**INTRODUCTION**

Problem Statement: Fruits and vegetables are usually not packaged in boxes. Most stores use plastic tags with identification numbers on the surface, which presents several problems.

✦ Unsustainable and high maintenance cost
✦ Not Environment-friendly
✦ More label-intensive for store

Goal: build a DL system that can efficiently classify and localize fruits and vegetables. We aim to apply the state-of-the-art technique Mask R-CNN.

**DATASET**

MvTec Densely Segmented Supermarket (D2S) Dataset

- 60 unique categories
- 1920x1440 RGB images
- Training/Dev/Test split: 2880/360/360
- Annotation format similar to COCO
- With various lighting, rotations, and backgrounds

**MODEL ARCHITECTURE (MASK-RCNN)**

Model Pipeline

- Backbone model (Resnet-50)
- Region Proposal Network (RPN)
- Region of Interest Classification and Bounding Box
- Segmentation Masks

Model Pipeline Diagram

- Backbone + RPN
- RoI Classifier and Bbox

**ERROR ANALYSIS**

- Analysis on mAP on each category and relationship between object count performed
- Carrot, Adelholzener water, and Kamillentee have lowest mAP
- Error appear to relate to difference in size of objects from the same category and object occlusion with limited label displaying

**RESULTS**

Transfer learning was applied and various of parameters were experimented. Model is initiated with COCO pre-trained weights.

Results: mAP of 84 is achieved at IoU 50:95, with tolerance to partial obstruction.

- **Backbone Network**
  - Backbone AP 
  - ResNet-50: 0.783
  - ResNet-101: 0.801

- **Image Size**
  - 800x1024: 0.877
  - 512x512: 0.839

- **Number of layers to re-train**
  - heads: 0.828
  - 2+ : 0.793
  - all: 0.773

**FUTURE WORK**

We show that Mask R-CNN can achieve state-of-art result in object instance segmentation with the right hyper-parameter tuning.

For future work, we could expand the model’s application to other tasks such as fruit/vegetable sorting and quality control. To build a robustness model, increase the level of complexity of dataset, including collecting more data with images from different angles, with different background, etc.