
CVPR

Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image

https://yinyunie.github.io/Total3D/

Yinyu Nie[1,2,3], Xiaoguang Han[2,3,*], Shihui Guo[4]

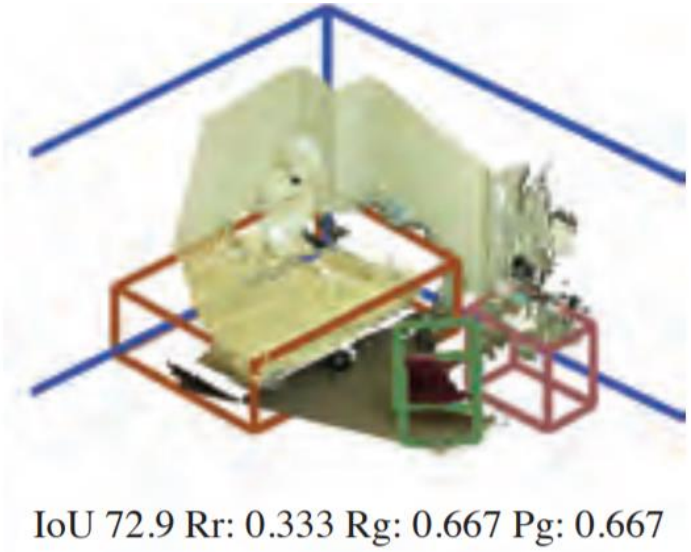
Yujian Zheng[2,3], Jian Chang[1], Jian J Zhang[1]

[1]Bournemouth University [2]The Chinese University of Hong Kong, Shenzhen

[3]Shenzhen Research Institute of Big Data [4]Xiamen University

Milestones (3D scenes)

S, Song , SUN-RGBD, CVPR'15



IoU 72.9 Rr: 0.333 Rg: 0.667 Pg: 0.667

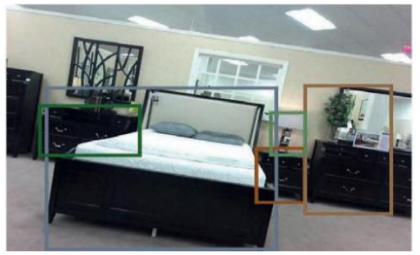
IoU: 77.0 Rr: 0.25 Rg: 0.25 Pg: 0.5

Holistic Scene Understanding Benchmark

Milestones (3D scenes)

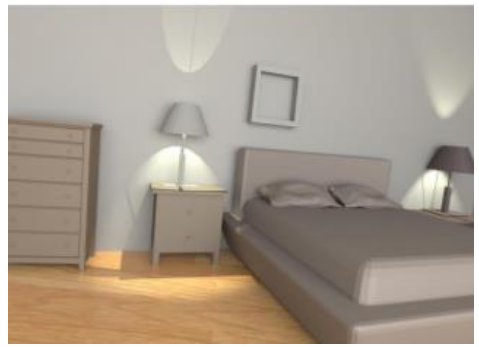
1963 | 1999 | 2009 | 2015 - Now → **Timeline**

CooP, S. Huang, NIPS'18



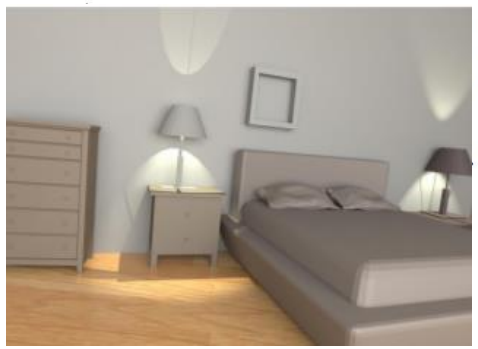

CooP, S. Huang, NIPS'19


IM2CAD, H. Izadinia, CVPR'17

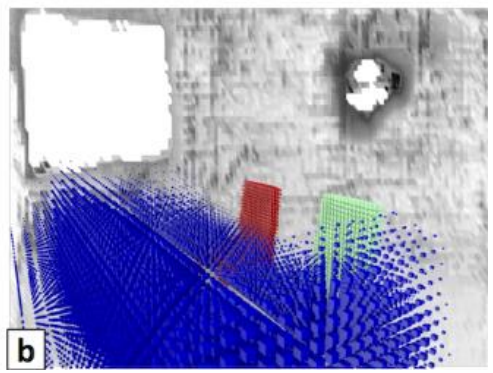



HSG, S. Huang, ECCV'18

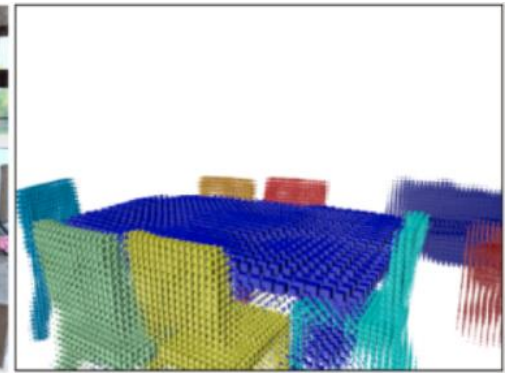

IM2CAD, H. Izadinia, CVPR'17

HSG, S. Huang, ECCV'18



Factored 3D, S. Tulsiani, CVPR'18

3D-RelNet, N. Kulkarni, ICCV'19

Thinking:

1. 3D detection has been developed for years.
2. Layout estimation has been researched for decades.
3. **Indoor object geometry** is still underdeveloped.

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Motivation: Total 3D Understanding

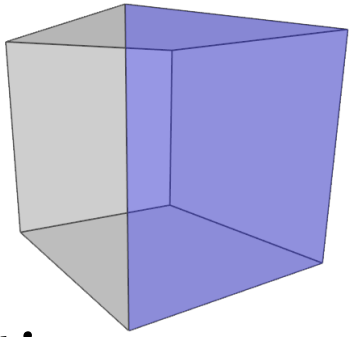
A single RGB image

Layout, Bounding boxes & Meshes

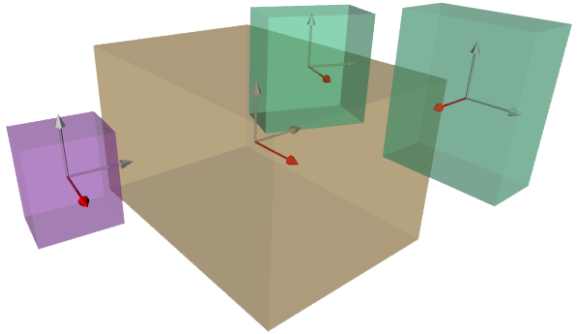
Overview

An image with
2D detections

Room layout



3D detections

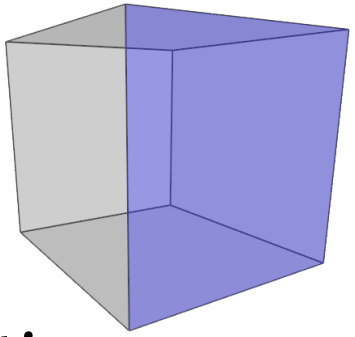


Object meshes

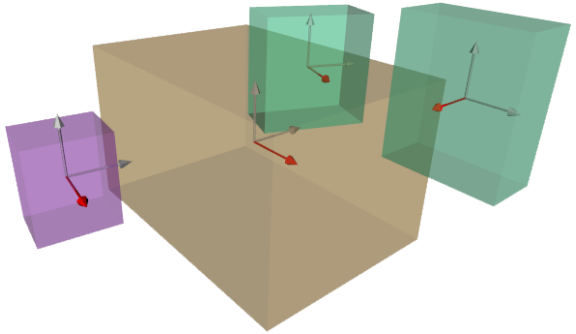
Embedding

Embedding

Overview



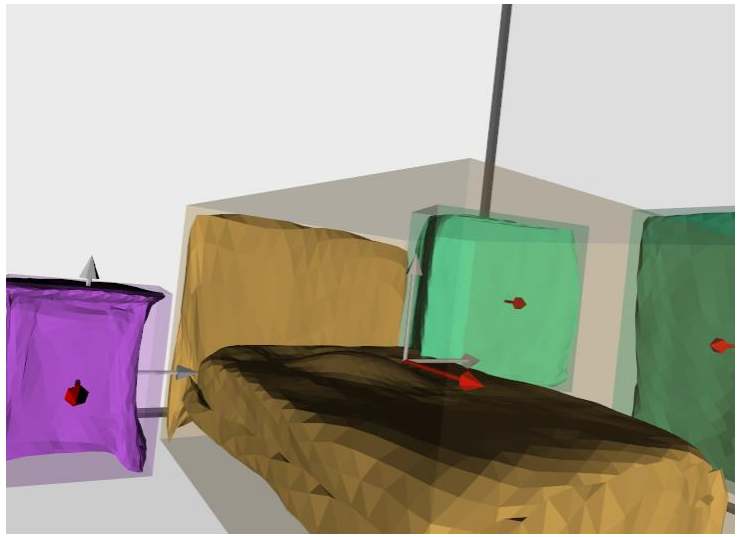
3D detections



Embedding

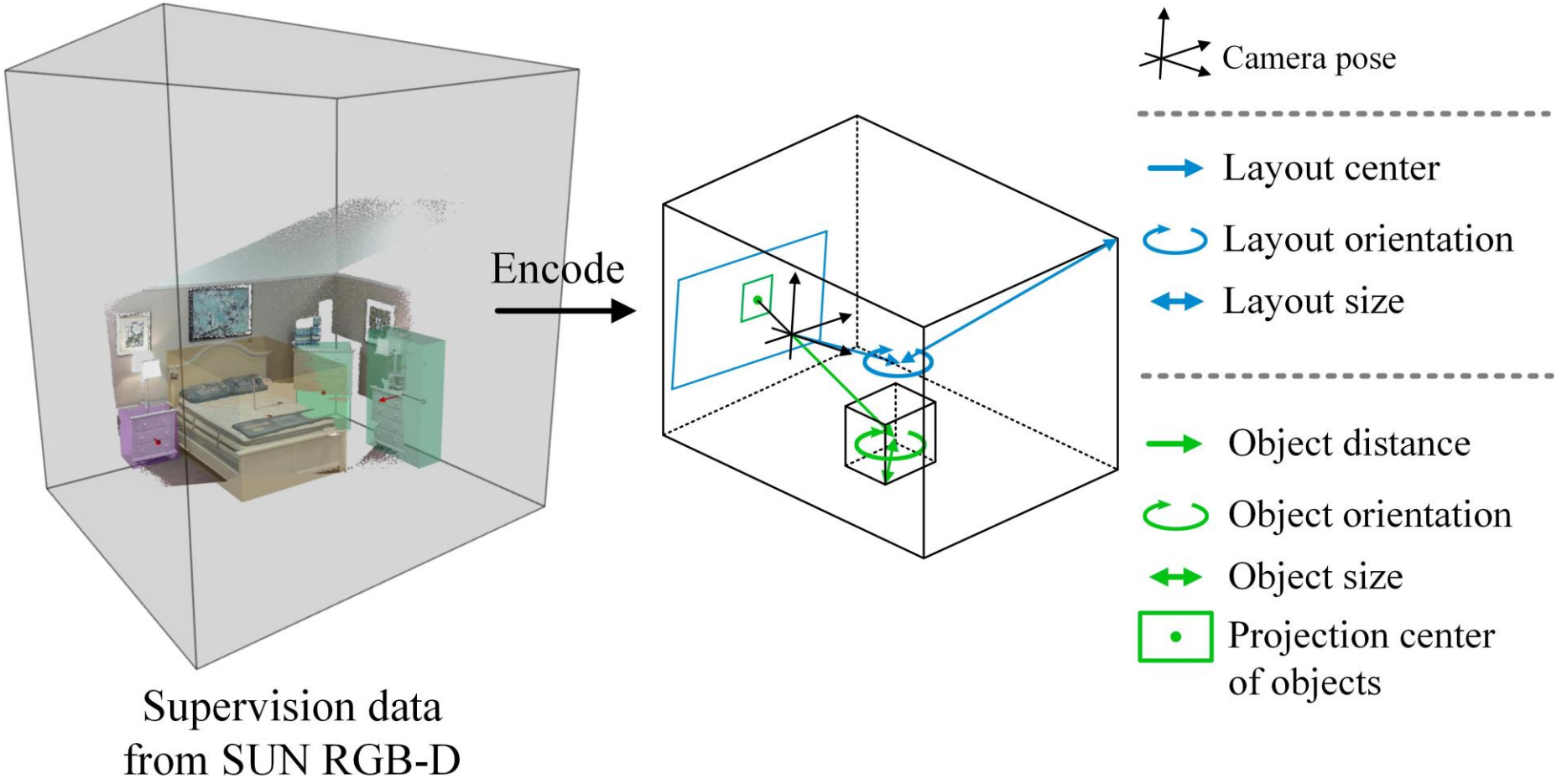
Object meshes

Embedding



Method

Target Parameterization



Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N. and Zhu, S.C., 2018. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In *Advances in Neural Information Processing Systems* (pp. 207-218).

3D detector

2D detections

ResNet

Target

Appearance feature

Attention sum

Element-wise sum

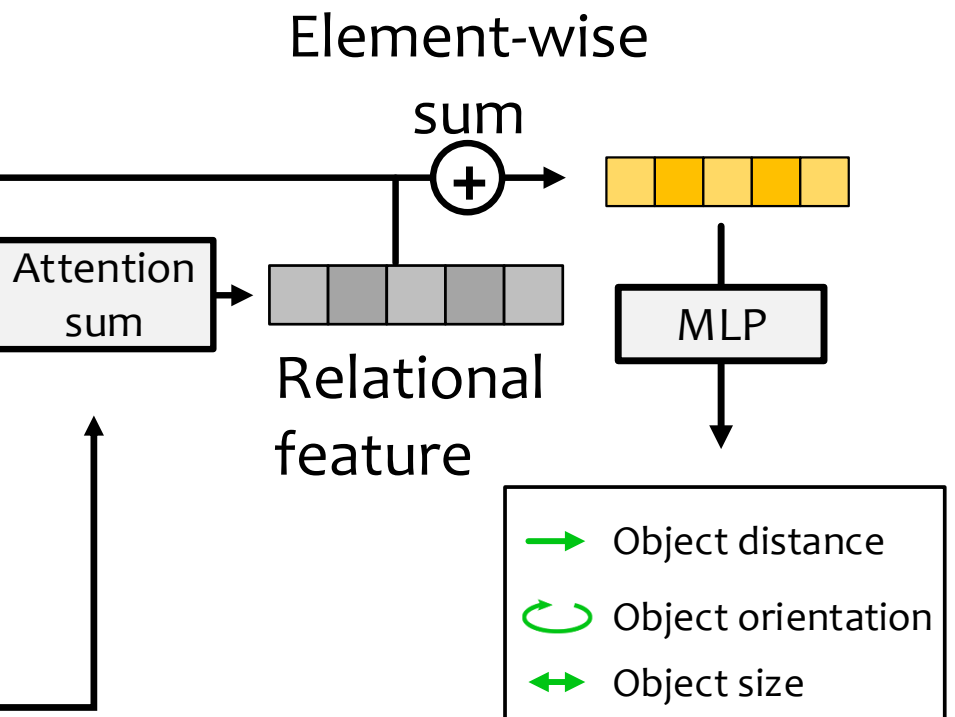
Relational feature

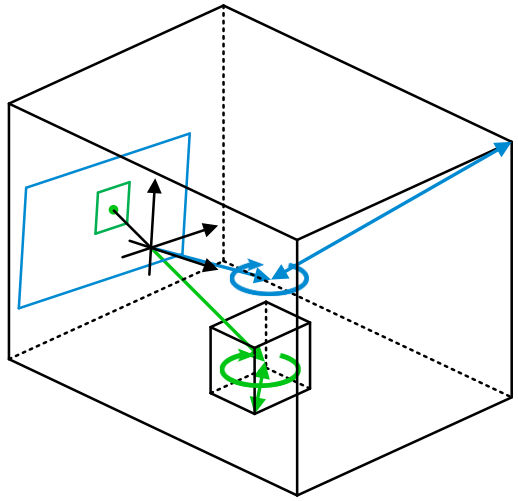
MLP

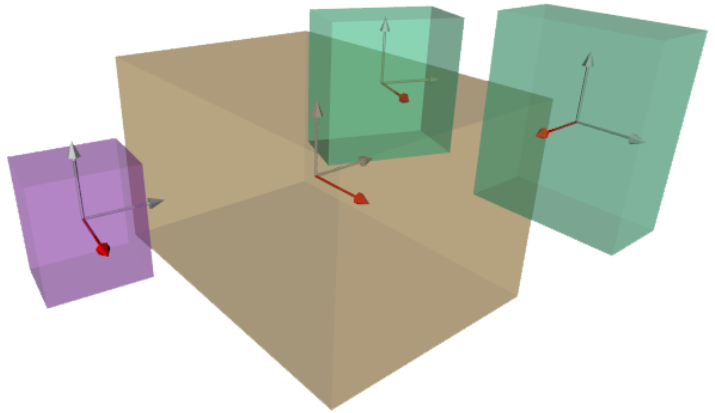
Object distance
Object orientation
Object size
Projection center of objects

Geometry feature

3D detector

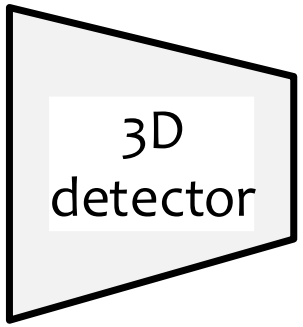

Element-wise
sum
Attention
sum
Relational
feature
MLP
Object distance
Object orientation
Object size
Projection center
of objects

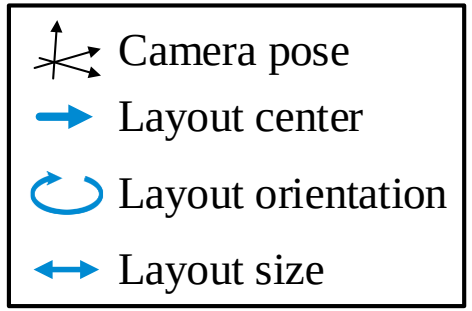


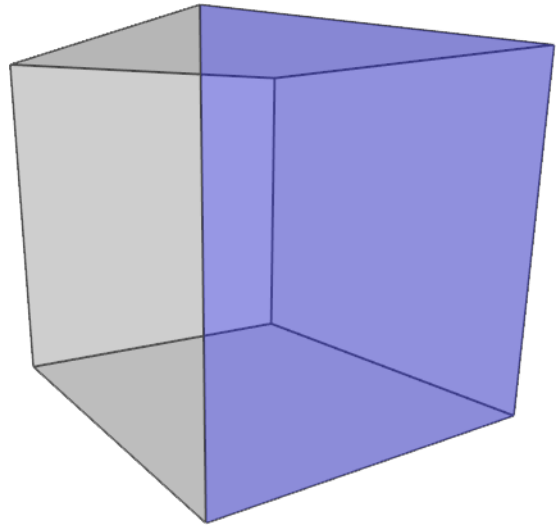


Layout estimation

Source image







Room layout

Mesh generation & modification

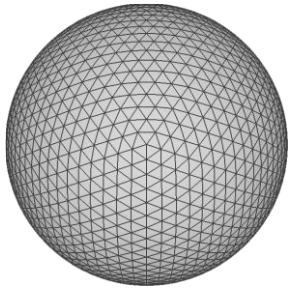
Input image

ResNet

Appearance feature

Category code

Cat

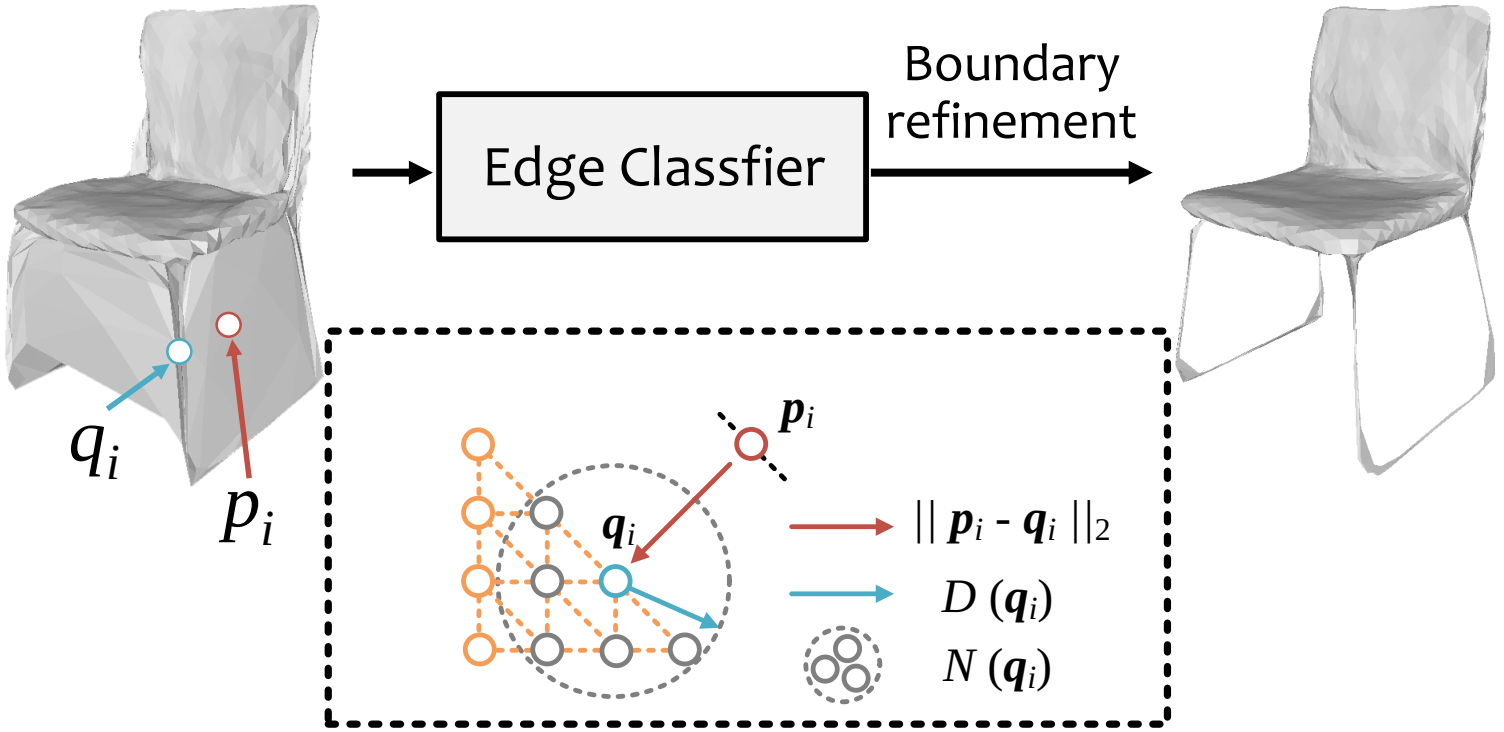


Template sphere

AtlasNet

Edge Classfier

Boundary refinement



$$f(\boldsymbol{p}_i) = \begin{cases} \text{False} & \|\boldsymbol{p}_i - \boldsymbol{q}_i\|_2 > D(\boldsymbol{q}_i) \\ \text{True} & \text{otherwise} \end{cases}, \qquad (2)$$

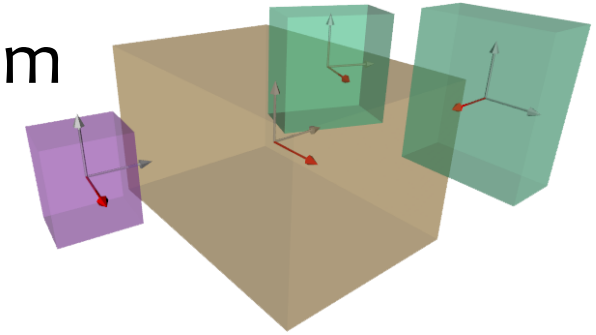
$$D(\boldsymbol{q}_i) = \max_{\boldsymbol{q}_m, \boldsymbol{q}_n \in N(\boldsymbol{q}_i)} \min \|\boldsymbol{q}_m - \boldsymbol{q}_n\|_2, m \neq n$$

Joint training & inference

Canonical system

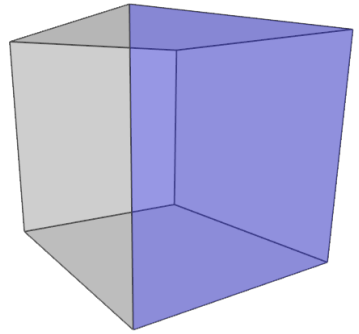
Object meshes

Camera system

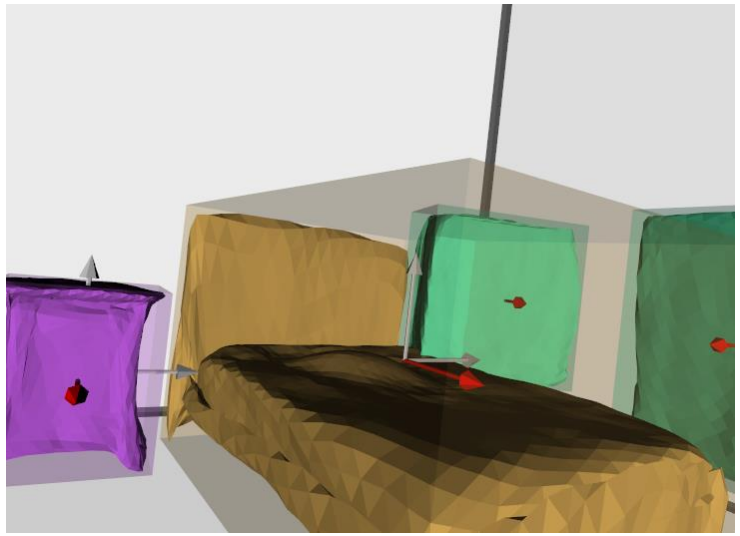


3D detections

World system



Room layout



Results

Our Results on Pix3D (single objects)

Input | Mesh R-CNN | AtlasNet-sphere | TMNet (t=0.1) | TMNet (t=0.05) | ours

Our Results on Pix3D (single objects)













Input

Mesh R-CNN

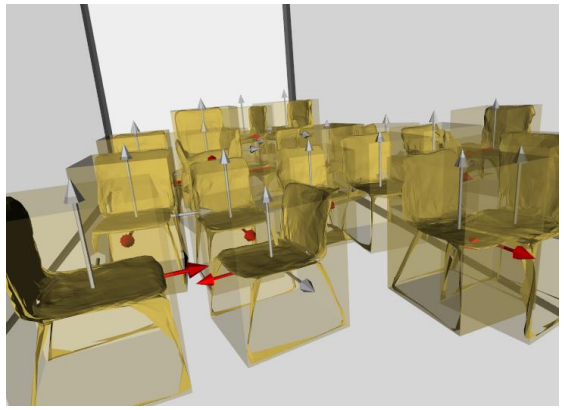
AtlasNet-sphere

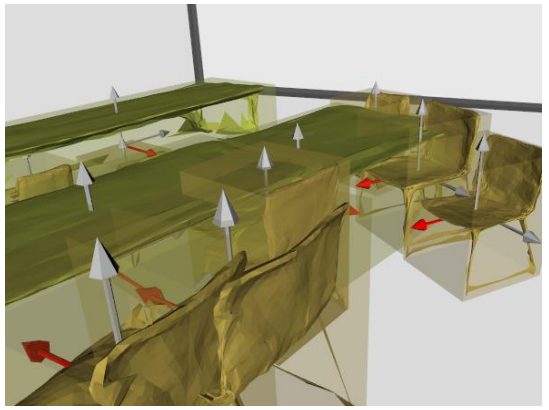
TMNet (t=0.1)

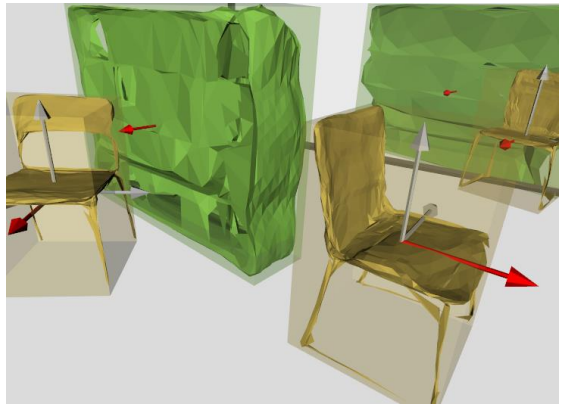
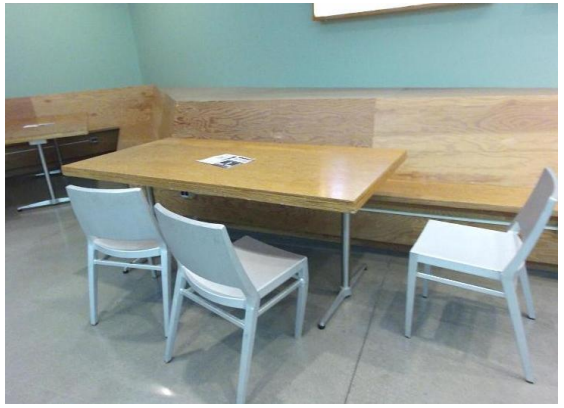
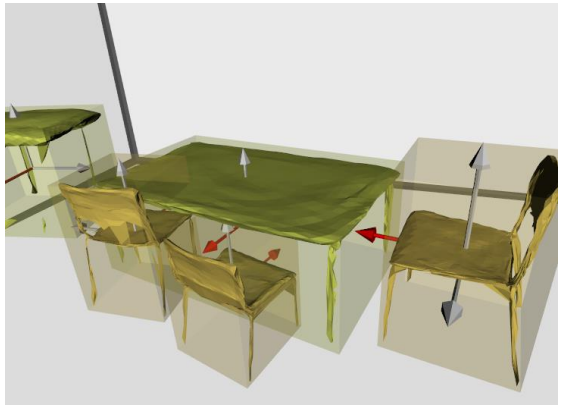
TMNet (t=0.05)

ours

Our Results on SUN-RGBD (scenes)

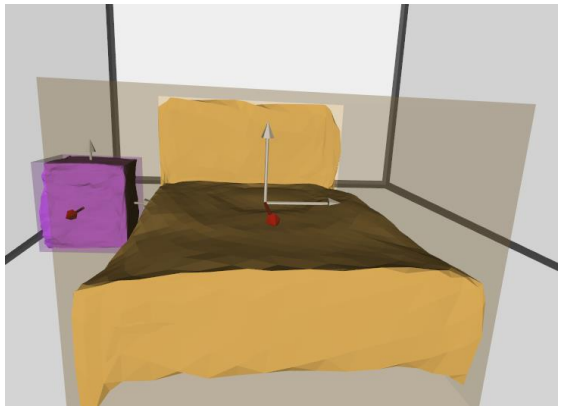



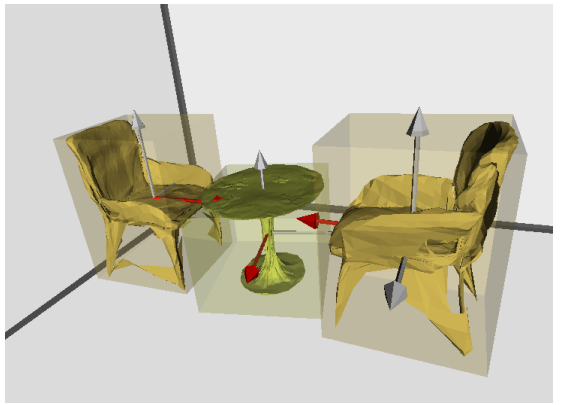


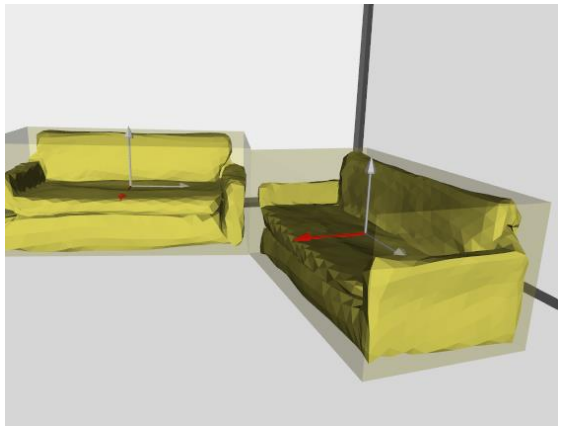




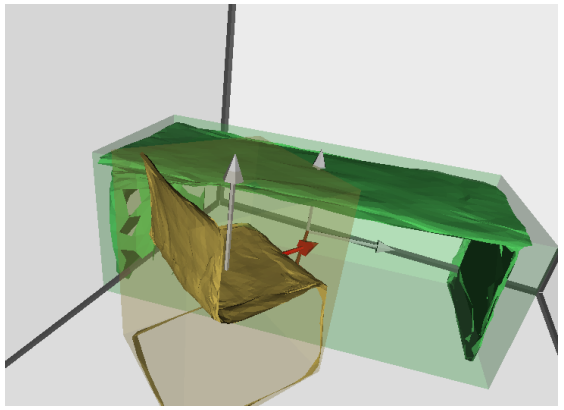
Input 3d scene Input 3d scene

Our Results on SUN-RGBD (scenes)







Input | 3d scene | Input | 3d scene

Evaluations

Layout estimation (on SUN RGB-D)

Method	3D IoU
3DGP [Choi et al. CVPR'2013]	19.2
HoPR [Huang et al. ECCV'2018]	54.9
CooP [Huang et al. NeurIPS 2018]	56.9
Ours (w/o. joint)	57.6
Ours (w. joint)	**59.2**

3D detection (on SUN RGB-D)

Method	mAP
HoPR [Huang et al. ECCV'2018]	14.47
CooP* [Huang et al. NeurIPS 2018]	17.80
CooP** [Huang et al. NeurIPS 2018]	21.77
Ours (w/o. joint)	23.32
Ours (w. joint)	**26.38**

Evaluations

Object pose (on NYU v2)

Method	Translation (Err≤0.5m) %	Rotation (Err≤30°) %	Scale (Err≤0.2)%
Tulsiani et al. CVPR'2018	51.0	63.8	18.9
Ours (w/o. joint)	49.2	64.1	42.1
Ours (w. joint)	**51.8**	**66.5**	**43.7**

Object mesh (on Pix3D)

Method	Chamfer distance
AtlasNet [Groueix et al. CVPR'2018]	12.26
TMN [Pan et al. ICCV'2019]	9.03
Ours	**8.36**

Effects of joint learning

Version	Layout (IoU) (higher is better)	3D detection (mAP) (higher is better)	Scene mesh (L_g) (lower is better)
Baseline (w/o. joint)	57.63	20.19	2.10
Baseline + relation feature	57.63	23.32	1.89
Baseline + joint losses	58.87	25.62	1.52
Baseline + relation feature + joint losses (full version)	**59.25**	**26.38**	**1.43**

Summary

- A solution to end-to-end reconstruct room layout, object bounding boxes, and meshes from a single image.
- This joint learning shows the complementary role of each component and reaches the state-of-the-art on each task.
- A novel topology modifier for object mesh generation. It prunes mesh edges to approximate the target shape by progressively modifying mesh topology.

Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image

Thanks for watching !

https://yinyunie.github.io/Total3D/

- Yinyu Nie[1,2,3], Xiaoguang Han[2,3,*], Shihui Guo[4]
- Yujian Zheng[2,3], Jian Chang[1], Jian J Zhang[1]

[1]Bournemouth University [2]The Chinese University of Hong Kong, Shenzhen

[3]Shenzhen Research Institute of Big Data [4]Xiamen University