Holistic^{++} Scene Understanding: Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-Object Interaction and Physical Commonsense

Yixin Chen*[1], Siyuan Huang*[1], Tao Yuan[1], Siyuan Qi[1,2], Yixin Zhu[1,2], and Song-Chun Zhu[1,2]

* Equal Contributors

[1] University of California, Los Angeles (UCLA)

[2] International Center for AI and Robot Autonomy (CARA)

{ethanchen,huangsiyuan,taoyuan,syqi,yixin.zhu}@ucla.edu,sczhu@stat.ucla.edu

Abstract

We propose a new 3D 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. The intuition behind is to leverage the coupled nature of these two tasks to improve the granularity and performance of scene understanding. We propose to exploit two critical and essential connections between these two tasks: (i) human-object interaction (HOI) to model the fine-grained relations between agents and objects in the scene, and (ii) physical commonsense to model the physical plausibility of the reconstructed scene. The optimal configuration of the 3D scene, represented by a parse graph, is inferred using Markov chain Monte Carlo (MCMC), which efficiently traverses through the non-differentiable joint solution space. Experimental results demonstrate that the proposed algorithm significantly improves the performance of the two tasks on three datasets, showing an improved generalization ability.

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 [41], that facilitates the 3D holistic scene understanding tasks. Consider a typical indoor scene shown in Figure 1 where a person sits in an office. We can effortlessly extract rich knowledge from the static scene, including 3D room layout, 3D position of all the objects and agents, and correct human-object interaction (HOI) relations in a physically plausible manner. In fact, psychology studies have established that even infants employ at least two constraints—HOI and physical commonsense—in perceiving occlusions [43, 20], tracking small objects even if contained by other objects [10], realizing object permanence [2], recognizing rational HOI [46, 37], understanding intuitive physic [11, 29, 1], and using exploratory play to understand the environment [42]. All the evidence calls for a treatment to integrate HOI and physical commonsense with a modern computer vision system for scene understanding.

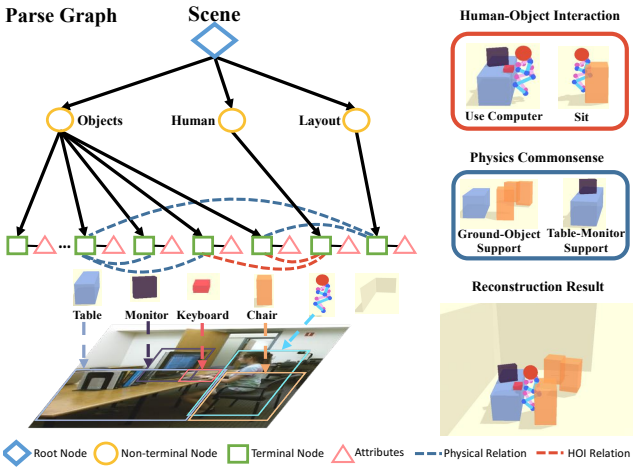


Figure 1. **holistic^{++} scene understanding** task requires to jointly recover a parse graph that represents the scene, including human poses, objects, camera pose, and room layout, all in 3D. Reasoning human-object interaction (HOI) helps reconstruct the detailed spatial relations between humans and objects. Physical commonsense (*e.g.*, physical property, plausibility, and stability) further refines relations and improves predictions.

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 [16, 62, 3, 40] or 3D human pose estimation [53, 32, 9]. 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 [12].

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 Figure 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 Figure 1). By learning HOI priors and using the inferred HOI as visual cues to adjust the fine-grained spatial relations between human and scene (objects and room layout), the geometric ambiguity (3D estimation solution space) in the single-view reconstruction would be largely eased, and the reconstruction performances of both tasks would be improved.

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

To jointly parse 3D human pose and 3D scene, we represent the configuration of an indoor scene by a parse graph shown in Figure 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 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 [34], Watch-n-Patch [47], and SUN RGB-D [38] 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 paper 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.

1.1. Related Work

Single-view 3D Human Pose Estimation: Previous methods on 3D pose estimation can be divided into two streams: (i) directly learning 3D pose from a 2D image [36, 23], and (ii) cascaded frameworks that first perform 2D pose estimation and then reconstruct 3D pose from the estimated 2D joints [53, 27, 32, 48, 6, 44]. 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*. [51] 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 [62, 40, 17, 61]. (ii) Align object proposals to RGB or depth image by treating objects as geometric primitives or CAD models [3, 39, 57]. (iii) Joint estimation of the room layout and 3D objects with contexts [40, 54, 7, 52, 62]. A more recent work by Huang *et al*. [16] 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 is essential for a more holistic indoor scene understanding as it provides crucial cues for human activities and physical interactions. In robotics and computer vision, prior work has exploited human-object relations in event, object, and scene modeling, but most work focuses on human-object relation detection in images [5, 30, 25, 21], probabilistic modeling from multiple data sources [45, 33, 13], and snapshots generation or scene synthesis [34, 24, 31, 18]. 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 [26, 22]. 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 [49, 28] and objects [60], understanding the support relationships and stability of objects [56], volumetric and occlusion reasoning [35, 55], inferring the hidden force [59], and reconstructing the 3D scene [15, 8] and 3D pose [51]. 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. Representation

The configuration of an indoor scene is represented by a parse graph $pg=(pt, E)$; see Figure 1. It combines a parse tree pt and contextual relations E among the leaf nodes. Here, a parse tree $pt=(V, R)$ includes the vertex set with a three-level hierarchical structure $V=V_r \cup V_m \cup V_t$ and the decomposing rules R, where the root node V_r represents the overall scene, the middle node V_m has three types of nodes (objects, human, and room layout), and the terminal nodes V_t contains child nodes of the middle nodes, representing the detected instances of the parent node in this scene. $E \subset V_t \times V_t$ is the set of contextual relations among the terminal nodes, represented by horizontal links.

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

- 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 use the same 3D bounding box parameterization as the one for the 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 $C^H \in \mathbb{R}^3$ (*i.e.*, hip), 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 provides strong and fine-grained constraints for holistic scene understanding. For instance, if a person is detected as sitting on a chair, we can constrain the relative 3D positions between this person and chair using a pre-learned spatial relation of "sitting."

3. Probabilistic Formulation

The parse graph pg is a comprehensive interpretation of the observed image I [58]. The goal of the holistic^{++} scene understanding is to infer the optimal parse graph pg^* given I by an MAP estimation:

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

We model the joint distribution by a Gibbs distribution, where the prior probability of parse graph can be decomposed into physical prior $\mathcal{E}_{phy}(pg)$ and HOI prior $\mathcal{E}_{hoi}(pg)$; balancing factors are neglected for simplicity.

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 $\mathcal{E}_{phy}(pg) = \mathcal{E}_s(pg) + \mathcal{E}_c(pg)$. Specifically:

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

$$\mathcal{E}_s(pg) = \sum_{(v_i, v_j)\in E_s} \mathcal{E}_o(v_i, v_j) + \mathcal{E}_{\text{height}}(v_i, v_j), \tag{2}$$

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 $\mathcal{E}_{\text{height}}(v_i, v_j)$ 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; $\mathcal{E}_o(v_i, v_j)=0$ when the supporting object is the floor and $\mathcal{E}_{\text{height}}(v_i, v_j)=0$ when the supporting object is the 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:

$$\mathcal{E}_c(pg)=\sum_{v\in(V_{\text{object}}\cup V_{\text{human}})}\mathcal{C}(v, V_{\text{layout}})+\sum_{\substack{v_i\in V_{\text{object}}\\ v_j\in V_{\text{human}}\\ (v_i,v_j)\notin E_{hoi}}}\mathcal{C}(v_i, v_j)+\sum_{\substack{v_i,v_j\in V_{\text{object}}\\ v_i,v_j\notin V_{\text{container}}}}\mathcal{C}(v_i, v_j), \quad (3)$$

where $\mathcal{C}()$ denotes the volume of intersection between entities. $V_{\text{container}}$ 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 by the interactions between human and objects:

$$\mathcal{E}_{hoi}(pg)=\sum_{(v_i,v_j)\in E_{hoi}}\mathcal{K}(v_i, v_j, a_{v_j}), \quad (4)$$

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

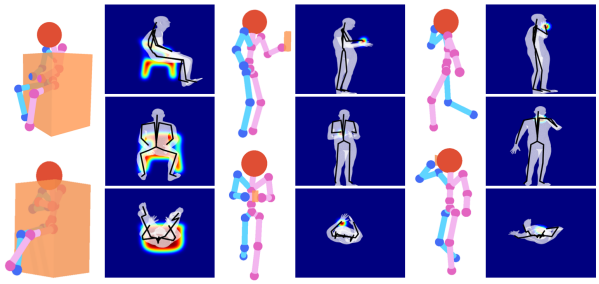
$$\mathcal{K}(v_i, v_j, a_{v_j})=-\log l(v_i, v_j|a_{v_j}), \quad (5)$$

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. We formulate the action detection as a *multi-label classification*; see Section 5.3 for details. The likelihood $l(\cdot)$ models the distance between key joints and the center of the object; *e.g.*, for "sitting," it models the relative spatial relation between the hip and the center of a chair. The likelihood can be learned from 3D HOI datasets with a multivariate Gaussian distribution $(\Delta x, \Delta y, \Delta z)\sim\mathcal{N}_3(\mu, \Sigma)$, where 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 poses with projected ones from inferred 3D results:

$$\mathcal{E}(I|pg)=\sum_{v\in V_{\text{object}}}\cdot\mathcal{D}_o(B(v), B'(v))+\sum_{v\in V_{\text{human}}}\cdot\mathcal{D}_h(Po(v), Po'(v)), \quad (6)$$

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.



(a) Sitting (b) Holding (c) Making phone call

Figure 2. Examples of typical HOIs and examples from the SHADE dataset. The heatmap indicates the probable locations of HOI.

4. SHADE Dataset

We collect SHADE (Synthetic Human Activities with Dynamic Environment), a self-annotated dataset that consists of dynamic 3D human skeletons and objects, to learn the prior model for each HOI. It is collected from a video game Grand Theft Auto V with various daily activities and HOIs. Currently, there are over 29 million frames of 3D human poses, where 772,229 frames are annotated. On average, each annotated frame is associated with 2.03 action labels and 0.89 HOIs. The SHADE dataset contains 19 fine-grained HOIs for both indoor and outdoor activities. By selecting most frequent HOIs and merging similar HOIs, we choose 6 final HOIs: *read [phone, notebook, tablet], sit-at [human-table relation], sit [human-chair relation], make-phone-call, hold, use-laptop*. Figure 2 shows some typical examples and relations in the dataset.

5. Joint Inference

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

5.1. 3D Scene Initialization

Following [15], 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 $v_i\in V_{\text{object}}$, we find its supporting object/layout by minimizing the supporting energy:

$$v_j^*=\arg\min_{v_j}\mathcal{E}_o(v_i, v_j)+\mathcal{E}_{\text{height}}(v_i, v_j)-\lambda_s\log p_{spt}(v_i, v_j), \quad (7)$$

where $v_j\in(V_{\text{object}}, V_{\text{layout}})$ 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.

5.2. 3D Human Pose Initialization

We take 2D poses as the input and predict 3D poses in a local 3D coordinate following [44], where the 2D poses are detected and estimated by [4]. The local 3D coordinate is centered at the human hip joint, and the z-axis is aligned with the up direction of the world coordinate.

To transform this local 3D pose into the world coordinate, we find the 3D world coordinate $\mathbf{v_{3D}} \in \mathbb{R}^3$ of one visible 2D joint $\mathbf{v_{2D}} \in \mathbb{R}^2$ (*e.g.*, head) by solving a linear equation with the camera intrinsic parameter K and estimated camera pose R. Per the pinhole camera projection model, we have

$$\alpha \begin{bmatrix} \mathbf{v_{2D}} \\ 1 \end{bmatrix} = K \cdot R \cdot \mathbf{v_{3D}}, \tag{8}$$

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 $\mathbf{v_{3D}}$.

5.3. Concurrent Action Detection

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

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 4. We project the 3D human poses of different HOIs into 2D poses with random camera poses. (ii) The dataset proposed and collected by [19], which also contains 3D poses of multiple persons in social interactions. We project 3D poses into 2D following the same method as in (i). (iii) The 2D poses in an action recognition dataset [50]. Our results show that the synthetic data can significantly expand the training set and help to avoid overfitting in concurrent action detection.

5.4. Inference

Given an initialized parse graph, we use MCMC with simulated annealing to jointly optimize the room layout, 3D objects, and 3D human poses through the non-differentiable energy space; see 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 agent in the scene, and search the interacting objects of each agent. (iii) Optimize objects, room layout, and human poses jointly with HOIs. (iv) Generate possible miss-detected objects by top-down sampling.

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

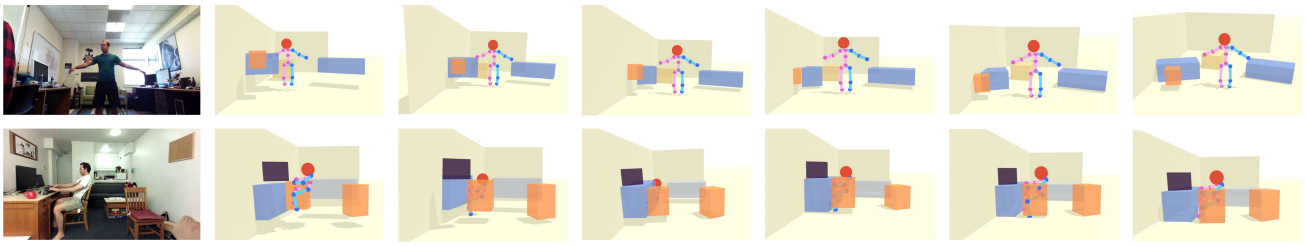
Dynamics: In Phase (i) and (iii), we use distinct MCMC processes. To traverse non-differentiable energy spaces, we design Markov chain dynamics q_1^o, q_2^o, q_3^o for objects, q_1^l, q_2^l for room layout, and q_1^h, q_2^h, q_3^h for human poses.

• 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 the object center. Each dynamic can diffuse in two directions: 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 rotates 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.

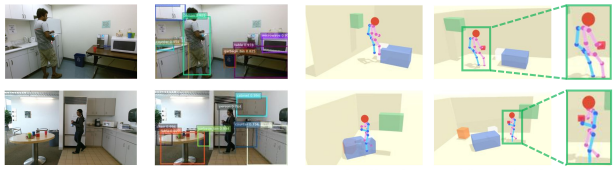
• Layout Dynamics: Dynamics q_1^l translates the wall towards or away from the layout center. Dynamics q_2^l adjusts the floor height, equivalent to changing the camera height.

In each sampling iteration, the algorithm proposes a new pg' from current pg under the proposal probability of $q(pg \to pg'|I)$ by applying one of the above dynamics. The generated proposal is accepted with respect to an acceptance rate



(a) Input (b) Initialization (c) Step 30 (d) Step 60 (e) Step 90 (f) Step 120 (g) Final output

Figure 3. The optimization process of the scene configuration by simulated annealing MCMC. Each step is the number of accepted proposal.



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

Figure 4. Illustration of the top-down sampling process. The object detection module misses the detection of the bottle held by the person, but our model can still recover the bottle by reasoning HOI.

$\alpha(\cdot)$ as in the Metropolis-Hastings algorithm [14]:

$$\alpha(pg \rightarrow pg') = \min(1, \frac{q(pg' \rightarrow pg) \cdot p(pg'|I)}{q(pg \rightarrow pg') \cdot p(pg|I)}), \quad (9)$$

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

Top-down sampling: By top-down sampling objects from HOIs relations, 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; we minimize the HOI energy in Equation 4 to determine the category and location of the object; see examples in Figure 4.

Implementation Details: In Phase (ii), we search the interacting objects for each agent involved in HOI by minimizing the energy in Equation 4. In Phase (iii), after matching each agent with their interacting objects, we can jointly optimize objects, room layout, and human poses with the constraint imposed by HOI. Figure 3 shows examples of the simulated annealing optimization process.

6. Experiments

Since the proposed task is new and challenging, limited data and state-of-the-art methods are available for the proposed problem. For fair evaluations and comparisons, we evaluate the proposed algorithm on three types of datasets: (i) Real data with full annotation on PiGraphs dataset [34] with limited 3D scenes. (ii) Real data with partial annotation on daily activity dataset Watch-n-Patch [47], 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 [38] and project the sampled skeletons back onto the 2D image plane.

6.1. Comparative methods

To the best of our knowledge, no previous algorithm jointly optimizes the 3D scene and 3D human pose from a single image. Therefore, we compare our model against state-of-the-art methods for each task. Particularly, we compare with [15] for single-image 3D scene reconstruction and VNect [27] for 3D pose estimation in the world coordinate.

Since VNect can only estimate a single person, we design an additional baseline for 3D multi-person human pose estimation in the world coordinate. We first extract a 2048-D image feature vector using the Global Geometry Network (GGN) [15] 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 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.

6.2. Dataset

PiGraphs [34] 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 with no human appearance or annotations, resulting in 36,551 test images.

Watch-n-Patch (WnP) [47] 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. We test our algorithm on 1,210 randomly selected frames.

SUN RGB-D [38] 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 augmented from SUN RGB-D dataset by sampling human poses in the scenes. Following methods of sampling imaginary human poses in [16], we extend the sampling to more generalized settings for various poses. The augmented human is represented by a 6-tuple

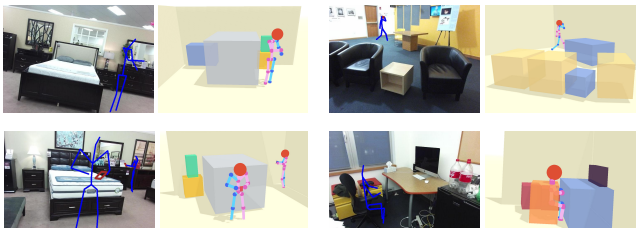


Figure 5. Augmenting SUN RGB-D with synthetic human poses.

$\langle a, \mu, t, r, s, \hat{\mu} \rangle$, where a is the action type, μ the pose template, t translation, r rotation, s scale, and $\hat{\mu}{=}\mu{\cdot}r{\cdot}s{+}t$ the imagined human skeleton. For each action label, we sample an imagined human pose inside a 3D scene: $\langle t^*, r^*, s^* \rangle = \arg\min_{t,r,s} \mathcal{E}_{phy}{+}\mathcal{E}_{hoi}$. If a is involved with any HOI unit, we further augment the 3D bounding box of the object. After sampling a human pose, we project the augmented 3D scenes back onto the 2D image plane using the ground truth camera matrix and camera pose; see examples in Figure 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 [15] 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 [27] 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.

6.3. Quantitative and Qualitative Results

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

Scene Reconstruction: We compute the 3D IoU and 2D IoU of object bounding boxes to evaluate the 3D scene reconstruction and the consistency between the 3D world and 2D image. Following the metrics described in [15], 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 1, the proposed method improves the state-of-the-art 3D scene reconstruction results on all three datasets without specific training on each of them. More importantly, it significantly improves the results on PiGraphs and Watch-n-Patch compared with [15]. The most likely reason is: [15] 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 combines such knowledge with 2D-3D consistency (likelihood) for joint inference, avoiding the over-fitting caused by the direct 3D estimation from 2D. Figure 6 shows the qualitative results on all three datasets.

Table 1. Quantitative Results of 3D Scene Reconstruction

Methods	Huang *et al.* [15]			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	67.3	-	0.375	**73.6**	-	**0.162**

Table 2. Quantitative Results of Global 3D Pose Estimation

Methods	VNect[27]		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 3. Ablative results of HOI on 3D object IoU (%), 3D pose estimation error (m), and miss-detection rate (MR, %)

Methods	*w/o hoi*			*Full model*		
HOI Type	Object ↑	Pose ↓	MR ↓	Object ↑	Pose ↓	MR ↓
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**

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 the 2D image plane and compute the pixel distance against the ground truth. See Table 2 for quantitative results. 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 a fair comparison. Although the baseline model achieves better performances since the baseline model fits well for the 3D human poses synthesized with limited templates, 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 performs consistently well, demonstrating an outstanding generalization ability across various datasets.

Ablative Analysis: To analyze the contributions of HOI and physical commonsense, we compare 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. Note that for the consistency of the ablative analysis across three different datasets, we merge the *sit* and *sit-at* into *sit*, and eliminate the *make-phone-call*. As shown in Table 3, the performances of both scene reconstruction and human pose estimation are

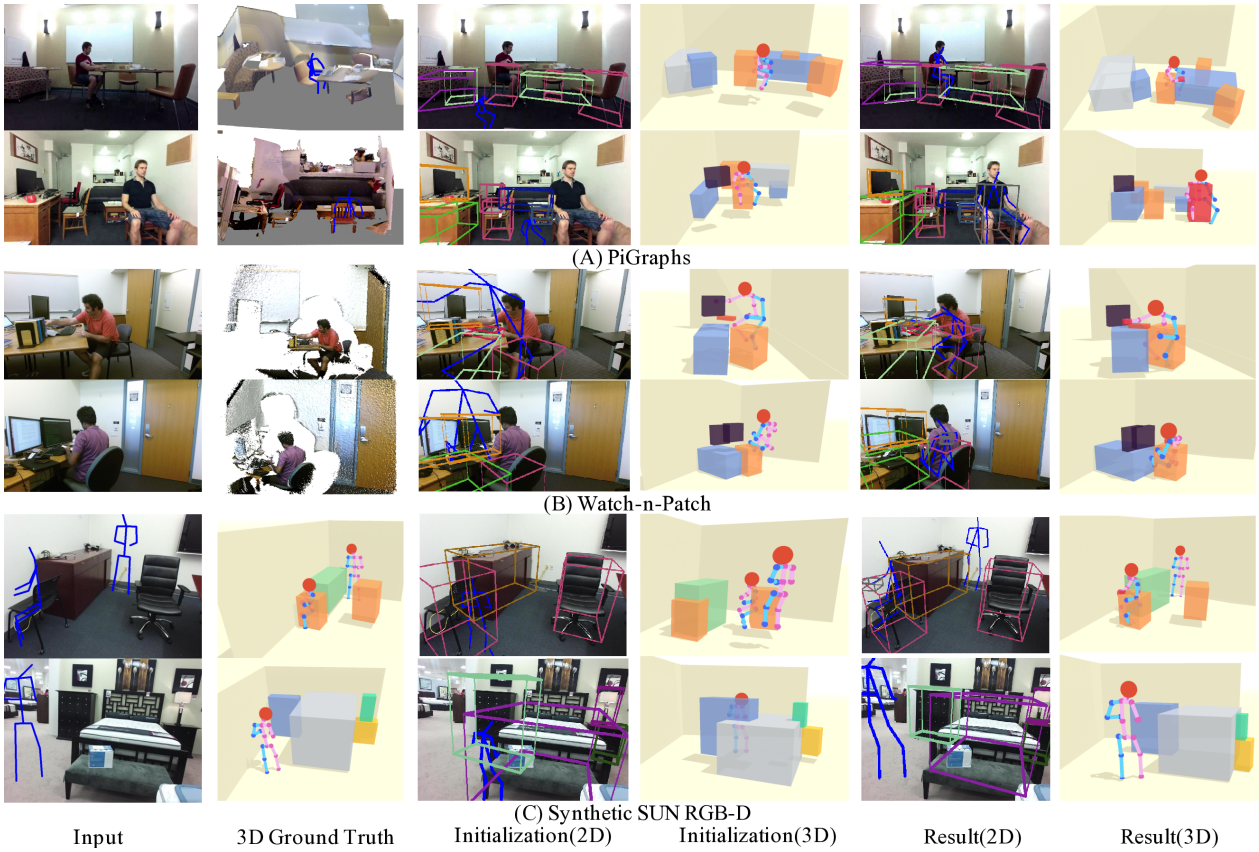


Figure 6. Qualitative results of the proposed method on three datasets. The proposed model improves the initialization with accurate spatial relations and physical plausibility and demonstrates an outstanding generalization across various datasets.

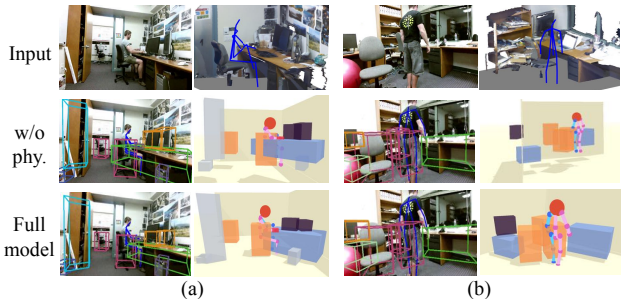


Figure 7. Qualitative comparison between (a) model *w/o phy.* and (b) the full model on PiGraphs dataset.

hindered without reasoning HOI, indicating HOI helps to infer the relative spatial relationship between agents and objects to improve the performance of both two tasks further. 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 test 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.* in two aspects: (i) 3D object detection IoU (from 23.5% to 24.9%), and (ii) physical violation (from 0.223m to 0.150m); see qualitative comparisons in Figure 7. The physical violation is computed as the distance between the lower surface of an object and the upper surface of its supporting object. Objects detected by model *w/o phy.* may float in the air or penetrate each other, while the full model yields physically plausible results.

7. Conclusion

This paper 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.

Acknowledgments: We thank Tengyu Liu from UCLA CS department for providing the SHADE dataset. The work reported herein was supported by DARPA XAI grant N66001-17-2-4029, ONR MURI grant N00014-16-1-2007, ONR robotics grant N00014-19-1-2153, ARO grant W911NF-18-1-0296, and an NVIDIA GPU donation grant.

References

[1] Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004. 1

[2] Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985. 1

[3] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[5] 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. 3

[6] Jungchan Cho, Minsik Lee, and Songhwai Oh. Complex nonrigid 3d shape recovery using a procrustean normal distribution mixture model. *International Journal of Computer Vision (IJCV)*, 117(3):226–246, 2016. 2

[7] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Understanding indoor scenes using 3d geometric phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[8] 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 (NIPS)*, 2018. 3

[9] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[10] Lisa Feigenson and Susan Carey. Tracking individuals via object-files: evidence from infants' manual search. *Developmental Science*, 6(5):568–584, 2003. 1

[11] György Gergely, Harold Bekkering, and Ildikó Király. Developmental psychology: Rational imitation in preverbal infants. *Nature*, 415(6873):755, 2002. 1

[12] James Jerome Gibson. *The ecological approach to visual perception*. Houghton, Mifflin and Company, 1979. 2

[13] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(10):1775–1789, 2009. 3

[14] W Keith Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970. 6

[15] 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 (NIPS)*, 2018. 3, 4, 6, 7

[16] 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. 2, 3, 6

[17] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[18] 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. 3

[19] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 5

[20] Philip J Kellman and Elizabeth S Spelke. Perception of partly occluded objects in infancy. *Cognitive psychology*, 15(4):483–524, 1983. 1

[21] 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. 3

[22] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017. 3

[23] 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. 2

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

[25] 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. 3

[26] Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: the straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4):636, 1983. 3

[27] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 2, 6, 7

[28] Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[29] Amy Needham. Factors affecting infants' use of featural information in object segregation. *Current Directions in Psychological Science*, 6(2):26–33, 1997. 1

[30] 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. 3

[31] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

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

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

[34] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)*, 35(4), 2016. 2, 3, 6

[35] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012. 3

[36] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[37] 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. 1

[38] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 6

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

[40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

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

[42] Aimee E Stahl and Lisa Feigenson. Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230):91–94, 2015. 1

[43] Nancy Termine, Timothy Hrynick, Roberta Kestenbaum, Henry Gleitman, and Elizabeth S Spelke. Perceptual completion of surfaces in infancy. *Journal of Experimental Psychology: Human Perception and Performance*, 13(4):524, 1987. 1

[44] Denis Tome, Christopher Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5

[45] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 3

[46] Amanda L Woodward. infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, 22(2):145–160, 1999. 1

[47] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 6

[48] 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. 2

[49] 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 (NIPS)*, 2015. 3

[50] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2011. 5

[51] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[52] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[53] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):3059–3066, 2017. 2

[54] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[55] 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. 3

[56] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[57] 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 (NIPS)*, 2014. 3

[58] Song-Chun Zhu, David Mumford, et al. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2007. 3

[59] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[60] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[61] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[62] Chuhang Zou, Zhizhong Li, and Derek Hoiem. Complete 3d scene parsing from single rgbd image. *International Journal of Computer Vision (IJCV)*, 2018. 2, 3