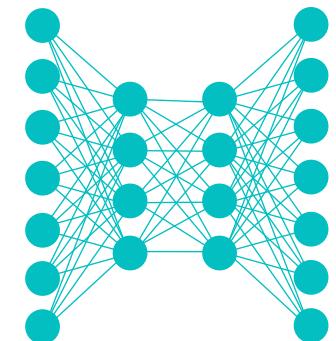


Lecture Notes for **Neural Networks** **and Machine Learning**



Fully Convolutional Learning
Instance Segmentation

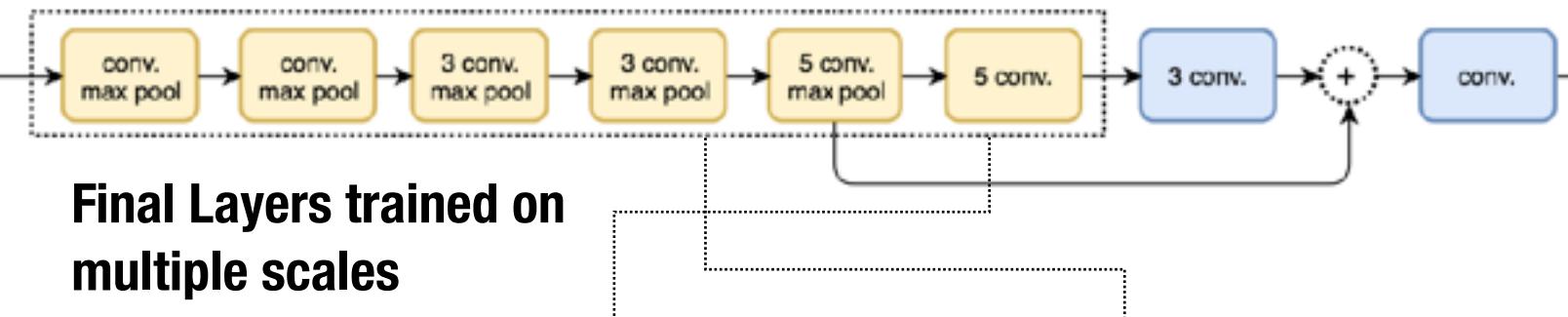
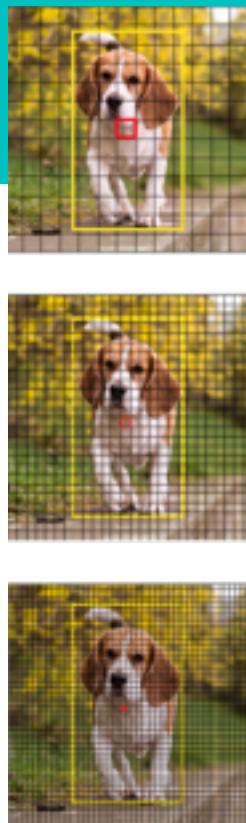


Logistics and Agenda

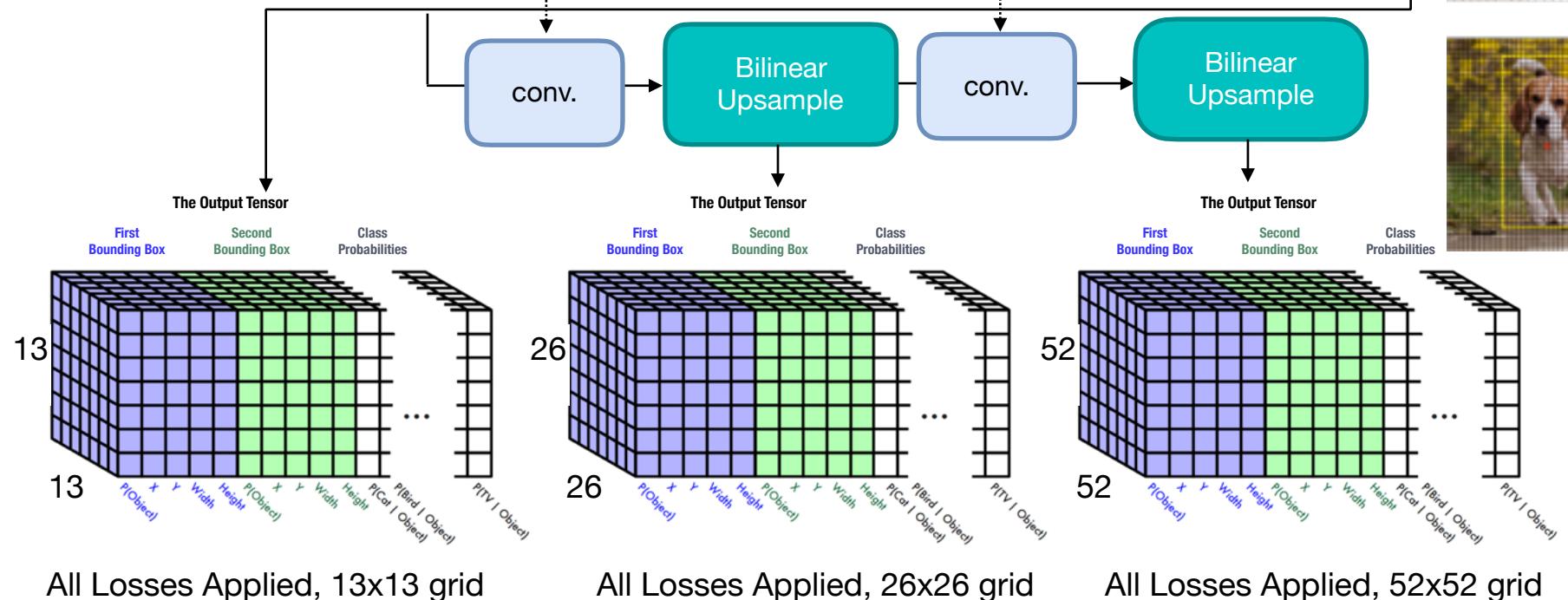
- Logistics
 - Labs due soon
- Agenda
 - Fully Convolutional Learning
 - ◆ Semantic Segmentation (last last time)
 - ◆ Object Detection (mostly last time, finish today)
 - ◆ Instance Segmentation (this time)



Last Time: YOLO up to v3



Final Layers trained on multiple scales



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



More YOLO Variants



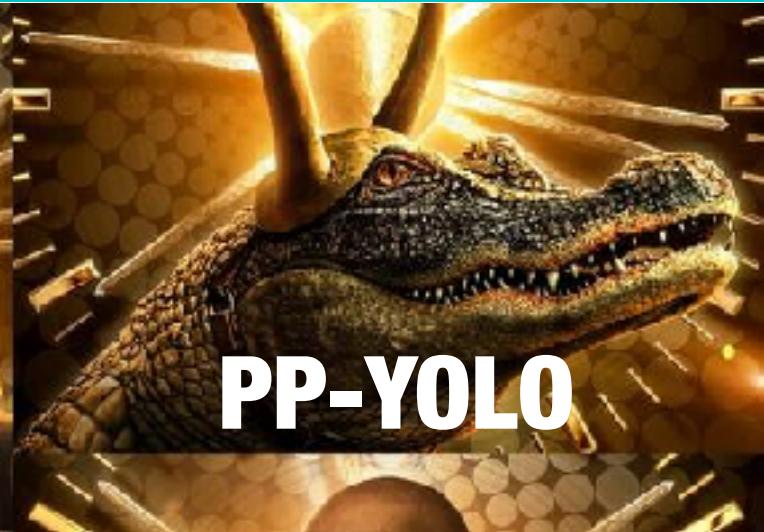
YOLO



YOLOv3



YOLOv4



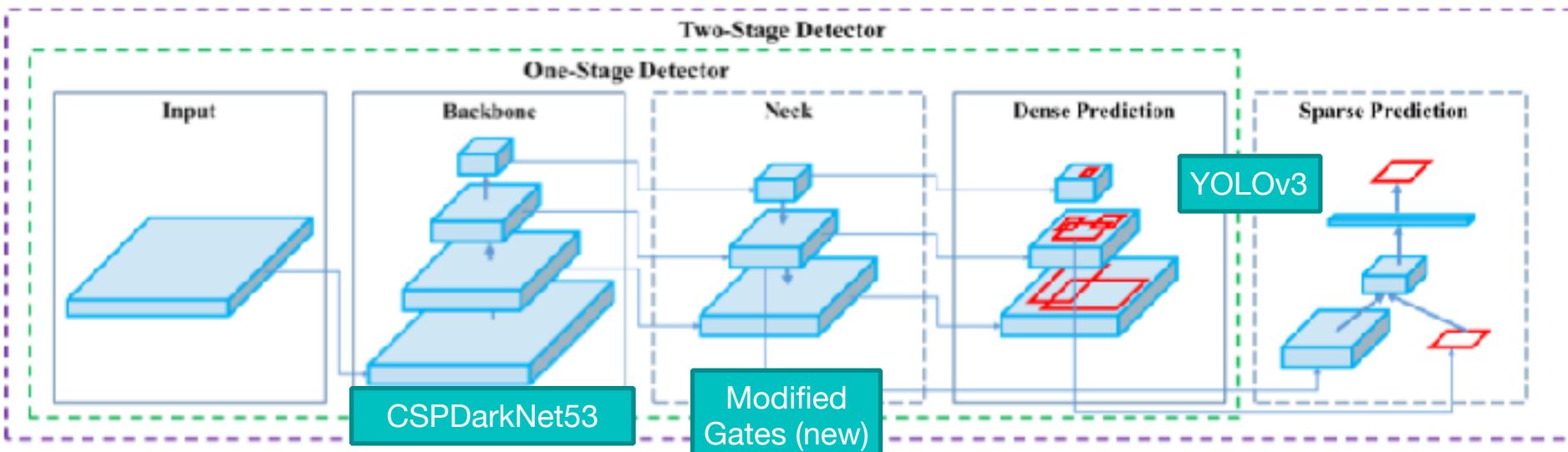
PP-YOLO



PP-YOLOv2



The Basic Elements

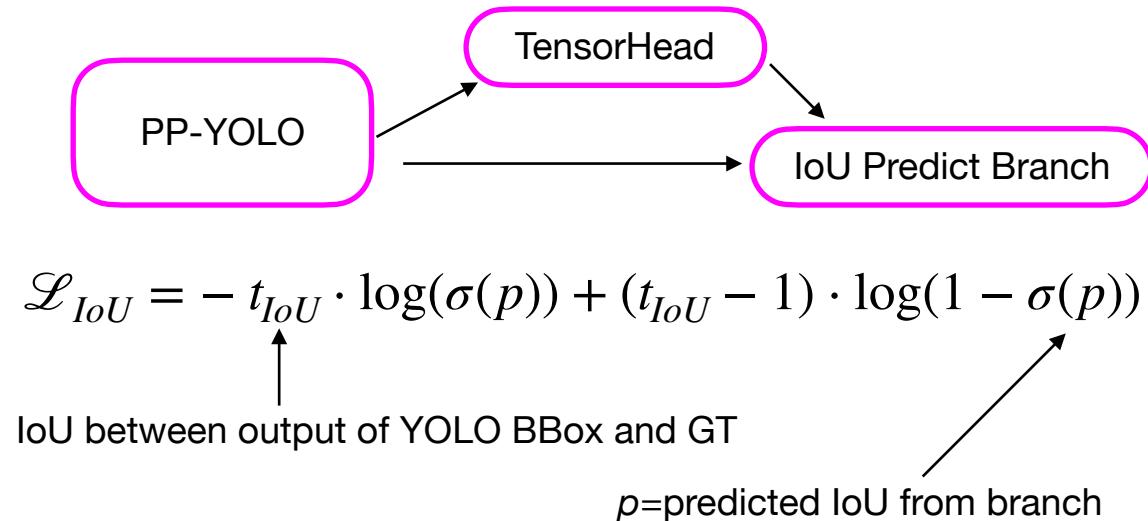
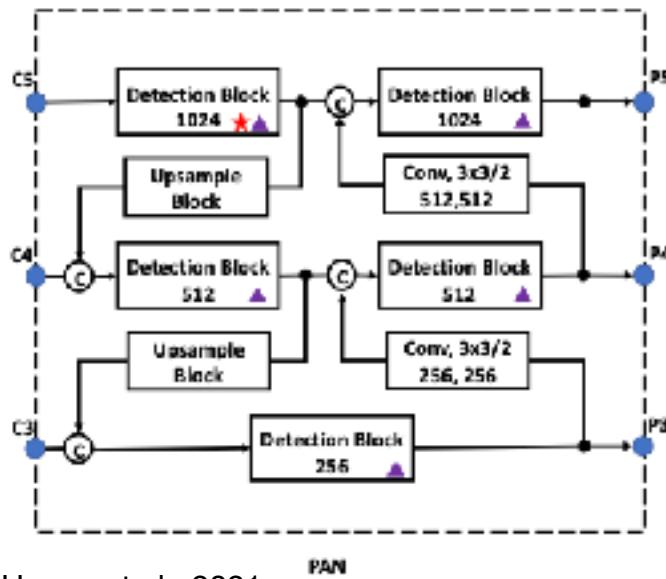


- **YOLOv4** made improvements based on some simple elements
- There is also a “YOLOv5” that caused a lot of controversy because it was essentially a port of existing YOLO to PyTorch
 - Torch-YOLO?
- And then a PP-YOLO, and PP-YOLOv2
 - from Baidu using their deep learning framework PaddlePaddle



PP-YOLOv2: Methods

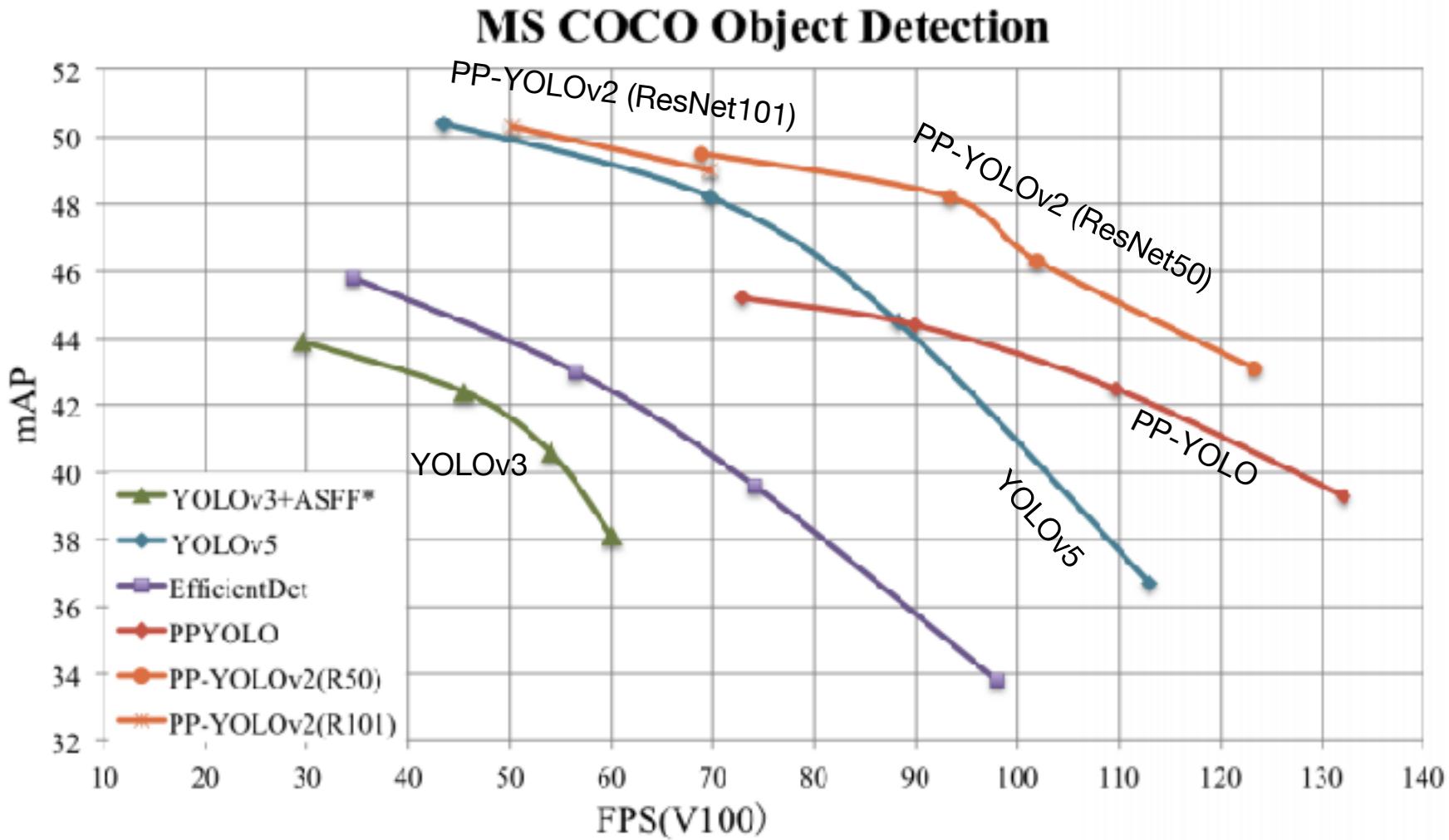
- Path Aggregation Network (PAN) for multi-scale processing
- Varied Input Sized Images and “Mish” activation
- “IoU Aware Branch” that uses the IoU prediction in the back propagation



Main idea is to predict the IoU in a separate multi-task branch, which makes the network better at finding the correct bounding boxes



PP-YOLOv2: Results



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

Replies 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

Replies to @skoulaidou

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, YOLOv3 Report

"What are we going to do with them?" A lot of Google and Facebook. Technology is in good hands. Protect your personal information saying that's exactly

ding vision research are anything horrible like technology oh wait....

the people using computers good stuff with it, like national park [13], or defend their house [19]. But to questionable use and liability to at least consider and think of ways to mitigate.

(I finally quit Twitter).

val Research and Google.



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.

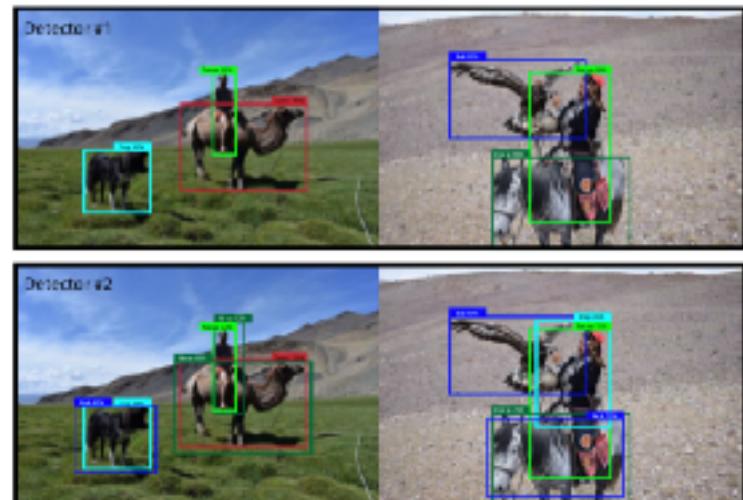
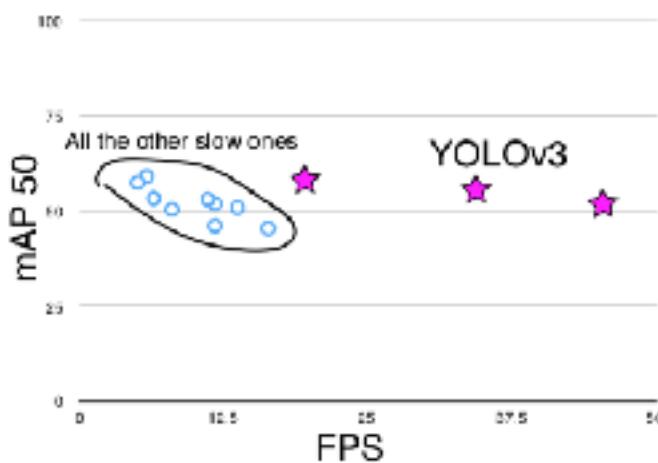


Figure 5. These two hypothetical detectors are perfect according to mAP over these two images. They are both perfect. Totally equal.

Now this is OBVIOUSLY an over-exaggeration of the problems with mAP but I guess my newly retconned 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 calculation per image and averaging over that?

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]

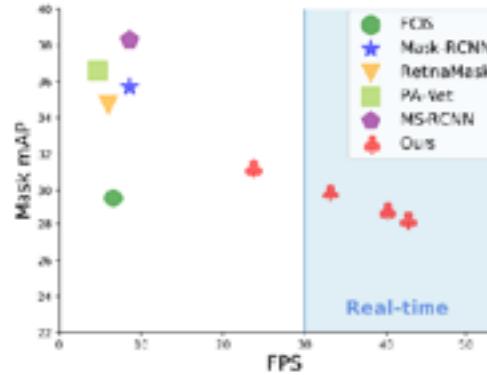


Figure 1: Speed-performance trade-off for various instance segmentation methods on COCO. To our knowledge, ours is the first *real-time* (above 30 FPS) approach with around 30 mask mAP on COCO test-dev.

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

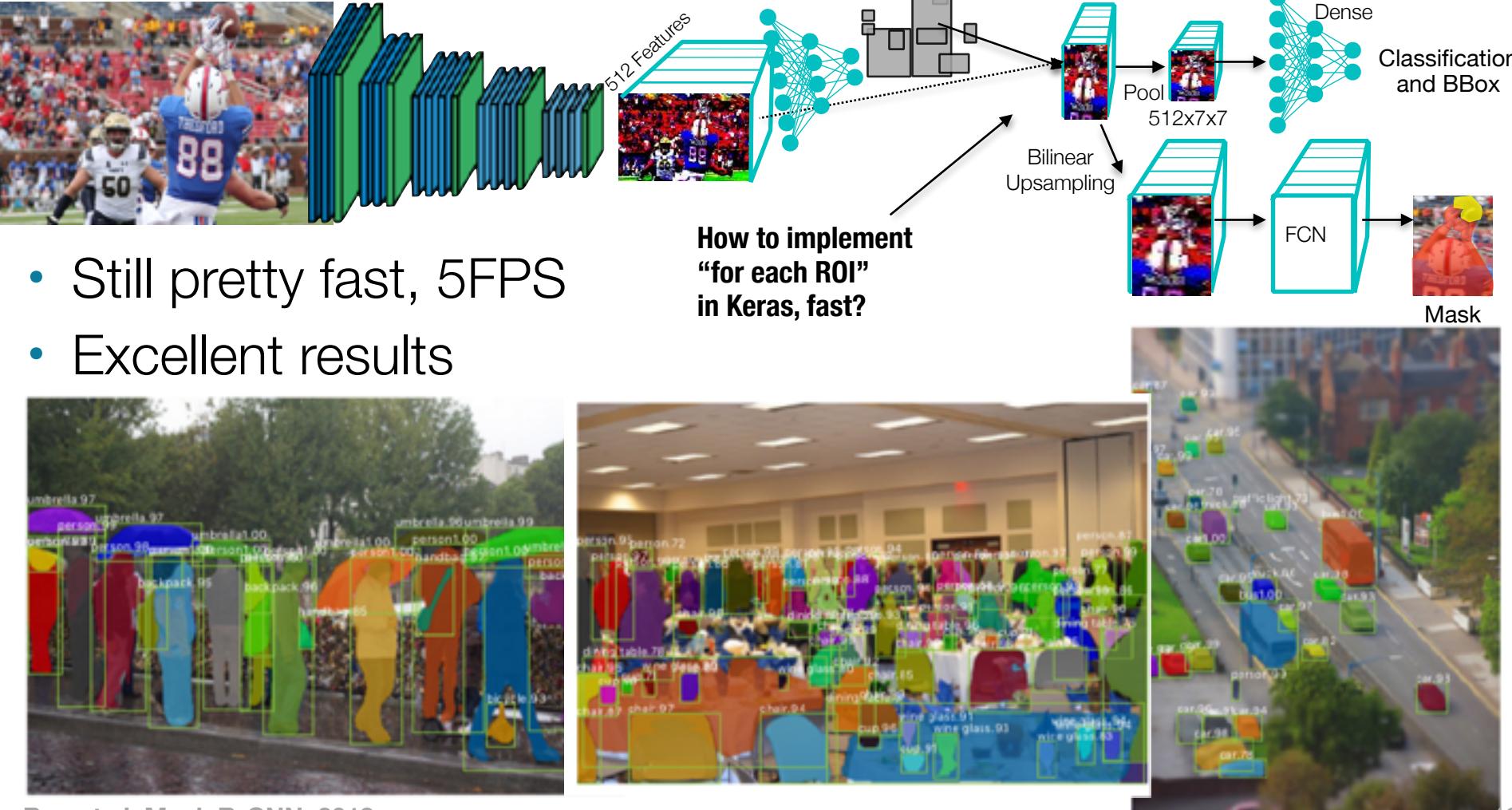
Bolya, Daniel, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. "YOLACT: real-time instance segmentation." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9157-9166. October 2019.



Instance Segmentation



2018: Mask R-CNN



Ren et al. Mask R-CNN, 2018

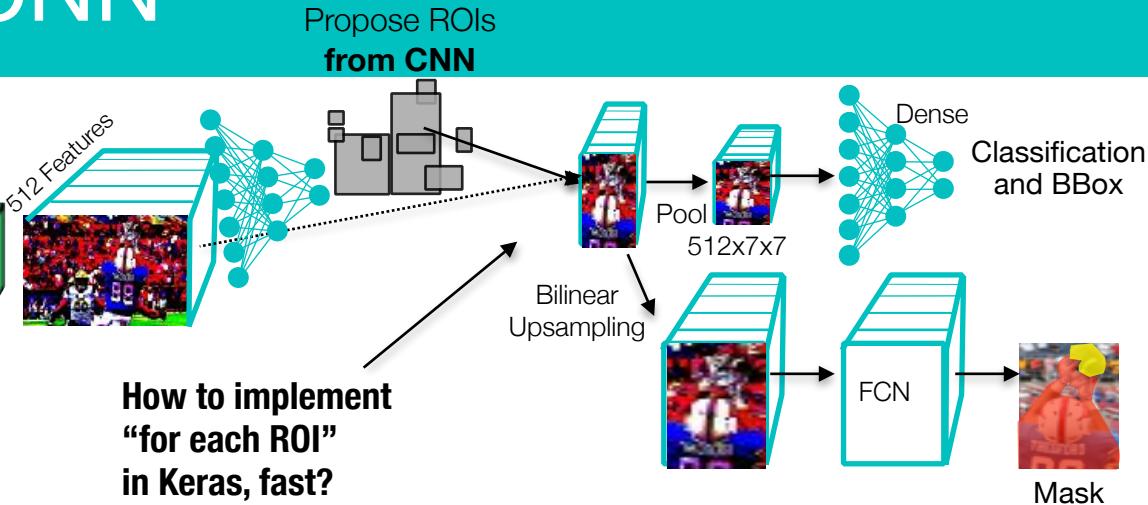


2018: Mask R-CNN



- Still pretty fast, 5FPS
- Excellent results

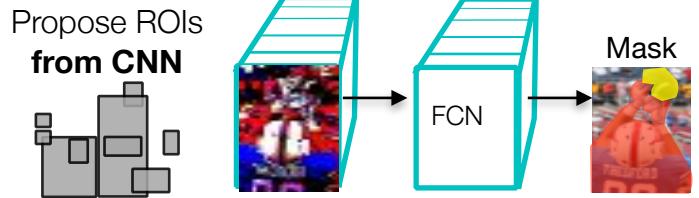
An Excellent, well documented Implementation here:
[https://github.com/matterport/Mask RCNN](https://github.com/matterport/Mask_RCNN)



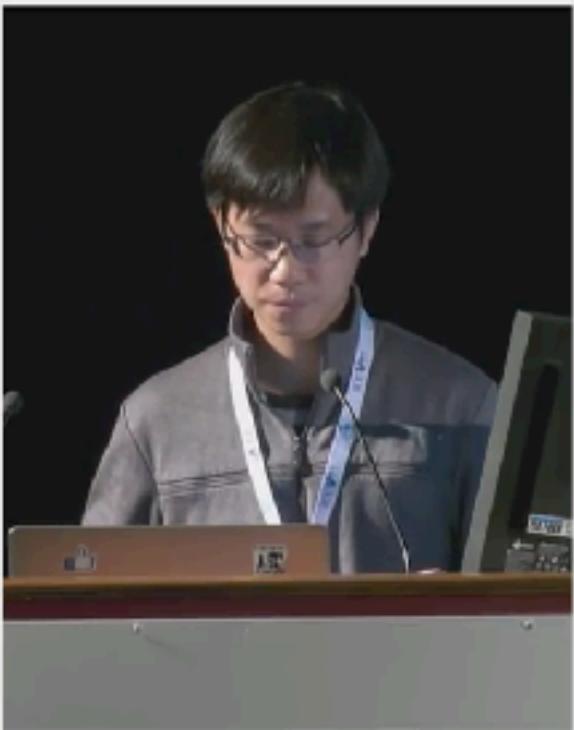
```
# Use shape of first image. Images in a batch must have the same size.  
image_shape = parse_image_meta_graph(image_meta)['image_shape'][0]  
# Equation 1 in the Feature Pyramid Networks paper. Account for  
# the fact that our coordinates are normalized here.  
# e.g. a 224x224 ROI (in pixels) maps 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(
```



2018: Mask R-CNN



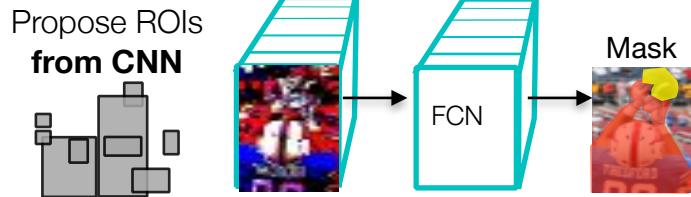
Can also provide **key point detection** from same FCN features (not real time, post processed)



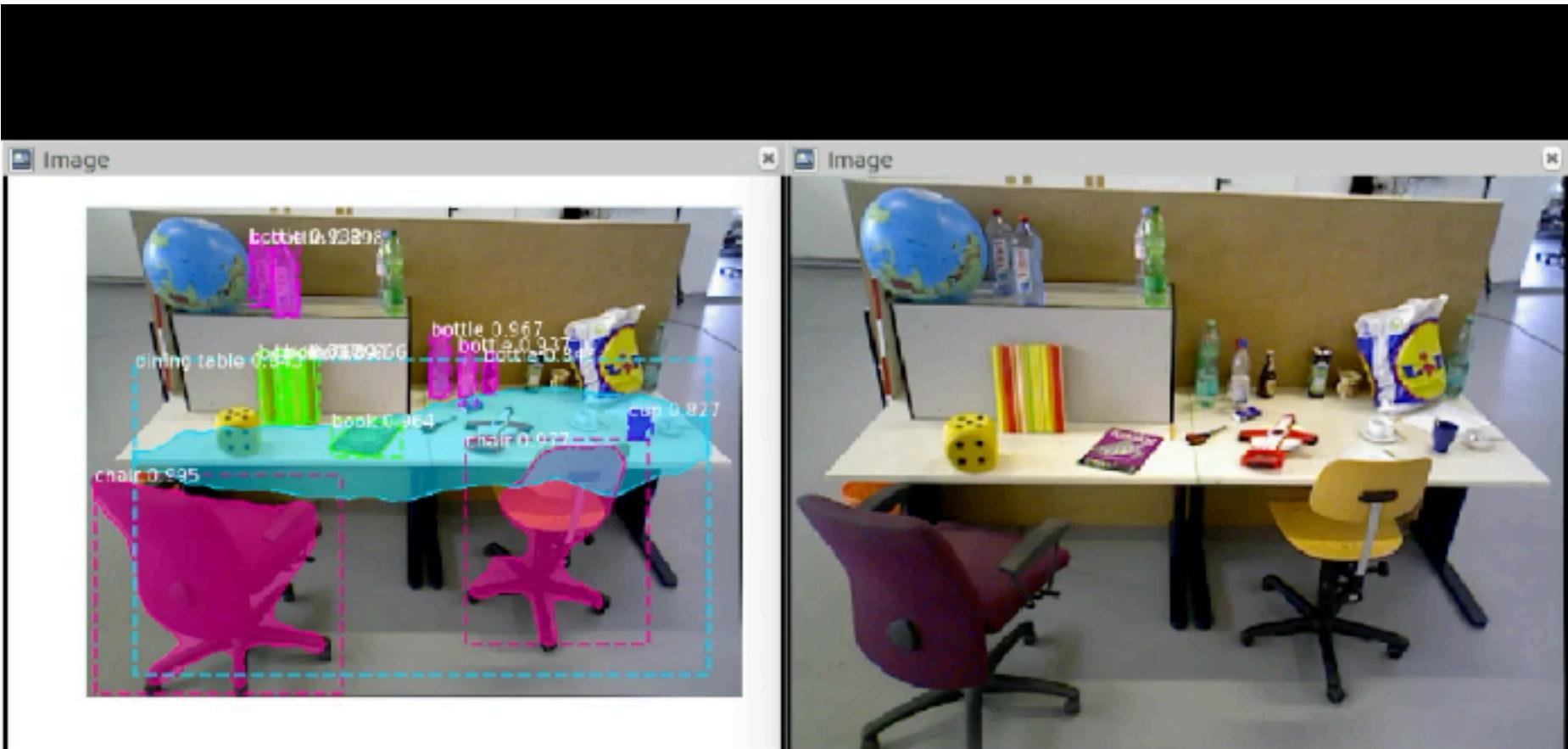
ICCV'17



2018: Mask R-CNN



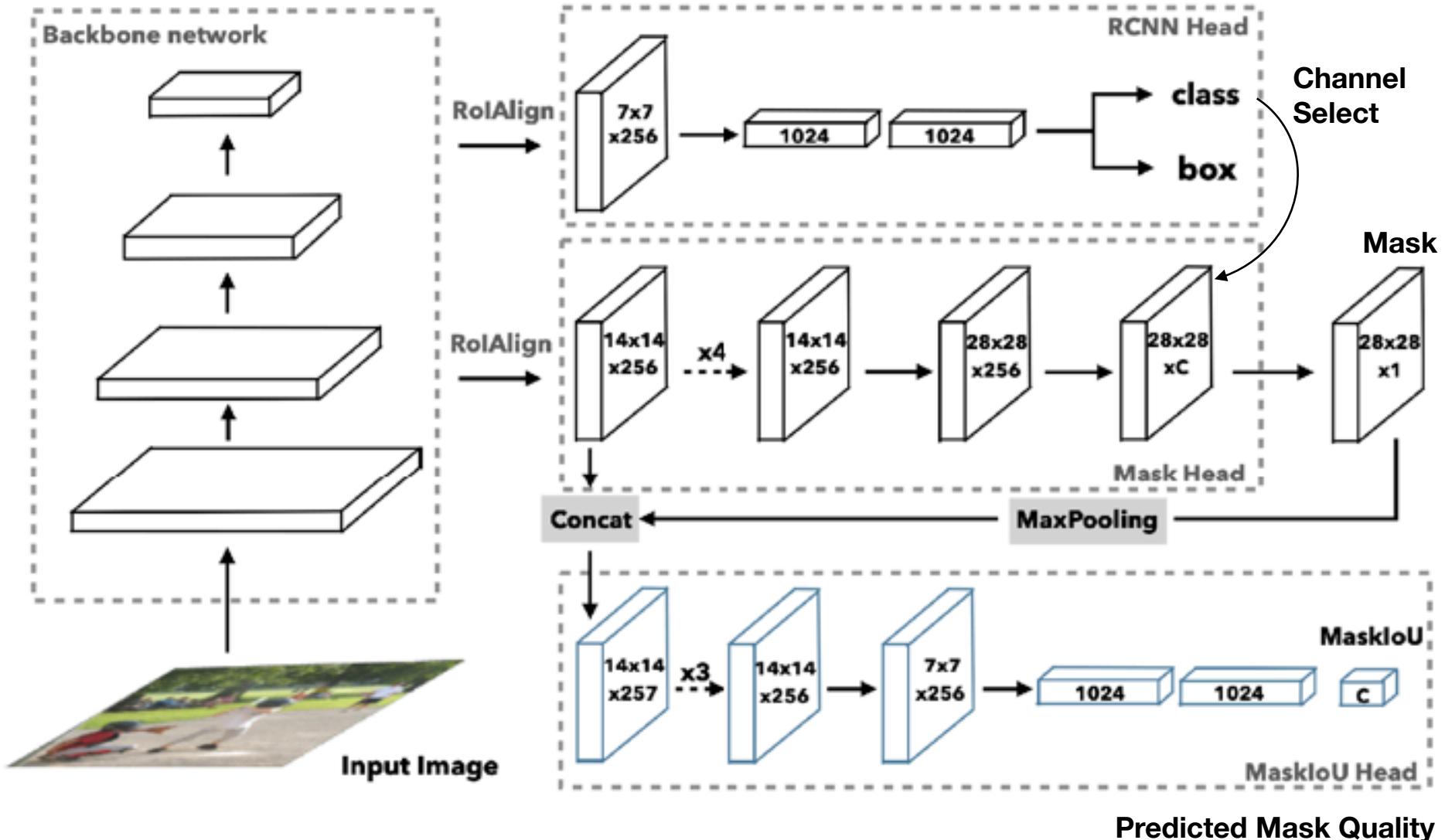
- Real time, Mask R-CNN



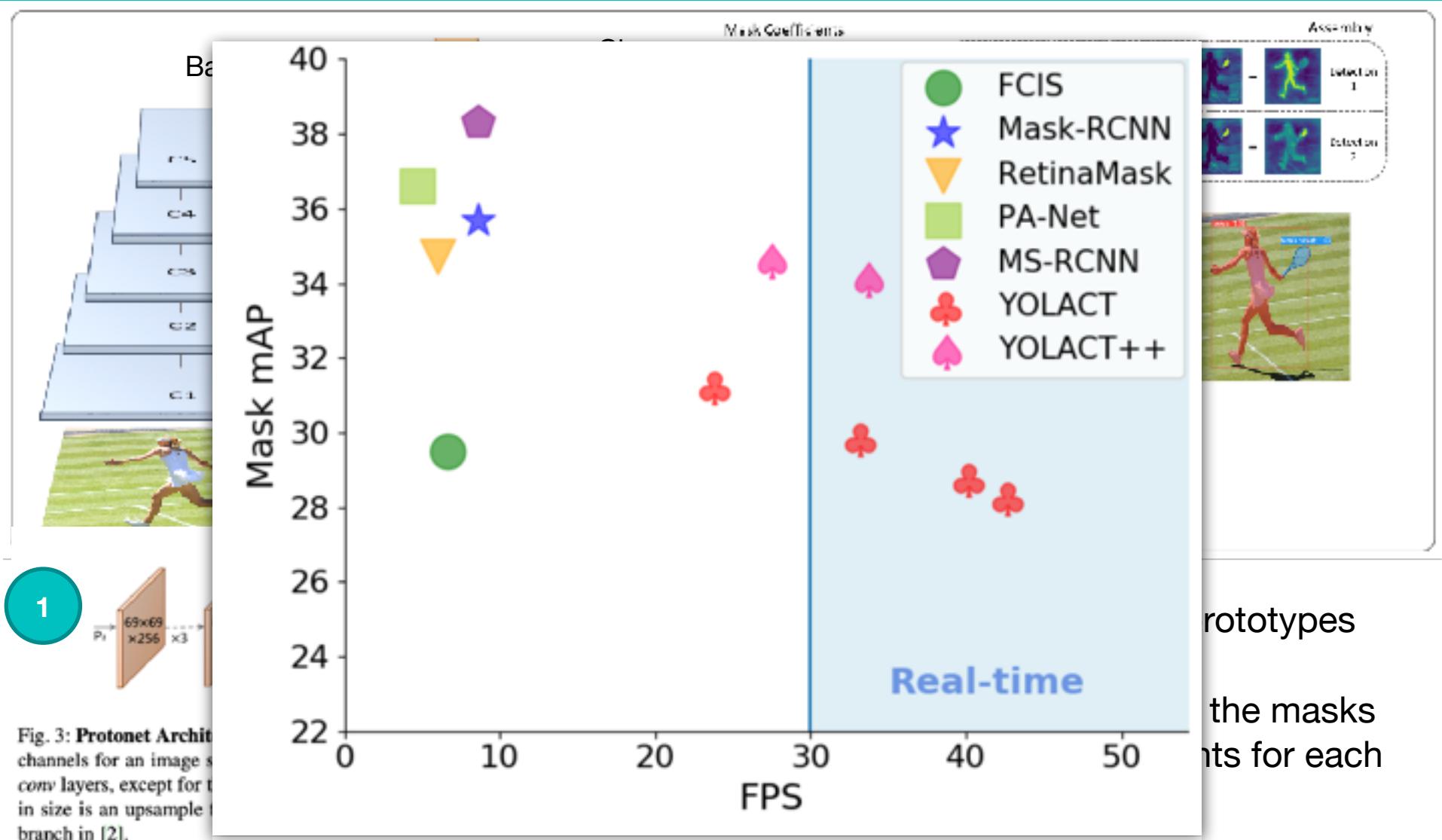
<https://www.youtube.com/watch?v=nEug0-pD0Ms>



March 2019: Mask Scoring RCNN, MS-RCNN



December 2019: YOLACT++

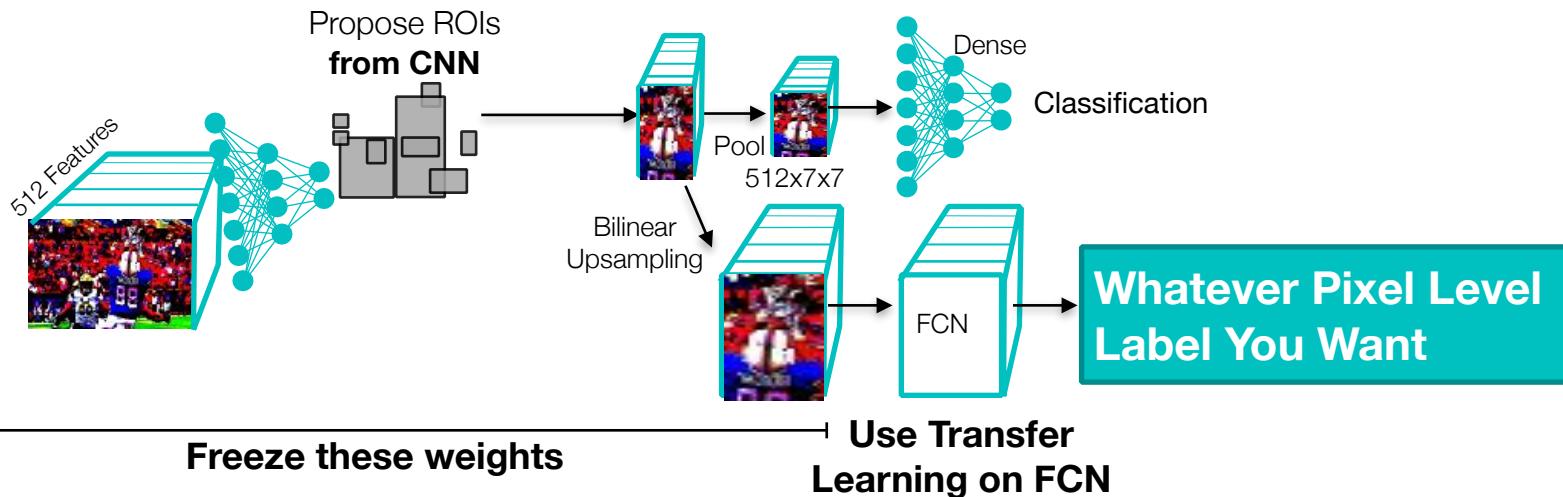


YOLACT++ Example Video (Real time)



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 connect 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)



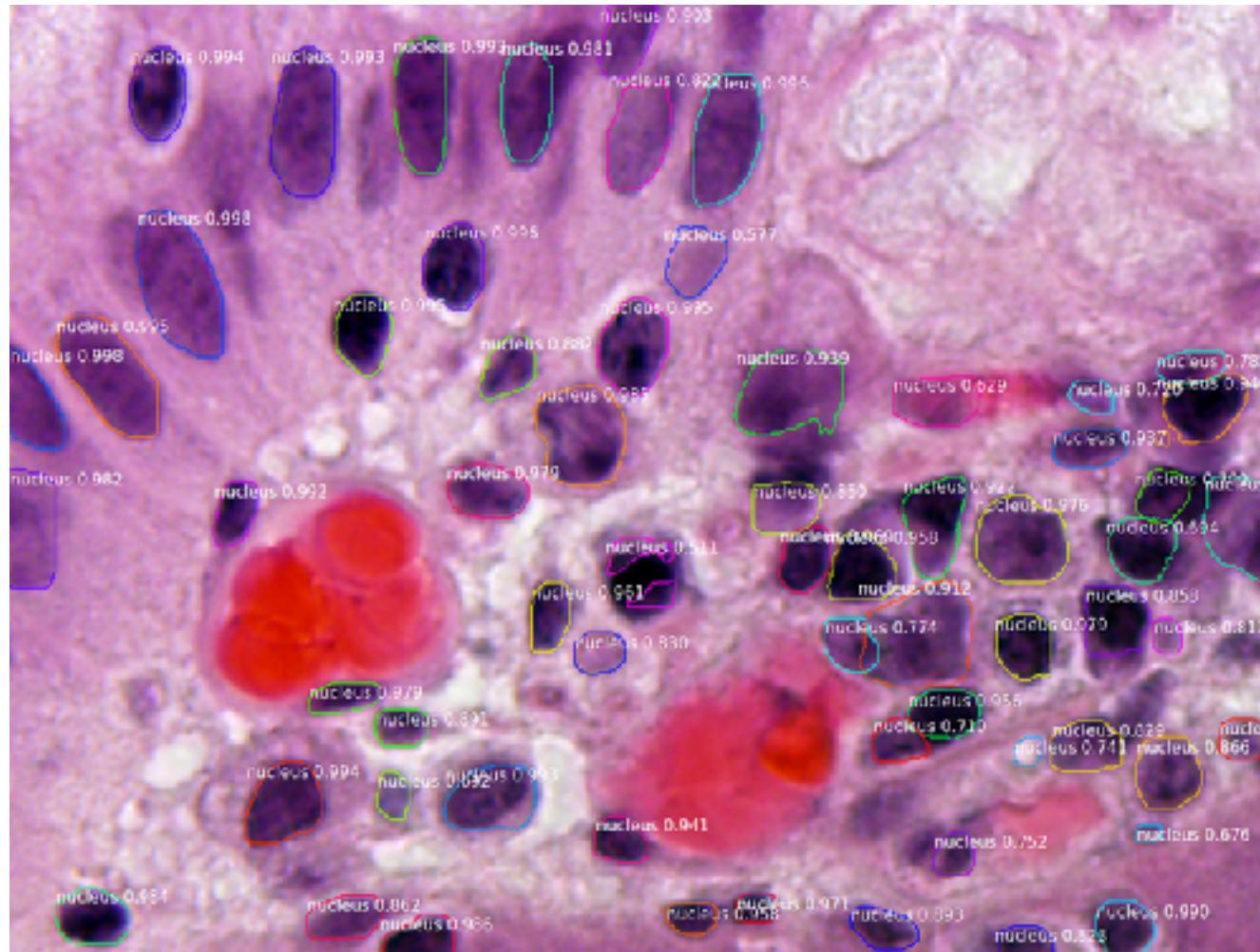
https://github.com/matterport/Mask_RCNN

21



Retraining Masks

Segmenting Nuclei

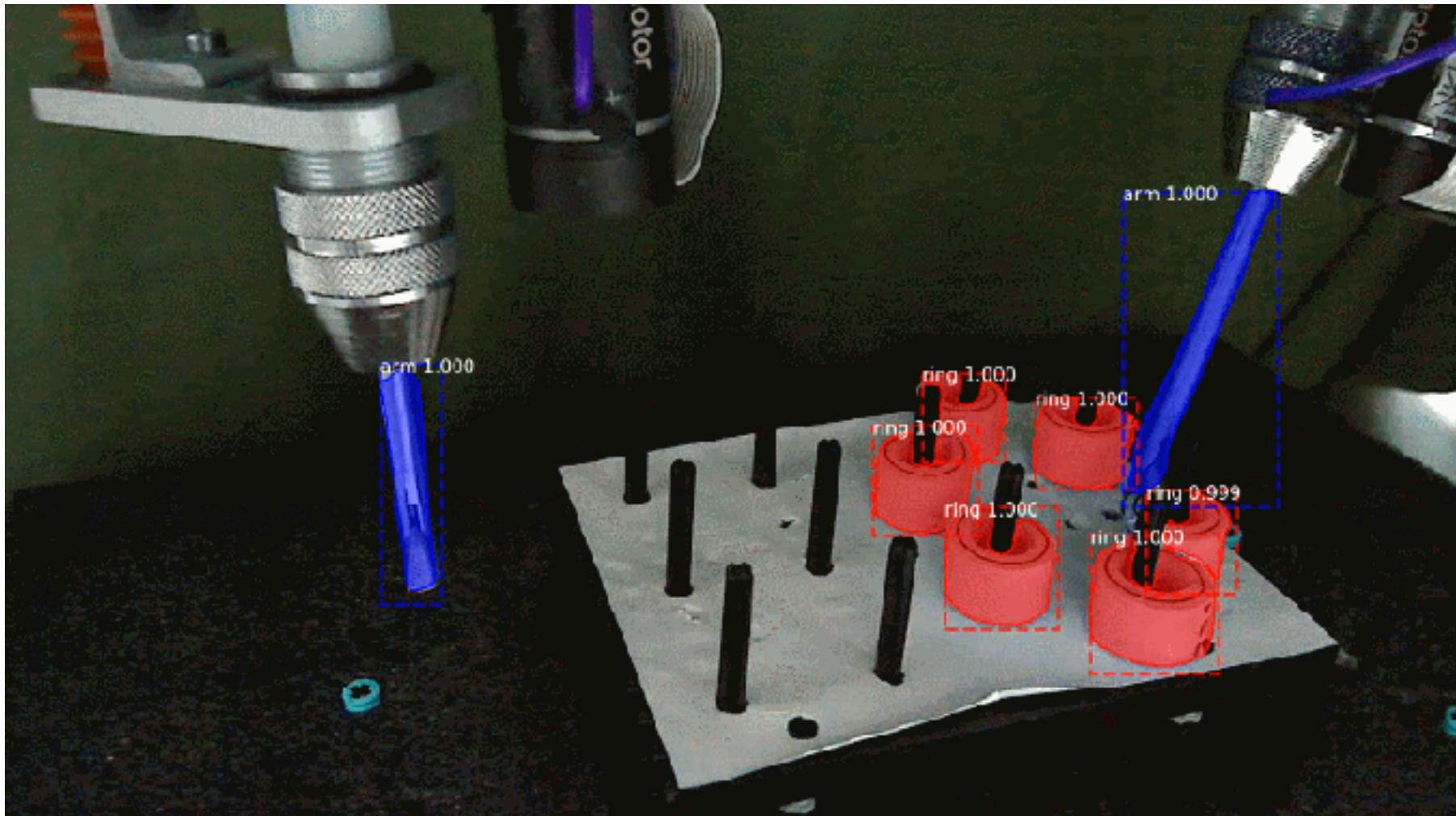


https://github.com/matterport/Mask_RCNN



Repurposing for Robotics

Robotic Movement (like surgery)

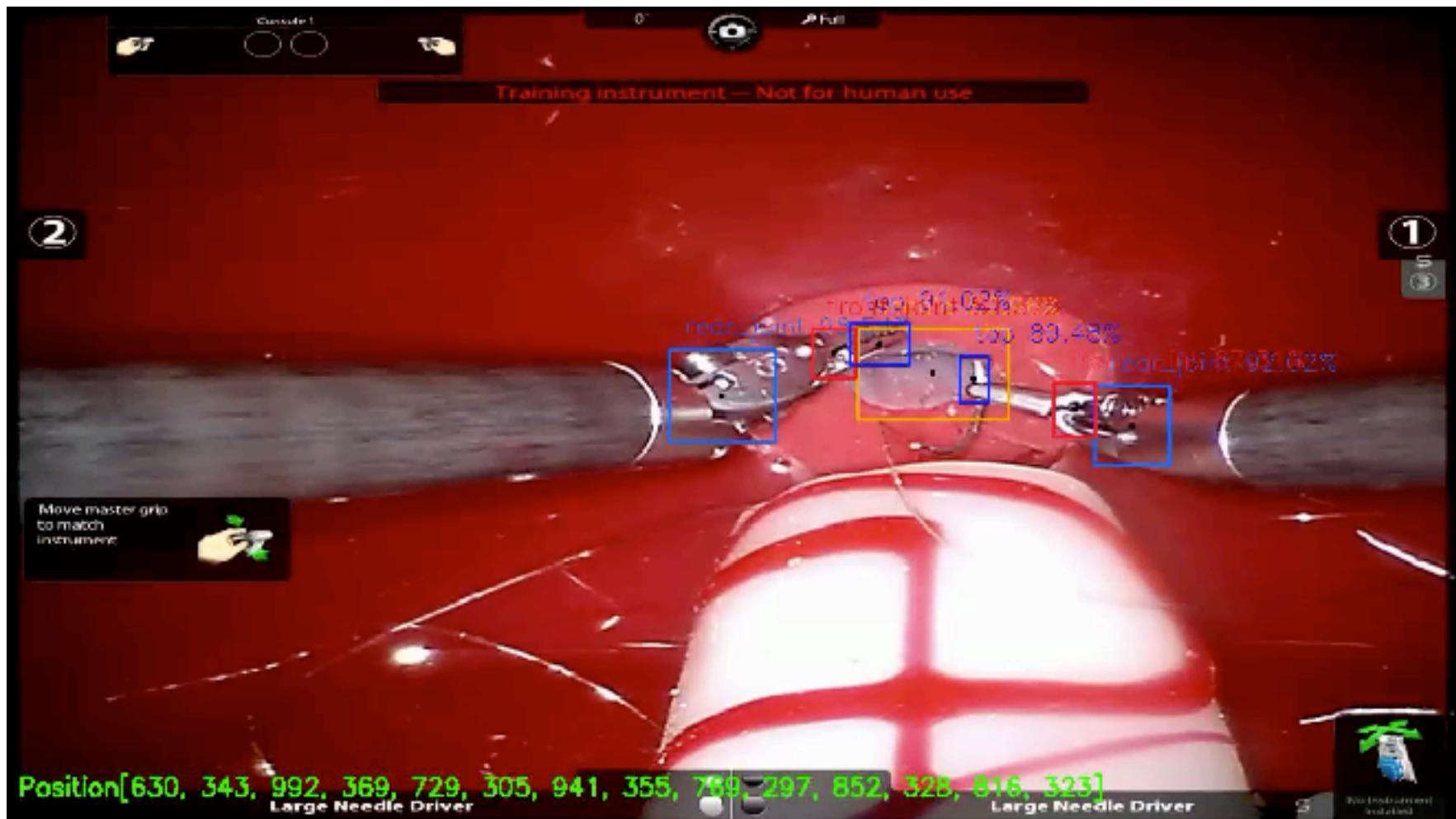


https://github.com/matterport/Mask_RCNN

23



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.

Y. Wang, J. Dai, T. Morgan, M. Elsaid, A. Garbens, X. Qu, R. Steinberg, J. Gahan, and E.C. Larson (2021). "Evaluating Robotic-Assisted Surgery Training Videos with Multi-task Convolutional Neural Networks." Journal of Robotic Surgery (JORS), 2021. Doi: 10.1007/s11701-021-01316-2

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/>

25

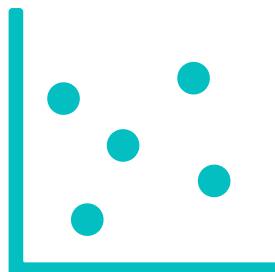
In summary

- **Semantic segmentation** through FCN is active research area
 - DeepLabV3+ or GSCNN are excellent choices, but have performance issues that make instance/object more desirable alternatives
- **Object detection** 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 10 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**

FCN: Instance Segmentation



Next Time:
Image Style Transfer
Reading: Chollet 8.1– 8.3

