde{g}) = \sum_{l \in g} -\log p(l|\tilde{p}), \tag{3.7}$$

where l is an action node in scene-centric parse graphs and $\tilde{p}$ are positions of all human parts in the view-centric parse graph. In practice, we separately train a action classifier that predicts action classes with joint positions of human parts and uses the classification score to approximate this probability.

Attribute consistency. In cross-view sequences, entities observed from multiple cameras shall have a consistent set of attributes. This energy term models the commonsense constraint that scene-centric human attributes shall agree with the observation in individual views:

$$\mathcal{E}_{Attr}(g, \tilde{g}) = \sum_{a \in g} \mathbf{1}(a \neq \tilde{a}) \cdot \xi, \tag{3.8}$$

where $\mathbf{1}(\cdot)$ is an indicator function and ξ is a constant energy penalty introduced when the two predictions mismatch.

3.5 Inference

The inference process consists of two sub-steps: (i) matching object nodes Φ in scene-centric and view-centric parse graphs (i.e. the structure of parse graph hierarchy) and (ii) estimating optimal values of parse graphs $\{g, \tilde{g}^{(1)}, \ldots, \tilde{g}^{(M)}\}$.

The overall procedure is as follows: we first obtain view-centric objects, actions, and attributes proposals from pre-trained detectors on all video frames. This forms the initial view-centric predictions $\{\tilde{g}^{(1)}, \ldots, \tilde{g}^{(M)}\}$. Next we use a Markov Chain Monte Carlo (MCMC) sampling algorithm to optimize the parse graph structure Φ. Given a fixed parse graph hierarchy, variables within the scene-centric and view-centric parse graphs $\{g, \tilde{g}^{(1)}, \ldots, \tilde{g}^{(M)}\}$ can be efficiently estimated by a dynamic programming algorithm. These two steps are performed iteratively until convergence.

3.5.1 Inferring Parse Graph Hierarchy

We use a stochastic algorithm to traverse the solution space of the parse graph hierarchy Φ. To satisfy the detailed balance condition, we define three reversible operators $\Theta = \{\Theta_1, \Theta_2, \Theta_3\}$ as follows.

Merging. The merging operator Θ_1 groups a view-centric parse graph with an other view-centric parse graph by creating a scene-centric parse graph that connects the two. The operator requires the two operands to describe two objects of the same type either from different views or in the same view but with non-overlapping time intervals.

Splitting. The splitting operator Θ_2 splits a scene-centric parse graph into two parse graphs such that each resulting parse graph only connects to a subset of view-centric parse graphs.

Swapping. The swapping operator Θ_3 swaps two view-centric parse graphs. One can view the swapping operator as a shortcut of merging and splitting combined.

We define the proposal distribution $q(G \to G')$ as an uniform distribution. At each iteration, we generate a new structure proposal Φ' by applying one of the three operators Θ_i with respect to probability 0.4, 0.4, and 0.2, respectively. The generated proposal is then accepted with respect to an acceptance rate $\alpha(\cdot)$ as in the Metropolis-Hastings algorithm [MRR53]:

$$\alpha(G \to G') = \min\left(1, \frac{q(G' \to G) \cdot p(G'|I)}{q(G \to G') \cdot p(G|I)}\right), \tag{3.9}$$

where $p(G|I)$ the posterior is defined in Eqn. (3.1).

3.5.2 Inferring Parse Graph Variables

Given a fixed parse graph hierarchy, we need to estimate the optimal value for each node within each parse graph. As illustrated in Fig. 3.3, for each frame, the scene-centric node g_t and the corresponding view-centric nodes $\tilde{g}_t^{(i)}$ form a star model, and the whole scene-centric nodes are regarded as a Markov chain in the temporal order. Therefore the proposed model is essentially a Directed Acyclic Graph (DAG). To infer the optimal node values, we can simply apply the standard factor graph belief propagation (sum-product) algorithm.

3.6 Experiments

3.6.1 Setup and Datasets

We evaluate our scene-centric joint-parsing framework in tasks including object detection, multi-object tracking, action recognition, and human attributes recognition. In object detection and multi-object tracking tasks, we compare with published results. In action recognition and human attributes tasks, we compare the performance of view-centric proposals without joint parsing and scene-centric predictions after joint parsing as well as additional baselines. The following datasets are used to cover a variety of tasks.

The **CAMPUS** dataset [XLL16] [1] contains video sequences from four scenes each captured by four cameras. Different from other multi-view video datasets focusing solely on multi-object tracking task, videos in the CAMPUS dataset contains richer human poses and activities with moderate overlap in the fields of views between cameras. In addition to the tracking annotation in the CAMPUS dataset, we collect new annotation that includes 5 action categories and 9 attribute categories for evaluating action and attribute recognition.

[1] bitbucket.org/merayxu/multiview-object-tracking-dataset

The **TUM Kitchen** dataset [TBB09][2] is an action recognition dataset that contains 20 video sequences captured by 4 cameras with overlapping views. As we only focusing on the RGB imagery inputs in our framework, other modalities such as motion capturing, RFID tag reader signals, magnetic sensor signals are not used as inputs in our experiments. To evaluate detection and tracking task, we compute human bounding boxes from motion capturing data by projecting 3D human poses to the image planes of all cameras using the intrinsic and extrinsic parameters provided in the dataset. To evaluate human attribute tasks, we annotate 9 human attribute categories for every subject.

In our experiments, both the CAMPUS and the TUM Kitchen datasets are used in all tasks. In the following subsection, we present isolated evaluations.

3.6.2 Evaluation

Object detection & tracking. We use FasterRCNN [RHG15] to create initial object proposals on all video frames. The detection scores are used in the likelihood term in Eqn. (3.2). During joint parsing, objects which are not initially detected on certain views are projected from object's scene-centric positions with the camera matrices. After joint parsing, we extract all bounding boxes that are grounded by object nodes from each view-centric parse graph to compute multi-object detection accuracy (DA) and precision (DP). Concretely, the accuracy measures the faction of correctly detected objects among all ground-truth objects and the precision is computed as fraction of true-positive predictions among all output predictions. A predicted bounding box is considered a match with a ground-truth box only if the intersection over union (IoU) score is greater than 0.5. When more than one prediction overlaps with a ground-truth box, only the one with the maximum overlap is counted as true positive.

When extracting all bounding boxes on which the view-centric parse graphs are grounded and grouping them according to the identity correspondence between different views, we obtain object trajectories with identity matches across multiple videos. In the evaluation, we compute four major tracking metrics: multi-object tracking accuracy (TA), multi-object

[2] ias.in.tum.de/software/kitchen-activity-data

CAMPUS-S1	DA (%)	DP (%)	TA (%)	TP (%)	IDSW	FRAG
Fleuret et al.	24.52	64.28	22.43	64.17	2269	2233
Berclaz et al.	30.47	62.13	28.10	62.01	2577	2553
Xu et al.	49.30	72.02	56.15	72.97	320	141
Ours	**56.00**	72.98	**55.95**	72.77	310	138
CAMPUS-S2	DA (%)	DP (%)	TA (%)	TP (%)	IDSW	FRAG
Fleuret et al.	16.51	63.92	13.95	63.81	241	214
Berclaz et al.	24.35	61.79	21.87	61.64	268	249
Xu et al.	27.81	71.74	28.74	71.59	1563	443
Ours	**28.24**	71.49	**27.91**	71.16	1615	418
CAMPUS-S3	DA (%)	DP (%)	TA (%)	TP (%)	IDSW	FRAG
Fleuret et al.	17.90	61.19	16.15	61.02	249	235
Berclaz et al.	19.46	59.45	17.63	59.29	264	257
Xu et al.	49.71	67.02	49.68	66.98	219	117
Ours	**50.60**	67.00	**50.55**	66.96	212	113
CAMPUS-S4	DA (%)	DP (%)	TA (%)	TP (%)	IDSW	FRAG
Fleuret et al.	11.68	60.10	11.00	59.98	828	812
Berclaz et al.	14.73	58.51	13.99	58.36	893	880
Xu et al.	24.46	66.41	24.08	68.44	962	200
Ours	**24.81**	66.59	**24.63**	68.28	938	194
TUM Kitchen	DA (%)	DP (%)	TA (%)	TP (%)	IDSW	FRAG
Fleuret et al.	69.88	64.54	69.67	64.76	61	57
Berclaz et al.	72.39	63.27	72.20	63.51	48	44
Xu et al.	86.53	72.12	86.18	72.37	9	5
Ours	**89.13**	72.21	**88.77**	72.42	12	8

Table 3.1: Quantitative comparisons of multi-object tracking on CAMPUS and TUM Kitchen datasets.

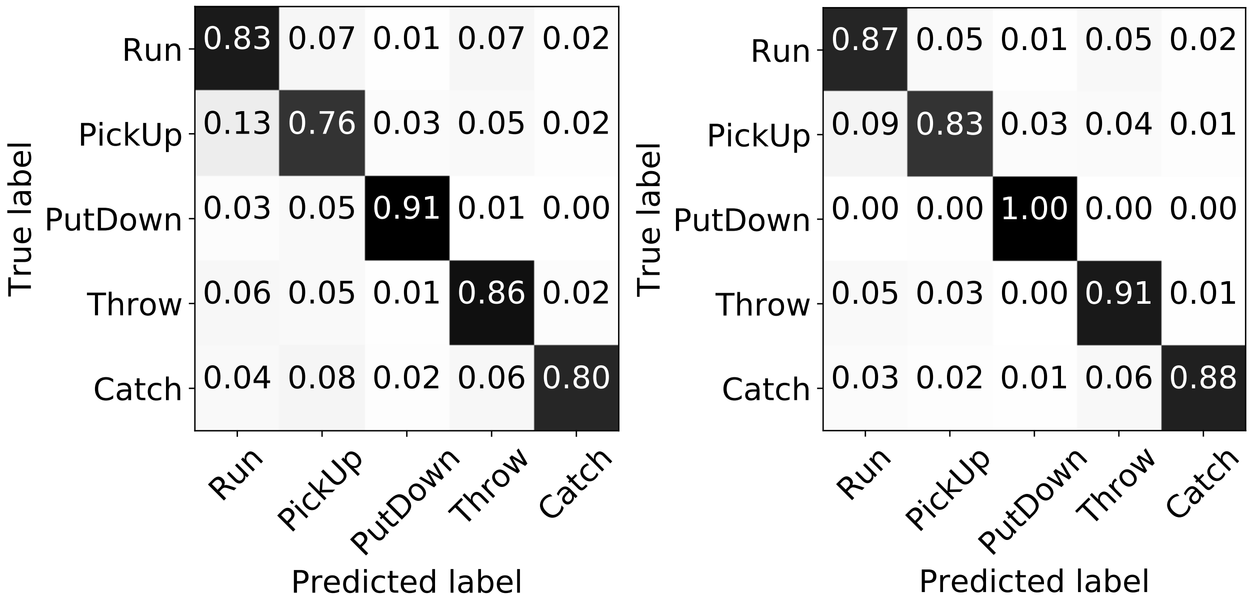


Figure 3.4: Confusion matrices of action recognition on view-centric proposals (left) and scene-centric predictions (right).

track precision (TP), the number of identity switches (IDSW), and the number of fragments (FRAG). A higher value of TA and TP and a lower value of IDSW and FRAG indicate the tracking method works better. We report quantitative comparisons with several published methods [XLL16, BFT11a, FBL08] in Table 3.1. From the results, the performance measured by tracking metrics are comparable to published results. We conjecture that the appearance similarity is the main drive for establish cross-view correspondence while additional semantic attributes proved limited gain to the tracking task.

	CAMPUS						**TUM Kitchen**								
Methods	Run	PickUp	PutDown	Throw	Catch	Overall	Reach	Taking	Lower	Release	OpenDoor	CloseDoor	OpenDrawer	CloseDrawer	Overall
view-centric	0.83	0.76	0.91	0.86	0.80	0.82	0.78	0.66	0.75	0.67	0.48	0.50	0.50	0.42	0.59
baseline-vote	0.85	0.80	0.71	0.88	0.82	0.73	0.80	0.63	0.77	0.71	0.72	0.73	0.70	0.47	0.69
baseline-mean	0.86	0.82	1.00	0.90	0.87	0.88	0.79	0.61	0.75	0.69	0.67	0.67	0.66	0.45	0.66
scene-centric	0.87	0.83	1.00	0.91	0.88	**0.90**	0.81	0.67	0.79	0.71	0.71	0.73	0.70	0.50	**0.70**

Table 3.2: Quantitative comparisons of human action recognition on CAMPUS and TUM Kitchen datasets.

	Methods	Gender	Long hair	Glasses	Hat	T-shirt	Long sleeve	Shorts	Jeans	Long pants	mAP
CAMPUS	view-centric	0.59	0.77	0.56	0.76	0.36	0.59	0.70	0.63	0.35	0.59
	baseline-mean	0.63	0.82	0.55	0.75	0.34	0.64	0.69	0.63	0.34	0.60
	baseline-vote	0.61	0.82	0.55	0.75	0.34	0.65	0.69	0.63	0.35	0.60
	scene-centric	0.76	0.82	0.62	0.80	0.40	0.62	0.76	0.62	0.24	**0.63**
	Methods	Gender	Long hair	Glasses	Hat	T-shirt	Long sleeve	Shorts	Jeans	Long pants	mAP
TUM Kitchen	view-centric	0.69	0.93	0.32	1.00	0.50	0.89	0.91	0.83	0.73	0.76
	baseline-mean	0.86	1.00	0.32	1.00	0.54	0.96	1.00	0.83	0.81	0.81
	baseline-vote	0.64	1.00	0.32	1.00	0.32	0.93	1.00	0.83	0.76	0.76
	scene-centric	0.96	0.98	0.32	1.00	0.77	0.96	0.94	0.83	0.83	**0.84**

Table 3.3: Quantitative comparisons of human attribute recognition on CAMPUS and TUM Kitchen datasets.

Action recognition. View-centric action proposals are obtained from a fully-connected neural network with 5 hidden layers and 576 neurons which predicts action labels using human pose. For the CAMPUS dataset, we collect additional annotations for 5 human action classes: Run, PickUp, PutDown, Throw, and Catch in total of 8,801 examples. For the TUM Kitchen dataset, we evaluate on the 8 action categories: Reaching, TakingSomething, Lowering, Releasing, OpenDoor, CloseDoor, OpenDrawer, and CloseDrawer. We measure both individual accuracies for each category as well as the overall accuracies across all categories. Table 3.2 shows the performance of scene-centric predictions with view-centric proposals, and two additional fusing strategies as baselines. Concretely, the *baseline-vote* strategy takes action predictions from multiple views and outputs the label with majority voting, while the *baseline-mean* strategy assumes equal priors on all cameras and outputs the label with the highest averaged probability. When evaluating scene-centric predictions, we project scene-centric labels back to individual bounding boxes and calculate accuracies following the same procedure as evaluating view-centric proposals. Our joint parsing framework demonstrates improved results as it aggregates marginalized decisions made on individual views while also encourages solutions that comply with other tasks. Fig. 3.4 compares the confusion matrix of view-centric proposals and scene-centric predictions after joint parsing for CAMPUS dataset. To further understand the effect of multiple views, we break down classification accuracies by the number of cameras where persons are observed (Fig. 3.5). Observing an entity from more cameras generally leads to better performance, while too many conflicting observations may also cause degraded performance. Fig. 3.6 shows some success and failure examples.

Human attribute recognition. We follow the similar procedure as in the action recognition case above. Additional annotations for 9 different types of human attributes are collected for both CAMPUS and TUM Kitchen dataset. View-centric proposals and score are obtained from an attribute grammar model as in [PNZ16]. We measure performance with average precisions for each attribute categories as well as mean average precision (mAP) as in human attribute literatures. Scene-centric predictions are projected to bounding boxes in each views when calculating precisions. Table 3.3 shows quantitative comparisons be-

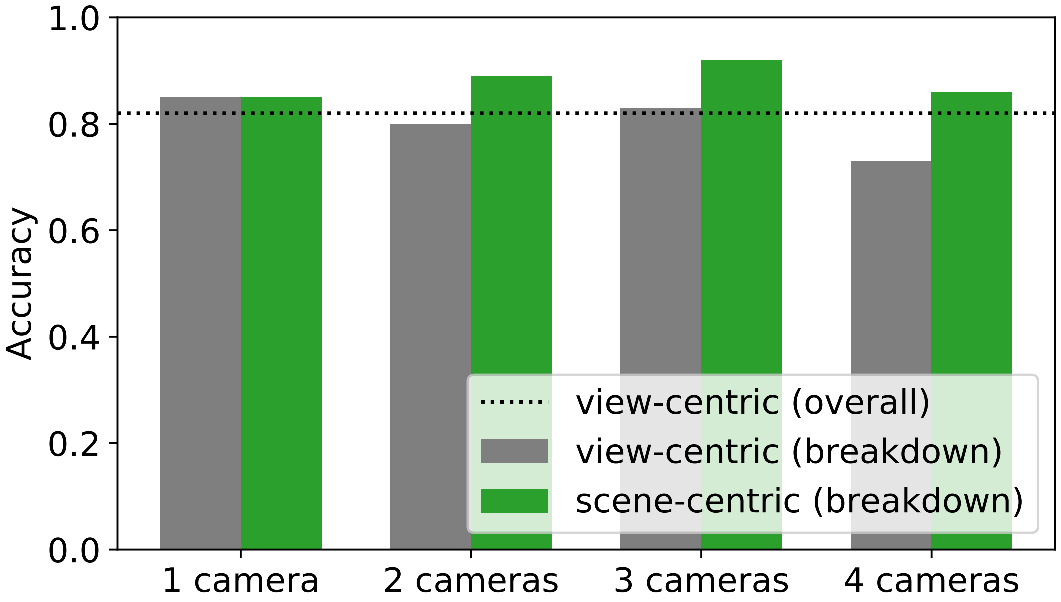


Figure 3.5: The breakdown of action recognition accuracy according to the number of camera views in which each entity is observed.

tween view-centric and scene-centric predictions. The same baseline fusing strategies as in the action recognition task are used. The scene-centric prediction outperforms the original proposals in 7 out of 9 categories while remains comparable in others. Notably, the CAMPUS dataset is harder than standard human attribute datasets because of occlusions, limited scales of humans, and irregular illumination conditions.

3.6.3 Runtime

With initial view-centric proposals precomputed, for a 3-minute scene shot by 4 cameras containing round 15 entities, our algorithm performs at 5 frames per second on average. With further optimization, our proposed method can run in real-time. Note that although the proposed framework uses a sampling-based method, using view-based proposals as initialization warm-starts the sampling procedure. Therefore, the overall runtime is significantly less than searching the entire solution space from scratch. For problems of a larger size, more efficient MCMC algorithms may be adopted. For example, the mini-batch acceptance testing technique [CSP16] has demonstrated several order-of-magnitude speedups.

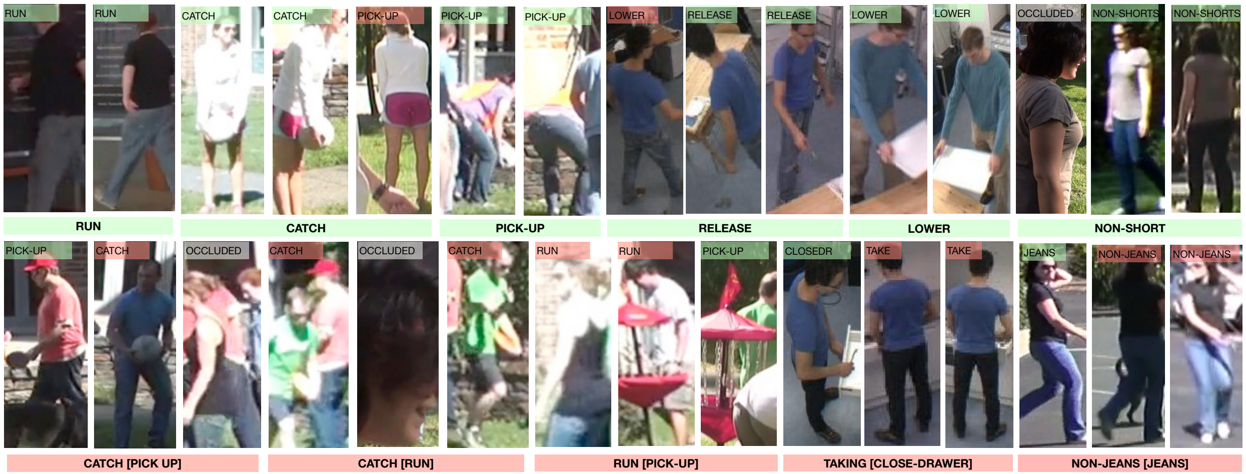


Figure 3.6: Success (1st row) and failure examples (2nd row) of view-centric (labels overlaid on the images) and scene-centric predictions (labels beneath the images) of action and attribute recognition tasks. For failure examples, true labels are in the bracket. "Occluded" means that the locations of objects or parts are projected from scene locations and therefore no view-centric proposals are generated. Better viewed in color.

3.7 Conclusion

We represent a joint parsing framework that computes a hierarchy of parse graphs which represents a comprehensive understanding of cross-view videos. We explicitly specify various constraints that reflect the appearance and geometry correlations among objects across multiple views and the correlations among different semantic properties of objects. Experiments show that the joint parsing framework improves view-centric proposals and produces more accurate scene-centric predictions in various computer vision tasks.

We briefly discuss advantages of our joint parsing framework and potential future directions from two perspectives.

3.7.0.1 Explicit Parsing

While the end-to-end training paradigm is appealing in many *data-rich* supervised learning scenarios, as an extension, leveraging loosely-coupled pre-trained modules and exploring commonsense constraints can be helpful when large-scale training data is not available or too expensive to collect in practice. For example, many applications in robotics and human-robot interaction domains share the same set of underlying perception units such as scene understanding, object recognition, etc. Training for every new scenarios entirely could end up with exponential number of possibilities. Leveraging pre-trained modules and explore correlation and constraints among them can be treated as a factorization of the problem space. Therefore, the explicit joint parsing scheme allows practitioners to leverage pre-trained modules and to build systems with an expanded skill set in a scalable manner.

3.7.0.2 Interpretable Interface.

Our joint parsing framework not only provides a comprehensive scene-centric understanding of the scene, moreover, the sence-centric spatio-temporal parse graph representation is an interpretable interface of computer vision models to users. In particular, we consider the following properties an explainable interface shall have apart from the correctness of answers:

- *Relevance*: an agent shall recognize the intent of humans and provide information relevant to humans' questions and intents.

- *Self-explainability*: an agent shall provide information that can be interpreted by humans as how answers are derived. This criterion promotes humans' trust on an intelligent agent and enables sanity check on the answers.

- *Consistency*: answers provided by an agents shall be consistent throughout an interaction with humans and across multiple interaction sessions. Random or non-consistent behaviors cast doubts and confusions regarding the agent's functionality.

- *Capability*: an explainable interface shall help humans understand the boundary of capabilities of an agent and avoid blinded trusts.

Potential future directions include quantifying and evaluating the interpretability and user satisfaction by conducting user studies.

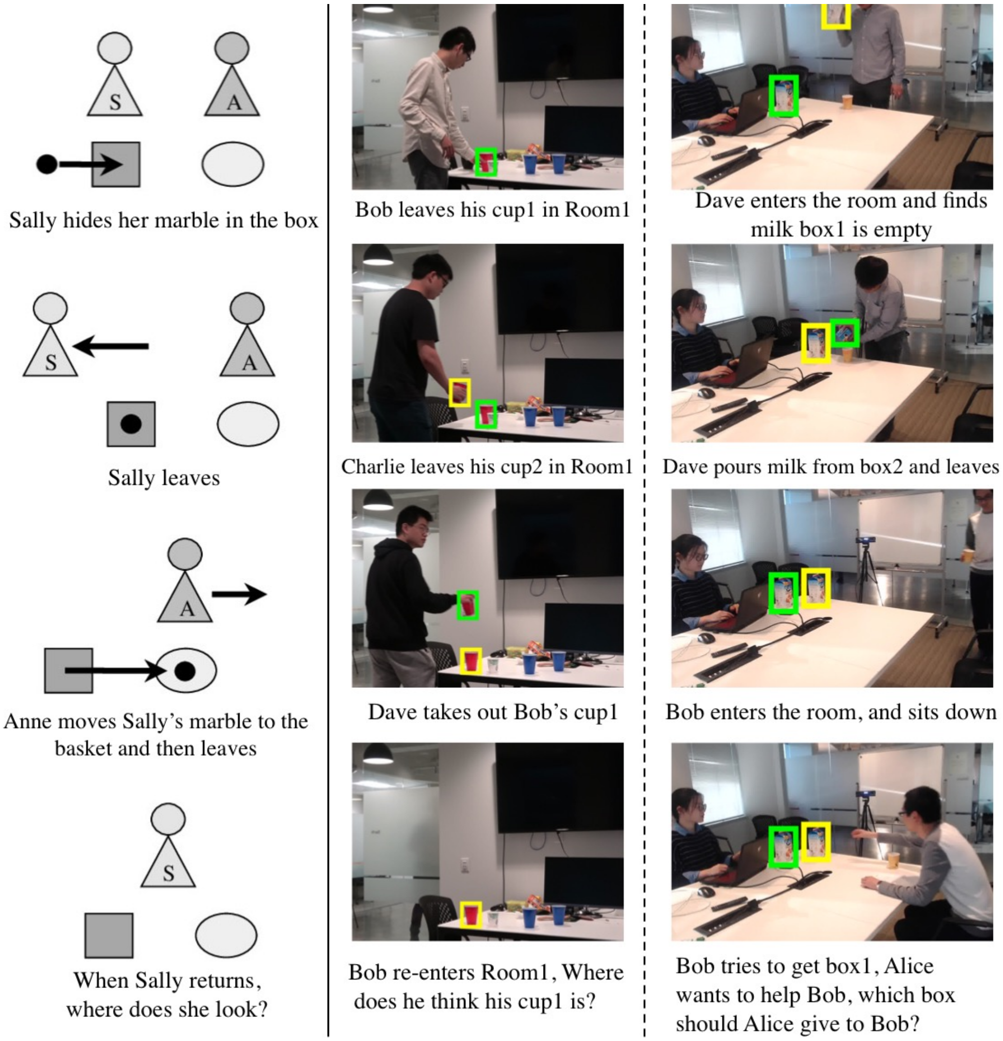
CHAPTER 4

Understanding False-Belief by Joint Inference of Object States, Robot Knowledge, and Human (False-)Beliefs

4.1 Introduction

Since its publication in 1985, the seminal Sally-Anne [BLF85] study has spawned a vast research literature in developmental psychology regarding Theory of Mind (ToM); in particular, human's social cognition in understanding *false-belief*—the ability to understand other's belief about the world may contrast with the true reality. A cartoon version of the Sally-Anne test is shown in the left of Figure 4.1: Sally puts her marble in the box and left. While Sally is out, Anne moves the marble from the box to a basket. The test would ask a human participant where Sally would look for her marble when she is back. In this experiment, the marble would still be inside the box according to Sally's false-belief, even though the marble is actually inside the basket. To answer this question correctly, a subject or an algorithm should understand and disentangle the object state (observation from the current frame), the (accumulated) knowledge, the belief of other agents, the ground-truth/reality of the world, and most importantly, the concept of false-belief.

Previous study suggests that at the age of 4 years old, children start to develop the capability to understand the concept of false-belief [GA88]. Such abilities to ascribe the mental belief to the human mind, to differentiate belief from the physical reality, and even to recognize false-belief and perform psychological reasoning, is a significant milestone in the acquisition of ToM [Sar14, WP83]. Further research also suggests that the cognitive



Sally hides her marble in the box

Sally leaves

Anne moves Sally's marble to the basket and then leaves

When Sally returns, where does she look?

Bob leaves his cup1 in Room1

Charlie leaves his cup2 in Room1

Dave takes out Bob's cup1

Bob re-enters Room1, Where does he think his cup1 is?

Dave enters the room and finds milk box1 is empty

Dave pours milk from box2 and leaves

Bob enters the room, and sits down

Bob tries to get box1, Alice wants to help Bob, which box should Alice give to Bob?

Figure 4.1: Left: Illustration of the classic Sally-Anne test [BLF85]. Middle and Right: Two different types of false-belief scenarios in our dataset: belief test and helping test.

capability to understand human false-belief plays a vital role in helping to explain and predict others' behavior, as well as in motivating human's pro-social helping and cooperative behavior since early childhood [BCT09, Tom18, PRG18]. In fact, humans are found to be ultra-cooperative both cognitively and motivationally stemming from the genetic trait, which in turn also motivates taking and coordinating different belief perspectives simultaneously for mental coordination and shared intentionality [Tom18]. Taking together, all these evidence emerged from developmental psychology in the past few decades call for integrating such socio-cognitive aspects into a modern anthropomorphic robot [BGB09].

In fact, false-belief is not rare in our daily life. Two examples are depicted in Figure 4.1:

(i) Where does Bob think his cup1 is after Charlie put cup2 (visually identical to cup1) on the table while Dave took cup1 away? (ii) Which milk box should Alice give to Bob if she wants to help? The one closer to Bob but empty, or the one further to Bob but full?

Although such false-belief tasks are primal examples for social and cognitive intelligence, current state-of-the-art intelligent systems are still facing challenges in acquiring such a capability. One fundamental challenge is the lack of proper representation for modeling the false-belief from visual input; it has to be able to handle the heterogeneous information of a system's current states, its accumulated knowledge, agent's belief, and the reality/ground-truth of the world. Without a unified representation, the information across all these domains cannot be easily interpreted, and the cross-domain reasoning of the events is infeasible.

Largely due to this difficulty, prior work can only solve a sub-problem in understanding false-belief. For instance, sensor fusion techniques are mainly used to obtain better state estimation by filtering the measurements from multiple same or different sensors [LHL17]. Similarly, the Multiple View Tracking (MVT) in computer vision is designed to combine the observations across camera views to better track an object. Visual cognitive reasoning (*e.g.*, human intention/attention predictions [QHW17, FCW18, WLS18]) only targets to model human mental states. These three lines of work are all crucial ingredients but developed independently; a cross-domain unified representation is still largely missing.

In order to endow such an ability to understand the concept of false-belief to a robot system, this work proposes to use a graphical model represented by a parse graph (*pg*) [ZM07] to serve as the unified representation of a robot's knowledge structure, fused knowledge across all robots, and the (false-)beliefs of human agents. A *pg* is learned from the spatiotemporal transition of humans and objects in the scene perceived by a robot in a distributed robot system. A joint *pg* can be induced by merging and fusing the individual *pg* from each robot to support inference for the human beliefs and object states. In particular, our system enables the following three capabilities with increasing depth in cognition:

1. *Tracking small objects with occlusions across different views.* The objects in an indoor environment (*e.g.*, cups) are usually small and have a similar appearance. Tracking such

small objects are even more challenging due to frequent occlusions with human interactions. Additionally, each robot's camera view has minimal overlaps with others. The proposed method can address such challenging multi-view multi-object tracking problem by properly maintaining cross-view object states and robot knowledge using the unified representation.

2. *Inferring human beliefs.* The state of an object normally does not change unless a human interacts with it. By identifying the interactions between human and objects, our system also supports the high-level cognitive capability; *e.g.*, knowing which object is interacted with which person, whether a person knows the state of the object has been changed, *etc.*
3. *Helping agents by recognizing false-belief.* Giving the above object tracking and cognitive reasoning of human beliefs, the proposed algorithm can further infer whether and why the person has false-belief, thereby to better assist the person given a specific context.

4.1.1 Related Work

Multi-view Visual Analysis is widely applied to 3D reconstruction [HWR13b], object detection [LS10b, UB11b], cross-view tracking [BFT11b, XLL16], and joint parsing [QXY18]. Built on top of these modules, Multiple Object Tracking (MOT) usually utilizes tracking-by-detection techniques [XLL16, WLY14, DTT15, DSY17]. This line of work primarily focuses on combining different camera views to obtain more comprehensive tracking of parsing results, lacking the capability of cognitive reasoning and human (false-)belief understanding.

Visual Cognitive Reasoning is an emerging field in computer vision. Related work includes predicting human activities [QHW17], recovering incomplete trajectories [LZZ18], learning utility and affordance [ZJZ16, ZZZ15], inferring human intention and attention [FCW18, WLS18], *etc.* As to understanding (false-)belief, despite many psychological experiments and theoretical analysis [CT99, WAS18, BBP16], very few attempts have been made to solve (false-)belief in man-made scenes with visual input; handcrafted constraints are usually required for specific problems in prior work. In contrast, our work utilizes a unified representation across different domains with heterogeneous information to model human mental states.

Robot ToM, aiming at understanding human beliefs and intents, receives increasing research attentions in human-robot interaction and collaboration [Sca02, THC16]. Several false-belief tasks akin to the classic Sally-Anne test were designed. For instance, Warnier *et al*. [WGL12] introduced a belief management algorithm, and the reasoning capability is subsequently endowed to a robot to pass the Sally-Anne test [MWC14] successfully. More sophisticated human-robot collaboration is achieved by maintaining a human partner's mental state [DA16]. More formally, Dynamic Epistemic Logic is introduced to represent and reason about belief and false-belief [Bol18, LR19]. These successes are, however, limited to the symbolic-based belief representations, requiring handcrafted variables and structures, making the logic-based reasoning approaches brittle in practice to handle noises and errors. To address this deficiency, our work utilizes a unified representation by *pg*, a probabilistic graphical model that have been successfully applied to various robotics tasks [EGX17, LZS18]; it not only can accumulate the observations over time to form a knowledge graph, but also can account for noises and errors to properly handle visual input during the learning and inference.

4.1.2 Contribution

Our work makes three contributions:

1. We adopt a unified graphical model *pg* to represent and maintain heterogeneous knowledge about object states, robot knowledge, and human beliefs.

2. On top of the unified representation, we propose an inference algorithm to merge individual *pg* from different domains across time and views into a joint *pg*, supporting human belief inference from multi-view to overcome the noises and errors merely from a single view.

3. With the inferred *pg*s, our system can keep track of the state and location of each object, infer human beliefs, and further discover false-belief to assist human better.

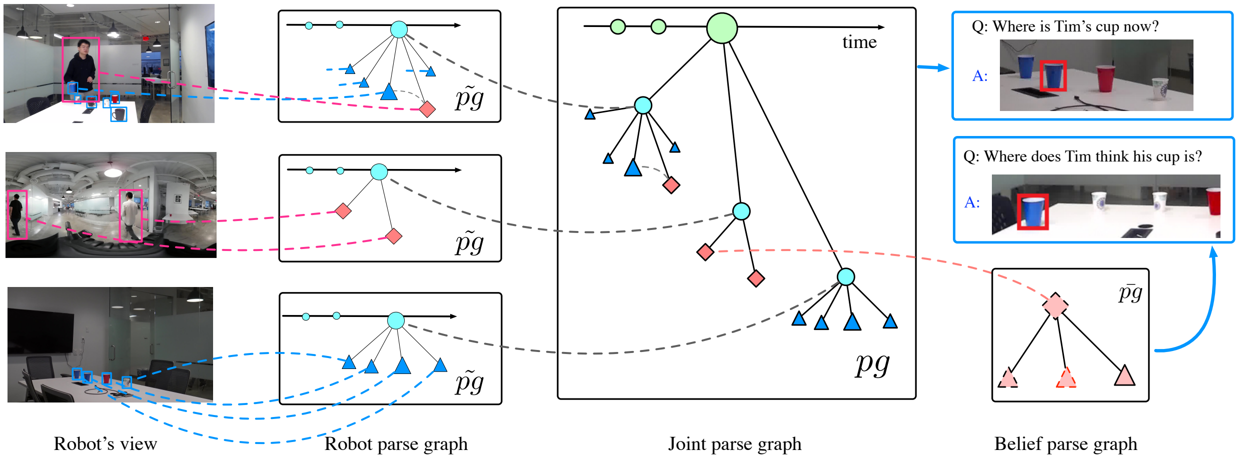


Figure 4.2: System overview. The robot *pg*s are obtained from each individual robot's view. The joint *pg* can be obtained by fusing all robots' *pg*s. The belief *pg*s can be inferred from the joint *pg*. All the *pg*s are optimized simultaneously under the proposed joint parsing framework to enable the queries about the object states and human beliefs.

4.1.3 Overview

The remainder of the chapter is organized as follows. section 4.2 and section 4.3 describes the representation and the detailed probabilistic formulation, respectively. We demonstrate the efficacy of the proposed method in section 4.4 and conclude the chapter with discussions in section 4.5.

4.2 Representation

In this work, we use the parse graph (*pg*)—a unified graphical model [ZM07]—to represent (i) the location of each human and object, (ii) the interaction between human and object, (iii) the beliefs of human, and (iv) the attributes and states of objects; see Figure 4.2 for an example. Specifically, three different types of *pg*s are utilized:

- *Robot pg* shown as blue circles maintains the knowledge structure of an individual robot, which is extracted from its visual observation, *i.e.*, an image. Notice that it also contains attributes that are grounded to the observed agents and objects.
- *Belief pg* shown as red diamonds represents the *inferred* human knowledge by each robot. Each robot maintains the parse graph for each agent it observed.
- *Joint pg* fuses all the information and views across a set of distributed robots.

Notations and Definitions The input of our system can be represented by M synchronized video sequences $I = \{I_{t=1..T}^{k=1..M}\}$ with length T captured from M robots. Formally, a scene $\mathcal{R}$ is expressed as

$$\begin{aligned} \mathcal{R} &= \{(O_t, H_t) : t = 1, 2, \ldots, T\}, \\ O_t &= \{o_t^i : i = 1, 2, \ldots, N_o\}, \\ H_t &= \{h_t^j : j = 1, 2, \ldots, N_h\}, \end{aligned} \tag{4.1}$$

where O_t and H_t denote the set of all the tracked objects (N_o objects in total) and the set of all the tracked agents (N_h agents in total) at time t, respectively.

Object o_t^i is represented by its bounding box location b_t^i, appearance feature ϕ_t^i, states s_t^i, and attributes a_t^i as a tuple,

$$o_t^i = (b_t^i, \phi_t^i, s_t^i, a_t^i), \tag{4.2}$$

where s_t^i is an index function: $s_t^i = j, j \neq 0$ indicates the object o_i is held by the agent h_j at time t, and $s_t^i = 0$ means it is not held by any agent at time t.

The agent h_t^j is represented by its body key points position p_t^i and appearance feature ϕ_t^j

$$h_t^j = (p_t^i, \phi_t^j). \tag{4.3}$$

Robot Parse Graph is formally expressed as

$$\tilde{pg}_t^k = \{(o_t^i, h_t^j) : o_t^i, h_t^j \in I_t^k\}, \tag{4.4}$$

where I_t^k is the area where kth robot can observe at time t.

Belief Parse Graph is formally expressed as

$$\bar{pg}_t^{k,j} = \{o_{t'}^i : o_{t'}^i \in I_{t'}^k\}, \tag{4.5}$$

where a $\bar{pg}_t^{k,j}$ represents the inferred belief of agent h_j under robot k's view; t' is the last time that the robot k observes the human h_j. We assume that the agent h_j only keeps the objects s/he observed last time in this area in mind, which satisfies the *Principle of Inertia*:

an agent's belief is preserved over time unless the agent gets information to the contrary.

Joint Parse Graph keeps track of all the information across a set of distributed robots, formally expressed as

$$pg_t = \{(o_t^i, h_t^j : i = 1, 2, ..., N_o; j = 1, 2, ..., N_h)\}. \tag{4.6}$$

Objective The objective of the system is to jointly infer all the parse graphs $PG = \{pg, \tilde{pg}, \bar{pg}\}$ so that it can (i) track all the agents and objects across scenes at any time by fusing the information collected by a distributed system, and (ii) infer human (false-)beliefs to provide assistance.

4.3 Probabilistic Formulation

We formulate the joint parsing problem as a MAP inference problem

$$\begin{aligned} PG^* &= \arg\max_{PG} p(PG|I) \\ &= \arg\max_{PG} p(I|PG) \cdot p(PG), \end{aligned} \tag{4.7}$$

where $p(PG)$ is the prior, and $p(I|PG)$ is the likelihood.

4.3.1 Prior

The prior term $p(PG)$ models the compatibility of the robot pgs and the joint pg, and the compatibility of the joint pg over time. Formally, we can decompose the prior as

$$p(PG) = p(pg_1) \prod_{t=1}^{T-1} p(pg_{t+1}|pg_t) \prod_{k=1}^{M} \prod_{t=1}^{T} p(\tilde{pg}_t^k|pg_t), \tag{4.8}$$

where the first term $p(pg_{t+1}|pg_t)$ is the transition probability of the joint pg over time, further decomposed as

$$p(pg_{t+1}|pg_t) = \frac{1}{Z} \exp\{-\mathcal{E}(pg_{t+1}|pg_t)\}, \tag{4.9}$$

$$\begin{aligned}\mathcal{E}(pg_{t+1}|pg_t) = &\sum_{i=1}^{N_o} \mathcal{E}_{L_o}(b^i_{t+1}, b^i_t, s^i_t) + \mathcal{E}_{ST}(s^i_{t+1}, s^i_t) \\ &+ \sum_{j=1}^{N_h} \mathcal{E}_{L_h}(p^j_{t+1}, p^j_t).\end{aligned} \tag{4.10}$$

The second term $p(\tilde{pg}^k_t|pg_t)$ is the probability models the compatibility of individual pgs and the joint pg. Its energy can be decomposed into three energy terms

$$\begin{aligned}p(\tilde{pg}^k_t|pg_t) &= \frac{1}{Z}\exp\{-\mathcal{E}(pg_t, \tilde{pg}^k_t)\} \\ &= \frac{1}{Z}\exp\{-\mathcal{E}_A(pg_t, \tilde{pg}^k_t) - \mathcal{E}_S(pg_t, \tilde{pg}^k_t) \\ &\qquad - \mathcal{E}_{Attr}(pg_t, \tilde{pg}^k_t)\}.\end{aligned} \tag{4.11}$$

Below, we detail the above six energy terms in Equation 4.10 and Equation 4.11.

Motion Consistency The term $\mathcal{E}_L$ measures the motion consistency of objects across time and is defined as

$$\begin{aligned}\mathcal{E}_{L_o}(b^i_{t+1}, b^i_t, s^i_t) &= \begin{cases}\delta(\mathcal{D}(b^i_{t+1}, b^i_t) > \tau)) & \text{if } s^i_t = 0 \\ \delta(\mathcal{D}(p^j_{t+1}, p^j_t) > \tau)) & \text{if } s^i_t = j\end{cases} \\ \mathcal{E}_{L_h}(p^j_{t+1}, p^j_t) &= \delta(\mathcal{D}(p^j_{t+1}, p^j_t) > \tau)),\end{aligned} \tag{4.12}$$

where $\mathcal{D}$ is the distance between two bounding boxes or human poses, τ is the speed threshold, and δ is the indicator function. If an object i is held by human j, we use the human's location to calculate $\mathcal{E}_L$ of the object.

State Transition Consistency The term $\mathcal{E}_{ST}$ is the state transition energy and is defined as

$$\mathcal{E}_{ST}(s^i_{t+1}, s^i_t) = -\log p(\delta(s^i_{t+1} = 0)|\delta(s^i_t = 0)), \tag{4.13}$$

where the state transition probability $p(\delta(s^i_{t+1} = 0)|\delta(s^i_t = 0))$ is learned from the training data.

Appearance Consistency $\mathcal{E}_A$ measures appearance consistency. In robot pgs, the appearance feature vector ϕ is extract by a deep person re-identification network [ZYC19]. In the joint pg, the feature vector is calculated by the mean pooling of all features for the same entity from all robot pgs

$$\mathcal{E}_A(pg_t, \tilde{pg}_t) = \sum_{e \in O_t \cup H_t} ||\phi_t^e - \phi_t^{\tilde{e}}||_2. \tag{4.14}$$

Spatial Consistency Each object and agent in robot's view should also have a corresponding location in the real-world coordinate system; *i.e.*, the bounding box of the object or key points position of the agent should remain consistent when projected from robot image plane back to the real-world coordinate. $\mathcal{E}_S$ captures such a spatial consistency

$$\mathcal{E}_S(pg_t, \tilde{pg}_t) = \sum_{e \in O_t \cup H_t} ||p_t^e - f(p_t^{\tilde{e}})||_2, \tag{4.15}$$

where p represents the real-world coordinate of s, and f is the project function from the robot's view to the real world, obtained by the intrinsic and extrinsic camera parameters of the robot's camera.

Attribute Consistency Attributes of each entity should remain the same across different time and views. The term $\mathcal{E}_{Attr}$ models such an attribute consistency of all the objects

$$\mathcal{E}_{Attr}(pg_t, \tilde{pg}_t) = \sum_{i=1}^{N_o} \delta(a_t^i \neq \tilde{a}_t^i). \tag{4.16}$$

4.3.2 Likelihood

The likelihood term $p(I|PG)$ models how well each robot can ground the knowledge in its pg to the visual data it captures. Formally, the likelihood is defined as

$$p(I|PG) = \prod_{k=1}^{M} \prod_{t=1}^{T} p(I_t^k | \tilde{pg}_t^k). \tag{4.17}$$

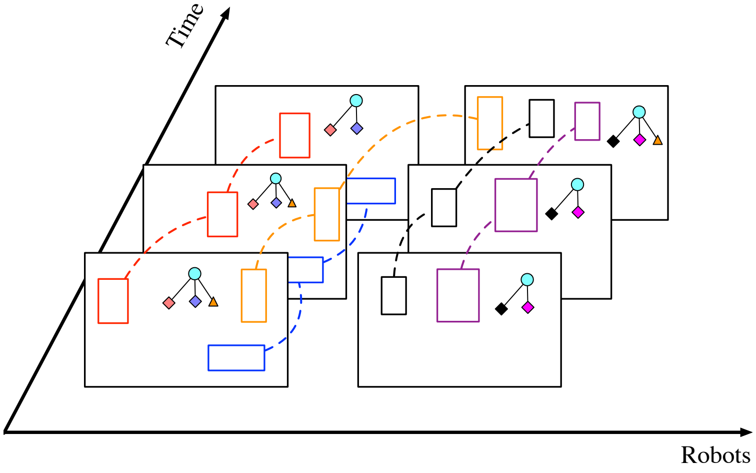
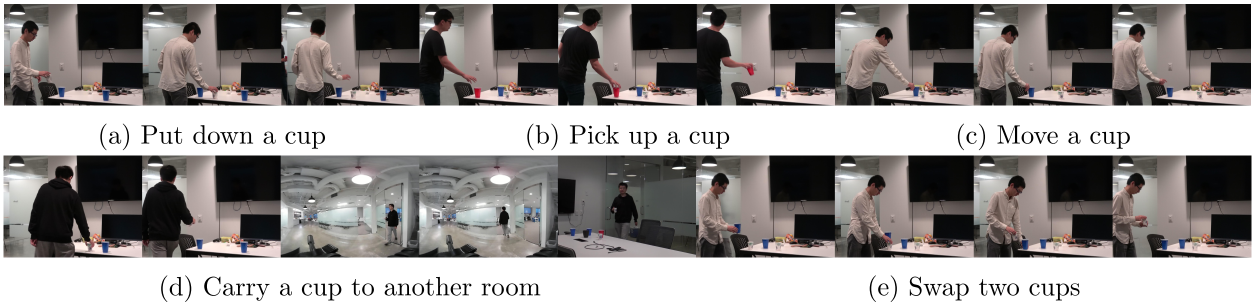


Figure 4.3: Illustration of the inference process. The maximizing a posterior (MAP) estimate can be transformed as an assignment problem that links initial proposals over time and different views.



(a) Put down a cup

(b) Pick up a cup

(c) Move a cup

(d) Carry a cup to another room

(e) Swap two cups

Figure 4.4: Examples of human-object interactions in the cross-view subset of the proposed dataset. Each scenario contains at least one kind of false-belief test or helping test recorded with four robot camera views.

The energy of term $p(I_t^k|\tilde{pg}_t^k)$ can be further decomposed as

$$p(I_t^k|\tilde{pg}_t^k) = \frac{1}{Z}\exp\{-\mathcal{E}(I_t^k|\tilde{pg}_t^k)\}, \tag{4.18}$$

$$\begin{aligned}\mathcal{E}(I_t^k|\tilde{pg}_t^k) = &\sum_{i=1}^{N_o}\mathcal{E}_D(b_t^i,\phi_t^i) + \mathcal{E}_{CA}(b_t^i,\phi_t^i,a_t^i)\\ &+\sum_{j=1}^{N_h}\mathcal{E}_D(p_t^j,\phi_t^j),\end{aligned} \tag{4.19}$$

where $\mathcal{E}_D$ can be calculated by the score of object detection or human pose estimation, and $\mathcal{E}_{CA}$ can be obtained by the object attributes classification scores.

4.3.3 Inference

Given the above probabilistic formulation, we can infer the best $\{pg^*, \tilde{pg}^*\}$ by an MAP estimate. It can be solved by two steps: (i) Each robot individually processes the frame image it received with vision modules, such as object detection and human pose estimation; the raw results can be used as the proposals for the second step. (ii) The MAP estimate can be transformed to an assignment problem given the proposals; see Figure 4.3. We can solve this assignment problem using the *Kuhn-Munkres* algorithm [Kuh55, Kuh56] in polynomial time.

Based on Equation 4.5, after we have the robot parse graphs $\tilde{pg}^k$ for robot k, we can generate belief parse graphs $\bar{pg}^{k,j}$ for human j in robot k's view.

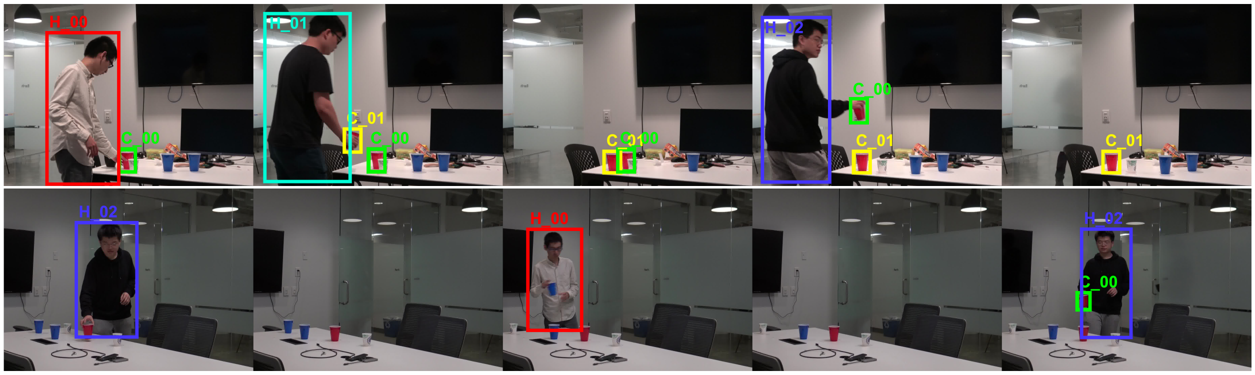
4.4 Experiment

We evaluate the proposed method in cross-view object tracking and human (false-)belief understanding with two experiments. The first experiment evaluates the accuracy of object localization by the inference algorithms using the robot parse graphs $\tilde{pg}$ and the joint parse graph pg. The second experiment evaluates the inference of the belief parse graphs $\bar{pg}$, *i.e.*, human beliefs regard the object states (*e.g.*, locations) in both single-view and multi-view settings.

4.4.1 Dataset

The dataset includes two subsets, a multi-view subset and a single-view subset.

- The single-view subset includes 5 different false-belief scenarios with 12532 frames. Each scenario contains at least one kind of false-belief test or helping test. In this subset, objects are not limited to the cups.
- The multi-view subset consists of 8 scenes, each shot with 4 robot camera views, making a total number of 72720 frames. Each scenario contains at least one kind of false-belief test. The objects in each scene are, however, limited to the cups: 12-16 different cups made with 3 different materials (plastic, paper, and ceramic) and 4 colors (red, blue, white, and



(a) Observation t_0 (b) Observation t_1 (c) Observation t_2 (d) Observation t_3 (e) Observation t_4

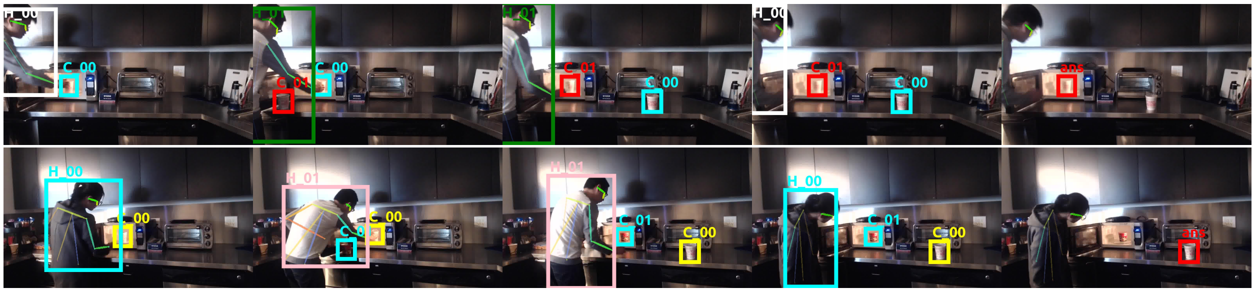
Figure 4.5: An example of tracking multiple objects across multiple views with human interactions in Experiment 1. Two rows shows sample synchronized visual inputs from two different rooms: (Top) Room 1, and (Bottom) Room 2. Here, for better interpretations, we only highlight two red cups marked by green and yellow bounding boxes. At time t_0, agent H_{00} put a red cup C_{00} in Room 1. At time t_1, agent H_{01} put another red cup C_{01} in Room 1. At time t_3, human H_{02} took away cup C_{00} and put it in Room 2 at time t_4. Our system can not only robustly perform such complex multi-room multi-view tracking, but also is able to reason about agent's belief. For instance, at time t_4, the tracking system knows that the cup C_{00} appear at time t_0 is in Room 2 and infers that human H_{00} thinks the cup is still in Room 1.

black). In each scene, three agents interact with cups by performing following actions (not necessarily in this specific action sequence): (i) picking up a cup, (ii) putting down a cup on the table, (iii) moving a cup, (iv) carrying a cup to another room, and (v) swapping a cup in hand and a cup on the table; see Figure 4.4 for some examples. Ground-truth tracking results of cups and agents, states of cups, and attributes of cups are all annotated in this dataset.

4.4.2 Implementation Details

Below, we provide some details about the specific implementations of the system.

- Object detection: we use the RetinaNet model [LGG17] pre-trained on the MS COCO dataset [LMB14]. We keep all the bounding boxes with a score higher than the threshold 0.2; these bounding boxes serve as the proposals for object detection. For the single-view subset, which includes more categories than the multi-view subset, we further fine-tuned the pre-trained model on the training set since some object categories are not included in MS COCO dataset.



(a) Observation t_0 (b) Observation t_1 (c) Observation t_2 (d) Observation t_3 (e) Prediction

Figure 4.6: Two sample results of the Sally-Anne false-belief task in Experiment 2. (Top) with false-belief. (Bottom) without false-belief. (a) The first person (H_{00}) puts their cup noodle (C_{00}) into the microwave and leaves. (b) The second person (H_{01}) takes out the cup noodle (C_{00}) of the first person from the microwave. (c) The second person puts their own cup noodle (C_{01}) in the microwave and leaves. (d) The first person (H_{00}) returns to the room. Question: where does the first person (H_{00}) think their cup noodle (C_{00}) is? (e) Our system can successfully predict the person in the bottom row will not have the false-belief due to the different color attributes of the cups.

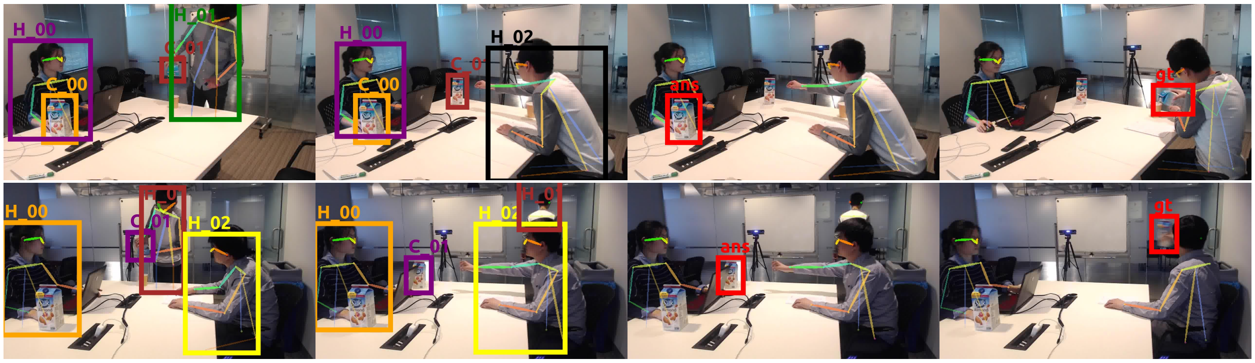
- Human pose estimation: we apply the AlphaPose [FXT17].
- Object attribute classification: A VGG16 network was trained to classify the color and the material of the objects.
- Appearance feature: A deep person re-id model [ZYC19] was fine-tuned on the training set.
- Due to the lack of multi-view in the single-view setting, we locate the object that an agent plan to interact by simply finding the object closest to the direction the agent points at according to the keypoints on the arm.

4.4.3 Experiment 1: Cross-view Object Localization

To test the overall cross-view tracking performance, 2000 queries are randomly sampled from the ground-truth tracks. Each query q can be formally described as

$$q = (k, t, b, t_q), \tag{4.20}$$

where the first three terms (k, t, b) indicate the cup shown in robot k's view located in bounding box b at time t. Such a form of the query can be very flexible. For instance, if we ask about the location of that cup at time t_q, the system should return an answer in the



(a) Observation t_0 (b) Observation t_1 (c) Prediction (d) Ground-truth

Figure 4.7: Two sample results of the helping task in Experiment 2. (Top) with false-belief. (Bottom) without false-belief. (a) The second person (H_{01}) enters the room and finds one box (C_{01}) is empty and then leaves. (b) Later, the third person (H_{02}) wants to get a box. Question: which box should the first person (H_{00}) give to the third person (H_{02}) if the first person (H_{00}) wants to help? (c) Answer returned by the system that correctly infers the third person (H_{02}) on the top row has false-belief, thus the first person (H_{00}) should not give the empty box (C_{01}) that the third person (H_{02}) is pointing to; rather, the first person (H_{00}) should give another box (C_{00}). Conversely, since the person (H_{02}) observes the entire process in the bottom row, there is no false-belief; in this case, the person (H_{02}) is trying to throw away the empty box (C_{01}).

form of (k_a, b_a), meaning that the system predicts the cup is shown in robot k_a's view at b_a.

The system generates the answer in two steps. It firstly locates the query of the object by searching the object i in $\tilde{pg}_t^k$ with the smallest distance to the bounding box b. Then it returns the location $b_{t_q}^i$ from $\tilde{pg}_{t_q}^{k'}$. The accuracy of model M can be calculated as

$$acc(M) = \frac{1}{N_q} \sum_{i=1}^{N_q} \delta(\mathrm{IoU}(b_{gt}^i, b_a^i) > \xi) \cdot \delta(k_{gt} = k_a), \tag{4.21}$$

where N_q is the number of queries, b_{gt} is the ground-truth bounding box, and b_a is the inferred bounding box returned by model M. We calculate the Intersection over Union (IoU) between the answer and the ground-truth bounding boxes; the answer is correct if and only if the answer predicts the right view and the IoU is larger than $\xi = 0.5$.

Table 4.1 shows the ablative study by turning on and off the joint parsing component that models human interactions, *i.e.*, whether the model parses and tracks objects by reasoning about the interaction with agents. #interactions means how many times the object was interacted by agents. The result shows that our system achieves an overall 88% accuracy.

Table 4.1: Accuracy of cross-view object tracking

# interactions	0	1	2	3	Overall
Parsing w/o humans acc.	0.98	0.82	0.78	0.75	0.82
Joint parsing acc.	0.98	0.86	0.85	0.82	0.88

Table 4.2: Accuracy of belief queries on single view subset

	True Belief	False-Belief	Overall
Joint parsing acc.	0.94	0.93	0.94
Random guessing acc.	0.45	0.53	0.46

Even without parsing humans, our system still possesses the ability to reason about object location by maintaining other consistencies, such as spatial consistency and appearance consistency. However, its performance drops significantly if the object was moved to different rooms. Figure 4.5 shows some qualitative results.

4.4.4 Experiment 2: (False-)Belief Inference

In this experiment, we evaluate the performance of belief and false-belief inference, *i.e.*, whether an agent's belief pg is the same as the true object states. The evaluations were conducted on both single-view and multi-view scenarios.

Multi-view For the multi-view setting, we collect 200 queries by annotators focused on the Sally-Anne false belief task in the form

$$q = (k_o, t_o, b_o, k_h, t_h, b_h, t_q). \tag{4.22}$$

The first three terms (k_o, t_o, b_o) define the objects in robot k_o's view located at b_o at time t_o. Similarly, another three terms (k_h, t_h, b_h) define an agent in robot k_h's view located at b_h at time t_h. The question is: where does the agent (k_h, t_h, b_h) think the object (k_o, t_o, b_o) is at time t_q?

For this task, our system generates the answer in three step. The first step is to search the object i and the agent j in robot parse graphs $\tilde{pg}_{t_o}^{k_o}$ and $\tilde{pg}_{t_h}^{k_h}$. In the second step, it retrieves all the belief parse graphs $\bar{pg}_{t_q}^{k',j}$ at time t_q to find the object i's location $\bar{b}_{t_q}^{i}$ in human j's belief. In the last step, the system finds an object i' in robot parse graph which has the same attributes as i's and has smallest distance to $\bar{b}^i_{t_q}$; the system finally returns i''s location $b^{i'}_{t_q}$ as the answer.

Since there is no publicly available code on this task, we compare our inference algorithm with a random baseline model as the reference for future benchmark; it simply returns an object with the same attributes as the query object at t_q. The result shows that our system achieves 81% accuracy while the baseline model only has 39% accuracy.

Single-view For the single-view setting, We collect 100 queries in total, including two types of belief inference task: the Sally-Anne false-belief task and the helping task as shown in Figure 4.1. The queries have two forms

$$q = (t_o, b_o, b_h, t_q), \tag{4.23}$$

$$q = (b_h, t_q), \tag{4.24}$$

indicating two different types of questions: (i) where does the agent b_h think the object (t_o, b_o) is at time t_q? and (ii) which object will you give to the agent (t_q, b_h) at time t_q if you would like to help? For the first type of questions, *i.e.*, the Sally-Anne false-belief task, similar to the multi-view setting, the system should return the object bounding box as the answer. For the second types of questions, *i.e.*, the helping task, the system first infers whether the agent has false-belief. If not, the system returns the object the person wants to interact based on their current pose; otherwise, the system returns another suitable object closest to them. Qualitative results are shown in Figure 4.6 and Figure 4.7, and quantitative results are provided in Table 4.2.

4.5 Conclusion and Discussions

In this work, we describe the idea of using *pg* as a unified representation for tracking object states, accumulating robot knowledge, and reasoning about human (false-)beliefs. From the spatiotemporal information observed from multiple camera views of one or more robots, robot *pg* and belief *pg* are induced and can be merged to a joint *pg* to facilitate more advanced reasoning and inference. Based on this representation, a joint inference algorithm is proposed, which has the capabilities of tracking small occluded objects across different views and infer human beliefs and false-beliefs. In experiments, we first demonstrate that the joint inference over the merged *pg* produced better tracking accuracy. We further evaluate the inference on human true and false-belief regarding objects' locations by jointly parsing the *pg*s. The high recognition accuracy demonstrates that our system is capable of modeling and understanding human (false-)beliefs, with the potential of helping capability as demonstrated in developmental psychology.

ToM and Sally-Anne test are interesting and difficult problems in the area of social robotics. For a service robot to interact with humans in an intuitive manner, it must be able to maintain a model of the belief states of the agents it interacts with. We hope the proposed method using a graphical model has demonstrated a different perspective compared to prior methods in terms of the flexibility and generalization. In future, a more interactive and active set up would be more practical and compelling. For instance, by integrating activity recognition modules, our system should be able to perceive, recognize, and extract richer semantic information from the observed visual input, thereby providing more subtle (false-)belief applications. Communication, gazes, and gestures are also crucial in intention expression and perception in collaborative interactions. By incorporating these essential ingredients and taking the advantage of the flexibility and generalization of the model, our system should be able to go from the current passive query to active response and helping in real-time.

CHAPTER 5

Conclusion

In this dissertation, we propose a cognition platform for joint inference of 3D Geometry, Object States, and Human Belief under single-view or multi-view scenarios.

At the engineering level, we build a hierarchical container-based system that can process image sequences from various types of devices. With the isolation, compatibility, and scalability of the system, it supports rapid development and can be dynamically deployed on different machines according to compute resources and requests. The web-based visualization module can show the results realtime on user browsers.

We jointly solve the 3D scene reconstruction and 3D human pose estimation for holistic++ scene understanding. Human-object interaction and physical commonsense are incorporated in the algorithm. And our MCMC inference algorithm can effectively optimize the scene configuration. The performance of the two tasks gets significantly improved on several datasets.

Under the cross-view setting, we propose the scene-centric parsing framework to utilize the appearance and geometry correlations. Experiments show that the joint parsing framework can produce more accurate results in various computer vision tasks. With the parse graphs, our system can also provide an interpretable interface to users.

We also describe the idea of using parse graphs as a unified representation for tracking object states, accumulating robot knowledge, and reasoning about human belief. Based on the representation, a joint inference algorithm is proposed. The experiments demonstrate that our system is capable of modeling and understanding human belief.

In the future, we hope more and more cognition functions can be integrated into the platform by taking advantage of the flexibility and generalization of our platform. So that our platform can extract richer semantic information form the visual input, and support more complex applications with the flexibility and generalization of our algorithms.

REFERENCES

[AAL15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "VQA: Visual Question Answering." In *IEEE International Conference on Computer Vision*, 2015.

[ABY16] S. Aditya, C. Baral, Y. Yang, Y. Aloimonos, and C. Fermuller. "DeepIU: An Architecture for Image Understanding." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[Bai04] Renée Baillargeon. "Infants' physical world." *Current directions in psychological science*, **13**(3):89–94, 2004.

[BBP16] Torben Braüner, Patrick Blackburn, and Irina Polyanskaya. "Second-order false-belief tasks: Analysis and formalization." In *International Workshop on Logic, Language, Information, and Computation*, 2016.

[BCT09] David Buttelmann, Malinda Carpenter, and Michael Tomasello. "Eighteen-month-old infants show false belief understanding in an active helping paradigm." *Cognition*, **112**(2):337–342, 2009.

[BFT11a] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. "multiple object tracking using K-Shortest Paths optimization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(9):1806–1819, 2011.

[BFT11b] 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.

[BGB09] Cynthia Breazeal, Jesse Gray, and Matt Berlin. "An embodied cognition approach to mindreading skills for socially intelligent robots." *International Journal of Robotics Research*, **28**(5):656–680, 2009.

[BLF85] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. "Does the autistic child have a "theory of mind"?" *Cognition*, **21**(1):37–46, 1985.

[BM14] Or Biran and Kathleen McKeown. "Justification narratives for individual classifications." In *IEEE International Conference on Machine Learning Workshops*, 2014.

[Bol18] Thomas Bolander. "Seeing Is Believing: Formalising False-Belief Tasks in Dynamic Epistemic Logic." In *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pp. 207–236. Springer, 2018.

[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.

[BSW85] Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. "Object permanence in five-month-old infants." *Cognition*, **20**(3):191–208, 1985.

[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.

[CLL18] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. "Learning to detect human-object interactions." 2018.

[CLO16] Jungchan Cho, Minsik Lee, and Songhwai Oh. "Complex non-rigid 3d shape recovery using a procrustean normal distribution mixture model." *International Journal of Computer Vision (IJCV)*, **117**(3):226–246, 2016.

[CLV06] Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. "Building explainable artificial intelligence systems." In *AAAI Conference on Artificial Intelligence*, 2006.

[CSP16] Haoyu Chen, Daniel Seita, Xinlei Pan, and John Canny. "An Efficient Minibatch Acceptance Test for Metropolis-Hastings." *arXiv preprint arXiv:1610.06848*, 2016.

[CSW17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[CT99] Josep Call and Michael Tomasello. "A nonverbal false belief task: The performance of children and great apes." *Child development*, **70**(2):381–395, 1999.

[DA16] Sandra Devin and Rachid Alami. "An implemented theory of mind to improve human-robot shared plans execution." In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[DLB18] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. "Learning to Exploit Stability for 3D Scene Parsing." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[DSY17] Xingping Dong, Jianbing Shen, Dajiang Yu, Wenguan Wang, Jianhong Liu, and Hua Huang. "Occlusion-aware real-time object tracking." *IEEE Transactions on Multimedia*, **19**(4):763–771, 2017.

[DTT15] Afshin Dehghan, Yicong Tian, Philip HS Torr, and Mubarak Shah. "Target identity-aware network flow for online multiple target tracking." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.

[FBL08] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. "Multi-camera people tracking with a probabilistic occupancy map." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(2):267–282, 2008.

[FC03] Lisa Feigenson and Susan Carey. "Tracking individuals via object-files: evidence from infants' manual search." *Developmental Science*, **6**(5):568–584, 2003.

[FCW18] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. "Inferring Shared Attention in Social Scene Videos." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[FXT17] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. "Rmpe: Regional multi-person pose estimation." In *International Conference on Computer Vision (ICCV)*, 2017.

[FXW18] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. "Learning pose grammar to encode human body configuration for 3D pose estimation." In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[GA88] Alison Gopnik and Janet W Astington. "Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction." *Child development*, pp. 26–37, 1988.

[GBK02] György Gergely, Harold Bekkering, and Ildikó Király. "Developmental psychology: Rational imitation in preverbal infants." *Nature*, **415**(6873):755, 2002.

[GGH15] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. "Visual Turing test for computer vision systems." *Proceedings of the National Academy of Sciences*, **112**(12):3618–3623, 2015.

[Gib79] James Jerome Gibson. *The ecological approach to visual perception.* Houghton, Mifflin and Company, 1979.

[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.

[HAR16] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. "Generating visual explanations." In *European Conference on Computer Vision*, 2016.

[Has70] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, **57**(1):97–109, 1970.

[HQX18] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. "Cooperative Holistic Scene Understanding: Unifying 3D Object, Layout, and Camera Pose Estimation." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[HQZ18] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. "Holistic 3D scene parsing and reconstruction from a single RGB image." In *European Conference on Computer Vision (ECCV)*, 2018.

[HWR13a] M. Hofmann, D. Wolf, and G. Rigoll. "Hypergraphs for joint multi-view reconstruction and multi-object tracking." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[HWR13b] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. "Hypergraphs for joint multi-view reconstruction and multi-object tracking." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[HZ09] F. Han and S.C. Zhu. "Bottom-up/top-down Image Parsing with Attribute Grammar." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(1):59–73, 2009.

[ISS17] Hamid Izadinia, Qi Shan, and Steven M Seitz. "Im2cad." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[JQZ18] Chenfanfu Jiang, Siyuan Qi, Yixin Zhu, Siyuan Huang, Jenny Lin, Lap-Fai Yu, Demetri Terzopoulos, and Song-Chun Zhu. "Configurable 3D Scene Synthesis and 2D Image Rendering with Per-pixel Ground Truth Using Stochastic Grammars." *International Journal of Computer Vision (IJCV)*, **126**(9):920–941, 2018.

[JSL17] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, et al. "Panoptic studio: A massively multiview system for social interaction capture." *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[KHL17] James R Kubricht, Keith J Holyoak, and Hongjing Lu. "Intuitive physics: Current research and controversies." *Trends in cognitive sciences*, **21**(10):749–759, 2017.

[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.

[KS83] Philip J Kellman and Elizabeth S Spelke. "Perception of partly occluded objects in infancy." *Cognitive psychology*, **15**(4):483–524, 1983.

[KS16] Hema S Koppula and Ashutosh Saxena. "Anticipating human activities using object affordances for reactive robotic response." *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **38**(1):14–29, 2016.

[Kuh55] Harold W Kuhn. "The Hungarian method for the assignment problem." *Naval research logistics quarterly*, **2**(1-2):83–97, 1955.

[Kuh56] Harold W Kuhn. "Variants of the Hungarian method for assignment problems." *Naval Research Logistics Quarterly*, **3**(4):253–258, 1956.

[KZG17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.” *International Journal on Computer Vision*, **123**(1):32–73, 2017.

[LC14] Sijin Li and Antoni B Chan. “3D human pose estimation from monocular images with deep convolutional neural network.” In *Asian Conference on Computer Vision (ACCV)*, 2014.

[LCC12] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. “Explaining robot actions.” In *ACM/IEEE International Conference on Human-Robot Interaction*, 2012.

[LCV05] H Chad Lane, Mark G Core, Michael Van Lent, Steve Solomon, and Dave Gomboc. “Explainable artificial intelligence for training and tutoring.” Technical report, Defense Technical Information Center, 2005.

[LGG17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. “Focal Loss for Dense Object Detection.” In *International Conference on Computer Vision (ICCV)*, 2017.

[LHL17] Martin Liggins II, David Hall, and James Llinas. *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[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.

[LPR12] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. “Branch-and-price global optimization for multi-view multi-object tracking.” In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[LR19] Emiliano Lorini and Fabian Romero. “Decision Procedures for Epistemic Logic Exploiting Belief Bases.” In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2019.

[LS04] Hugo Liu and Push Singh. “ConceptNet—a practical commonsense reasoning tool-kit.” *BT technology journal*, **22**(4):211–226, 2004.

[LS10a] Joerg Liebelt and Cordelia Schmid. “Multi-view object class detection with a 3D geometric model.” In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[LS10b] Joerg Liebelt and Cordelia Schmid. “Multi-view object class detection with a 3d geometric model.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[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.

[LZZ14] Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. "Single-View 3D Scene Parsing by Attributed Grammar." In *IEEE Conference Computer Vision and Pattern Recognition*, 2014.

[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.

[MBR16] Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. "Newtonian scene understanding: Unfolding the dynamics of objects in static images." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[MLZ16] Rui Ma, Honghua Li, Changqing Zou, Zicheng Liao, Xin Tong, and Hao Zhang. "Action-driven 3D indoor scene evolution." *ACM Transactions on Graphics (TOG)*, **35**(6):173–1, 2016.

[MRR53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics*, **21(6)**:1087–1092, 1953.

[MSS17] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. "Vnect: Real-time 3d human pose estimation with a single rgb camera." *ACM Transactions on Graphics (TOG)*, **36**(4):44, 2017.

[MWC14] Grégoire Milliez, Matthieu Warnier, Aurélie Clodic, and Rachid Alami. "A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management." In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2014.

[Nee97] Amy Needham. "Factors affecting infants' use of featural information in object segregation." *Current Directions in Psychological Science*, **6**(2):26–33, 1997.

[PBH13] L.D. Pero, J. Bowdish, E. Hartley, B. Kermgard, and K. Barnard. "Understanding bayesian rooms using composite 3D object models." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[PNZ16] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. "Attribute And-Or Grammar for Joint Parsing of Human Attributes, Part and Pose." *arXiv preprint arXiv:1605.02112*, 2016.

[PRF11] H. Pirsiavash, D. Ramanan, and C. Fowlkes. "Globally-optimal greedy algorithms for tracking a variable number of objects." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[PRG18] Beate Priewasser, Eva Rafetseder, Carina Gargitter, and Josef Perner. "Helping as an early indicator of a theory of mind: Mentalism or Teleology?" *Cognitive Development*, **46**:69–78, 2018.

[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.

[QWJ18] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. "Learning human-object interactions by graph parsing neural networks." In *European Conference on Computer Vision (ECCV)*, 2018.

[QWL15] Hang Qi, Tianfu Wu, Mun-Wai Lee, and Song-Chun Zhu. "A Restricted Visual Turing Test for Deep Scene and Event Understanding." *arXiv preprint arXiv:1512.01715*, 2015.

[QXY18] Hang Qi, Yuanlu Xu, Tao Yuan, Tianfu Wu, and Song-Chun Zhu. "Scene-centric Joint Parsing of Cross-view Videos." In *AAAI Conference on Artificial Intelligence (AAAI)*, 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.

[RHG15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In *Annual Conference on Neural Information Processing Systems*, 2015.

[RKS12] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. "Reconstructing 3d human pose from 2d image landmarks." In *European Conference on Computer Vision (ECCV)*, 2012.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

[Sar14] Olivia N Saracho. "Theory of mind: children's understanding of mental states." *Early Child Development and Care*, **184**(6):949–961, 2014.

[Sca02] Brian Scassellati. "Theory of mind for a humanoid robot." *Autonomous Robots*, **12**(1):13–24, 2002.

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

[SCH16] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. "PiGraphs: Learning Interaction Snapshots from Observations." *ACM Transactions on Graphics (TOG)*, **35**(4), 2016.

[SCS13] Amy E Skerry, Susan E Carey, and Elizabeth S Spelke. "First-person action experience reveals sensitivity to action efficiency in prereaching infants." *Proceedings of the National Academy of Sciences (PNAS)*, 2013.

[SF15] Aimee E Stahl and Lisa Feigenson. "Observing the unexpected enhances infants' learning and exploration." *Science*, **348**(6230):91–94, 2015.

[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.

[SK07] Elizabeth S Spelke and Katherine D Kinzler. "Core knowledge." *Developmental Science*, **10**(1):89–96, 2007.

[SLX15] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. "Sun rgb-d: A rgb-d scene understanding benchmark suite." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[SRA12] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. "Single image 3D human pose estimation from noisy observations." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[SX14] Shuran Song and Jianxiong Xiao. "Sliding shapes for 3d object detection in depth images." In *European Conference on Computer Vision (ECCV)*, 2014.

[SYZ17] 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.

[TBB09] Moritz Tenorth, Jan Bandouch, and Michael Beetz. "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition." In *IEEE International Conference on Computer Vision Workshop*, 2009.

[THC16] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. "Computational human-robot interaction." *Foundations and Trends® in Robotics*, **4**(2-3):105–223, 2016.

[THK87] Nancy Termine, Timothy Hrynick, Roberta Kestenbaum, Henry Gleitman, and Elizabeth S Spelke. "Perceptual completion of surfaces in infancy." *Journal of Experimental Psychology: Human Perception and Performance*, **13**(4):524, 1987.

[TML14] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. "Joint Video and Text Parsing for Understanding Events and Answering Queries." *IEEE MultiMedia*, **21**(2):42–70, 2014.

[Tom18] Michael Tomasello. "How children come to understand false beliefs: A shared intentionality account." *Proceedings of the National Academy of Sciences (PNAS)*, **115**(34):8491–8498, 2018.

[TRA17] Denis Tome, Christopher Russell, and Lourdes Agapito. "Lifting from the deep: Convolutional 3d pose estimation from a single image." *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[UB11a] A. Utasi and C. Benedek. "A 3-D marked point process model for multi-view people detection." In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[UB11b] Akos Utasi and Csaba Benedek. "A 3-D marked point process model for multi-view people detection." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[VFM04] Michael Van Lent, William Fisher, and Michael Mancuso. "An explainable artificial intelligence system for small-unit tactical behavior." In *National Conference on Artificial Intelligence*, 2004.

[WAS18] Yang Wu, Jennah A Haque, and Laura Schulz. "Children can use others' emotional expressions to infer their knowledge and predict their behaviors in classic false belief tasks." In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2018.

[WGL12] Mathieu Warnier, Julien Guitton, Séverin Lemaignan, and Rachid Alami. "When the robot puts itself in your shoes. managing and exploiting human and robot beliefs." In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2012.

[WLS18] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. "Where and why are they looking? jointly inferring human attention and intentions in complex tasks." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[WLY14] Longyin Wen, Wenbo Li, Junjie Yan, Zhen Lei, Dong Yi, and Stan Z Li. "Multiple target tracking based on undirected hierarchical relation hypergraph." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[WNX14] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. "Cross-view Action Modeling, Learning and Recognition." In *IEEE Confererence on Computer Vision and Pattern Recognition*, 2014.

[Woo99] Amanda L Woodward. "infants' ability to distinguish between purposeful and non-purposeful behaviors." *Infant Behavior and Development*, **22**(2):145–160, 1999.

[WP83] Heinz Wimmer and Josef Perner. "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception." *Cognition*, **13**(1):103–128, 1983.

[WXL16] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. "Single image 3d interpreter network." In *European Conference on Computer Vision (ECCV)*, 2016.

[WYL15] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[WZS15] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. "Watch-n-patch: Unsupervised understanding of actions and relations." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[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.

[WZZ16] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. "Modeling 4D Human-Object Interactions for Joint Event Segmentation, Recognition, and Object Localization.", 2016.

[XLL16] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. "Multi-view people tracking via hierarchical trajectory composition." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[XLQ17] Yuanlu Xu, Xiaobai Liu, Lei Qin, and Song-Chun Zhu. "Multi-view people tracking via hierarchical trajectory composition." In *AAAI Conference on Artificial Intelligence*, 2017.

[XLZ13] Yuanlu Xu, Liang Lin, Wei-Shi Zheng, and Xiaobai Liu. "Human Re-identification by Matching Compositional Template with Cluster Sampling." In *IEEE International Conference on Computer Vision*, 2013.

[XMH14] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. "Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness." In *ACM Multimedia Conference*, 2014.

[YJK11] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. "Human action recognition by learning bases of action attributes and parts." In *International Conference on Computer Vision (ICCV)*, 2011.

[YYL10] Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. "I2t: Image parsing to text description." *Proceedings of the IEEE*, **98**(8):1485–1508, 2010.

[ZGB16] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. "Visual7W: Grounded Question Answering in Images." In *Advances of Cognitive Systems*, 2016.

[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.

[ZLH17] Chuhang Zou, Zhizhong Li, and Derek Hoiem. "Complete 3D Scene Parsing from Single RGBD Image." *arXiv preprint arXiv:1710.09490*, 2017.

[ZLX14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. "Learning deep features for scene recognition using places database." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[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.

[ZMS18] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. "Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes–The Importance of Multiple Scene Constraints." In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[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.

[ZWM17] Ruiqi Zhao, Yan Wang, and AM Martinez. "A Simple, Fast and Highly-Accurate Algorithm to Recover 3D Shape from 2D Landmarks on a Single Image." *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **40**(12):3059–3066, 2017.

[ZYC19] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. "Omni-Scale Feature Learning for Person Re-Identification." *arXiv preprint arXiv:1905.00953*, 2019.

[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.

[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.

[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)*, 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.