

# Masking bits and boxing entities to detect objects

Mask R-CNN: A fast deep neural network for Recognition and Localization of objects

Shivvrat Arya  
sxa180157@utdallas.edu  
May 1, 2013

# Motivation

## Uses of Instance Segmentation

Autonomous vehicles

Robotics

Open Street Map

Smart video surveillance

Tracking objects

Facial detection and recognition

Medical imaging

OCR, and many more

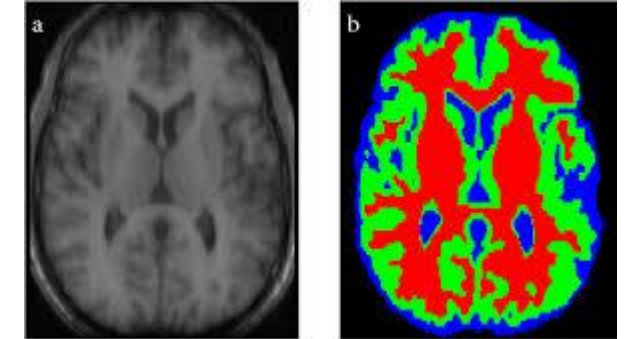
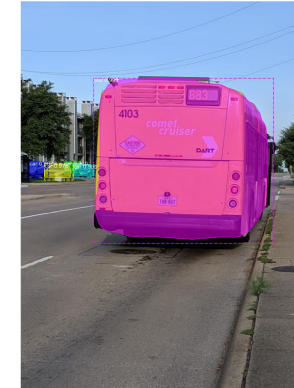


Image courtesy :- Daniel J. Withey and Zoltan J. Koles



Uses of Object Detection (Image Courtesy - Google)

# Motivation (continued)

## Common issues with current methods

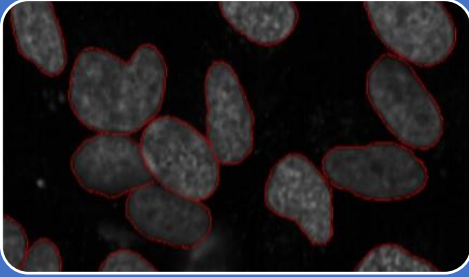
Current state of the art algorithms for Object detection output the anchor boxes for objects.

Bounding/anchor boxes give a general idea where the object is and not the exact pixels for the object.

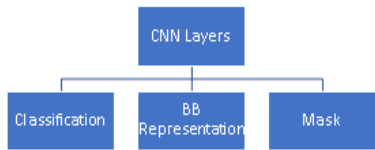
Decades of extensive research and still we don't have a state of the art algorithm for image segmentation, which precisely demonstrates how difficult the problem is.

In image segmentation we don't have the control over environment which makes it hard. Thus it is still an open area of research.

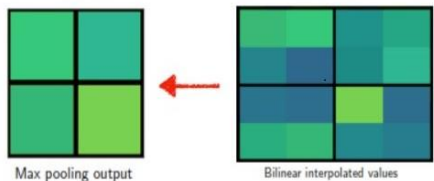
# Key Insights



Each pixel can be the part of the object or not. Thus the problem can be seen as classifying each pixel into a pre-defined class without differentiating object instances.

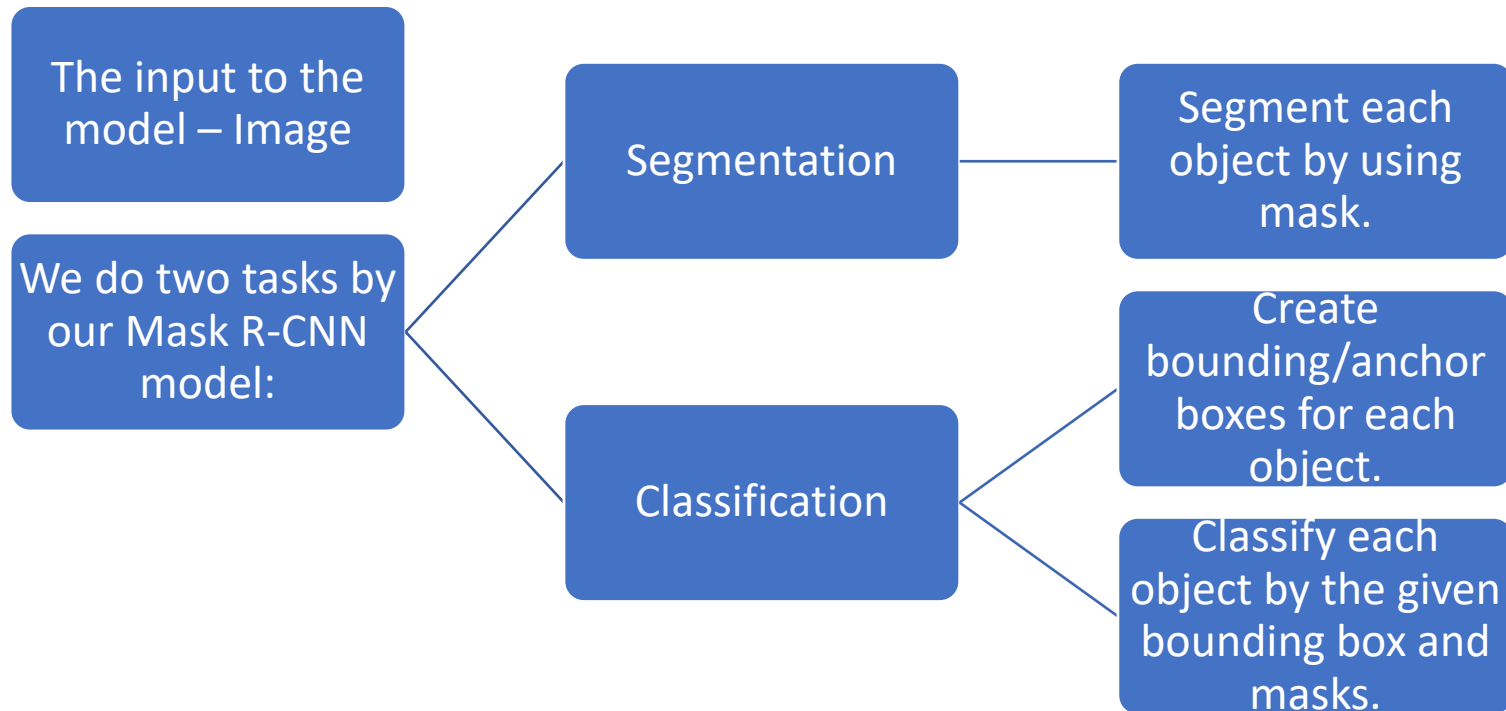


Added a branch for predicting an object mask in parallel and using it for classification.

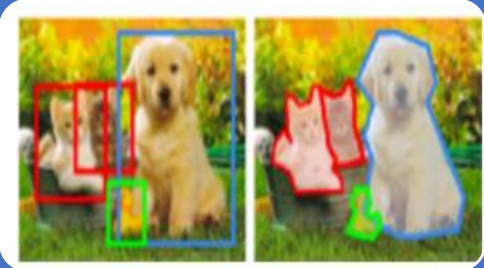


We have use RoIAlign layer, which helped to increase the accuracy by a relative amount of 10% to 50%.

# Image segmentation as a Classification Problem



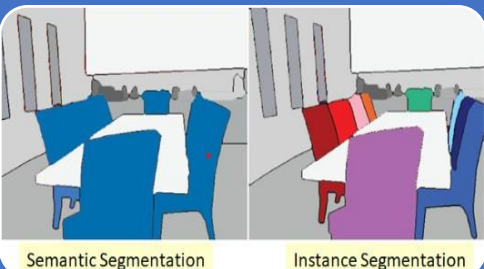
# What is Instance Segmentation?



Combines object detection and semantic segmentation.



Classifying objects, localizing by bounding/anchor boxes and classifying each pixel.



Each object of the same class/different class will be classified as a different colored object in the final output. (Helps to distinguish objects of same class)

# Proposed Approach for Mask R-CNN

Using ResNet as tail and body and using new head. (Backbone)



Using RoIAlign in place of other forms of pooling to eliminate the harsh quantization.



Adding a branch for creating segmentation masks



Using the computed masks to classify each pixel for different objects



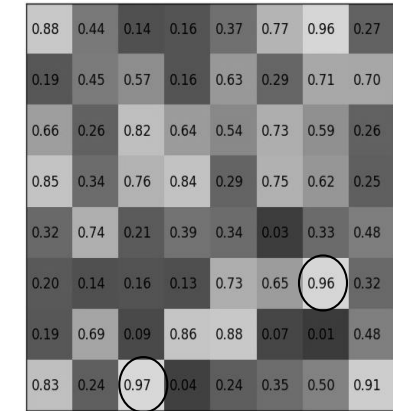
# RoIAlign

## What is RoIAlign

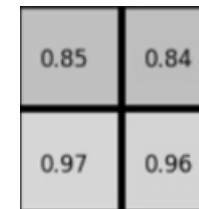
- RoIAlign is the proposed layer which removes the harsh quantization of RoIPool.
- It aligns extracted strong features with input to the layer.

## Why do we use RoIAlign

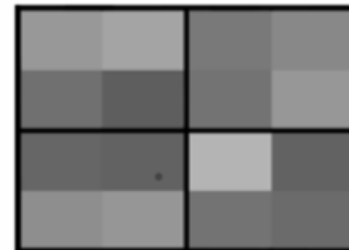
- Improves the mask accuracy by 10% to 50% relatively.
- As we can see in the images on right two high values (the circled values) caused high bias when we used MaxPool only, while RoIAlign normalizes the effect of outliers.
- If we use average pool the extremes will still cause problems. We can see RoIAlign as a middle ground solution between MaxPool and Average Pool.
- It does not break pixel-to-pixel translation equivariance.



Input to both the algorithms



Max pooling output



Bilinear interpolated values

Output without RoIAlign & output with RoIAlign



# Network Architecture

(how conversion to classification is addressed)

## Tail and body of the network

- In the proposed network we have used the Residual Network Structure for the tail and body of the network.

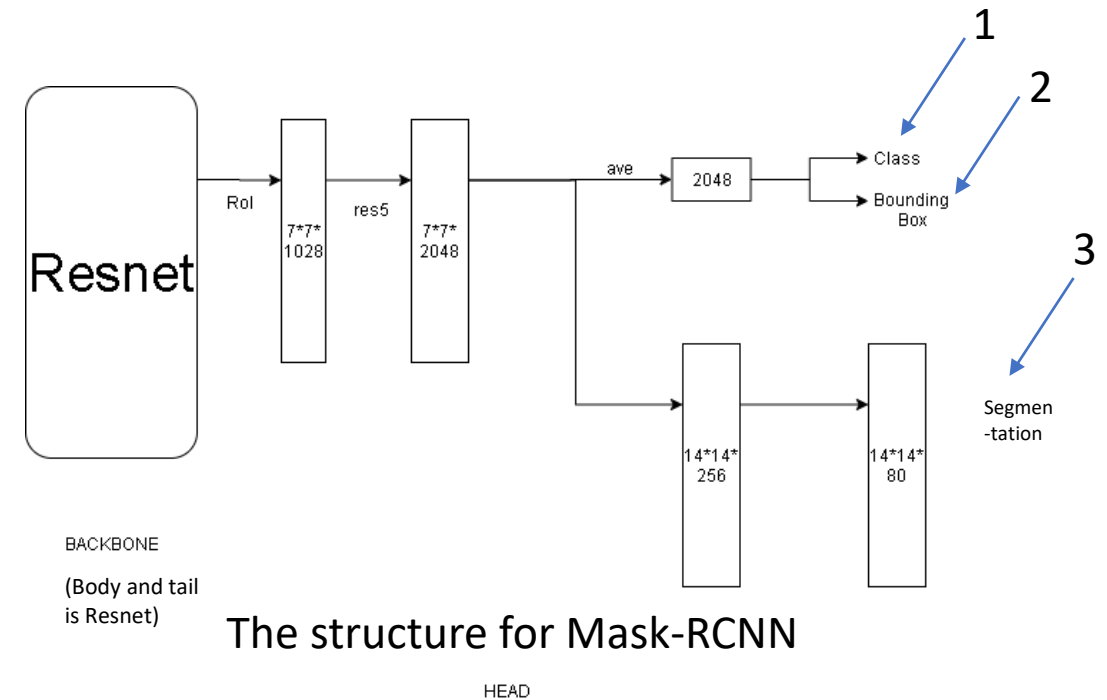
## Head of the network

- The output of ResNet is first passed through a RoIAlign layer and then through a CNN block.
- The upper branch is used to find the class and the bounding/anchor box for the object.
- The lower branch is used for the Image Segmentation.

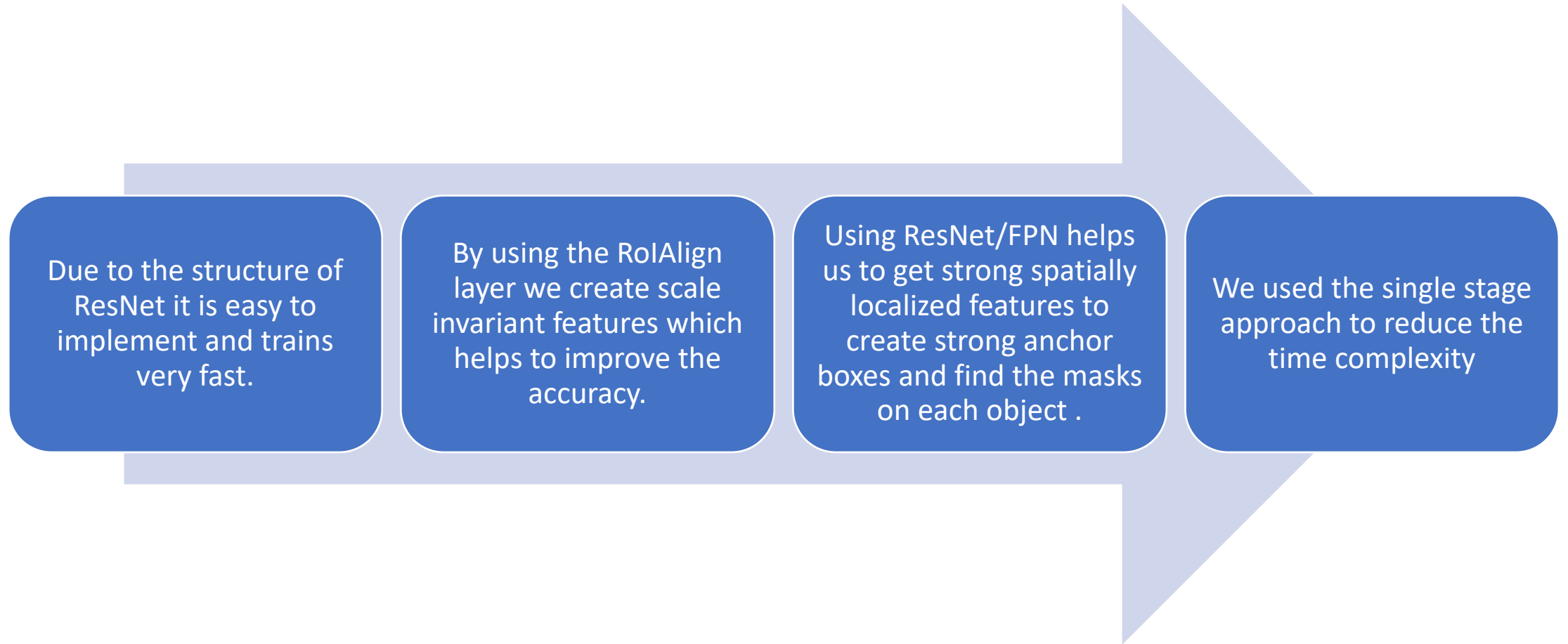
# Network Architecture (Continued)

## Heads for the network

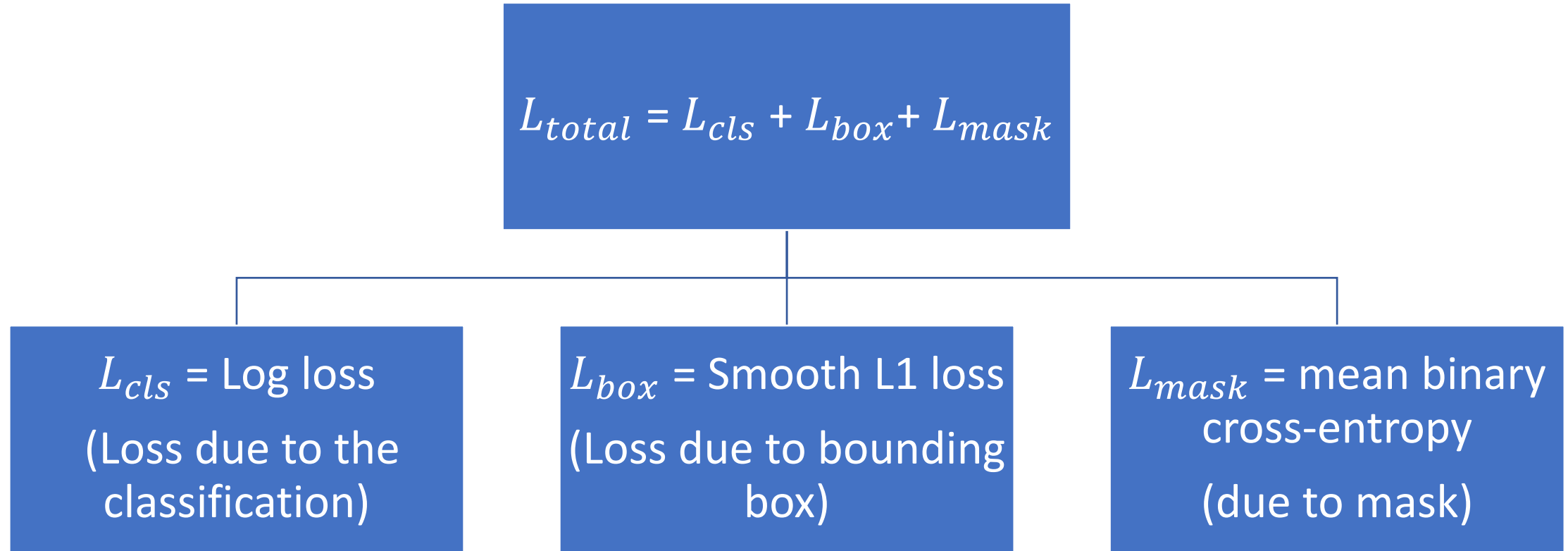
- In the case of proposed algorithm we have 3 parallel heads :-
  1. Classification head
  2. Bounding box reg head
  3. Mask Head



# Advantages of using the proposed architecture



# Loss Function for training



- This loss will then be used to update the parameters for the network

# Step by Step Algorithmic Approach

## Anchor sorting and filtering

**Find the anchor boxes and segregate them into +ve and -ve boxes.**



## Bounding/Anchor Box Refinement

**Refine the anchor boxes and find the final bounding/anchor box.**



## Mask Generation

**Generate the mask and place them onto the objects in the image.**



## Weight Histograms

**This tool is used to see and debug the algorithm**



## Composing the different pieces into a final result

**Then we impose the bounding boxes and masks on the image**

# Experimentation Details

## Dataset Used to train the model :- Coco Dataset

- 123,287 images and 886,284 instances.
- 80 thing classes, 91 stuff classes and 1 class 'unlabeled'.
- For testing we used some generic images which contained people, vehicles, umbrella and many more things.

## Details about the output images

- Anchor boxes are represented as dashed line.
- Segmented images are highlighted with different colors for different objects.
- The name of the object and its probability is mentioned on the top left corner of the bounding box.

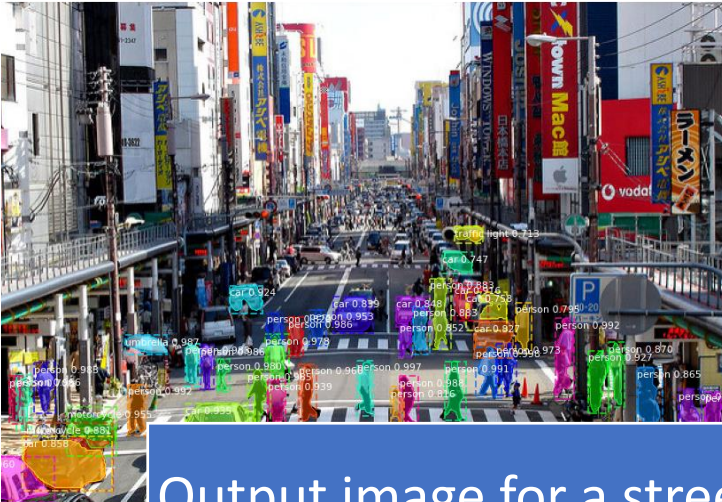
# Demo Results



Input image of a street



Image of Comet Cruiser



Output image for a street



Output for Comet Cruiser





Image of UTD Parking




Image of Sun Temple Modera, India



Output for UTD Parking



Output for Sun Temple



## Results Analysis

Ablation	Mask AP		Box AP	
	AP – 50	AP – 75	AP – 50	AP – 75
Max Pooling	46.5	21.6	52.7	26.9
RoIAlign	51.8	32.1	55.3	36.4
	+5.3	+10.5	+2.6	+9.5

- Here as we can see that using the RoIAlign layer gives huge improvements in the values.

	backbone	AP <sup>bb</sup>	AP <sup>bb</sup> <sub>50</sub>	AP <sup>bb</sup> <sub>75</sub>	AP <sup>bb</sup> <sub>S</sub>	AP <sup>bb</sup> <sub>M</sub>	AP <sup>bb</sup> <sub>L</sub>
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	<b>52.1</b>
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3	19.8	40.2	48.8
<b>Mask R-CNN</b>	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
<b>Mask R-CNN</b>	ResNeXt-101-FPN	<b>39.8</b>	<b>62.3</b>	<b>43.4</b>	<b>22.1</b>	<b>43.2</b>	51.2

- The Mask R-CNN algorithm gives the best results in 5 out of 6 cases.
- The use of RoIAlign and an extra head has helped to increase the accuracy this much.
- We can also see that using RoIAlign in Faster R-CNN also gives better results, which indeed proves that RoIAlign is a factor in improving the performance.



# Failure cases (recognition)





Car not Represented

Inconsistent Mask

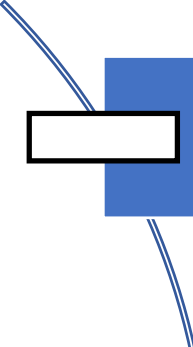
car 0.869 car 0.989 car 0.964 car 0.787 car 0.985 car 0.801 car 0.974 car 0.715 car 0.981 car 0.855 car 0.991 car 0.991 car 0.718 car 0.960 car 0.969 car 0.727 car 0.919 car 0.990 car 0.990 car 0.990

20

Car not Represented

## Inconsistent Mask


# Observations




As we can see in the previous slides, we have some false positives, false negatives, and some inconsistencies in the masks.



These errors can be led to a death if we use a model like this in automated cars, robots, medical imaging, etc.

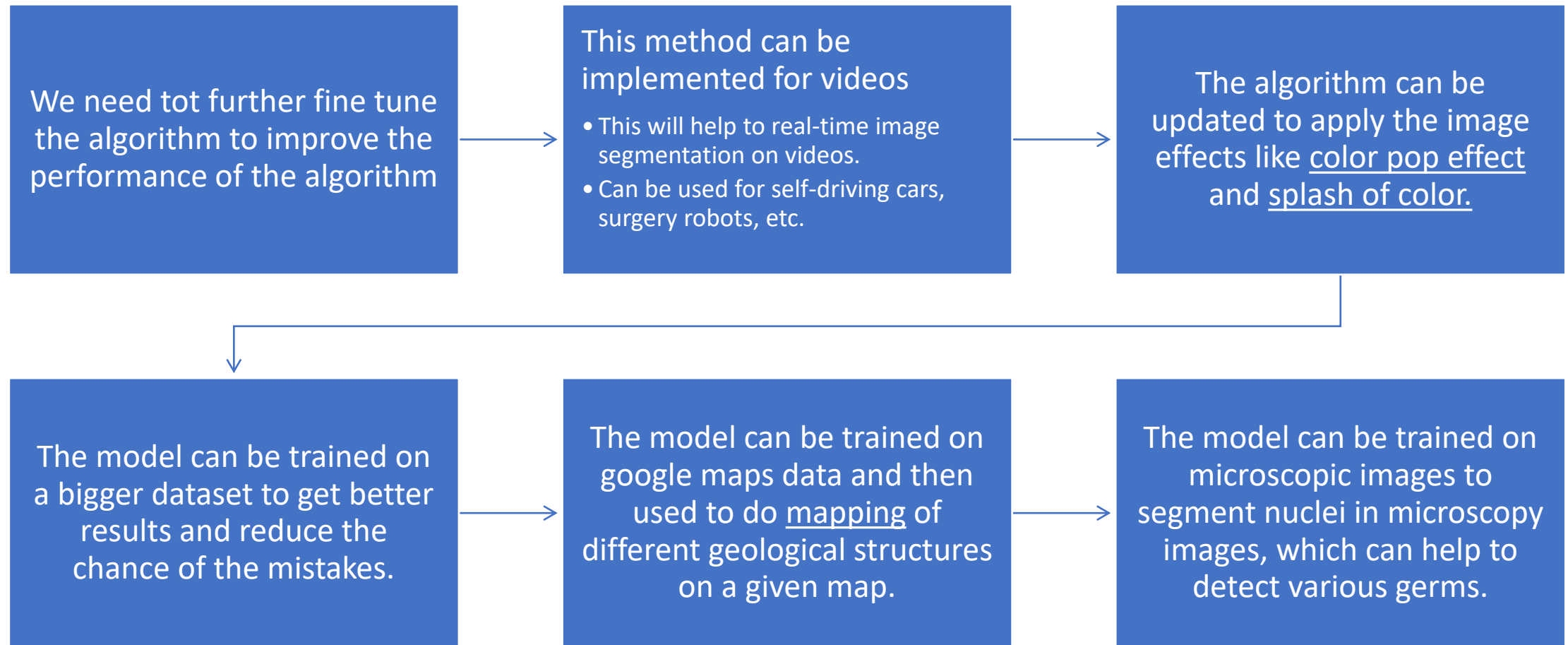


As we can see in the images, objects of small size are also labeled with high probability, which strengthens the scale-invariant mask statement we made earlier.



In the street image we saw, there were many objects of same and different classes and the model was able to perform very well and thus it can be used in various difficult tasks.

# Future plans



Thus we can say that everything depends on what we give as the input to the algorithm. And the same structure can be use for different applications



# References

- Note: As stated above, this presentation is a work of fiction; the following are the actual inventors of the ideas described in this presentation
- R. Girshick. “Fast R-CNN“. In ICCV, 2015.
- J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation“. In CVPR, 2015.
- S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks.” In NIPS, 2015.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. “Deep Residual Learning for Image Recognition.” In arXiv : 1512.03385
- Jaderberg, Max, et al. "Spatial Transformer Networks." arXiv preprint arXiv:1506.02025
- Kaiming He and Georgia Gkioxari and Piotr Dollár and Ross Girshick, et. al., “Mask R-CNN,” In arXiv:1703.06870, 2017.
- Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. “Fully convolutional instance-aware semantic segmentation.” In CVPR, 2017.

# Thank You!