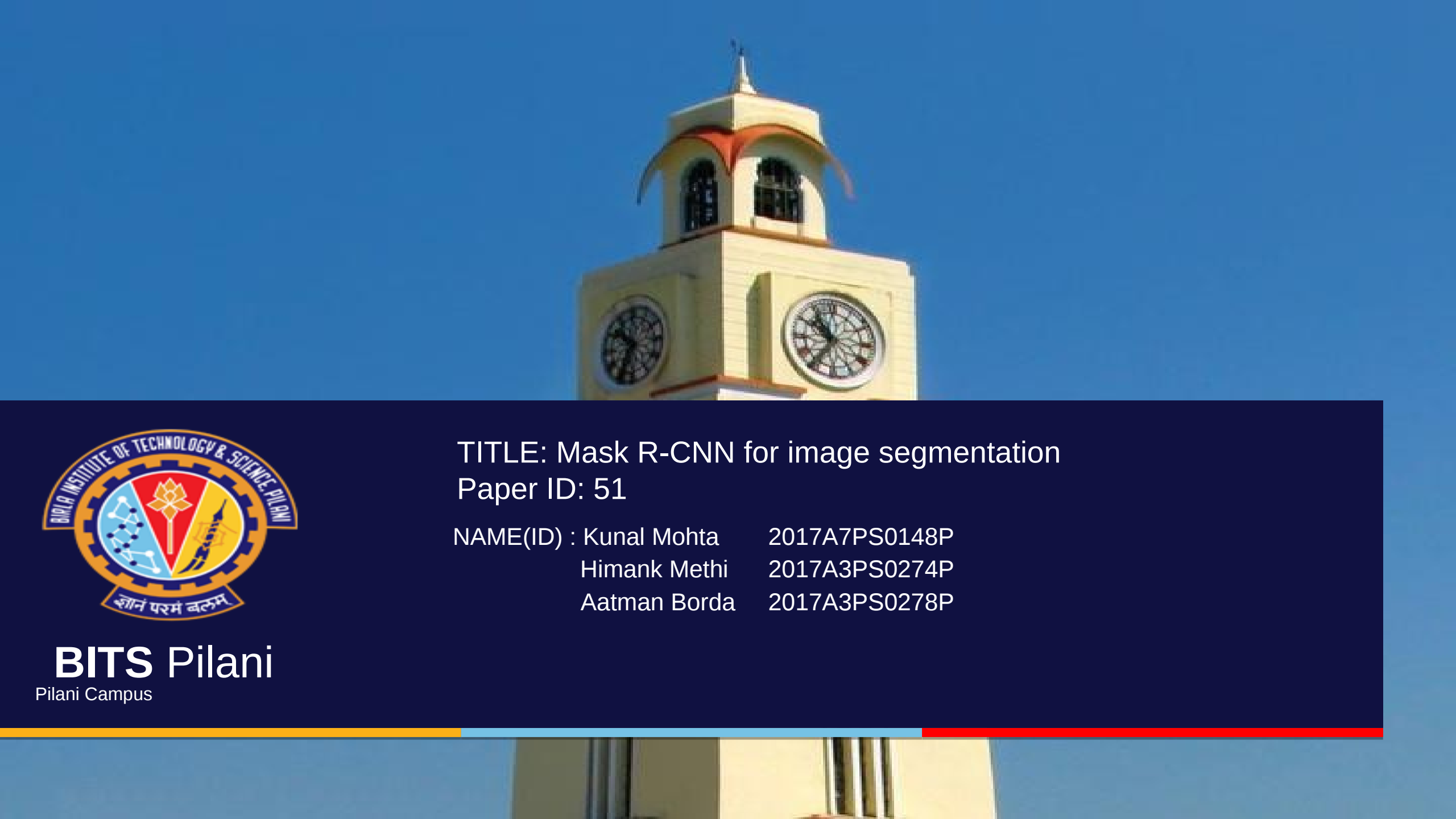


TITLE: Mask R-CNN for image segmentation
Paper ID: 51

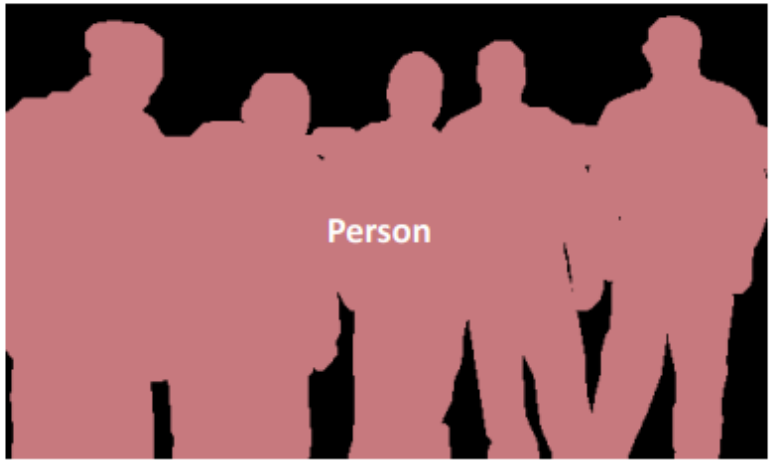
NAME(ID) : Kunal Mohta 2017A7PS0148P
Himank Methi 2017A3PS0274P
Aatman Borda 2017A3PS0278P

BITS Pilani
Pilani Campus

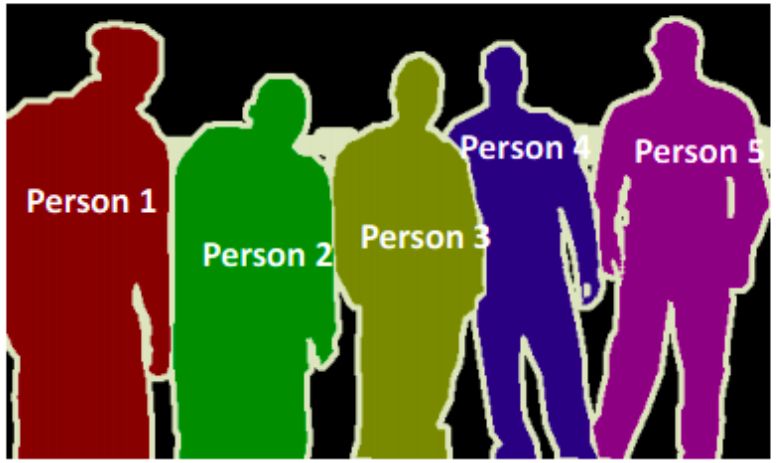
Problem

3 tracks of challenges in computer vision

Object Detection



Semantic Segmentation

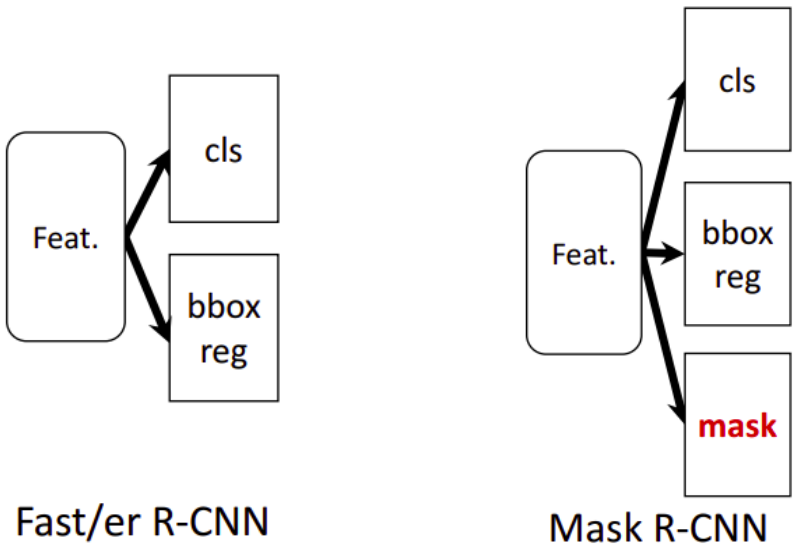


Instance Segmentation

Mask R-CNN

Problem

- **Instance segmentation :-** Identification of boundaries around object at **pixel** level.
- It is difficult because it combines 2 classical tasks - object detection & semantic segmentation.
- Mask R-CNN is a leading model for the purpose of instance segmentation.
- It is an extension of Faster R-CNN, which is a fast and accurate model for object detection.



Dataset - Coco 2014

- Standard dataset for object detection and segmentation.
- Size - 82,783 train images, 40,504 validation images.
- Clean dataset with annotations (label, bounding box, mask) available on official website (http://cocodataset.org/#home).
- Annotations given in JSON format:-

```
{
    "info": {...},
    "licenses": [...],
    "images": [...], // relevant
    "annotations": [...], // relevant
    "categories": [...]
}
```

Structure of annotations

Image metadata

```
{
    ...
    "file_name":
    "COCO_val2014_000000*.jpg",
    "height": ...,
    "width": ...,
    "id": 514508
    ...
}
```

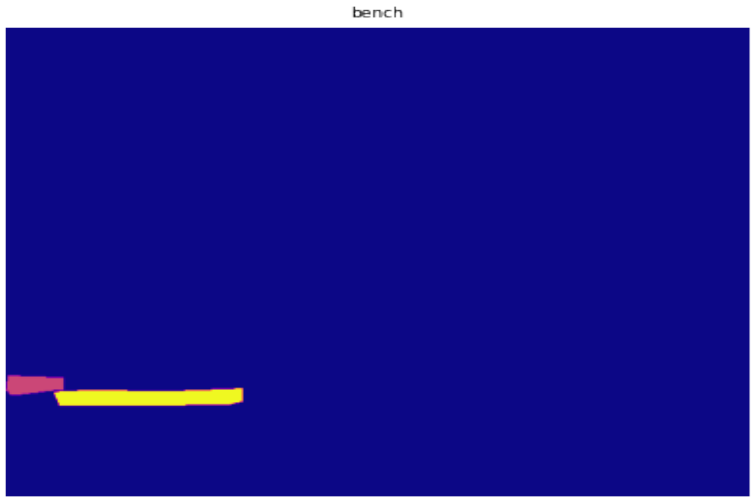
Bounding box & Mask

```
{
    "segmentation":[
        [coordinates of polygon
        representing mask]
    ],
    "image_id": ...,
    "bbox": [coords of bounding box] //**
}
```

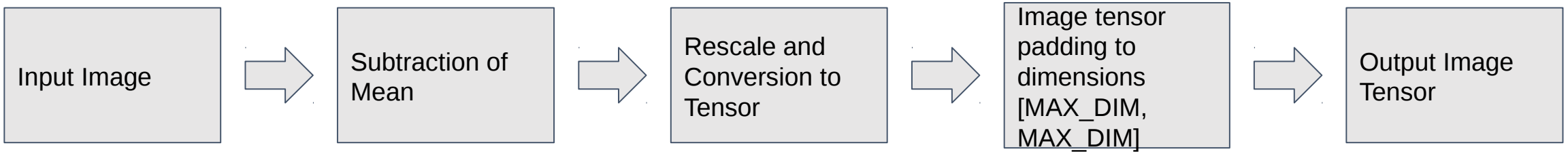
**NOTE: not using coordinates of bbox from annotations because they are subject to change on image modification (cropping, augmenting) - instead calculating these from the extremities of the mask itself

Visual appearance of masks

H x W=480x640



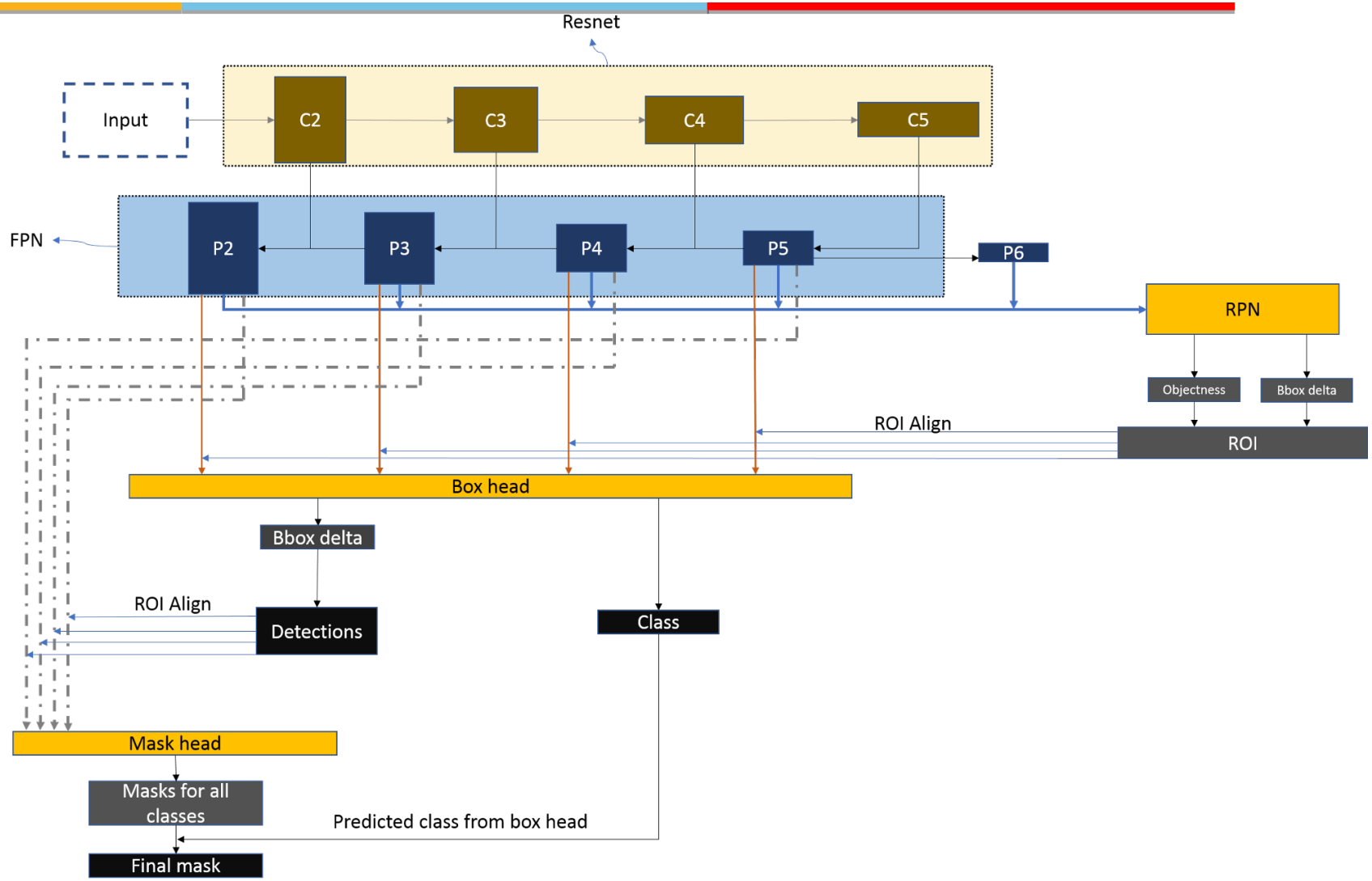
Pre-processing



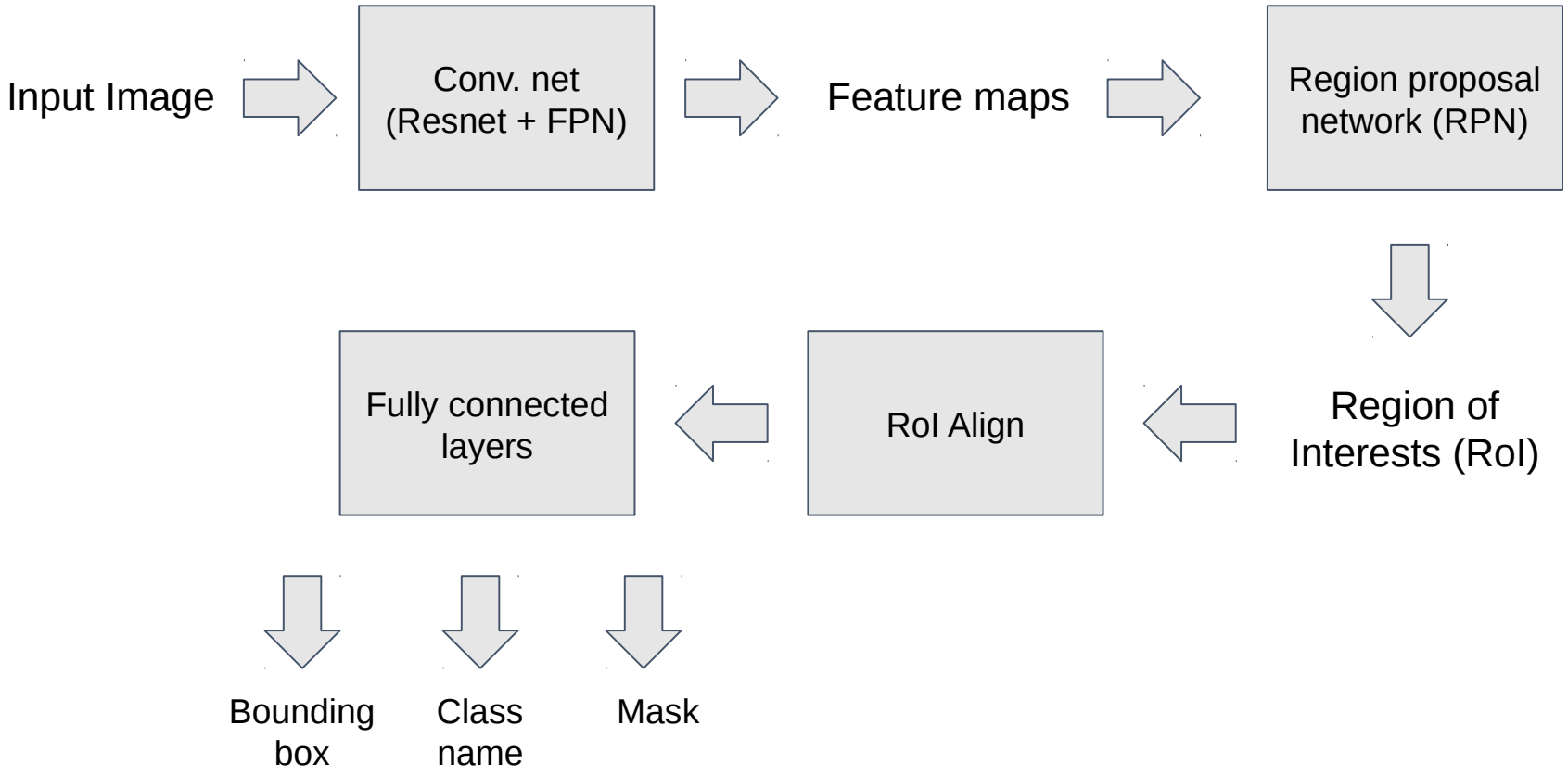
Pre-processing

- An important task before starting the training process was to store dataset information into appropriate data structures for model to work on.
- Dataset is provided to the model with following info:-
  - Image metadata - id, shape (height x width)
  - Ground truth bounding box, mask and detected class
  - RPN ground truths - for training of RPN

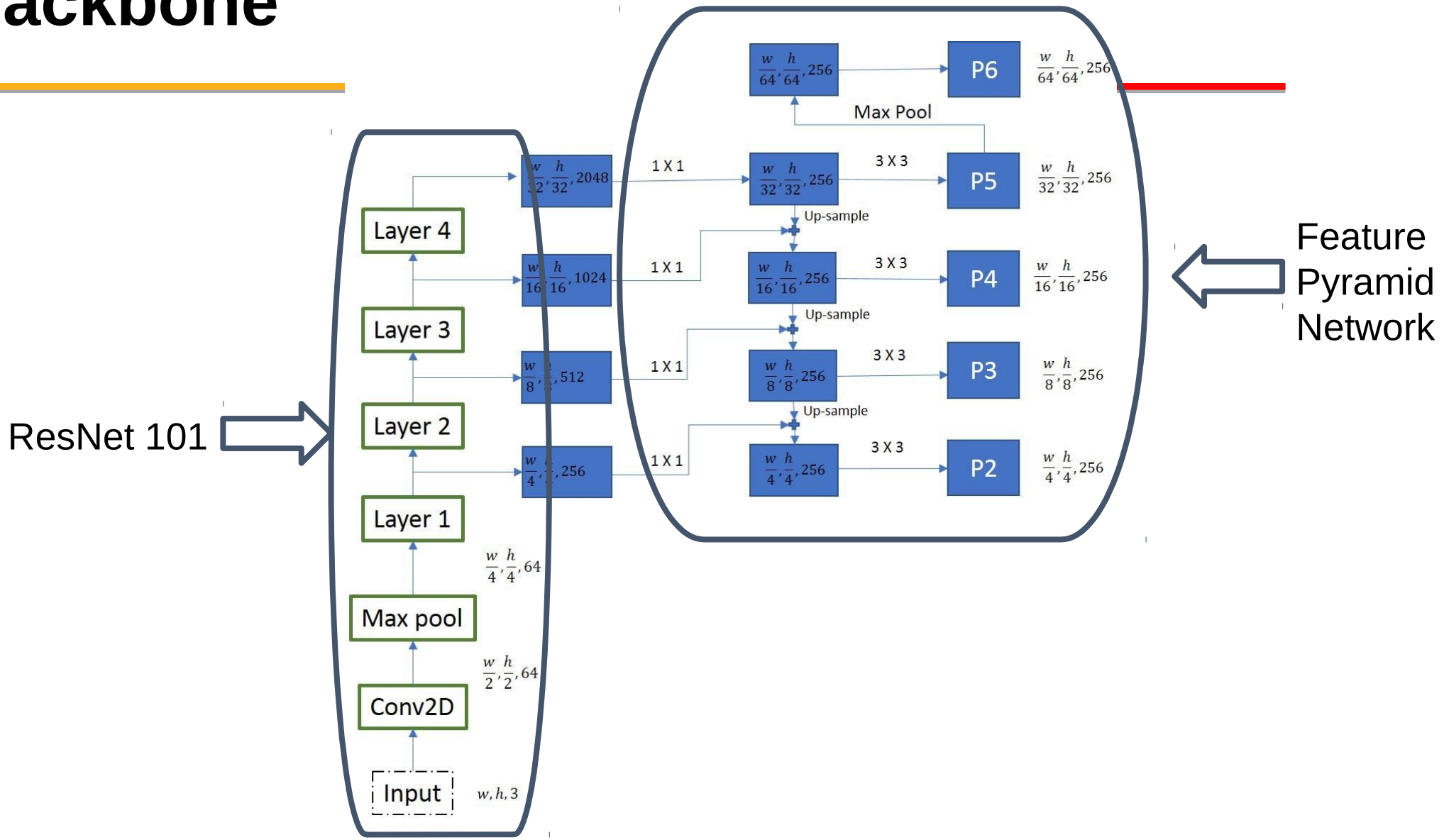
Complete structure



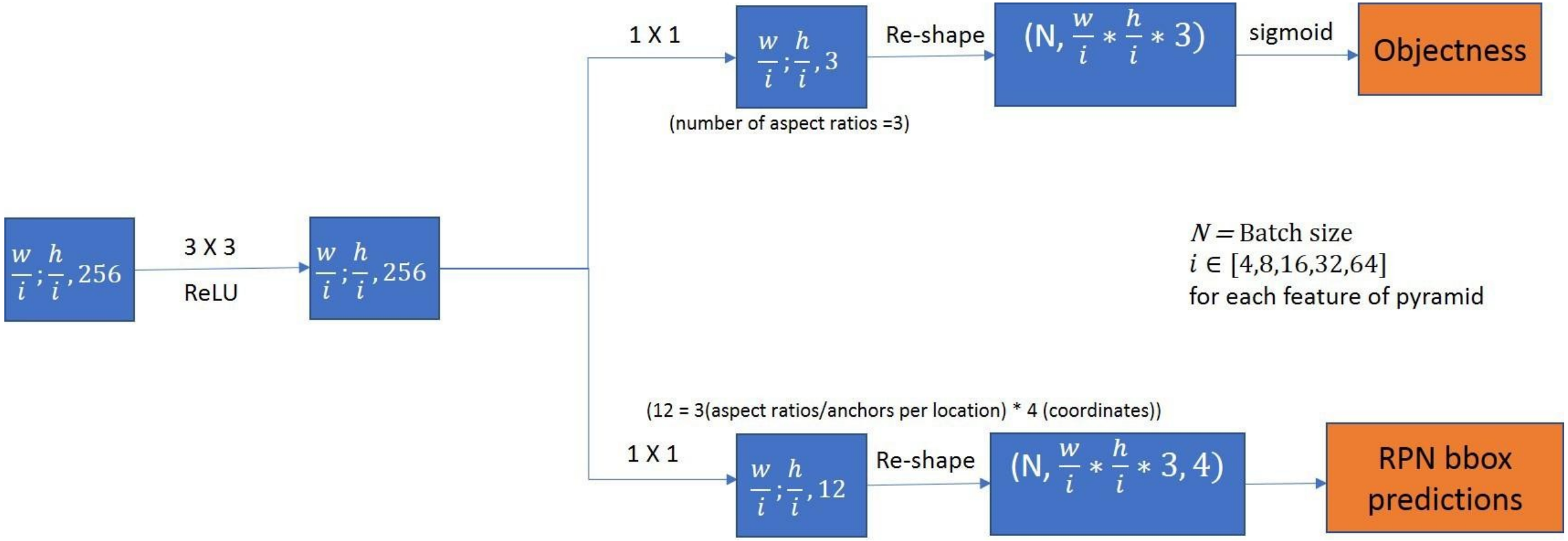
Steps (in a nutshell)



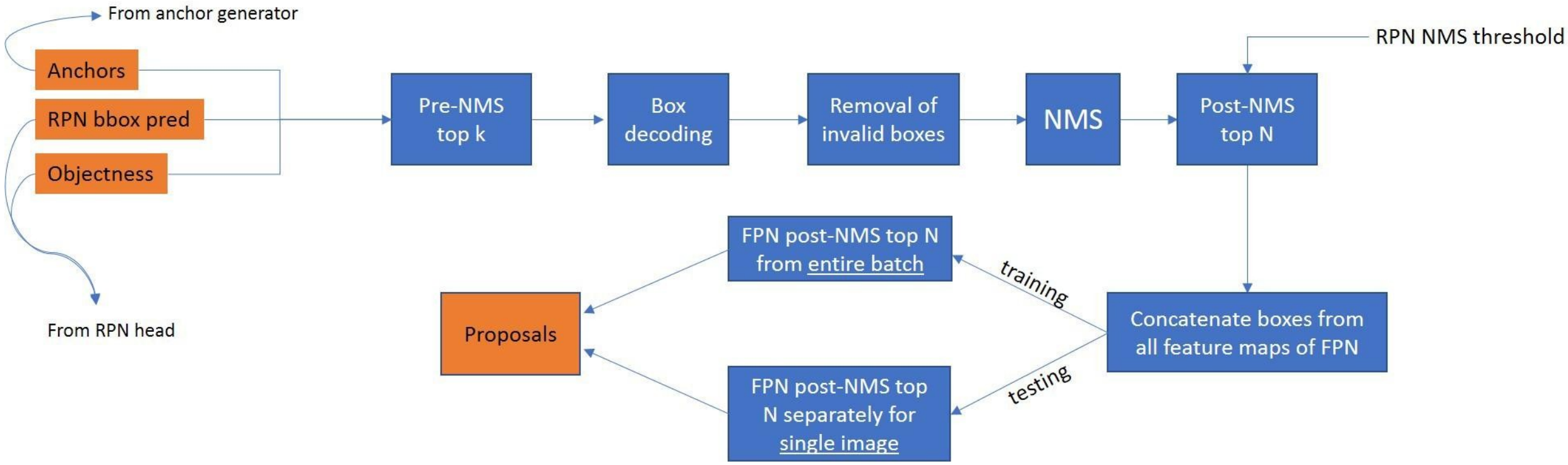
Backbone


ResNet 101
Layer 4
Layer 3
Layer 2
Layer 1
Max pool
Conv2D
Input
$w, h, 3$
$\frac{w}{4}, \frac{h}{4}, 64$
$\frac{w}{2}, \frac{h}{2}, 64$
$\frac{w}{32}, \frac{h}{32}, 2048$
$\frac{w}{16}, \frac{h}{16}, 1024$
$\frac{w}{8}, \frac{h}{8}, 512$
$\frac{w}{4}, \frac{h}{4}, 256$
1 X 1
1 X 1
1 X 1
1 X 1
Up-sample
Up-sample
Up-sample
Max Pool
3 X 3
3 X 3
3 X 3
3 X 3
$\frac{w}{32}, \frac{h}{32}, 256$
$\frac{w}{16}, \frac{h}{16}, 256$
$\frac{w}{8}, \frac{h}{8}, 256$
$\frac{w}{4}, \frac{h}{4}, 256$
$\frac{w}{64}, \frac{h}{64}, 256$
P6
P5
P4
P3
P2
$\frac{w}{64}, \frac{h}{64}, 256$
$\frac{w}{32}, \frac{h}{32}, 256$
$\frac{w}{16}, \frac{h}{16}, 256$
$\frac{w}{8}, \frac{h}{8}, 256$
$\frac{w}{4}, \frac{h}{4}, 256$
Feature Pyramid Network

Region Proposal Network



RPN (conti.)


From anchor generator
Anchors
RPN bbox pred
Objectness
From RPN head
Pre-NMS top k
Box decoding
Removal of invalid boxes
NMS
Post-NMS top N
RPN NMS threshold
FPN post-NMS top N from entire batch
training
Concatenate boxes from all feature maps of FPN
Proposals
testing
FPN post-NMS top N separately for single image

Loss Functions

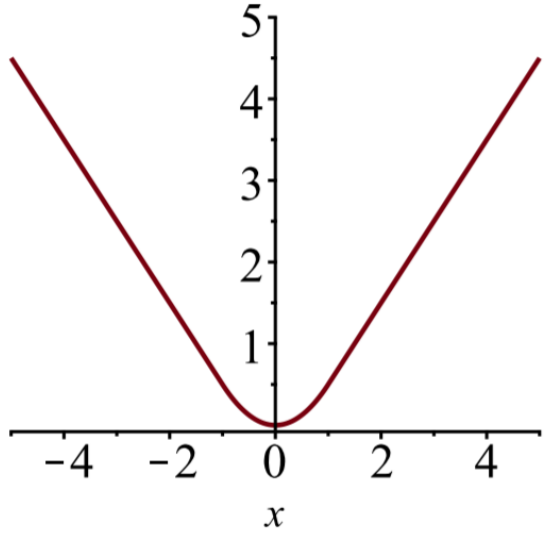
Smooth L1 loss: used for training bounding box heads of RPN and final Mask RCNN

Softmax Crossentropy: Used for multiclass classification

$$smooth_{L1_{plot}} := piecewise(\text{abs}(x) < 1, 0.5 \cdot x^2, \text{abs}(x) - 0.5)$$

$$\begin{cases} 0.5\,x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases}$$

→



Quantitative results

- Training for 160-200 epochs with learning rate varying between 0.001 and 0.0001 gave the following results:-
  - Mean Average precision (AP) for object detection => **0.288**
  - Mean Average precision (AP) for image segmentation => **0.275**
- For reference, results mentioned in the paper are:-
  - Mean Average precision (AP) for object detection => **0.382**
  - Mean Average precision (AP) for image segmentation => **0.357**

****NOTE**: Mean Average Precision is the mean of AP over all classes and IoU thresholds. It is considered as the primary metric for COCO challenges.

Quantitative results

Object Detection

```
 Average Precision  (AP) @[ IoU=0.50:0.50 | area=   all | maxDets=100 ] = 0.288
 Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ] = 0.288
 Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ] = -1.000
 Average Precision  (AP) @[ IoU=0.50:0.50 | area= small | maxDets=100 ] = 0.113
 Average Precision  (AP) @[ IoU=0.50:0.50 | area=medium | maxDets=100 ] = 0.321
 Average Precision  (AP) @[ IoU=0.50:0.50 | area= large | maxDets=100 ] = 0.392
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=   all | maxDets=  1 ] = 0.245
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=   all | maxDets= 10 ] = 0.329
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=   all | maxDets=100 ] = 0.332
 Average Recall     (AR) @[ IoU=0.50:0.50 | area= small | maxDets=100 ] = 0.120
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=medium | maxDets=100 ] = 0.360
 Average Recall     (AR) @[ IoU=0.50:0.50 | area= large | maxDets=100 ] = 0.467
Prediction time: 3711.2487921714783. Average 0.749444424913465/image
Total time:  3812.668703317642
```

Image Segmentation

```
 Average Precision  (AP) @[ IoU=0.50:0.50 | area=   all | maxDets=100 ] = 0.275
 Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ] = 0.275
 Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ] = -1.000
 Average Precision  (AP) @[ IoU=0.50:0.50 | area= small | maxDets=100 ] = 0.102
 Average Precision  (AP) @[ IoU=0.50:0.50 | area=medium | maxDets=100 ] = 0.309
 Average Precision  (AP) @[ IoU=0.50:0.50 | area= large | maxDets=100 ] = 0.381
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=   all | maxDets=  1 ] = 0.237
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=   all | maxDets= 10 ] = 0.319
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=   all | maxDets=100 ] = 0.321
 Average Recall     (AR) @[ IoU=0.50:0.50 | area= small | maxDets=100 ] = 0.111
 Average Recall     (AR) @[ IoU=0.50:0.50 | area=medium | maxDets=100 ] = 0.350
 Average Recall     (AR) @[ IoU=0.50:0.50 | area= large | maxDets=100 ] = 0.455
Prediction time: 3705.060626268387. Average 0.7481947952884465/image
Total time:  3804.186886072159
```

Qualitative results

Good prediction

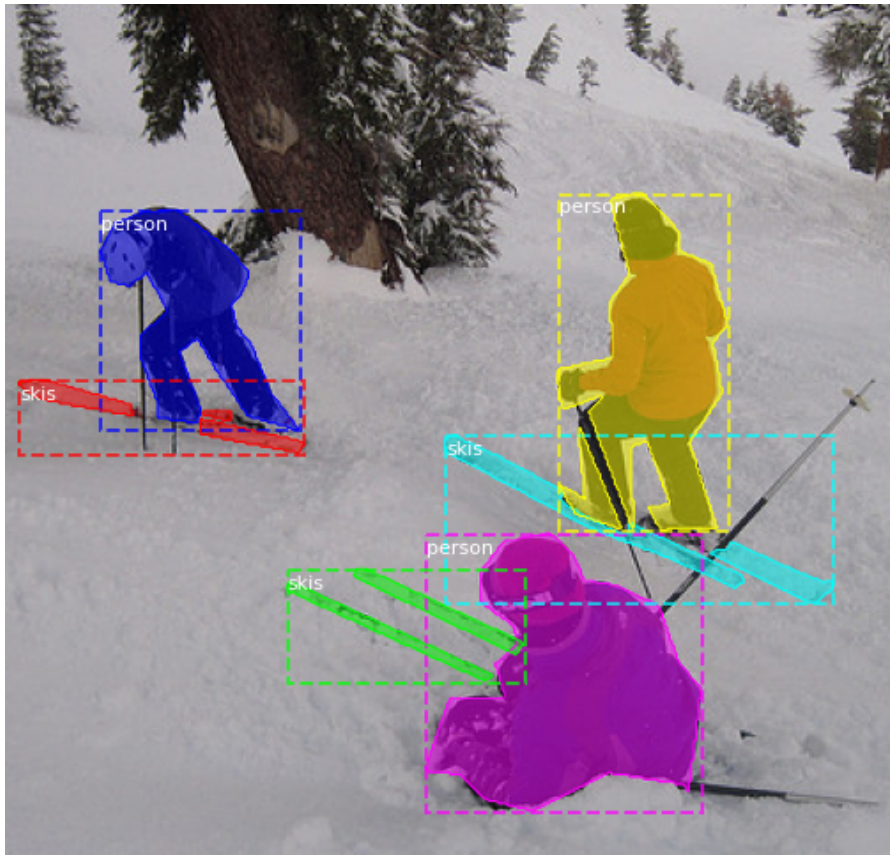
Our model

COCO Masks

Qualitative results

Bad prediction

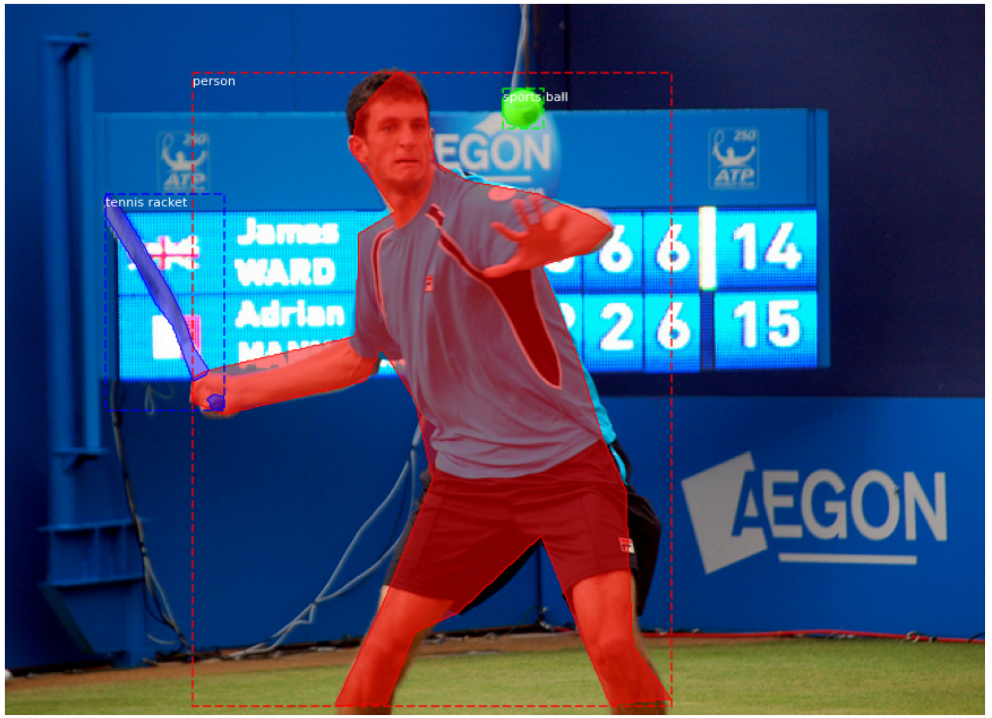
Our model



COCO Masks

Qualitative results

Missed objects

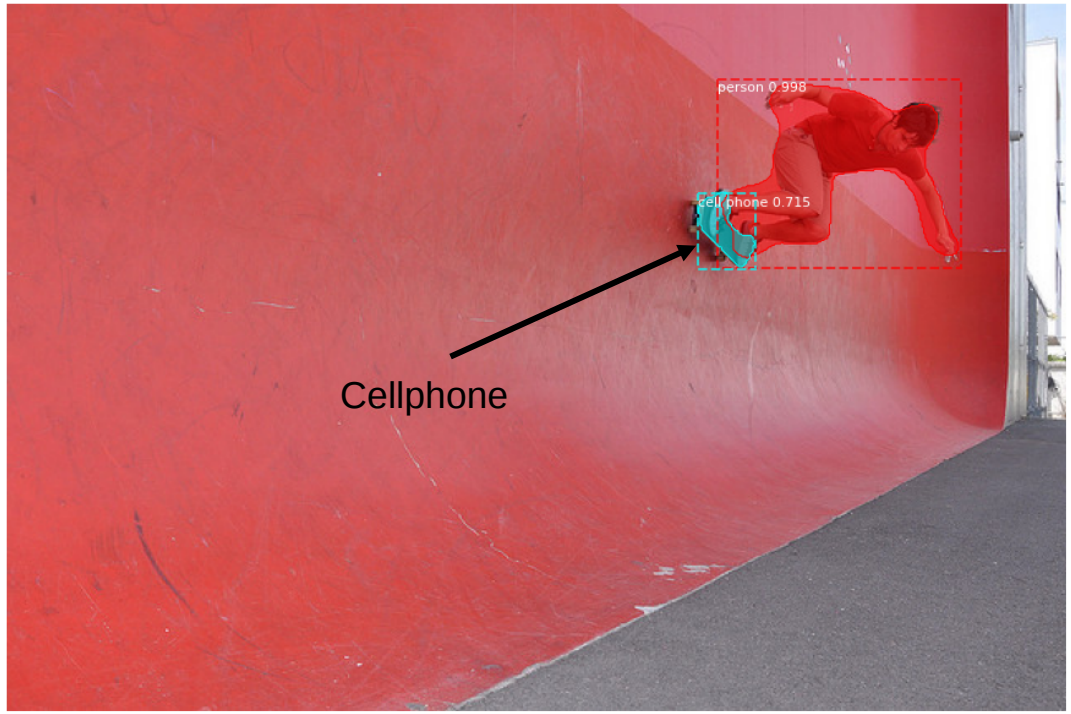


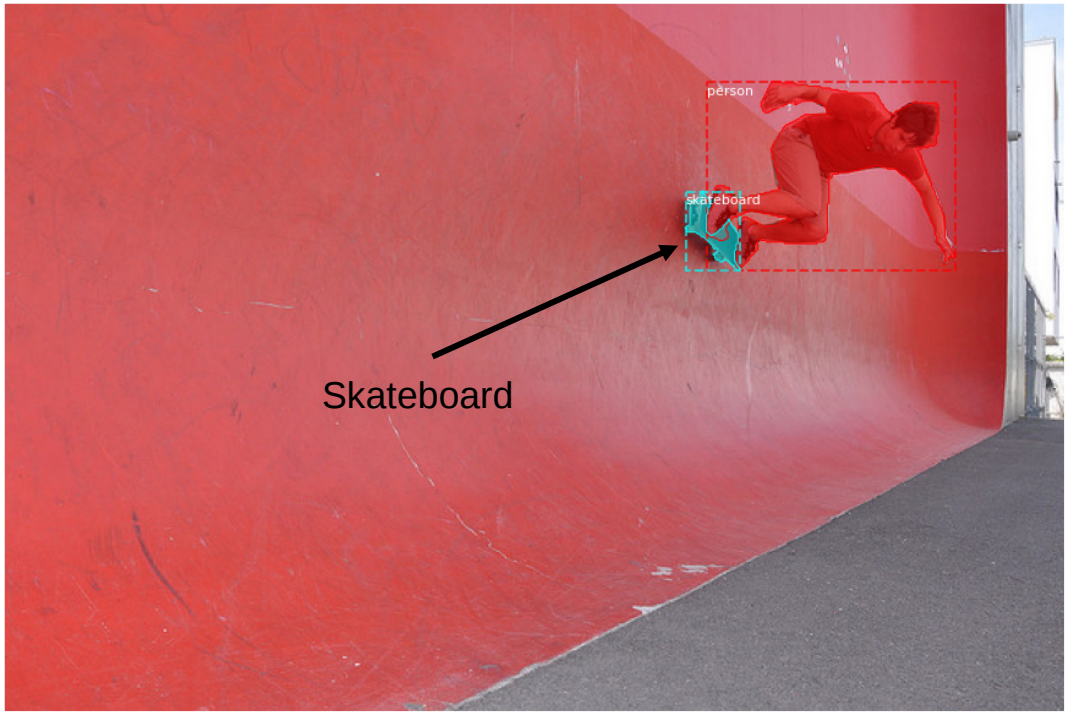
Our model

COCO Masks

Qualitative results

Wrong class prediction

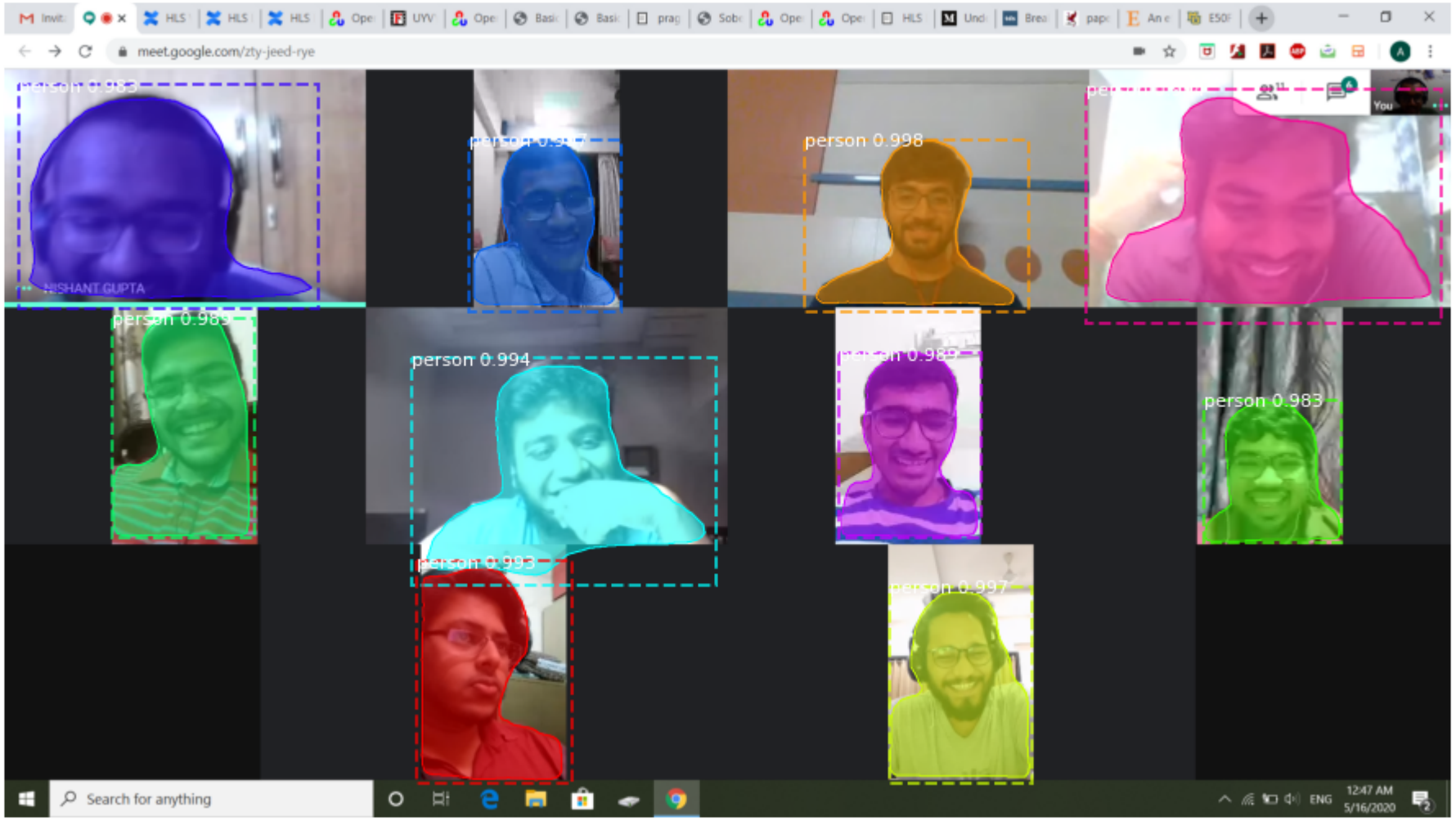




Our model

COCO Masks

Custom Image



Thank You