Cooperative Holistic Scene Understanding: Unifying 3D Object, Layout, and Camera Pose Estimation

Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, Song-Chun Zhu

University of California, Los Angeles

Objective

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.

We aim to recover a **geometrically consistent** and **physically plausible** 3D scene and jointly solve all three tasks in an **efficient** and **cooperative** way, only from a single RGB image.

Motivation

- Humans are capable of performing such tasks effortlessly within 200ms.
- Most current methods are inefficient or only tackle the problem partially.

Problems

- **2D-3D consistency.** How to maintain a high consistency between the 2D image plane and the 3D world coordinate?
- **Cooperation.** How to solve the three tasks cooperatively and make different modules reinforce each other?
- **Physically Plausible.** How to model a 3D scene in a physically plausible fashion?

We solve these problems by cooperative training.

Contribution

1. Formulate an **end-to-end** model for 3D holistic scene understanding tasks.
2. Propose a novel **parametrization of the 3D bounding boxes** and **integrate physical constraint**, enabling the cooperative training.
3. **Bridge the gap** between the 2D image plane and the 3D world by introducing a differentiable objective function between the 2D and 3D bounding boxes.
4. Our method significantly outperforms previous methods and runs in real-time.

Framework

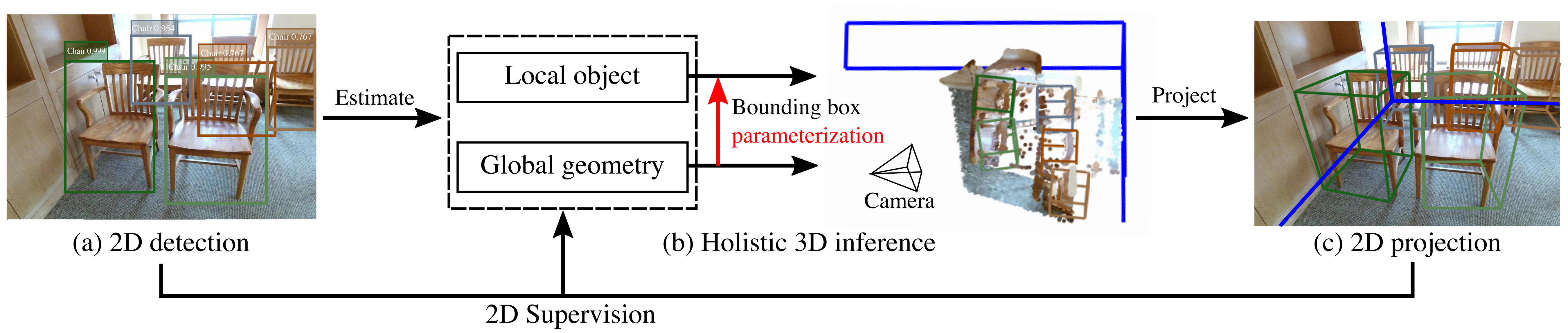

Figure 1: Overview of the proposed framework for cooperative holistic scene understanding.

(a) We first detect 2D objects and generate their bounding boxes, given a single RGB image as the input, from which (b) we can estimate 3D object bounding boxes, 3D room layout, and 3D camera pose. (c) We project 3D objects to the image plane with the learned camera pose, forcing the projection from the 3D estimation to be consistent with 2D estimation.

Parametrization

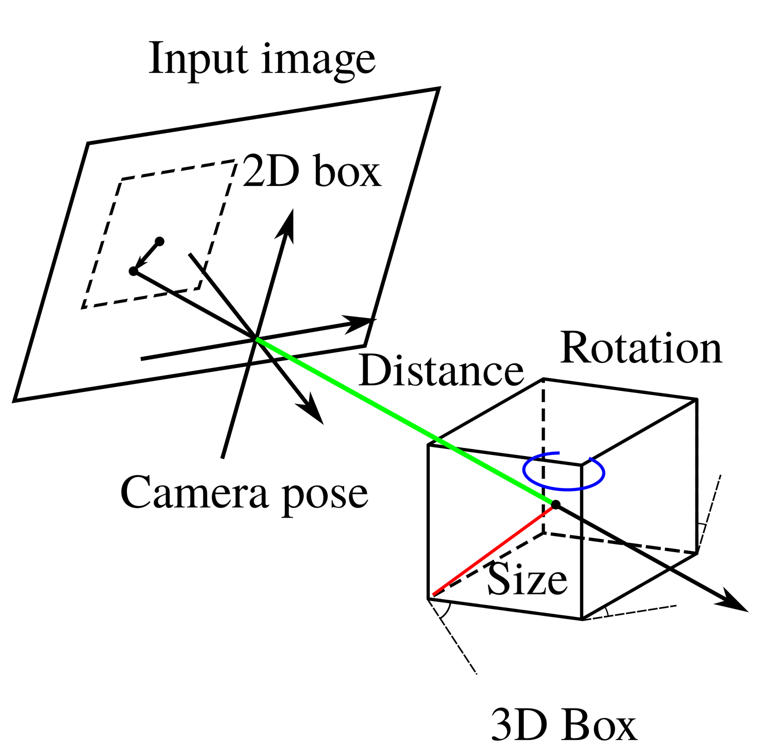

Figure 2: 3D Object Parametrization.

Network

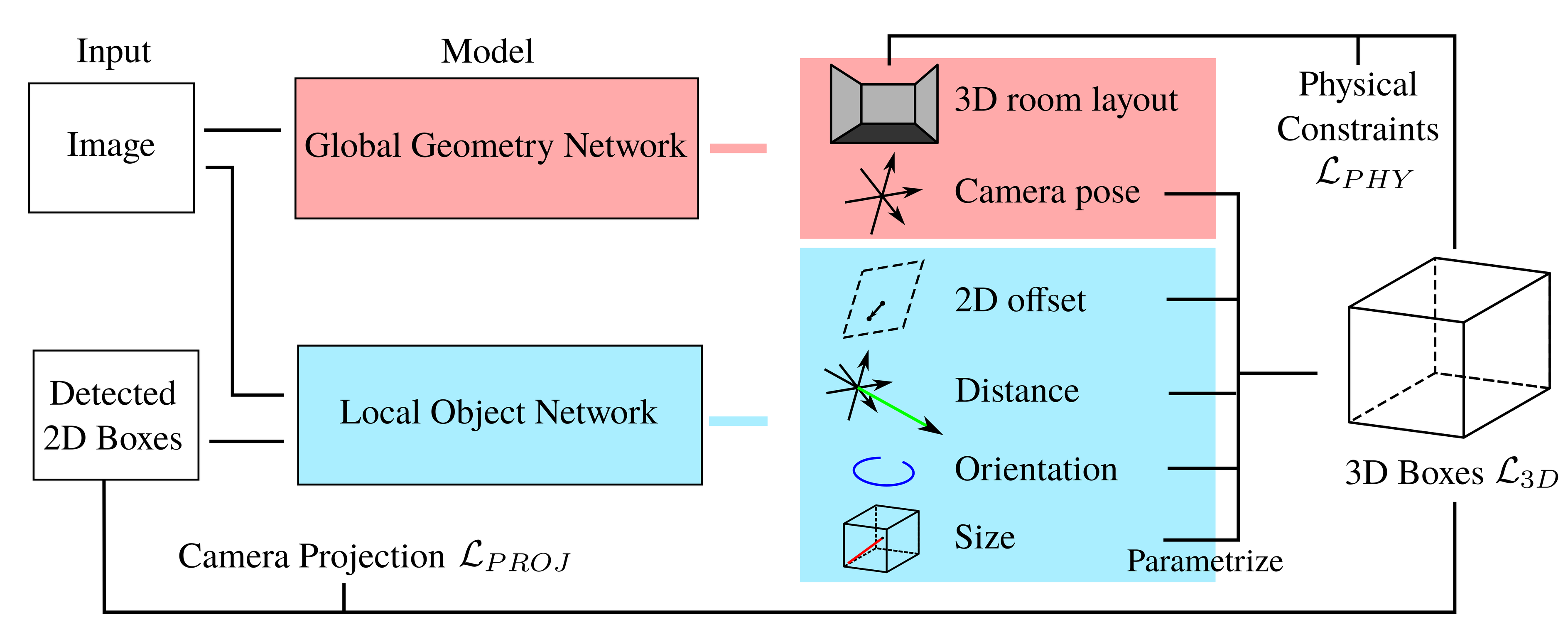

Figure 3: Illustration of the network architecture.

Cooperative Training

We propose three **cooperative losses** which jointly provide supervisions and makes a physically plausible estimation.

- **3D bounding box loss**: optimizes the GGN and LON cooperatively by constraining the corners of each bounding box.

$$\mathcal{L}_{\text{3D}} = \frac{1}{N}\sum_{j=1}^{N}\left|h(C_j^W, R(\theta_j), S_j) - X_j^{W*}\right|_2^2$$

- **2D projection loss**: maintains the coherence between the 2D bounding boxes and the 3D bounding boxes.

$$\mathcal{L}_{\text{PROJ}} = \frac{1}{N}\sum_{j=1}^{N}\left|f(X_j^W, R, K) - X_j^{I*}\right|_2^2$$

- **Physical loss**: penalizes the physical violations between 3D objects and 3D room layout.

$$\mathcal{L}_{\text{PHY}} = \frac{1}{N}\sum_{j=1}^{N}\left(\text{ReLU}(\text{Max}(X_j^W) - \text{Max}(X^L)) + \text{ReLU}(\text{Min}(X^L) - \text{Min}(X_j^W))\right)$$

Qualitative Results

Figure 4: Qualitative results on SUN RGB-D dataset.

Ablative Study

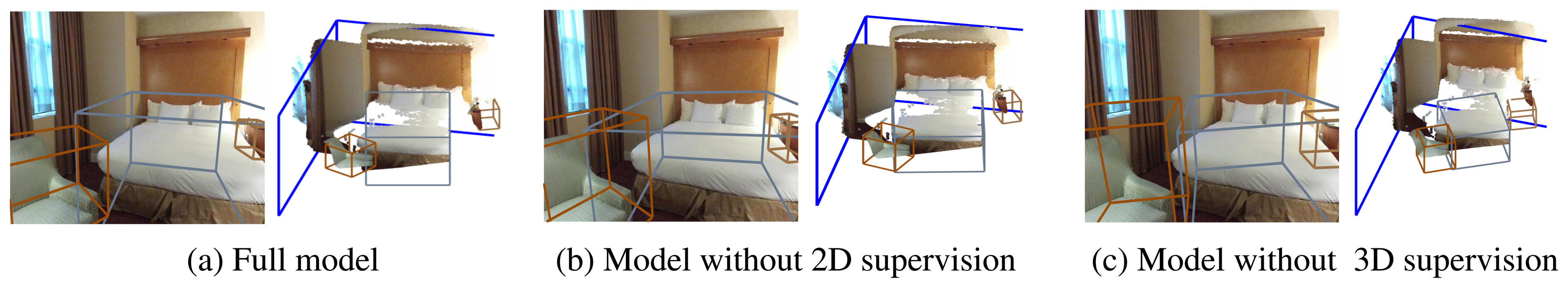

Figure 5: Comparison with two variants of our model.

Quantitative Results

Table 1: Comparison of 3D room layout estimation and holistic scene understanding on SUN RGB-D.

Method	3D Layout Estimation	Holistic Scene Understanding			
	IoU	P_g	R_g	R_r	IoU
3DGP [Choi et al., 2013]	19.2	2.1	0.7	0.6	13.9
HoPR [Huang et al., 2018]	54.9	37.7	23.0	18.3	40.7
Ours (individual)	55.4	36.8	22.4	20.1	39.6
Ours (cooperative)	**56.9**	**49.3**	**29.7**	**28.5**	**42.9**

Table 2: Comparisons of 3D object detection on SUN RGB-D.

Method	bed	chair	sofa	table	desk	toilet	bin	sink	shelf	lamp	mAP
Choi et al. [2013]	5.62	2.31	3.24	1.23	-	-	-	-	-	-	-
Huang et al. [2018]	58.29	13.56	28.37	12.12	4.79	16.50	0.63	2.18	1.29	2.41	14.01
Ours (individual)	53.08	7.7	27.04	22.80	5.51	28.07	0.54	5.08	2.58	0.01	15.24
Ours (cooperative)	**63.58**	**17.12**	**41.22**	**26.21**	**9.55**	**58.55**	**10.19**	**5.34**	**3.01**	**1.75**	**23.65**