# 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans

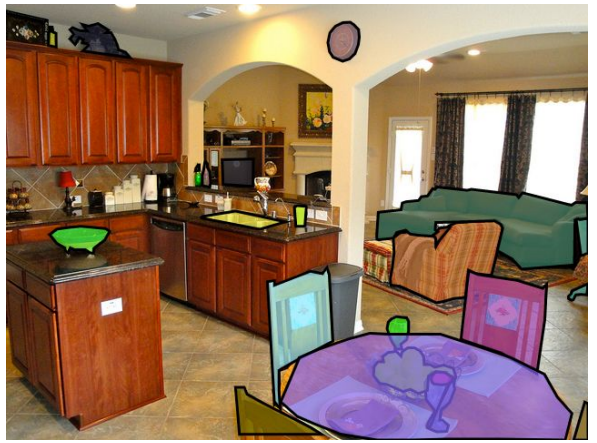
2020/02/07

Presenter: Haoping Xu

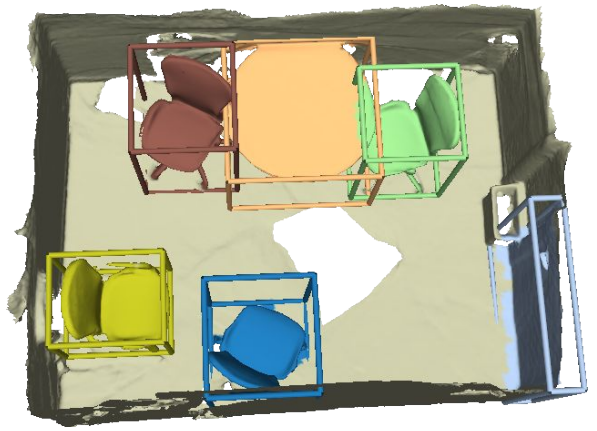
Authors: Ji Hou, Angela Dai, Matthias Nießner

UNIVERSITY OF TORONTO

# Motivation and Main Problem



- Semantic instance segmentation
  - Extend the 2D task to 3D
    - Bonding box
    - Class
    - 3D mask
  - Fuse both color and depth feature
    - Others only focus on 3D geometry



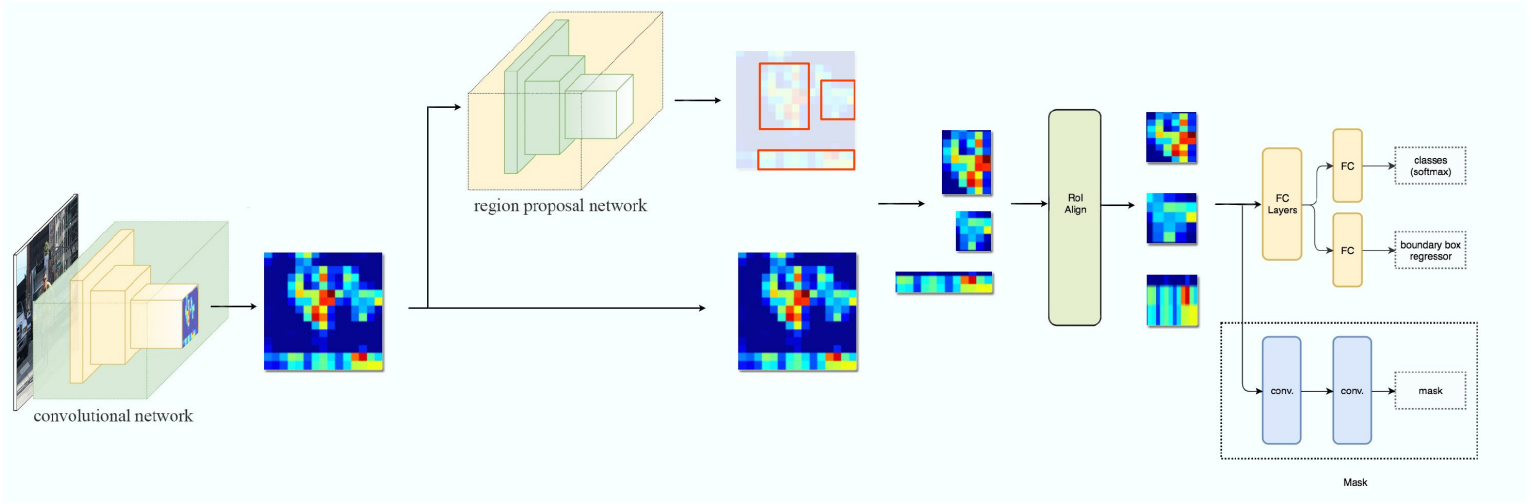
# Why RGB-D?

- Sensors readily available
  - Low price tag ( $300 – 700)
  - Available as consumer product
  - Mature product eco-system
    - Stable code base and utilities
    - Integration with application (i.e. ROS)
- Color signal cannot be ignored
  - Development in 2D CV
- Depth map should be the focus

# Contributions

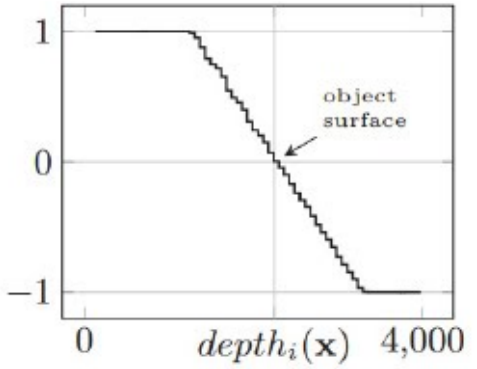
- Proposed a novel network to preform 3D semantic instance segmentation in RGB-D scenes.
  - New SOTA, improve mAP by 13.5 on ScanNet
- Unlike previous works that solely focuses on one feature
  - Either 2D frames (mulit-view fusion)
  - Or 3D geometry (PointNet)
- Fuse both features together
  - Proposed a back-projection method for color to geometry mapping
  - Extend the 2D Mask-RCNN pipeline to 3D

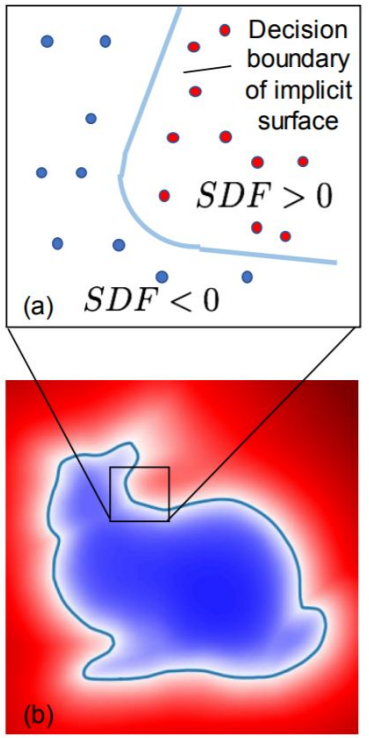
# General Background : Mask RCNN



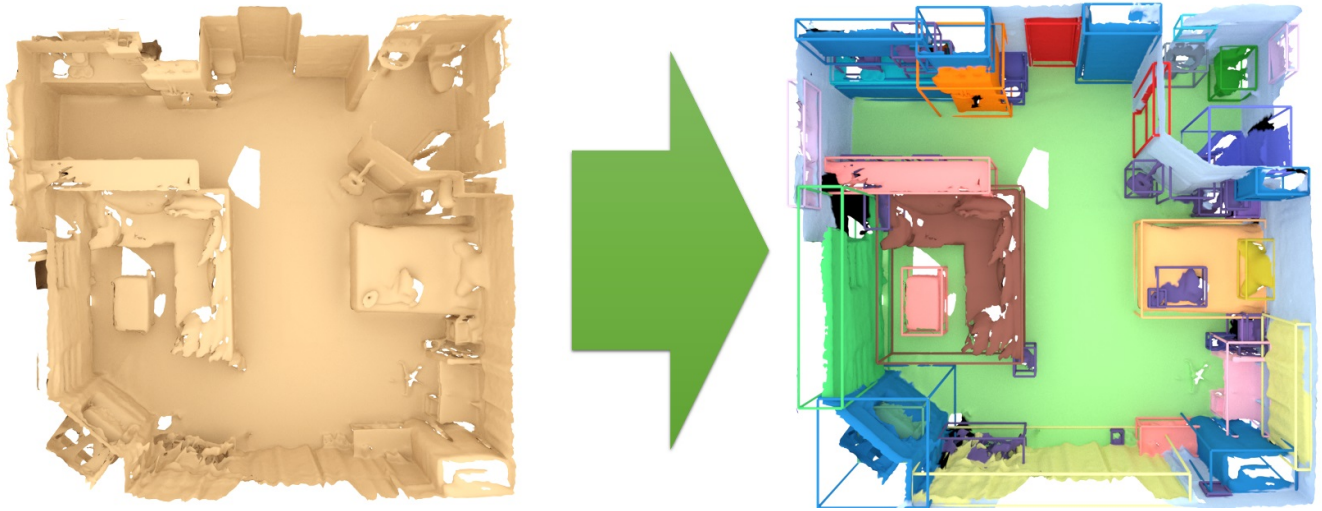

- Backbone
  - Extract features via conv
- RPN & RoI
  - Prepose instance region
  - Crop the feature region for down strain tasks
- Classification
- Bounding box
- Mask prediction
  - Pre pixel map of the instance

# Problem Setting




- Input: RGB-D scan
  - Depth: truncated sign distance field (TSDF)
  - RGB image: map to geometry based on 6DoF pose
- Output: semantic instance segmentation
  - Bounding Box
  - Class prediction
  - Pre-voxel 3D mask

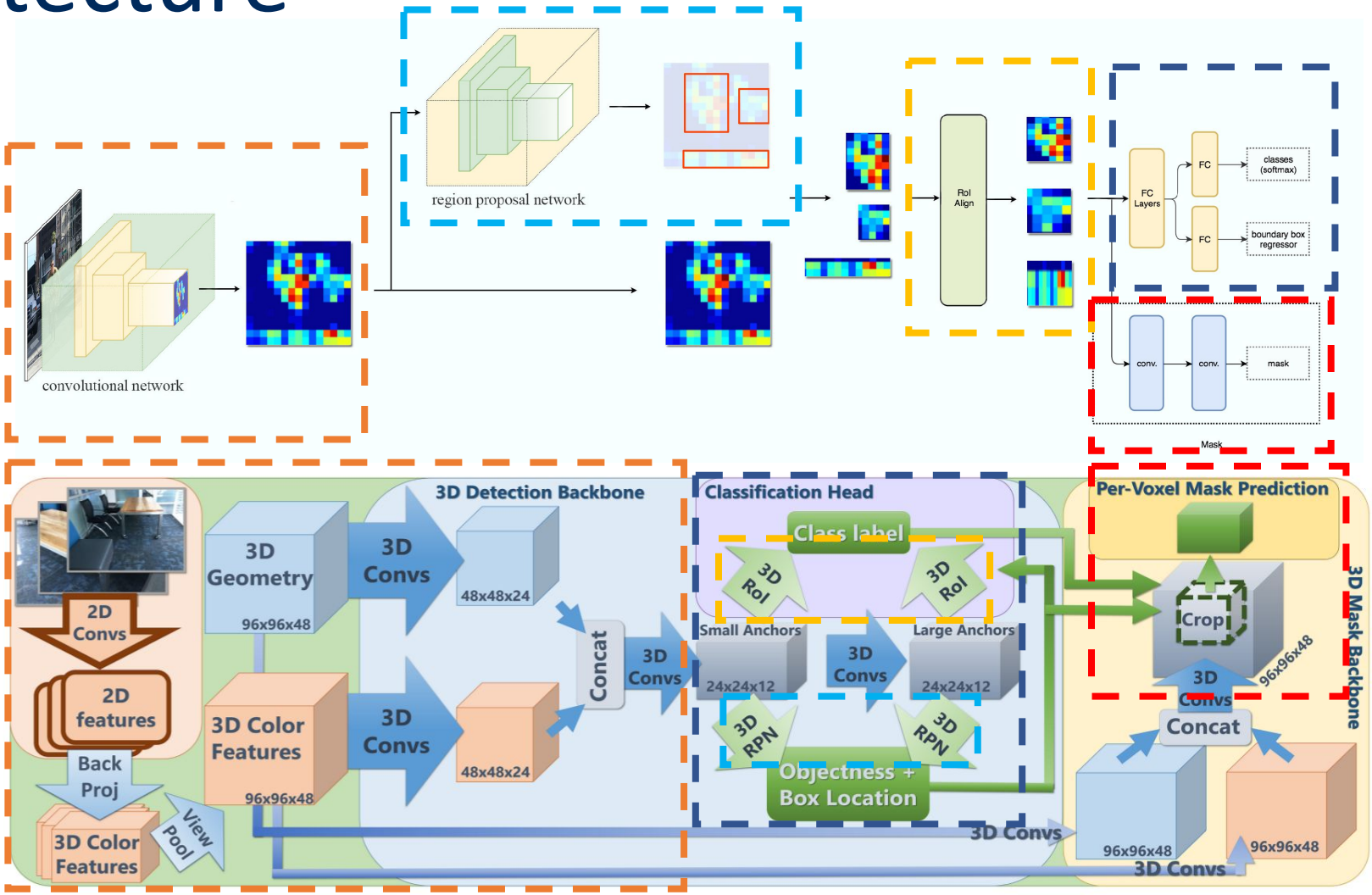


# Network architecture

## Mask RCNN

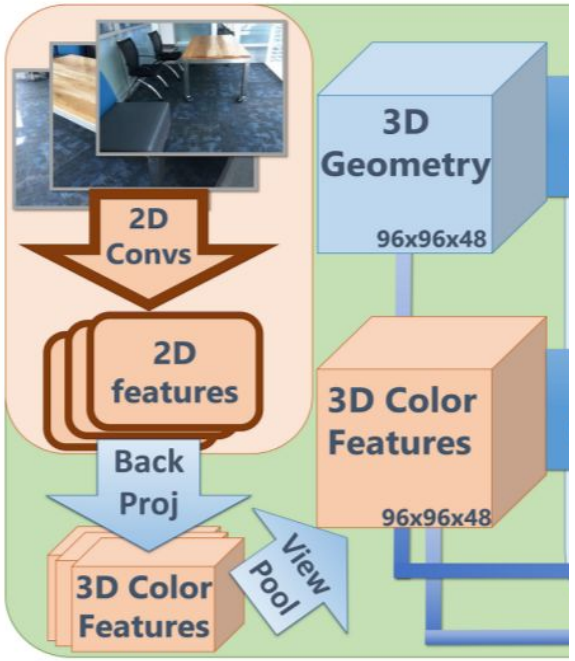
1. Backbone
2. RPN & RoI
3. Classification
4. Bounding box
5. Mask prediction

## 3D-SIS

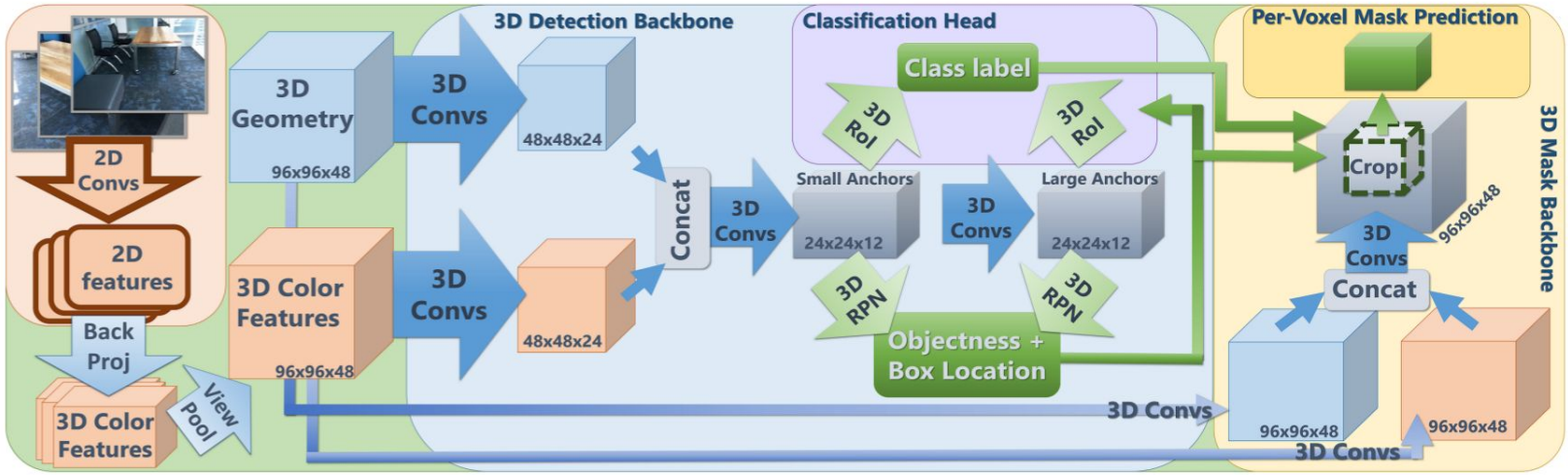


# 2D Backbone

- Back-project 2D to 3D:
  - Naive way: direct projection leads to down sampling
  - Better way: extract color feature before projection
- Pre-trained model:
  - ENet trained on NYUv2 (agnostic to network choice)
- Multi-view:
  - max pooling on 3D color feature
- Fuse with 3D:
  - do a 3D conv before concat with 3D geometry



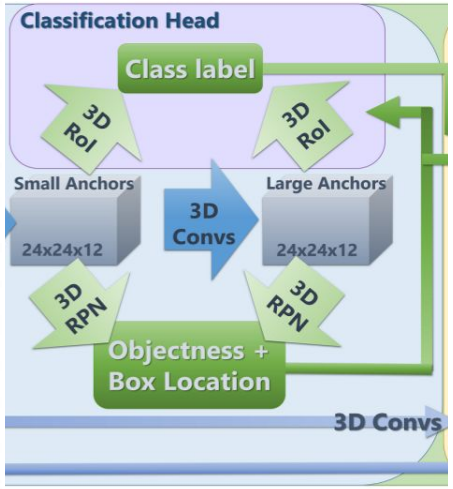
# 3D Backbone




- Voxel dimension:
  - size = 4.69 cm^3
  - 96 * 96 * 48 voxels in 4.5m * 4.5m * 2.25m scene
  - At test time: fixed height dimension and voxel size, variable xy size.
- 3D convolution:
  - 3D ResNet structure
  - Color and 3D process separately before join together
- Two branches backbone
  - Detaction branch reduce spatial dimension by 4
  - Segmentation branch maintain the dimension
  - Better result and train time compared to single branch, why?

# RPN & RoI

| small anchors | big anchors |
|---|---|
| (8, 8, 9) | (21, 7, 38) |
| (14, 14, 11) | (7, 21, 39) |
| (14, 14, 20) | (32, 15, 18) |
| | (15, 31, 17) |
| | (53, 24, 22) |
| | (24, 53, 22) |
| | (28, 4, 22) |
| | (4, 28, 22) |
| | (18, 46, 8) |
| | (46, 18, 8) |
| | (9, 9, 35) |

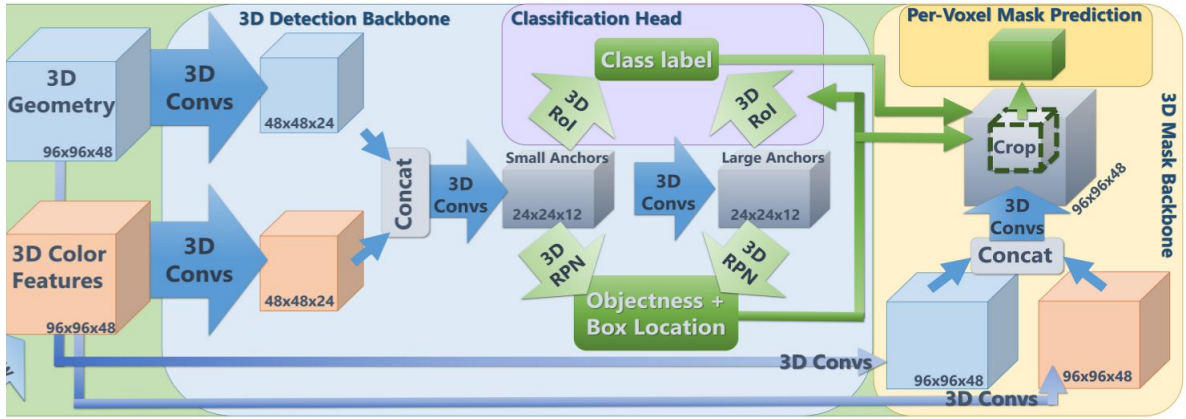



- Anchor selection:
  - K-mean (K=14) on training data
  - Split into small(1m^3) and large, ratio= 3:11
  - Process separately at RPN and RoI
- 3D RPN -> bounding box
  - Objectiveness = CE of two class. Positive(IoU>0.35), negative(IoU<0.15).
  - Δx, Δy, Δz, Δw, Δh, Δl for each anchor
- 3D RoI -> class label
  - Crop out feature based on bounding box
  - RoI max pooling to 4*4*4 blocks
  - Class prediction with MLP

$$\Delta_x = \frac{\mu - \mu_{anchor}}{\phi_{anchor}} \qquad \Delta_w = \ln(\frac{\phi}{\phi_{anchor}})$$

where $\mu$ is the box center point and $\phi$ is the box width.

# 3D Instance Segmentation




- Crop feature
  - Use bounding box to get fused feature
- Mask prediction
  - Predict mask for each class
  - Select the mask channel with class prediction
- Voxel resolution unchanged
  - Unlike detection task, resolution unchanged along backbone
  - Better performance than lower res
  - May due to the low res nature of voxel rigid

# Experimental Results

## Two datasets: ScanNetV2 (real) and SUNCG (synthetic)

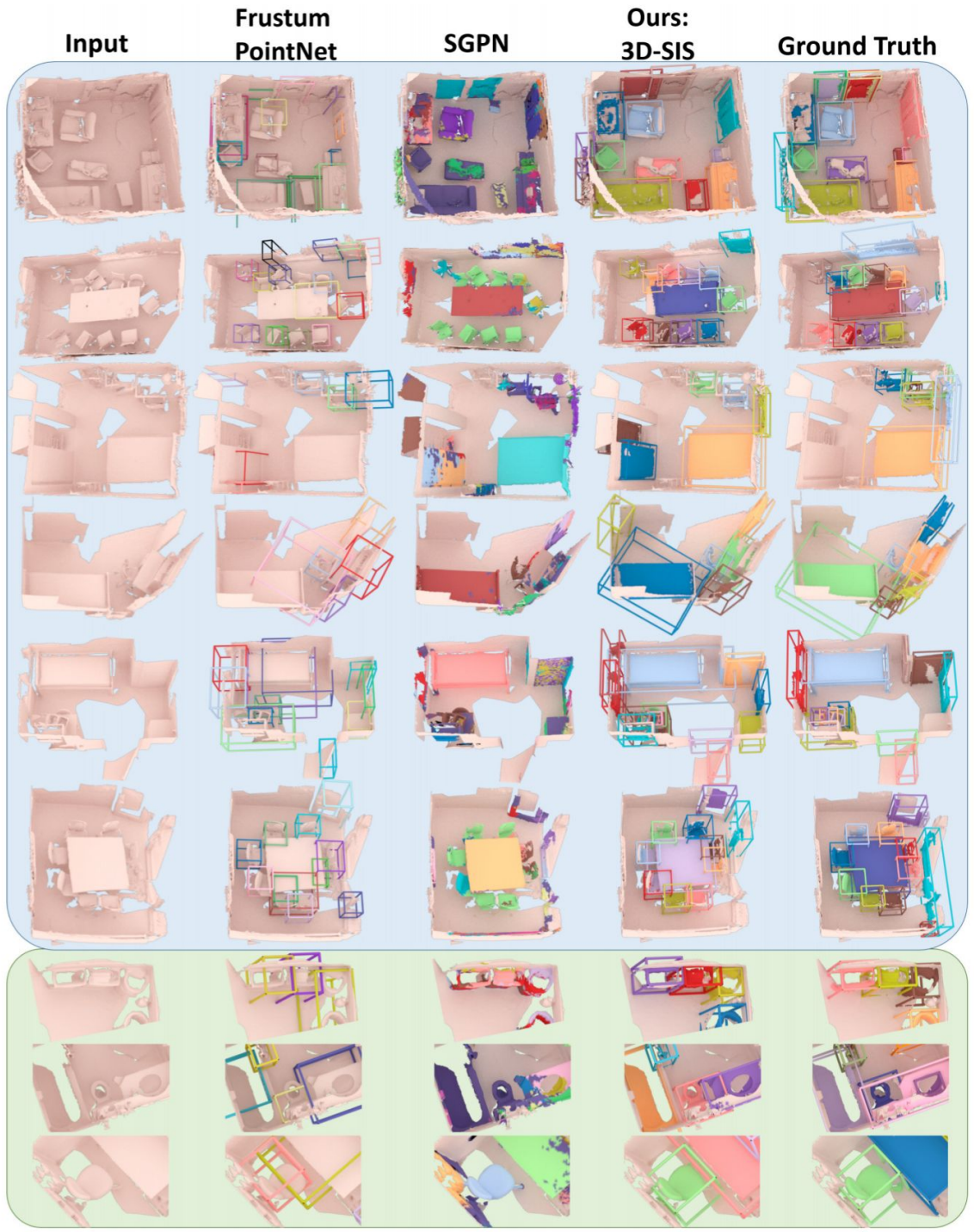
| | cab | bed | chair | sofa | tabl | door | wind | bkshf | pic | cntr | desk | curt | fridg | showr | toil | sink | bath | ofurn | **avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mask R-CNN [12] | 5.3 | 0.2 | 0.2 | 10.7 | 2.0 | 4.5 | 0.6 | 0.0 | **23.8** | 0.2 | 0.0 | 2.1 | 6.5 | 0.0 | 2.0 | 1.4 | 33.3 | 2.4 | 5.8 |
| SGPN [35] | 6.5 | 39.0 | 27.5 | 35.1 | 16.8 | 8.7 | 13.8 | 16.9 | 1.4 | 2.9 | 0.0 | 6.9 | 2.7 | 0.0 | 43.8 | 11.2 | 20.8 | 4.3 | 14.3 |
| MTML | 2.7 | **61.4** | 39.0 | 50.0 | 10.5 | 10.0 | 0.3 | **33.7** | 0.0 | 0.0 | 0.1 | **11.8** | 16.7 | 14.3 | 57.0 | 4.6 | 66.7 | 2.8 | 21.2 |
| 3D-BEVIS [10] | 3.5 | 56.6 | 39.4 | 60.4 | 18.1 | 9.9 | 17.1 | 7.6 | 2.5 | 2.7 | 9.8 | 3.5 | 9.8 | 37.5 | 85.4 | 12.6 | 66.7 | 3.0 | 24.8 |
| R-PointNet [37] | **34.8** | 40.5 | **58.9** | 39.6 | **27.5** | 28.3 | **24.5** | 31.1 | 2.8 | **5.4** | **12.6** | 6.8 | 21.9 | 21.4 | 82.1 | **33.1** | 50.0 | **29.0** | 30.6 |
| 3D-SIS (Ours) | 13.4 | 55.4 | 58.7 | **72.8** | 22.4 | **30.7** | 18.1 | 31.9 | 0.6 | 0.0 | 12.1 | 0.0 | **54.1** | **100.0** | **88.9** | 4.5 | **66.7** | 21.0 | **36.2** |

ScanNetV2

| | cab | bed | chair | sofa | tabl | door | wind | bkshf | cntr | desk | shlf | curt | drsr | mirr | tv | nigh | toil | sink | lamp | bath | ostr | ofurn | oprop | **avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Seg-Cluster | 16.8 | 16.2 | 15.6 | 11.8 | 14.5 | 10.0 | 11.7 | 27.2 | 20.0 | 25.7 | 10.0 | 0.0 | 15.0 | 0.0 | 20.0 | 27.8 | 39.5 | 22.9 | 10.7 | 38.9 | 10.4 | 0.0 | 12.3 | 16.4 |
| Mask R-CNN [12] | 14.9 | 19.0 | 19.5 | 13.5 | 12.2 | 11.7 | 14.2 | 35.0 | 15.7 | 18.3 | 13.7 | 0.0 | 24.4 | **23.1** | **26.0** | **28.8** | 51.2 | 28.1 | 14.7 | 32.2 | 11.4 | 10.7 | 19.5 | 19.9 |
| SGPN [35] | 18.6 | 39.2 | 28.5 | 46.5 | 26.7 | **21.8** | **15.9** | 0.0 | 24.9 | 23.9 | 16.3 | **20.8** | 15.1 | 10.7 | 0.0 | 17.7 | 35.1 | 37.0 | **22.9** | 34.2 | 17.7 | 31.5 | 13.9 | 22.5 |
| Ours(geo only) | 23.2 | **78.6** | 47.7 | 63.3 | 37.0 | 19.6 | 0.0 | 0.0 | 21.3 | 34.4 | 16.8 | 0.0 | 16.7 | 0.0 | 10.0 | 22.8 | **59.7** | 49.2 | 10.0 | 77.2 | 10.0 | 0.0 | **19.3** | 26.8 |
| Ours(geo+1view) | 22.2 | 70.8 | 48.5 | **66.6** | **44.4** | 10.0 | 0.0 | **63.9** | 25.8 | 32.2 | 17.8 | 0.0 | 25.3 | 0.0 | 0.0 | 14.7 | 37.0 | 55.5 | 20.5 | 58.2 | **18.0** | 20.0 | 17.9 | 29.1 |
| Ours(geo+3views) | **26.5** | 78.4 | 48.2 | 59.5 | 42.8 | 26.1 | 0.0 | 30.0 | 22.7 | **39.4** | 17.3 | 0.0 | **36.2** | 0.0 | 10.0 | 10.0 | 37.0 | 50.8 | 16.8 | **59.3** | 10.0 | **36.4** | 17.8 | 29.4 |
| Ours(geo+5views) | 20.5 | 69.4 | **56.2** | 64.5 | 43.8 | 17.8 | 0.0 | 30.0 | **32.3** | 33.5 | **21.0** | 0.0 | 34.2 | 0.0 | 10.0 | 20.0 | 56.7 | **56.2** | 17.6 | 56.2 | 10.0 | 35.5 | 17.8 | **30.6** |

SUNCG

# Experimental Results



# Discussion of results

- Overall improved result
  - SOTA when it released on mAP@0.5
  - Beats pure geometry and color based methods
- Show effectiveness of color and geometry fusion
  - Show improved mAP compared to geo only version
  - Also show multi-view boosts mAP

SUNCG

| | mAP@0.25 | mAP@0.5 |
|---|---|---|
| Deep Sliding Shapes [30] | 12.8 | 6.2 |
| Mask R-CNN 2D-3D [12] | 20.4 | 10.5 |
| Frustum PointNet [22] | 24.9 | 10.8 |
| Ours – 3D-SIS (geo only) | 27.8 | 21.9 |
| Ours – 3D-SIS (geo+1view) | 30.9 | 23.8 |
| Ours – 3D-SIS (geo+3views) | 31.3 | 24.2 |
| Ours – 3D-SIS (geo+5views) | **32.2** | **24.7** |

ScanNetV2

| | mAP@0.25 | mAP@0.5 |
|---|---|---|
| Deep Sliding Shapes [30] | 15.2 | 6.8 |
| Mask R-CNN 2D-3D [12] | 17.3 | 10.5 |
| Frustum PointNet [22] | 19.8 | 10.8 |
| Ours – 3D-SIS (geo only) | 27.6 | 16.0 |
| Ours – 3D-SIS (geo+1view) | 35.1 | 18.7 |
| Ours – 3D-SIS (geo+3views) | 36.6 | 19.0 |
| Ours – 3D-SIS (geo+5views) | **40.2** | **22.5** |

# Critique / Limitations / Open Issues

- RoI align better than RoI pooling in 3D?
  - In 2D mask rcnn, RoI align can increase performance
  - Would be good to compare them in 3D
- Discuss more about voxel resolution
  - Two branch backbone shows the impact of voxel resolution
  - Lack of discussion about that in paper
- Marginal effect of multi-view?
  - As ScanNet only has 5 images per data point, result stops at 5 view
  - Use SUNCG can study performance change for more images

# Contributions (Recap)

- 3D semantic instance segmentation has wide application and hard due to high dimension
- Previous networks focus on either color or geometry
- Propose a network 3D-SIS
  - fuse two features together ( RGB mutli-view and depth)
  - Extend mask rcnn pipeline to 3D
- 3D-SIS achieves STOA on ScanNetV2 and good result on SUNCG
- Problem the reading is discussing