Computer Vision
Assignment 4

Object Detection Models

Prepared By:

Eman Mohamed Salah El-Din	**6072**
Nael Mostafa Mohamed	**6099**
Ahmad Abdelrahman Marouf	**6543**

Contents

1. Introduction

In this assignment, we worked on COCO-2017 dataset, which is a large-scale object detection, segmentation, and captioning dataset.
We were required to run 3 different object detection models on the COCO-2017 dataset and compare the results. Then verify model generalization using VOC-2012 dataset.

2. Different Models Tested

2.1 Model 1: Faster R-CNN (using resnet50)

- Faster R-CNN is a state-of-the-art object detection model. It is based on the R-CNN model, **which is a two-stage object detection model**.
- The first stage of R-CNN is a **selective search algorithm**, which generates a large number of region proposals.
- The second stage is a **classifier that classifies each region proposal into one of the 20 object classes**.
- The classifier is a **linear SVM trained on the features extracted from the region proposals**.
- The R-CNN model is slow because it requires running the selective search algorithm on each image.
- Faster R-CNN addresses this problem by using a deep convolutional network to generate region proposals.
- The deep network is a fully convolutional network that takes an image as input and outputs a set of region proposals.
- The region proposals are then classified by the same linear SVM classifier used in R-CNN.
- The Faster R-CNN model is **faster than R-CNN because it only requires running the deep network once per image**.
- The deep network is trained end-to-end, which means that it is trained jointly with the linear SVM classifier.

2.2 Model 2: FCOS (Fully Convolutional One-Stage)

- FCOS is a **one-stage object detection model**.
- Computes **per-pixel prediction** in a fully convolutional manner: the number of detection prediction equals the size of feature maps spatially.
- It is based on the FPN model, which is a feature pyramid network.
- The FPN model is a deep convolutional network that takes an image as input and outputs a set of feature maps.

- The **feature maps are used to generate region proposals**.
- The region proposals are then classified by a linear SVM classifier.
- The FPN model is trained end-to-end, which means that it is trained jointly with the linear SVM classifier.
- The FPN architecture helps in high BPR (Best Possible Recall) and better prediction recall.
- FCOS is **faster than Faster R-CNN** because it is a one-stage model.
- FCOS is also **more accurate than Faster R-CNN because it uses a feature pyramid network to generate region proposals**.

2.3 Model 3: SSD (Single Shot Detector)

- SSD is a **one-stage object detection model**.
- SSD is **faster than Faster R-CNN** because it is a one-stage model.
- SSD has two components: a **backbone** model and **SSD head**.
- **Backbone** model usually is a pre-trained image classification network as a feature extractor.
- The **SSD head** is just one or more convolutional layers added to this backbone and the outputs are interpreted as the bounding boxes and classes of objects in the spatial location of the final layer's activations.
- SSD divides the image using a **grid** and have each grid cell be responsible for detecting objects in that region of the image.
- Each grid cell in SSD can be assigned with multiple **anchor/prior boxes**. These anchor boxes are pre-defined and each one is responsible for a size and shape within a grid cell.
- SSD uses a matching phase while training, to match the appropriate anchor box with the bounding boxes of each ground truth object within an image.

2.4 References

- Faster R-CNN
- FCOS
- SSD

3. Strengths and Weaknesses of Selected Models

Model	Strengths	Weaknesses
Faster R-CNN	- Faster than Fast R-CNN and R-CNN (still slower than single stage models) - Simplified structure and robustness	- Slower than FCOS and SSD as two-stage detection pipeline is too slow - Network may take a lot of time until reaching convergence
FCOS	- Avoids the complicated computation - Avoids all hyper-parameters related to anchor boxes	
SSD	- Lower localization error comparing with R-CNN - Makes more predictions, better coverage on location, scale, and aspect ratios - Eliminates region proposal approach	- Performs worse than Faster R-CNN for small-scale objects - Higher classification errors due to using same boundary box to make multiple class predictions - High cost

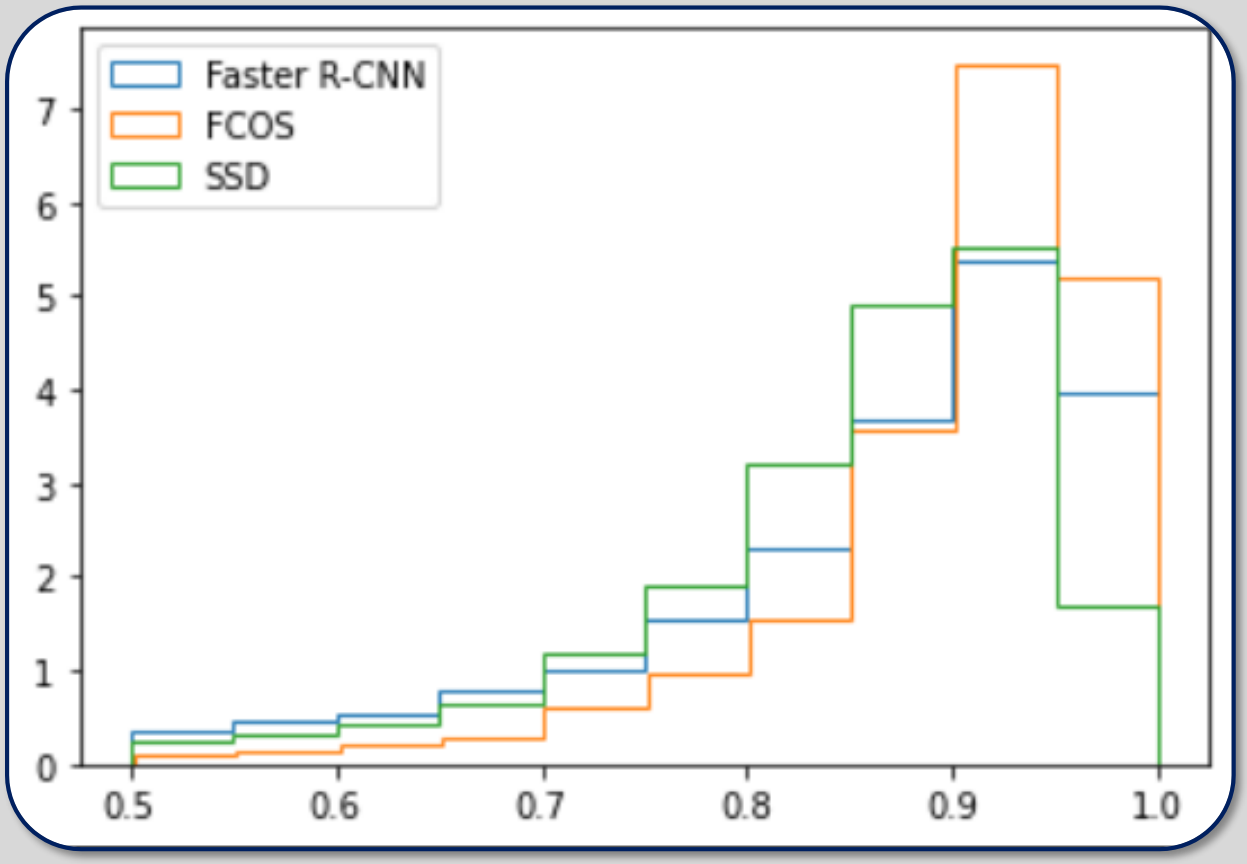
4. Comparing Evaluation Results

4.1 COCO-17 Validation set:

Inference Results:

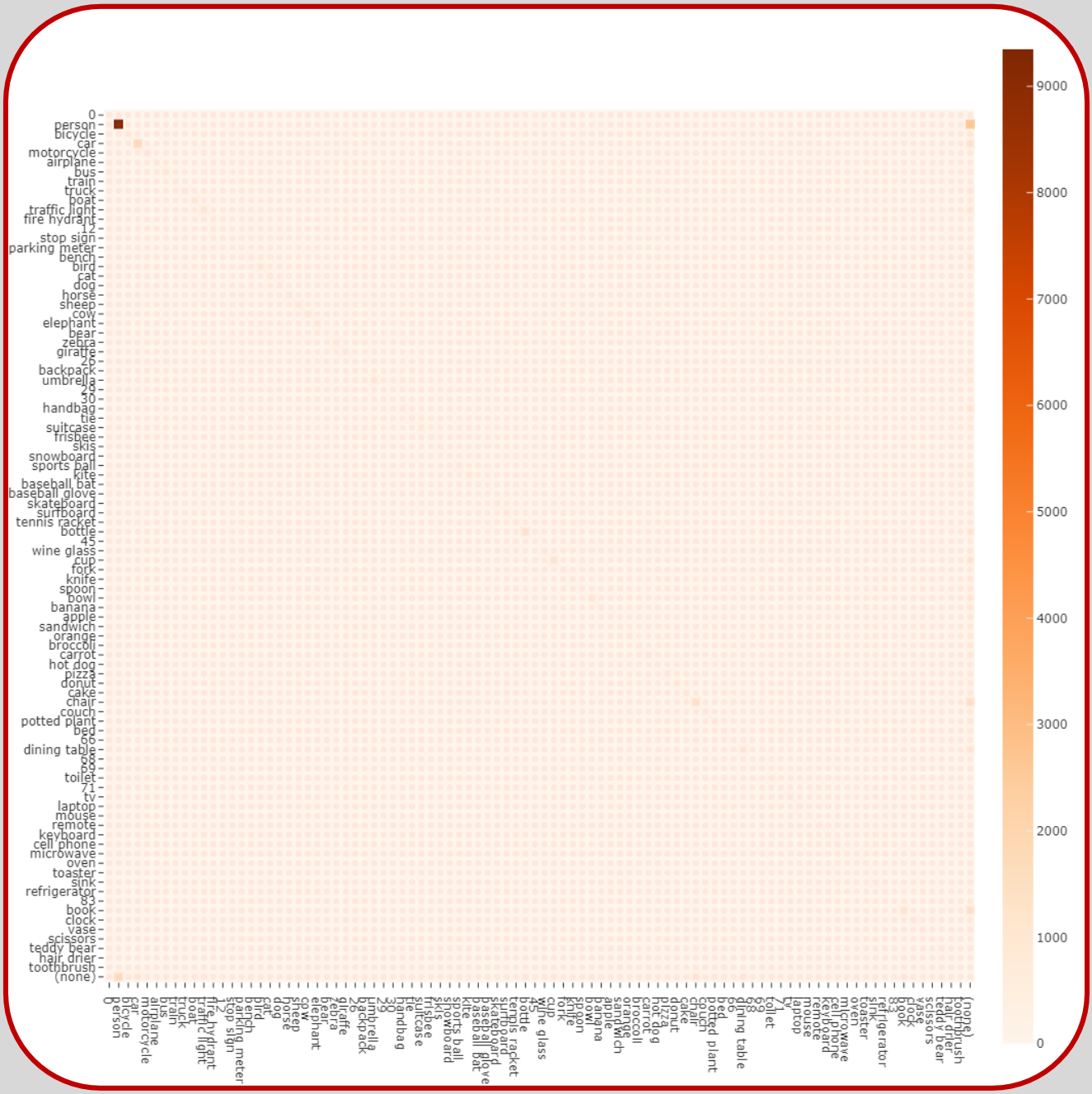
Model	mAP	FPS (samples/sec)	Average IOU
Faster R-CNN	0.414	5.5	0.861
FCOS	0.151	6.7	0.897
SSD	0.144	6.2	0.850

IOU Distribution:

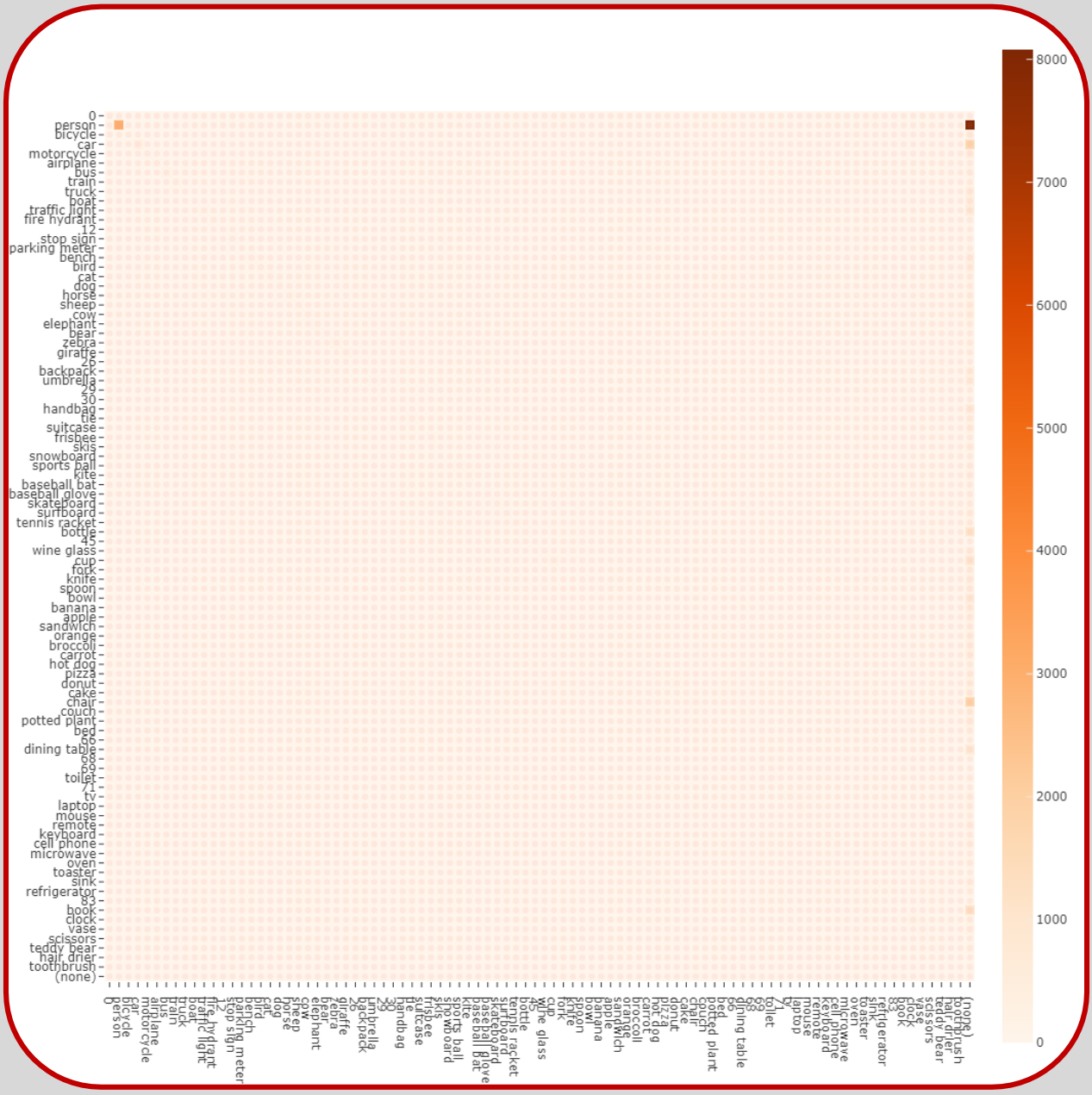


Confusion Matrices:

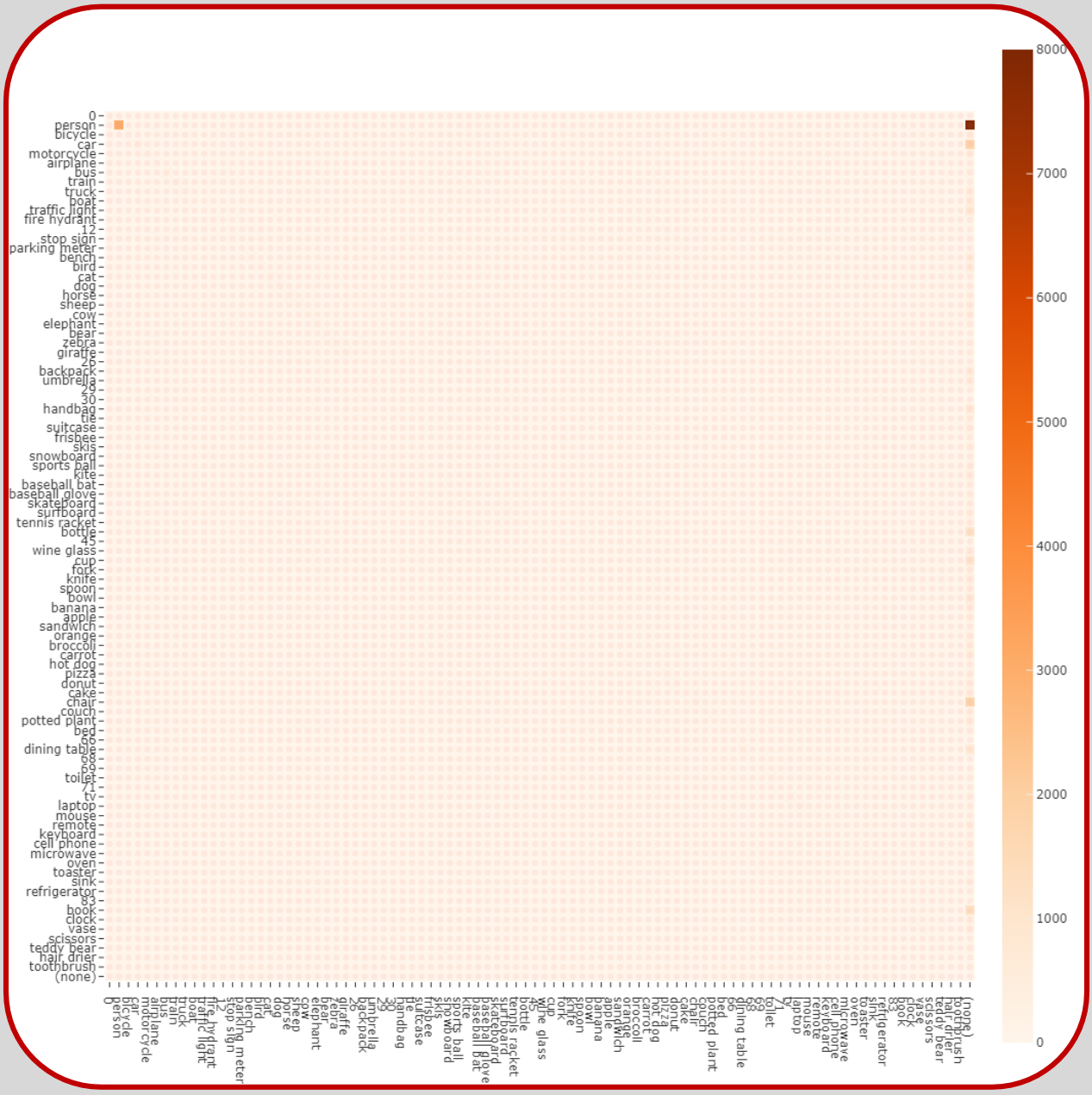
Faster R-CNN:



FCOS:


0
person
bicycle
car
motorcycle
airplane
bus
train
truck
boat
traffic light
fire hydrant
12
stop sign
parking meter
bench
bird
cat
dog
horse
sheep
cow
elephant
bear
zebra
giraffe
26
backpack
umbrella
29
30
handbag
tie
suitcase
frisbee
skis
snowboard
sports ball
kite
baseball bat
baseball glove
skateboard
surfboard
tennis racket
bottle
45
wine glass
cup
fork
knife
spoon
bowl
banana
apple
sandwich
orange
broccoli
carrot
hot dog
pizza
donut
cake
chair
couch
potted plant
bed
66
dining table
68
69
toilet
71
tv
laptop
mouse
remote
keyboard
cell phone
microwave
oven
toaster
sink
refrigerator
83
book
clock
vase
scissors
teddy bear
hair drier
toothbrush
(none)
8000
7000
6000
5000
4000
3000
2000
1000
0

SSD:

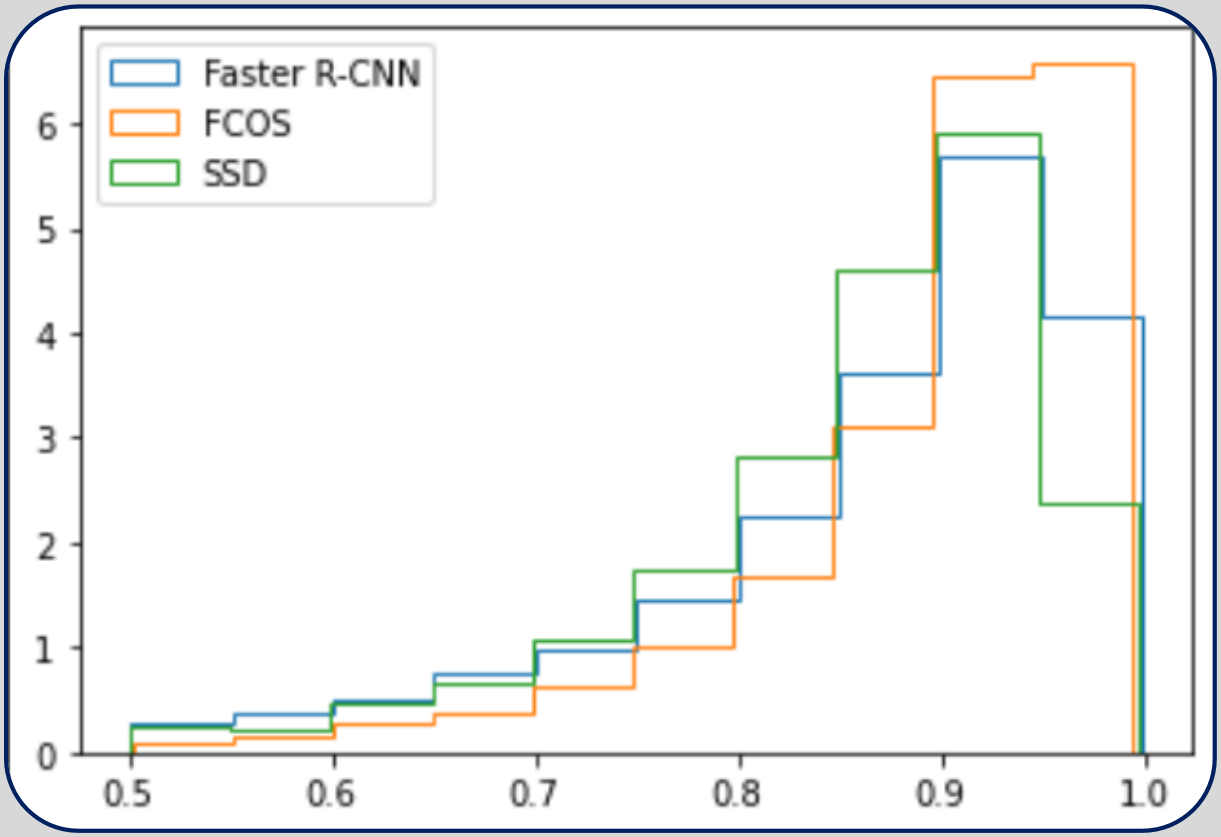

0
person
bicycle
car
motorcycle
airplane
bus
train
truck
boat
traffic light
fire hydrant
12
stop sign
parking meter
bench
bird
cat
dog
horse
sheep
cow
elephant
bear
zebra
giraffe
26
backpack
umbrella
29
30
handbag
tie
suitcase
frisbee
skis
snowboard
sports ball
kite
baseball bat
baseball glove
skateboard
surfboard
tennis racket
bottle
45
wine glass
cup
fork
knife
spoon
bowl
banana
apple
sandwich
orange
broccoli
carrot
hot dog
pizza
donut
cake
chair
couch
potted plant
bed
66
dining table
68
69
toilet
71
tv
laptop
mouse
remote
keyboard
cell phone
microwave
oven
toaster
sink
refrigerator
83
book
clock
vase
scissors
teddy bear
hair drier
toothbrush
(none)
8000
7000
6000
5000
4000
3000
2000
1000
0

4.2 VOC-2012 Validation set:

Inference Results:

Model	mAP	FPS (samples/sec)	Average IOU
Faster R-CNN	0.580	6.4	0.865
FCOS	0.237	7.9	0.894
SSD	0.318	7.6	0.855

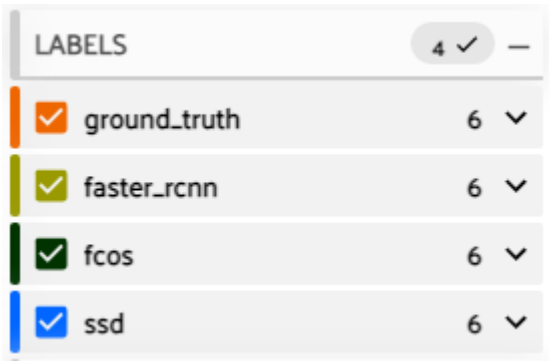
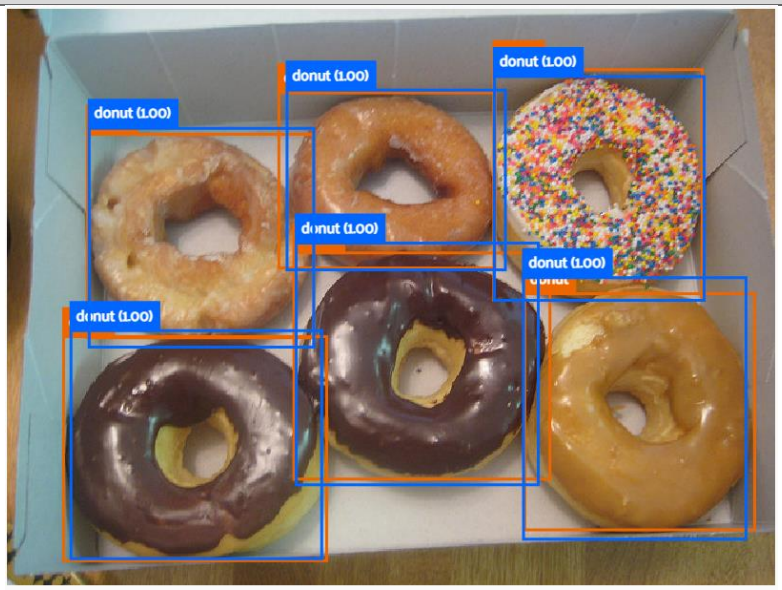
IOU Distribution:



5. Success and Failure Cases

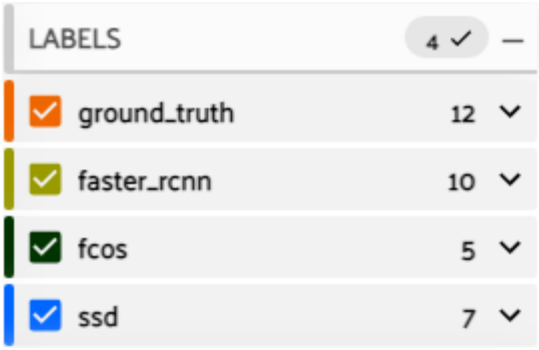
5.1 General Success Cases:

1)

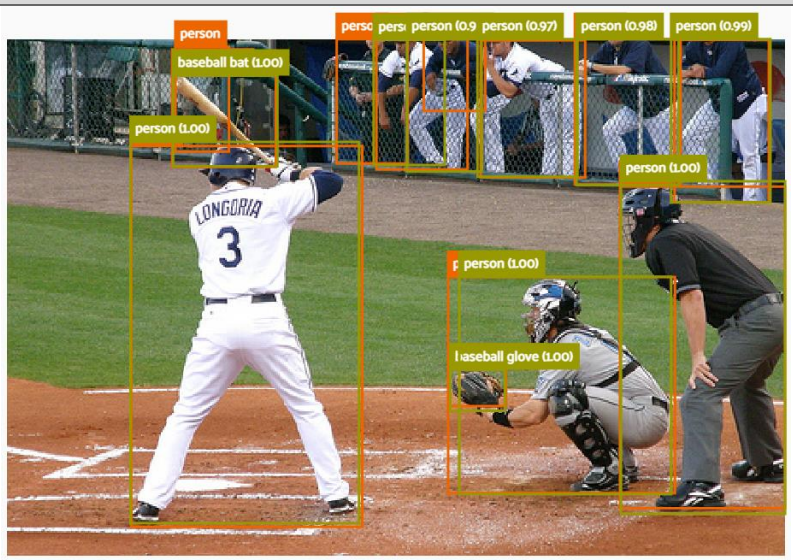
Labels	Faster R-CNN
	
FCOS	**SSD**
	

2)

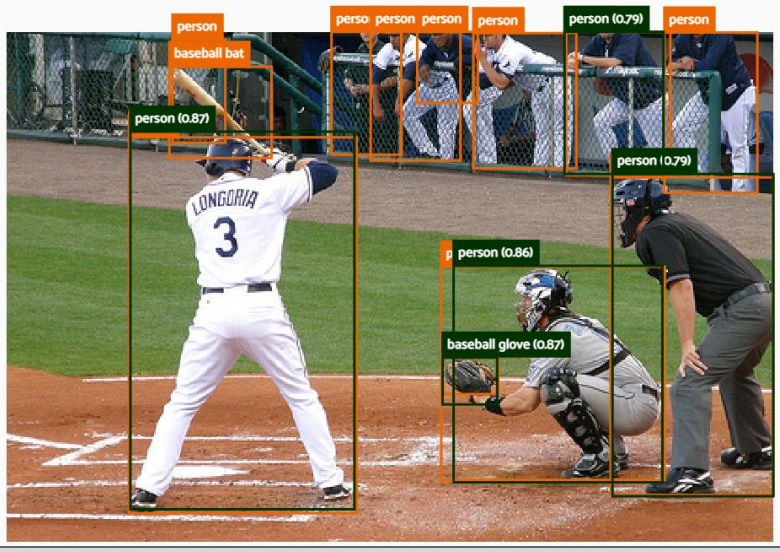
Labels



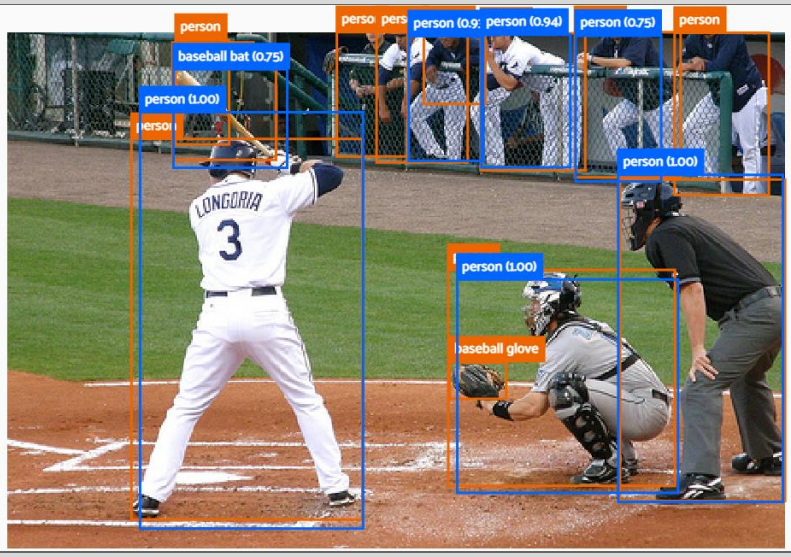
Faster R-CNN



FCOS



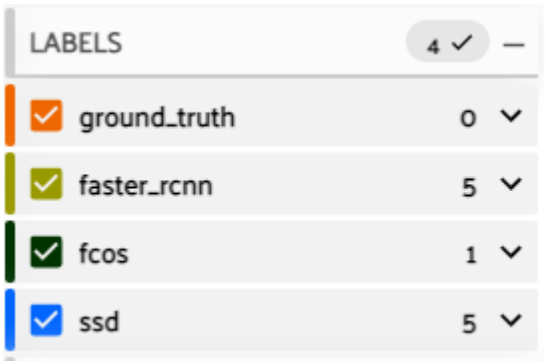
SSD



5.2 General Failure Cases:

1) Objects not in dataset

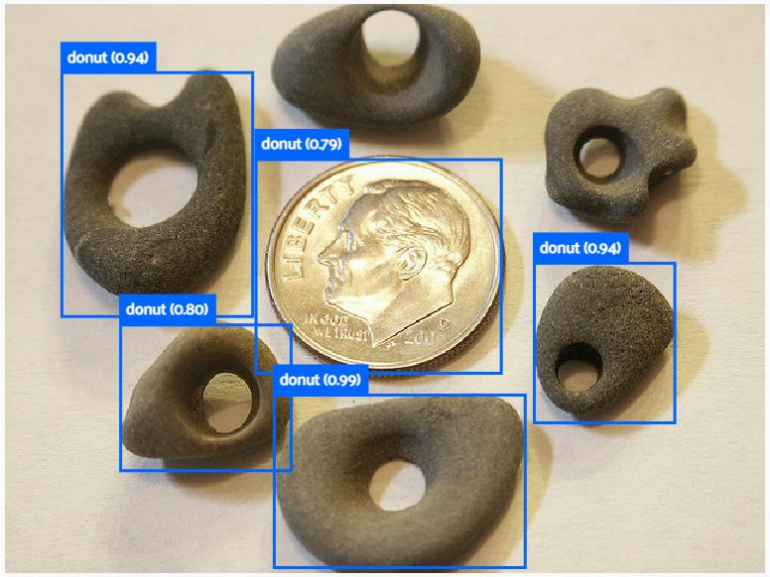
Labels



Faster R-CNN

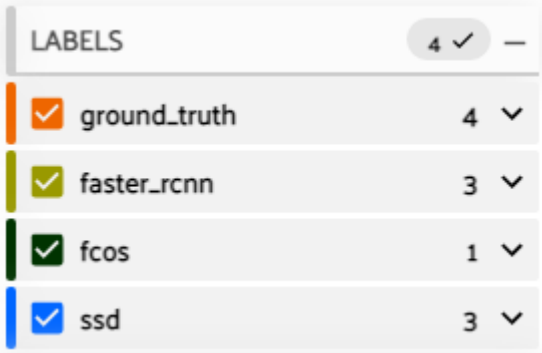
FCOS

SSD



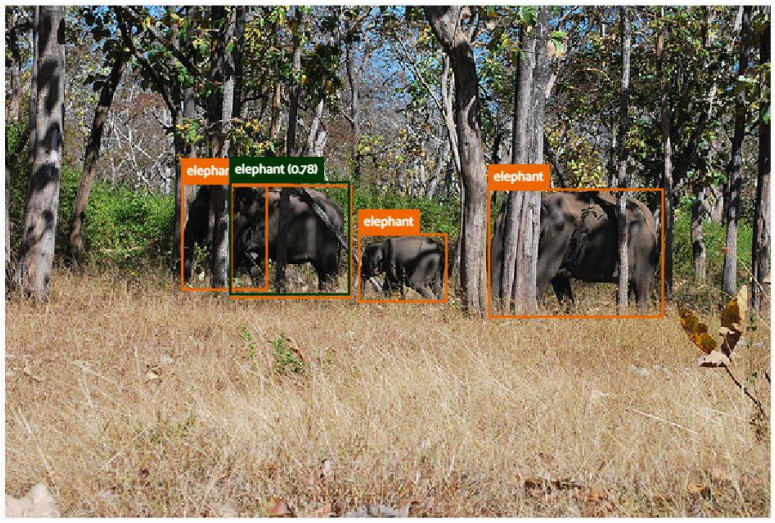
2) Objects Stacked and Occluded + Lighting Differences

Labels



Faster R-CNN

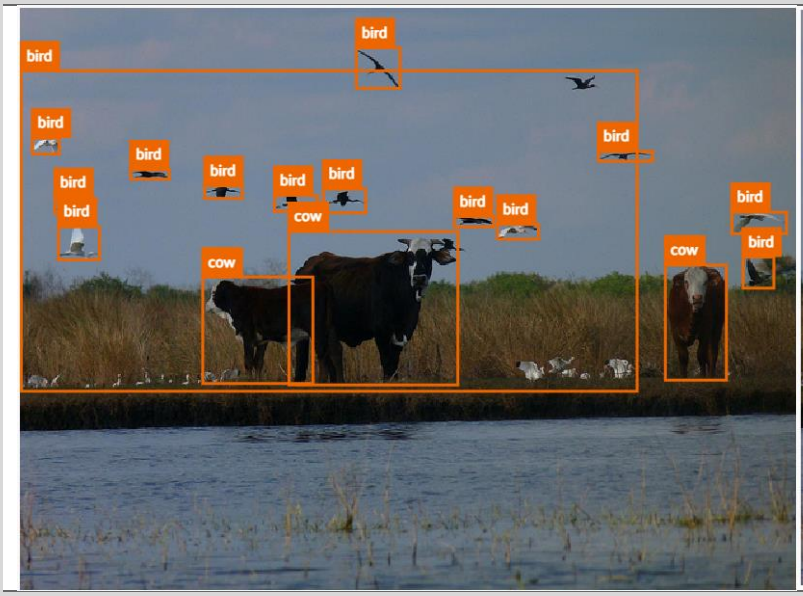
FCOS

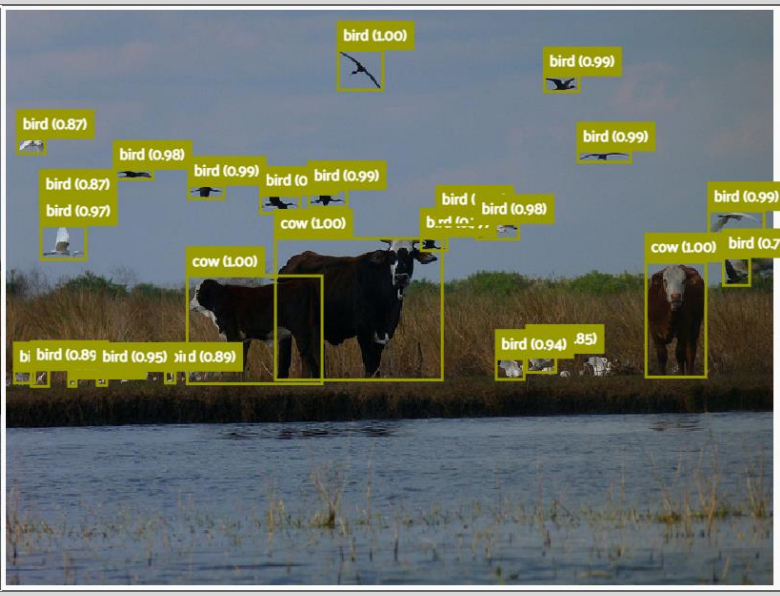


SSD

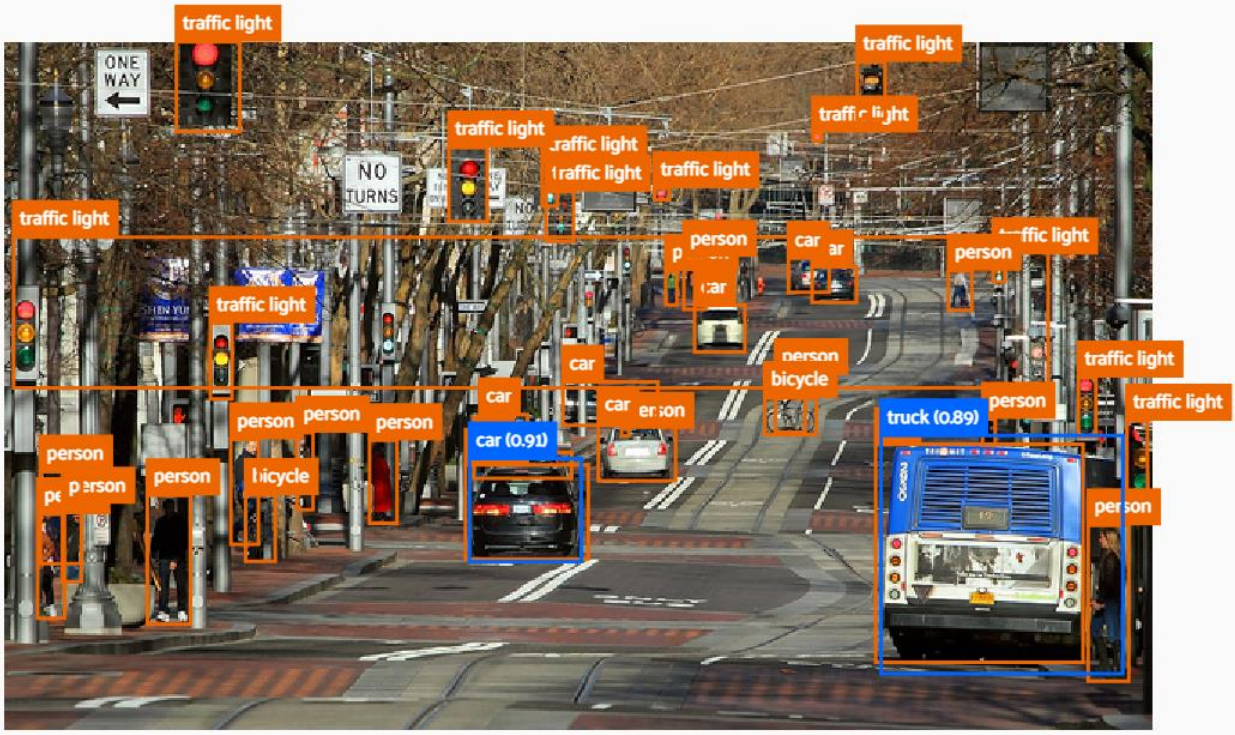
5.3 Model Success/Failure Cases:

1) Faster R-CNN in some cases detects can correctly detect smaller objects that are not labelled





2) SSD fails to detect objects that are small compared to the image scale



6. Analysis of Results

- As expected, both FCOS and SSD were faster than Faster R-CNN due to their single stage architecture.
- Region based detectors like Faster R-CNN demonstrate an accuracy advantage if real-time speed is not needed.
- SSD failed to detect in most cases where the object size was small.
- Performance of models was generally on VOC-2012 dataset as images contain less objects and the images have a lower resolution.
- The speed of SSD could have been higher if it were implemented with ResNet backbone instead of VGG16 (the pretrained model available in the Pretrained Pytorch Models was sdd-vgg16).

7. Visualizing Feature Maps

We extracted the convolution layers from the Faster R-CNN model then passed a sample image through them and visualized the output of each layer to help understand the levels of the features as the network becomes deeper.