Holistic 3D Scene Parsing and Reconstruction from a single RGB Image

Siyuan Huang[1,2] Siyuan Qi[1,2] Yixin Zhu[1,2] Yinxue Xiao[1] Yuanlu Xu[1,2] Song-Chun Zhu[1,2]

[1]University of California, Los Angeles

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

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Task

Given a single RGB Image as input, we aim to parse the 3D structure and reconstruct the 3D scene composed by a set of **CAD** models.

Framework

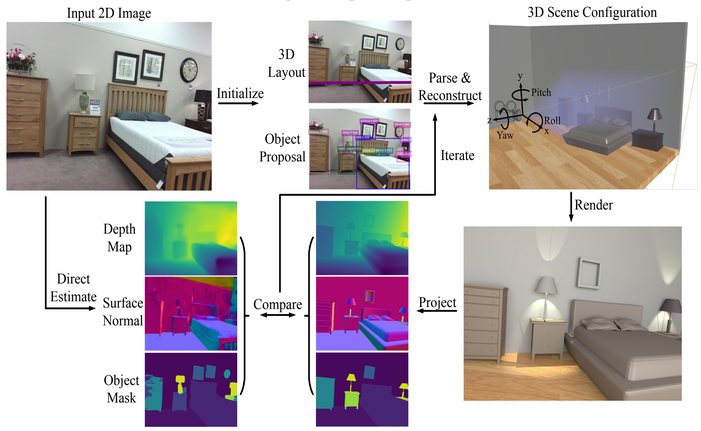
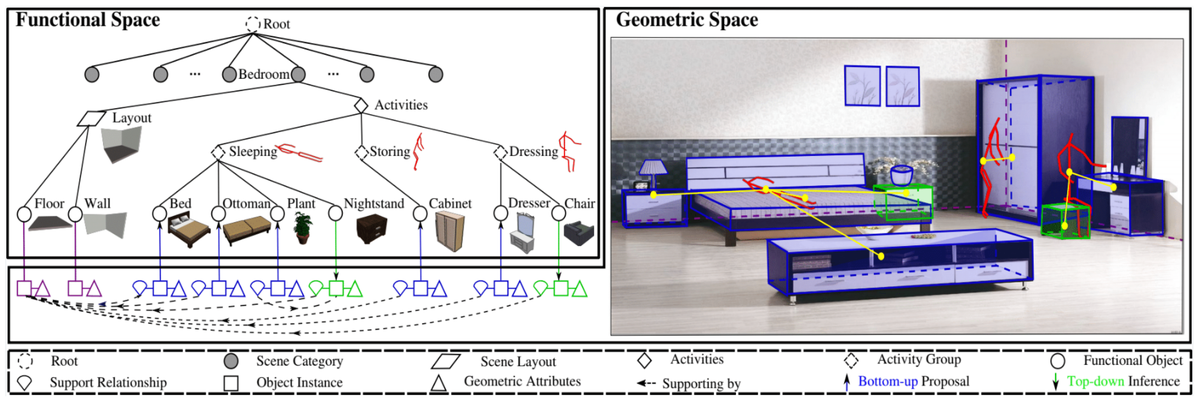


Illustration of the proposed holistic 3D indoor scene parsing and reconstruction in an **analysis-by synthesis fashion**. A 3D configuration is initialized by individual vision modules (e.g. object detection, 2D layout estimation). A **joint inference** algorithm compares the differences between the rendered **normal, depth, and segmentation map** with the ones estimated directly from the input RGB image, and adjust the 3D structure iteratively.

Contribution

- We integrate geometry and physics to interpret and reconstruct indoor scenes with CAD models. We **jointly optimize 3D room layouts and object configurations**, improving the performance of scene parsing and reconstruction on SUN RGB-D dataset.
- We incorporate **hidden human context** into our model, enabling to imagine latent human pose in each activity group by grouping and sampling. In this way, we can optimize the joint distribution of both visible and invisible components of the scene.
- We propose a **complete computational framework** to combine generative model, discriminative models and graphics engines in scene parsing and reconstruction.
- We model the **supporting relations** among objects, eliminating the widely adopted assumption that all objects must stand on the ground. Such flexibility provides better parsing and reconstruction of the real-world scenes with complex object relations.

Representation



In our representation, the functional space (pg_f) characterizes the **hierarchical structure** and the geometric space (pg_g) encodes the spatial entities with **contextual relations**.

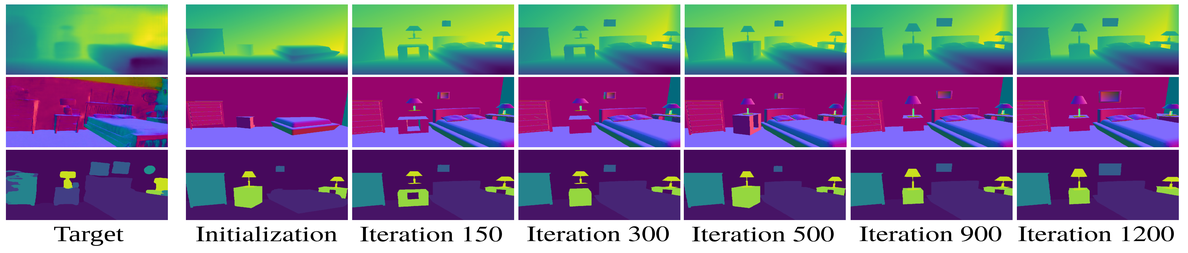
$$p(pg|I) \propto p(pg_f) \cdot p(pg_g|pg_f) \cdot p(I|pg_g) = \frac{1}{Z}\exp\{-\mathcal{E}(pg_f) - \mathcal{E}(pg_g|pg_f) - \mathcal{E}(I|pg_g)\}$$

$\mathcal{E}(I|pg_g)$ characterizes the similarity between the observed image and the rendered image.

Inference

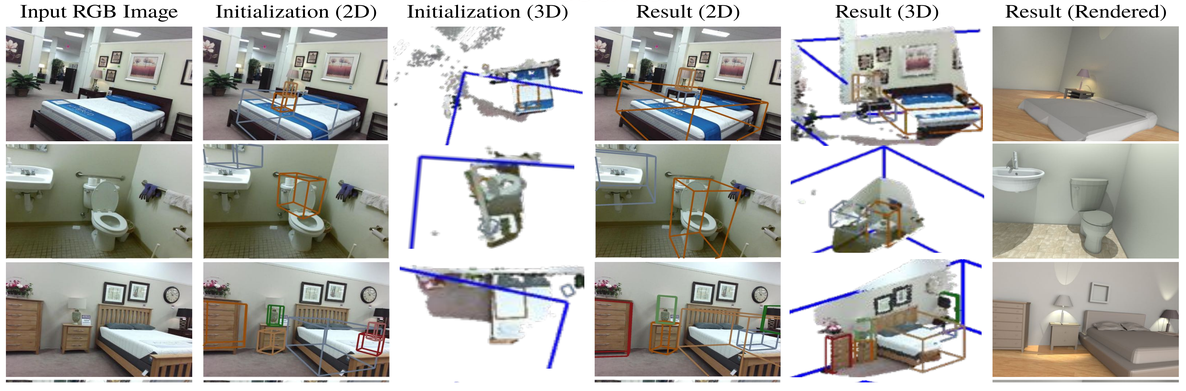
Bottom-Up Initialization

Joint Inference

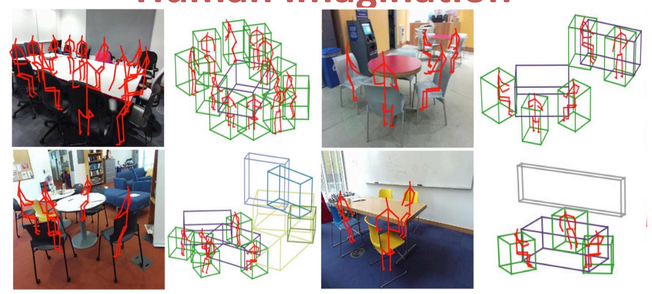


We optimize the objects, layout and hidden human context by maximizing the posterior probability of parse graph through Markov chain Monte Carlo sampling.

Results



Human Imagination



Evaluation

Method	# of image	3D Layout Estimation IoU	Holistic Scene Understanding P_g	R_g	R_r	IoU
3DGP	5050	19.2	2.1	0.7	0.6	13.9
Ours (init.)	5050	46.7	25.9	15.5	12.2	36.6
Ours (joint.)	5050	**54.9**	**37.7**	**23.0**	**18.3**	**40.7**
3DGP	749	33.4	5.3	2.7	2.1	34.2
IM2CAD	484	62.6	-	-	-	49.0
Ours (init.)	749	61.2	29.7	17.3	14.4	47.1
Ours (joint.)	749	**66.4**	**40.5**	**26.8**	**21.7**	**52.1**

More Results