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

Yixin Chen*, Siyuan Huang*, Tao Yuan, Siyuan Qi, Yixin Zhu, Song-Chun Zhu

University of California, Los Angeles

ICCV 2019 Seoul, Korea

Holistic^{++} Scene Understanding

We propose a new task **Holistic^{++} Scene Understanding** which simultaneously solve both s 3D holistic scene parsing task and a human pose estimation task from a single RGB image.

- **Holistic 3D scene understanding**
  - The estimation of the 3D camera pose.
  - The estimation of the 3D room layout.
  - The estimation of the 3D object bounding boxes.
- **Global 3D human pose estimation**

Motivation

- Psychology studies have established that even infants employ human-object interaction (HOI) and physical commonsense in perceiving occlusions, tracking small objects, realizing object permanence, recognizing rational HOI, and understanding intuitive physics.
- Scene reconstruction and human pose estimation are intertwined tightly since the indoor scenes are invented and constructed by human designs to support daily activities.

Contribution

1. Propose a new **holistic^{++} scene understanding** task with a computational framework to jointly infer human poses, object poses, room layout, and camera pose, all in 3D.
2. Integrate **HOI** to bridge the human pose estimation and the scene reconstruction, reducing geometric ambiguities (solution space) of the single-view reconstruction.
3. Incorporate **physical commonsense**, which helps to predict physically plausible scenes and improve the 3D localization of both humans and objects.
4. Our method demonstrates 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.

Pipeline

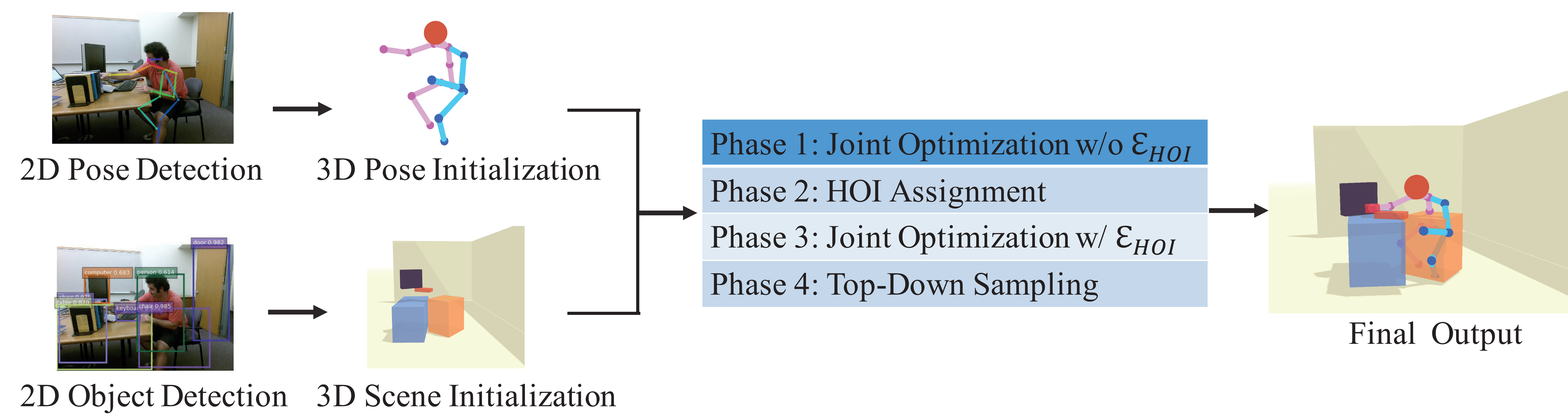
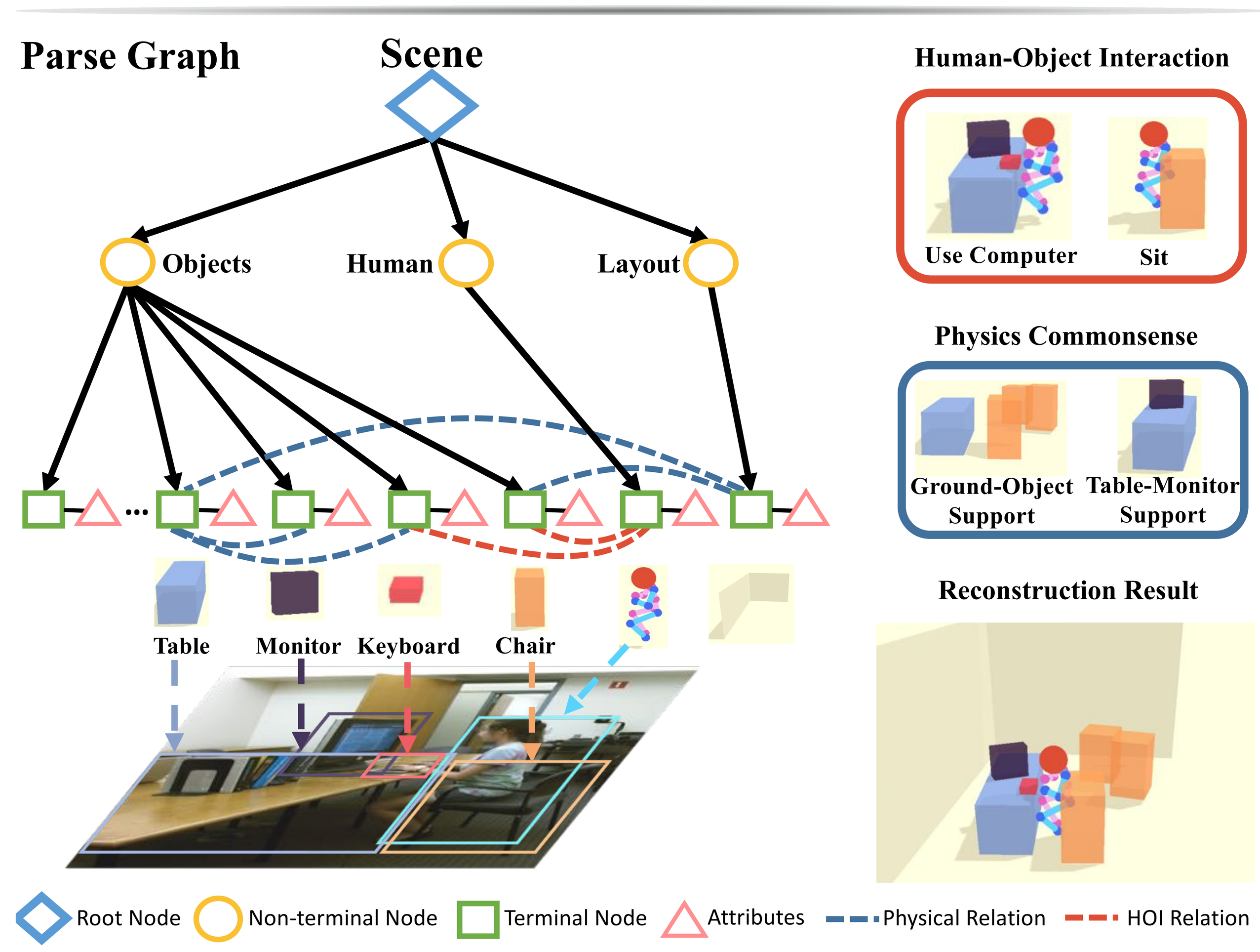


Figure 1:Overview of the proposed framework for holistic^{++} scene understanding.

Representation



Algorithm

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

Probabilistic Formulation and Inference

The configuration of an indoor scene is represented by a parse graph $pg = (pt, E)$, which combines a parse tree pt and contextual relations E among the terminal nodes defined on a Markov Random Field (MRF). The optimal parse graph pg^* given an image I is inferred by a maximum a posteriori (MAP) estimation:

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

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. **Human-object Interaction Prior** $\mathcal{E}_{hoi}(pg)$ evaluates the interaction between an object and a human given an action label. **Likelihood** $\mathcal{E}(I|pg)$ characterizes the consistency between the observed 2D image and the inferred 3D result.

Given an initial parse graph, we use **Markov chain Monte Carlo (MCMC)** with simulated annealing to jointly optimize the room layout, 3D object poses, and 3D human poses through the non-differentiable energy space.

Optimization Process

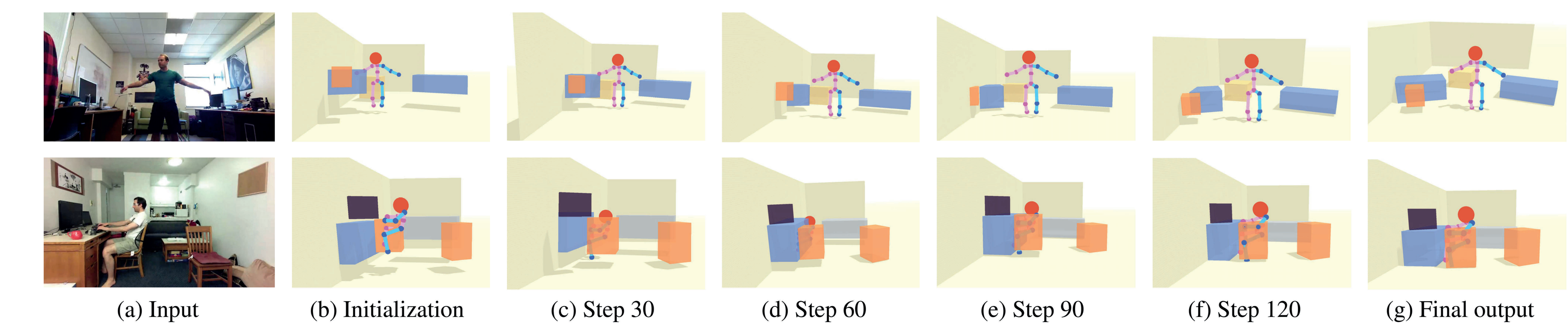


Figure 2:The optimization process of the scene configuration by simulated annealing MCMC.

Qualitative Results

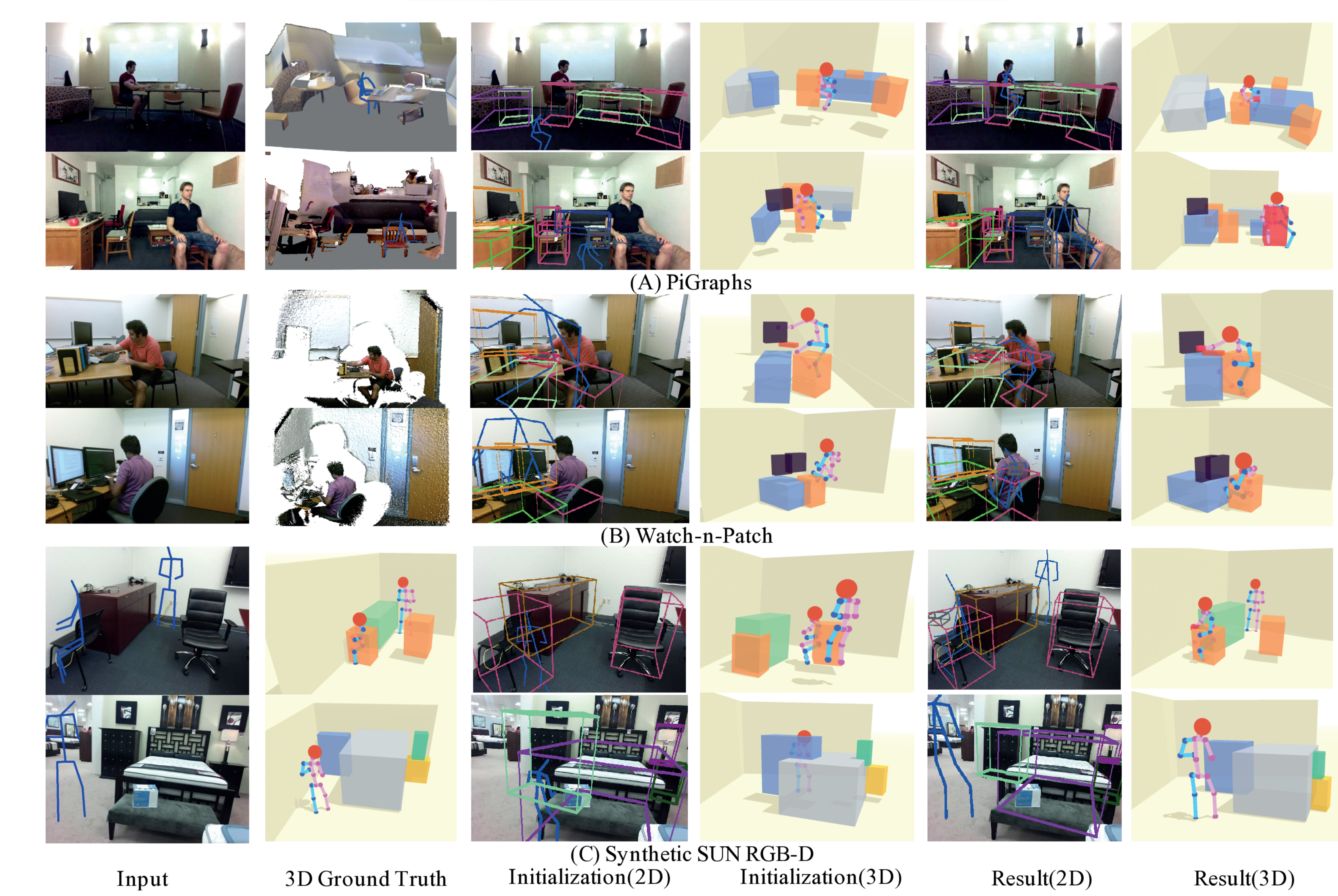


Figure 3:Qualitative results of the proposed method on three datasets.

Quantitative Results

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**

Ablative Study

Figure 4:Comparison of model *w/o phy* and full model.

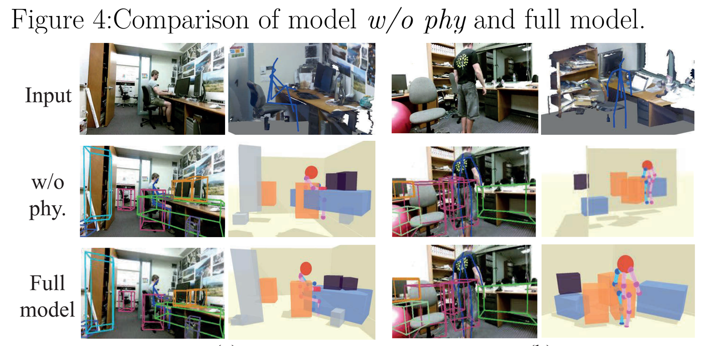


Table 3. Ablative results of HOI on 3D object IoU (%), 3D pose estimation error (m), and miss-detection rate (MR, %)

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**