# Evaluating Mask R-CNN Performance on Indoor Scene Understanding

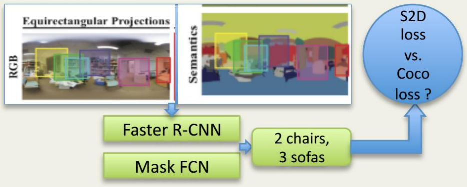
Shiva Badruswamy shivalgo@stanford.edu

## Abstract

Indoor robotics and AR are fast becoming the fundamental building blocks of future home living. However, evaluations on indoor images are non-existent due to privacy issues and acquisition cost . A variety of fast R-CNN instance segmentation NNs exist for outdoor scene understanding. Here, we use a modified, state-of-the-art Mask R-CNN on indoor 3D projected to 2.5D images to predict instances of 12 foreground classes on indoor, high-def images. A smaller MR-CNN performs well on Class loss and comparably on Mask and Box Loss.

## Introduction

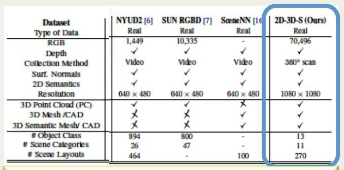
Stanford 2.5D Dataset



- Mask R-CNN has detection speeds in 4fps to 45fps (depending on use case) but is less accurate than Faster R-CNN, which is slower than 4fps even.
- Baseline MR-CNN performance highly dependent on quality of mask annotations
- Hard network to gauge performance: requires both bounding box and mask annotations.

## Data

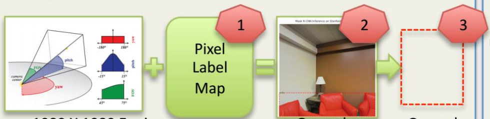
At most 1 instance per image



| Dataset | NYUD2 [6] | SUN RGBD [7] | SceneNN [16 | 2D-3D-S (Ours) |
|---|---|---|---|---|
| Type of Data | Real | Real | Real | Real |
| RGB | 1,449 | 10,335 | - | 70,496 |
| Depth | ✓ | ✓ | ✓ | ✓ |
| Collection Method | Video | Video | Video | 360° scan |
| Surf. Normals | ✓ | ✓ | ✓ | ✓ |
| 2D Semantics | ✓ | ✓ | ✓ | ✓ |
| Resolution | 640 × 480 | 640 × 480 | 640 × 480 | 1080 × 1080 |
| 3D Point Cloud (PC) | ✓ | ✓ | ✗ | ✓ |
| 3D Mesh /CAD | ✗ | ✗ | ✓ | ✓ |
| 3D Semantic Mesh/ CAD | ✗ | ✗ | ✓ | ✓ |
| # Object Class | 894 | 800 | - | 13 |
| # Scene Categories | 26 | 47 | - | 11 |
| # Scene Layouts | 464 | - | 100 | 270 |

Split: 5000 Train, 1000 Dev, 100 Test

## Ground Truth Generation



1080 X 1080 Equi-rectangular RGB projections

Ground Truth Masks

Ground Truth Boxes

(1) Each pixel color is encoded as 256-base number indexing to an instance label map.

**Index = 256*256*RGB(0)+256*RGB(1)+ 1*RGB(2)**

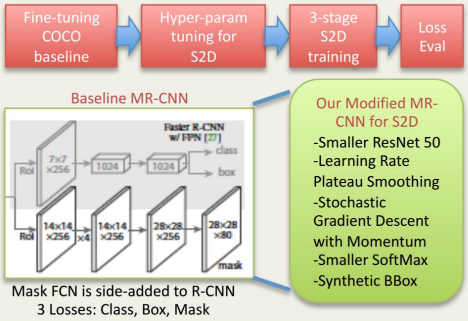
(2) Pixel location is tagged with label. Semantically colored pixels translated to **binary masks**. Then, compressed and stored using **RLE byte encoding**.

(3) GT Bounding Boxes are **extracted from masks**. Images with no foreground labeled masks are discarded.

Training, Validation, and Test Data from the same distribution - 100% of locations in Area 3 and 50% random in the large Area 5

## Models

Transfer Learning from MR-CNN pre-trained on COCO Weights



Baseline MR-CNN

Our Modified MR-CNN for S2D

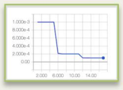
-Smaller ResNet 50
-Learning Rate Plateau Smoothing
-Stochastic Gradient Descent with Momentum
-Smaller SoftMax
-Synthetic BBox

Mask FCN is side-added to R-CNN
3 Losses: Class, Box, Mask

## Results

TensorBoard – LR Chart

**Robust transfer learning**, smaller feature maps, un-crowded, large instances lets us **decrease LR to 1e-04** (0.02 in MR-CNN Paper) for fast runs.



| Evaluation Performance | mAP | mAR | F1 |
|---|---|---|---|
| Fine-tuned COCO Baseline | 0.03 | 0.53 | 0.06 |
| Transfer-learnt S2D MR-CNN | 0.31 | 0.73 | 0.43 |

**Modified** MR-CNN Eval mAP is **10X higher** than Baseline

| Loss Comparison | MR-CNN Box | MR-CNN Class | MR-CNN Mask |
|---|---|---|---|
| Fine-tuned Coco baseline Training | 0.5817 | 0.5912 | 0.5680 |
| Fine-tuned Coco baseline Val | 0.5869 | 1.039 | 0.5583 |
| Transfer-learnt S2D MR-CNN training | 0.7574 | 0.093 | 0.6964 |
| Transfer-learnt S2D MR-CNN Val | 0.7445 | 0.1103 | 0.6925 |

- Our modified MR-CNN Sparse Cross Entropy SoftMax Class Loss is **lower** than Baseline, as we have low occlusion, simpler feature maps, and just 12 classes.
- Both training and validation losses **converge** in several hyper-param tuning runs indicating model's robustness for transfer learning to indoor features.

## Summary

- Our model demonstrates that the state-of-the-art Mask R-CNN **gains in accuracy** in **less occluded, less dynamic scenes**.
- **1/6th reduction in training loss** and almost **1/10th reduction in validation loss** is a promising result to investigate further.
- Adding RGB depth via z-axis distance or via surface normals could lead to further accuracy improvements.

## Reference

1. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girschick, Mask R-CNN, 2017, arXiv:1703.06870v3 [cs.CV]
2. Iro Armeni, Alexander Sax, Amir R. Zamir, Silvio Savarese, Stanford University, UC Berkeley, Joint 2D-3D-Semantic Data for Indoor Scene Understanding.
3. Thomas M. Breuel, DFKI and U.Kaiserslautern, Efficient Binary and Run Length Morphology and its Application to Document Image Processing, 2 Dec 2007, arXiv:0712.0121v1[cs.GR]

# YouTube Video Link

https://www.youtube.com/watch?v=1QsR8IcVV50&feature=youtu.be