Evaluating Mask R-CNN Performance on Indoor Scene Understanding
Shiva Badrurswamy  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 2D 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 2D 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

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

- Fine-tuning COCO baseline
- Hyper-param tuning for 2D
- 3-stage training
- Loss Eval

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

Ground Truth Generation

- 1080 X 1080 Equirectangular RGB projections
- Pixel Label Map
- Truth Masks
- Ground

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

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

- 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 less occluded, less dynamic scenes.
- 1/10th 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
2. T. Alvarez, Alexios Gnecco, Anir Ram, David Hogg, Stanford University. UCI Berkeley Indoor 2D-3D Semantic Data for Indoor Scene Understanding.
YouTube Video Link

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