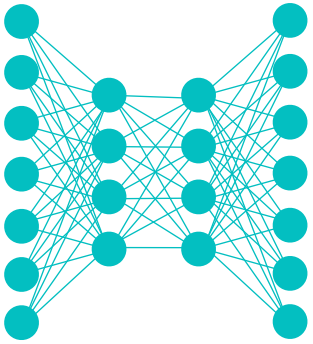
Lecture Notes for

Neural Networks and Machine Learning

Fully Convolutional Learning Instance Segmentation

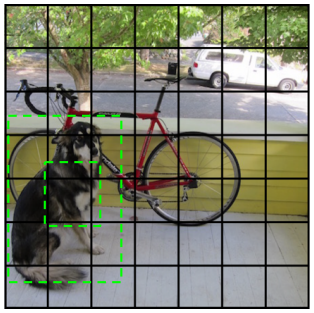


Logistics and Agenda

- Logistics
  - Lab One Updates
- Agenda
  - Segmentation
    - Semantic (last last time)
    - Object (finish this time)
    - Instance (this time)

The YOLO Loss Function

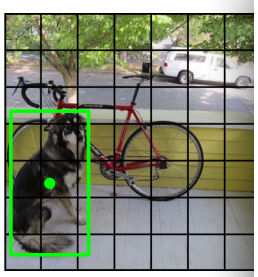
Update Bounding Box



$$\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_{ij})^2 + (y_i - \hat{y}_{ij})^2 \right]$$

$$+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_{ij}} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_{ij}} \right)^2 \right]$$

Localization Loss



SxS cells, i^{th} cell

B boxes per cell, jth box

$\mathbb{1}^{\text{obj}}$ indicator function, from GT

$\hat{C}$ is confidence per box

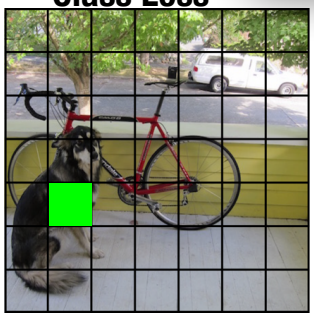
$\hat{p}(c)$ softmax output, per class

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_{ij} \right)^2$$

$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_{ij} \right)^2$$

Object Detection Loss

Class Loss



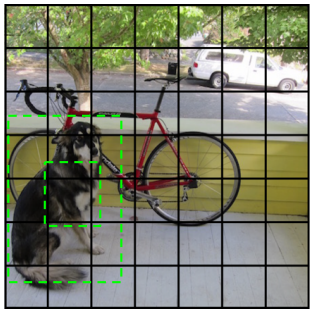
$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2$$

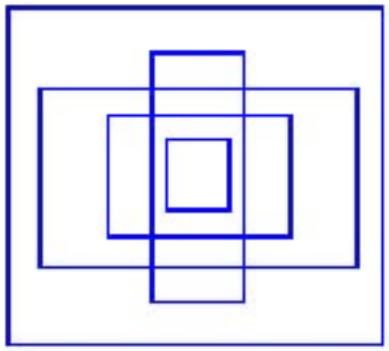
Classification Loss

https://medium.com/@jonathan_hui/real-time-object-detection-with-yolo-yolov2-28b1b93e2088

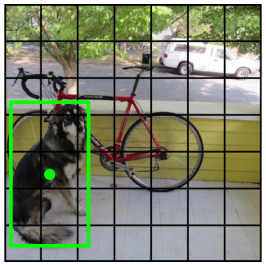
Updated YOLO Localization

Update Bounding Box

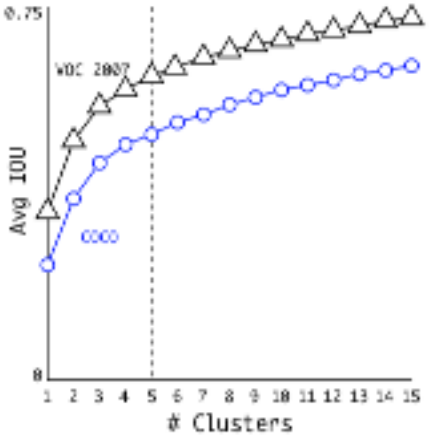


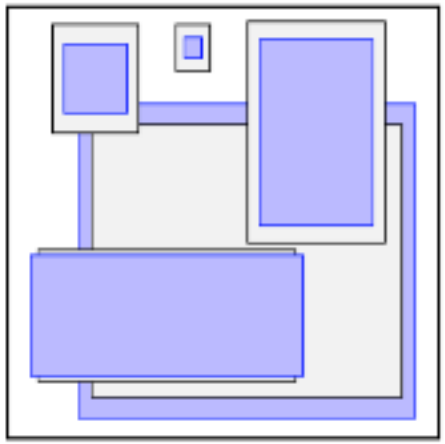


- Define 5 pre-defined box shapes (based on data)
- Regress x, y offset from cell center
- Bound x and y to the bounds of the cell
- And w and h scaling of predefined shape

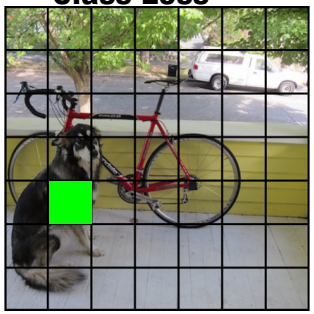


"Good" Shape Priors Found via Clustering





Class Loss



$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{\text{obj}}\left(C_i-\hat{C}_{ij}\right)^2$$

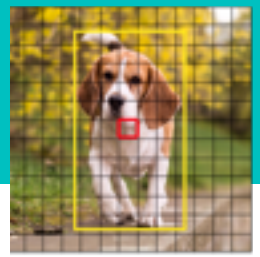
$$+\lambda_{\text{noobj}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{\text{noobj}}\left(C_i-\hat{C}_{ij}\right)^2$$

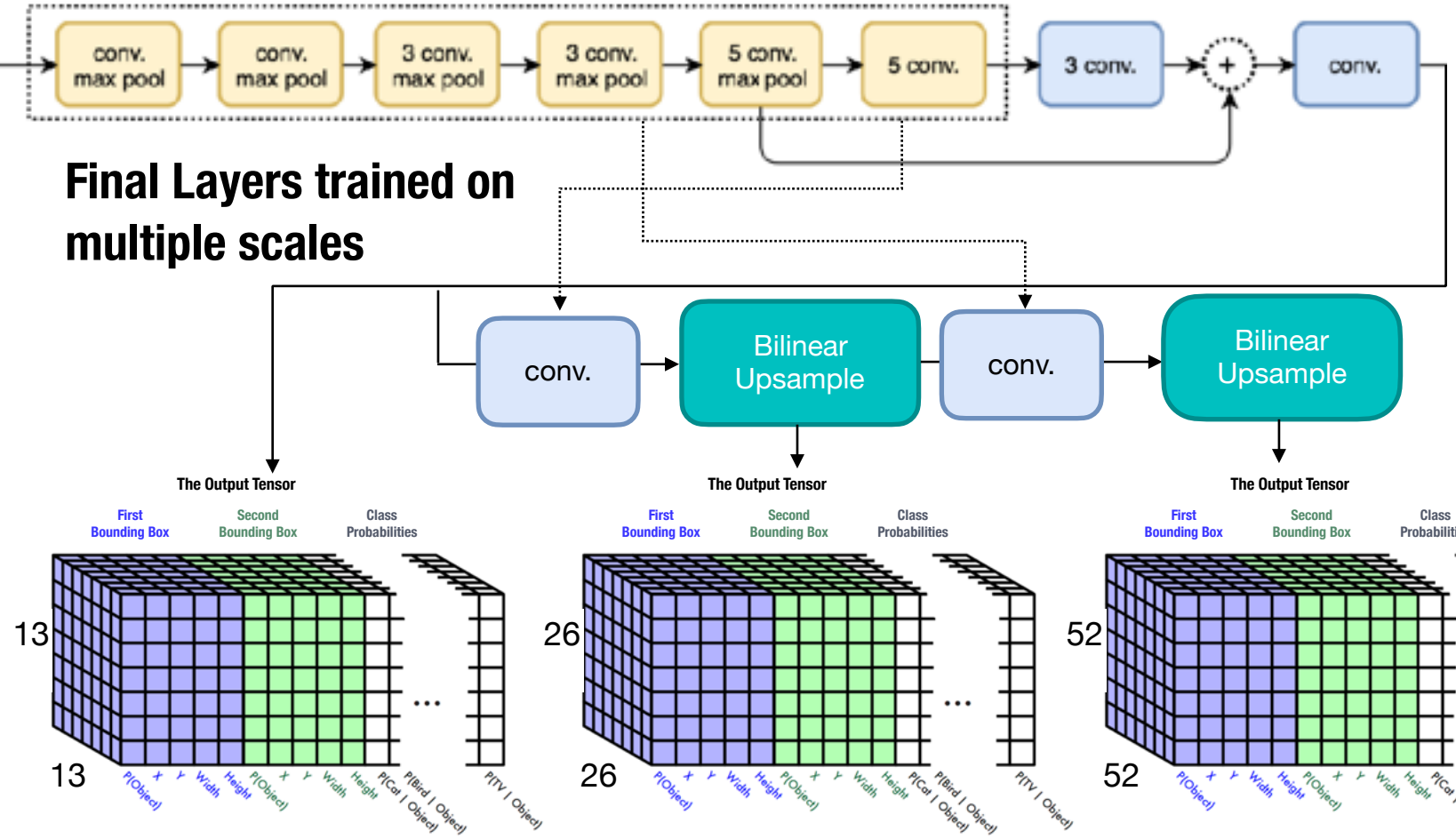
$$+\sum_{i=0}^{S^2}\mathbb{1}_{i}^{\text{obj}}\sum_{c\in\text{classes}}\left(p_i(c)-\hat{p}_i(c)\right)^2$$

Classification Loss

https://medium.com/@jonathan_hui/real-time-object-detection-with-yolo-yolov2-28b1b93e2088

The YOLOv3 Architecture





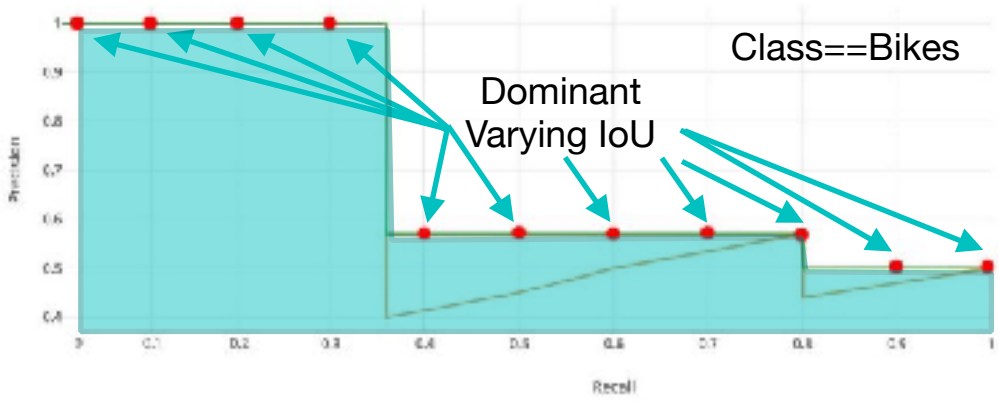
All Losses Applied, 13x13 grid

All Losses Applied, 26x26 grid

All Losses Applied, 52x52 grid

https://medium.com/@jonathan_hui/real-time-object-detection-with-yolo-yolov2-28b1b93e2088

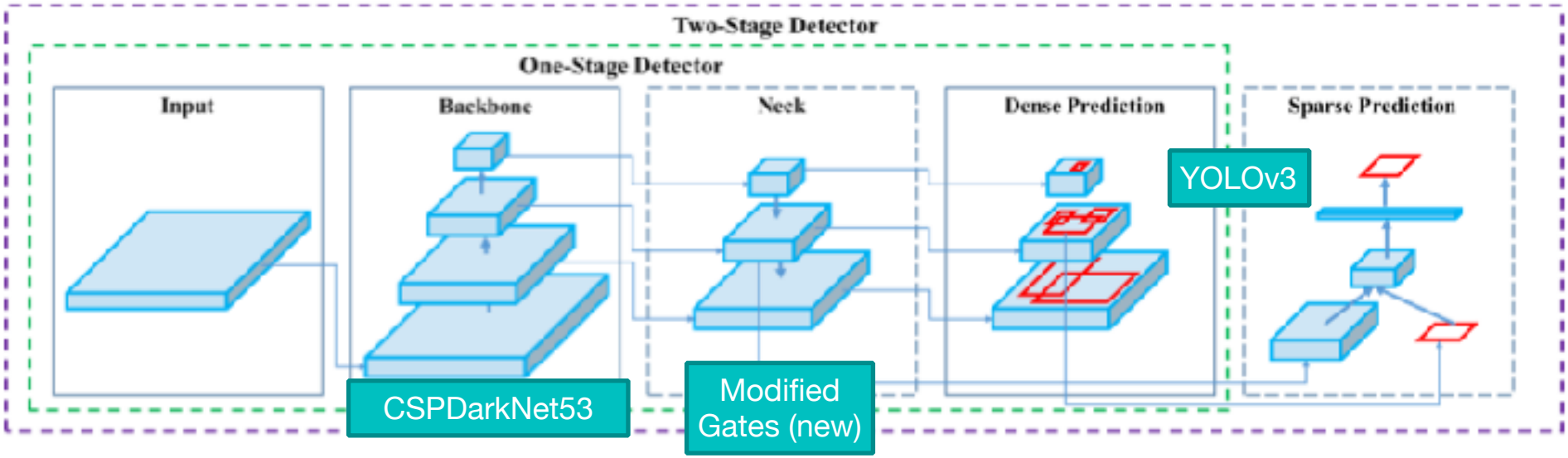
Measuring Performance



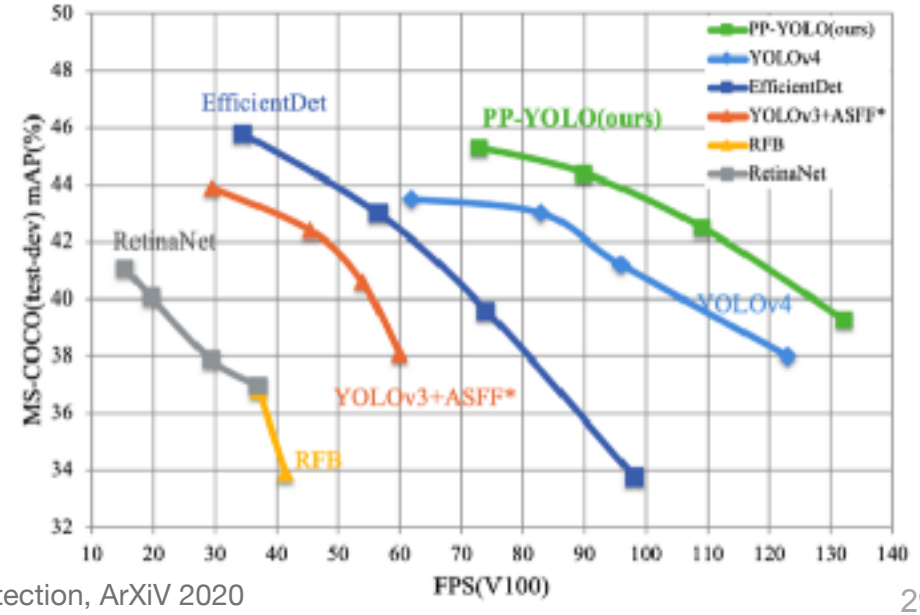
- mAP(IoU=x%)
  - if IoU > X%, correct
  - else not
  - Usually: 50% and 75%
  - Define precision for each class, take average
- mAP(%), *sometimes just AP*
  - Formulate precision/recall curve for a class at varying levels of IoU (50%-95%)
  - Calculate dominating points
  - Take area under curve
  - Take average area over all classes (macro or micro averaging can be done, usually macro)

https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Is there a YOLOv4?



- **Yes** and it has improvements based on some simple elements
- There is also a "YOLOv5" that caused a lot of controversy and then a PP-YOLO.



Bochkovskiy et al. YOLOv4: Optimal Speed and Accuracy of Object Detection, ArXiV 2020

A closing thought from YOLOv3 Report

Joseph Redmon @pjreddie · Feb 20, 2020

"We shouldn't have to think about the societal impact of our work because it's hard and other people can do it for us" is a really bad argument.

> **Roger Grosse** @RogerGrosse
> Replying to @kevin_zakka and @hardmaru
> To be clear, I don't think this is a positive step. Societal impacts of AI is a tough field, and there are researchers and organizations that study it professionally. Most authors do not have expertise in the area and won't do good enough scholarship to say something meaningful.

Joseph Redmon
@pjreddie

I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore.

> **Roger Grosse** @RogerGrosse
> Replying to @skoularidou
> What's an example of a situation where you think someone should decide not to submit their paper due to Broader Impacts reasons?

10:09 AM · Feb 20, 2020

Redmon and Farhadi, Y

A closing thought from YOLOv3 Report

The Rebuttal I wish I could Write:

Reviewer #2 AKA Dan Grossman (lol blinding who does that) insists that I point out here that our graphs have not one but two non-zero origins. You're absolutely right Dan, that's because it looks way better than admitting to ourselves that we're all just here battling over 2-3% mAP. But here are the requested graphs. I threw in one with FPS too because we look just like super good when we plot on FPS.

Reviewer #4 AKA JudasAdventus on Reddit writes "Entertaining read but the arguments against the MSCOCO metrics seem a bit weak". Well, I always knew you would be the one to turn on me Judas. You know how when you work on a project and it only comes out alright so you have to figure out some way to justify how what you did actually was pretty cool? I was basically trying to do that and I lashed out at the COCO metrics a little bit. But now that I've staked out this hill I may as well die on it.

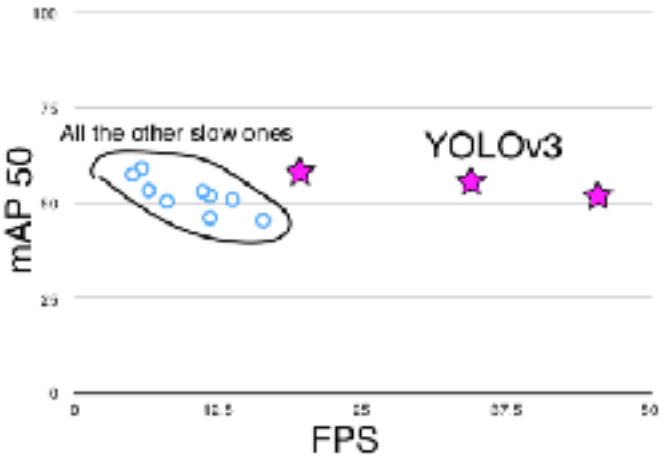


Figure 5. These two hypothetical detectors are perfect according to mAP over these two images. They are both perfect. Totally equal.

Now this is OBVIOUSLY an over-exaggeration of the problems with mAP but I guess my newly reteconned point is that there are such obvious discrepancies between what people in the "real world" would care about and our current metrics that I think if we're going to come up with new metrics we should focus on these discrepancies. Also, like, it's already mean average precision, what do we even call the COCO metric, average mean average precision?

Here's a proposal, what people actually care about is given an image and a detector, how well will the detector find and classify objects in the image. What about getting rid of the per-class AP and just doing a global average precision? Or doing an AP calcu-

Boxes are stupid anyway though, I'm probably a true believer in masks except I can't get YOLO to learn them.

YOLACT, one pass mask generation

YOLACT
Real-time Instance Segmentation

Daniel Bolya Chong Zhou Fanyi Xiao Yong Jae Lee
University of California, Davis
{dbolya, czhou, fyxiao, yongjaelee}@ucdavis.edu

Abstract

*We present a simple, fully-convolutional model for real-time instance segmentation that achieves 29.8 mAP on MS COCO at 33.5 fps evaluated on a single Titan Xp, which is significantly faster than any previous competitive approach. Moreover, we obtain this result after training on **only one GPU**. We accomplish this by breaking instance segmentation into two parallel subtasks: (1) generating a set of prototype masks and (2) predicting per-instance mask coefficients. Then we produce instance masks by linearly combining the prototypes with the mask coefficients. We find that because this process doesn't depend on repooling, this approach produces very high-quality masks and exhibits temporal stability for free. Furthermore, we analyze the emergent behavior of our prototypes and show they learn to localize instances on their own in a translation variant manner, despite being fully-convolutional. Finally, we also propose Fast NMS, a drop-in 12 ms faster replacement for standard NMS that only has a marginal performance penalty.*

1. Introduction

"Boxes are stupid anyway though, I'm probably a true believer in masks except I can't get YOLO to learn them."
– Joseph Redmon, YOLOv3 [36]

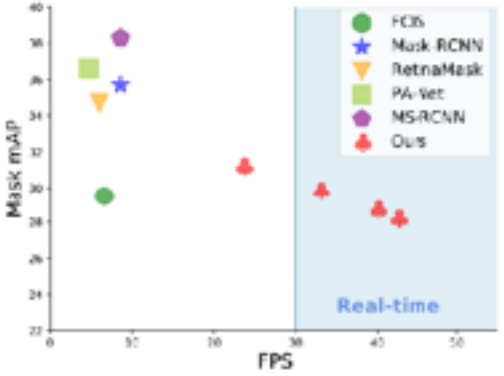


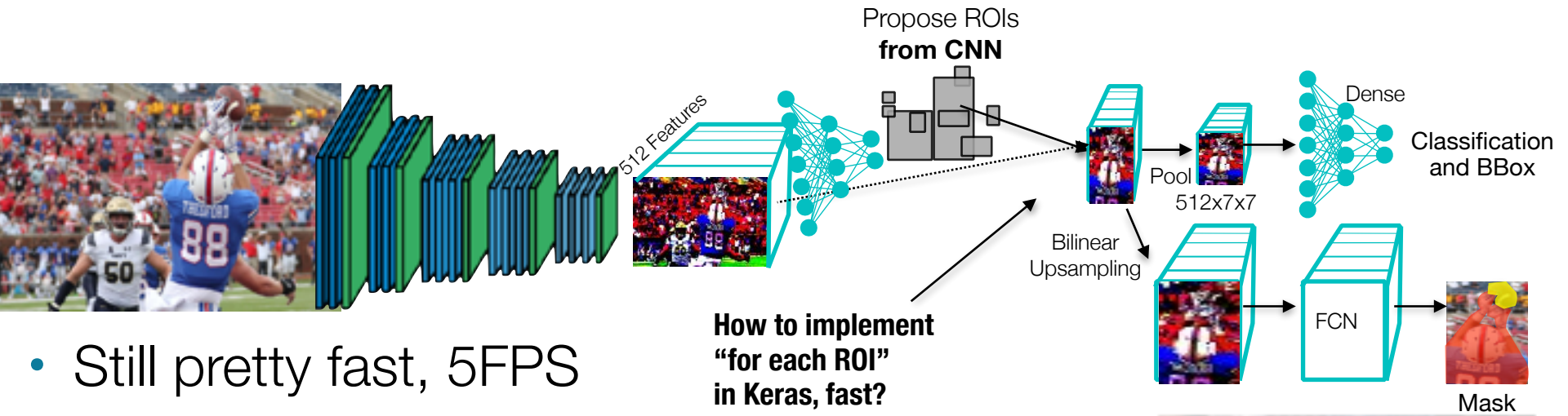
Figure 1: Speed-performance trade-off for various instance segmentation methods on COCO. To our knowledge, ours is the first *real-time* (above 30 FPS) approach with around 30 mask mAP on COCO `test-dev`.

However, instance segmentation is hard—much harder than object detection. One-stage object detectors like SSD and YOLO are able to speed up existing two-stage detectors like Faster R-CNN by simply removing the second stage and making up for the lost performance in other ways. The same approach is not easily extendable, however, to instance segmentation. State-of-the-art two-stage

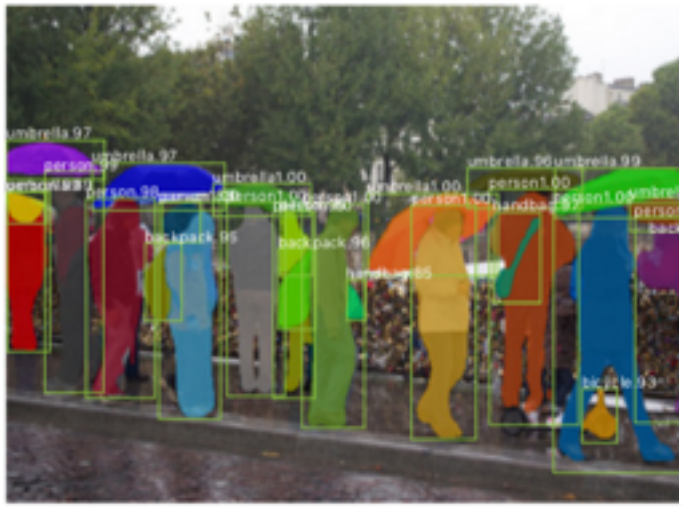
Bolya, Daniel, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. "YOLACT: real-time instance segmentation." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9157-9166. October 2019.

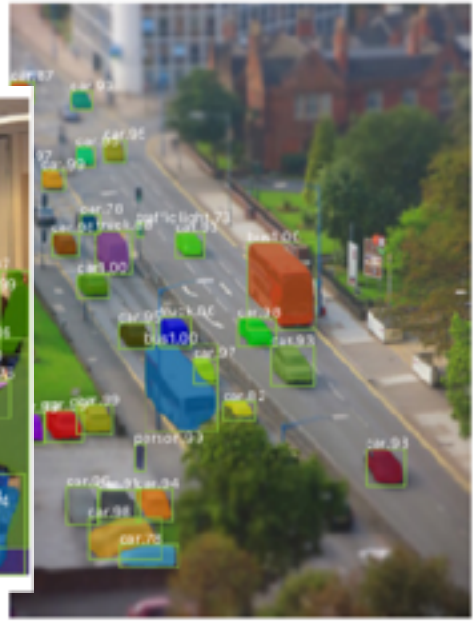
Instance Segmentation

2018: Mask R-CNN



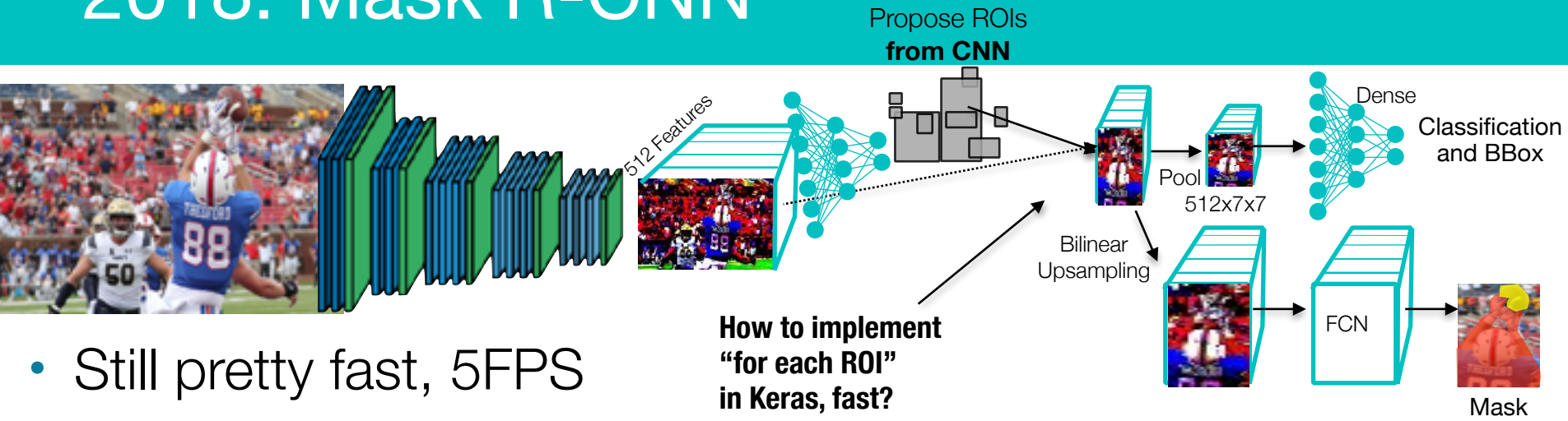
- Still pretty fast, 5FPS
- Excellent results





Ren et al. Mask R-CNN, 2018

2018: Mask R-CNN



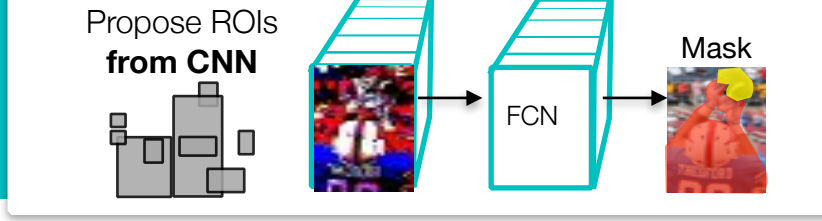
- Still pretty fast, 5FPS
- Excellent results

An Excellent, well documented Implementation here:
https://github.com/matterport/Mask_RCNN

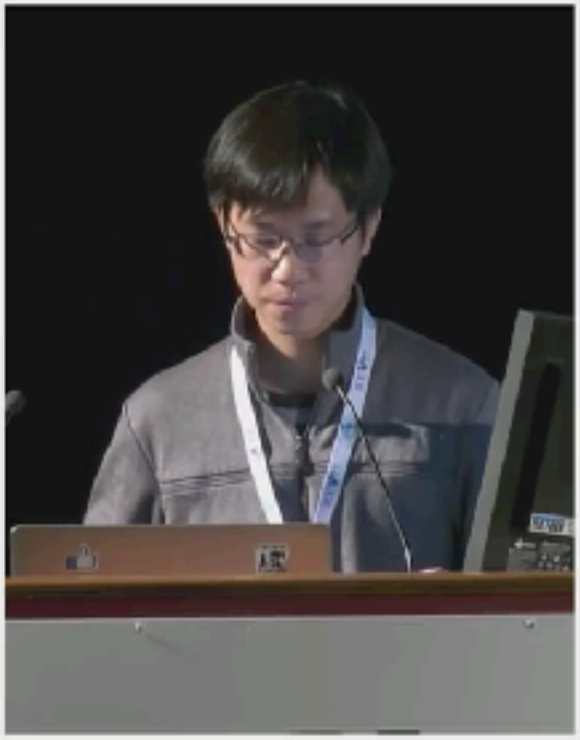
```
# Use shape of first image. Images in a batch must have the same size.
image_shape = parse_image_meta_graph(image_meta)['image_shape'][0]
# Equation 1 in the Feature Pyramid Networks paper. Account for
# the fact that our coordinates are normalized here.
# e.g. a 224x224 ROI (in pixels) maps to P4
image_area = tf.cast(image_shape[0] * image_shape[1], tf.float32)
roi_level = log2_graph(tf.sqrt(h * w) / (224.0 / tf.sqrt(image_area)))
roi_level = tf.minimum(5, tf.maximum(
```

Ren et al. Mask R-CNN, 2018

2018: Mask R-CNN

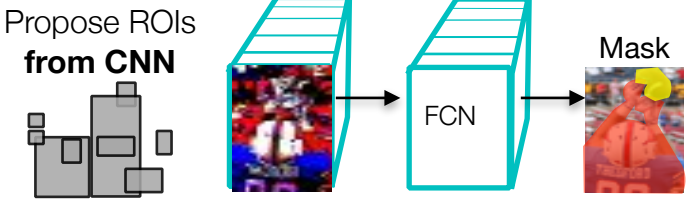


Can also provide **key point detection** from same FCN features (not real time, post processed)



ICCV17

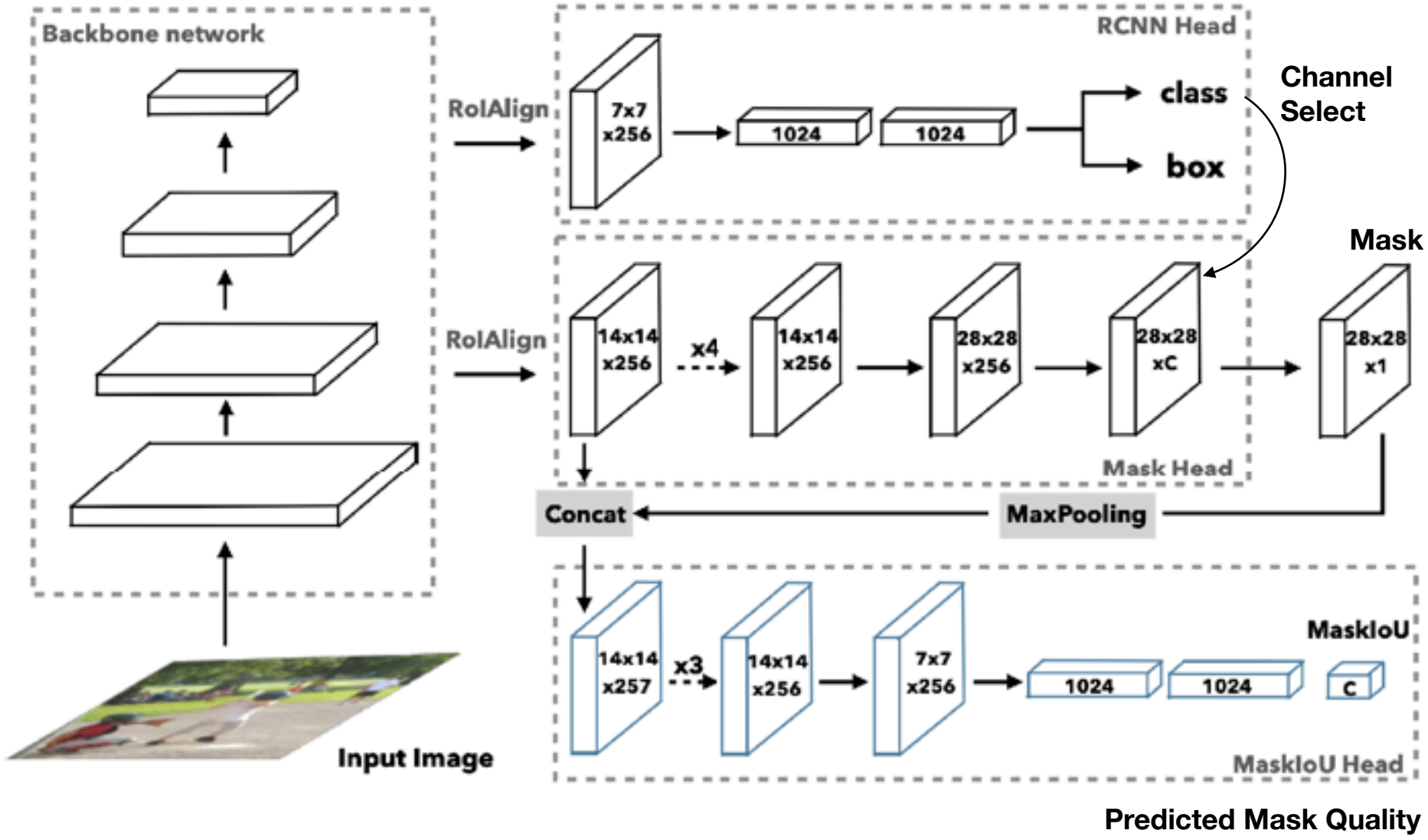
2018: Mask R-CNN



- Real time, Mask R-CNN

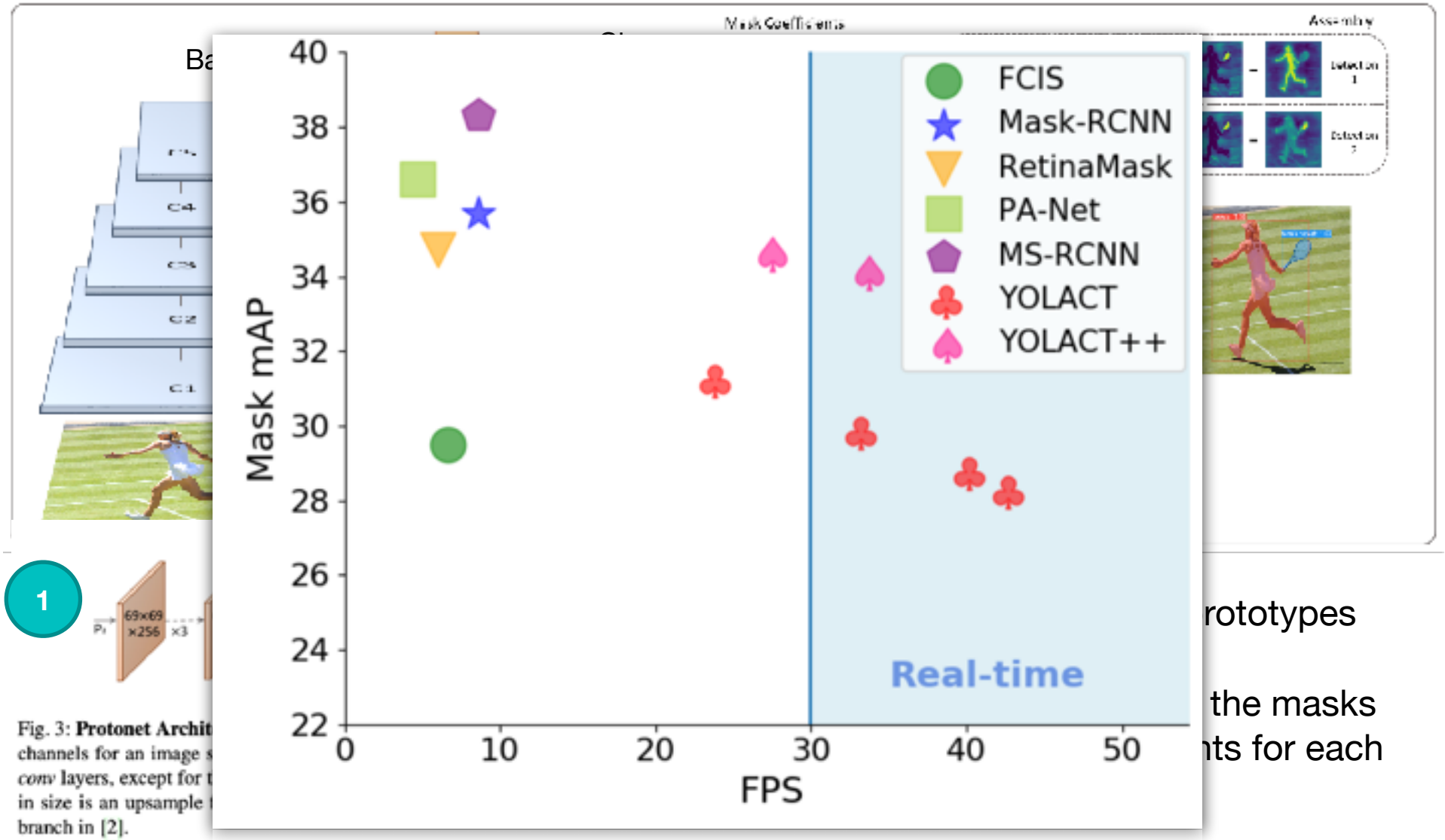
https://www.youtube.com/watch?v=nEug0-pD0Ms

March 2019: Mask Scoring RCNN, MS-RCNN



Zhaojin Huang *et al.* Mask Scoring RCNN. arXiv pre-print, 2019

December 2019: YOLACT++

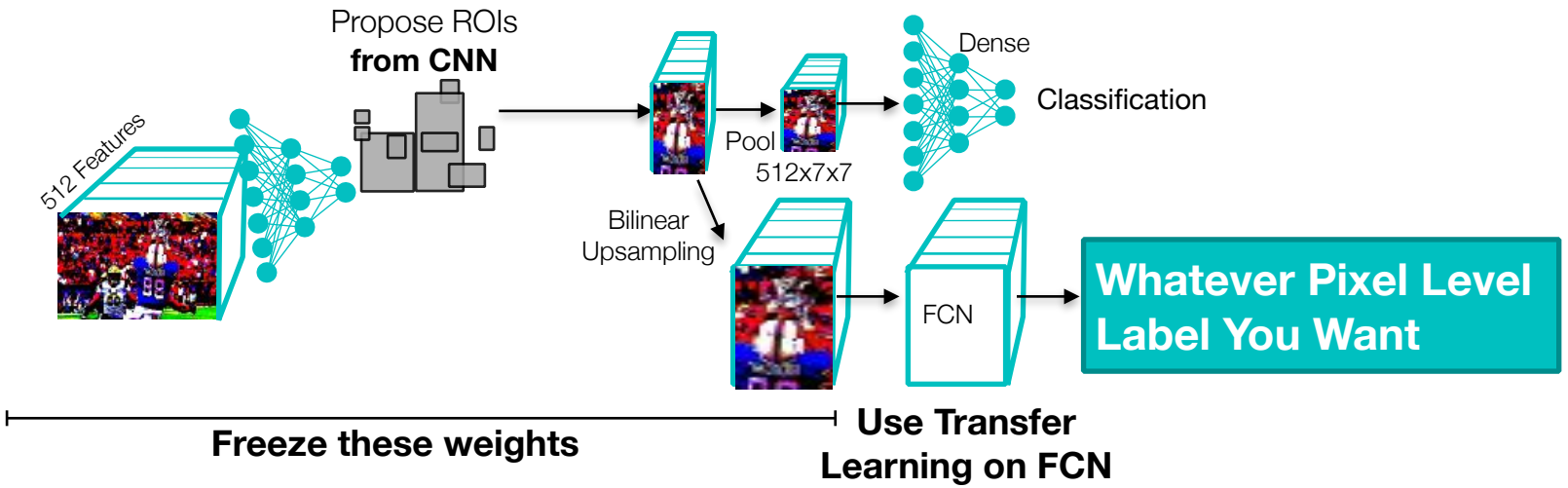


rototypes

the masks
nts for each

Expanding Masking

- Key insight: features that can be used for getting mask of object, are good at doing other things:
  - Like human pose estimation
  - …Depth processing *and more*
- Just connected FCN to image features and learn any label



Operationalizing Masks: Ripeness Detection

Fruit Sorting: YOLO at 130FPS, ripe versus not ripe

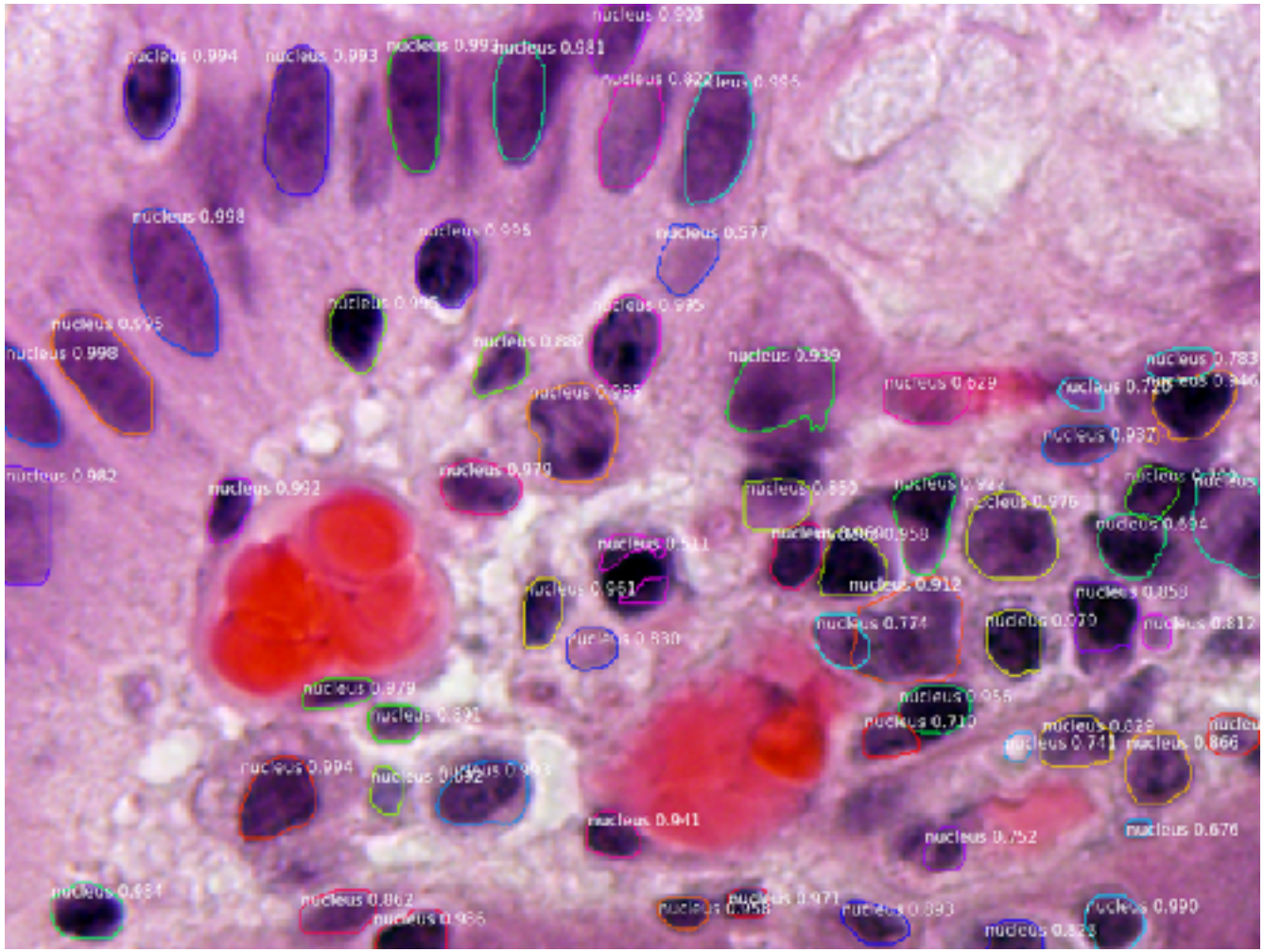
Expanding Masks

3D Building Reconstruction (mask becomes 3D point cloud)

https://github.com/matterport/Mask_RCNN

Retraining Masks

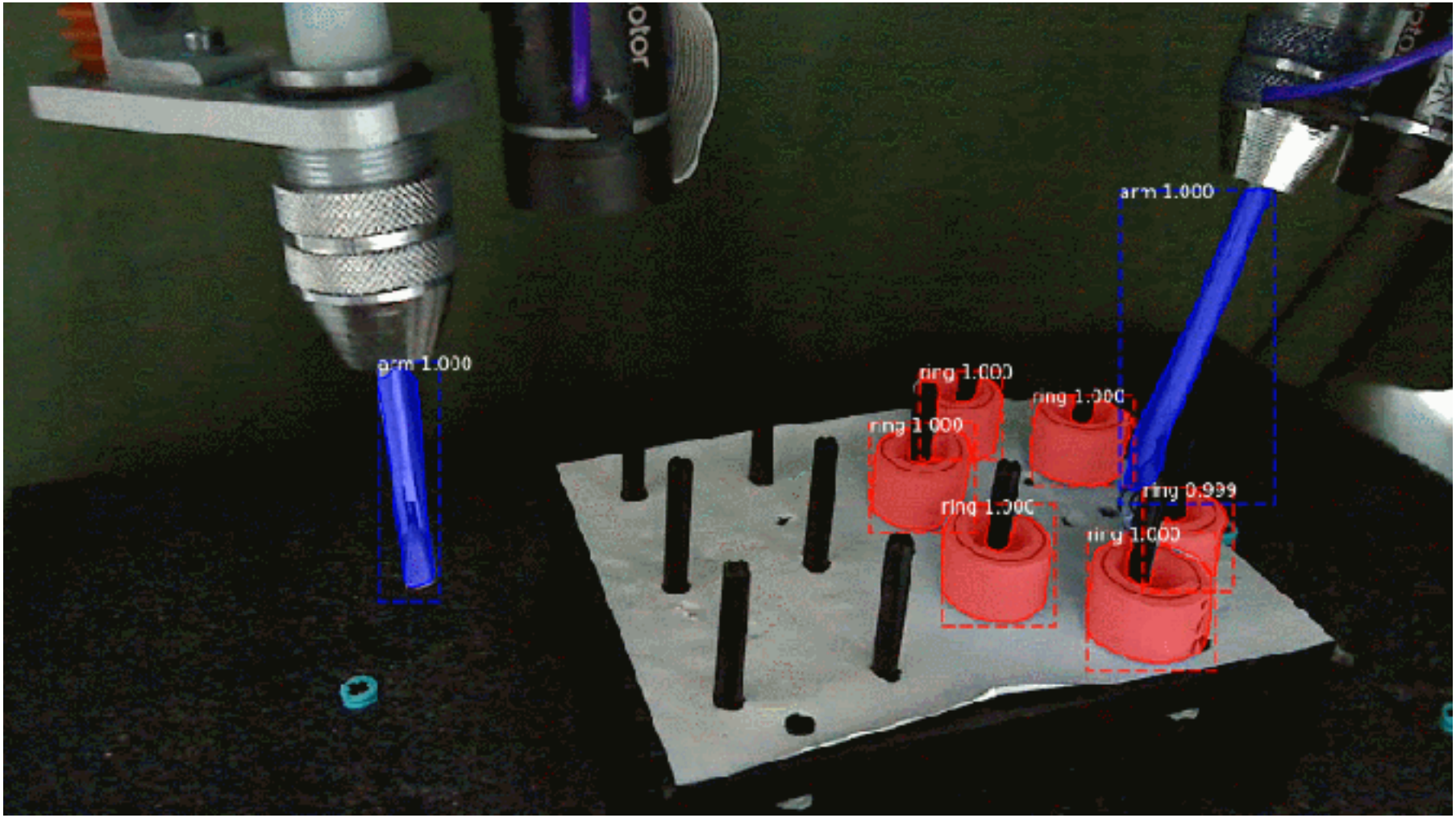
Segmenting Nuclei



https://github.com/matterport/Mask_RCNN

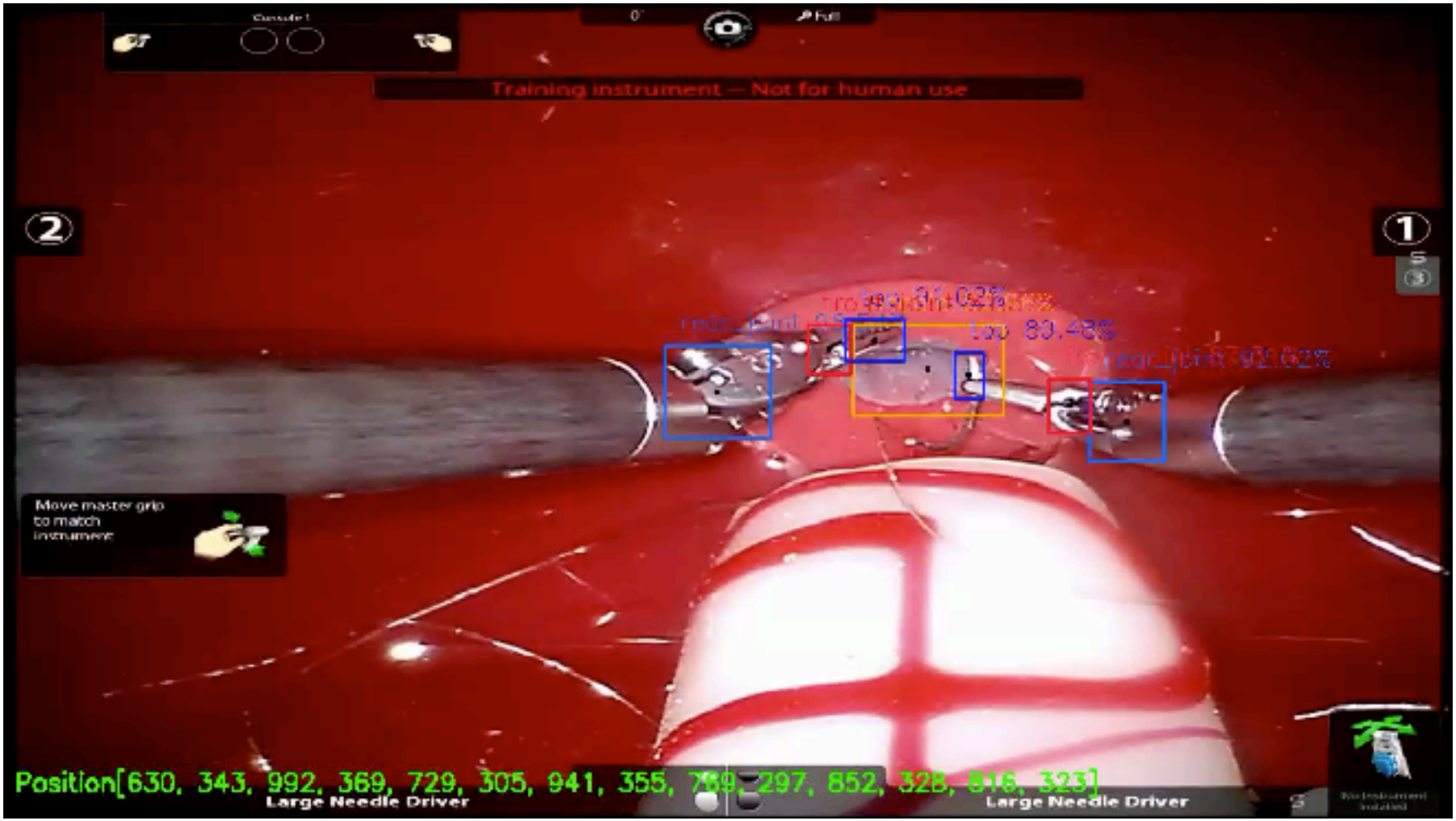
Repurposing for Robotics

Robotic Movement (like surgery)



https://github.com/matterport/Mask_RCNN

Robotic Surgery Assessment



X. Qu, M. El-Saied, J. Gahan, R. Steinberg, and E.C. Larson (2019). "Machine Learning using a Multi-task Convolutional Neural Networks Can Accurately Provide Robotic Skills Assessment." 2019 World Congress of Endourology.

Learning Depth and 3D Shapes

Mesh R-CNN, Facebook AI January 2020

https://ai.facebook.com/blog/pushing-state-of-the-art-in-3d-content-understanding/

In summary

- Semantic segmentation through FCN is active research area
  - DeepLabV3+ or GSCNN are excellent choices

- Object segmentation is excellent, ready for use in industry (Apple's ObjectDetector uses YOLO variant)
  - Already deployed in a many of Apps
  - At 60 FPS, supports tracking applications and AR
  - Can backoff to CPU only at about 5 FPS (on phone)

- Instance Segmentation is ready for deployment in a number of areas, and is now better than realtime with good performance

Lecture Notes for

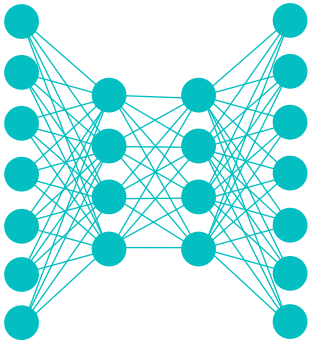
Neural Networks and Machine Learning

Fully Convolutional Learning

Next Time:

Image Style Transfer

Reading: Chollet 8.1– 8.3



Backup slides

Title Between Topics

Example Slide

Title

Subtitle

`Follow Along: Notebook Name`