3D Spatial Recognition without Spatially Labeled 3D

Jason Ren[1, 2], Ishan Misra[1],

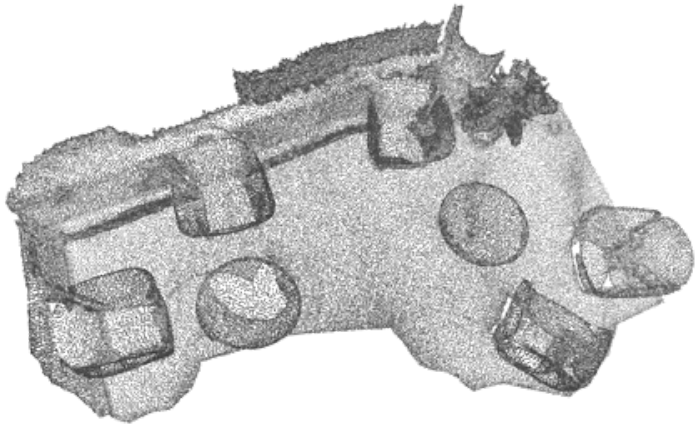
Alex Schwing[2] , Rohit Girdhar[1]

[1] FAIR [2] UIUC

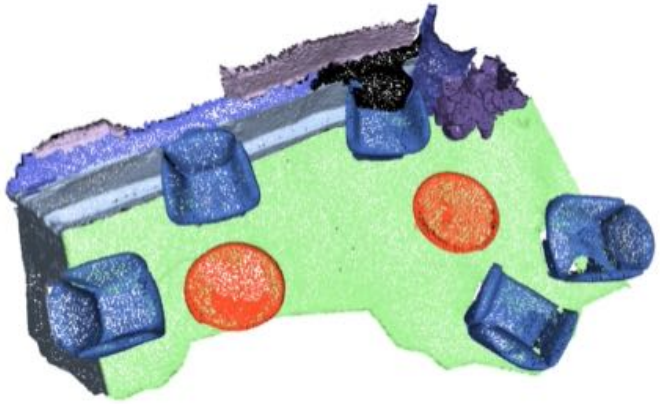
FACEBOOK

Motivating example -- ScanNet

- Collecting 3D scans is easy: an iPad is all you need
- Labeling strong labels: ~22.3 min/scan



Point cloud



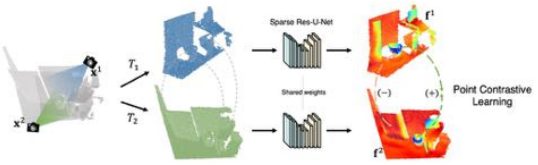
Strong labels

(* ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. Dai et al., CVPR 2017)

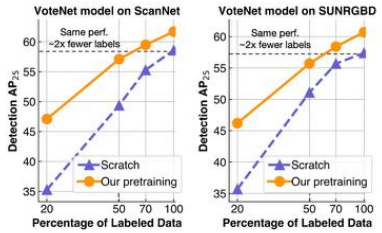
Related work

Weak → Strong

No supervision

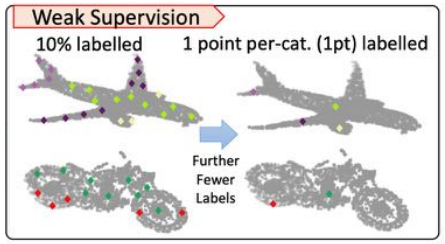


PointContrast
Xie et al., 2020 [FAIR]

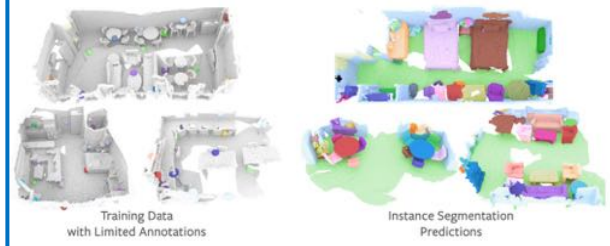


DepthContrast
Zhang et al., 2020 [FAIR]

Sparse point label

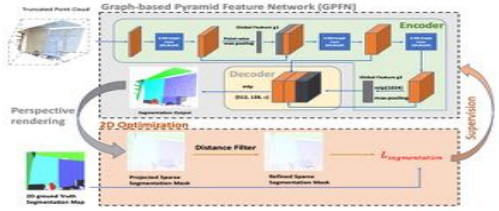


Xu and Lee, 2020



Hou et al., 2020 [FAIR]

2D instance label



Wang el al., 2019
(2D segmentation)

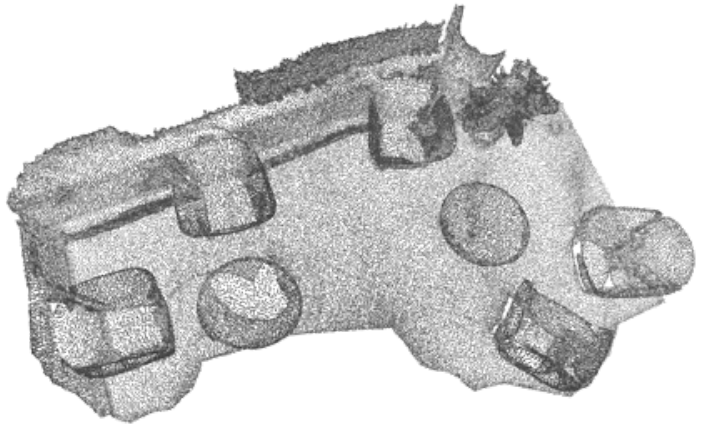
Tang and Lee, 2019
(2D bounding box)

●●●

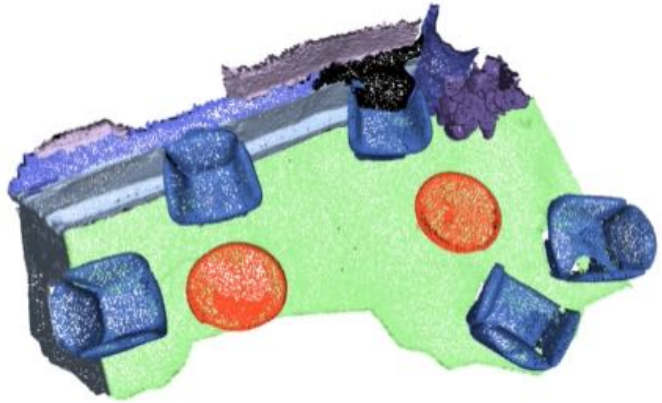
class tags

Weak label: class tags

- Collecting 3D scans is easy: an iPad is all you need
- Labeling strong labels: ~22.3 min/scan
- Labeling weak labels: ~15 sec/scan (~90× faster)



Point cloud

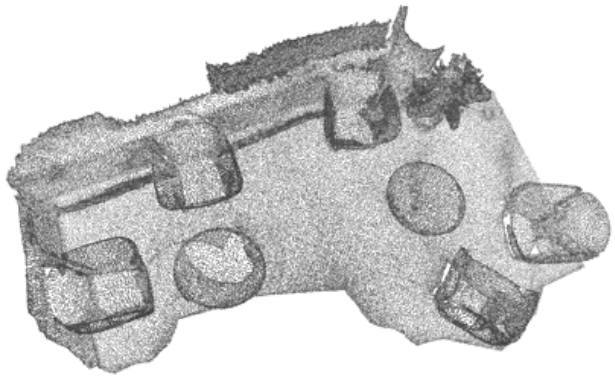


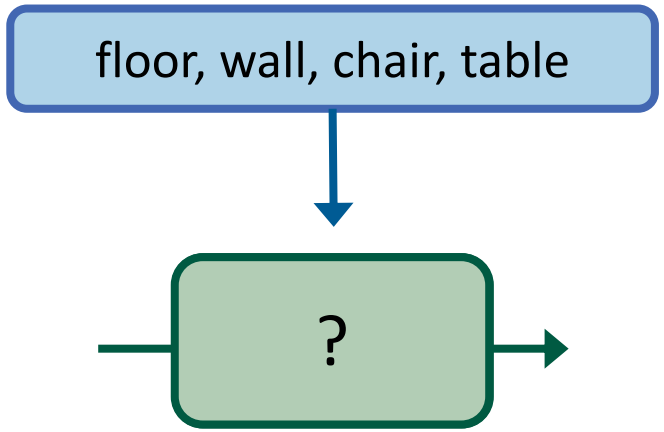
Strong labels

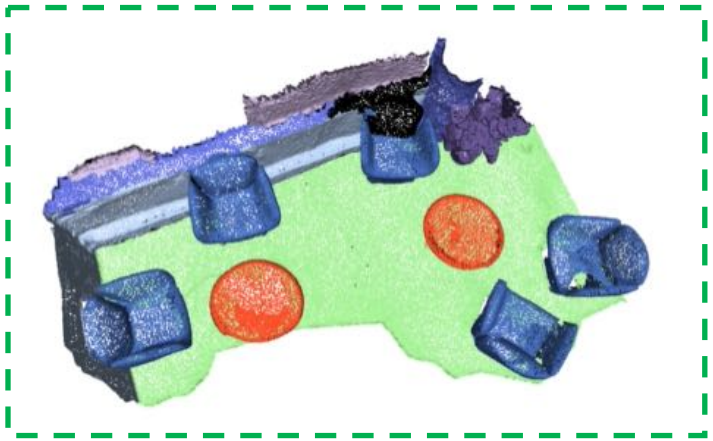
floor, wall,
chair, table…

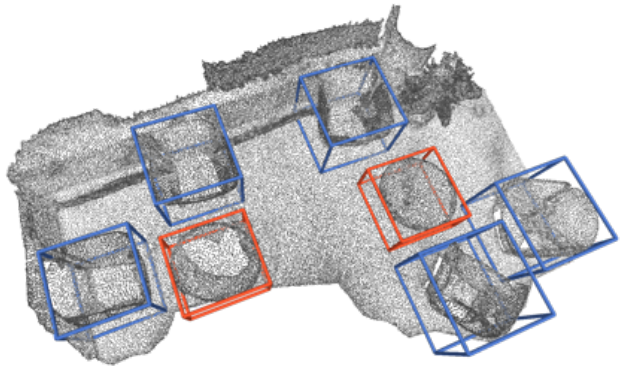
Weak labels

Goal: Spatial Recognition









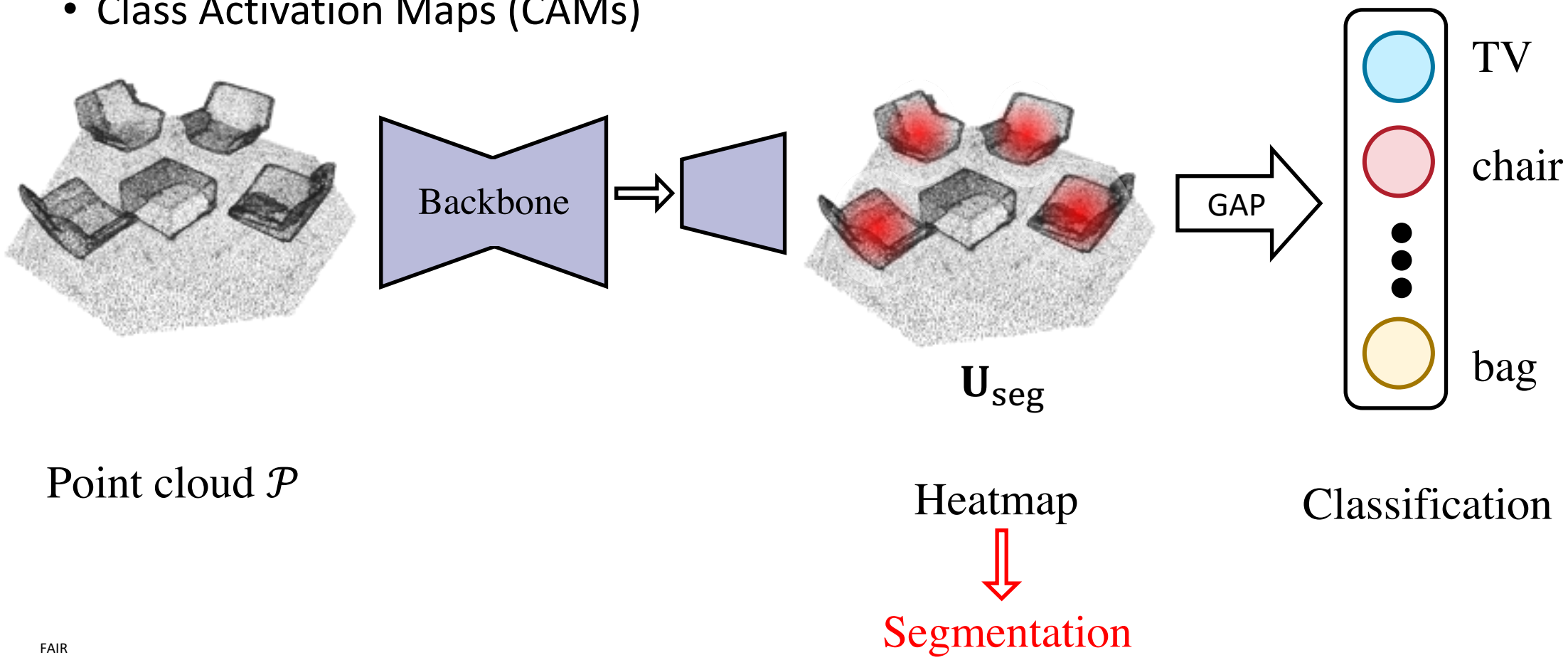
Input: point cloud

Goal: localizing each object

The “what” problem: segmentation

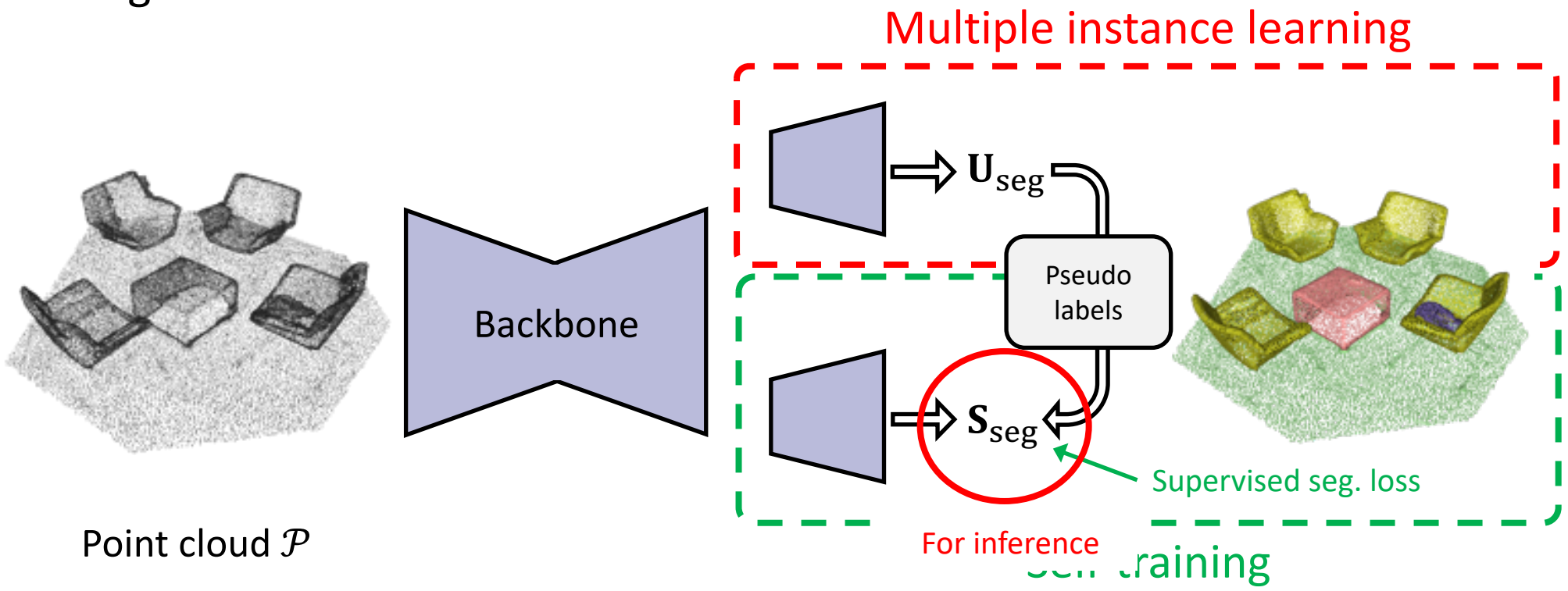
Naïve solution: Multiple Instance Learning (MIL)

- Class Activation Maps (CAMs)



The “what” problem: segmentation

Self-training

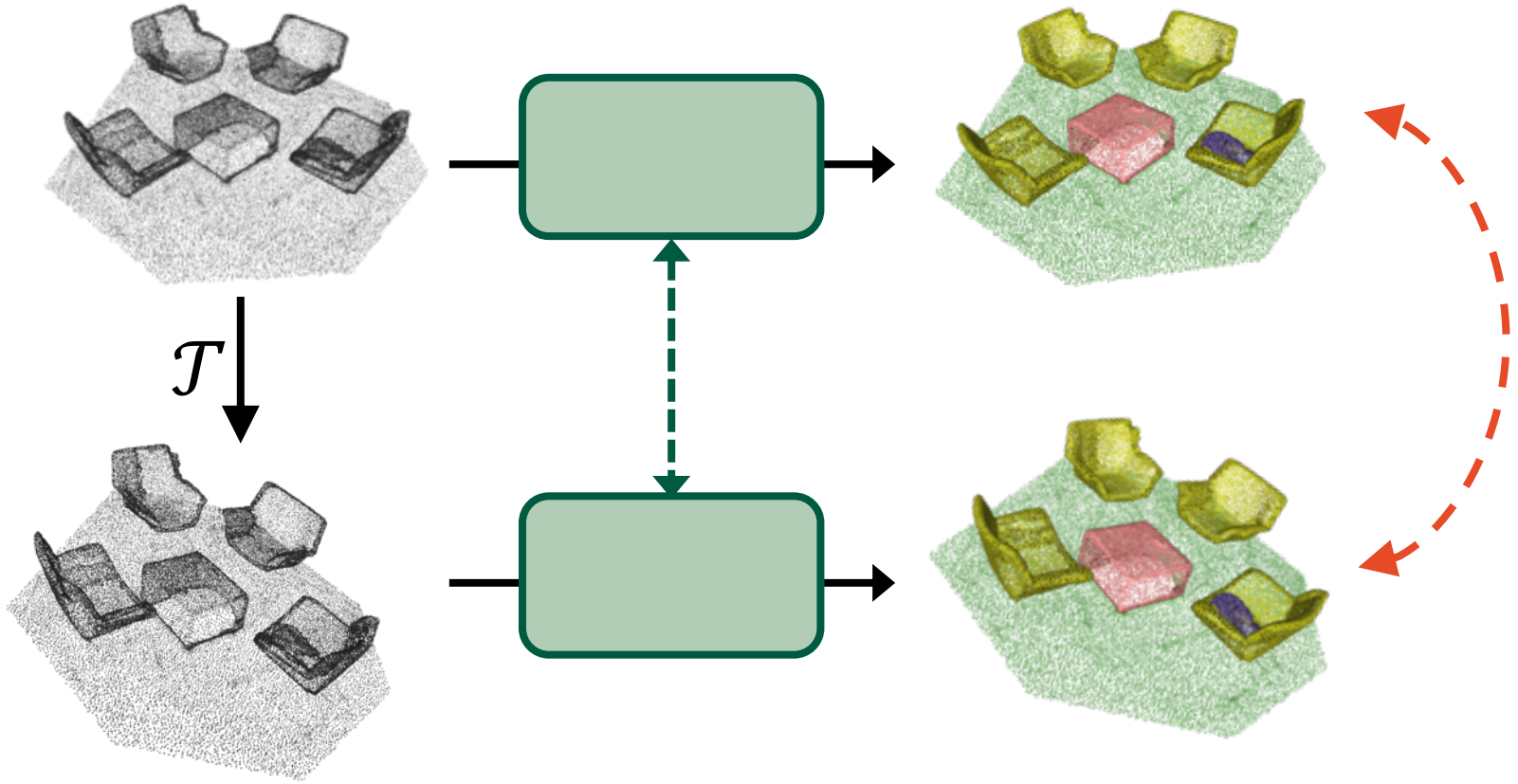


[1] Zoph et al., Rethinking Pre-training and Self-training, 2020

[2] Wei et al., Object region mining with adversarial erasing: a simple classification to semantic segmentation approach, 2017

Cross-transformation consistency

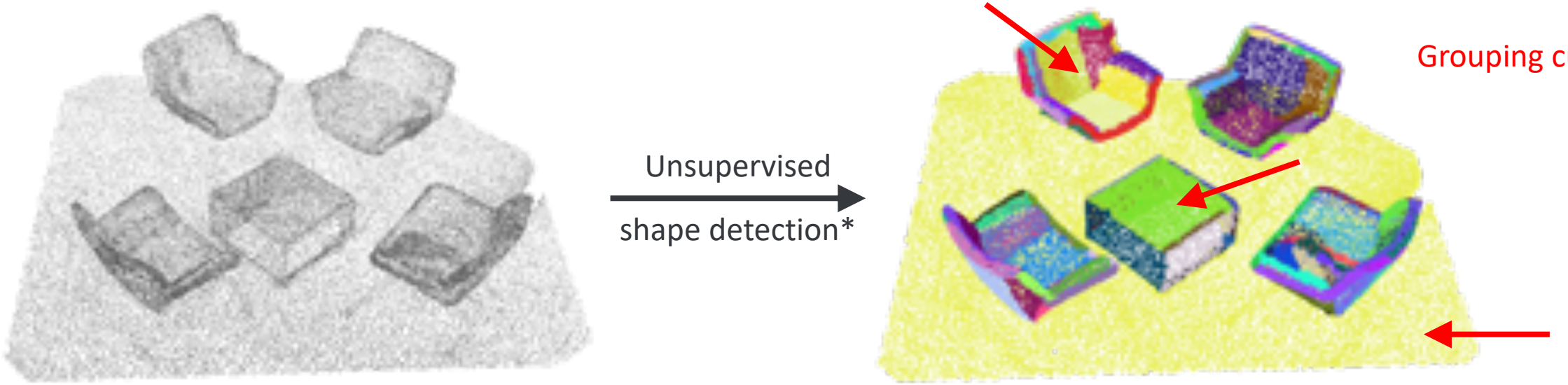
Standard technique used in Semi-/Self-supervised learning



Local smoothness

Unsupervised Shape Detection:

- Encourage segmentation to be consistent within shapes



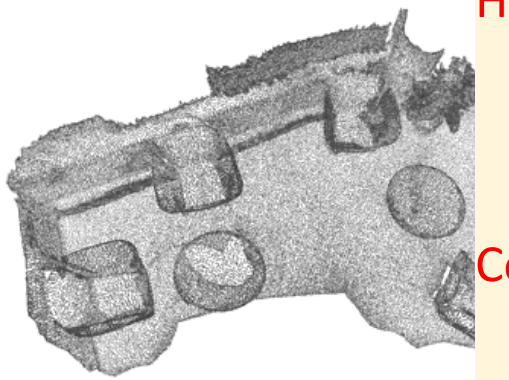
Point cloud

Detected shapes

🚨🚨Spoiler alert: detected shapes will be re-used later!

* Region growing algorithm: https://cgal.geometryfactory.com/CGAL/doc/master/Shape_detection/index.html

Goal: Spatial Recognition



How to predict bbox without bbox?

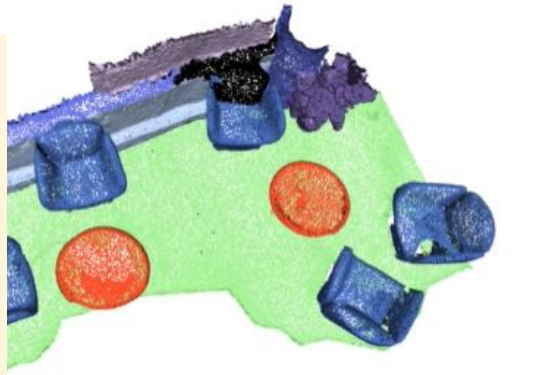
✓ No issues if we have proposals!

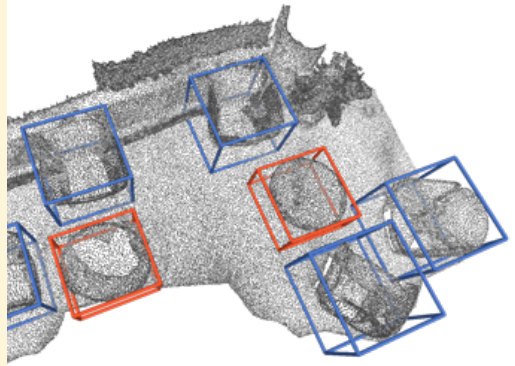
Compute proposals using weak labels?

✓ No issues!

✓ Unsupervised is also fine!

Geometric Selective Search (GSS)





Input: point cloud

Goal: localizing each object

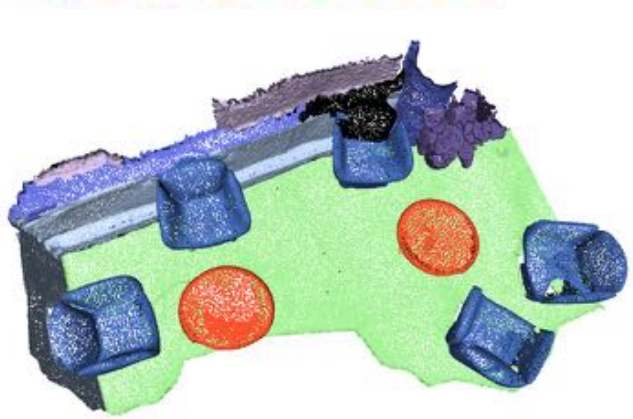
Recap: Selective Search

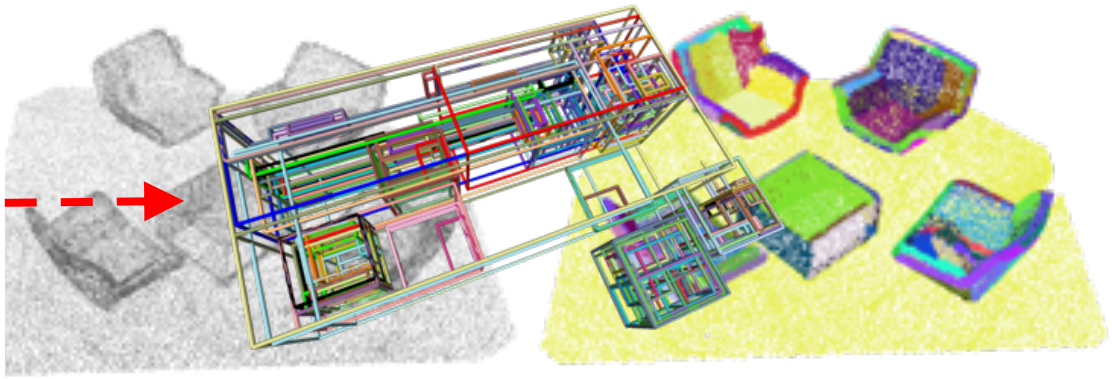
Tl; dr: grouping super-pixels using low-level cues (color, size, shape…)

GSS: Geometric Selective Search

	Input	Color	Size	Shape	Texture	Segmentation
SS	Super-pixel	✓	✓	✓	✓	
GSS	Shapes	✓	✓	✓		✓

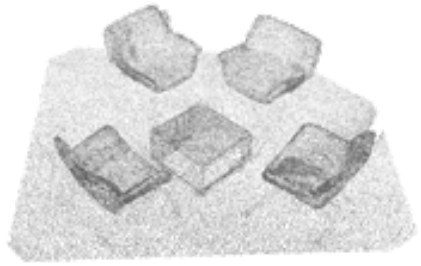
Unsupervised cues → Weakly-supervised cues



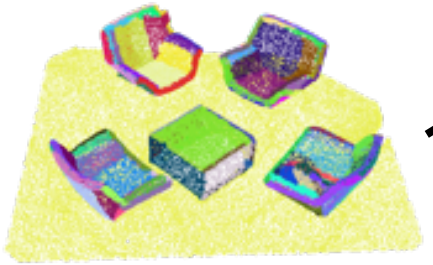


Tl; dr: grouping primitive shapes using geometric + semantic cues (size, seg…)

GSS: Geometric Selective Search

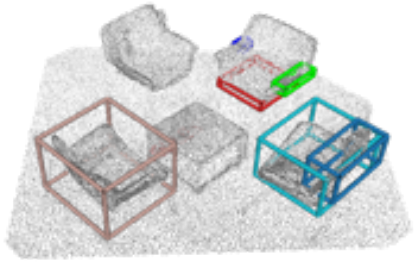


Point cloud

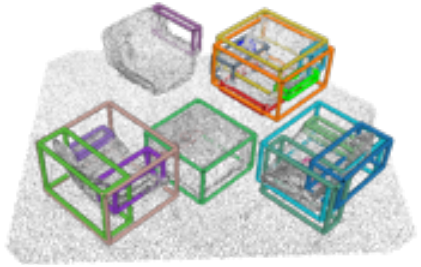


Detected shapes

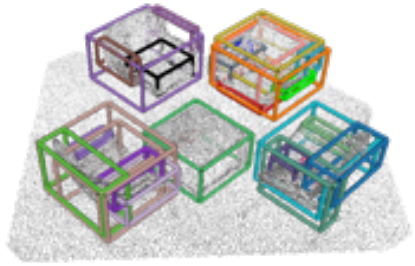
Hierarchical Agglomerative Clustering (HAC)



Iteration 20

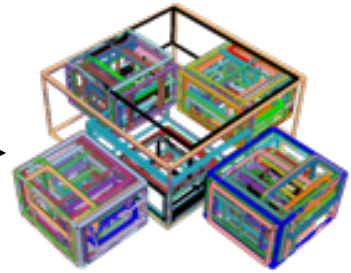


Iteration 40



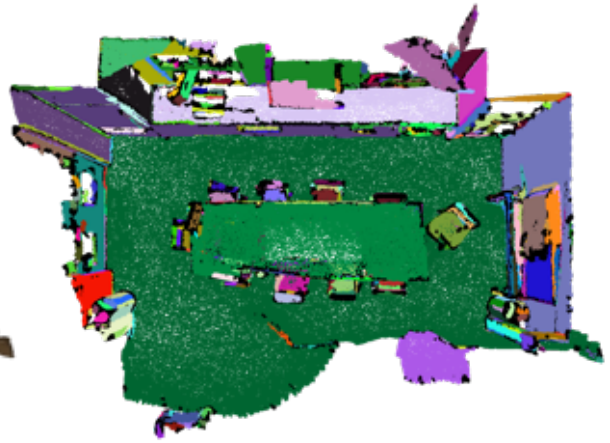
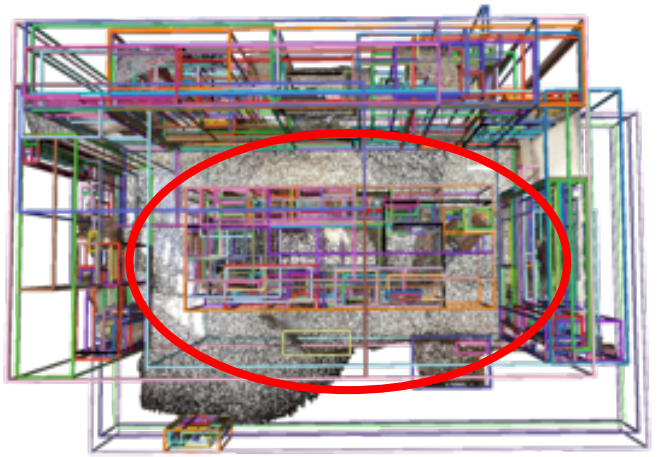
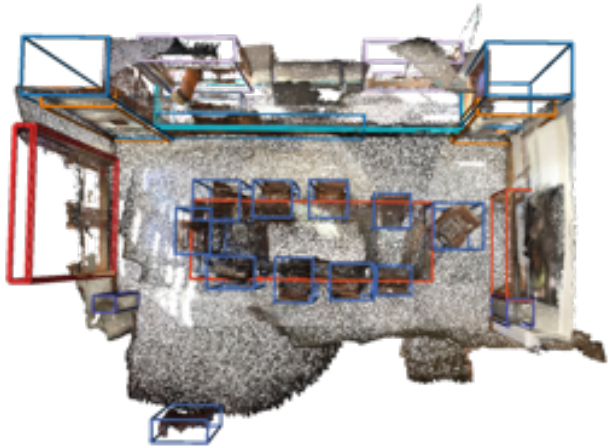
Iteration 60

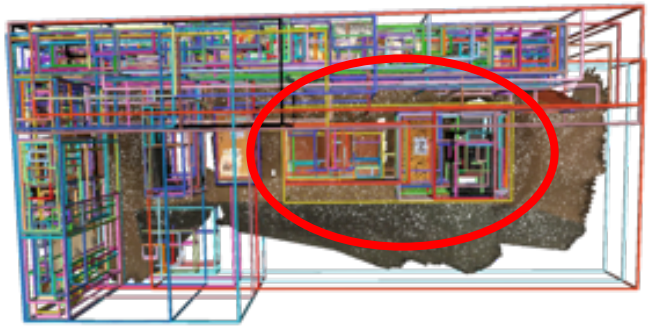
Ensemble



GSS proposals

GSS: visualization

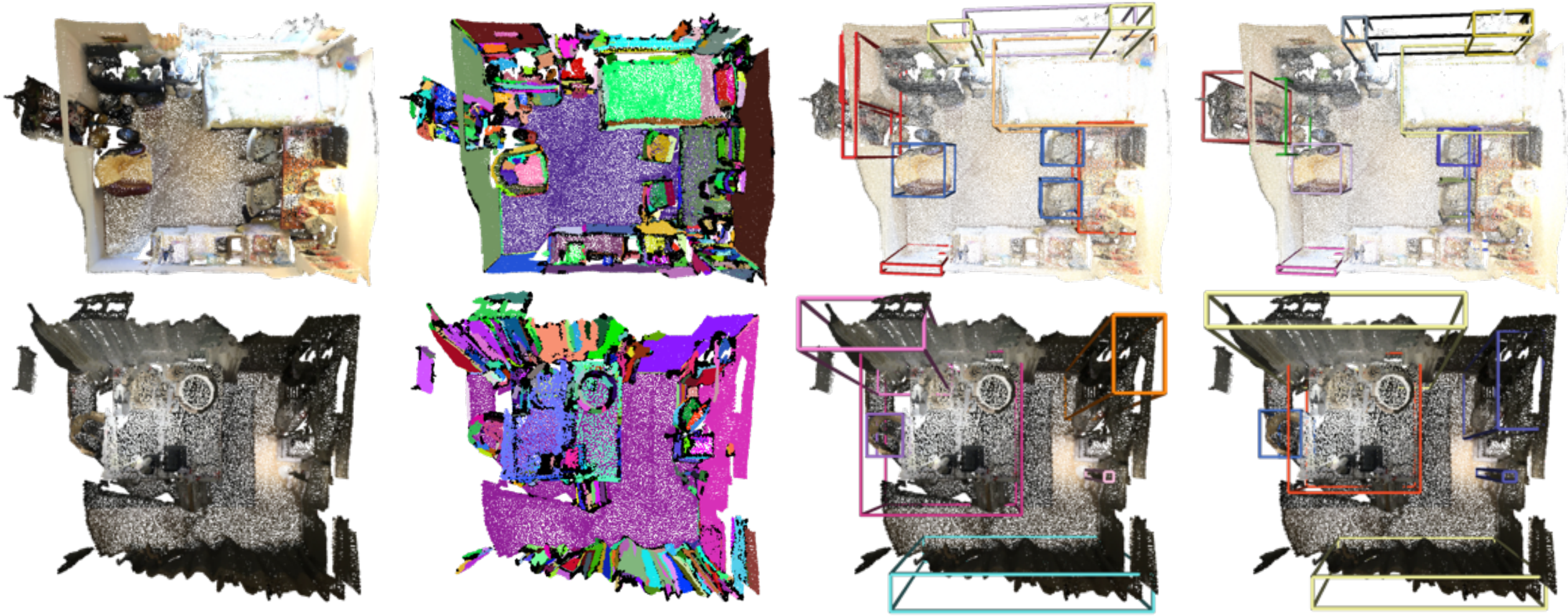

Point cloud

Detected shapes

GSS output

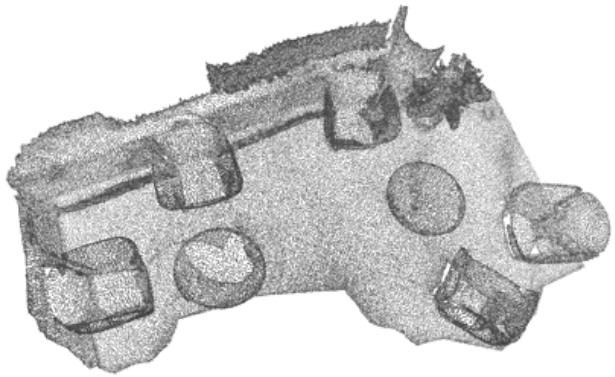
Ground-truth

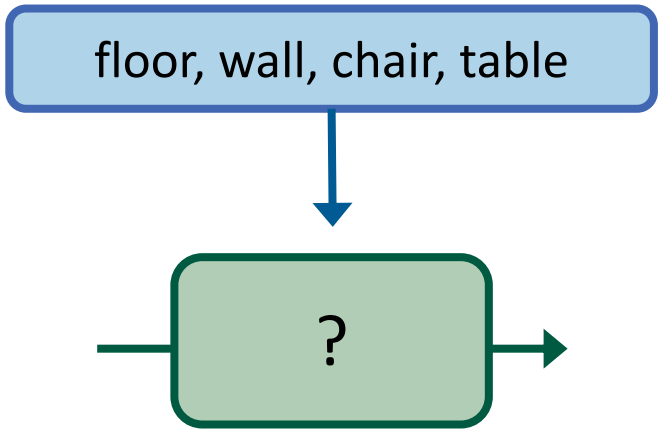
GSS: Geometric Selective Search

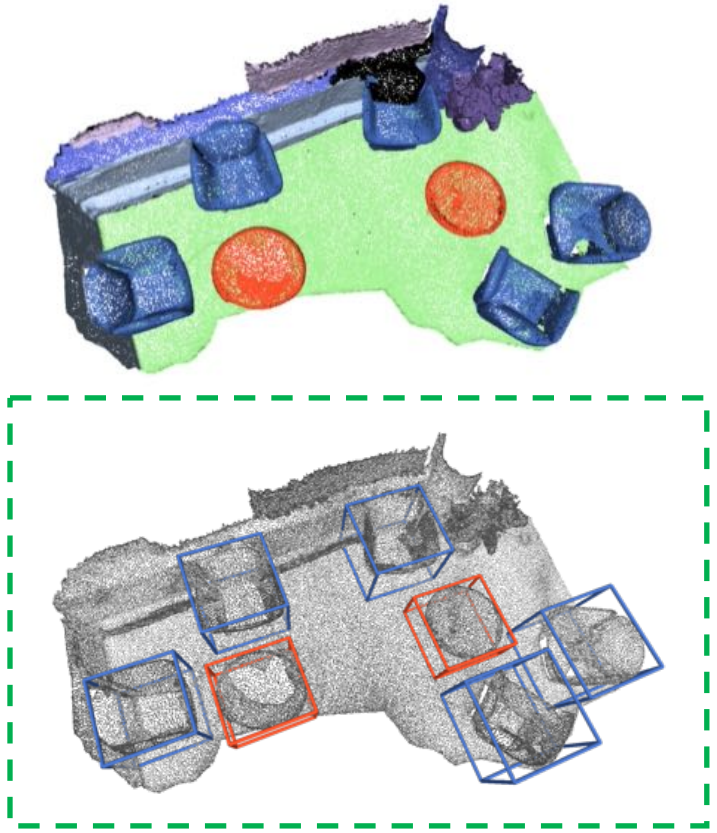


Point cloud | Detected shapes | GSS output | Ground-truth

Goal: Spatial Recognition





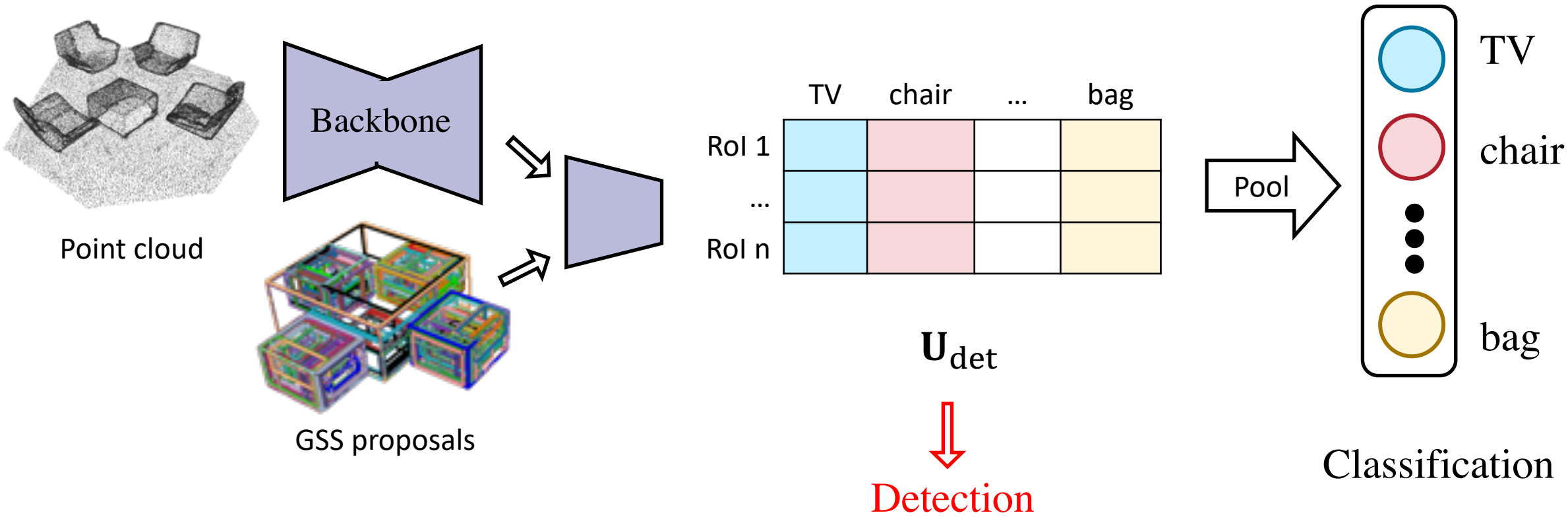


Input: point cloud

Goal: localizing each object

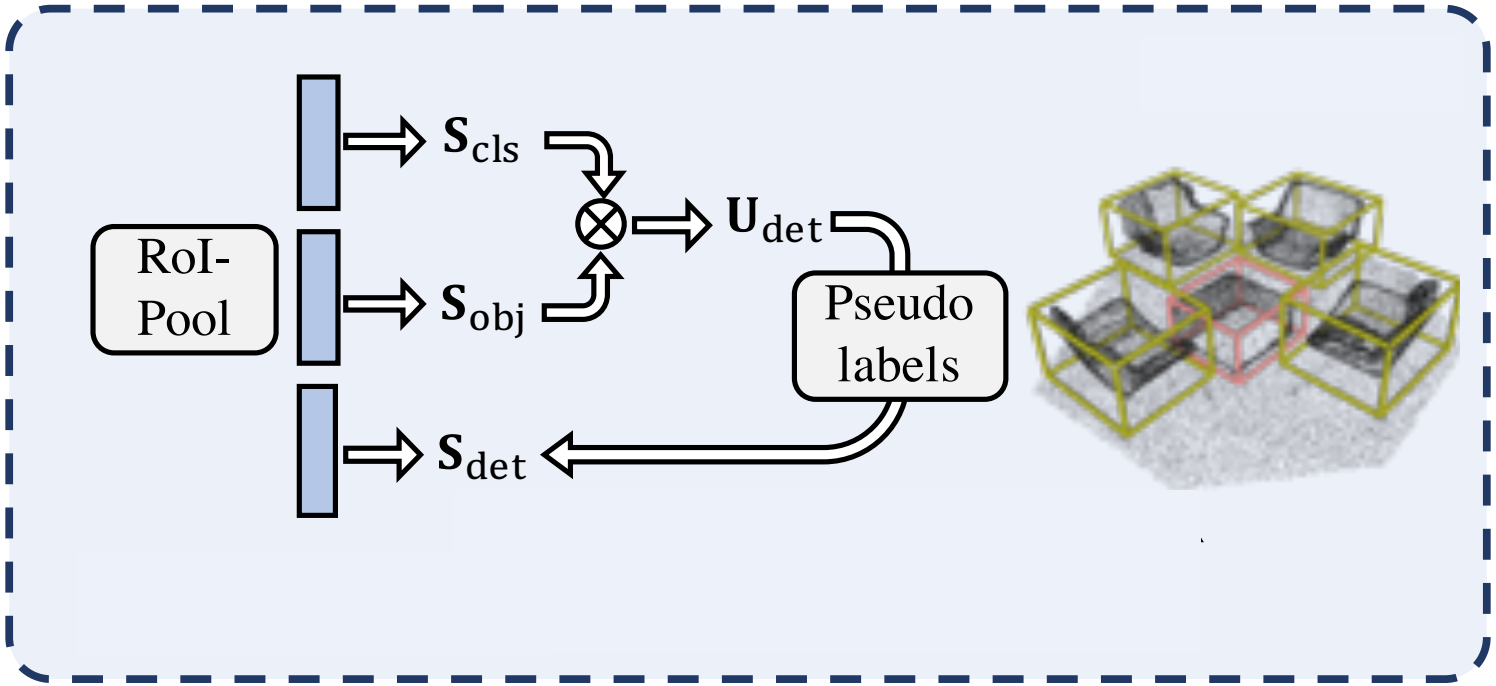
The “where” problem: detection

RoI Multiple Instance Learning (MIL)



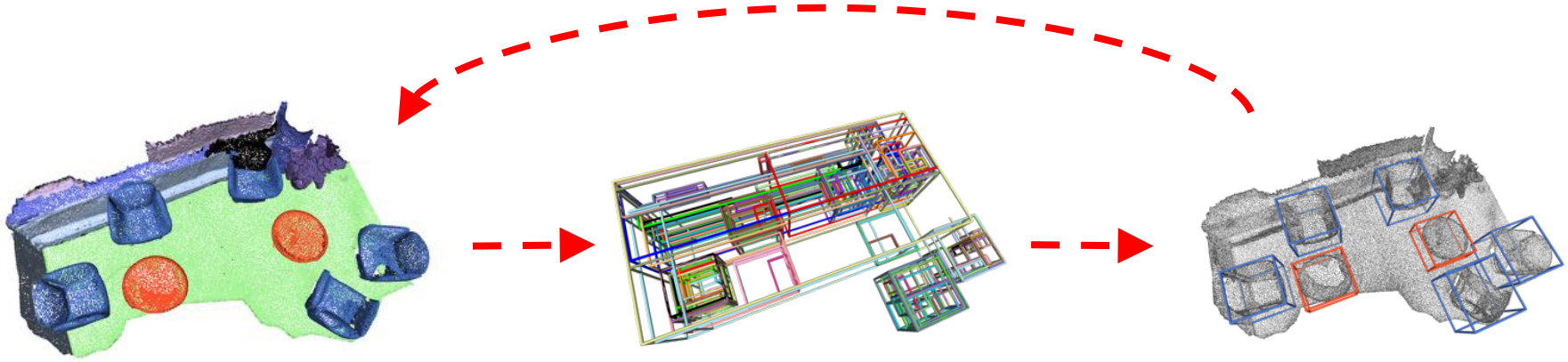
The “where” problem: detection

- **RoI** Self-training



- Cross-transformation consistency

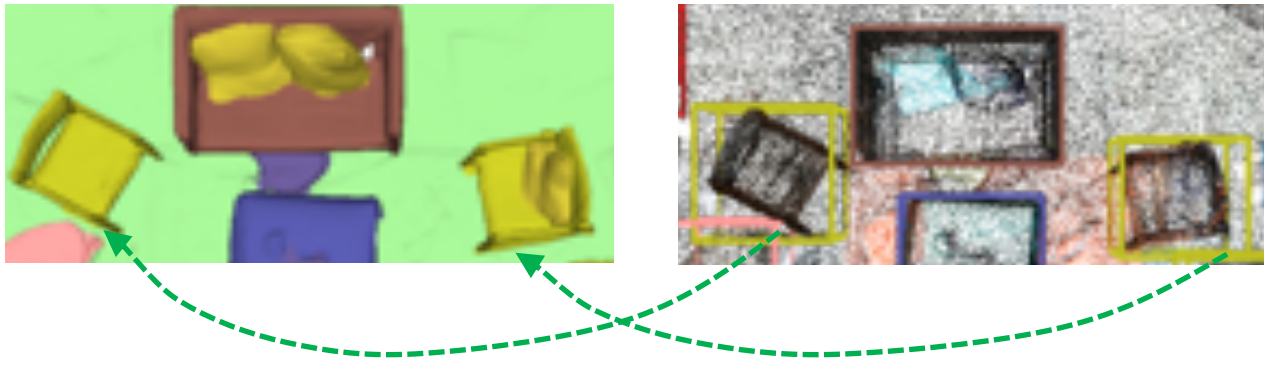
Bridging “what” & “where”: joint-training



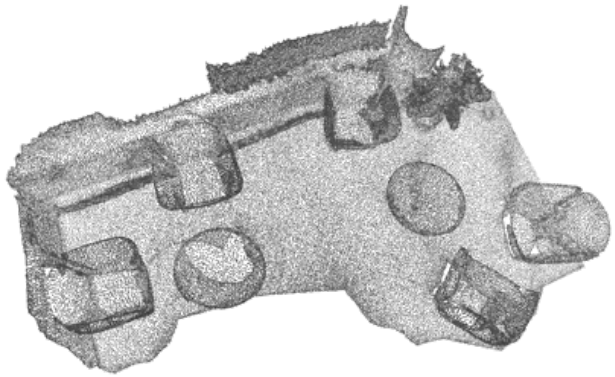
1. Better representation learning
2. Forward consistency
   - seg → proposal → det
3. Backward consistency
   - seg ← det

Backward consistency

Idea: label propagation from “confident” box to the points within it

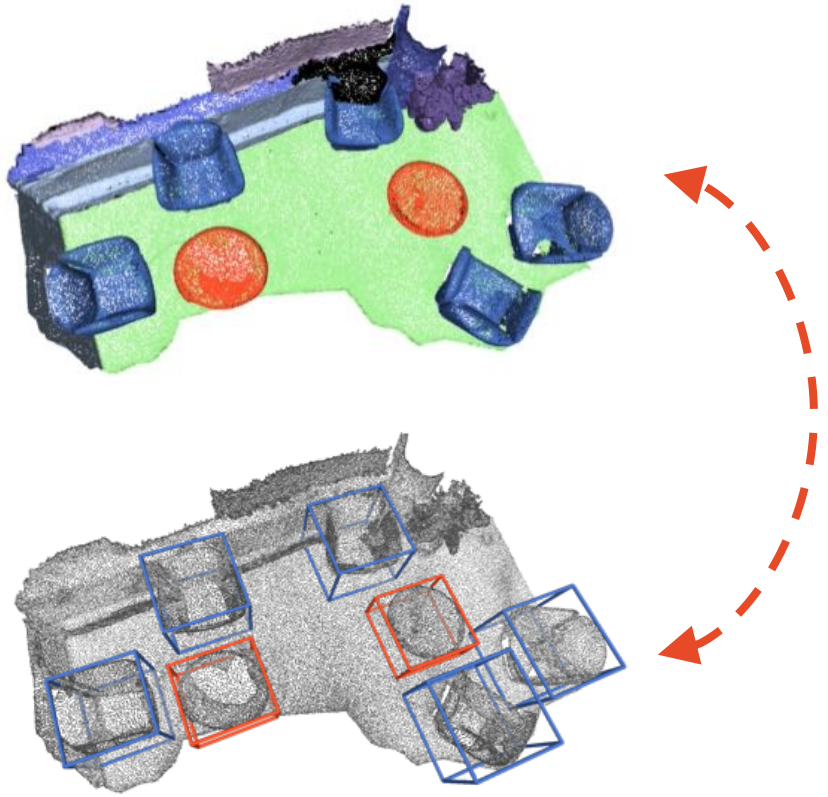


WyPR: Weakly-sup. Point Cloud Recognition



floor, wall, chair, table

WyPR



Input: point cloud

Goal: localizing each object

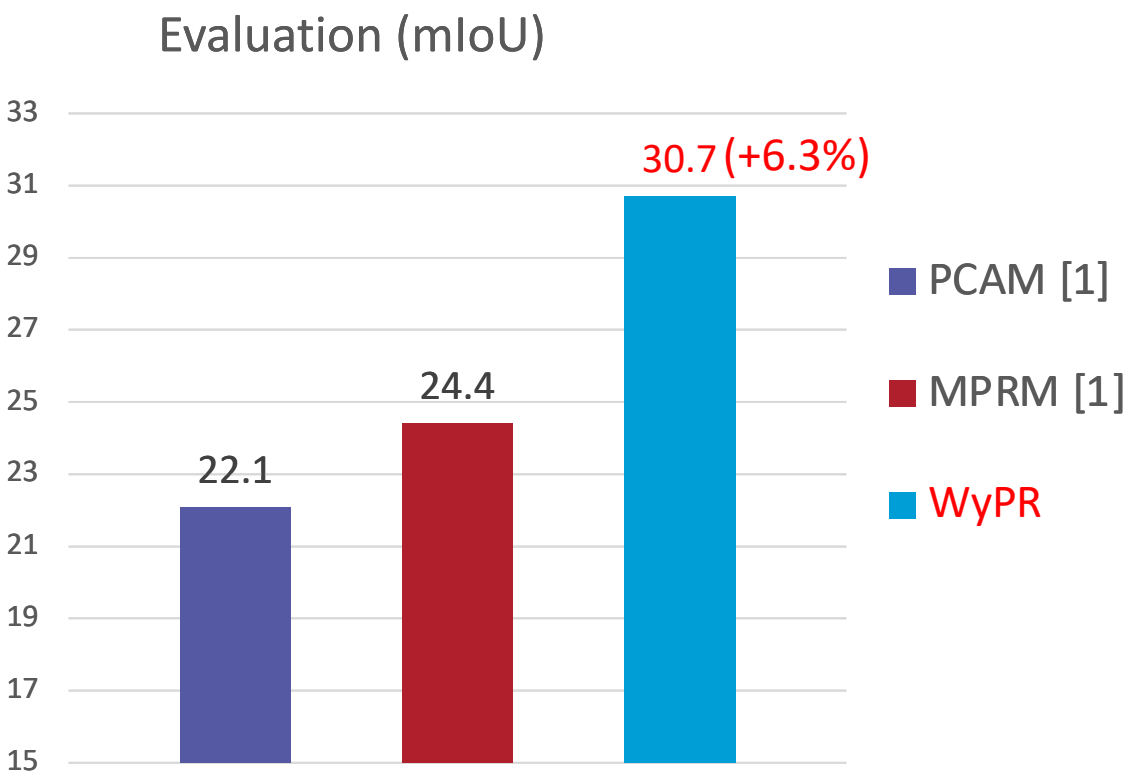
Experiments

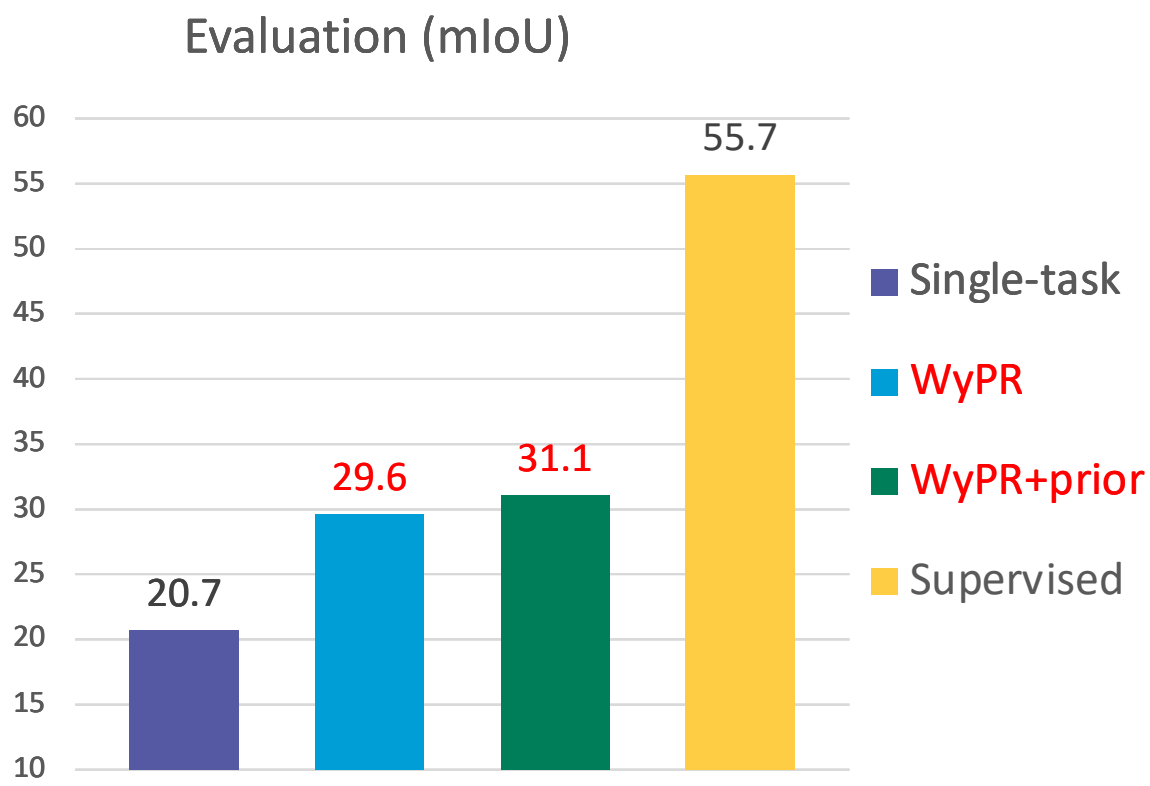
1. Backbone
   - PointNet++
2. Dataset
   - ScanNet, S3DIS
3. Metrics
   - mIoU / AR / mAP

Baselines

1. Single-task baseline
   - MIL-seg
   - MIL-det
2. External Prior (“WyPR+prior”)
   - Object shape (easily accessible from synthetic data)
   - Floor height
3. Prior work

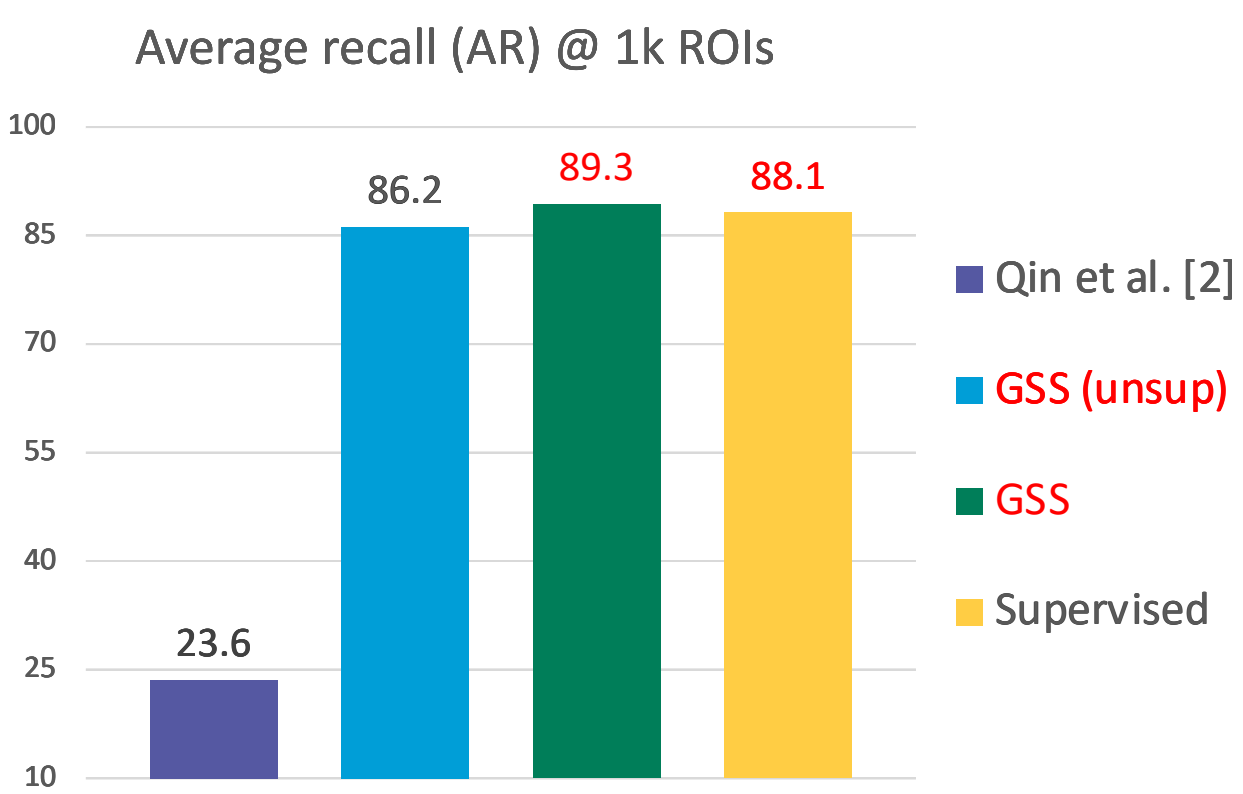
Semantic segmentation (ScanNet)

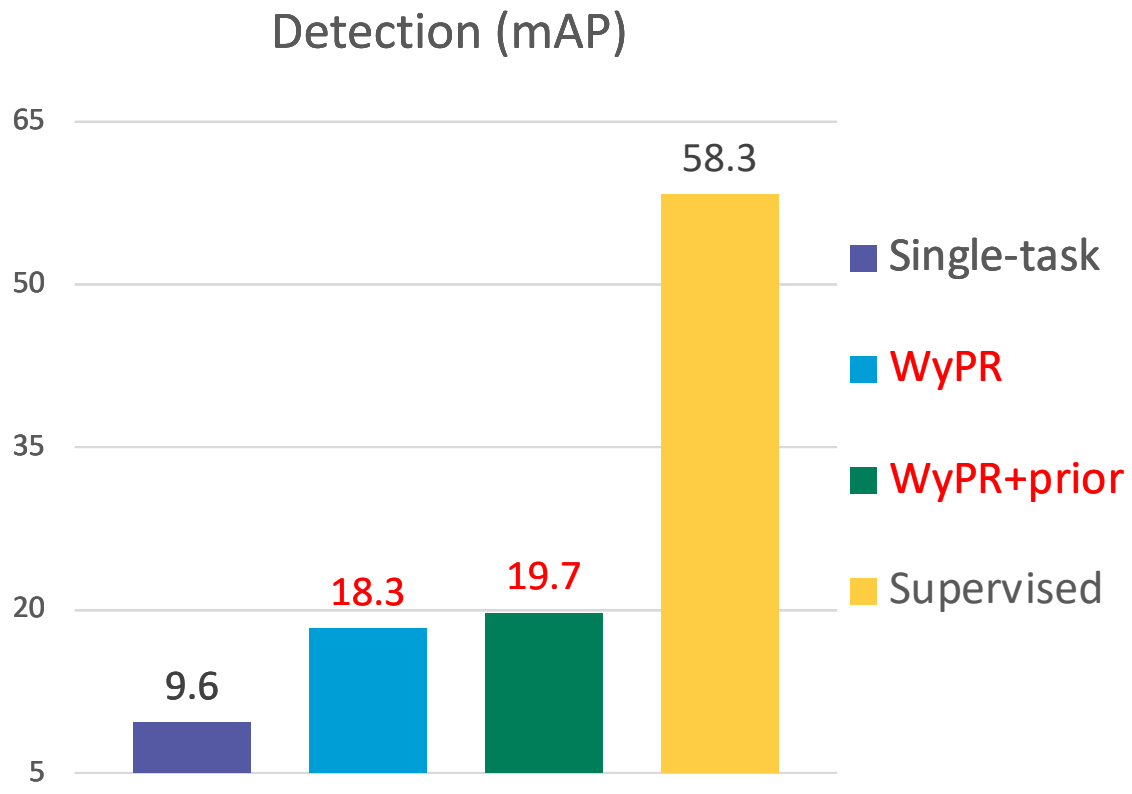




[1] Wei et al., Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds, 2020

Detection (ScanNet)



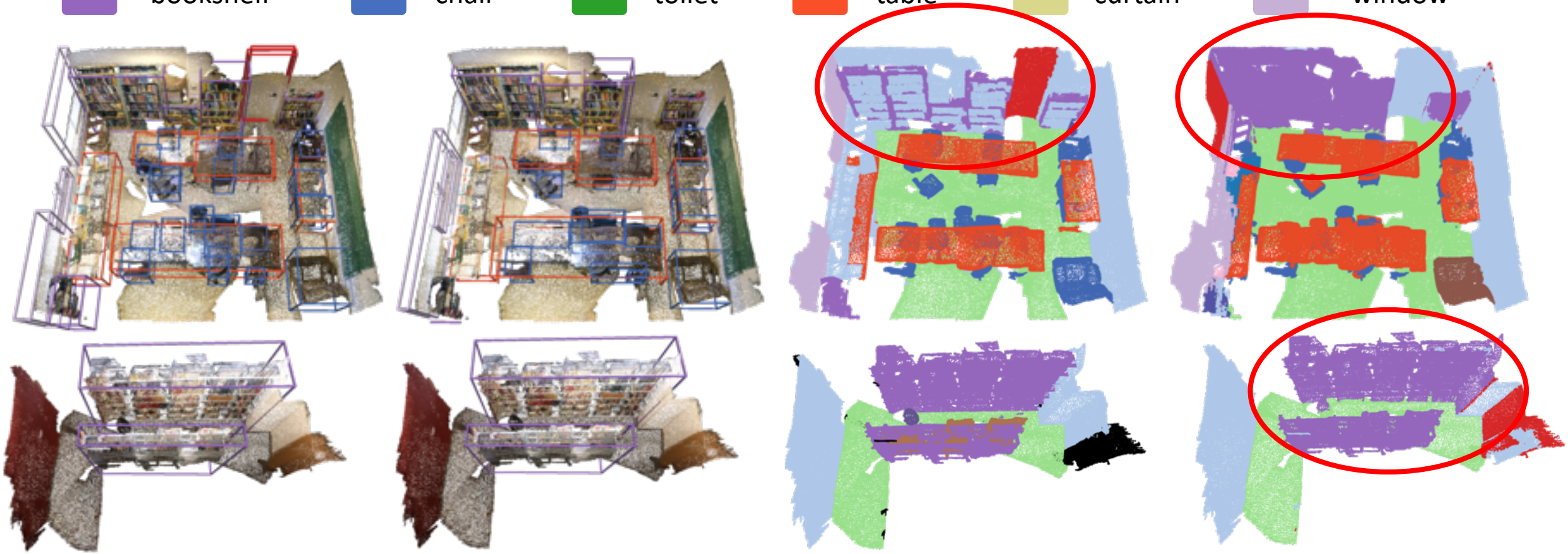


[2] Qin et al., Weakly Supervised 3D Object Detection from Point Clouds, 2020

Visualization

floor wall sofa door bed sink desk

bookshelf chair toilet table curtain window

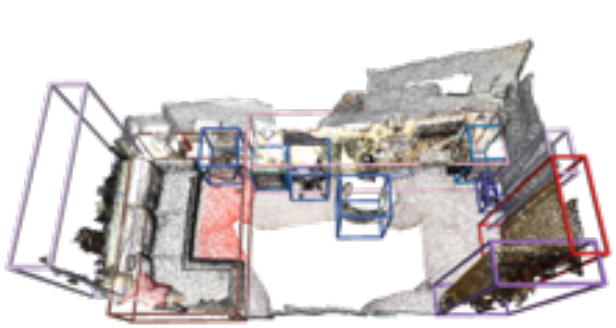
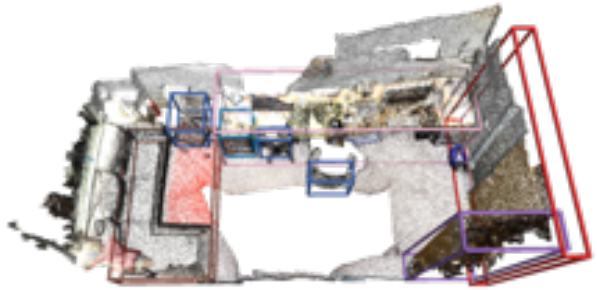
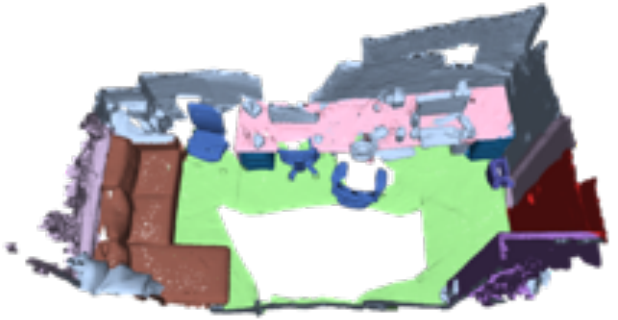


GT detection | Pred detection | GT segmentation | Pred segmentation

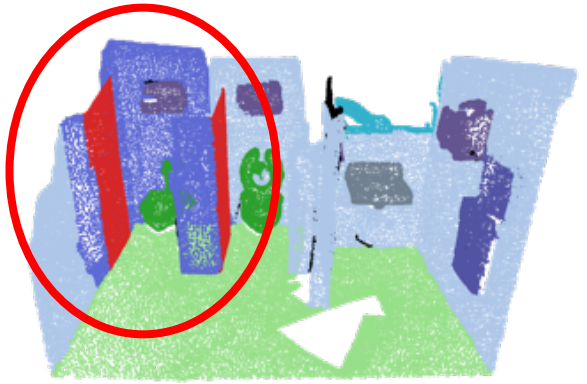
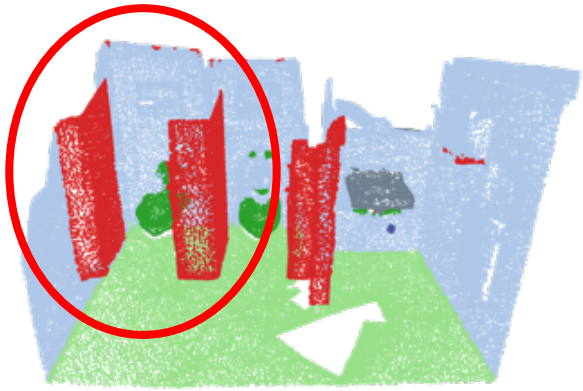
Visualization

floor wall sofa door bed sink desk

bookshelf chair toilet table curtain window

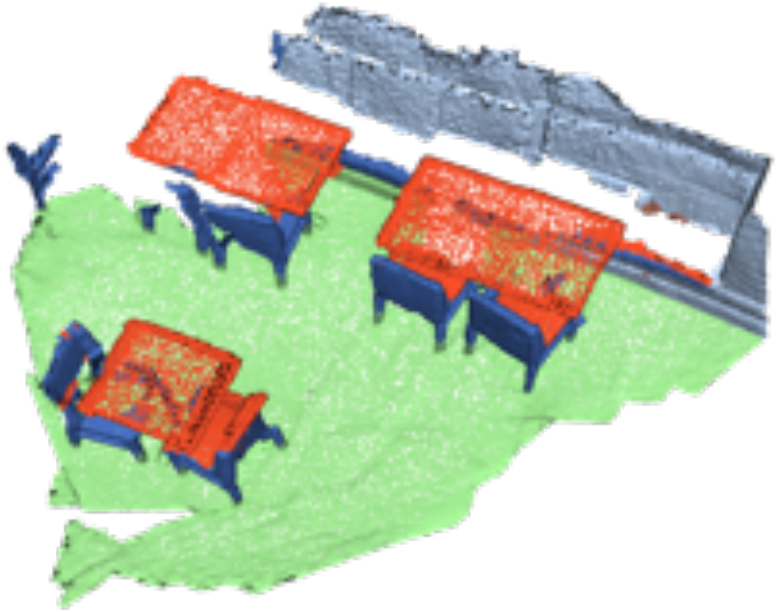




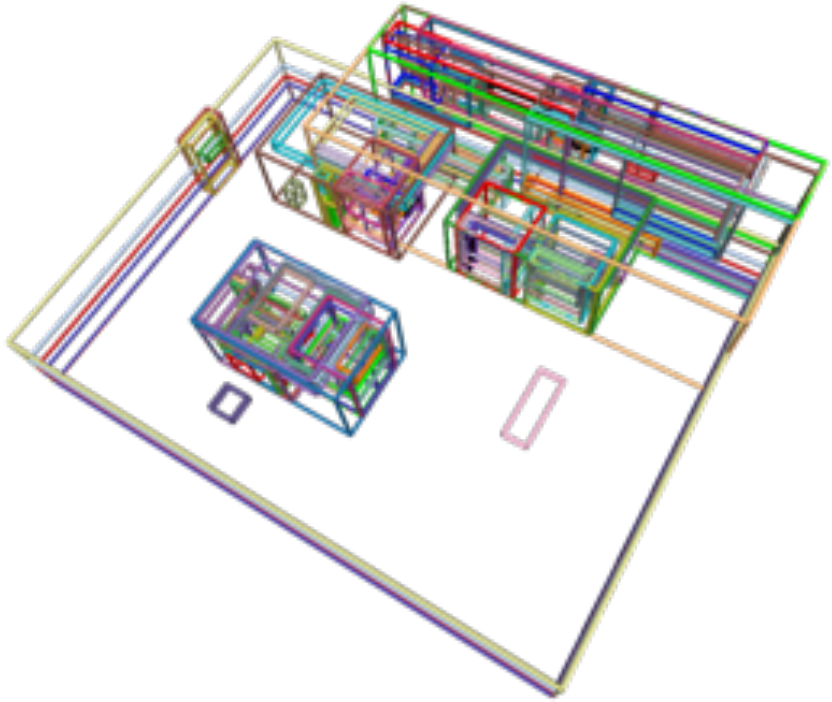


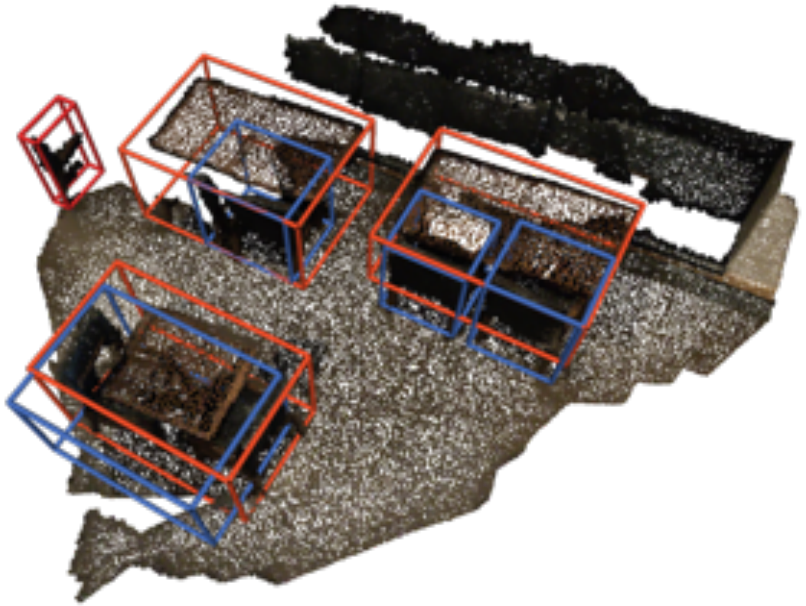



GT detection Pred detection GT segmentation Pred segmentation

Questions?







FACEBOOK