

UCLA

UCLA Electronic Theses and Dissertations

Title

Holistic Scene Understanding and Goal-directed Multi-agent Event Parsing

Permalink

<https://escholarship.org/uc/item/3bb429zf>

Author

Chen, Yixin

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Holistic Scene Understanding and Goal-directed Multi-agent Event Parsing

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

by

Yixin Chen

2022

© Copyright by

Yixin Chen

2022

ABSTRACT OF THE DISSERTATION

Holistic Scene Understanding and Goal-directed Multi-agent Event Parsing

by

Yixin Chen

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2022

Professor Song-Chun Zhu, Chair

Humans, even young infants, are adept at perceiving and understanding complex indoor scenes and events. Holistic scene understanding involves abundant aspects, including 3D human pose, objects, physical relations, functionality, etc. Besides the physical and functional configuration of the scene, interpreting human actions and goal-oriented tasks is a higher-level goal, and requires reasoning about the complex structures in activities along the temporal dimension. When multiple people are in the scene, collaborations and communications inevitably happen, in both verbal and non-verbal forms. Despite the recent remarkable progress in artificial intelligence, building an intelligent machine with human-like perception and reasoning capability for the aforementioned complex tasks remains a significant and challenging problem.

In this dissertation, we study the holistic scene understanding and goal-directed multi-agent event parsing by identifying the critical problems from various perspectives. We first propose a framework for holistic 3D scene parsing and human pose estimation, with a particular focus on human-object interaction and physical commonsense reasoning. Contact information is critical in modeling the fine-grained human-object relations from visual cues.

We demonstrate how to extract meaningful contact information from 2D images and its usefulness in 3D human pose estimation. Then we introduce our efforts in understanding goal-directed actions, concurrent multi-tasks, and collaborations among multi-agents. Finally, we investigate the two typical types of human communications by proposing a spatial and temporal model for shared attention and examining the power of both language and gesture under the embodied reference setting.

The dissertation of Yixin Chen is approved.

Ying Nian Wu

Demetri Terzopoulos

Hongjing Lu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2022

To those who are always curious to learn, understand, and experience.

TABLE OF CONTENTS

1	Introduction	1
2	Holistic⁺⁺ Scene Understanding and Human-Object Contact	5
2.1	Holistic ⁺⁺ Scene Understanding	5
2.1.1	Introduction	5
2.1.2	Related Work	9
2.1.3	Representation of the Scene	11
2.1.4	Probabilistic Formulation	12
2.1.5	SHADE Dataset	14
2.1.6	Joint Inference	15
2.1.7	Experiments	20
2.1.8	Conclusion	26
2.2	Detecting Human-Object Contact in Images	27
2.2.1	Introduction	27
2.2.2	Related Work	30
2.2.3	Human-Object conTact (HOT) Dataset	32
2.2.4	Method	37
2.2.5	Experiments	39
2.2.6	Conclusion	46
3	Goal-directed, Multi-agent and Multi-task Event Parsing	47
3.1	Introduction	47

3.2	Related Work	50
3.3	The LEMMA Dataset	52
3.4	Benchmarks	57
3.5	Experiments	59
3.6	Conclusions	65
4	Human Communication in Shared Attention and Embodied Reference .	66
4.1	YouRefIt: Embodied Reference Understanding with Language and Gesture .	66
4.1.1	Introduction	66
4.1.2	Related Work	69
4.1.3	The YouRefIt Dataset	72
4.1.4	Embodied Reference Understanding (ERU)	76
4.1.5	Conclusion and Future Work	83
4.2	Inferring Shared Attention in Social Scene Videos	84
4.2.1	Introduction	84
4.2.2	Related Work	87
4.2.3	VideoCoAtt Dataset	89
4.2.4	Model	91
4.2.5	Experiments	96
4.2.6	Conclusion	100
5	Conclusion	102
	References	104

LIST OF FIGURES

1.1	In-depth understanding of a scene or event through joint parsing [ZGF20]. . . .	2
1.2	Focus shift from low-level tasks to high-level tasks.	4
2.1	Holistic ⁺⁺ scene understanding.	6
2.2	Typical Human-Object Interaction (HOI)s from the SHADE dataset.	14
2.3	The optimization process of the scene configuration.	18
2.4	Illustration of the top-down sampling process.	19
2.5	Augmenting SUN RGB-D with synthetic human poses.	21
2.6	Qualitative results of the proposed method on three datasets.	25
2.7	Qualitative comparison between model without physics and the full model. . . .	26
2.8	Contact estimations for images taken in the wild.	28
2.9	Images and contact annotations for our HOT dataset.	29
2.10	HOT dataset statistics.	35
2.11	Overview of the contact detection framework.	36
2.12	Attention visualization for all human parts.	40
2.13	Qualitative results on HOT dataset.	41
2.14	Representative failure examples.	43
2.15	Comparison of full-body contact detector against part-specific detectors.	43
2.16	Example applications of contact detection.	45
3.1	Illustrations of the LEMMA dataset with annotations.	48
3.2	An exemplar task instruction of making juice for two agents.	53
3.3	Statistics of the LEMMA dataset.	55

3.4	The co-occurrence statistics for verbs, nouns, and tasks in LEMMA.	56
3.5	Compositional action recognition benchmark on LEMMA.	59
3.6	Qualitative results of compositional action recognition on LEMMA.	62
4.1	Embodied reference in daily deictic-interaction scenario.	67
4.2	Illustration of the YouRefIt dataset collection procedure.	73
4.3	Statistics of the YouRefIt dataset.	75
4.4	The proposed multimodal framework for the ERU task.	77
4.5	Qualitative results in Image ERU.	79
4.6	Qualitative results in Video ERU	82
4.7	ROC Curve for canonical frame detection.	83
4.8	Examples of shared attention in daily life.	85
4.9	Example frames from VideoCoAtt dataset.	87
4.10	Illustration of shared attention location.	91
4.11	Illustration of VideoCoAtt model architecture.	91
4.12	Illustration of gaze heatmap generation procedure.	92
4.13	Illustration of shared attention inference process.	94
4.14	Quantitative evaluation results with ROC Curve.	97
4.15	Shared attention detection results on example frames.	99

LIST OF TABLES

2.1	Quantitative Results of 3D Scene Reconstruction	23
2.2	Quantitative Results of Global 3D Pose Estimation	24
2.3	Ablative results of HOI.	24
2.4	Evaluation of contact detection accuracy on the HOT dataset.	41
2.5	Contact-driven human pose estimation on PROX’s Quantitative set.	44
3.1	Comparisons between LEMMA and relevant indoor activity datasets.	50
3.2	Comparisons of compositional action recognition on LEMMA.	62
3.3	Comparisons of the action and task anticipations on LEMMA.	64
4.1	Comparisons between YouRefIt and other reference datasets. happens.	70
4.2	Comparisons of Image ERU performances on the YouRefIt dataset.	80
4.3	Video ERU performance comparisons on the YouRefIt dataset.	81
4.4	Canonical frame detection performance.	82
4.5	Comparison of VideoCoAtt related datasets.	88
4.6	Distributions of culture and scenario settings in VideoCoAtt dataset.	90
4.7	Statistics of the shared attentions and people involved in VideoCoAtt dataset.	91
4.8	Quantitative evaluation results with Prediction Accuracy and L_2 Distance.	98

ACKNOWLEDGMENTS

First, I would like to express my deepest thanks and gratitude to my advisor Song-Chun Zhu, for introducing me to the field of computer vision and artificial intelligence and the opportunity to be a member of the Center for Vision, Cognition, Learning, and Autonomy (VCLA). His lectures *Statistical Modeling in Computer Vision and Cognition* open the door for me to this amazing path. I still remember the afternoon when I was first exposed to those ideas and the excitements still inspire me to this day. His passion for AI research and visionary guidance have encouraged me to challenge the topics on the edge of the area.

I also treasure the tremendous support from my committee members, Prof. Ying Nian Wu, Prof. Demetri Terzopoulos, and Prof. Hongjing Lu. Their insightful ideas and expertise help shape my research attitude and intuition, and expand my knowledge in statistical thinking, computational artificial life, and human cognition.

I'm very fortunate to be in the family of VCLA, where I enjoyed the vibrant and cooperative working atmosphere. I especially thank Siyuan Huang for his help in introducing me to my current research topic, and I learned so much from him in every way. I thank Yixin Zhu for his acute research guidance and for being a leader who takes care of so many things for us. I want to thank my close collaborators, Tao Yuan, Baoxiong Jia, Qing Li, Lifeng Fan, Siyuan Qi, Yining Hong, Deqian Kong, Yik Lun Kei, Chao Xu, Prof. Tao Gao, and Prof. Ping Wei, for their contributions in the works towards this dissertation. I also want to thank other members of VCLA, Chi Zhang, Feng Gao, Liang Qiu, Tengyu Liu, Hangxin Liu, Zeyu Zhang, Muzhi Han, Mark Edmonds, Arjun R. Akula, Xiaofeng Gao, Ziyuan Jiao, Xiaojian Ma, Jonathan Mitchell, Erik Nijkamp, Shuwen Qiu, Yuxin Qiu, Feng Shi, Shu Wang, Sirui Xie, Xu Xie, Yifei Xu, Luyao Yuan, Yizhou Zhao, Zilong Zheng, Ruiqi Gao, Yaxuan Zhu, Daniel Ciao, Pan Lu, Ran Gong. The time with you is the most entertaining.

I also appreciate the opportunities to work with some brilliant minds in this field. I thank Li Erran Li and Du Zheng for hosting my internship at Amazon, where we explored the area

of natural language and speech recognition with close collaboration with Alejandro Mottini and Weiyi Lu. I also thank Prof. Michael Black and Dimitrios Tzionas for the chance to work remotely in Perceiving Systems, Max Planck Institute for Intelligent Systems, where I learn tremendously from both the expertise, profession, and passion people are devoting into research. I especially thank Hongwei Yi and Sai Kumar Dwivedi, who offer me the most help and warm discussions during this time.

I always appreciate the joy and happiness from my time with my friends, whether close or in the distance. All the experiences with you add the best color to my life, and I look forward to the adventures yet to come. Life is a journey, and my friends have kept me accompanied the whole time.

Finally, I thank my parents and my family for their love and support throughout my life, without whom none of this would be possible.

VITA

- 2017–2022 Graduate Research Assistant, Department of Statistics, UCLA.
- 2021–2022 Visiting Student Researcher, Perceiving Systems, Max Planck Institute for Intelligent Systems.
- 2020 Research Scientist Intern, Amazon.
- 2012–2016 B.E. in Communication Engineering, Harbin Institute of Technology.

PUBLICATIONS

(* indicates equal contribution.)

YouRefIt: Towards Embodied Reference with Language and Gesture. **Y. Chen**, Q. Li, D. Kong, Y. Kei, S. Huang, Y. Zhu, S. Zhu. International Conference on Computer Vision (ICCV), 2021

Top-down Attention in End-to-end Spoken Language Understanding. **Y. Chen**, W. Lu, A. Mottini, E. Li, J. Droppo, Z. Du, B. Zeng. International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2021

LEMMA: A Multi-view Dataset for Learning Multi-agent Multi-task Activities. B. Jia, **Y. Chen**, S. Huang, Y. Zhu, S. Zhu. European Conference on Computer Vision (ECCV), 2020

Closed Loop Neural-Symbolic Learning via Integrating Neural Perception, Grammar Parsing, and Symbolic Reasoning. Q. Li, S. Huang, Y. Hong, **Y. Chen**, Y. Wu, S. Zhu. International Conference on Machine Learning (ICML), 2020

3D Object Detection from a Single RGB Image via Perspective Points. S. Huang, **Y.**

Chen, T. Yuan, S. Qi, Y. Zhu, S. Zhu. Advances in Neural Information Processing Systems (NeurIPS), 2019

Holistic⁺⁺ Scene Understanding: Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-Object Interaction and Physical Commonsense. **Y. Chen***, S. Huang*, T. Yuan, S. Qi, Y. Zhu, S. Zhu. IEEE International Conference on Computer Vision (ICCV), 2019

Inferring Shared Attention in Social Scene Videos. L. Fan*, **Y. Chen***, P. Wei, W. Wang, S. Zhu. IEEE Computer Vision and Pattern Recognition (CVPR), 2018

CHAPTER 1

Introduction

Humans, even young infants, are adept at perceiving and understanding complex indoor scenes and events. Holistic scene understanding involves abundant aspects, including low-level tasks such as scene, human poses reconstruction, action recognition and human-object interaction, mid-level functionality, affordance and physics reasoning, and high-level event parsing like human activity and human communications. However, facts that are obvious to the average human adult—that describe how our physical and social worlds work, are still far from being understood by the current computer vision systems. For such complex and delicate intellectual skills, how can we build a machine or robot with similar capabilities, or even more importantly, where shall we start?

One can achieve an impressive performance in a single low-level task by training with an enormous amount of annotated data with the recent progress in Deep Neural Networks(DNN). Most existing work focuses only on 3D holistic scene understanding [HQZ18, ZLH18], 3D human pose estimation [ZWM17, RKS12], atomic action recognition [SZS12, KTS14, CEG15], visual grounding [KOM14, YPY16] or gaze direction prediction [RVK17]. But how to jointly infer such information from simple input with generalization ability is largely missing from current computer vision research. Take Fig. 1.1 [ZGF20] as example. A computer vision system should be able to jointly (i) reconstruct the 3D scene; (ii) estimate camera parameters, materials, and illumination; (iii) parse the scene hierarchically with attributes, fluents, and relationships; (iv) reason about the intentions and beliefs of agents (e.g., the human and dog in this example); (v) predict their actions in time; and (vi)

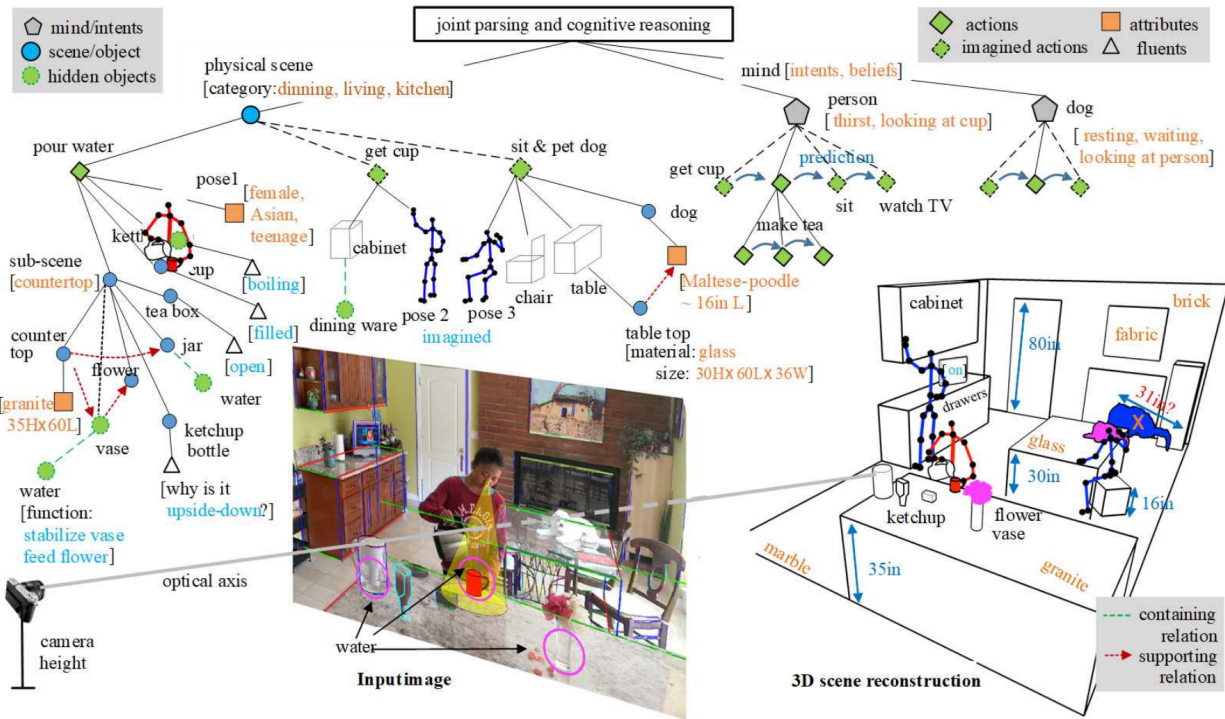


Figure 1.1: In-depth understanding of a scene or event through joint parsing [ZGF20].

recover invisible elements such as water, latent object states, and so forth. To reach such a comprehensive understanding requires efforts in several aspects.

First, the machine needs the core knowledge, e.g., functionality, physics, perceived intent and causality, to perform joint reasoning. Physical commonsense affords humans the ability to understand the physical world we live in and functionality is a further understanding of the physical environment humans use when they interact with it, performing appropriate actions to change the world in service of activities. These core knowledge can be learned separately and further act as general prior in the reasoning process to ensure the generalization ability. This is especially useful when end-to-end training fails (e.g., there is not enough data or the task is of high complexity [CHY19]).

Second, it calls for focus shift from low-level vision tasks to high-level event parsing. Current DNNs have shown advantages in single low-level tasks by feeding the model huge

amount of data. But human-like common sense can only be approximated or validated by using limited data to achieve generalizations across a variety of tasks. Such tasks would include a mixture of both low-level tasks (i.e., classification, localization, and reconstruction), and high-level problems, including but not limited to causal reasoning, learning functionality and affordance, intent prediction, collaboration and communication. It's important to introduce new high-level problems that are representative in the goal-directed events and human communications, so that these human-like capabilities can be evaluated and examined.

In this dissertation, we take steps further from the low-level and single task, towards more holistic scene understanding and goal-directed event parsing, as can be seen from Fig. 1.2. More specifically, we study this complex problem from three dimensions: physically and functionally accurate 3D reconstruction, long-term human activity understanding and human communications. We first propose a framework that jointly optimizes the 3D objects, 3D human poses, camera parameters from RGB images with HOI and physical stability constraints. This goes beyond object detection and human pose estimation, and provides generalization ability in various scenes for this complex task. We then propose to promote the machines' ability in understanding long-term, multi-tasked, collaborative activities rather than the atomic action recognition. As previous efforts of understanding reference have been primarily devoted in localizing a particular object in an image with only natural language expressions and human gaze interaction has been limited in gaze direction prediction, we further bring the human communication understanding into the computer vision community by introducing embodied reference and shared attention.

The remainder of the dissertation is organized as follows:

In Chapter 2, we will talk about the framework we propose to solve the holistic⁺⁺ scene understanding problem, which jointly reconstruct the 3D scene and 3D human pose from a single image. In this framework, one crucial factor to reconstruct the physically plausible and stable 3D configuration is the human-object interaction and contact. We will show how to effectively detect contact from 2D images and further utilize it in downstream tasks.

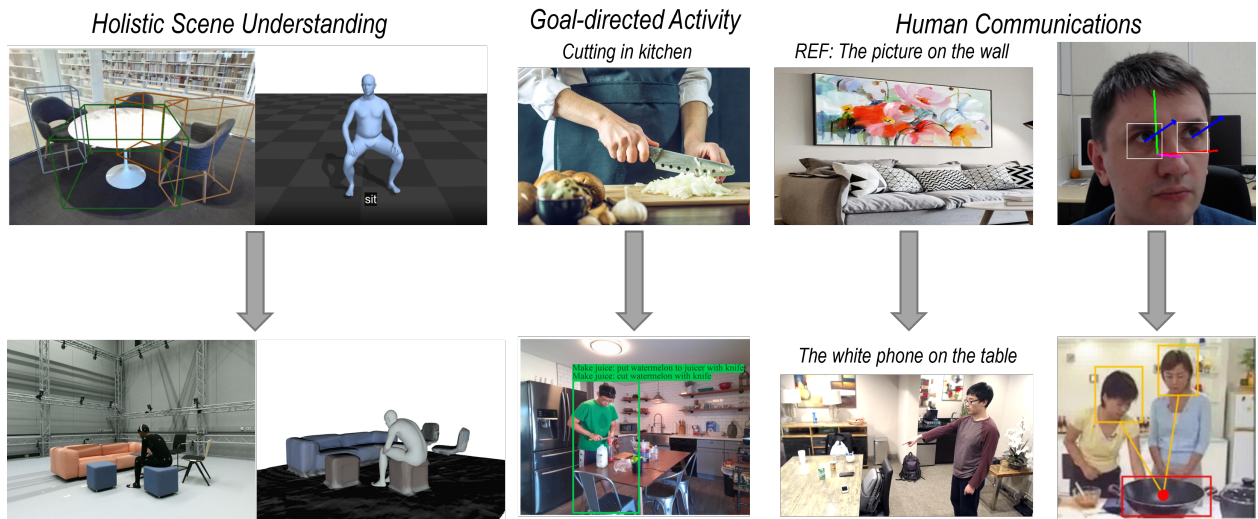


Figure 1.2: Focus shift from low-level tasks to high-level tasks.

In Chapter 3, we introduce our first step to understand and interpret human actions under the context of goal-directed actions, concurrent multi-tasks, and collaborations among multi-agents. We provide a multi-view dataset and benchmarks compositional action recognition and action/task anticipation to simulate the machine’s capability to consider the features mentioned above.

In Chapter 4, we present how to understand two typical forms of human communications: shared attention and embodied reference. A spatial-temporal model is proposed to explicitly leverage human gaze direction, target region candidates, and temporal inter-frame constraints for identifying shared attention. We also formulate a new multimodal framework to tackle the embodied reference understanding tasks by incorporating both language and gestural cues. Both require dealing with unique information sources, verbal and non-verbal, and they convey vivid and complex messages.

Finally, we summarize our work and point out promising future directions in the last chapter.

CHAPTER 2

Holistic⁺⁺ Scene Understanding and Human-Object Contact

In this chapter, we will first talk about the framework we propose to solve the holistic⁺⁺ scene understanding problem, which jointly tackles two tasks from a single-view image: (i) holistic scene parsing and reconstruction—3D estimations of object bounding boxes, camera pose, and room layout, and (ii) 3D human pose estimation. In this framework, one important factor to reconstruct the physically plausible and stable 3D configuration is to make sure the human and the 3D scene are in correct contact. We then show how to detect contact from 2D images and its important role in the optimization process.

2.1 Holistic⁺⁺ Scene Understanding

2.1.1 Introduction

Humans, even young infants, are adept at perceiving and understanding complex indoor scenes. Such an incredible 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

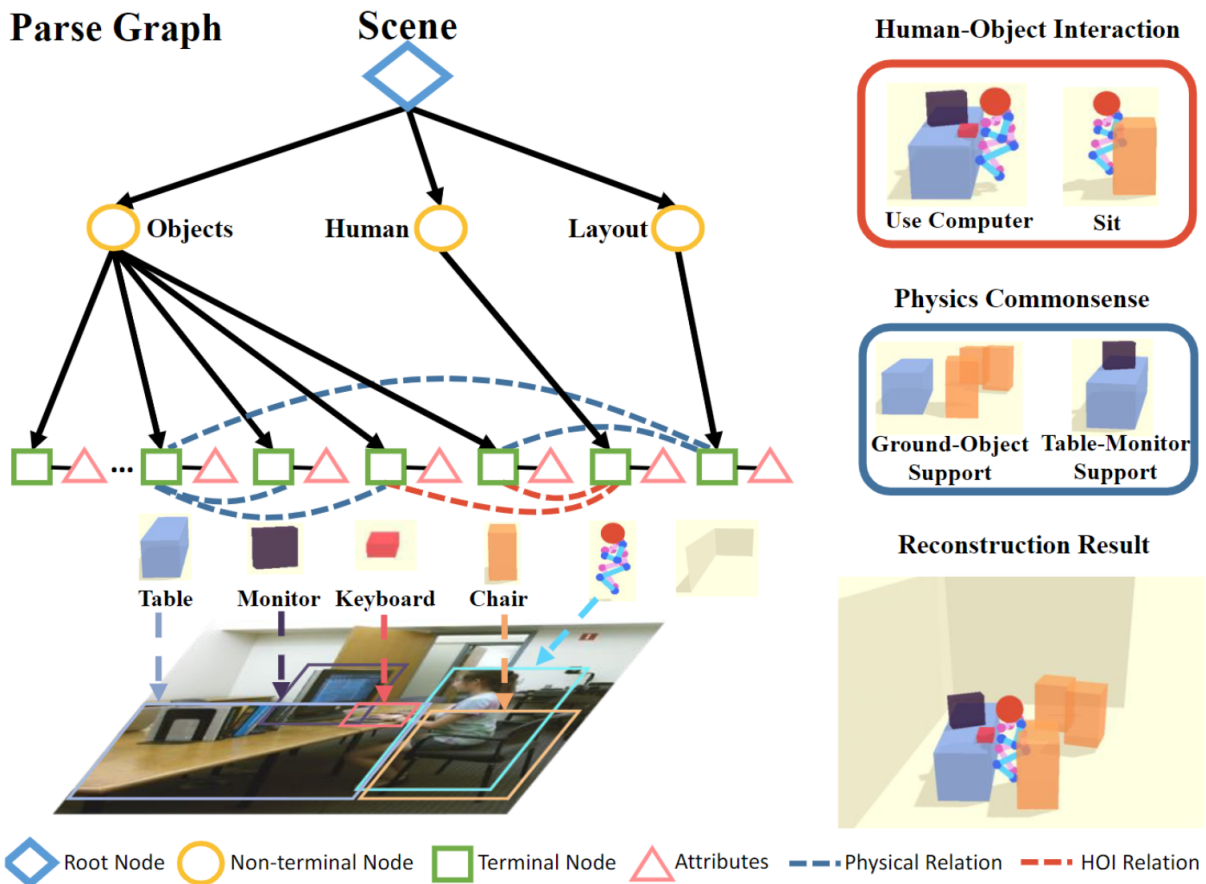


Figure 2.1: Holistic⁺⁺ scene understanding.

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 physics [GBK02a, Nee97, Bai04], and using exploratory play to understand the environment [SF15]. All the evidence calls for a treatment to integrate HOI and physical commonsense with a modern computer vision system for scene understanding.

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.

Based on one single-view image, existing work either focuses only on 3D holistic scene understanding [HQZ18, ZLH18, BRG16, SYZ17] or 3D human pose estimation [ZWM17, RKS12, FXW18]. 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 rich tasks and human activities [Gib79].

To solve the proposed *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 understanding, 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 2.1 as an example; humans are able to impose a constraint and infer the relative position and orientation between the girl and chair by recognizing the girl is sitting in the chair. Similarly, such a constraint can help to recover the small objects (e.g., recognizing keyboard by detecting the girl is using a computer in 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 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 maximum a posteriori (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 Markov chain Monte Carlo (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 are 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. 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 to bridge the human pose estimation and the scene reconstruction, reducing geometric ambiguities (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.

2.1.2 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 [ZLH18, 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, ZLH18]. 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] and objects [ZZC15], 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.

2.1.3 Representation of the Scene

We represent the configuration of an indoor scene by a parse graph $pg = (pt, E)$ as shown in Fig. 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 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”.

2.1.4 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:

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

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:

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

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 :

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

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 Section 2.1.6.3 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 Δx , Δy , 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

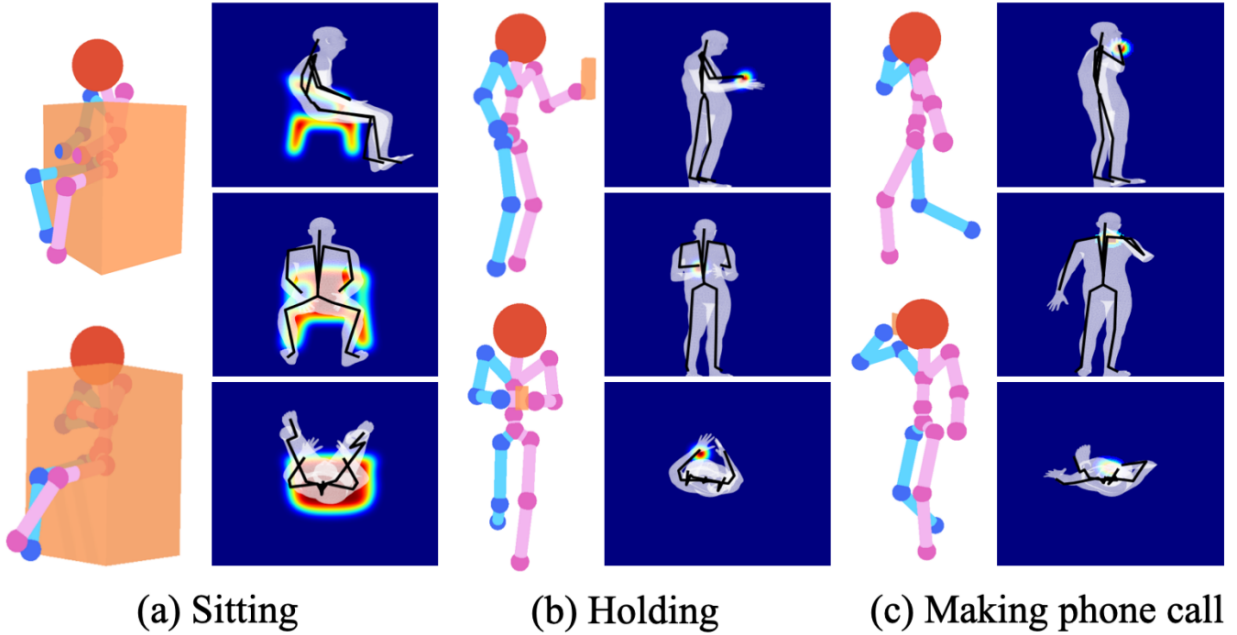


Figure 2.2: Typical HOIs from the SHADE dataset.

poses with projected ones from inferred 3D results:

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

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 Intersection over Union (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.

2.1.5 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. Fig. 2.2 shows some typical examples and relations in the dataset.

2.1.6 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.

2.1.6.1 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:

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

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.

2.1.6.2 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

$$\alpha \begin{bmatrix} \mathbf{v}_{2D} \\ 1 \end{bmatrix} = K \cdot R \cdot \mathbf{v}_{3D}, \quad (2.8)$$

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} .

2.1.6.3 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).

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 2.1.5. 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

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 synthetic data can significantly expand the training set and help to avoid overfitting in concurrent action detection.

2.1.6.4 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 Algorithm 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.

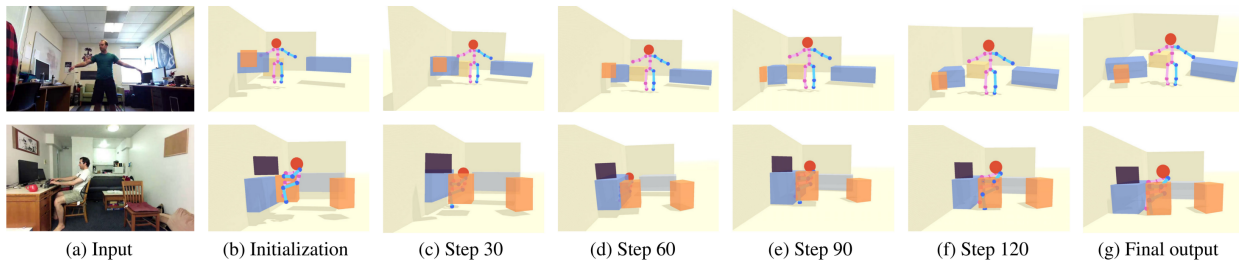


Figure 2.3: The optimization process of the scene configuration.

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 $q(pg \rightarrow pg'|I)$ 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]:

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

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 Eq. (2.4) to determine the category and location of the object; see examples in Fig. 2.4.

Implementation Details. In Phase (ii), we search the interacting objects for each agent involved in HOI by minimizing the energy in Eq. (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. Fig. 2.3 shows examples of the simulated annealing optimization process.

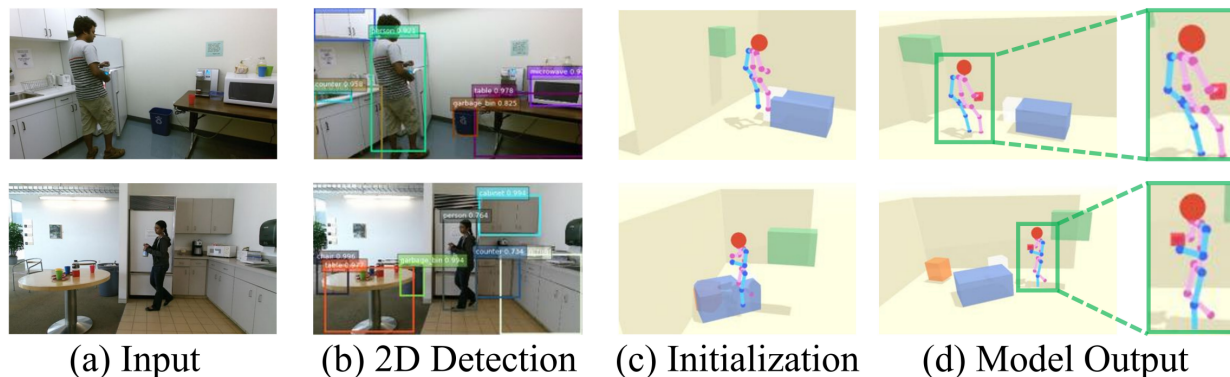


Figure 2.4: Illustration of the top-down sampling process.

2.1.7 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 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.

2.1.7.1 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.

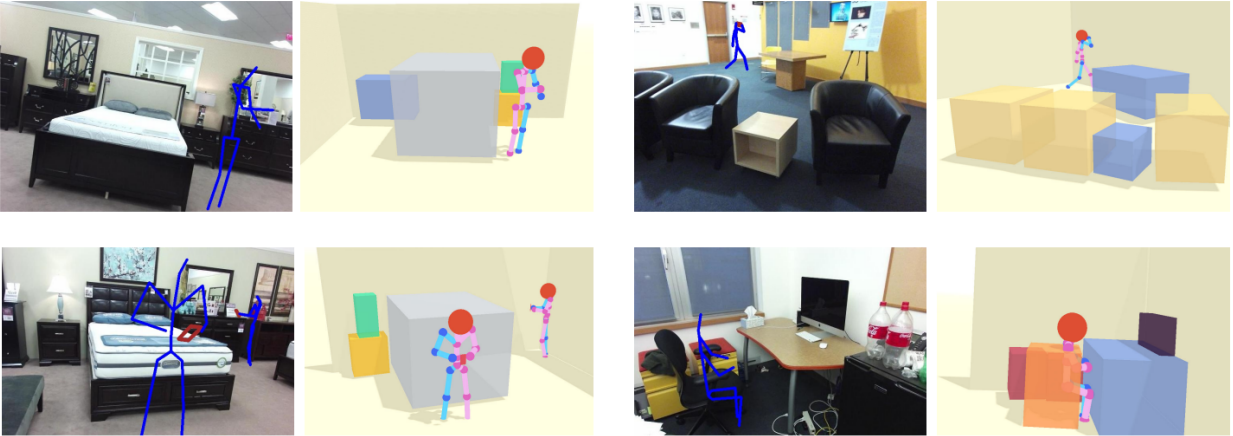


Figure 2.5: Augmenting SUN RGB-D with synthetic human poses.

2.1.7.2 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 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 Fig. 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.

2.1.7.3 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

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	-	75.1	24.9	-
SUN RGB-D	63.9	17.7	-	72.9	18.2	-
WnP	-	-	0.375	-	-	0.162

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. Fig. 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 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.

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	15.9	0.472
SUNRGBD	-	-	45.81	0.435	14.03	0.517
WnP	50.51	0.646	325.2	2.14	20.5	0.330

Table 2.3: Ablative results of HOI.

Methods	<i>w/o hoi</i>			<i>Full model</i>		
HOI Type	Object \uparrow	Pose \downarrow	MR \downarrow	Object \uparrow	Pose \downarrow	MR \downarrow
Sit	26.9	0.590	15.2	27.8	0.521	13.1
Hold	17.4	0.517	78.9	17.6	0.490	54.6
Use Laptop	14.1	0.544	58.8	15.0	0.534	43.3
Read	14.5	0.466	65.3	14.3	0.453	41.9

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 $\mathcal{E}_{phy}(pg)$.

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

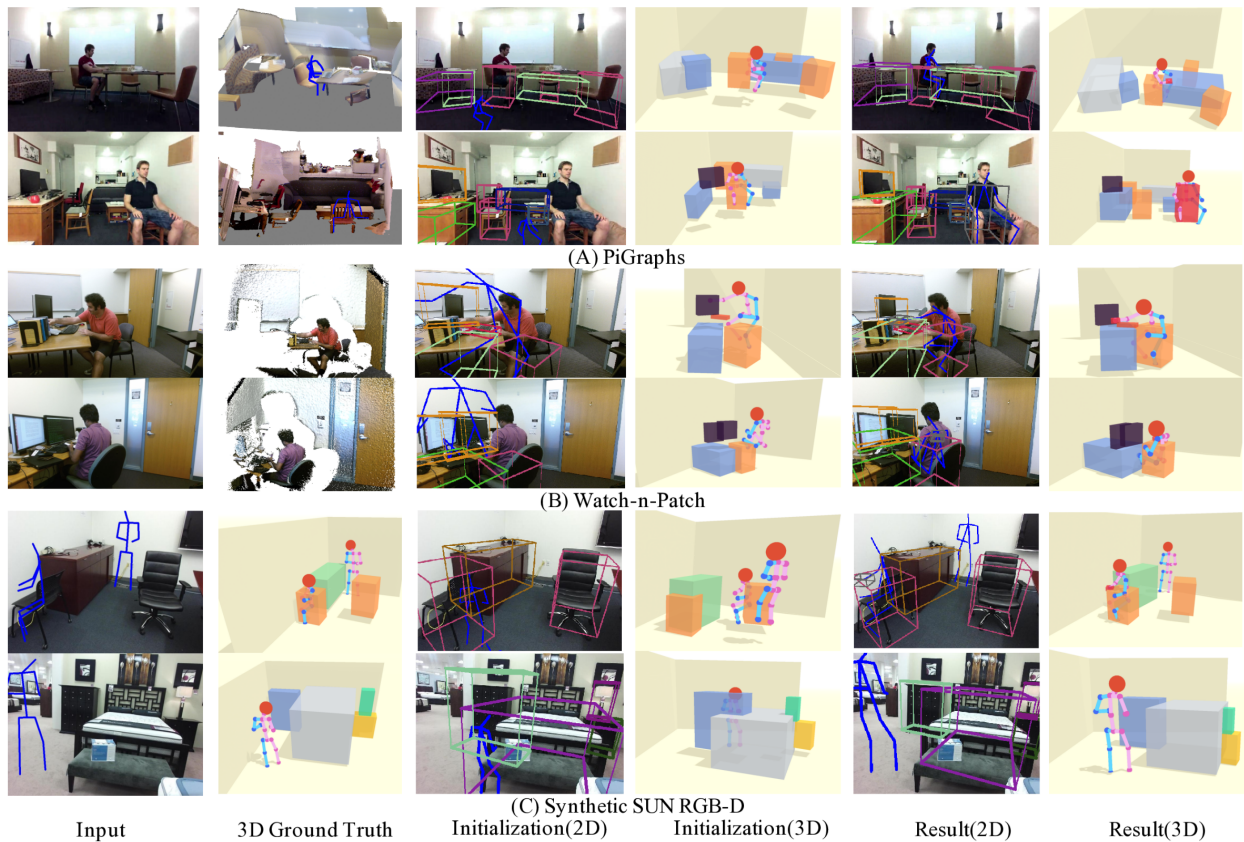


Figure 2.6: Qualitative results of the proposed method on three datasets.

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 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 Fig. 2.7. Objects detected by model *w/o*

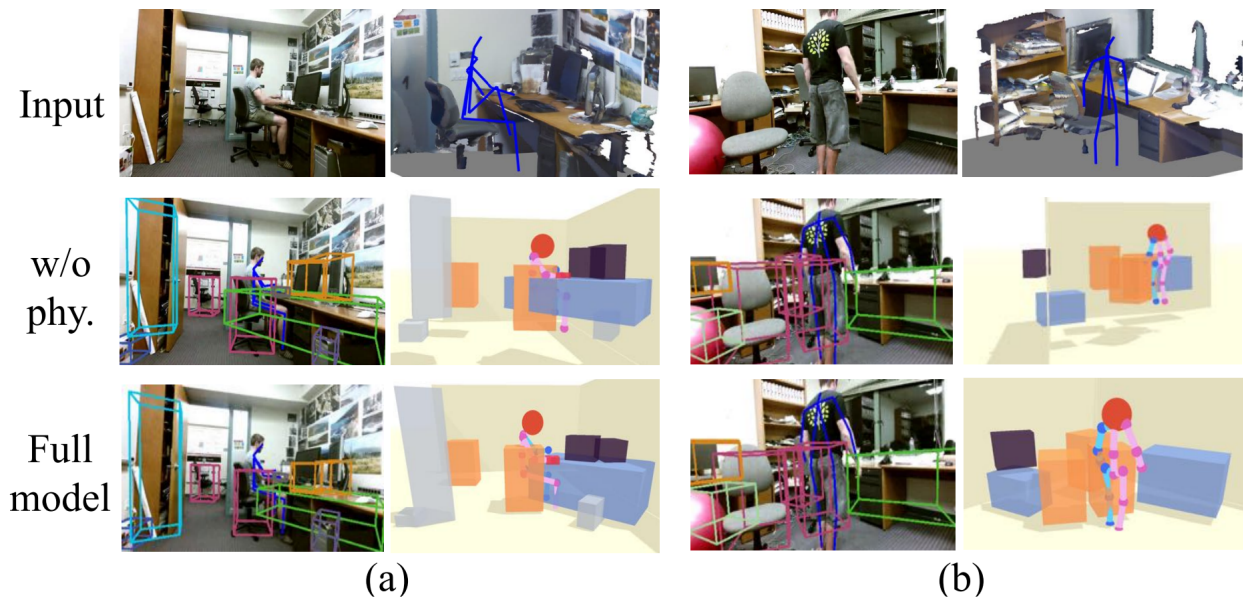


Figure 2.7: Qualitative comparison between model without physics and the full model.

phy. may float in the air or penetrate each other, while the full model yields physically plausible results.

2.1.8 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.

2.2 Detecting Human-Object Contact in Images

2.2.1 Introduction

Contact is an important part of people’s everyday lives. We constantly contact objects to move and perform tasks. We walk by contacting the ground with our feet, we sit by contacting chairs with our buttocks, hips and back, we grasp and manipulate tools by contacting them with our hands. Therefore, estimating contact between humans and objects is useful for human-centered AI, especially for applications such as AR/VR [ACV09, GSJ21, KT15, LFR17], activity recognition [JCH20, RCJ21, SCH16], affordance detection [FDG12, KS14, NG20, ZJZ16], fine-grained human-object interaction detection [LXL20, QWJ18, WAK19, XWL19], imitation learning [RWP20, TS10, ZMJ18], populating scenes with interacting avatars [HGT21, ZZM20, ZHN20], and sanitization or contamination prevention. Maybe surprisingly, there exists no detector for contact available online, similarly to off-the-shelf detectors for segmenting humans in images, or estimating their 2D joints or 3D shape and pose. Some work exists for detecting hand-object contact as bounding boxes, but hands are only part of the story. Moreover, bounding boxes are only a rough representation that can not capture the precise contact area. What we need, instead, is a new detector for the *entire body* that estimates detailed, body-part-specific, contact *maps* in images. To train this, we need data, but no suitable dataset exists at the moment. We account for this in this work, with a novel dataset and model for detecting contact between whole-body humans and objects in color images taken in the wild.

Annotating contact is challenging, as contact areas are ipso facto occluded. Think of a person standing on the floor; the sole of the shoe, and the floor area it contacts, can not be observed. A naive approach is to instrument a human with contact sensors, however, this is intrusive, cumbersome to set up and does not scale. Instead, we use two alternative data sources, with different but complementary properties. (1) We use the PROX [HCT19] dataset, which has pseudo ground-truth 3D human meshes for real humans moving in 3D

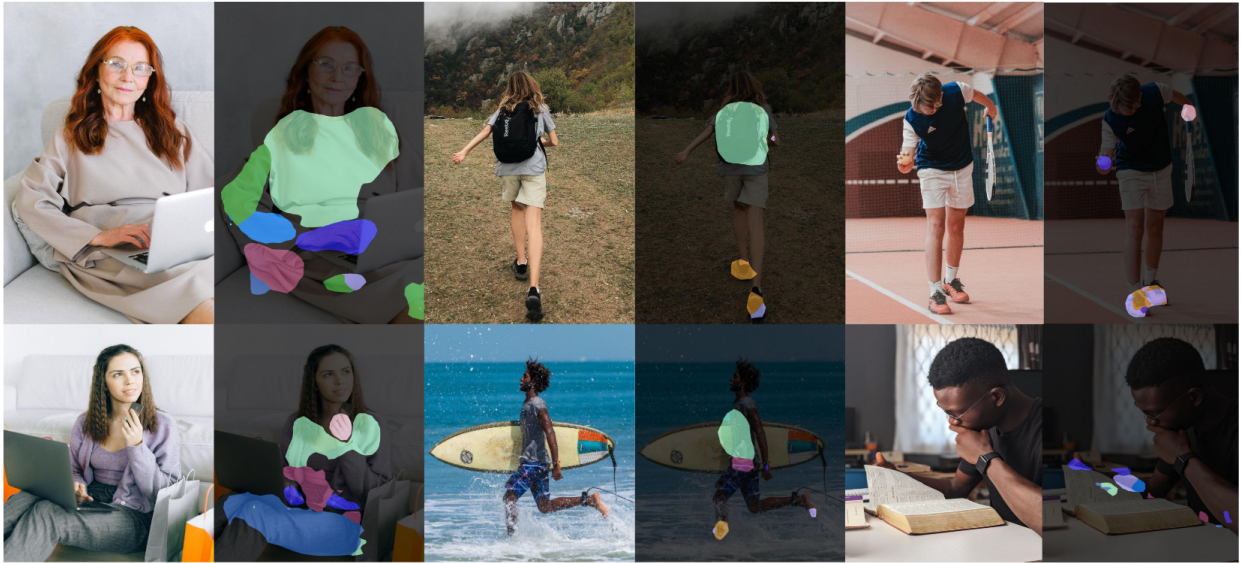


Figure 2.8: Contact estimations for images taken in the wild.

scanned scenes. By having 3D meshes for both humans and scenes, we can *automatically* annotate contact areas, by computing the proximity between the 3D meshes. (2) We use the V-COCO [GM15] and HAKE [LXL20] datasets, which contain images taken in the wild. We then hire professional annotators, and train them to annotate contact areas as 2D polygons in images. Although *manual* annotation is only approximate, 2D annotations are important because they allow scaling to large, varied, and natural datasets. This improves generalization. Note that in both cases we also annotate the body part that is involved in contact, corresponding to the body parts of the SMPL(-X) [LMR15, PCG19] human model.

We thus present HOT (“Human-Object conTact”), a new dataset of images with human-object contact. Examples from HOT are shown in Fig. 2.9. The first part of HOT, called “HOT-Generated”, has automatic annotations, but lacks variety for human subjects and scenes. The second part, called “HOT-Annotated”, has manual annotations, but has a huge variety of people, scenes and interactions. HOT has 35,750 images with 163,534 contact annotations.

We then train a new contact detector on our HOT dataset. Given a single color image as

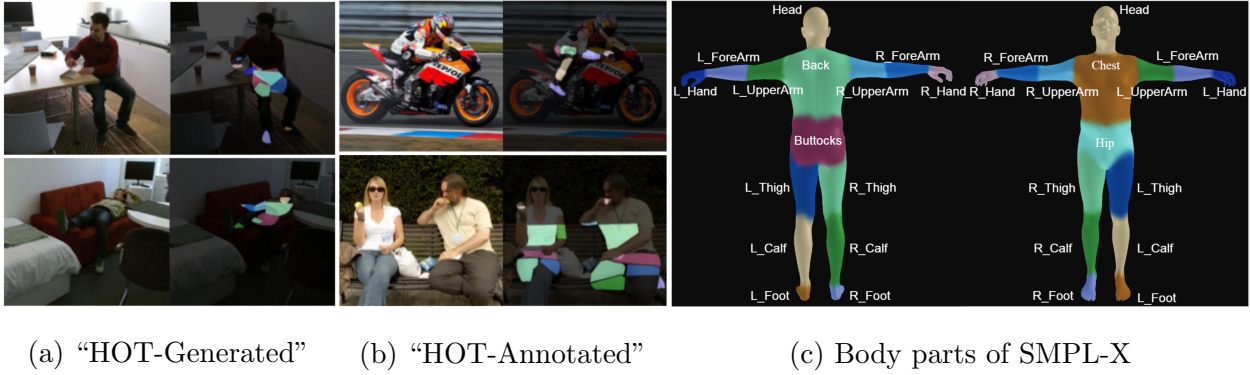


Figure 2.9: Images and contact annotations for our HOT dataset.

input, we want to know that, if contact takes place in the image, the area it occurs, as well as the body part that is involved. Specifically, we detect 2D heatmaps in an image, encoding the contact location and likelihood, and classify each pixel in contact to one of SMPL(-X)’s body parts. However, training directly with HOT annotations leads to “bleeding” heatmaps and false detections. We observe that humans reason about contact by looking at body parts and their proximity to objects in their local vicinity. Therefore, we use a body-part-driven attention module that significantly boosts performance.

We evaluate our detector on withheld parts of our HOT dataset. Quantitative evaluation and ablation studies show that our model outperforms the baselines, and that all components contribute to detection performance. Our body-part attention module is key for this, and a visual analysis shows that it attends to meaningful image locations, i.e., on body parts and their vicinity. Qualitative results show reasonable detections on in-the-wild images. By applying our detector on datasets unseen during training, we show that our model generalizes reasonably well; see Fig. 2.8. Note that there exists no other model for detecting contact between *whole bodies* and the scene in an image. We further show that our general-purpose full-body contact detector performs on par with existing part-specific contact detectors for the foot [RGH20] or hand [NNH20], having the prospect to work as a drop-in replacement for these. We also discuss that HOT has the potential to support several applications. Therefore,

we think that the community will find good use for our data and models.

In summary, HOT takes a step towards automatic contact detection between humans and objects in color images and our contribution is three-fold: (1) We introduce the new task of full-body human-object contact detection in images. (2) To facilitate machine learning for this, we introduce the HOT dataset with 2D contact area heatmaps and the associated human part labels as annotations, using both auto-generated and manual annotations. (3) We develop a new contact detector that incorporates a body-part attention module. Experiments and ablation studies demonstrate the benefits of the proposed model and its components.

2.2.2 Related Work

2.2.2.1 Contact Modeling

Modeling contact information between humans and objects has been studied for different aspects and scale, namely for body-object and hand-object contact.

Body-object contact: Several works model the contact between the human body and object. Clever et al. [CEK20] propose a synthetic dataset with special focus on the lying pose where contact takes place between a human body and a pressure-sensing mat. Li et al. [LSC19] reconstruct the 3D motion of a person interacting with an object by estimating the 3D pose of the person and object, the joint-level contact, and forces and torques actuated by the human limbs. Rempe et al. [RBH21, RGH20] estimate joint-level foot-ground contact from a video, and use it to constrain the human pose with trajectory optimization. Others use HOI relationships to reconstruct [CHY19, WY21] or generate [ZZM20, ZHN20] 3D human and object pose by enforcing contact and penalizing collision.

In prior work, contact information is often used as prior knowledge, but is often oversimplified as either body-ground contact at the skeleton-joint level, or hand-object contact at a rough bounding-box level, or manually-annotated point contacts of other human parts.

In this work, we seek to automatically estimate contact heatmaps for the whole body in a bottom-up manner directly from the 2D image. We also predict the associated human body part label, which provides a more systematic understanding of human-object contact.

Hand-object contact: People interact with objects using their hands, so contact plays an important role in hand-object interaction, grasping and hand pose estimation. Contact information is often captured as a byproduct in grasp datasets [BTT20, LJX21, TGB20] through hand-object proximity or thermal information. Hand-object grasp reconstruction also employs contact to refine the hand and object pose estimation [CRK21, GTT21, HVT19]. In addition, some works [NNH20, SGS20] detect hands and classify their physical contact state into self-contact, person-person contact, and person-object contact. Although they consider the relationship between hands and other objects in the scene, they detect only a rough bounding box for the hand, instead of a finer-grained contact area. In this work, we take a step further to estimate general-purpose full-body contact from 2D images at a finer scale.

2.2.2.2 Human-Object Interaction (HOI)

The task of HOI understanding [QWJ18, WAK19, XWL19] aims to infer the interaction relationships between humans and objects. While both humans and objects are located in the image, often in form of 2D bounding boxes, the literature seldomly focuses on how the interaction takes place, whether the interaction requires contact, and what human part is involved in the contact. These limitations make the current HOI detection less relevant for downstream scene understanding tasks. Recently, Li et al. [LXL20] provide more detailed body-part state annotations in the context of HOI, and offer contact information in the form of action labels (e.g., hold, paddle) and the involved human body part (e.g., hand, foot). However, they do not annotate 2D contact areas in images, and their predefined human parts are not fine-grained enough to capture everyday HOI scenarios for which contact plays an important role. On the contrary, our new dataset contains 2D contact areas that are

also associated with the involved human body parts following the part segmentation of the popular SMPL(-X) [LMR15, PCG19] statistical 3D body model.

2.2.2.3 Affordance Learning

Contact and HOI are closely related with object affordances, which reflect the functional aspects of an object. Recent work explores object affordance learning from human actions and object manipulations [FDG12, KS14, NG20, ZJZ16]. More specifically, Fang et al. [FWY18] and Nagarajan et al. [NFG19] learn to predict the interaction region with the corresponding action label on a target object from human demonstration videos. Savva et al. [SCH14, SCH16] capture physical contact and visual attention links between 3D geometry and human body parts from RGB-D videos. Deng et al. [DXW21] collect a 3D visual affordance dataset with potential interaction areas on 3D objects for various actions. Affordance learning is object-centric; the final product does not capture much about the human actor. On the contrary, detecting interaction areas (e.g. contact heatmaps) reflects how people interact with objects, and considers both the human and the object.

2.2.3 Human-Object conTact (HOT) Dataset

To facilitate research in contact estimation, we introduce HOT, a new dataset with 2D contact areas and the associated human part labels as annotations. Annotating and detecting contact in images is challenging, as contact depends on the scene and its objects, the humans, the camera view and the occlusions arising from all these factors. To create a well-varied dataset, we collect images from two different sources and gather contact annotations for these. Below we discuss the creation of HOT and provide a comprehensive analysis of it.

2.2.3.1 Data Sources

First we collect data from the PROX dataset [HCT19], which contains people reconstructed as 3D SMPL-X [PCG19] meshes interacting with static 3D scenes; this involves actions like sitting, walking, lying down, etc. Recent work [RBH21, ZZB21] improves on the quality of reconstructed meshes in PROX, facilitating the automatic generation of contact heatmaps by simply using 3D proximity metrics between the 3D human mesh and the static 3D scene mesh. We sub-sample frames from the qualitative set of PROX, and form the “HOT-Generated” part of HOT.

Another source for images with human-object contact is HOI datasets like V-COCO [GM15] and HAKE [LXL20]. As they are collected from Flickr, these datasets contain very diverse HOI interactions in complex and cluttered scenes. Existing HOI datasets contain activity labels and bounding boxes for humans and objects, but boxes are too coarse for understanding contact. Therefore, we select a subset from the V-COCO [GM15] and HAKE [LXL20] datasets and use these to gather new contact annotations. To keep the task tractable, we first remove images with indirect human-object interaction, heavily cropped humans, motion blur, distortion or extreme lighting conditions. Other interesting datasets are indoor action recognition datasets like Watch-n-Patch [WZS15], which contain several daily human activities like “fetch-from-fridge”, “put book back”, etc. We sample image frames from video clips where human subjects and objects are clearly visible. We eventually combine images selected from V-COCO [GM15], HAKE [LXL20] and Watch-n-Patch [WZS15], and form the “HOT-Annotated” part of HOT.

2.2.3.2 Contact Generation for “HOT-Generated”

The PROX dataset [HCT19] captures people interacting with static scenes. Using the reconstructed 3D human and scene meshes, we can first compute the vertices that are in close 3D proximity as contact vertices, and then render these onto the 2D image to get automatic con-

tact area annotations, as well as the associated body labels. More specifically, we represent the human pose and shape with the SMPL-X [PCG19] body model, which captures the body surface including the hands and face. The SMPL-X model represents the human body with pose parameters, θ , and shape parameters, β , and outputs a posed 3D mesh, $\mathcal{M}_b \in \mathbb{R}^{10475 \times 3}$. Each vertex, $v \in \mathbb{R}^3$, has a surface normal, n^v , and a human part label, c , associated with it. We divide the parametric human body model SMPL-X into 17 parts c_i , with $i \in \{1, 2, \dots, 17\}$. For this, we are based on the original part segmentation of SMPL-X, but for simplicity we unite certain parts (e.g., parts of the back across the spine), that even human annotators cannot easily differentiate. See Fig. 2.9c for the color-coded segmentation of SMPL-X and the corresponding body-part labels. For each frame, given the reconstructed SMPL-X mesh, \mathcal{M}_b , and the scene mesh, \mathcal{M}_s , we first calculate human-to-scene mesh distances. Then, all human vertices with a distance to the scene below a threshold, and with compatible normals to the scene ones, are annotated as contact vertices. Finally, for the contact vertices we find the respective triangles on the 3D body mesh, and render these separately per body part to get dense 2D contact areas. In this way, we automatically create pseudo ground truth for contact. Examples are shown in Fig. 2.9a.

2.2.3.3 Contact Annotation for “HOT-Annotated”

In the following, we describe how we annotate contact for in-the-wild images. The annotation process includes two steps: (1) we draw a polygon around the image area containing human-object contacts, and (2) assign a human body-part label associated each contact. See Fig. 2.9b for some annotation examples.

Determining the exact contact area between a human and an object is non-trivial, especially in the image space, because contact areas are always occluded. Thus, we hire professional annotators and ask them to “detect” the image areas in which contact takes place, and draw a polygon around each of these. We take a number of steps to ensure good quality and consistency for the annotations. In particular, we have two rounds of quality

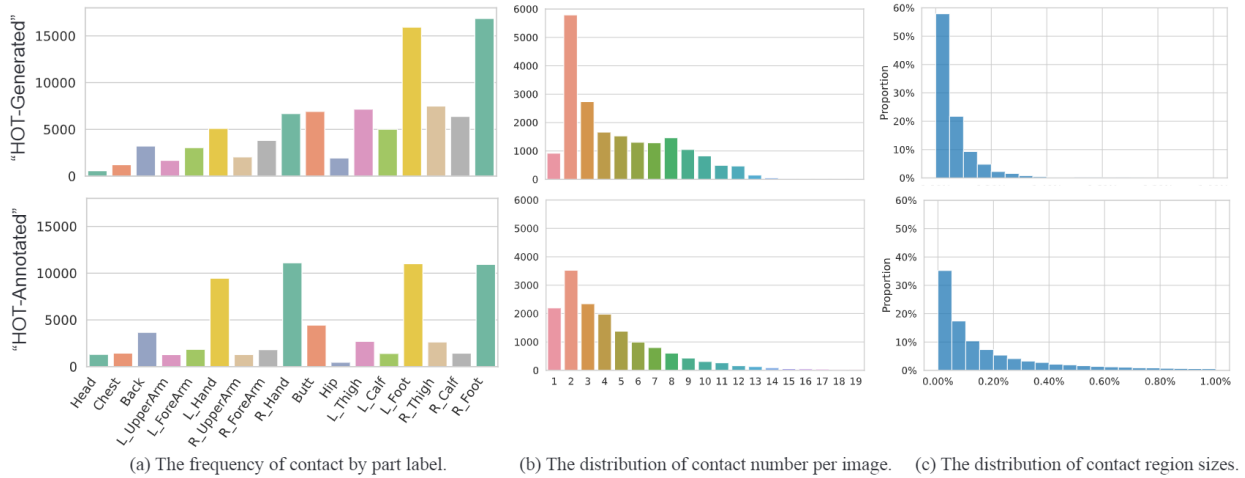


Figure 2.10: HOTA dataset statistics.

check after the initial annotation; for every 3 annotators there is 1 extra annotator that only conducts quality checks.

Compared to the automatic annotations of “HOT-Generated”, the manual annotations of “HOT-Annotated” are only approximate. However, capturing people in scenes and accurately reconstructing them in 3D is hard and does not scale. Thus, manual 2D annotations are important because they allow scaling to large, varied, and natural datasets with images taken in the wild. For data-driven models, this helps towards improving generalization and making them robust.

2.2.3.4 HOTA Dataset Statistics

The HOTA dataset has a total of 35,750 images and 163,534 contact area annotations, along with a body-part label for each area. Specifically, for “HOT-Annotated” we collect 5,459 images and 20,898 contact areas for V-COCO [GM15], 9,761 images and 46,287 contact areas for HAKE [LXL20], and 325 images and 1,170 contact areas for the Watch-n-Patch [WZS15] dataset. For “HOT-Generated”, we auto-generate 95,179 contact areas in 20,205 images using the PROX dataset. More statistics of “HOT-Annotated” and “HOT-Generated” are

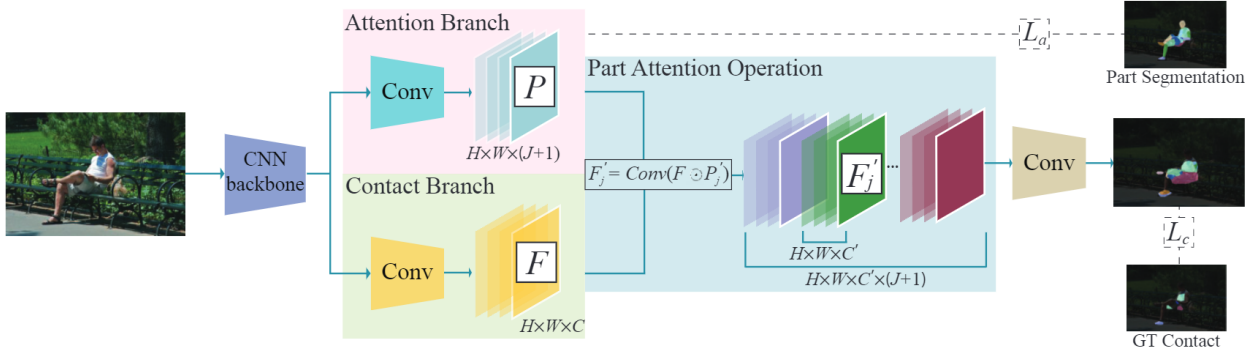


Figure 2.11: Overview of the contact detection framework.

shown in Fig. 2.10.

Figure 2.10a shows the distribution of body-part labels for contact. We see that “HOT-Annotated” has noticeably more contacts than “HOT-Generated” for both hands. The reason is that PROX captures humans interacting with static scenes, i.e., without grasping and moving objects with their hands, while “HOT-Annotated” contains a lot of images with interactions between hands and objects.

Figure 2.10b shows the number of contact area annotations per image. We see that “HOT-Annotated” has generally more contacts per image than “HOT-Generated”. This is possibly because HOI datasets also contain images of multiple interacting persons, while PROX only has a single person in every image.

Figure 2.10c shows the distribution of contact area size. We observe that the areas are generally smaller for “HOT-Generated” than “HOT-Annotated”. This is potentially because images in PROX are captured with the camera away from the body to include more scene context, whereas images in “HOT-Annotated” are taken in the wild, including close-ups, as well as more object grasps.

2.2.4 Method

To estimate contact areas in images, humans use the global context of the image, but also focus on regions around body parts to examine if there is contact and how the area around it looks like. Based on these insights, we design our contact detector to extract global features with attention to human body parts.

2.2.4.1 Model Architecture

Figure 2.11 shows the overall architecture of our proposed contact detector. Given an image, we first use a CNN backbone to extract the image features. Then we use an image decoder with two branches: an *Attention Branch* for predicting attention masks for each human part and a *Contact Branch* to extract contact features. We denote the attention branch as $P \in \mathbb{R}^{H \times W \times (J+1)}$, where J is the number of human parts, with one extra channel for the background. The symbols H and W are the height and width of the feature map. We denote the contact prediction branch as $F \in \mathbb{R}^{H \times W \times C}$, with the same spatial dimensions $H \times W$ as the attention branch P , but with a different number of channels, C .

In the attention branch, the j_{th} channel $P_j \in \mathbb{R}^{H \times W}$ represents the likelihood that each pixel is associated with contact of the j_{th} body part. This is used to guide the model to focus around different human parts in the feature space F of the contact branch. By applying a channel-wise softmax normalization $\sigma(\cdot)$ on P , we get the attention mask. $P' = \sigma(P)$, with $P' \in \mathbb{R}^{H \times W \times (J+1)}$.

We then use a *Part Attention Operation* to combine the attention and contact branches, i.e., use P'_j as an attention mask to extract part-related features:

$$F'_j = Conv(F \odot P'_j), \quad \text{with } F'_j \in \mathbb{R}^{H \times W \times C'}, \quad (2.10)$$

where \odot is the element-wise product between all channels in F and the attention mask P'_j . We concatenate F'_j for all j parts and the background along the channel dimension as

$F' \in \mathbb{R}^{H \times W \times C^*}$, where $C^* = C'(J + 1)$, and later feed it into a convolution layer to get the final per-pixel prediction.

We supervise the attention branch with part-segmentation maps, and the contact branch with contact area annotations; see “dataset splits” in Section 2.2.5.1 for details on the part-segmentation supervision source. Both branches conduct pixel-wise classification to a certain human part or the background. The “background” label for the contact branch indicates “no contact”. Our joint loss is:

$$L = \lambda_a L_a + \lambda_c L_c, \tag{2.11}$$

where L_a is a cross-entropy loss between the estimated attention maps and ground-truth part-segmentation maps, L_c is a cross-entropy loss between the estimated and ground-truth contact maps, and λ_a and λ_c are steering weights.

2.2.4.2 Implementation Details

During training, body-part supervision for the attention branch is applied only in the initial stages, following Kocabas et al. [KHH21]; λ_a is set to 0 at later stages. As Fig. 2.12 shows, the learned attention mask attends not only regions on each body part, but also on the surrounding area to explore contextual features.

We use a pre-trained dilated ResNet-50 [YKF17] as image encoder backbone. For the attention branch we use 3×3 convolutional layers with batch-norm and ReLU as image decoder, followed by another convolutional layer with kernel size 1 to make pixel-wise human part label classification. For the contact branch, we apply 3×3 convolutional layers with batch-norm and ReLU on the part-specific features, which we further concatenate along the channel axis. Note that the weights of convolution layers are different across human parts, so that the contact branch learns part-specific features under the attention guidance. Another convolutional layer with kernel size 1 is used to make pixel-wise contact label prediction. Since the background dominates the label ground truth for both human part segmentation

and contact estimation, we assign a smaller weight for the background label in the cross-entropy loss.

During training, we re-scale all images to have their bigger side 400 pixels long, and then pad, if necessary. Random flipping is applied for data augmentation. We use a batch size of 24 and the SGD [RM51] optimizer, with an initial learning rate of 0.02 with polynomial decay following [ZZP18].

2.2.5 Experiments

In this section, we first benchmark the contact estimation task on a withheld test set of our HOT dataset. Then, we compare our full-body contact detector with existing part-specific contact detectors. Finally we showcase examples of contact detection for several tasks interesting for the broader community.

2.2.5.1 Contact Detection

Dataset splits: For the “HOT-Annotated” part of HOT, we randomly split the collected images into a training and test set with a ratio of 8:2, resulting in 12,436 images for training and 3,109 images for testing. For the “HOT-Generated” part, we split the training and testing set based on the scene, to make sure that scenes in the test set are unseen during training. This results in 14,143 images for training and 6,052 images for testing. We evaluate each baseline (ours and competitors) on the test set for several checkpoints, and report the best performer for each case. For supervising the attention branch, we obtain pseudo ground truth for human part segmentation by rendering part-segmented SMPL(-X) meshes; we use LEMO’s [ZZB21] SMPL-X fits for the PROX dataset and use FrankMocap [RSJ21] to estimate SMPL-X for the images of “HOT-Annotated”.

Evaluation protocol: We adopt the evaluation protocol of Zhou et al. [ZZP18]; this is originally for semantic segmentation. We add one metric for contact area prediction to eval-

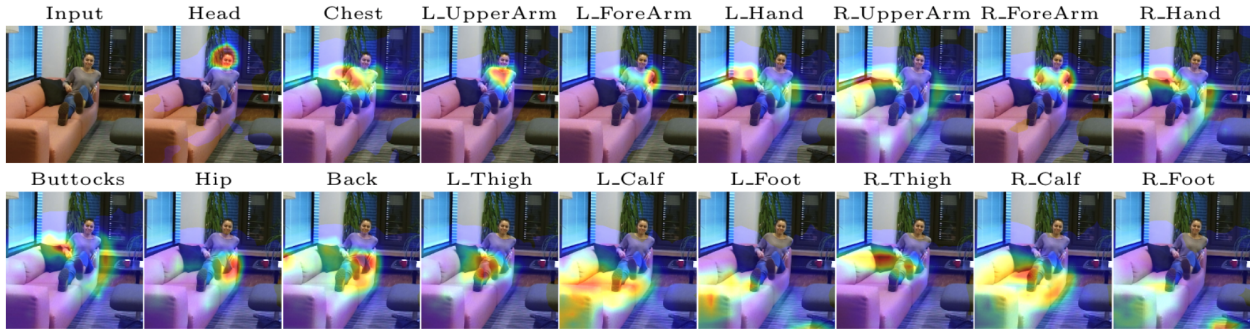


Figure 2.12: Attention visualization for all human parts.

uate whether the model distinguishes between contact and non-contact, i.e., “background”.

We use the following metrics:

- *Semantic contact accuracy (SC-Acc.)*: The proportion of pixels that are correctly classified as in contact and associated with the correct body-part label.
- *Contact accuracy (C-Acc.)*: The proportion of correctly classified pixels for binary contact labels; this ignores the body-part label compared to “SC-Acc.”.
- *Mean IoU (mIoU)*: The IoU between the predicted and the ground-truth contact pixels, averaged over all the body-part labels.
- *Weighted IoU (wIoU)*: mIoU weighted by the pixel ratio of each contact label.

Background labels are not considered for computing “SC-Acc.” and “IoU”.

To study the influence of the “HOT-Annotated” and “HOT-Generated” sets of the HOT dataset, we report performance by training and testing models separately on these, as well as on their combination that we denote as “Full Set”. For the “Full Set”, we randomly choose images from the “HOT-Generated” so that the number of training and testing images from both sets are the same.

Baselines: We evaluate contact estimation for two baselines, ResNet+PPM [ZSQ17] and ResNet+UperNet [XLZ18] originally developed for semantic segmentation. Note that there exists no other model that does full-body contact detection in images.

Ablations: We evaluate two variants of our proposed model to ablate the contribution of the



Figure 2.13: Qualitative results on HOT dataset.

attention branch: (i) $\text{Ours}_{\text{wo/att}}$: without the attention branch and (ii) $\text{Ours}_{\text{pure.att}}$: without supervision for the attention branch, which functions as an unsupervised pure soft-attention module.

Results & discussion: Quantitative results for contact detection are shown in Table 2.4, and qualitative results are shown in Fig. 2.13. Below we discuss key findings:

1. Our model outperforms state-of-the-art (SOTA) methods [XLZ18, ZSQ17] developed for semantic segmentation. This is due to the different nature of semantic scene understanding and contact estimation. The former relies on dense pixel annotations of the entire scene and global context-

Table 2.4: Evaluation of contact detection accuracy on the HOT dataset.

Model	“HOT-Annotated”				“HOT-Generated”				“Full Set”			
	<i>SC-Acc</i> ↑	<i>C-Acc</i> ↑	<i>mIoU</i> ↑	<i>wIoU</i> ↑	<i>SC-Acc</i> ↑	<i>C-Acc</i> ↑	<i>mIoU</i> ↑	<i>wIoU</i> ↑	<i>SC-Acc</i> ↑	<i>C-Acc</i> ↑	<i>mIoU</i> ↑	<i>wIoU</i> ↑
ResNet+UperNet [XLZ18]	36.2	62.9	0.199	0.229	21.6	43.5	0.085	0.117	35.3	65.4	0.200	0.227
ResNet+PPM [ZSQ17]	35.8	61.0	0.205	0.243	21.3	40.9	0.078	0.121	33.5	57.6	0.191	0.232
$\text{Ours}_{\text{wo/att}}$	25.2	43.5	0.148	0.195	12.2	25.4	0.053	0.101	18.3	30.3	0.128	0.152
$\text{Ours}_{\text{pure.att}}$	35.3	59.3	0.194	0.242	20.8	41.6	0.081	0.116	33.0	57.0	0.178	0.226
$\text{Ours}_{\text{Full}}$	42.9	70.1	0.235	0.263	31.9	56.1	0.143	0.177	39.2	68.4	0.231	0.270

tual features. The latter relies on sparser annotations and needs an attention mechanism to focus around humans.

2. Our attention mechanism guides our model to learn better features that improve contact estimation. $\text{Ours}_{\text{pure_att}}$, which uses unsupervised pure soft-attention, outperforms $\text{Ours}_{\text{wo/ att}}$ which has no attention branch. By adding supervision on human-part segmentation in early training stages, the attention focuses on areas around each human part; intuitively, this helps reasoning about contact by using both human-body and surrounding-object information. Figure 2.12 provides evidence for this by visualizing the learned attention maps.

3. Learning on “HOT-Generated” is more difficult than on “HOT-Annotated”. This is partially because, even though we generate contact annotations from relatively “clean” SMPL-X fits by the involved LEMO [ZZB21] method, which reasons about temporal continuity and occlusion, these are still a bit noisy. Some reasons are the strong occlusions during interactions, motion blur, the low resolution for people observed by indoor-monitoring cameras, and the imperfect “hallucination” of SOTA methods [RBH21, ZZB21] against these ambiguities. Fine-grained contact detection is sensitive to such errors. This shows the value of “HOT-Annotated”, i.e. the collection of a well-defined and high-quality dataset of in-the-wild images with rich manual contact annotations, and points to important future work.

4. Figure 2.14 shows some examples of failure cases. We see that our model might struggle with occlusions, multiple persons or fine-grained contact areas. We also observe that the model sometimes fails in distinguishing left and right for the body parts. These point out that contact detection may benefit from future work on adding human pose information, reasoning from multi-resolution and differentiating human-object contact with self-contact and person-contact, but these are currently out of our scope.

2.2.5.2 Comparison with Part-specific Contact Detectors

To evaluate the robustness of our general-purpose full-body contact detector, we compare against two existing part-specific contact detectors, as shown in Fig. 2.15:

(i) Foot contact: “ContactDynamics” [RGH20] estimates *joint-level* foot-ground contact from a



Figure 2.14: Representative failure examples.

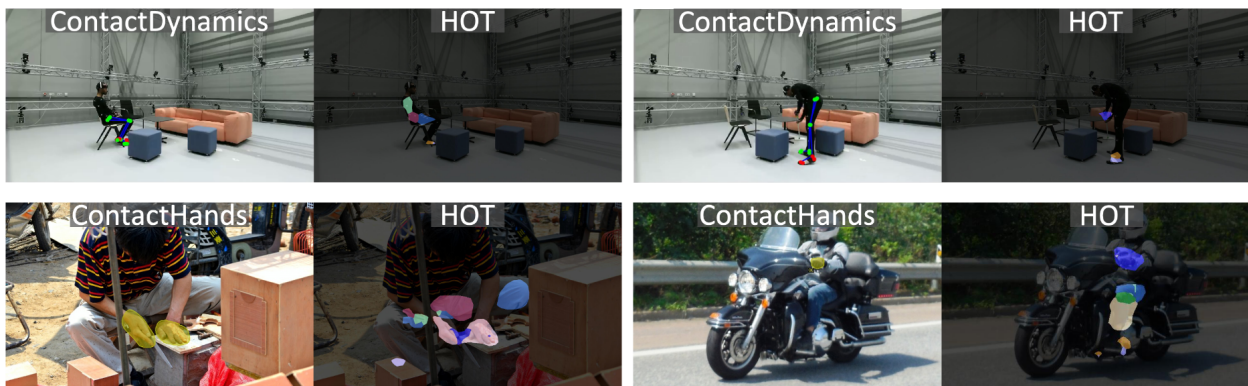


Figure 2.15: Comparison of full-body contact detector against part-specific detectors.

video, while we detect 2D contact areas in single images. We evaluate our model and “ContactDynamics” against the ground-truth foot contact from PROX’s “quantitative set”. Our detector achieves similar performance (ours **59.2%** vs “ContactDynamics” 58.6%), thus, it could be a drop-in replacement contact detector for the 3D pose estimation application of Rempe et al. [RGH20]. Note that, as Fig. 2.15 shows, “ContactDynamics” simply classifies foot joints as in contact or not, while we generalize to the full body and detect richer heatmaps.

(ii) Hand contact: “ContactHands” [NNH20] detects hands and classifies their contact state into “self-contact”, “person-person”, and “person-object” (hand-object) contact. We evaluate our model and “ContactHands” on hand-object contact on a subset of the “HOT-Annotated” test set. We report contact recognition accuracy under an IoU threshold of 0.4; our detector achieves similar performance (ours **63.5%** vs [NNH20] 62.2%). Note that, as Fig. 2.15 shows, “ContactHands” detects hands as bounding boxes, while we generalize to the full body with heatmaps.

The fact that our full-body contact detector performs on par with existing part-expert ones shows good prospects towards developing a general purpose contact detector for diverse human-object and human-scene interactions.

2.2.5.3 Comparison with Heuristic Contact

The PROX dataset of Hassan et al. [HCT19] enjoys popularity for developing and evaluating HOI methods. This dataset contains 3D SMPL-X meshes of real humans moving and interacting with static 3D scenes. The human meshes look physically plausible, and have been reconstructed with an optimization method that fits SMPL-X to images, with an a-priori known 3D scene. The method encourages the contact vertices on the human body to be close to the scene while not penetrating it, where potential contact vertices are manually annotated.

Table 2.5: Contact-driven human pose estimation on PROX’s Quantitative set.

Method	No Cont.	PROX [HCT19]	All Cont.	Pred. Cont.	GT Cont.
V2V ↓	183.3	174.0	176.3	172.3	163.0

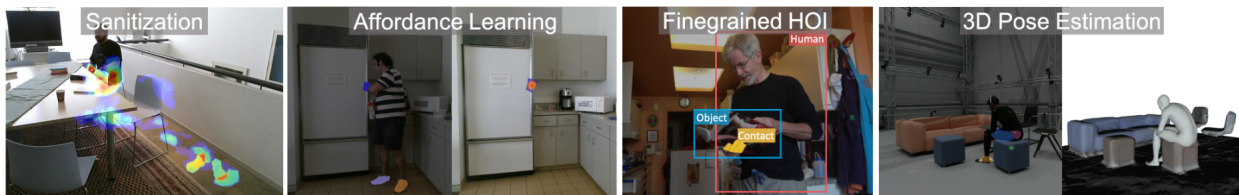


Figure 2.16: Example applications of contact detection.

We replace PROX’s manually annotated contact vertices with vertices of the SMPL-X human parts our detector suggests that are in contact, given the input image. We call this setup “Predicted Contact” and evaluate this on PROX’s “quantitative set” via the Vertex-to-Vertex ($V2V$) error. We also compare with a baseline with “No Contact” constraints. We use the same optimization process in PROX [HCT19] for fair comparison. Results in Table 2.5 show that our “Predicted Contact” is on par with “PROX”, indicating that detecting contact in images is promising for replacing PROX’s handcrafted heuristics. We also simulate a perfect contact detector using PROX’s ground truth (“GT Contact”). This shows that there is room and merit for improving image-based contact detection as future work.

2.2.5.4 Contact Detection Applications

Contact detection is important for applications in many domains such as AR/VR, activity recognition, affordance detection, fine-grained human-object interaction detection (beyond bounding boxes), 3D human pose estimation and populating scenes with interacting avatars. Here we showcase several examples in Fig. 2.16. For instance, one possible future direction is to extend the triplet definition of HOI $\langle \text{human}/\text{action}/\text{object} \rangle$ by adding contact as $\langle \text{human-part}/\text{contact-area}/\text{object} \rangle$, which supports finer-grained HOI reasoning. Another application is detecting in videos the areas that people contact, and guiding human cleaners (AR) or robots with heatmaps for sanitization or contamination prevention.

2.2.6 Conclusion

Here we focus on human-object contact detection for images. To this end, we introduce the HOT dataset and propose a new contact estimation method with human-part guided attention. Our model outperforms the baseline models and shows reasonable generalizability for in-the-wild images. Experiments provide empirical evidence that human-part attention is critical for contact estimation. Importantly, our data and model go significantly beyond existing work towards a general-purpose contact detector for the full body. We believe that this new task and dataset fill a gap in the literature, and will help the community for several applications.

CHAPTER 3

Goal-directed, Multi-agent and Multi-task Event Parsing

In the Chapter, we will introduce how to understand and interpret human actions under the context of goal-directed actions, concurrent multi-tasks, and collaborations among multi-agents. We introduce the LEMMA dataset to provide a single home to address these missing dimensions in prior literature with meticulously designed settings, wherein the number of tasks and agents varies to highlight different learning objectives. We densely annotate the atomic-actions with human-object interactions to provide ground-truths of the compositionality, scheduling, and assignment of daily activities. We further devise challenging compositional action recognition and action/task anticipation benchmarks with baseline models to measure the capability of compositional action understanding and temporal reasoning.

3.1 Introduction

Activity understanding is one of the most fundamental problems in artificial intelligence and computer vision. As the most readily available learning source, videos of daily human activities could be used to train intelligent agents and, in turn, to assist humans. However, compared to recent progress in learning from static images [AAL15, HZR16, HGD17, RHG15], current machine vision’s ability to understand activities from videos still falls short. Admittedly, activity understanding is inherently more challenging, which requires reason about the complex structures in activities along the additional temporal dimension; but we argue there are more profound reasons that we must look back to the origin of activity understanding.

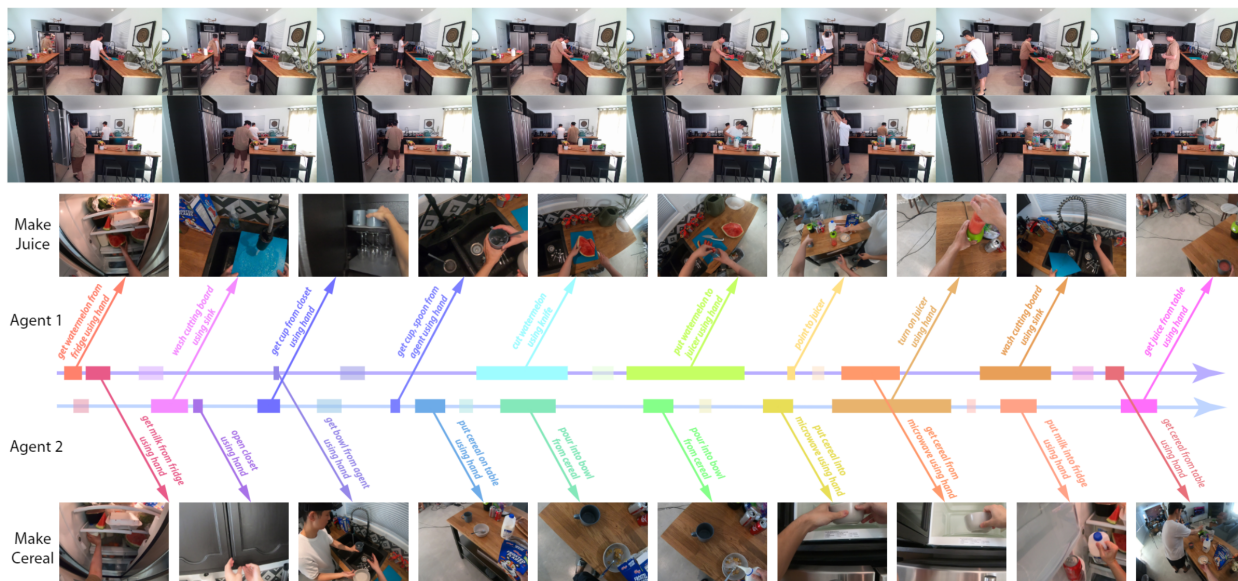


Figure 3.1: Illustrations of the LEMMA dataset with annotations.

The study and analysis of human motion perception are rooted in the field of neuroscience [TCS08]. Using a dot-representation of human motions, Johansson [Joh73] adopted a method to produce proximal patterns (i.e., the moving light display experiment), which demonstrated that human perception of activities does not tightly couple with *pixel-based features*; human subjects can still perceive the semantics of activities from *sparse* representations of motions. Evidence from developmental psychology, the classic Heider-Simmel experiment, further suggests that we perceive human activities from as *goal-directed* behaviors [Woo98, BBS01, GBK02b, CG07]; it is the underlying intent, rather than the surface pixels or behavior, that matters when we observe motions [BB01]. Such a **goal-directed [LMR99] perspective** of activity understanding has been largely left untouched in computer vision.

Daily human activities are intrinsically multi-tasked [Mon03, RME01]; understanding activity naturally demands a learning system to interpret concurrent interactions. As agents' decision-making processes are deeply affected by their unique social values, task scheduling is significantly affected by interactions (e.g., cooperation, competition, subordination) among multi-agents [KHA16]. These observations implicate that the machine vision system must objectively understand how a given task should be decomposed into atomic-actions, how multi-tasks should be executed and coor-

minated in parallel among multi-agents, and take the perspective from human agents to understand why the observed human activities are optimal solutions. Such a **decompositional, multi-task, multi-agent, diagnostic-driven, social perspective** of activity understanding is critical for an intelligent agent to understand human behavior and team with humans collaboratively; yet it is broadly missing in activity understanding literature.

The semantics of human actions are intrinsically ambiguous when described in natural language. For instance, although both “opening the fridge” and “opening a book” use the action verb “open,” their semantics of the actions are utterly different. In this work, we take the stance of Grice’s influential work on language act [Gri75]—technical tools for reasoning about rational action should elucidate linguistic phenomena [GF16]. Specifically, the compositional relations between the verbs and nouns could reveal the functionality of the object and the patterns of human-object interactions, which subsequently facilitate the understanding of the observed human activities and the language that describes them. Though the previous work [GKM17] attempted to address this issue, more general and flexible **compositional relations for describing human actions interacting with objects** are requisite for a goal-directed activity understanding.

Motivated by these deficiencies in prior work, we introduce the LEMMA dataset to explore the essence of complex human activities in a goal-directed, multi-agent, multi-task setting with ground-truth labels of compositional atomic-actions and their associated tasks. By quantifying the scenarios to up to two multi-step tasks with two agents, we strive to address human multi-task and multi-agent interactions in four scenarios: single-agent single-task (1×1), single-agent multi-task (1×2), multi-agent single-task (2×1), and multi-agent multi-task (2×2). Task instructions are only given to one agent in the 2×1 setting to resemble the robot-helping scenario, hoping that the learned perception models could be applied in robotic tasks (especially in HRI) in the near future.

Both the third-person views (TPVs) and the first-person views (FPVs) were recorded to account for different perspectives of the same activities; see Fig. 3.1. We densely annotate atomic-actions (in the form of compositional verb-noun pairs) and tasks of each atomic-action, to facilitate the learning of multi-agent multi-task task scheduling and assignment; see more details in Section 3.3.

3.2 Related Work

In this section, we review and compare prior indoor activity datasets on the basis of tasks and captured video contents; see a detailed summary in Table 3.1.

Crowd-sourced from online videos and movie sharing platforms, typical large-scale video datasets [SZS12, KTS14, CEG15, CZ17, FKE18] focus on **video-level summarization and classification**. Although activity classes exhibit a large inter-class variability, spanning from outdoor sports activities to indoor household activities, they generally lack sequential, goal-directed activities. Notably, they suffer from a major drawback [GR20]; activities are highly correlated to the general scene and object context, possessing a strong dataset bias for activity understanding.

Some datasets tackle the **human atomic-actions** using short clips or limited tasks, with a focus on the semantics of action verbs and objects [GKM17], 3D action analysis [LZL10, IPO13, SCH16], and action grounding with multi-modality inputs [MAZ19]. Although such datasets are suitable for atomic-actions, they are intrinsically impaired at studying the long-term reasoning of goal-directed human activities.

Recently, **concurrent actions** have been taken into consideration. For instance, Charades [SVW16] is a large-scale benchmark for household activities, and Charades-Ego [SGS18] steps further with

Table 3.1: Comparisons between LEMMA and relevant indoor activity datasets.

Dataset	Task Annotation	Multi-agent	Multi-task	Multi-view	Samples	Frames	Action Classes	Action Segments	Actions per Video	Modality	Year
MPII Cooking [RAA12]	✓	✗	✗	✗	273	2.9M	88	14,105	51.7	RGB	2012
ADL [PR12]	✗	✗	✓	✗	20	1.0M	32	436	13.6	RGB	2012
50Salads [SM13]	✓	✗	✗	✗	50	0.5M	17	966	19.3	RGB-D	2013
CAD-120 [KGS13]	✗	✗	✗	✗	120	0.1M	10	1,175	9.8	RGB-D	2013
Breakfast [KAS14]	✓	✗	✗	✓	433	3.0M	50	3,078	7.1	RGB	2014
Watch-n-Patch [WZS15]	✓	✗	✗	✗	458	0.1M	21	2978	6.5	RGB-D	2015
Charades [SVW16]	✗	✗	✓	✗	9,848	7.4M	157	67,000	6.8	RGB	2016
Something-Something [GKM17]	✗	✗	✗	✗	108,499	-	174	108,499	1.0	RGB	2017
EGTEA GAZE+ [LLR18]	✓	✗	✗	✗	86	2.4M	106	10,325	120.1	RGB	2018
EPIC-KITCHENS [DDM18]	✗	✗	✓	✗	432	11.5M	149	39,596	91.7	RGB	2018
LEMMA (proposed)	✓	✓	✓	✓	324	4.6M	641	11,781	36.4	RGB-D	2020

both FPVs and TPVs. However, the activities involved are mostly unrelated to specific goals due to the crowdsourced script generation process. Similarly, although Multi-THUMOS [YRJ18] and AVA [GSR18] focus on highly paralleled activities, and some datasets look at the temporal order of activities [BLB14, TZS16], the unnaturally scripted activities result in the lack of meaningful goal-directed tasks exhibited in our daily life.

Conversely, **instructional video** datasets [ABA16, SM13, KAS14, KGS13, RRR16] tackle goal-directed multi-step tasks, mostly in cooking, repairing, and assembling activities. In spite of their relevance, they fail to account for multi-agent or multi-task problems. EPIC-KITCHENS [DDM18] is perhaps the only exception; it records naturally paralleled task execution of agents in kitchen environments, but with no task specification or multi-agent interactions. Additionally, prior instructional video datasets have either drastic view perspective changes [ZXC18, ABA16, TDR19, TCH17] or limited egocentric view with severe occlusions [PR12, LLR18], hindering the activity understanding.

Another related stream of work is the learning of group-level activities in a **multi-agent** setting [IMD16], such as detecting key actors [RHA16], predicting future trajectories [PES09, LCL07], and recognizing collective activities [CSS09, OHP11, SXR15]. However, such coarse-grained multi-agent interactions leave the latent subtlety of collaboration and task assignment untouched. Although simulation-based multi-agent environments [BKM20, VBC19, BBC19] can partially address such an issue, learning from noisy and real visual input in physical work is still essential for understanding collaborative planning behaviors of agents in the context of complex daily tasks.

The collected LEMMA dataset strives to address the shortcomings of the aforementioned works, capturing goal-directed, decompositional, multi-task activities with multi-agent collaborations. As shown in Table 3.1, the size, annotation, and actions per video of LEMMA are at a comparable scale to state-of-the-art benchmarks. We hope such a design will boost the study of human activity understanding and potentially motivate new cross-disciplinary research insights.

Contributions This work’s contribution is three-fold. (i) We design and collect a multi-view video dataset, capturing multi-agent, multi-task activities with goal-directed daily tasks. (ii) We annotate the dataset, focusing on the compositionality of actions and the governing task for each

atomic-action. (iii) We provide compositional action recognition and action/task anticipation benchmarks by considering the aforementioned features; we also compare and analyze multiple baseline models to promote future research on human activity understanding.

3.3 The LEMMA Dataset

This section describes the design, data collection, and data annotation process of the LEMMA dataset. The dataset is profiled by various statistics from diversified perspectives to highlight its potentials in activity understanding.


Activities and Scenarios

We first build a task pool of 15 common tasks in the kitchen (e.g., “make juice,” “make cereal”) and living room (e.g. “watch TV,” “water plant”). On top of these tasks, we design four types of scenarios (with a different focus) to study goal-directed multi-step multi-task indoor activities in multi-agent settings.

1. **Single-agent Single-task (1×1):** Each participant was first asked to perform all tasks from the task pool independently; this ensures participants are clear with the goal of each task and could schedule and assign tasks efficiently in later multi-task or multi-agent scenarios. Participants were asked to read the instructions and walk around to get familiarized with the new environments.
2. **Single-agent Multi-task (1×2):** Each participant was then asked to simultaneously perform two tasks, randomly sampled from the task pool. The participants determined the order of task executions without any restrictions.
3. **Multi-agent Single-task (2×1):** Two participants were asked to perform a single task cooperatively; the task is randomly selected from the task pool. To emulate human-robot teaming accurately, only one participant (leader) was provided with task instructions; the other participant (helper), with no knowledge of the task, was asked to collaborate with the leader agent to finish the task efficiently. Only nonverbal communications (e.g., gestures) were allowed between two participants; this design would open up new venues on nonverbal communications and the emergence

In this task, you are asked to **make watermelon juice**. Here are things to know before your start: **Leader**

- All the items needed for this task can be found either in the **fridge**, on the **table**, or in one of the **drawers** or **closets**.
- Please **cut** the **watermelon** into pieces before blending it with the **juicer**.
- Please keep the kitchen clean; wash all the **tools/objects** you used.
- You will have an additional **helper** to collaborate with you.
- Do **Not** speak with them. They do **NOT** know anything about the task you are working on.
- Feel free to ask them for help, but only using **non-verbal** communication (e.g., gestures). For instance, you may point to something, or any other gestures you think may help instruct them.



In this task, you are asked to **collaborate** with your friend to finish a task in the kitchen. **Helper**
 Here are things to know before your start:

- All the items needed for this task can be found either in the **fridge**, on the **table**, or in one of the **drawers** or **closets**.
- Please keep the kitchen clean; wash all the **tools/objects** you used.
- As only your friend knows the task instruction, please try to infer what the task is and offer helps.
- **You may not speak with your friend**. You can only use **non-verbal** communication (e.g., gestures).

Figure 3.2: An exemplar task instruction of making juice for two agents.

of language in real-world environments.

4. **Multi-agent Multi-task** (2×2): Both participants were provided with task instructions. Since both participants were asked to accomplish two complex multi-step tasks collaboratively, this scenario has the most natural activity/task patterns and richest mechanisms for learning task scheduling and assignment.

In total, the LEMMA dataset includes 37 unique task combinations in the multi-task scenarios. Participants were explicitly instructed to perform tasks efficiently and provided with a brief task instruction with basic environment information. Except for the specification of the goal states for each task, we add no additional constraint to the order of task execution; participants perform tasks naturally and freely. Fig. 3.2 shows a sample instruction for the 2×1 scenario.

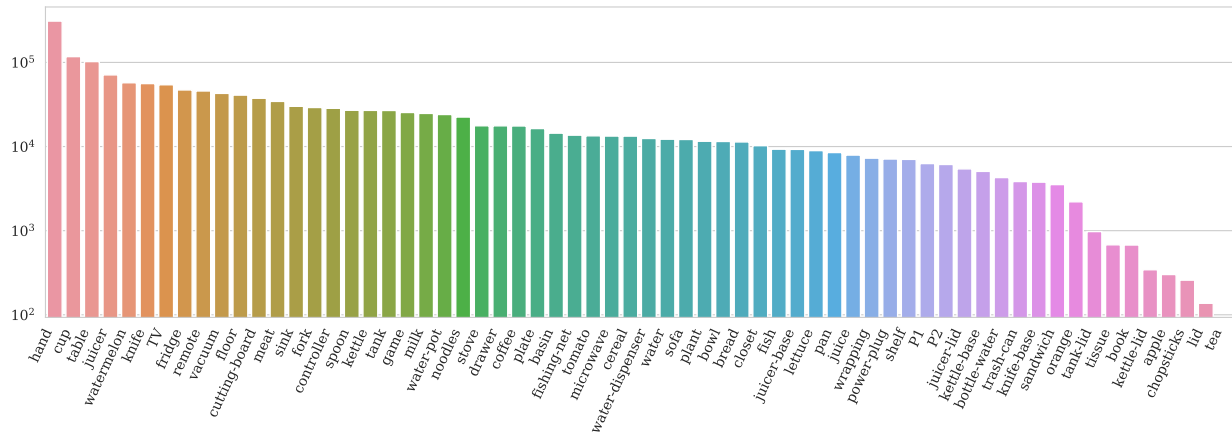
Data Collection

We recorded the data in 7 different Airbnb houses, performed by 8 individuals in 14 unique kitchens/living rooms. To provide different views of performing the daily activities and avoid occlusion in narrow spaces, we set up two Kinect Azure cameras to capture the RGB-D videos of the global scene and human bodies. In addition, each participant was instructed to wear a head-mounted GoPro camera to capture detailed agent-specific actions in an egocentric view. In post-processing, we synchronize the camera recordings of all views at a frame rate of 24 FPS. Fig. 3.2 shows an example of a scene with a point cloud merged from two Kinects and four RGB views from both Kinects and GoPros. Combining TPVs and FPVs captures most of the details of performing daily activities, provides sufficient data for understanding human activities, and benefits future research in embodied vision. The additional depth information and 3D human skeletons captured by Kinects can also be adopted for future 3D understanding tasks.

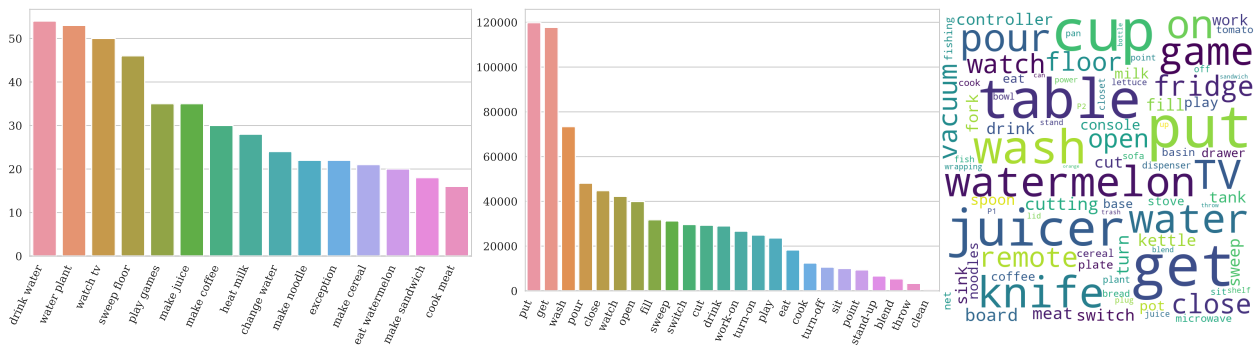
Ground-truth Annotation

We used the Amazon Mechanical Turk (AMT) to annotate both human bounding boxes and action information in the synchronized recordings. Specifically, action information includes the temporal localization of segments, semantic labels, and the governing task of each atomic-action. The semantic labels of atomic-actions are composed of verbs and nouns, representing flexible compositional relations to describe human actions. Additional details are provided below.

Bounding Boxes and Segments: Bounding boxes of humans are annotated on the primary view of TPVs. Skeletons captured by Kinects are used to provide initial estimations of bounding boxes. Next, we use Vatic [VPR13b] to adjust bounding boxes and annotate the segments of atomic-actions. The segments of atomic-actions are defined by verbs without corresponding nouns, for example, “put __ to __ using __,” “pour into __ from __.” Each video was first annotated by two AMT workers; task-irrelevant actions (e.g., “walking,” “holding”) are ignored. We then compute the Intersection over Union (IoU) of both bounding boxes and temporal segments. A third AMT worker is asked to fine-tune the annotations if the IoU of bounding boxes or segments annotated is



(a) Frequency of annotated noun classes across all frames



(b) Frequency of recorded tasks

(c) Frequency of annotated verb

(d) Action wordle

Figure 3.3: Statistics of the LEMMA dataset.

lower than 0.5.

Atomic-actions and Activities: Given the verbs of the atomic-action segments, two AMT workers were asked to fill in the blanks of the verb patterns and annotate the governing tasks in multi-task scenarios with a self-developed interactive annotation tool (see *supplementary material*). We allow concurrent actions for each agent with multiple nouns for the same verb; for example, “get spoon, cup from table using hand.” As there might exist ambiguities in describing the atomic-actions with natural languages, such as the possible annotations of “wash cup using water” *vs.* “wash cup using sink,” we manually go through all the annotations and resolve the ambiguous action annotations following a uniform criterion. Examples of annotation results are shown in

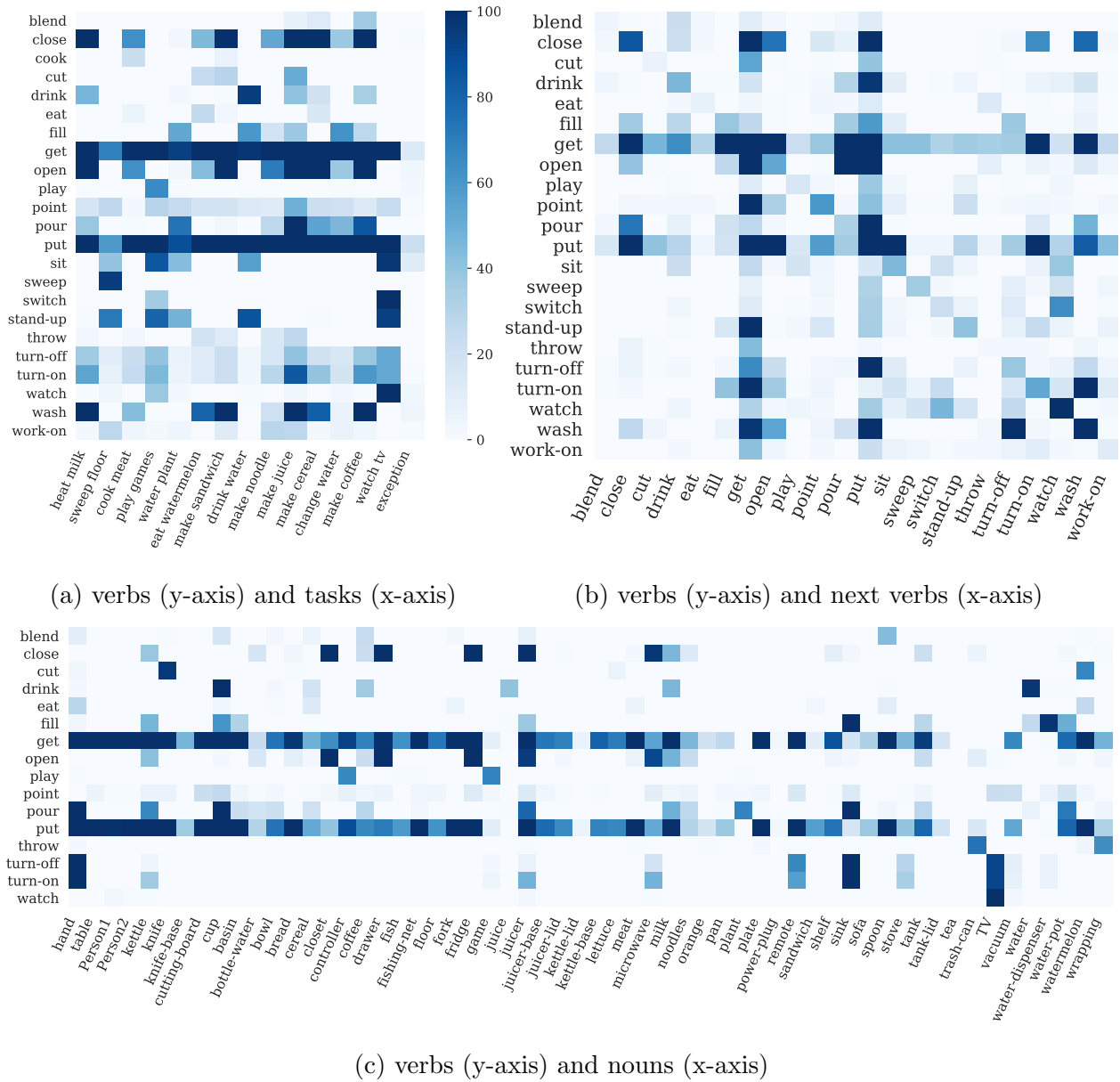


Figure 3.4: The co-occurrence statistics for verbs, nouns, and tasks in LEMMA.

supplementary.

Dataset Statistics

In total, we recorded 324 activities, generating 324×2 TPV videos (from both Kinects) and 445 FPV videos. Among them, 136 activities were performed in kitchens and the remaining 188 in the living rooms. The collected LEMMA dataset consists of 127 1×1 activities, 76 1×2 activities, 66 2×1 activities, and 55 2×2 activities. The frequency of the recorded tasks is shown in Fig. 3.3b. The total duration of all the activities is 10.1 hours, with an average duration of 2 minutes per video and the longest activity of 7 minutes.

We retrieved a total of 4.6 million images during post-processing, including 2.9 million RGB images captured by both GoPros and Kinects and 1.7 million depth images captured by Kinects. We annotated 0.9 million RGB frames captured by the primary view Kinect and gathered 0.8 million annotated frames with one or more actions performed by each of the agents (if multiple).

After resolving annotation ambiguities, we collected 24 verb classes and 64 noun classes, resulting in 862 compositional atomic-action labels, of which 641 appear more than 50 times. We show the frequencies of annotated verbs and nouns in Figs. 3.3a and 3.3c; both distributions roughly follow the Zipf’s law.

Co-occurrence relations among annotated verbs, nouns, and tasks are shown in Fig. 3.4. As we can see from Figs. 3.4a and 3.4c, verbs like “get” and “put” co-occur with various nouns in almost all of the tasks, which aligns with our intuition that moving objects around consists a large portion of our daily activities. Interactive actions between participants are captured by verbs (e.g., “point-to”) and nouns (e.g., “P1,” short for “participant 1”) in the form of annotations like “get knife from P1 using hand” or “point-to sink.”

3.4 Benchmarks

Aligned with our motivations, two general goals are constructed to evaluate indoor human activity understanding on the collected LEMMA dataset: (i) recognize atomic-actions and their semantics; and (ii) understand the goal-directed activities and monitor multiple concurrent tasks, especially in multi-agent scenarios. Specifically, we define two challenging benchmarks to test the capability

of understanding complex goal-directed activities for computer vision algorithms.

Compositional Action Recognition

Human indoor activities are composed of fine-grained action segments with rich semantics. As mentioned by Goyal et al. [GKM17], interactions with objects are highly purposive. From the simplest verb of “put,” we can generate a plethora of combinations of objects and target places, such as “put cup onto table,” “put fork into drawer.” Situations could become even more challenging when objects were used as tools; for example, “put meat into pan using fork.”

Motivated by the above observation, we propose the compositional action recognition benchmark on the collected LEMMA dataset with each object attributed to a specific semantic position in the action label. Specifically, we build 24 compositional action templates; see Fig. 3.5a for some examples. In these action templates, each noun could denote an interacting object, a target or a source location, or a tool used by a human agent to perform certain actions.

The proposed compositional action recognition benchmark is challenging; it requires computational models to correctly detect the ongoing concurrent action verbs as well as the nouns at their correct semantic positions. We evaluate model performances by metrics on compositional action recognition in both FPVs and TPVs. Specifically, the model is asked to predict (i) multiple labels in verb recognition for concurrent actions (e.g., “watch tv” and “drink with cup” at the same time), and (ii) multiple labels in noun recognition for each semantic position given verbs, representing the interactions with multiple objects using the same action (e.g., “wash spoon, cup using sink”). Fig. 3.5b shows the schematics of the evaluation process. For training and testing on TPVs, we provide ground-truth bounding boxes of humans as additional information on spatial localization.

Action and Task Anticipation

As emphasized throughout the work, the most significant factor of human activities is the goal-directed, teleological stand. An in-depth understanding of goal-directed tasks demands a predictive ability of latent goals, action preferences, and potential outcomes. To tackle these challenges, we propose the action and task anticipation benchmark on the collected LEMMA dataset. Specifically,

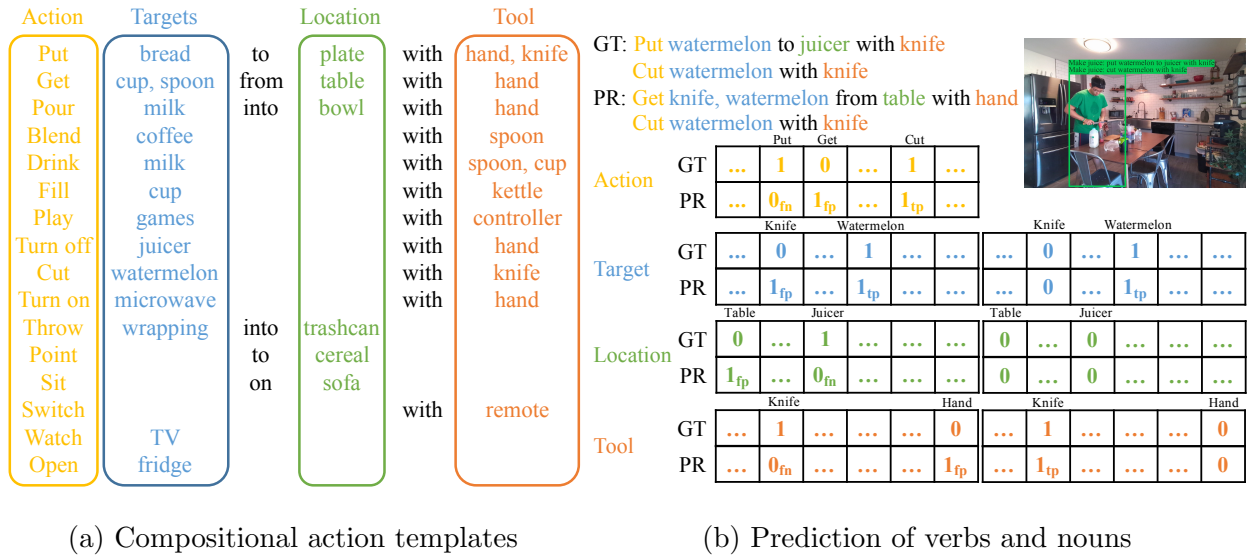


Figure 3.5: Compositional action recognition benchmark on LEMMA.

we evaluate model performances for the anticipation (i.e., predictions for the next action segment) of action and task with both FPV and TPV videos.

This benchmark provides both the training and testing data in all four scenarios of activities to study the goal-directed multi-task multi-agent problem. As there is an innate discrepancy of prediction difficulties among these four scenarios, we gradually increase the overall prediction difficulty, akin to a curriculum learning process, by setting the percentage of training videos to be 3/4, 1/4, 1/4, and 1/4 for 1×1 , 1×2 , 2×1 and 2×2 scenarios, respectively. Intuitively, with sufficient clean demonstrations of tasks in 1×1 scenario, interpreting tasks in more complex settings (i.e., 1×2 , 2×1 , and 2×2) should be easier, thus requiring less learning samples; such a design encourages the model to generalize. The model performance is evaluated individually for each scenario.

3.5 Experiments

In this section, we conduct experiments on the two proposed benchmarks with details on evaluation metrics, experimental settings, and baseline results. We further discuss the results to highlight the

underlying challenges of each task.

Compositional Action Recognition

Experimental Setup: We randomly split all the video samples into training and test sets with a ratio of 3:1, resulting in 243 recorded activities for training and the remaining 81 for testing. Due to the multi-agent setup, each activity may have multiple FPVs; 333 (out of 445) FPV videos are split into training. In TPVs, the recordings of the primary view with the ground-truth human bounding box annotations are given for both training and testing videos. Results are evaluated on two separate sources of inputs: FPVs and TPVs.

Evaluation Metrics: Model performances are evaluated separately for verbs, nouns, and compositional action recognition. Verb and compositional action recognition are treated as multi-label classifications with 25 verb classes and 863 compositional action classes (including a “null” action). After generating multi-hot labels for each semantic position in the presented verb, noun recognition is evaluated as multi-label classification (64 object classes). Average precision, recall, and F1-score for all predictions are reported on testing sets. During the evaluation, we sample image frames at 5 FPS and evaluate on these frames.

Methods: We adopt two recent 3D-CNN networks, I3D [CZ17] and SlowFast Network [FFM19], as the baseline models. The baseline models predict the compositional action directly. Considering compositionality of verbs and nouns, we propose two variants of the baseline models: (i) a multi-branch network (branching model) that builds on the bottleneck layer of the backbone models to leverage both verb and noun supervision, and (ii) a multi-step inference model (sequential model), wherein verbs are first inferred with a beam search and then fed into object inference with their verb embeddings for joint learning.

Implementation Details: The training procedure utilizes all annotated segments in the training set. Additionally, we re-scale all the images with the short side to 256 pixels. To feed data into 3D-CNN models, 4 frames are first sampled for each action segment as center frames, and

an additional 8 frames are then uniformly sampled around center frames with a window length of 32. We train each model on 8 Titan RTX GPUs on a single computing node for 50 epochs (20k iterations) with a batch size of 96. We use warm-up strategy and perform large mini-batch batch normalization, as suggested in [GDG17]. The learning rate is initially set to 0.0125 for each parallel branch and decays with a cosine annealing. Other settings of the backbone models are the same as in [FFM19]. For the proposed sequential model, we use the beam search with a size of 5 for action inference. We extract bounding box features of humans with ROIAlign [HGD17] for frames in TPVs. More implementation details are provided in *supplementary material*.

Results and Discussion: Table 3.2 shows quantitative results of predicting verbs, nouns, and compositional actions for the compositional action recognition task. For FPVs, rather than directly predicting the compositional actions (baseline models), predicting the verbs and nouns with their semantic positions boosts the performance on all metrics, indicating that understanding the compositional structures of human actions indeed supports the prediction. We also observe that the results of compositional action recognition in the sequential models are slightly lower than the branching model due to the aggregated error brought in by a relatively low precision ($\sim 25\%$) of the verb recognition.

In comparison, the results of compositional action recognition in TPVs are significantly lower than those in the FPVs due to severe occlusion. It also shows that predicting the composition of verbs and nouns makes no significant improvement compared with predicting compositional action directly. Such a result implies that current models could not capture the details of compositions between verbs and nouns from TPVs. Taken together, the results indicate that fusion among the representations of visual embodiment between TPVs and FPVs might be a crucial ingredient to tackle this problem in the future.

Fig. 3.6 shows qualitative results for the composed action recognition task.

Table 3.2: Comparisons of compositional action recognition on LEMMA.

View Type	Method	Verb			Noun			Compositional Action		
		Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1
FPV	I3D	17.09	43.89	24.60	3.42	16.15	5.72	11.07	39.49	17.30
	Slowfast	22.27	56.42	31.94	4.31	20.60	7.13	18.68	50.65	27.3
	I3D sequential	25.04	57.00	34.80	19.36	75.29	30.80	18.00	50.04	26.47
	Slowfast sequential	24.30	49.71	32.64	17.95	59.11	27.54	26.80	38.41	31.57
	I3D branching	25.73	55.62	35.8	18.63	69.76	29.41	22.29	48.46	30.53
	Slowfast branching	26.16	56.33	35.73	18.18	73.46	29.15	27.97	48.87	35.58
TPV	I3D	14.18	36.34	20.40	2.29	11.05	3.79	6.85	23.82	10.64
	Slowfast	14.28	37.38	20.66	2.32	11.14	3.83	7.76	23.25	16.31
	I3D sequential	16.17	30.17	21.05	7.79	25.41	11.93	2.23	12.67	3.79
	Slowfast sequential	15.31	28.84	20.00	6.37	22.39	9.92	3.27	9.16	4.82
	I3D branching	12.92	32.09	18.43	12.75	17.70	14.82	4.67	20.76	7.6
	Slowfast branching	16.64	33.40	22.21	17.29	18.36	17.81	6.52	21.55	10.01

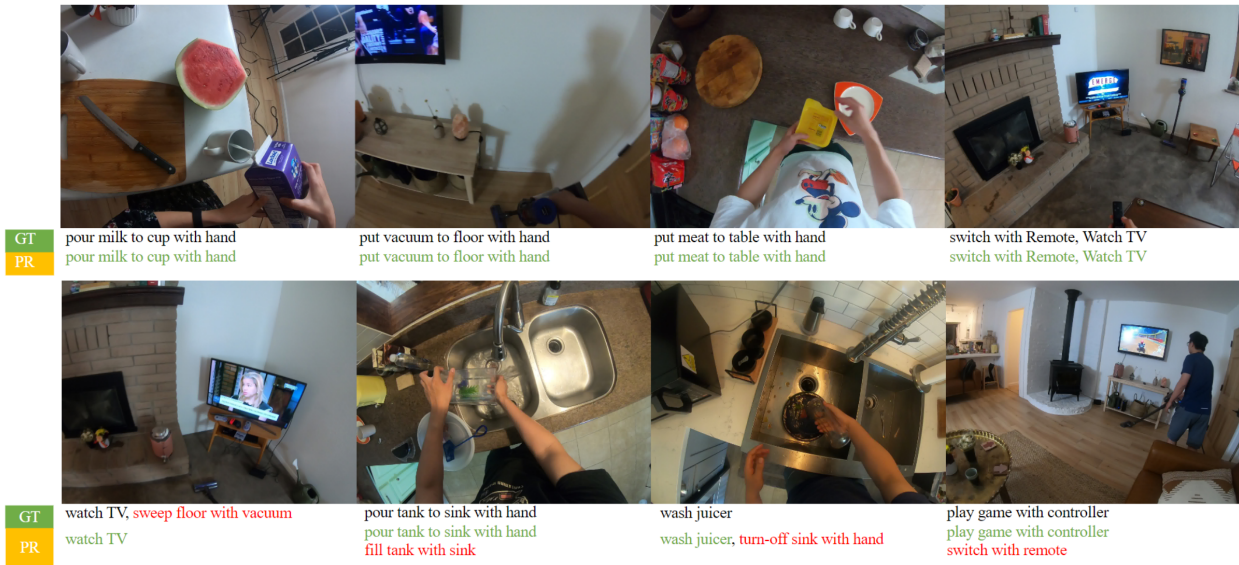


Figure 3.6: Qualitative results of compositional action recognition on LEMMA.

Action and Task Anticipations

Experimental Setup: We split the training and test sets with ratios 3 : 1, 1 : 3, 1 : 3, 1 : 3 for the four scenarios 1×1 , 1×2 , 2×1 , 2×2 , respectively. Such a split results in training set with

(96, 19, 16, 13) activities and a test set with (31, 57, 50, 42) activities in four scenarios. During training and testing, the computational models have access to both FPVs and TPVs, together with the ground-truth human bounding boxes annotations of the TPV primary view.

Evaluation Metrics: Model performances are evaluated individually (per agent) for the action and task anticipations task. Specifically, both action and task anticipations are evaluated as multi-label classifications with 863 compositional action classes (including a “null” action) and 15 task classes. Average precision, recall, and F1-score are reported individually for each of the four scenarios on the testing sets. Similar to the protocol used in the above compositional action recognition task, we re-sample image frames at 5 FPS and evaluate these sub-sampled frames during the testing phase.

Methods: We leverage the visual features extracted by the pre-trained SlowFast model in compositional action recognition for baseline models. Specifically, we compare two backbone models: (i) using segment-level recognition feature (SF) directly by adding an MLP on top of the features, and (ii) using long-term feature bank (LFB) with max pooling [WFF19]. For activities with multi-agent interactions, we use the other agent’s FPV features together with their own’s to capture the joint task execution progress for learning and inference; these variants are denoted as M-SF (FPV) and M-LFB (FPV) For comparison, we also use the concatenation of the FPV feature and primary TPV feature as the input; the corresponding models are denoted as M-SF (TPV) and M-LFB (TPV).

Implementation Details: For the LFB model, we use a history window size of 10 and aggregate the features using max-pooling, as described in [WFF19]. For the multi-agent variants, we use max-pooling to fuse features of two views and process them with a different branch as another temporal inference module. We train models on a single Titan Xp GPU for 50 epochs with a learning rate of 0.001. See *supplementary material* for more details on network architectures.

Results and Discussion: Table 3.3 shows quantitative results of action and task anticipation. The proposed multi-agent variants (M-) of baseline models perform the best among all models.

Table 3.3: Comparisons of the action and task anticipations on LEMMA.

Scenario	Method	1×1			1×2			2×1			2×2		
		Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1
Compositional action	SF	23.42	22.25	22.82	20.13	20.06	20.10	18.89	19.22	19.05	18.31	16.67	17.45
	LFB	23.03	28.67	25.54	20.48	25.4	22.67	18.31	22.30	20.11	18.53	20.97	19.68
	M-SF (TPV)	24.22	28.05	25.99	20.10	24.48	22.08	19.15	16.71	17.85	19.64	15.18	17.12
	M-LFB (TPV)	23.54	37.81	29.01	21.10	31.86	25.39	19.67	21.03	20.33	20.11	20.30	20.15
	M-SF (FPV)	23.30	25.41	24.31	21.34	23.18	22.22	19.70	17.46	18.51	19.82	15.8	17.58
	M-LFB (FPV)	23.26	31.07	26.60	20.78	27.40	23.63	19.42	21.73	20.51	19.49	20.12	19.8
Task	SF	50.53	79.08	61.66	48.07	67.78	56.25	39.05	57.43	46.49	44.88	62.09	52.1
	LFB	57.57	84.31	68.42	52.12	68.94	59.36	38.40	53.08	44.56	48.17	64.61	55.19
	M-SF (TPV)	58.61	79.96	67.05	55.45	67.24	60.78	45.73	58.98	51.51	49.66	64.47	56.10
	M-LFB (TPV)	60.27	82.19	69.54	56.2	72.46	63.30	43.94	61.41	51.23	48.85	67.48	56.67
	M-SF (FPV)	51.12	79.18	62.13	48.42	69.04	56.92	41.00	58.11	48.08	46.04	65.97	54.24
	M-LFB (FPV)	55.56	82.83	66.51	52.22	70.01	59.82	41.33	64.49	50.38	46.65	69.59	55.86

For single-agent activities (1×1 , 1×2), we have the following crucial observations. First, models that consider temporal relations between frames generally perform better than the models using segment features. Second, adding additional TPV features to single-agent activities slightly helps interpret the task being executed and therefore promotes anticipation. This result matches the intuition that computational models having access to both FPVs and TPVs would perceive more holistic scene information. We also find that the performances of task anticipation in the 1×1 single-task scenario are better than the one in the 1×2 multi-task scenario, matching what we would expect from more complicated task execution patterns.

For multi-agent activities (2×1 , 2×2), we observe that the aggregation of FPV and TPV features generally performs better. It supports our hypothesis that observing the other agents’ actions helps the computational models to “understand” task scheduling and assignment. We also observe that, models’ performances in 2×1 activities are slightly worse than in 2×2 activities. We hypothesize that task plans in the 2×2 scenarios change less frequently, with a clear task assignment coordinates the individual tasks. In comparison, in the 2×1 scenarios, the sequential ordering of the task requires more frequent communications between agents to coordinate. Such a performance gap calls for better modeling of multi-agent task assignments. Due to the page limit, we show qualitative results of action and task anticipation in the *supplementary material*.

3.6 Conclusions

In this work, we introduce the LEMMA dataset with a focus on natural multi-agent multi-task daily activities. Dense annotations are provided on both compositional action and task for learning and inference on four different activity scenarios with increasing difficulty. Additionally, we propose two challenging tasks on LEMMA to measure existing models' competence in action understanding and temporal reasoning: (i) compositional action recognition, and (ii) action/task anticipations. We hope this effort would attract the computer vision community to look into natural and realistic goal-directed human activities and further study the task scheduling and assignment in real-world scenarios.

CHAPTER 4

Human Communication in Shared Attention and Embodied Reference

In this chapter, we will talk another important aspect in scene and activity understanding: human communications comprehension, in which shared attention and referential behavior are two typical form involving both verbal and non-verbal signals. Both conveys vivid and complex messages and play critical role towards social interaction and Theory of Mind (TOM). We first study the machine’s understanding of embodied reference where one agent uses both language and gesture to refer to an object to another agent in a shared physical environment. Then we address the problem of inferring shared attention in third-person social scene videos.

4.1 YouRefIt: Embodied Reference Understanding with Language and Gesture

4.1.1 Introduction

Human communication [Tom10] relies heavily on establishing common ground [TSZ20, SZZ20] by referring to objects in a shared environment. This process usually takes place in two forms: language (abstract symbolic code) and gesture (unconventionalized and uncoded). In the computer vision community, efforts of understanding reference have been primarily devoted in the first form through an artificial task, Referring Expression Comprehension (REF) [YPY16, HRA17, YLS18, LWS19, YRL19, YLY19, YLY20], which localizes a particular object in an image with a natural language expression generated by the annotator. Evidently, the second form, gesture, has been



Figure 4.1: Embodied reference in daily deictic-interaction scenario.

left almost untouched. Yet, this nonverbal (gesture) form is more profound in the communication literature compared to the pure verbal (language) form with ample evolutionary evidence [ALP08, McN12, HRT13]; it is deeply rooted in human cognition development [LCH04, LCS06] and learning process [CMG08], and tightly coupled with the language development [Kit03, CSK10, IG05].

Fundamentally, most modern literature deviates from the natural setting of reference understanding in daily scenes, which is **embodied**: An agent refers to an object to another in a *shared* physical space [QLZ20, WWZ21, FQZ21], as exemplified by Fig. 4.1. Embodied reference possesses two distinctive characteristics compared to REF. First, it is **multimodal**. People often use both natural language and gestures when referring to an object. The gestural component and language component are semantically coherent and temporally synchronous to coordinate with one another, creating a concise and vivid message [Ken04] while elucidating the overloaded meaning if

only one modality is presented [JSW21]. Second, recognizing embodied reference requires visual **perspective-taking** [KF91, BES97, QLZ20], the awareness that others see things from different viewpoints and the ability to imagine what others see from their perspectives. It requires both the message sender and receiver to comprehend the immediate environments [FQZ21], including the relationship between the interlocutors and the relationships between objects, in the shared perceptual fields for effective communication.

To address the deficiencies in prior work and study reference understanding at a full spectrum, we introduce a new dataset, **YouRefIt**, for embodied reference understanding. The reference instances in *YouRefIt* are crowd-sourced with diverse physical scenes from Amazon Mechanic Turk (AMT). Participants are instructed to film videos in which they reference objects in a scene to an imagined person (i.e., a mounted camera) using both language and gestures. Minimum requirements of the scenes, objects, and words are imposed to ensure the naturalness and the variety of collected videos. Videos are segmented into short clips, with each clip containing an exact one reference instance. For each clip, we annotate the reference target (object) with a bounding box. We also identify **canonical frames** in a clip: They are the “keyframes” of the clip and contain sufficient information of the scene, human gestures, and referenced objects that can truthfully represent the reference instance. Fine-grained semantic parsing of the transcribed sentences is further annotated to support a detailed understanding of the sentences. In total, the *YouRefIt* dataset includes 4,195 embodied reference instances from 432 indoor scenes.

To measure the machine’s ability in Embodied Reference Understanding (ERU), we devise two benchmarks on top of the proposed *YouRefIt* dataset. (i) **Image ERU** takes a canonical frame and the transcribed sentence of the reference instance as the inputs and predicts the bounding box of the referenced object. Image ERU adopts the settings from the well-studied REF but is inherently more challenging and holistic due to its requirement on a joint and coherent understanding of human gestures, natural language, and objects in the context of human communication. (ii) **Video ERU** takes the video clip and the sentence as the input, identifies the canonical frames, and locates the reference target within the clip. Compared to Image ERU, Video ERU takes one step further and manifests the most natural human-robot communication process that requires distinguishing the initiation, the canonical frames, and the ending of a reference act while estimating the reference

target in a temporal order.

Incorporating both language and gestural cues, we formulate a new multimodal framework to tackle the ERU tasks. In experiments, we provide multiple baselines and ablations. Our results reveal that models with explicit gestural cues yield better performance, validating our hypothesis that gestural cues are as critical as language cues in resolving ambiguities and overloaded semantics with cooperation (perspective-taking) in mind [JSW21, JCH20, QLZ20, YLF20, ZGF20], echoing a recent finding in the embodied navigation task [WWZ21]. We further verify that temporal cues are essential in canonical frame detection, necessitating understanding embodied reference in dynamic and natural sequences.

This work makes three major contributions. (i) We collect the first video dataset in physical scenes, *YouRefIt*, to study the reference understanding in an *embodied* fashion. We argue this is a more natural setting than prior work and, therefore, further understanding human communications and multimodal behavior. (ii) We devise two benchmarks, Image ERU and Video ERU, as the protocols to study and evaluate the embodied reference understanding. (iii) We propose a multimodal framework for ERU tasks with multiple baselines and model variants. The experimental results confirm the significance of the joint understanding of language and gestures in embodied reference.

4.1.2 Related Work

Our work is related to two topics in modern literature: (i) Referring Expression Comprehension (REF) studied in the context of Vision and Language, and (ii) reference recognition in the field of Human-Robot Interaction. Below, we compare our work with prior arts on these two topics.

4.1.2.1 Referring Expression Comprehension (REF)

REF is a visual grounding task. Given a natural language expression, it requires an algorithm to locate a particular object in a scene. Several datasets, including both images of physical scenes [KOM14, YPY16, MHT16, PWC15, DSC17, CWM20, CBM20, AAX20] and synthetic images [LLB19], have been constructed by asking annotators or algorithms to provide utterances describing regions of images. To solve REF, researchers have attempted various approaches [YRL19,

Table 4.1: Comparisons between YouRefIt and other reference datasets. happens.

Datasets	Lang.	Gest.	Embo.	Type	Source	No. of images	No. of instances	No. of object categories	Ave. sent. length
PointAt [SRF10]	✗	✓	✓	image	lab	220	220	28	-
ReferAt [SF10]	✓	✓	✓	video	lab	-	242	28	-
IPO [SEP15]	✗	✓	✓	image	lab	278	278	10	-
IMHF [SEP16]	✗	✓	✓	image	lab	1716	1,716	-	-
RefIt [KOM14]	✓	✗	✗	image	image CLEF	19,894	130,525	238	3.61
RefCOCO [YPY16]	✓	✗	✗	image	MSCOCO	19,994	142,209	80	3.61
RefCOCO+ [YPY16]	✓	✗	✗	image	MSCOCO	19,992	141,564	80	3.53
RefCOCOg [MHT16]	✓	✗	✗	image	MSCOCO	26,711	104,560	80	8.43
Flickr30k entities [PWC15]	✓	✗	✗	image	Flickr30K	31,783	158,915	44,518	-
GuessWhat? [DSC17]	✓	✗	✗	image	MSCOCO	66,537	155,280	-	-
Cops-Ref [CWM20]	✓	✗	✗	image	COCO/Flickr	75,299	148,712	508	14.40
CLEVR-Ref+ [LLB19]	✓	✗	✗	image	CLEVR	99,992	998,743	3	22.40
<i>YouRefIt</i>	✓	✓	✓	video	crowd-sourced	497,348	4,195	395	3.73

LWS19, YLY19, YLY20]. Representative methods include (i) localizing a region by reconstructing the sentence using an attention mechanism [RRH16], (ii) incorporating contextual information to ground referring expressions [ZNC18, YPY16], (iii) using neural modular networks to better capture the structured semantics in sentences [HRA17, YLS18], and (iv) devising a one-stage approach [YGW19, YCW20]. In comparison, our work fundamentally differs from REF at two levels.

Task-level REF primarily focuses on building correspondence between visual cues and verbal cues (natural language). In comparison, the proposed ERU task mimics the minimal human communication process in an embodied manner, which requires a mutual understanding of both verbal and nonverbal messages signaled by the sender. Recognizing references in an embodied setting also introduces new challenges, such as visual perspective-taking [GMG08]: The referrers need to consider the perception from the counterpart’s perspective for effective verbal and nonverbal communication, requiring a more holistic visual scene understanding both geometrically and semantically. In this work, to study the reference understanding that echoes the above characteristics, we collect a new dataset containing natural reference scenarios with both language and gestures.

Model-level Since previous REF approaches are only capable of comprehending communicative messages in the form of natural language and mostly ignore the gestural cues, it is insufficient in the ERU setting or to be applied in our newly collected dataset. To tackle this deficiency, we design a principled framework to combine verbal (natural language) and nonverbal (gestures) cues. The proposed framework outperforms prior single-modality methods, validating the significant role of the gestural cue in addition to the language cue in embodied reference understanding.

4.1.2.2 Reference in Human-Robot Interaction

The combination of verbal and nonverbal communication for reference is one of the central topics in Human-Robot Interaction. Compared with REF, this line of work focuses on more natural settings but with specialized scenarios. One stream of work emphasizes pointing direction and thus are not object-centric while missing language reference: The Innsbruck Pointing at Objects dataset [SEP15] investigates two types of pointing gestures with index finger and tool, and the Innsbruck Multi-View Hand Gesture Dataset [SEP16] records hand gestures in the context of human-robot interaction in close proximity. The most relevant prior arts are ReferAt [SF10] and PointAt [SRF10], wherein participants are tasked to point at various objects with or without linguistic utterance. Some other notable literature includes (i) a robotics system that allows users to combine natural language and pointing gestures to refer to objects on a display [KAR86], (ii) experiments that investigate the semantics and pragmatics of co-verbal pointing through computer simulation [LPR15], (iii) deictic interaction with a robot when referring to a region using pointing and spatial deixis [HSK10], and (iv) effects of various referential strategies, including talk-gesture-coordination and handshape, for robots interacting with humans when guiding attentions in museums [PW14].

Although related, the above literature is constrained in lab settings with limited sizes, scenarios, and expressions, thus insufficient for solving the reference understanding in natural, physical scenarios with both vision and language. In comparison, crowd-sourced by AMT, our dataset is much more diverse in environment setting, scene appearance, and types of utterance. Our dataset also collects videos instead of static images commonly used in prior datasets, opening new venues to study dynamic and evolutionary patterns that occurred during natural human communications.

4.1.3 The YouRefIt Dataset

To study the embodied reference understanding, we introduce a new dataset named *YouRefIt*, a video collection of people referring to objects with both natural language and gesture in indoor scenes. Table 4.1 tabulates a detailed comparison between *YouRefIt* against twelve existing reference understanding datasets. Compared to existing datasets collected either in laboratories or from the Internet (MSCOCO/Flickr) or simulators (CLEVR), *YouRefIt* has a clear distinction: It contains videos crowd-sourced by AMT, and thus the reference happens in a more natural setting with richer diversity. Compared with the datasets on referring expression comprehension, the referrers (human) and the receivers (camera) in our dataset share the same physical environment, with both language and gesture allowed for referring to objects; the algorithm ought to understand from an embodiment perspective to tackle this problem. Next, we discuss the data collection and annotation process details, followed by a comprehensive analysis.

4.1.3.1 Data Collection

Our dataset was collected via AMT; see the data collection process in Fig. 4.2. Workers were asked to record a video containing actions of referring to objects in the scene to an imagined person (i.e., the camera) using both natural languages (sentences) and pointing gestures. Most videos were collected in indoor scenes, such as offices, kitchens, and living rooms. Unlike existing datasets in which objects are usually put on a table with a clean background, all the objects in our collected videos were placed at their natural positions. Each video also included more than ten objects in the scene to avoid trivial scenarios and increase the reference difficulty. The camera was set up such that the referrer and all referred objects are within the field of view.

When referring to a specific object, participants were instructed to use arbitrary natural languages and gestures freely. However, they were also required to avoid potential ambiguities, such that the observer would be able to uniquely identify the referred object by merely observing the reference behaviors. After reference actions were finished, participants were instructed to tap the referred object; this extra step helps annotate the referred target. In addition to the voices recorded in the video, participants were also asked to write down the sentences after the recording.

Task: Refer to an object in the scene to an imagined person (camera)

Steps:

1. Refer to one object using both pointing gesture and language.
2. After the reference, tap the target object to confirm.
3. Repeat until no more objects.
4. Write down the sentences in the same order as during the recording.
5. Submit both the videos and sentences.

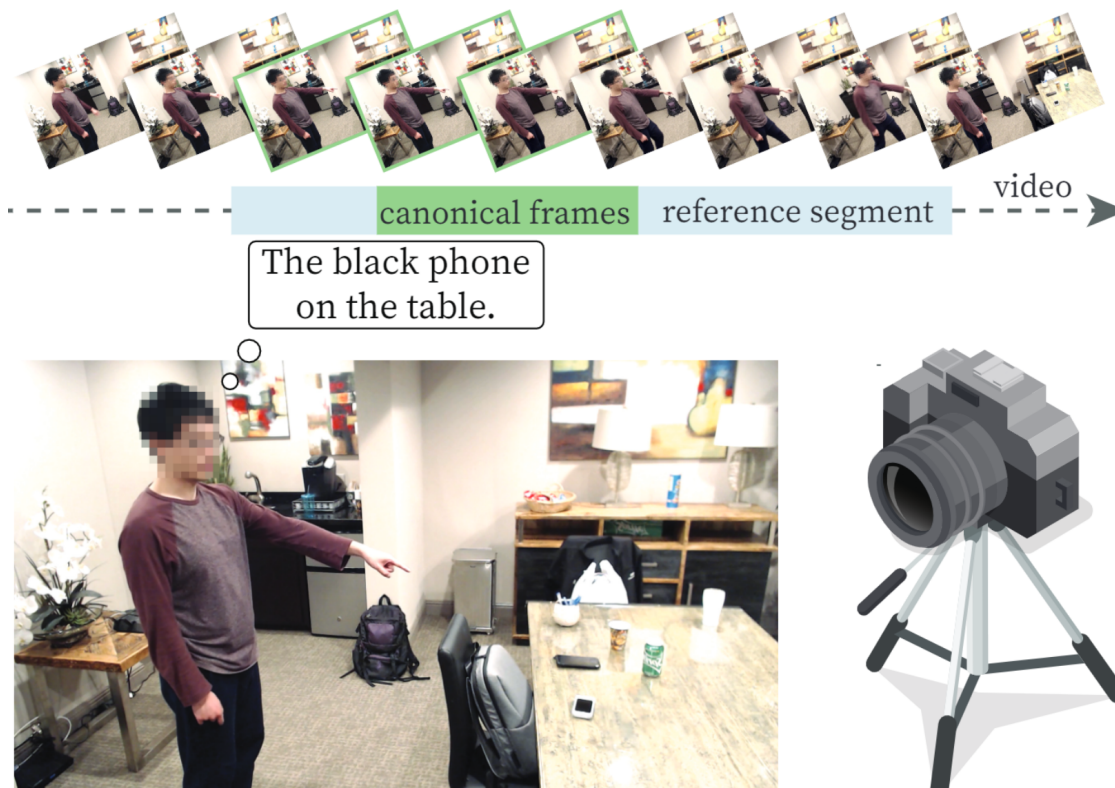


Figure 4.2: Illustration of the YouRefIt dataset collection procedure.

4.1.3.2 Data Annotation

The annotation process takes two stages: (i) annotation of temporal segments, canonical frames, and referent bounding boxes, and (ii) annotation of sentence parsing. Please refer to the *supplementary material* for more details of the data post-processing and annotation process.

Segments Since each collected video consists of multiple reference actions, we first segment the video into clips; each contains an exact one reference action. A segment is defined from the start of gesture movement or utterance to the end of the reference, which typically includes the raise of hand and arm, pointing action, and reset process, synchronized with its corresponding language description.

Canonical Frames In each segment, the annotators were asked to annotate further the canonical moments, which contain the “keyframes” that the referrer holds the steady pose to indicate what is being referred clearly. Combined with natural language, it is sufficient to use any canonical frame to localize the referred target.

Bounding Boxes Recall that participants were instructed to tap the referred objects after each reference action. Using this information, bounding boxes of the referred objects were annotated using Vatic [VPR13b], and the tapping actions were discarded. The object color and material were also annotated if identifiable. The taxonomy of object color and material is adopted from Visual Genome dataset [KZG17].

Sentence Parsing Given the sentence provided by the participants who performed reference actions, AMT annotators were asked to refine the sentence further and ensure it matches the raw audio collected from the video. We further provided more fine-grained parsing results of the sentence for natural language understanding. AMT annotators annotated target, target-attribute, spatial-relation, and comparative-relation. Take “The largest red bottle on the table” as an example: “the bottle” will be annotated as the target, “red” as target-attribute, “on the table” as spatial-relation, and “largest” as comparative-relation. For each relation, we further divided them into “relation” (e.g., “on”) and “relation-target” (e.g., “the table”).

4.1.3.3 Dataset Statistics

In total, *YouRefIt* includes 432 recorded videos and 4,195 localized reference clips with 395 object categories. We retrieved 8.83 hours of video during the post-processing and annotated 497,348

4.1.4 Embodied Reference Understanding (ERU)

In this section, we benchmark two tasks of embodied reference understanding on the *YouRefIt* dataset, namely, Image ERU and Video ERU. The first benchmark evaluates the performance of understanding embodied reference based on the canonical frame, whereas the second benchmark emphasizes how to effectively recognize the canonical moments and reference targets simultaneously in a video sequence. Below, we describe the detailed settings, baselines, analyses, and ablative studies in the experiments.

Dataset Splits We randomly split the dataset into the training and test sets with a ratio of 7:3, resulting in 2,950 instances for training and 1,245 instances for testing.

4.1.4.1 Image ERU

Given the canonical frame and the sentence from an embodied reference instance, Image ERU aims at locating the referred object in the image through both the human language and gestural cues.

Experimental Setup and Evaluation Protocol For each reference instance, we randomly pick one frame from the annotated canonical frames. We adopt the evaluation protocol similar to the one presented in Mao et al. [MHT16]: (i) predict the region referred by the given image and sentence, (ii) compute the IoU ratio between the ground-truth and the predicted bounding box, and (iii) count it as correct if the IoU is larger; otherwise wrong. We use accuracy as the evaluation metric. Following object detection benchmark [GLU12], we report the results under three different IoUs: 0.25, 0.5, and 0.75.

We also evaluate on subsets with various object sizes, i.e., *small*, *medium* and *large*. Object size is estimated using the ratio between the area of the ground-truth object bounding box and the area of the image. The size thresholds are 0.48% and 1.76% based on the size distribution in the dataset; see the size distribution in *supplementary material*.

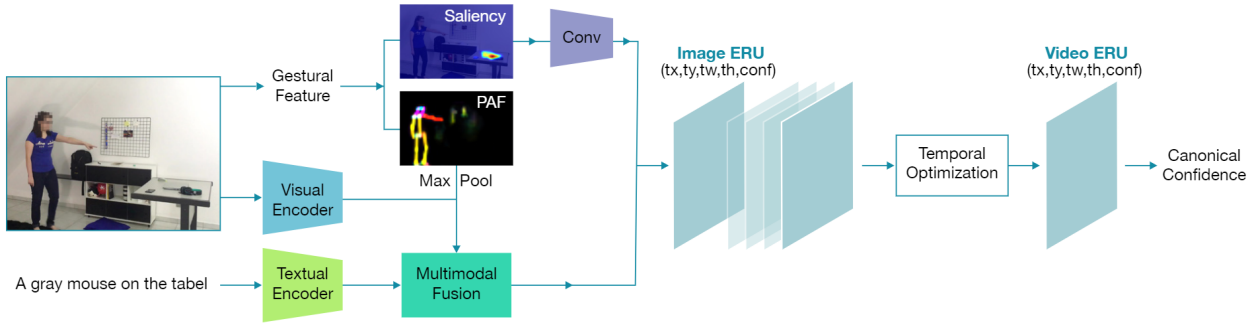


Figure 4.4: The proposed multimodal framework for the ERU task.

Methods We devise a novel multimodal framework for Image ERU that leverages both the language and gestural cues; see Fig. 4.4. At a high level, our framework includes both the visual and language encoder, similar to prior REF models [YGW19, YCW20, LZS20], as well as explicitly extracted gesture features. We utilize the features from three modalities to effectively predict the target bounding box.

Specifically, we use Darknet-53 [RF18] pre-trained on COCO object detection [LMB14] as the visual encoder. The textual encoder is the uncased base version of BERT [DCL18] followed by two fully connected layers. We incorporate two types of gestural features: (i) the Part Affinity Field (PAF) [CMS19] heatmap, and (ii) the pointing saliency heatmap. Inspired by the visual saliency prediction, we train MSI-Net [KSD20] on the *YouRefIt* dataset to predict the salient regions by considering both the latent scene structure and the gestural cues, generating more accurate guidance compared to the commonly used Region of Interests (RoIs); see some examples of predicted salient regions in Fig. 4.5. We aggregate the visual feature and PAF heatmaps by max-pooling and concatenation, fusing them with textual features by updating text-conditional visual features attended to different words through a sub-query module [YCW20]. Following convolution blocks, the saliency map feature is concatenated with the text-conditional visual feature as the high-level guidance to predict anchor boxes and confidence scores; we use the same classification and regression loss as in Yang et al. [YGW19] for anchor-based bounding box prediction.

Baselines and Ablations We first evaluate the Image ERU performance on FAOA [YGW19] and ReSC [YCW20], originally designed for the REF task. We also design baselines to test the

gestural cues in a two-stage architecture, similar to MAttNet [YLS18]. We generate the RoIss by Region Proposal Network from Faster R-CNN [RHG16] pre-trained on the MSCOCO dataset. To score the object proposal, we test two categories of heatmaps that reflect the gestural cues. (i) By pointing heatmap from the primary pointing direction characterized by arm, hand, and index finger. Following Fan et al. [FCW18], we generate the pointing heatmap by a Gaussian distribution to model the variation of a pointing ray w.r.t the primary pointing direction. We choose 15° and 30° as the standard deviations (i.e., $\text{RPN}_{\text{pointing}15}$ and $\text{RPN}_{\text{pointing}30}$). (ii) By pointing saliency map (i.e., $\text{RPN}_{\text{saliency}}$). The scores are computed according to the heatmap of average density.

We design ablation studies from two aspects: data and architecture. For the **data-wise** ablation, we first evaluate the MAttNet, FAOA, and ReSC models pre-trained on the REF datasets RefCOCO, RefCOCO+, and RefCOCOg, where the references are not embodied. Therefore, these three pre-trained models neglect the human gestural cues. Next, for a fair comparison without the gestural cues, we further generate an inpainted version of *YouRefIt*, where humans are segmented and masked by a pre-trained Mask R-CNN [HGD17], and the masked images are inpainted by DeepFill [YLY18b, YLY18a] pre-trained on the Places2 [ZLK17] dataset; see examples in Fig. 4.5. After the human gestural cues are masked out, we train FAOA and ReSC on the inpainted dataset, denoted as $\text{FAOA}_{\text{inpaint}}$ and $\text{ReSC}_{\text{inpaint}}$. For the **architecture-wise** ablation, we compare two variants of our proposed full model to evaluate the contribution of different components: (i) $\text{Ours}_{\text{no_lang}}$: without the language embedding module, and (ii) $\text{Ours}_{\text{PAF_only}}$: with the PAF heatmap as the only gestural cue; see the *supplementary material* for more details.

Results and Discussion Table 4.2 tabulates the quantitative results of the Image ERU, and Fig. 4.5 shows some qualitative results. We categorize the models based on their information sources: *Language-only*, *Gesture-only*, and *Language + Gesture*. Below, we summarize some key findings.

1. Gestural cues are essential for embodied reference understanding. As shown in Table 4.2, FAOA and ReSC models show significant performance improvement when trained on the original *YouRefIt* dataset compared to that on the inpainted version. Of note, in embodied reference, the referrer will adjust their own position to ensure the referred targets are not blocked by its body,



Figure 4.5: Qualitative results in Image ERU.

one of the main advantages introduced by perspective-taking. As such, the inpainted images always contain the reference targets with only gestural cues masked.

2. Language cues elucidate ambiguities where the gestural cues alone cannot resolve. As shown by the *Gesture-only* models, RPN+heatmap models possess ambiguities when presented with gestural cues alone; pointing gestures suppress the descriptions of target location and attend to spatial regions but are not object-centric. Without the referring expressions, the performance of $Ours_{no_lang}$ also deteriorates compared to $Ours_{Full}$.
3. Explicit gestural features are beneficial for understanding embodied reference. $Ours_{PAF_only}$, which incorporates PAF features that encode unstructured pairwise relationships between body parts, outperforms the original FAOA and ReSC models. By further adding the saliency heatmap, our full model $Ours_{Full}$ achieves the best performance in all baselines and ablations. Taken together, these results strongly indicate that the fusion of the language and gestural cues could be the crucial ingredient to achieving high model performance.

Human Performance We also conducted a human study of the embodied reference understanding task. We ask three Amazon Turkers to annotate the referred object bounding box in 1,000

Table 4.2: Comparisons of Image ERU performances on the YouRefIt dataset.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	<i>all</i>	<i>small</i>	<i>medium</i>	<i>large</i>	<i>all</i>	<i>small</i>	<i>medium</i>	<i>large</i>	<i>all</i>	<i>small</i>	<i>medium</i>	<i>large</i>
Language-only												
MAttNet _{pretrain}	14.2	2.3	4.1	34.7	12.2	2.4	3.8	29.2	9.1	1.0	2.2	23.1
FAOA _{pretrain}	15.9	2.1	9.5	34.4	11.7	1.0	5.4	27.3	5.1	0.0	0.0	14.1
FAOA _{impaint}	23.4	14.2	23.6	32.1	16.4	9.0	17.9	22.5	4.1	1.4	4.7	6.2
ReSC _{pretrain}	20.8	3.5	17.5	40.0	16.3	0.5	14.8	36.7	7.6	0.0	4.3	17.5
ReSC _{impaint}	34.3	20.3	38.9	44.0	25.7	8.1	32.4	36.5	9.1	1.1	10.1	16.0
Gesture-only												
RPN+Pointing ₁₅	15.3	10.5	16.9	18.3	10.2	7.2	12.4	11.0	6.5	3.8	9.1	6.6
RPN+Pointing ₃₀	14.7	10.8	17.0	16.4	9.8	7.4	12.4	9.8	6.5	3.8	8.9	6.8
RPN+Saliency[KSD20]	27.9	29.4	34.7	20.3	20.1	21.1	26.8	13.2	12.2	10.3	17.9	8.6
Ours _{no_lang}	41.4	29.9	48.3	46.3	30.6	17.4	37.0	37.4	10.8	1.7	13.9	16.6
Language + Gesture												
FAOA[YGW19]	44.5	30.6	48.6	54.1	30.4	15.8	36.2	39.3	8.5	1.4	9.6	14.4
ReSC[YCW20]	49.2	32.3	54.7	60.1	34.9	14.1	42.5	47.7	10.5	0.2	10.6	20.1
OursPAF _{only}	52.6	35.9	60.5	61.4	37.6	14.6	49.1	49.1	12.7	1.0	16.5	20.5
OursFull	54.7	38.5	64.1	61.6	40.5	16.3	54.4	51.1	14.0	1.2	17.2	23.3
Human	94.2±0.2	93.7±0.0	92.3±1.3	96.3±1.7	85.8±1.4	81.0±2.2	86.7±1.9	89.4±1.7	53.3±4.9	33.9±7.1	55.9±6.4	68.1±3.0

images randomly sampled from the test set. We report the average accuracy under different IoUs in Table 4.2. Humans achieve significantly higher accuracy than all current machine learning models, demonstrating the human’s outstanding capability to understand embodied references combined with language and gestural cues. The performance drops when the IoU threshold increases, especially for *small* and *medium* objects, indicating the difficulties in resolving the ambiguity in small objects.

4.1.4.2 Video ERU

Compared with Image ERU discussed above, Video ERU is a more natural and practical setting in human-robot interaction. Given a referring expression and a video clip that captures the whole dynamics of a reference action with consecutive body movement, Video ERU aims at recognizing the canonical frames and estimate the referred target at the same time.

Experimental Setup and Evaluation Protocol For each reference instance, we sample image frames with 5 FPS from the original video clip. Average precision, recall, and F1-score are reported for the canonical frame detection. For referred bounding box prediction, we report the averaged accuracy in all canonical frames.

Baselines To further exploit the temporal constraints in videos, we integrate a temporal optimization module to aggregate and optimize the multimodal feature extracted from the Image ERU. We test two designs of temporal optimization module: (i) ConvLSTM: a two-layer convolutional Long Short-Term Memory [SCW15], and (ii) Transformer: a three-layer Transformer encoder [VSP17] with four attention heads in each layer. After the temporal optimization module, we use the features of each frame to predict canonical frames and anchor bounding boxes simultaneously.

We further design a third *Frame-based* baseline that learns from the individual frame by adding two fully connected regression layers on top of our model in Image ERU. This *Frame-based* model takes all sampled frames from the video clip during training and testing.

During training, we add a binary cross-entropy loss for canonical frame detection on top of the loss function for bounding box prediction in the Image ERU framework. Please refer to the *supplementary material* for more details.

Table 4.3: Video ERU performance comparisons on the YouRefIt dataset.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	<i>all</i>	<i>small</i>	<i>medium</i>	<i>large</i>	<i>all</i>	<i>small</i>	<i>medium</i>	<i>large</i>	<i>all</i>	<i>small</i>	<i>medium</i>	<i>large</i>
Frame-based	55.2	42.3	58.9	64.8	41.7	22.7	53.4	48.8	16.9	1.6	21.8	27.0
Transformer	52.3	40.2	55.6	58.3	38.8	21.2	54.1	47.1	13.9	1.5	20.8	22.7
ConvLSTM	54.8	43.1	57.5	60.0	39.3	22.5	54.8	46.7	17.3	1.8	24.3	25.5
Ours _{Full}	54.7	38.5	64.1	61.6	40.5	16.3	54.4	51.1	14.0	1.2	17.2	23.3

Results and Discussion Table 4.3 shows quantitative results of predicting reference targets with the ground-truth canonical frames given a video. We observe that the frame-based method



Figure 4.6: Qualitative results in Video ERU

and the temporal optimization methods reach similar performance, comparable to the model that only trained on selected canonical frames (i.e., Ours_{Full}). This result indicates that the canonical frames can indeed provide sufficient language and gestural cues for clear reference purposes, and the temporal models may be distracted from non-canonical frames. This observation aligns with the settings of previous REF tasks. Meanwhile, as shown in Table 4.4 and Fig. 4.7, temporal information can significantly improve the performance of canonical frame detection; both the *ConvLSTM* and the *Transformer* model outperform the *Frame-based* method by a large margin. These results indicate the significance of distinguishing various stages of reference behaviors, e.g., initiation, canonical moment, and ending, for better efficacy in embodied reference understanding. Fig. 4.6 shows some qualitative results.

Table 4.4: Canonical frame detection performance.

Method	Avg. Prec	Avg. Rec	Avg. F1
Frame-based	31.9	37.7	34.5
Transformer	35.1	44.2	39.1
ConvLSTM	57.0	37.9	45.4

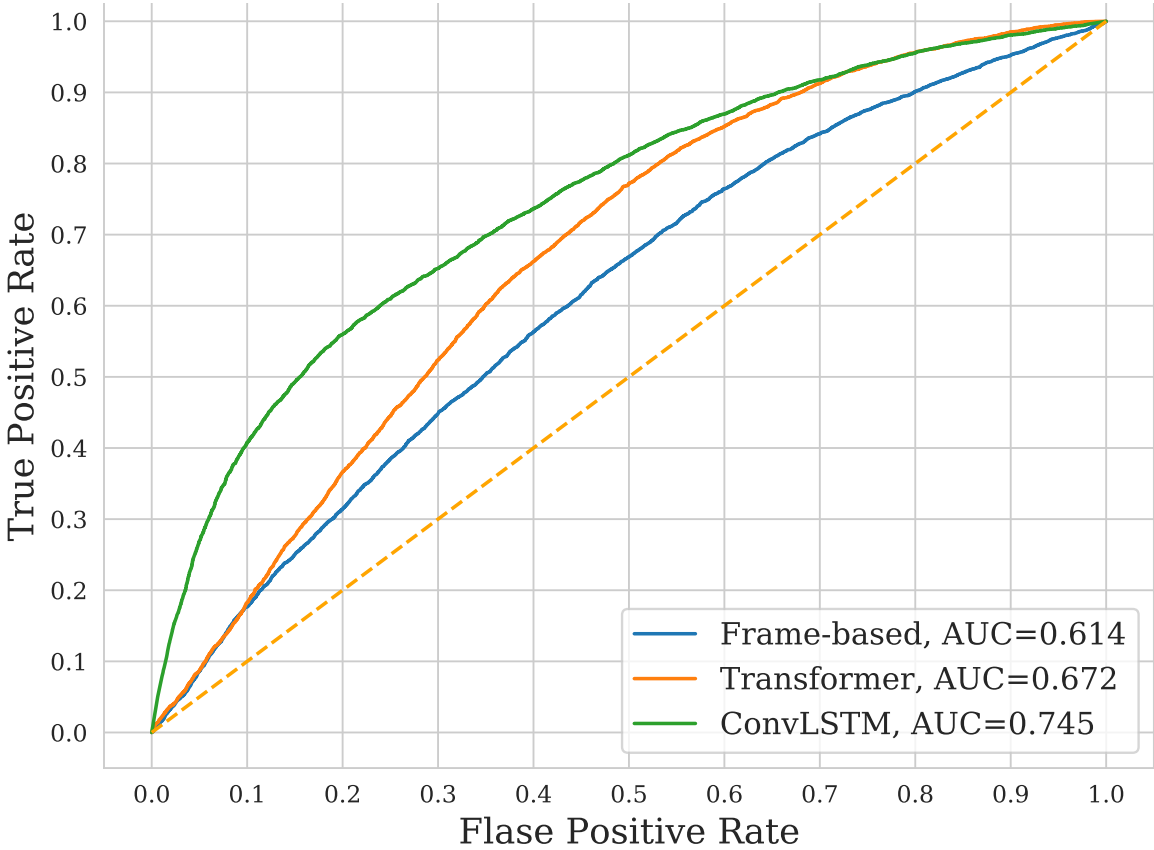


Figure 4.7: ROC Curve for canonical frame detection.

4.1.5 Conclusion and Future Work

We present the novel problem of embodied reference understanding. Such a setting with both language and gestural cues is more natural for understanding human communication in our daily activities. To tackle this problem, we crowd-source the *YouRefIt* dataset and devise two benchmarks on images and videos. We further propose a multimodal framework and conduct extensive experiments with ablations. The experimental results provide strong empirical evidence that language and gestural coordination is critical for understanding human communication.

Our work initiates the research on embodied reference understanding and can be extended to many aspects. For example, the difficulty in resolving reference ambiguity within a single-round communication, even for humans, calls for studying embodied reference using multi-round dialogues.

Human-robot interaction may benefit from referential behavior generation by considering scene contexts. We hope our work can inspire more future work on these promising directions, focusing on understanding human communication from multimodal (verbal/nonverbal) inputs.

4.2 Inferring Shared Attention in Social Scene Videos

4.2.1 Introduction

Shared attention is defined as the attention focus shared by two or more individuals on one object or human [Eme00]. Shared attention differs from joint attention in a subtle way and in the literature the two terms are used interchangeably [Eme00]. Shared attention is everywhere in our daily life and we can observe it every now and then in almost all social interactions. Imagine in a party, usually humans can easily recognize a group of people with shared attention and what exactly is their shared attention in the group at present. They can join the group and form shared attention with them naturally and instantly. However, patients with autism may feel it difficult to interact with people around them since they lack the ability to build shared attention with others [Bar95]. Fig. 4.8 shows some examples of shared attention in social scenes and how shared attention shifts temporally as well as who are currently involved in the shared attention.

Research in developmental psychology clearly states that the development of skills to understand, manipulate and coordinate attentional behavior plays a pivotal role for imitation, social cognition and the development of language [Hob02, MD95, TCC05]. And among the complicated cognitive functions of human minds, the ability to form, recognize and understand shared attention is pretty crucial in human social interactions [MC94, MD95, Nag04]. All human communication, even including linguistic communication, is only possible when the people involved in such communications have built a common conceptual ground consisting of shared attention, shared experience, common cultural knowledge, etc [Tom08]. Overall, shared attention is a crucial first step towards social interaction, as well as the primary basis of social intelligence and a precursor of theory of mind [MC94, MD95, Nag04], language learning [MD95, MMR98, MG98], the ability of imitation [KIU03] and so on. The study of shared attention is important because it helps a computer vision



Figure 4.8: Examples of shared attention in daily life.

system to better understand and interpret human activities in images or videos. Robotics equipped with the ability to detect and understand human shared attention can also be more intelligent when interacting with humans.

Despite the importance of this topic, works on shared attention are quite limited in the computer vision community. Some previous works address the problem by using special input data, such as first-person videos taken by multiple head-mounted cameras [APS14, PJS12, PS15, SHS16]. Some limited shared attention to the field of Human Robot Interaction [SPR09, KJD08, Nag05, NHM03, SGB05, SHY07]. Few works studied shared attention in human social interaction based on third-person social scene videos.

In order to be clarified in our work with the concept of shared attention, we formulate our problem as follows: shared attention is the gaze focus shared by two or more individuals on one

object or human; given a video clip, the task is to detect which frames contain shared attention and where is the shared attention in those frames. To tackle this problem, we collect a new dataset VideoCoAtt and build a deep spatial-temporal neural network with four modules: gaze estimation module, region proposal module, spatial detection module and temporal optimization module. The intuitions for building such a deep neural network architecture are as follows: 1) Firstly, gaze direction, which can be utilized to learn external environment state and internal mental state, is a key feature for shared attention detection. The strongest and most direct indication of human gaze direction is the closeup image patch of human head. We need to detect human heads in videos and predict gaze directions for each detected head. 2) Secondly, gaze direction is of course important, but still not the whole story. Shared attention is more than gaze intersection. According to our definition, there must be an object or human body part as the carrier of shared attention, which means the shared attention detection task is object-driven. Thus, bounding box proposals of object or human body parts, such as laptop, human face, etc, is another key feature for our task. We didn't use saliency models (like [PSG16, WS18]) because shared attention is more influenced by social group interaction instead of visual importance, and people engaged in shared attention are not free-viewing and may not look at the most salient object in the environment. We use a generic object proposal generation method to generate all potential bounding boxes independent of their categories. 3) Shared attention may last for a while before termination. Temporal information is a good constraint to make the detection results more accurate and robust. The input to our model is just a video clip without any other additional annotation, and the output is a shared attention heatmap for each video frame and the final shared attention prediction results can also be inferred based on the shared attention heatmap.

This work makes **three major contributions**: (i) It addresses a new problem - inferring shared attention in third-person social scene videos. To the best of our knowledge, this is the first work to deal with such problem in computer vision community. (ii) It proposes a spatial-temporal network to address the problem of inferring shared attention in videos. The proposed model explicitly leverages human gaze direction, target region candidates, and temporal inter-frame constraints for identifying shared attention. (iii) It presents a large-scale dataset covering diverse social scenes with full annotations, VideoCoAtt, and benchmark results on the dataset for shared



Figure 4.9: Example frames from VideoCoAtt dataset.

attention study.

4.2.2 Related Work

The problem of inferring shared attention from third-person videos is closely related to the following works:

Gaze Prediction: Recasens et al. proposed a deep learning based model for gaze prediction in images [RKV15] and contributed a dataset called GazeFollow. Given head location, their method extracts head pose and gaze orientation, follows the gaze of the person and identifies the object being looked at in the image. Then they further extended their work to gaze prediction in videos [RVK17] and contributed another new dataset VideoGaze. Given a video clip and the annotations of head and eye location, their model combines gaze pathway, saliency pathway and transformation pathway to predict where a person is looking even when the object being looked at is in a different frame. These works only focus on predicting single-person gaze, while do not consider the task of inferring attention shared by multiple persons in social activities.

Shared attention in Social Interaction: There are some inspiring studies of shared attention in human social interaction. Park et al. presented a method to construct a 3D social

Dataset	Year	Format	Size	Annotation	Goal	Shared Attention	Data Source
HMDB [KJG11]	2011	Video	7,000 clips, 51 action categories	Human action	Action recognition	-	Digitized movies, YouTube
TVHI [PMR12]	2012	Video	300 video clips, 30 to 600 frames per clip	Body btx, head orientation, interaction label	Human interaction learning	-	different TV shows
MPII-MD [RRT15]	2015	Video	94 videos, 68,337 clips	Video description	Automatic video description	-	British Amazon, Hollywood2
GazeFollow [RKV15]	2015	Image	122,143 images, 130,339 people	Eye loc. and gaze loc.	Gaze following in images	-	Actions 40, MS COCO, SUN, PASCAL
VideoGaze [RVK17]	2017	Video	140 movies, 6 frames per movie	Eye loc., head btx, gaze loc.	Gaze following in videos	-	MovieQA
Sitcom Affordance [WGG17]	2017	Image	11,449 indoor scenes, 28,882 human poses	Human pose	Affordance prediction	-	7 sitcoms
VideoCoAtt (Ours)	2018	Video	380 videos, 492,100 frames	Shared attention btx, involved head btx.	Shared attention detection in videos	✓	20 different TV shows

Table 4.5: Comparison of VideoCoAtt related datasets.

saliency field and locate multiple gaze concurrences that occur in a social scene from videos taken by head-mounted cameras [PJS12]. After that, they proposed a method to predict social saliency from images or videos captured by multiple first-person view cameras [PS15]. These works directly study social saliency, which by their definition represents the likelihood of shared attention in a social group. Besides, they also use shared attention as a constraint to predict social behavior in first-person videos, such as individuals’ future movements and future gaze directions in a social group. The predicted behaviors reflect an individual physical space that affords to take the next actions while conforming to social behaviors by engaging to shared attention [SHS16]. Generally, these work well explored and illustrated shared attention detection and application in social activities. However, they only focus on first-person videos without generalizing to ordinary third-person videos.

Shared attention in HRI: The field of Human-Robot Interaction (HRI) strives to enable easy, intuitive interactions between people and robots, which requires natural communication [AS17]. Many of the difficulties encountered in human-robot interaction and the communication between autonomous robots could be traced back to unsolved issues related to shared attention [KH06]. There are many works that try to realize gaze-following and shared attention between robot and human in HRI with or without external evaluation [SPR09, KJD08, Nag05, NHM03, SGB05, SHY07]. The key points of these work are inferring human gaze direction and then forming shared attention between robot and human by making the robot head turn to that direction. Our work is beneficial to improve the implementation of shared attention in HRI because robots can further detect, understand and learn to join in the on-going shared attention in the environment.

4.2.3 VideoCoAtt Dataset

In this section we describe our proposed VideoCoAtt dataset, which is specifically designed for studying shared attention in social scenes. Some example frames with annotations are presented in Fig. 4.9.

Dataset Collection. The following principles drive the collection of our dataset:

- *Natural social interaction.* Shared attention usually occur in daily life naturally. If we deliberately shoot videos for the purpose of shared attention study, then the social interactions performed by the volunteers may seem unnatural and not convincing. Instead, TV show is a good choice because social interactions in TV shows appear to be relatively more natural. As summarized in Table 4.5, there are some TV show datasets available in the computer vision community, e.g., HMDB [KJG11], TVHI [PMR12], etc. However, they are designed for different purposes, like action recognition, human interaction understanding, etc, and none of them offer annotations of shared attention. Differently, the proposed VideoCoAtt dataset is carefully collected for studying shared attention in human social activities. The videos are sourced from 20 different TV shows on Youtube.
- *Large scale and high quality.* Both scale and quality are essential to build a long-lifespan benchmark. We carefully collect 380 RGB video sequences from 20 different TV shows or movies. Each video sequence lasts for various time, from around 20s to more than 1 minute with a frame rate of 25 fps. In total, there are 492,100 frames at the spatial resolution of 320×480 .
- *Diversity and generality.* The videos in the VideoCoAtt dataset cover different countries and cultures, such as American, Chinese, Indian, European, etc. The appearances of actors/actresses, the costume and props vary a lot. There are also diverse scenario settings in VideoCoAtt, including living room, kitchen, restaurant, Cafe, office, outdoor, etc. See Table 4.6 for detailed statistics and Fig. 4.9 for example frames. Moreover, the number of shared attentions per frame and the number of involved people per shared attention can vary in different frames and videos, as can be seen from the sample frames in Fig. 4.9 and the statistics in Table 4.7. This generality in VideoCoAtt dataset is beneficial for the trained model to deal with multiple cases as in real life. Fig. 4.10 shows the shared attention location distribution averaged over the whole dataset. It appears that shared attention in our dataset tends to lie near the top part of the image frame, as is consistent with

Culture Distribution		Scenario Setting Distribution			
American	44.1 %	Living Room	29.4 %	Dining Room	4.7 %
Chinese	40.7 %	Kitchen	14.3 %	Office	4.7 %
Indian	9.1 %	Restaurant	7.0 %	Bathroom	2.3 %
European	4.1 %	Bedroom	6.8 %	Outdoor	16.4 %
(Others)	2.0 %	Cafe	5.8 %	(Others)	8.6 %

Table 4.6: Distributions of culture and scenario settings in VideoCoAtt dataset.

previously analyzed eye tracking datasets [ZTM08, Tat07, JED09].

Dataset Annotation. We manually annotate all the video frames using the online tool Vatican [VPR13a]. For each frame, we mark whether there is shared attention in the scene. If there is on-going shared attention in the scene, we mark all the shared attentions with bounding boxes. Only those shared attentions within the view of the scene will be annotated; those out of view or occluded will not be counted as shared attention. Furthermore, for each shared attention, we annotated all the heads that are currently engaged in the certain shared attention using bounding boxes and attributes related to the shared attention numbering.

Dataset Splitting. We split our VideoCoAtt dataset into three parts for training, validation and testing respectively. There are 181 videos (250,030 frames) in the training set, 90 videos (128,260 frames) in the validation set and 109 videos (113,810 frames) in the testing set. To avoid overfitting caused by similarities in human appearances and scenario settings, we split our videos by different sources. Videos for training, validation and testing come from different TV shows, which we believe is necessary and will require a strong generalization ability of our shared attention model.

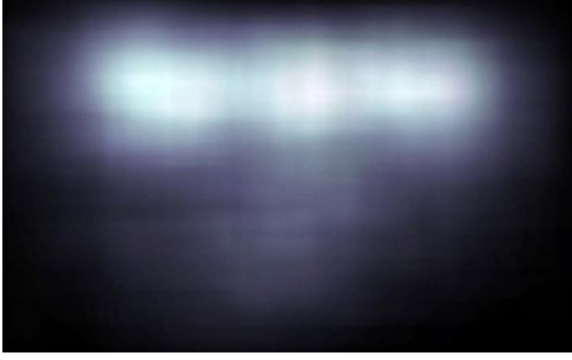


Figure 4.10: Illustration of shared attention location.

VideoCoAtt	#shared attentions per frame				
	0	1	≥ 2		
#frames	349,468	139,348	3,284		
VideoCoAtt	#people involved per S.A.				
	2	3	4	5	≥ 6
#S.A.	86,988	34,105	16,396	4,955	3,661

Table 4.7: Statistics of the shared attentions and people involved in VideoCoAtt dataset.

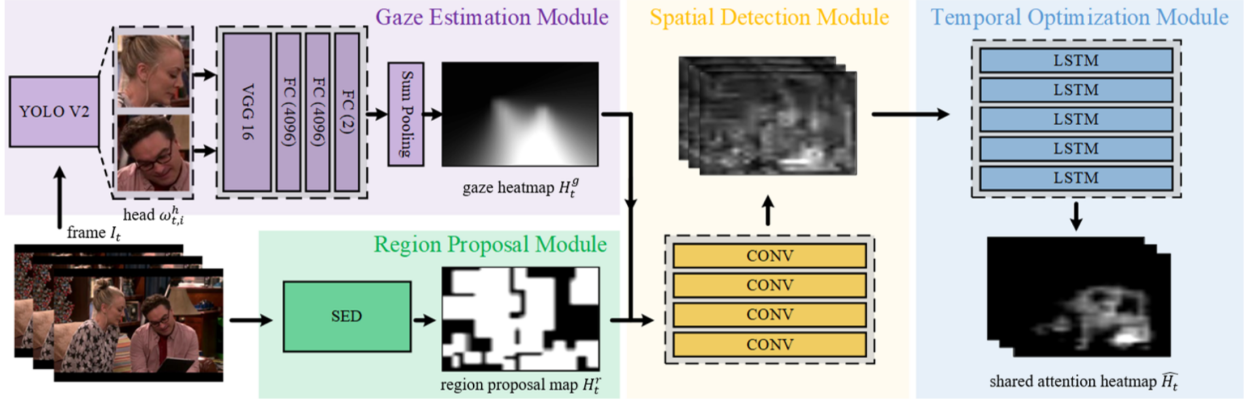


Figure 4.11: Illustration of VideoCoAtt model architecture.

4.2.4 Model

Shared attention usually locates at the objects or human body parts gazed by two or more people simultaneously. Obviously, human gaze and target objects in the context environment are essential for inferring shared attention in social scene videos. Thus our shared attention detection model comprises of four modules: 1) the gaze estimation module (§ 4.2.4.1) that extracts individual gaze directions to generate a gaze heatmap for the whole scene; 2) the region proposal module (§ 4.2.4.1) that extracts region proposals from the context environment; 3) the spatial detection module (§ 4.2.4.2) that combines the gaze heatmap and the region proposal map to detect shared attention in spatial space; and 4) the temporal optimization module (§ 4.2.4.2) that utilizes inter-



Figure 4.12: Illustration of gaze heatmap generation procedure.

frame correlation to optimize the predicted shared attention heatmap in temporal space. An illustration of our whole model architecture is presented in Fig. 4.11.

4.2.4.1 Gaze and Region Proposal Modules

Gaze Estimation Module. Suppose for an input frame I_t in a video sequence $\{I_t\}_{t=1,\dots,T}$, our head detector outputs a set of head locations $q_{t,i} = (x_{t,i}^{min}, y_{t,i}^{min}, x_{t,i}^{max}, y_{t,i}^{max})$, $i = 1, 2, \dots, n$, where n could be zero when no head is detected in frame I_t (see the red rectangles in Fig. 4.12 (a) and (c)). The corresponding closeup image patch for head location $q_{t,i}$ is cropped out from I_t and denoted as $w_{t,i}^h$, $i = 1, 2, \dots, n$. We then use a batch of neural network layers $\Psi(\cdot)$ to regress a gaze direction $d_{t,i} \in [-1, 1]^2$ (yellow arrows in Fig. 4.12 (a) and (c)) for the input image patch $w_{t,i}^h$:

$$d_{t,i} \triangleq (d_{t,i}^x, d_{t,i}^y) = \Psi(w_{t,i}^h). \quad (4.1)$$

We use a Gaussian distribution to model the variation of a gaze ray with respect to the predicted primary gaze direction $d_{t,i}$, and the probability distribution is

$$P(\theta_{t,i}|d_{t,i}) \propto \frac{1}{\sigma} \exp\left\{-\frac{\theta_{t,i}^2}{2\sigma^2}\right\}, \quad (4.2)$$

where $\theta_{t,i}$ is the angle between a gaze ray and the predicted primary gaze direction $d_{t,i}$. With detected head position $q_{t,i}$ and corresponding predicted gaze direction $d_{t,i}$, we compute $\theta_{t,i}$ for each grid in the image and then use Eq. 4.2 to get the probability for this grid to be gazed at by head $q_{t,i}$. After a gaze heatmap $H_{t,i}^g$ (see Fig. 4.12 (b) and (d)) for each head position $q_{t,i}$ is prepared, we generate the final gaze heatmap H_t^g (Fig. 4.12 (e)) of size $M \times N$ via Sum-Pooling $\{H_{t,i}^g\}_i$:

$$H_t^g = \sum_{i=1}^n H_{t,i}^g = \sum_{i=1}^n \phi(\Psi(w_{t,i}^h), q_{t,i}), \quad (4.3)$$

where $\phi(\cdot)$ indicates the gaze heatmap generator based on Eq. 4.2. More illustrations about the gaze heatmap generation procedure are shown in Fig. 4.12.

Region Proposal Module. To exploit context information, we use a region proposal module $Z(\cdot)$ to generate a binary region proposal map H_t^r of size $M \times N$ for input image I_t :

$$H_t^r = Z(I_t). \quad (4.4)$$

This module is implemented by Structured Edge Detector (SED) [ZD14] to get region bounding boxes $\{b_{t,i}, i = 1, 2, \dots, m\}$ for each frame I_t and then setting all the pixel values within the bbx proposals to 1 and all other pixel values outside to 0.

4.2.4.2 Spatio-temporal Shared Attention Network

The output feature maps of the gaze estimation module and the region proposal module are then fed to the subsequent spatial detection module and temporal optimization module for shared attention detection.

Spatial Detection Module. Shared attention detection is firstly conducted in a frame-by-frame style. We apply a spatial detection module $F(\cdot)$ that consists of several convolutional layers to combine the gaze heatmap H_t^g and region proposal map H_t^r for intra-frame shared attention detection:

$$\tilde{H}_t = F(H_t^g, H_t^r), \quad (4.5)$$

where \tilde{H}_t indicates the intermediate shared attention heatmap output from the spatial detection module.

Temporal Optimization Module. To further exploit the temporal inter-frame constraints in videos, we add a temporal optimization module $LSTM(\cdot)$ that consists of several convolutional Long Short-Term Memory (convLSTM) network [SCW15] layers to optimize the output shared attention heatmap \tilde{H}_t :

$$\{\hat{H}_t\}_t = LSTM(\{\tilde{H}_t\}_t), \quad (4.6)$$

where \hat{H}_t denotes the eventual shared attention heatmap.

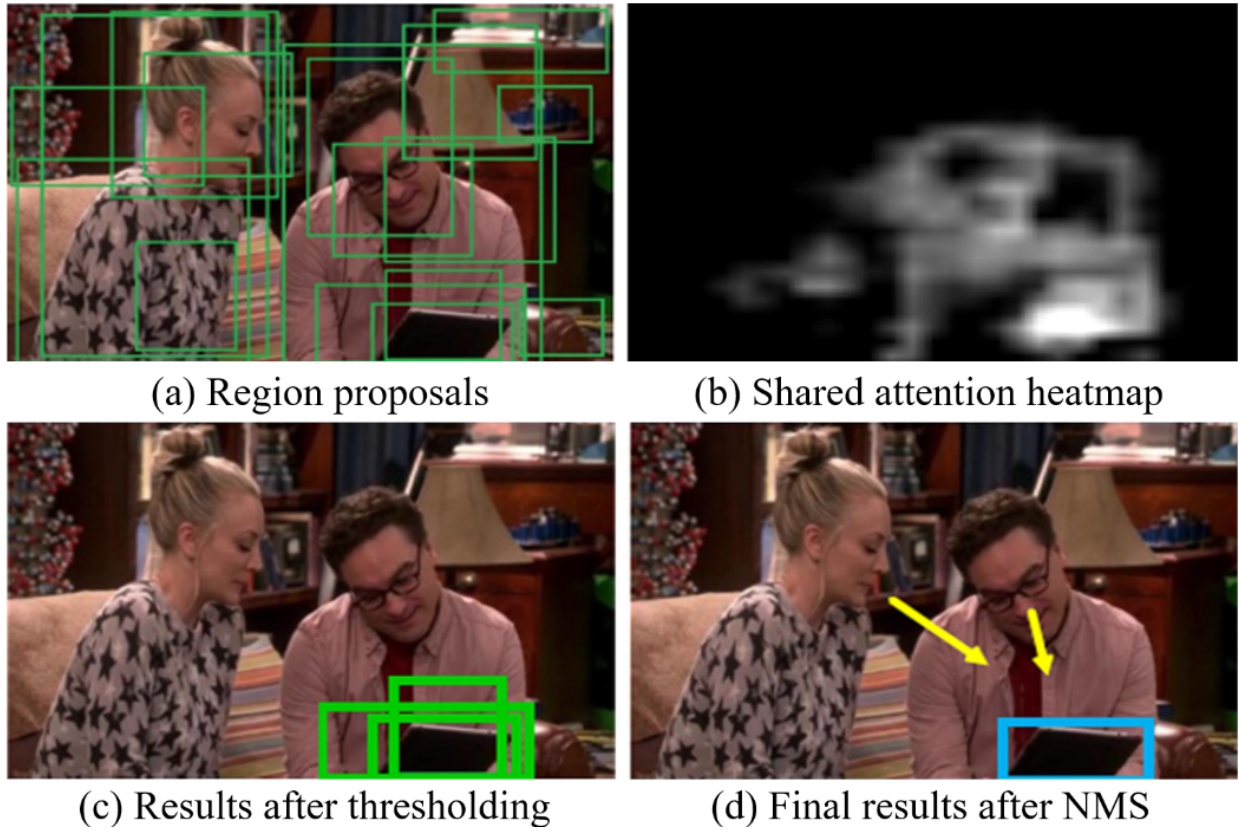


Figure 4.13: Illustration of shared attention inference process.

4.2.4.3 Learning and Inference

For the loss function, we apply the Mean Squared Error (MSE) between the predicted shared attention heatmap \hat{H}_t and the ground truth shared attention binary map H_t :

$$L(\hat{H}_t, H_t) = \frac{1}{M \cdot N} \|\hat{H}_t - H_t\|^2, \quad (4.7)$$

where both \hat{H}_t and H_t are of size $M \times N$.

The inference is possible given the predicted shared attention heatmap \hat{H}_t , based on which we can compute the cumulative score for each region proposal bounding box $b_{t,i}$. We only keep those proposal bounding boxes with a score higher than a threshold. Then we conduct a Non-Maximum Suppression (NMS) [FGM10] and treat the remaining bounding boxes as our final shared attention prediction for frame I_t . See Fig. 4.13 for more detailed illustration.

Since there may be no shared attention or more than one shared attention in a scene, our model is designed to support multimodal predictions instead of regressing a single shared attention location.

4.2.4.4 Implementation Details

We implement our model using Keras with Tensorflow as backend. For the gaze estimation module, we first fine-tuned YOLO V2 darknet [RF17] on our own training set. The re-trained YOLO V2 is applied as a head detector to generate human head image patches $\{\omega_{t,i}^h\}$ for the following gaze direction estimation. We apply the VGG16 network to regress gaze direction, and replace the last fully connected (fc) layer (1000) with a new fc layer of size 2. Then the tanh activation is used for generating a unit gaze direction vector and the gaze direction regression network is fine-tuned on our training set with mean-squared-error loss. To generate the gaze heatmap, we assume that the gaze cone projected from each head is subject to a gaussian distribution with standard deviation $\sigma = 0.5$. For the region proposal module, we use the Structured Edge Detection Toolbox [ZD14] to generate the bounding box proposals for each frame.

The outputs of the gaze estimation module and the region proposal module are of size 28×28 . We concatenate the gaze heatmap H_t^g and the region proposal map H_t^r as the input to the spatial detection module, which first contains three convolutional layers with kernel size 3×3 and output channel size 16, 16, 8 respectively, followed by the last convolutional layer with kernel size 1×1 , output channel size 1 and sigmoid activation. The output of spatial detection module is a 28×28 shared attention heatmap \tilde{H}_t for each frame. The subsequent temporal optimization module consists of five convLSTM layers with filter sizes 40, 40, 40, 40 and 1 respectively. The kernel size is 3×3 for the first four convLSTM layers and 1×1 for the last convLSTM layer. The final convLSTM layer uses sigmoid as activation function.

4.2.5 Experiments

4.2.5.1 Experimental Setup

We train and evaluate our model on disjoint training, validation and testing sets from VideoCoAtt in our experiments, as described in §4.2.3. The ground truth annotations of shared attention bounding boxes and relevant human faces’ bounding boxes are only used in training. For testing, the input to our model only includes the raw videos without any additional annotation.

Evaluation Metrics. We use several metrics to compare our model predicted shared attentions with the ground truth shared attention annotations across the testing videos. For the shared attention interval detection task, the percentage of frames with right shared attention existence prediction over all the video frames is applied as a metric *Prediction Accuracy*. For the shared attention location prediction task, we use the region proposal bounding boxes and shared attention heatmap to generate a *ROC Curve*, reflecting the precision and recall when predicting shared attention bounding boxes under different score thresholds. *AUC* refers to the area under the ROC curve (higher is better). Then given a certain score threshold, the *L² Distance* (measured in pixel) is the Euclidean distance between the predicted shared attention bbx and the annotated ground truth.

Baseline Methods. We compare our approach against several baselines ranging from simple ones (Random, Fixed Bias) to more complex ones (Gaze Follow, Gaze+Saliency, Gaze+Saliency+LSTM) as described below.

Random: A weak baseline that draws a Gaussian heatmap with random mean and variance. *Fixed Bias:* As visible in Fig. 4.10, there exists shared attention location bias in the TV shows. We use a fixed-biased heatmap subject to a 2D Gaussian Distribution with mean and variance learned from our dataset as a baseline to model such bias. *Gaze Follow:* We apply the gaze following model in [RKV15] to detect all the people’s gaze fixations and gaze concurrences in a frame as a baseline. *Gaze+Saliency* and *Gaze+Saliency +LSTM:* We replace our region proposal module with a top-performance saliency model [PSG16], and consider two baselines with and without the temporal optimization module respectively.

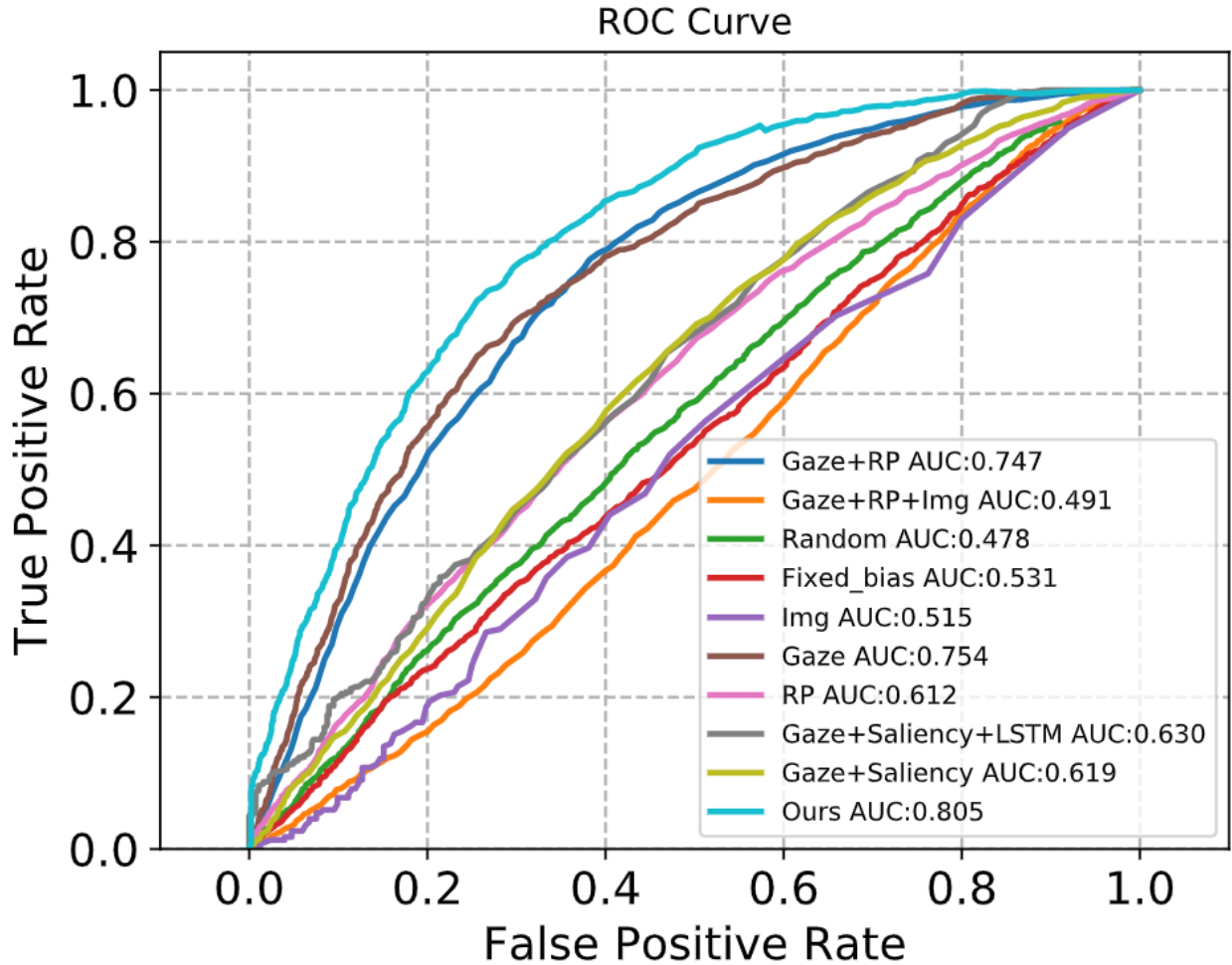


Figure 4.14: Quantitative evaluation results with ROC Curve.

Ablation Study. To better understand the importance of each module in our proposed model architecture, we also studied the model performance after removing some modules. *Raw Img.:* We first only use raw image as input to train an end-to-end model, which means we only keep the spatial detection module. *Only Gaze:* Then we try to augment the model by adding gaze estimation module to spatial detection module. *Only RP:* We also tested the architecture with only region proposal module and spatial detection module. *Gaze+RP:* We add both gaze estimation and region proposal modules before spatial detection module. *Gaze+RP+Img.:* This is a variation of our model that uses gaze, region proposal and raw image feature as input to spatial detection module without using temporal optimization module.

Model	Prediction Acc.	L^2 Dist.
Raw Img.	52.3 %	188
Only Gaze	64.0 %	108
Only RP	58.0 %	110
Gaze+RP	68.5 %	74
Gaze+RP+Img.	54.0 %	72
Fixed Bias	52.4 %	122
Random	50.8 %	286
Gaze Follow [RKV15]	58.7 %	102
Gaze+Saliency[PSG16]	59.4 %	83
Gaze+Saliency[PSG16]+LSTM	66.2 %	71
Ours (Gaze+RP+LSTM)	71.4 %	62

Table 4.8: Quantitative evaluation results with Prediction Accuracy and L_2 Distance.

4.2.5.2 Results and Analysis

Quantitative results. Table 4.8 shows the comparison of our model with baseline methods and several ablation models by two evaluation metrics *Prediction Accuracy* and L^2 *Distance*. Our model achieves the best performance in both the shared attention interval detection task (Prediction Acc.: 71.4%) and the shared attention location prediction task (L^2 Dist.: 62).

Among all the baseline models, the second best model is *Gaze+Saliency+LSTM* with a Prediction Acc. of 66.2% and a L^2 Dist. of 71. The replacement of region proposal module with a saliency model impairs our model performance because the shared attention of people in a social

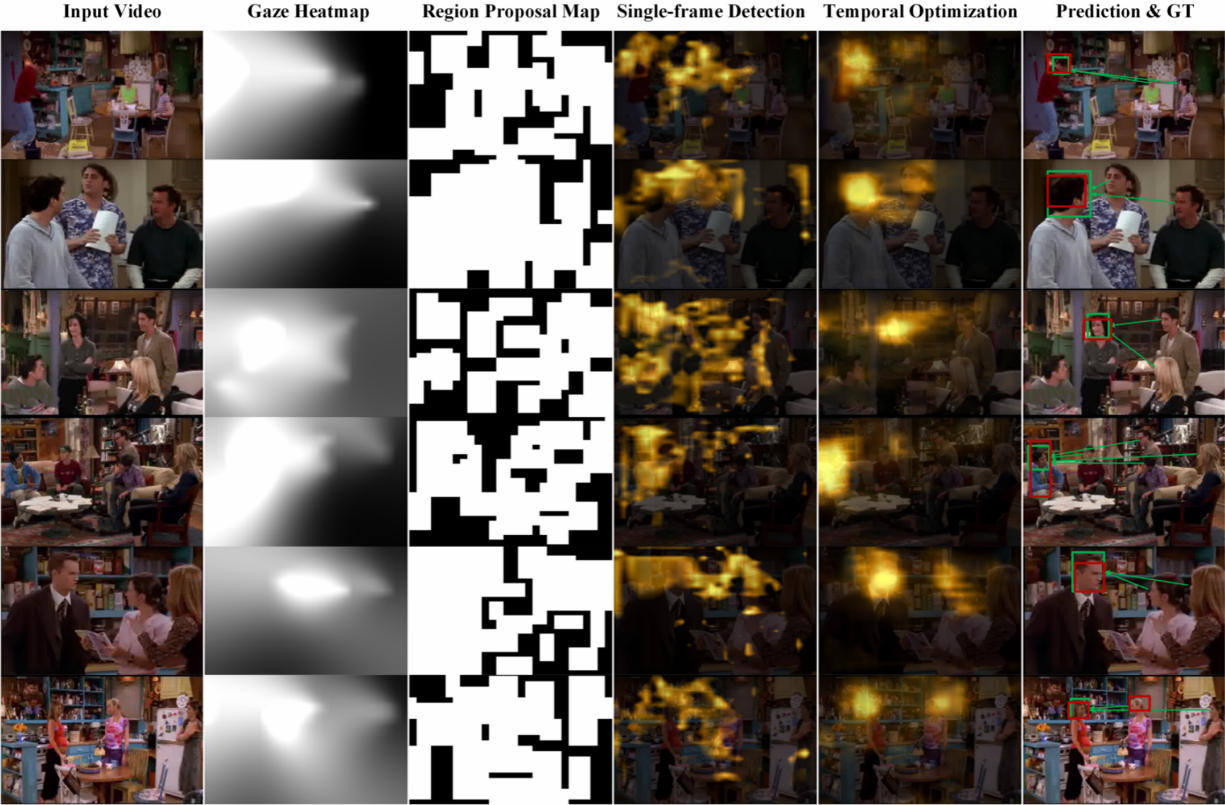


Figure 4.15: Shared attention detection results on example frames.

interaction may not be the most visually salient object in the scene, but more influenced by the on-going interaction. The performance of the Gaze Follow baseline in detecting shared attention is mediocre, which is mainly because that shared attention of a social group is goal-driven and object-related, not just the concurrence of human gazes.

Among all the ablation models, *Gaze+RP* shows a overall best performance (Prediction Acc.: 68.5% and L^2 Dist.: 74), but is still inferior to our full model with all the four modules. And overall *Only Gaze* performs better than *Only RP*, indicating the gaze estimation module plays a more important role than the region proposal module in shared attention detection, which is consistent with our intuitions. The simplest model without any module design *Raw Img.* performs worst. The ablation study shows that each of the four modules proposed by our model (§ 4.2.4) is important and necessary for shared attention detection in videos.

Fig. 4.14 shows the ROC Curve and AUC comparison results among our full model, baseline

models and ablation models. Our model has the best precision and recall performance and the largest AUC value than all the other models. *Gaze+RP* and *Gaze* also perform significantly better than the remaining models. The result further confirms the significance and effectiveness of our model architecture design. The gaze direction feature and the region proposal feature as well as the temporal constraints indispensably help our model to gain great performance improvements in the task of inferring shared attention in social scene videos.

Qualitative results. Fig. 4.15 exhibits an internal visualization of shared attention detection results by our full model on some example frames. The *Gaze Heatmap* roughly features the attention of each individual in the social scene and is not enough to accurately feature shared attention. The *Region Proposal Map* gives some potential shared attention proposals and provides the important spatial constraints. *Single-frame Detection* combines the *Gaze Heatmap* and the *Region Proposal Map* to generate a preliminary shared attention heatmap, which still has too much noises. After the *Temporal Optimization* by convLSTM, the shared attention heatmap is much clearer and can provide more accurate shared attention distribution information. The final column in Fig. 4.15 compares our eventual shared attention prediction results (depicted in red rectangles) with the ground truth shared attention annotations (depicted in green rectangles). As shown, there are good predictions that can exactly locate the shared attention in the social scenes, like the prediction in the first example. However, there are also some false alarms existing. For example, The scene in the last row actually has only one shared attention, but our model gives two predictions located near the two human faces. This is an interesting failure example since whether the third person on the right side is looking at the person on the left side or the person in the middle is somehow ambiguous for our model to distinguish. That’s why the shared attention heatmap gets two peaks for this example. But similar situation in the fifth scene is successfully solved by our model.

4.2.6 Conclusion

This work addresses a new problem of inferring shared attention in third-person social scene videos. Although shared attention is common in daily life and important for social interactions, relevant studies are quite limited in the computer vision community. We propose a dataset VideoCoAtt

and a model to detect shared attention in videos. Our model combines individual gaze features and context region proposal features from the raw video inputs. Based on the two bottom features, our model learns to spatially detect and temporally optimize shared attention in videos. Although we get some reasonable results in the experiments, we are still far from completely solving this problem. We hope our dataset and model will serve as important resources to facilitate future studies related to this topic.

CHAPTER 5

Conclusion

This dissertation introduces our efforts towards building human-like machine learning models in holistic scene understanding and goal-directed event parsing. More specifically, we focus on three aspects: 3D reconstruction of humans and scenes, long-term goal-directed activity, and human communications. These general tasks not only rely on the data-driven pattern recognition but also root from the visual reasoning system, known as the core knowledge of human intelligence. We identify and pinpoint several representative tasks and provide the following insights.

- The 3D scene reconstruction and 3D human pose estimation are two deeply coupled tasks. We propose to exploit physical commonsense and human-object interaction in the MCMC optimization, which traverses the non-differentiable solution space to reach the physically stable and action-aware scene configurations. These two critical information sources can act as general priors that significantly boost the generalization ability of the inference framework.
- Human-object contact is a key component in modeling the human-object interaction, and it's often neglected or simplified in prior work. The contact information from the 2D visual cues can act as effective proposals to reasoning about the actual contact in 3D space, which provides the physical stability required in both reconstruction and task planning.
- Daily human activities are intrinsically goal-oriented and multi-tasked; as agents' decision-making processes are deeply affected by their unique social values, understanding activity naturally demands a learning system to understand how a given task should be decomposed into atomic actions, how multi-tasks should be executed and coordinated in parallel among multi-agents, and take the perspective from human agents to understand why the observed human activities are optimal solutions. Current machine learning models still fall short on

these critical aspects.

- We benchmark machines' capability to understand human communications under both shared attention and reference settings. Shared attention goes beyond the gaze and relates human, object through a triadic dynamic interactions. Embodied reference, on the other hand, is inherently multi-modal, which requires reasoning jointly with both gestural and verbal information. Understanding human communication is currently under-explored and our work initiates more research in this area.

In conclusion, a joint framework that combines low-level vision tasks, mid-level scene understanding and high-level event parsing is essential for accomplishing various tasks in real-life scenarios, rather than specific models tailed for small tasks. The joint inference and learning can be incorporated in a closed loop of passive perception and active interaction to mimic human's excellence in learning and generalizing new concepts and knowledge. It requires interdisciplinary expertise in computer vision, natural language understanding, computer graphics, machine learning, robotics, and cognitive science to build up a machine system to reach human-like intelligence. We hope future efforts will be devoted to stimulating higher cognitive capabilities towards the more holistic scene and event understanding.

REFERENCES

- [AAL15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. “Vqa: Visual question answering.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [AAX20] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. “Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes.” In *European Conference on Computer Vision*, pp. 422–440. Springer, 2020.
- [ABA16] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. “Unsupervised learning from narrated instruction videos.” In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [ACV09] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. “A survey of robot learning from demonstration.” *Robotics and Autonomous Systems (RAS)*, **57**(5):469–483, 2009.
- [ALP08] Michael A Arbib, Katja Liebal, and Simone Pika. “Primate vocalization, gesture, and the evolution of human language.” *Current anthropology*, **49**(6):1053–1076, 2008.
- [APS14] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. “Automatic Editing of Footage from Multiple Social Cameras.” *ACM Trans. Graph.*, **33**(4), 2014.
- [AS17] H. Admoni and B. Scassellati. “Social Eye Gaze in Human-robot Interaction: A Review.” *J. Hum.-Robot Interact.*, **6**(1), 2017.
- [Bai04] Renée Baillargeon. “Infants’ physical world.” *Current directions in psychological science*, **13**(3):89–94, 2004.
- [Bar95] Simon Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1995.
- [BB01] Dare A Baldwin and Jodie A Baird. “Discerning intentions in dynamic human action.” *Trends in Cognitive Sciences*, **5**(4):171–178, 2001.
- [BBC19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. “Dota 2 with Large Scale Deep Reinforcement Learning.” *arXiv preprint arXiv:1912.06680*, 2019.
- [BBS01] Dare A Baldwin, Jodie A Baird, Megan M Saylor, and M Angela Clark. “Infants parse dynamic action.” *Child development*, **72**(3):708–717, 2001.

- [BES97] C Daniel Batson, Shannon Early, and Giovanni Salvarani. “Perspective taking: Imagining how another feels versus imagining how you would feel.” *Personality and social psychology bulletin*, **23**(7):751–758, 1997.
- [BKM20] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. “Emergent tool use from multi-agent autocurricula.” In *International Conference on Learning Representations (ICLR)*, 2020.
- [BLB14] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. “Weakly supervised action labeling in videos under ordering constraints.” In *European Conference on Computer Vision (ECCV)*, 2014.
- [BRG16] Aayush Bansal, Bryan Russell, and Abhinav Gupta. “Marr revisited: 2d-3d alignment via surface normal prediction.” In *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.
- [BTT20] Samarth Brahmhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. “ContactPose: A dataset of grasps with object contact and hand pose.” In *European Conference on Computer Vision (ECCV)*, volume 12358, pp. 361–378, 2020.
- [CBM20] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. “Refer360: A Referring Expression Recognition Dataset in 360: A Referring Expression Recognition Dataset in 360 Images Images.” In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [CCP13] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. “Understanding indoor scenes using 3d geometric phrases.” In *Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [CEG15] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. “ActivityNet: A large-scale video benchmark for human activity understanding.” In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [CEK20] Henry M. Clever, Zackory Erickson, Ariel Kapusta, Greg Turk, Karen Liu, and Charles C. Kemp. “Bodies at rest: 3D human pose and shape estimation from a pressure image using synthetic data.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 6215–6224, 2020.
- [CG07] Gergely Csibra and György Gergely. “‘Obsessed with goals’: Functions and mechanisms of teleological interpretation of actions in humans.” *Acta psychologica*, 2007.

- [CHY19] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. “Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense.” In *International Conference on Computer Vision (ICCV)*, pp. 8648–8657, 2019.
- [CLL18] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. “Learning to detect human-object interactions.” In *Winter Conference on Applications of Computer Vision (WACV)*, 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.
- [CMG08] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. “Gesturing makes learning last.” *Cognition*, **106**(2):1047–1058, 2008.
- [CMS19] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.
- [CRK21] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. “Reconstructing Hand-Object Interactions in the Wild.” In *International Conference on Computer Vision (ICCV)*, pp. 12417–12426, 2021.
- [CSK10] Cristina Colonnese, Geert Jan JM Stams, Irene Koster, and Marc J Noom. “The relation between pointing and language development: A meta-analysis.” *Developmental Review*, **30**(4):352–366, 2010.
- [CSS09] Wongun Choi, Khuram Shahid, and Silvio Savarese. “What are they doing?: Collective activity classification using spatio-temporal relationship among people.” In *International Conference on Computer Vision Workshops (ICCV Workshops)*, 2009.
- [CSW17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields.” In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [CWM20] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. “Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension.” In *Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [CZ17] Joao Carreira and Andrew Zisserman. “Quo vadis, action recognition? a new model and the kinetics dataset.” In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [DCL18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv preprint arXiv:1810.04805*, 2018.
- [DDM18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. “Scaling egocentric vision: The epic-kitchens dataset.” In *European Conference on Computer Vision (ECCV)*, 2018.
- [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 *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.
- [DSC17] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. “Guesswhat?! visual object discovery through multi-modal dialogue.” In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [DXW21] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. “3D AffordanceNet: A Benchmark for Visual Object Affordance Understanding.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1787, 2021.
- [Eme00] N.J. Emery. “The eyes have it: the neuroethology, function and evolution of social gaze.” *Neuroscience & Biobehavioral Reviews*, **24**(6):581 – 604, 2000.
- [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 *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [FDG12] David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, and Josef Sivic. “People watching: Human actions as a cue for single view geometry.” In *European Conference on Computer Vision (ECCV)*, volume 7576, pp. 732–745, 2012.
- [FFM19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. “Slowfast networks for video recognition.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [FGM10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. “Object Detection with Discriminatively Trained Part-Based Models.” *IEEE TPAMI*, **32**(9):1627–1645, 2010.

- [FKE18] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. “From lifestyle vlogs to everyday interactions.” In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [FQZ21] Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. “Learning Triadic Belief Dynamics in Nonverbal Communication from Videos.” In *Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [FWY18] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. “Demo2vec: Reasoning object affordances from online videos.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2139–2147, 2018.
- [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*, 2018.
- [GBK02a] György Gergely, Harold Bekkering, and Ildikó Király. “Developmental psychology: Rational imitation in preverbal infants.” *Nature*, **415**(6873):755, 2002.
- [GBK02b] György Gergely, Harold Bekkering, and Ildikó Király. “Rational imitation in preverbal infants.” *Nature*, **415**(6873):755–755, 2002.
- [GDG17] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. “Accurate, large minibatch sgd: Training imagenet in 1 hour.” *arXiv preprint arXiv:1706.02677*, 2017.
- [GF16] Noah D Goodman and Michael C Frank. “Pragmatic language interpretation as probabilistic inference.” *Trends in cognitive sciences*, **20**(11):818–829, 2016.
- [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.
- [GKM17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. “The ”Something Something” Video Database for Learning and Evaluating Visual Common Sense.” In *International Conference on Computer Vision (ICCV)*, 2017.

- [GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite.” In *Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [GM15] Saurabh Gupta and Jitendra Malik. “Visual Semantic Role Labeling.” *arXiv:1505.04474*, 2015.
- [GMG08] Adam D Galinsky, William W Maddux, Debra Gilin, and Judith B White. “Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations.” *Psychological Science*, **19**(4):378–384, 2008.
- [GR20] Rohit Girdhar and Deva Ramanan. “CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning.” *International Conference on Learning Representations (ICLR)*, 2020.
- [Gri75] Herbert P Grice. “Logic and conversation.” In *Speech acts*, pp. 41–58. Brill, 1975.
- [GSJ21] Shivam Grover, Kshitij Sidana, and Vanita Jain. “Pipeline for 3D reconstruction of the human body from AR/VR headset mounted egocentric cameras.” *arXiv:2111.05409*, 2021.
- [GSR18] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. “Ava: A video dataset of spatio-temporally localized atomic visual actions.” In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [GTT21] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmhatt, and Charles C. Kemp. “ContactOpt: Optimizing Contact to Improve Grasps.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021.
- [Has70] W Keith Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.
- [HCT19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. “Resolving 3D human pose ambiguities with 3D scene constraints.” In *International Conference on Computer Vision (ICCV)*, pp. 2282–2292, 2019.
- [HGD17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask r-cnn.” In *International Conference on Computer Vision (ICCV)*, 2017.
- [HGT21] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. “Populating 3D scenes by learning human-scene interaction.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 14708–14718, 2021.

- [Hob02] P. Hobson. *The cradle of thought: challenging the origins of thinking*. MacMillan: London, UK., 2002.
- [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 *Conference on 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.
- [HRA17] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. “Modeling relationships in referential expressions with compositional modular networks.” In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [HRT13] Marta Halina, Federico Rossano, and Michael Tomasello. “The ontogenetic ritualization of bonobo gestures.” *Animal cognition*, **16**(4):653–666, 2013.
- [HSK10] Yasuhiko Hato, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita. “Pointing to space: modeling of deictic interaction referring to regions.” In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.
- [HVT19] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. “Learning joint reconstruction of hands and manipulated objects.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 11807–11816, 2019.
- [HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [IG05] Jana M Iverson and Susan Goldin-Meadow. “Gesture paves the way for language development.” *Psychological science*, **16**(5):367–371, 2005.
- [IMD16] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. “A hierarchical deep temporal model for group activity recognition.” In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [IPO13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **36**(7):1325–1339, 2013.

- [ISS17] Hamid Izadinia, Qi Shan, and Steven M Seitz. “Im2cad.” In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [JCH20] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. “Lemma: A multi-view dataset for learning multi-agent multi-task activities.” In *European Conference on Computer Vision (ECCV)*, volume 12371, pp. 767–786, 2020.
- [JED09] T. Judd, K. Ehinger, F. Durand, and A. Torralba. “Learning to predict where humans look.” In *ICCV*, 2009.
- [Joh73] Gunnar Johansson. “Visual perception of biological motion and a model for its analysis.” *Perception & psychophysics*, **14**(2):201–211, 1973.
- [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.
- [JSW21] Kaiwen Jiang, Stephanie Stacy, Chuyu Wei, Adelpha Chan, Federico Rossano, Yixin Zhu, and Tao Gao. “Individual vs. Joint Perception: a Pragmatic Model of Pointing as Communicative Smithian Helping.” In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2021.
- [KAR86] Alfred Kobsa, Jurgen Allgayer, Carola Reddig, Norbert Reithinger, Dagmar Schmauks, Karin Harbusch, and Wolfgang Wahlster. “Combining deictic gestures and natural language for referent identification.” In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 1986.
- [KAS14] Hilde Kuehne, Ali Arslan, and Thomas Serre. “The language of actions: Recovering the syntax and semantics of goal-directed human activities.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 780–787, 2014.
- [Ken04] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.
- [KF91] Robert M Krauss and Susan R Fussell. “Perspective-taking in communication: Representations of others’ knowledge in reference.” *Social cognition*, **9**(1):2–24, 1991.

- [KGS13] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. “Learning human activities and object affordances from rgb-d videos.” *International Journal of Robotics Research (IJRR)*, **32**(8):951–970, 2013.
- [KH06] F. Kaplan and V. Hafner. “The challenges of joint attention.” *Interaction Studies*, **7**(2):135–169, 2006.
- [KHA16] Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. “Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction.” In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2016.
- [KHH21] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. “PARE: Part Attention Regressor for 3D Human Body Estimation.” In *International Conference on Computer Vision (ICCV)*, pp. 11127–11137, 2021.
- [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.
- [Kit03] Sotaro Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003.
- [KIU03] M. Kumashiro, H. Ishibashi, Y. Uchiyama, S. Itakura, A. Murata, and A. Iriki. “Natural imitation induced by joint attention in Japanese monkeys.” *International Journal of Psychophysiology*, **50**(1):81 – 99, 2003.
- [KJD08] Hyundo Kim, Hector Jasso, Gedeon Deák, and Jochen Triesch. “A robotic model of the development of gaze following.” In *2008 7th IEEE International Conference on Development and Learning*, pp. 238–243. IEEE, 2008.
- [KJG11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. “HMDB: A large video database for human motion recognition.” In *ICCV*, 2011.
- [KOM14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. “Referitgame: Referring to objects in photographs of natural scenes.” In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [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.

- [KS14] Hema S. Koppula and Ashutosh Saxena. “Physically grounded spatio-temporal object affordances.” In *European Conference on Computer Vision (ECCV)*, volume 8691, pp. 831–847, 2014.
- [KSD20] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. “Contextual encoder–decoder network for visual saliency prediction.” *Neural Networks*, **129**:261–270, 2020.
- [KT15] Vikash Kumar and Emanuel Todorov. “Mujoco haptix: A virtual reality system for hand manipulation.” In *International Conference on Humanoid Robots (HUMANOIDS)*, pp. 657–663, 2015.
- [KTS14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. “Large-scale video classification with convolutional neural networks.” In *Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [KZG17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations.” *International Journal of Computer Vision (IJCV)*, **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.
- [LCH04] Ulf Liszowski, Malinda Carpenter, Anne Henning, Tricia Striano, and Michael Tomasello. “Twelve-month-olds point to share attention and interest.” *Developmental science*, **7**(3):297–307, 2004.
- [LCL07] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. “Crowds by example.” In *Computer Graphics Forum (CGF)*, 2007.
- [LCS06] Ulf Liszowski, Malinda Carpenter, Tricia Striano, and Michael Tomasello. “12- and 18-month-olds point to provide information for others.” *Journal of cognition and development*, **7**(2):173–187, 2006.
- [LFR17] Jeffrey I Lipton, Aidan J Fay, and Daniela Rus. “Baxter’s homunculus: Virtual reality spaces for teleoperation in manufacturing.” *Robotics and Automation Letters (RA-L)*, **3**(1):179–186, 2017.
- [LJX21] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. “Semi-supervised 3D hand-object poses estimation with interactions in time.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 14687–14697, 2021.

- [LLB19] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. “Clevr-ref+: Diagnosing visual reasoning with referring expressions.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [LLR18] Yin Li, Miao Liu, and James M Rehg. “In the eye of beholder: Joint learning of gaze and actions in first person video.” In *European Conference on Computer Vision (ECCV)*, 2018.
- [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)*, volume 8693, pp. 740–755, 2014.
- [LMR99] Michael Land, Neil Mennie, and Jennifer Rusted. “The roles of vision and eye movements in the control of activities of daily living.” *Perception*, **28**(11):1311–1328, 1999.
- [LMR15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. “SMPL: A Skinned Multi-Person Linear Model.” *Transactions on Graphics (TOG)*, **34**(6):248:1–248:16, 2015.
- [LPR15] Andy Lücking, Thies Pfeiffer, and Hannes Rieser. “Pointing and reference reconsidered.” *Journal of Pragmatics*, **77**:56–79, 2015.
- [LSC19] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. “Estimating 3D motion and forces of person-object interactions from monocular video.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 8640–8649, 2019.
- [LWS19] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. “Improving referring expression grounding with cross-modal attention-guided erasing.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [LXL20] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. “Pastanet: Toward human activity knowledge engine.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 382–391, 2020.
- [LZL10] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. “Action recognition based on a bag of 3d points.” In *Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [LZS20] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. “Multi-task collaborative network for joint referring expression comprehension and segmentation.” In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

- [MAZ19] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfrueud, Carl Vondrick, et al. “Moments in Time Dataset: one million videos for event understanding.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.
- [MC94] C. Moore and V. Corkum. “Social Understanding at the End of the First Year of Life.” *Developmental Review*, **14**(4):349 – 372, 1994.
- [McN12] David McNeill. *How language began: Gesture and speech in human evolution*. Cambridge University Press, 2012.
- [MD95] C. Moore and P. J. Dunham. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.
- [MG98] P. Mundy and A. Gomes. “Individual differences in joint attention skill development in the second year.” *Infant Behavior and Development*, **21**(3):469 – 482, 1998.
- [MHT16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. “Generation and comprehension of unambiguous object descriptions.” In *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.” *Transactions on Graphics (TOG)*, **35**(6):173–1, 2016.
- [MMR98] M. Morales, P. Mundy, and J. Rojas. “Following the direction of gaze and language development in 6-month-olds.” *Infant Behavior and Development*, **21**(2):373 – 377, 1998.
- [Mon03] Stephen Monsell. “Task switching.” *Trends in cognitive sciences*, **7**(3):134–140, 2003.
- [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.” *Transactions on Graphics (TOG)*, **36**(4):44, 2017.
- [Nag04] Y. Nagai. *Understanding the Development of Joint Attention from a Viewpoint of Cognitive Developmental Robotics*. PhD thesis, Osaka University, 2004.

- [Nag05] Y. Nagai. “The Role of Motion Information in Learning Human-Robot Joint Attention.” In *ICRA*, 2005.
- [Nee97] Amy Needham. “Factors affecting infants’ use of featural information in object segregation.” *Current Directions in Psychological Science*, **6**(2):26–33, 1997.
- [NFG19] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. “Grounded human-object interaction hotspots from video.” In *International Conference on Computer Vision (ICCV)*, pp. 8688–8697, 2019.
- [NG20] Tushar Nagarajan and Kristen Grauman. “Learning affordance landscapes for interaction exploration in 3D environments.” In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 2005–2015, 2020.
- [NHM03] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. “A constructive model for the development of joint attention.” *Connect. Sci.*, **15**:211–229, 2003.
- [NNH20] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. “Detecting hands and recognizing physical contact in the wild.” In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7841–7851, 2020.
- [OHP11] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. “A large-scale benchmark dataset for event recognition in surveillance video.” In *Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [PCG19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. “Expressive Body Capture: 3D Hands, Face, and Body from a Single Image.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [PES09] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. “You’ll never walk alone: Modeling social behavior for multi-target tracking.” In *International Conference on Computer Vision (ICCV)*, 2009.
- [PJS12] H. S. Park, E. Jain, and Y. Sheikh. “3D Gaze Concurrences From Head-mounted Cameras.” In *NIPS*, 2012.
- [PMR12] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. “Structured Learning of Human Interactions in TV Shows.” *IEEE TPAMI*, **34**(12):2441–2453, 2012.
- [PR12] Hamed Pirsiavash and Deva Ramanan. “Detecting activities of daily living in first-person camera views.” In *Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [PS15] H. S. Park and J. Shi. “Social Saliency Prediction.” In *CVPR*, 2015.

- [PSG16] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. O’Connor. “Shallow and Deep Convolutional Networks for Saliency Prediction.” In *CVPR*, 2016.
- [PW14] Karola Pitsch and Sebastian Wrede. “When a robot orients visitors to an exhibit. Referential practices and interactional dynamics in real world HRI.” In *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2014.
- [PWC15] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [QLZ20] Shuwen Qiu, Hangxin Liu, Zeyu Zhang, Yixin Zhu, and Song-Chun Zhu. “Human-Robot Interaction in a Shared Augmented Reality Workspace.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2020.
- [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.
- [QZH18] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. “Human-centric indoor scene synthesis using stochastic grammar.” In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [RAA12] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. “A database for fine grained activity detection of cooking activities.” In *Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [RBH21] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. “HuMoR: 3D Human Motion Model for Robust Pose Estimation.” In *International Conference on Computer Vision (ICCV)*, pp. 11488–11499, 2021.
- [RCJ21] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. “Home action genome: Cooperative compositional action understanding.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 11184–11193, 2021.
- [RF17] J. Redmon and A. Farhadi. “YOLO9000: Better, Faster, Stronger.” In *CVPR*, 2017.
- [RF18] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement.” *arXiv preprint arXiv:1804.02767*, 2018.

- [RGH20] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. “Contact and human dynamics from monocular video.” In *European Conference on Computer Vision (ECCV)*, volume 12350, pp. 71–87, 2020.
- [RHA16] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. “Detecting events and key actors in multi-person videos.” In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [RHG15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks.” In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.
- [RHG16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: towards real-time object detection with region proposal networks.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **39**(6):1137–1149, 2016.
- [RKS12] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. “Reconstructing 3d human pose from 2d image landmarks.” In *European Conference on Computer Vision (ECCV)*, 2012.
- [RKV15] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. “Where are they looking?” In *NIPS*, 2015.
- [RM51] Herbert Robbins and Sutton Monro. “A stochastic approximation method.” *The annals of mathematical statistics*, pp. 400–407, 1951.
- [RME01] Joshua S Rubinstein, David E Meyer, and Jeffrey E Evans. “Executive control of cognitive processes in task switching.” *Journal of experimental psychology: human perception and performance*, **27**(4):763, 2001.
- [RRH16] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. “Grounding of textual phrases in images by reconstruction.” In *European Conference on Computer Vision (ECCV)*, 2016.
- [RRR16] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. “Recognizing fine-grained and composite activities using hand-centric features and script data.” *International Journal of Computer Vision (IJCV)*, **119**(3):346–373, 2016.
- [RRT15] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. “A Dataset for Movie Description.” In *CVPR*, 2015.
- [RSJ21] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. “FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration.” In *International Conference on Computer Vision Workshops (ICCVw)*, 2021.

- [RVK17] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba. “Following Gaze in Video.” In *ICCV*, 2017.
- [RWP20] Ilija Radosavovic, Xiaolong Wang, Lerrel Pinto, and Jitendra Malik. “State-only imitation learning for dexterous manipulation.” In *International Conference on Intelligent Robots and Systems (IROS)*, pp. 7865–7871, 2020.
- [SCH14] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. “SceneGrok: Inferring action maps in 3D environments.” *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.” *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.
- [SCW15] Xingjian Shi, Zhoung Chen, Hao Wang, Dit Yan Yeung, Wai Kin Wong, and Wang Chun Woo. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting.” *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.
- [SEP15] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. “Probabilistic detection of pointing directions for human-robot interaction.” In *International Conference on Digital Image Computing: Techniques and Applications*, 2015.
- [SEP16] Dadhichi Shukla, Özgür Erkent, and Justus Piater. “A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios.” In *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016.
- [SF10] Boris Schauerte and Gernot A Fink. “Focusing computational visual attention in multi-modal human-robot interaction.” In *International conference on multi-modal interfaces and the workshop on machine learning for multimodal interaction*, 2010.
- [SF15] Aimee E Stahl and Lisa Feigenson. “Observing the unexpected enhances infants’ learning and exploration.” *Science*, **348**(6230):91–94, 2015.
- [SGB05] A. P. Shon, D. B. Grimes, C. L. Baker, M. W. Hoffman, S. Zhou, and R. P. N. Rao. “Probabilistic Gaze Imitation and Saliency Learning in a Robotic Head.” In *ICRA*, 2005.

- [SGS18] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Kar-teek Alahari. “Charades-ego: A large-scale dataset of paired third and first person videos.” *arXiv preprint arXiv:1804.09626*, 2018.
- [SGS20] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. “Understanding human hands in contact at internet scale.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 9869–9878, 2020.
- [SHK12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. “Indoor segmentation and support inference from rgb-d images.” In *European Conference on Computer Vision (ECCV)*, 2012.
- [SHS16] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. “Social behavior prediction from first person videos.” *arXiv preprint arXiv:1611.09464*, 2016.
- [SHY07] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada. “Acquisition of joint attention through natural interaction utilizing motion cues.” *Advanced Robotics*, **21**(9):983–999, 2007.
- [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 *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [SM13] Sebastian Stein and Stephen J McKenna. “Combining embedded accelerometers with computer vision for recognizing food preparation activities.” In *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2013.
- [SPR09] R. R. da Silva, C. A. Policastro, and R. A. F. Romero. “Relational reinforcement learning applied to shared attention.” In *IJCNN*, pp. 2943–2949, 2009.
- [SRA12] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. “Single image 3D human pose estimation from noisy observations.” In *Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [SRF10] Boris Schauerte, Jan Richarz, and Gernot A Fink. “Saliency-based identification and recognition of pointed-at objects.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2010.
- [SVP98] John Sweller, Jeroen JG Van Merriënboer, and Fred GWC Paas. “Cognitive architecture and instructional design.” *Educational Psychology Review*, **10**(3):251–296, 1998.

- [SVW16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. “Hollywood in homes: Crowdsourcing data collection for activity understanding.” In *European Conference on Computer Vision (ECCV)*, 2016.
- [SX14] Shuran Song and Jianxiong Xiao. “Sliding shapes for 3d object detection in depth images.” In *European Conference on Computer Vision (ECCV)*, 2014.
- [SXR15] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. “Joint inference of groups, events and human roles in aerial videos.” In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [SYZ17] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. “Semantic scene completion from a single depth image.” In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. “UCF101: A dataset of 101 human actions classes from videos in the wild.” *arXiv preprint arXiv:1212.0402*, 2012.
- [SZZ20] Stephanie Stacy, Qingyi Zhao, Minglu Zhao, Max Kleiman-Weiner, and Tao Gao. “Intuitive signaling through an “Imagined We”.” In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020.
- [Tat07] B.W. Tatler. “The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions.” *Journal of Vision*, **7**(14):4–4, 2007.
- [TCC05] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. “Understanding and sharing intentions: The origins of cultural cognition.” *Behavioral and Brain Sciences*, **28**(5):675–691, 2005.
- [TCH17] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. “Human pose forecasting via deep markov models.” In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017.
- [TCS08] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. “Machine recognition of human activities: A survey.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **18**(11):1473–1488, 2008.
- [TDR19] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. “COIN: A large-scale dataset for comprehensive instructional video analysis.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

- [TGB20] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. “GRAB: A dataset of whole-body human grasping of objects.” In *European Conference on Computer Vision (ECCV)*, volume 12349, pp. 581–600, 2020.
- [THK87] Nancy Termine, Timothy Hrynck, 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.
- [Tom08] M. Tomasello. *Origins of Human Communication*. The MIT Press, 2008.
- [Tom10] Michael Tomasello. *Origins of human communication*. MIT press, 2010.
- [TRA17] Denis Tome, Christopher Russell, and Lourdes Agapito. “Lifting from the deep: Convolutional 3d pose estimation from a single image.” *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [TS10] Anand Thobbi and Weihua Sheng. “Imitation learning of hand gestures and its evaluation for humanoid robots.” In *International Conference on Information and Automation (ICIA)*, pp. 60–65, 2010.
- [TSZ20] Ning Tang, Stephanie Stacy, Minglu Zhao, Gabriel Marquez, and Tao Gao. “Bootstrapping an imagined We for cooperation.” In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020.
- [TZS16] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. “Movieqa: Understanding stories in movies through question-answering.” In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [VBC19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning.” *Nature*, **575**(7782):350–354, 2019.
- [VPR13a] C. Vondrick, D. Patterson, and D. Ramanan. “Efficiently Scaling up Crowdsourced Video Annotation.” *IJCV*, **101**(1):184–204, 2013.
- [VPR13b] Carl Vondrick, Donald Patterson, and Deva Ramanan. “Efficiently scaling up crowdsourced video annotation.” *International Journal of Computer Vision (IJCV)*, **101**(1):184–204, 2013.
- [VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” *arXiv preprint arXiv:1706.03762*, 2017.

- [WAK19] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. “Deep contextual attention for human-object interaction detection.” In *International Conference on Computer Vision (ICCV)*, pp. 5694–5702, 2019.
- [WFF19] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. “Long-term feature banks for detailed video understanding.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [WGG17] X. Wang, R. Girdhar, and A. Gupta. “Binge Watching: Scaling Affordance Learning from Sitcoms.” In *CVPR*, 2017.
- [Woo98] Amanda L Woodward. “Infants selectively encode the goal object of an actor’s reach.” *Cognition*, **69**(1):1–34, 1998.
- [Woo99] Amanda L Woodward. “infants’ ability to distinguish between purposeful and non-purposeful behaviors.” *Infant Behavior and Development*, **22**(2):145–160, 1999.
- [WS18] Wenguan Wang and Jianbing Shen. “Deep Visual Attention Prediction.” *IEEE TIP*, **27**(5):2368–2378, 2018.
- [WWZ21] Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. “Communicative Learning with Natural Gestures for Embodied Navigation Agents with Human-in-the-Scene.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2021.
- [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.
- [WY21] Zhenzhen Weng and Serena Yeung. “Holistic 3D Human and Scene Mesh Estimation from Single View Images.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 334–343, 2021.
- [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 *Conference on 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 *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.
- [XLZ18] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. “Unified perceptual parsing for scene understanding.” In *European Conference on Computer Vision (ECCV)*, volume 11209, pp. 418–434, 2018.
- [XWL19] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. “Learning to detect human-object interactions with knowledge.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2019–2028, 2019.
- [YCW20] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. “Improving One-stage Visual Grounding by Recursive Sub-query Construction.” *arXiv preprint arXiv:2008.01059*, 2020.
- [YGW19] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. “A fast and accurate one-stage approach to visual grounding.” In *International Conference on Computer Vision (ICCV)*, 2019.
- [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.
- [YKF17] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. “Dilated residual networks.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 472–480, 2017.
- [YLF20] Tao Yuan, Hangxin Liu, Lifeng Fan, Zilong Zheng, Tao Gao, Yixin Zhu, and Song-Chun Zhu. “Joint Inference of States, Robot Knowledge, and Human (False-)Beliefs.” In *International Conference on Robotics and Automation (ICRA)*, 2020.
- [YLS18] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. “Mattnet: Modular attention network for referring expression comprehension.” In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [YLY18a] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. “Free-Form Image Inpainting with Gated Convolution.” *arXiv preprint arXiv:1806.03589*, 2018.
- [YLY18b] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. “Generative image inpainting with contextual attention.” In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

- [YLY19] Sibeï Yang, Guanbin Li, and Yizhou Yu. “Cross-modal relationship inference for grounding referring expressions.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [YLY20] Sibeï Yang, Guanbin Li, and Yizhou Yu. “Graph-Structured Referring Expression Reasoning in The Wild.” In *Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [YPY16] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. “Modeling context in referring expressions.” In *European Conference on Computer Vision (ECCV)*, 2016.
- [YRJ18] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. “Every moment counts: Dense detailed labeling of actions in complex videos.” *International Journal of Computer Vision (IJCV)*, **126**(2-4):375–389, 2018.
- [YRL19] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. “Cross-modal self-attention network for referring image segmentation.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [ZD14] C. L. Zitnick and P. Dollár. “Edge Boxes: Locating Object Proposals from Edges.” In *ECCV*, 2014.
- [ZGF20] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Josh Tenenbaum, and Song-Chun Zhu. “Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense.” *Engineering*, **6**(3):310–345, 2020.
- [ZHN20] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. “Generating 3D people in scenes without people.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 6194–6204, 2020.
- [Zip49] George Kingsley Zipf. *Human Behaviour and the Principles of Least Effort*. Addison-Wesley, 1949.
- [ZJZ16] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. “Inferring forces and learning human utilities from videos.” In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [ZLH18] Chuhan Zou, Zhizhong Li, and Derek Hoiem. “Complete 3D Scene Parsing from Single RGBD Image.” *International Journal of Computer Vision (IJCV)*, 2018.
- [ZLK17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. “Places: A 10 million image database for scene recognition.” *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **40**(6):1452–1464, 2017.

- [ZLX14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. “Learning deep features for scene recognition using places database.” In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.
- [ZMJ18] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation.” In *International Conference on Robotics and Automation (ICRA)*, pp. 5628–5635, 2018.
- [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 *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [ZNC18] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. “Grounding referring expressions in images by variational context.” In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [ZSQ17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. “Pyramid scene parsing network.” In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.
- [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 *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [ZTM08] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. “SUN: A Bayesian framework for saliency using natural statistics.” *Journal of Vision*, **8**(7):32–32, 2008.
- [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.
- [ZXC18] Luowei Zhou, Chenliang Xu, and Jason J Corso. “Towards automatic learning of procedures from web instructional videos.” In *AAAI Conference on Artificial Intelligence*, 2018.
- [ZZ13] Yibiao Zhao and Song-Chun Zhu. “Scene parsing by integrating function, geometry and appearance models.” In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

- [ZZB21] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. “Learning Motion Priors for 4D Human Body Capture in 3D Scenes.” In *International Conference on Computer Vision (ICCV)*, October 2021.
- [ZZC15] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. “Understanding tools: Task-oriented object modeling, learning and recognition.” In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [ZZM20] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. “PLACE: Proximity learning of articulation and contact in 3D environments.” In *International Conference on 3D Vision (3DV)*, pp. 642–651, 2020.
- [ZZP18] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. “Semantic understanding of scenes through the ADE20K dataset.” *International Journal of Computer Vision (IJCV)*, 2018.
- [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 *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.