

# Human-Object Interaction in Retail

Max Tanski & Lauren Heintz  
6.869 Computer Vision Final Project

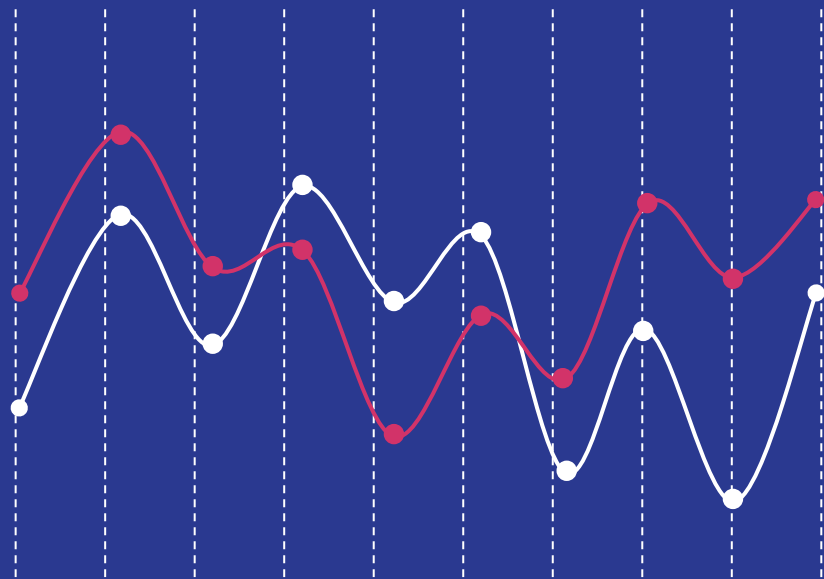


# Purpose

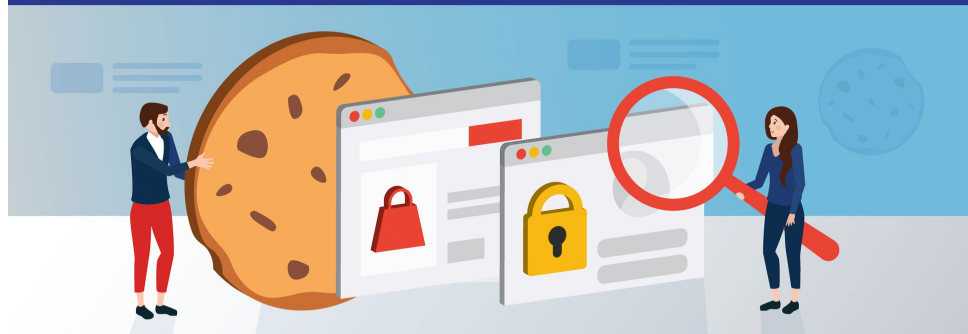


# The Rise of Retail Cookies

Capturing human object interaction in grocery stores



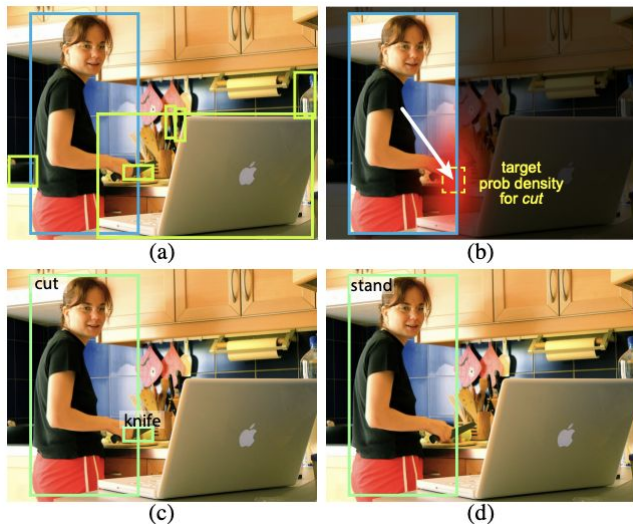
What are the others buying?



# State of the Art Approaches

< human, action, target >

## Instance Segmentation & Pose Estimation

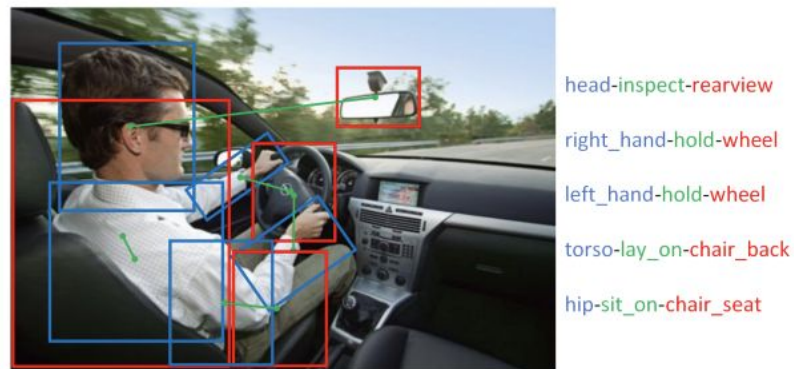


FacebookAI InteractNet

< Person, Cut, Knife >

< human, action, target >

## Keypoints & Part State Relationships

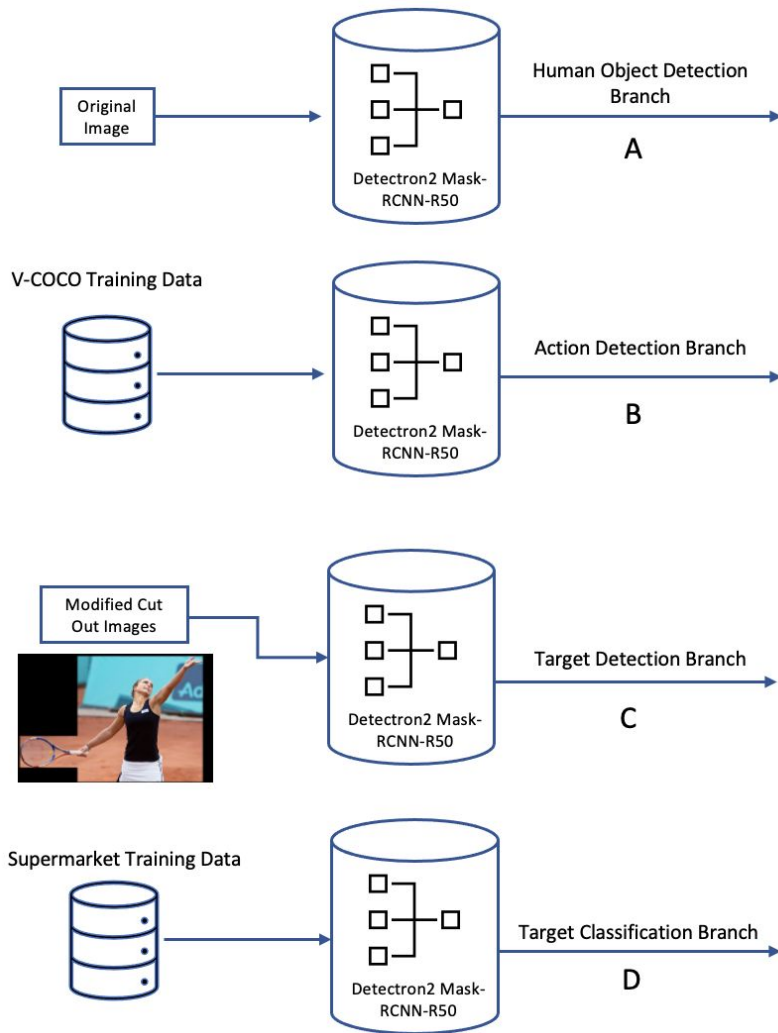


HAKE: Human Activity Knowledge Engine

< Person, Drives, Car >

< human, action, target >

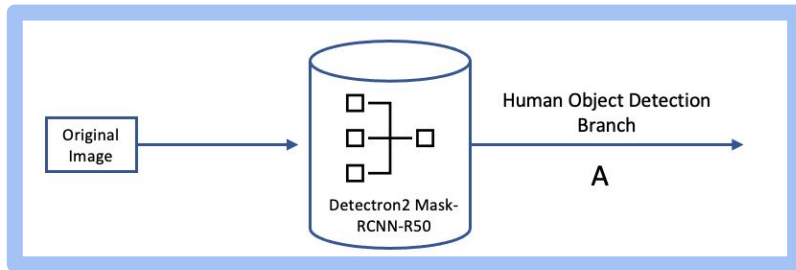
# Approach



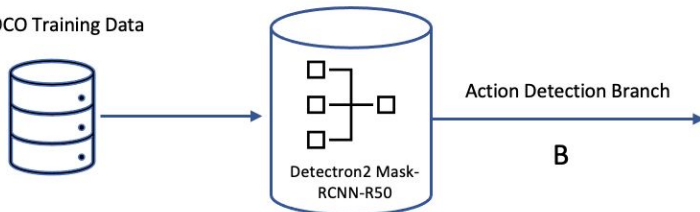
# Our Approach

< human, action, target >

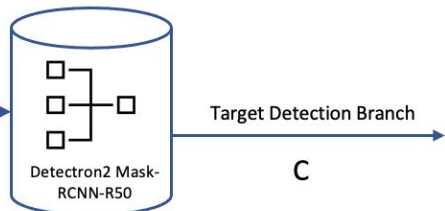
- A. Human Object Detection Branch
- B. Action Detection Branch
- C. Target Detection Branch
- D. Target Classification Branch



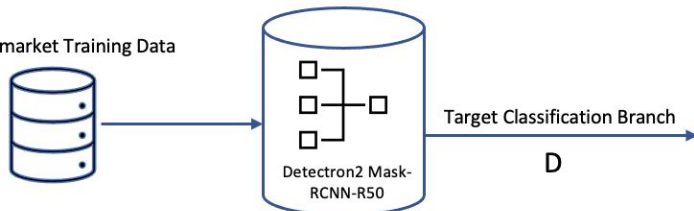
V-COCO Training Data



Modified Cut Out Images



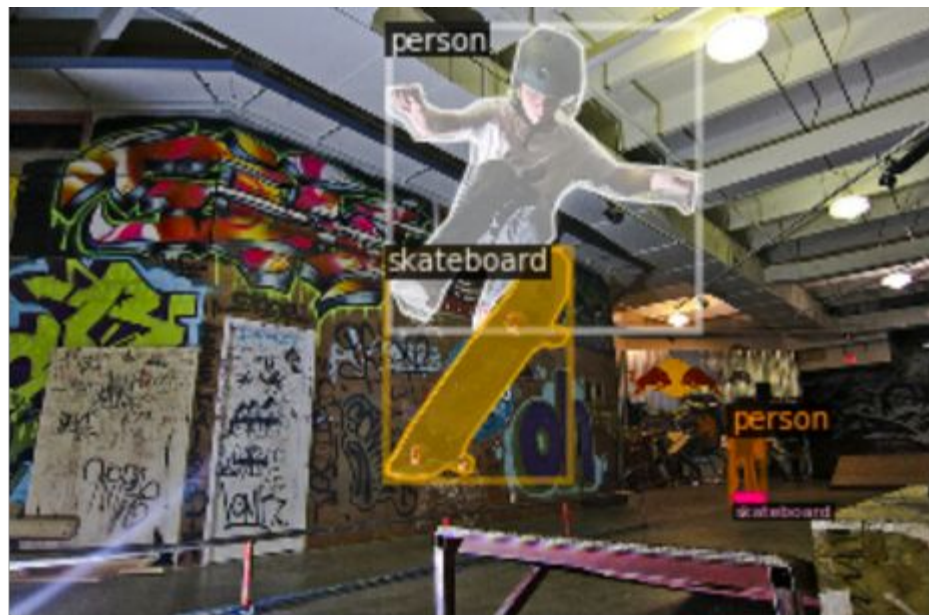
Supermarket Training Data



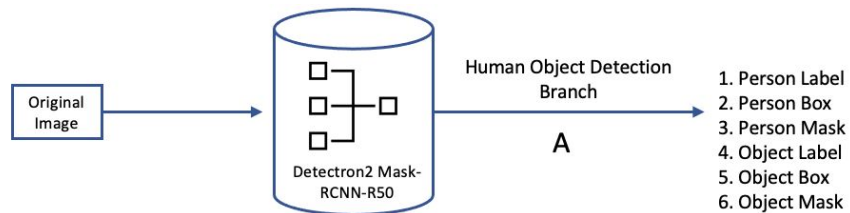
1. Person Label
2. Person Box
3. Person Mask
4. Object Label
5. Object Box
6. Object Mask

## Branch A

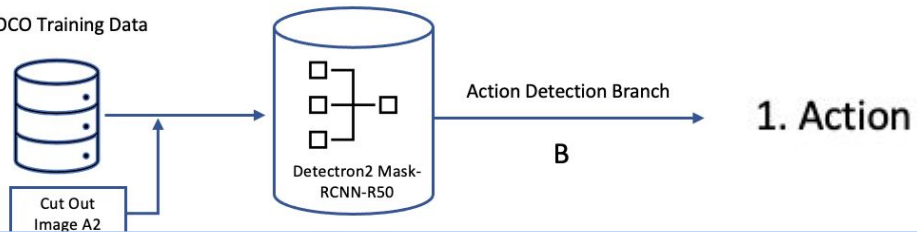
< human, ~~action~~, target >



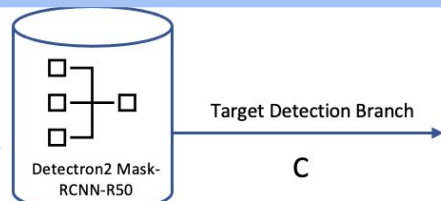




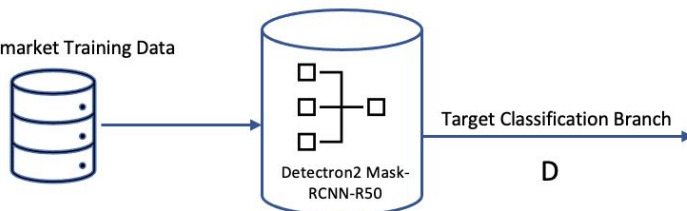
V-COCO Training Data



Modified Cut Out Images



Supermarket Training Data



# Branch B

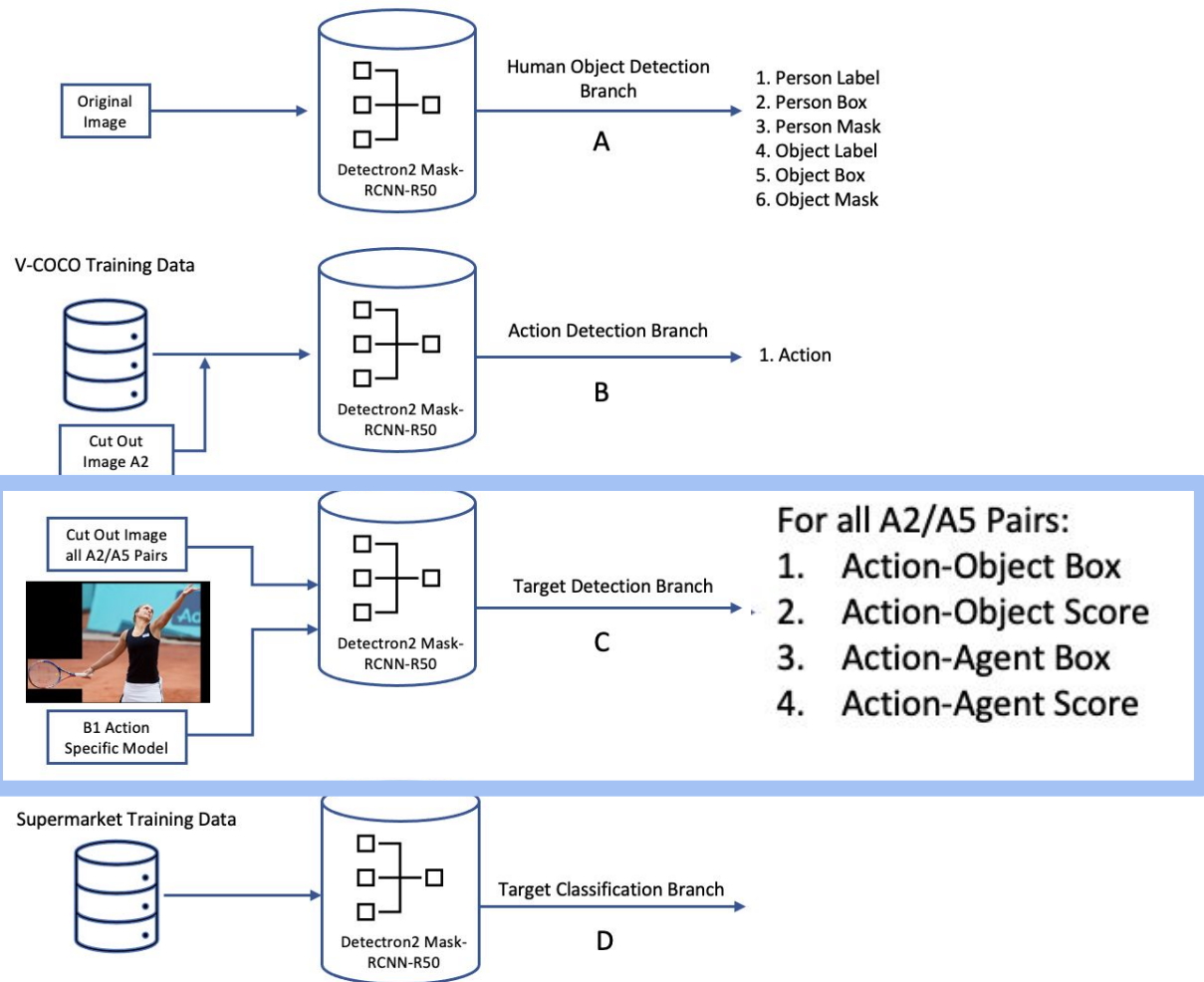
< human, action, target >

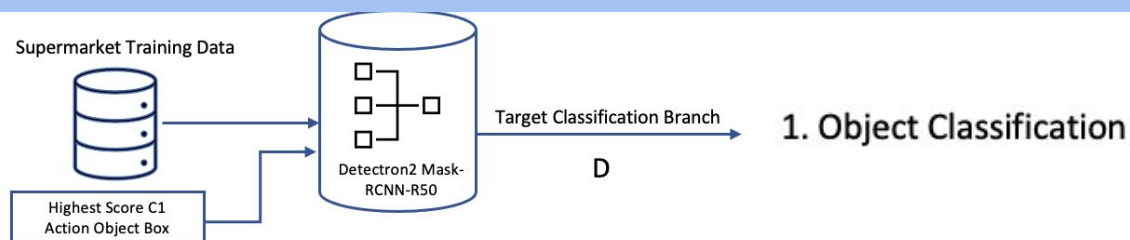
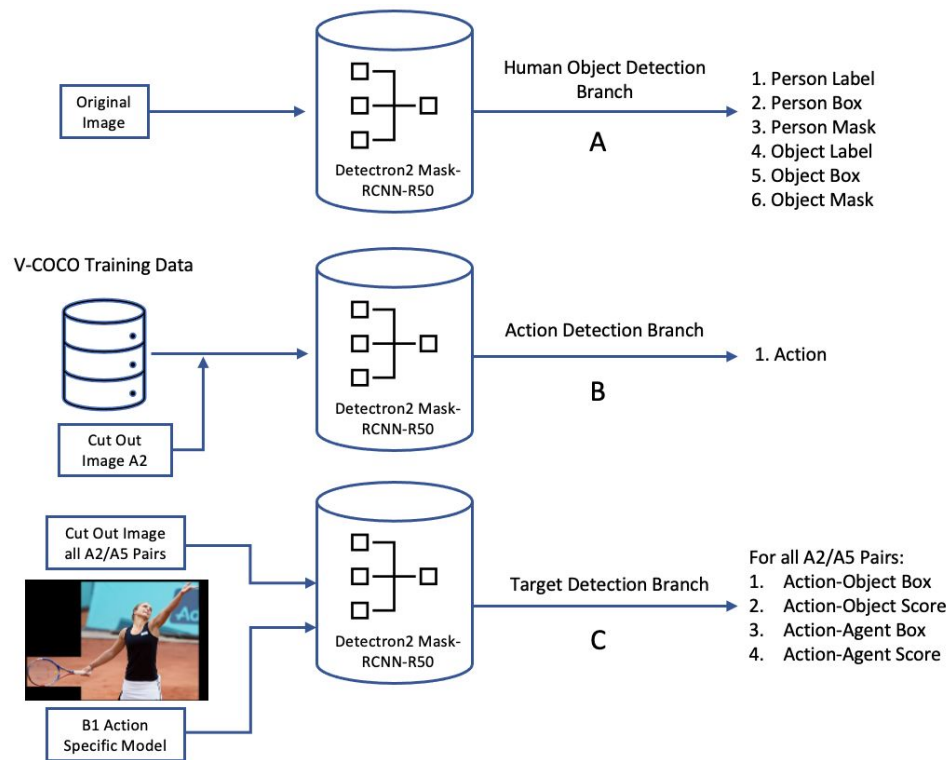




# Branch C

< human, ~~action~~, target >





# Branch D

< human, action, target >



# Results

# Experiment Parameters

## Backbones

Types tested:

`mask_rcnn_R_50`

`faster-rcnn_R_50`

`retinanet_R_101`

## Epochs

Values tested:

100

300

500

1000

1200

## Learning Rate

Values tested:

0.01

0.002

0.001

## Data Sets

Types:

COCO

V-COCO

V-COCO Agents

V-COCO Cutouts

Grocery Data

# Metrics for Action Detection and Object Detection

## Average Precision

$$Precision = \frac{TP}{TP + FP}$$

$TP$  = True positive

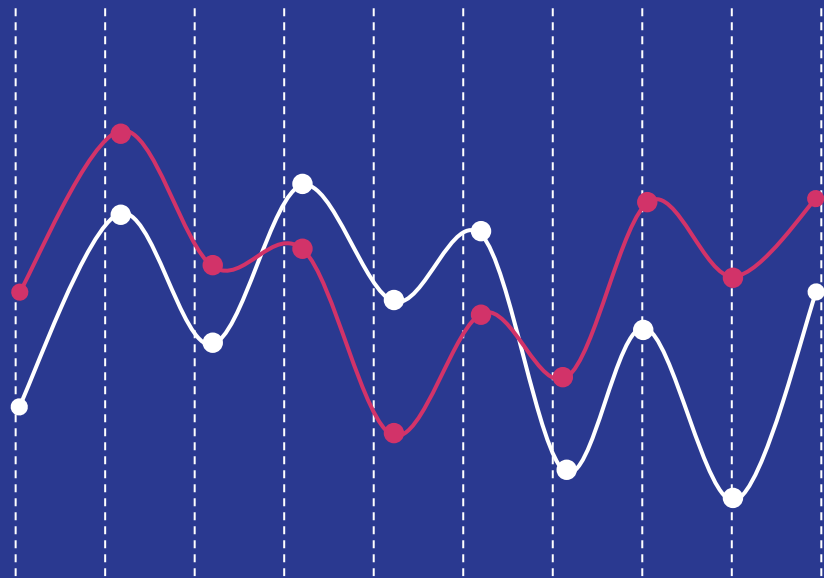
$TN$  = True negative

$FP$  = False positive

$FN$  = False negative

## Mean Average Precision

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{|TP_c|}{|FP_c| + |TP_c|}$$



# Branch mAP Best Results

## Branch A

Model:

**mask\_rcnn\_R\_50**

Data: COCO

Metrics:

AP = 53.6

mAP = 39.9

## Branch B -

Model:

**retinanet\_R\_101**

Data: V-COCO Agent

Metrics:

AP = 14.973

mAP = 18.309

“Hold” AP = 37.947

## Branch C

Model:

**faster-rcnn\_R\_50**

Data: V-COCO Cutout

Metrics:

AP = 12.776

mAP = 14.092

“Hold\_obj” AP =  
38.219

## Branch D

Model:

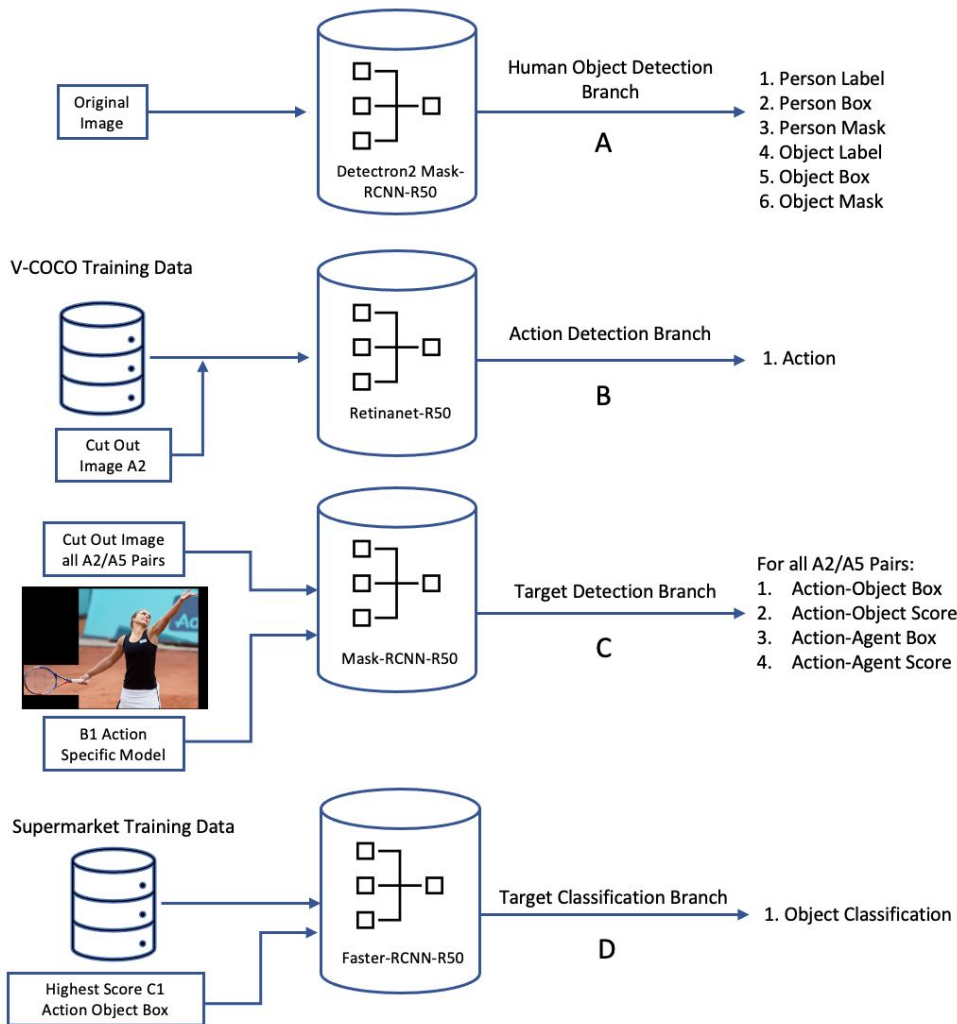
**faster-rcnn\_R\_50**

Data: Groceries

Metrics:

AP = 23.377

mAP = 9.414



# Final Architecture

< human, action, target >

- A. Human Object Detection Branch
- B. Action Detection Branch
- C. Target Detection Branch
- D. Target Classification Branch

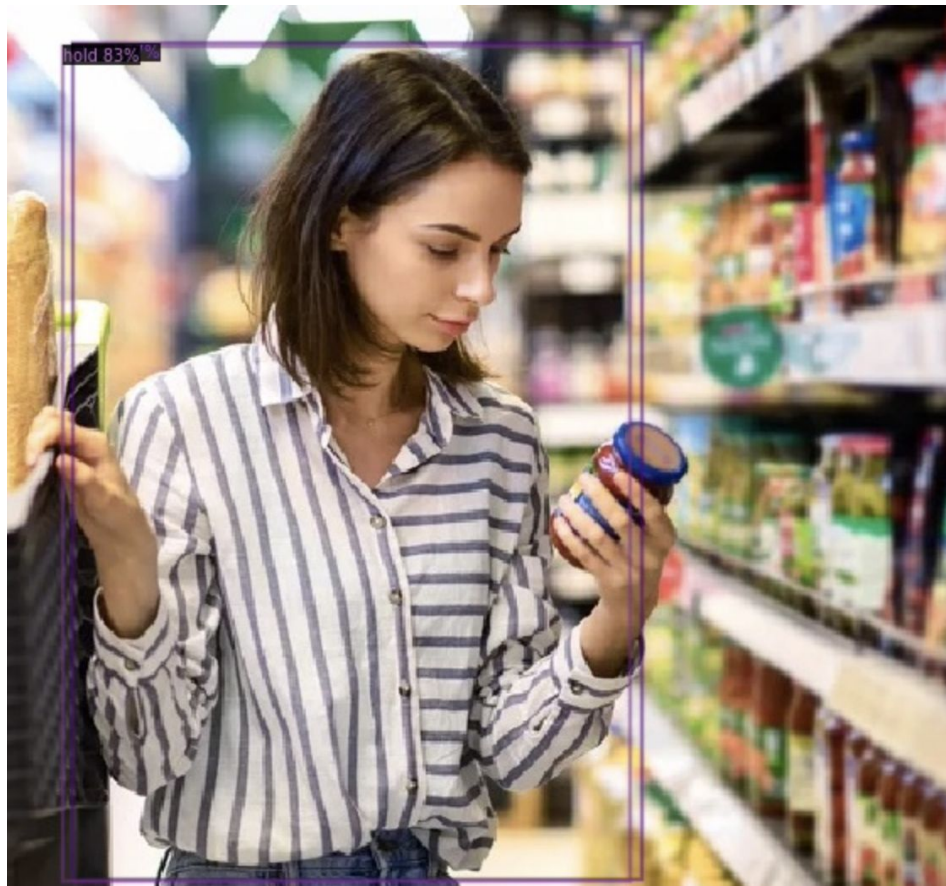


# Sample Results

# Branch A



# Branch B





# Branch C





# Branch D





Thank you! Questions?