**Motivation/Introduction:**
- Important part of applications such as automated driving
- Explore transfer learning to train a small dataset using a pretrained Mask R-CNN model
- Investigate whether incorporating depth enhances object detection part of instance segmentation

**Data:**
- NYU depth V2 dataset: 1449 densely labeled pairs of aligned Kinect depth and RGB images.
- Contains 895 object classes → limited to 80 classes (mapped to COCO dataset classes - for transfer learning baseline)

**Challenges:**
1. **Small** labeled dataset - challenge to train proposed architecture and baseline
2. The labels are being aggregated such that the neighboring objects of the same type are labeled together with a single label.

**Future:**
- The proposed model would benefit from using much larger and better annotated dataset. Princeton SUN-RGBD would be a viable alternative.
- Using a larger computational budget would help improve the scope and results of the study.
- With more data, we could train more than just the head layer of the network, so that it learns features more pertinent to the current data.

**RGB-D Model:**
- ResNet-FPN Backbone
- Convolution Feature Map
- RPN
- ResNet-FPN Backbone
- Convolution Feature Map
- ROI Align
- Concatenation
- 7x7x256, 1024, 2048
- Class
- Bounding Box
- 14x14x256, 14x14x256, 28x28x256, 28x28x256

**Results:**

**Discussion:**
- RGB-D results had more accurate class predictions than RGB on average.
- All three model results above show that bounding boxes try to mimic GT.
- Depth image contains no information about picture frames - no picture frames in RGB-D result.
- Loss curves show that RGB-D has marginally better loss over RGB.
- Transfer learning loss relatively flat (only network heads trained).
- RGB and RGB-D loss curves plateau due to small dataset: examples do not completely define multidimensional space.

**References:**

**Table above shows mAP scores for each model and experiment:** RGB-D achieves similar if not better scores in each category. Transfer learning scores are higher overall because of pretrained knowledge.