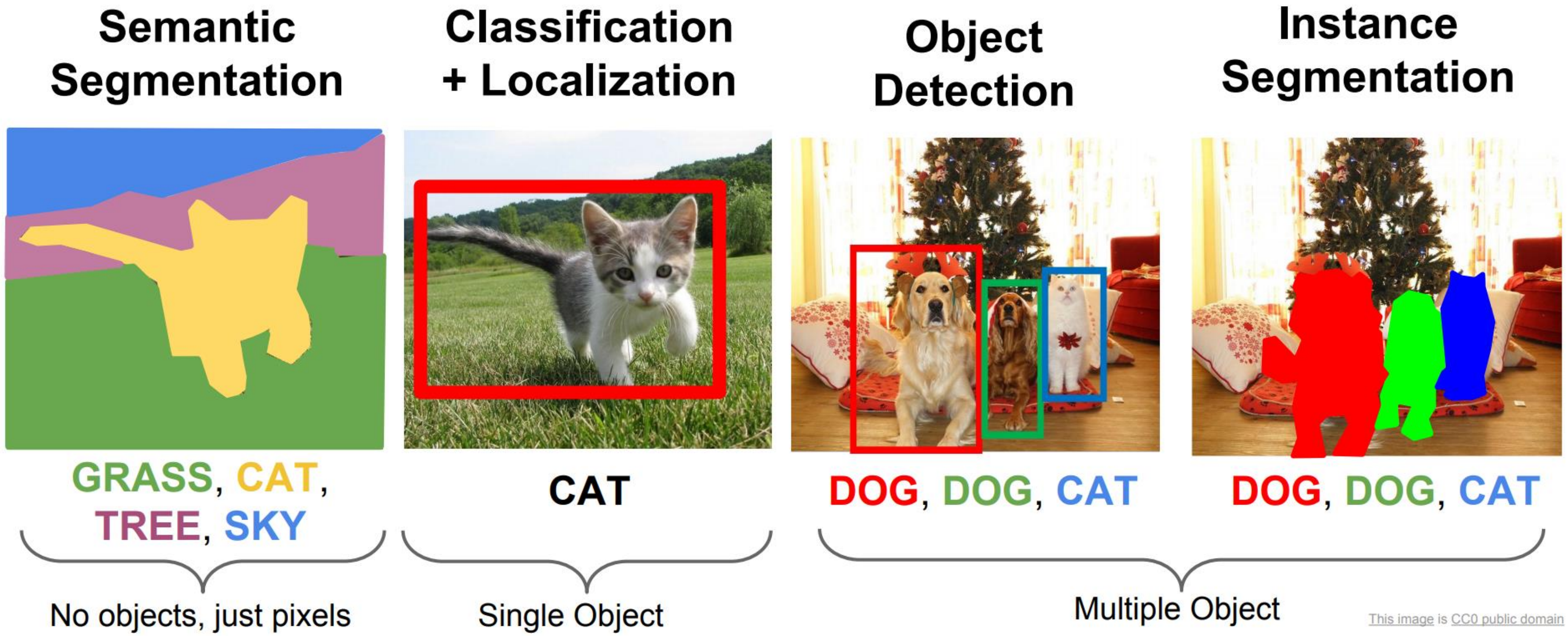
Mask R-CNN

By Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross Girshick

Presented By Aditya Sanghi

Types of Computer Vision Tasks



http://cs231n.stanford.edu/

Semantic vs Instance Segmentation

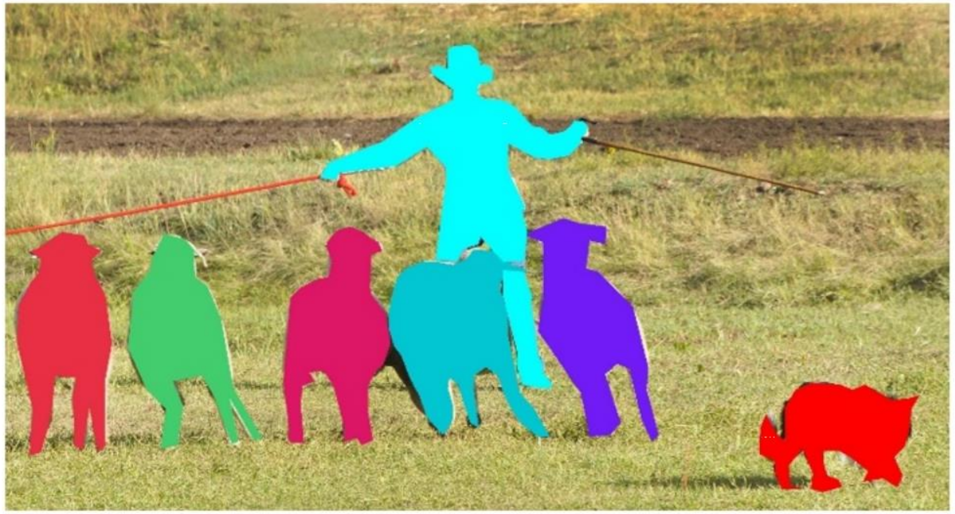
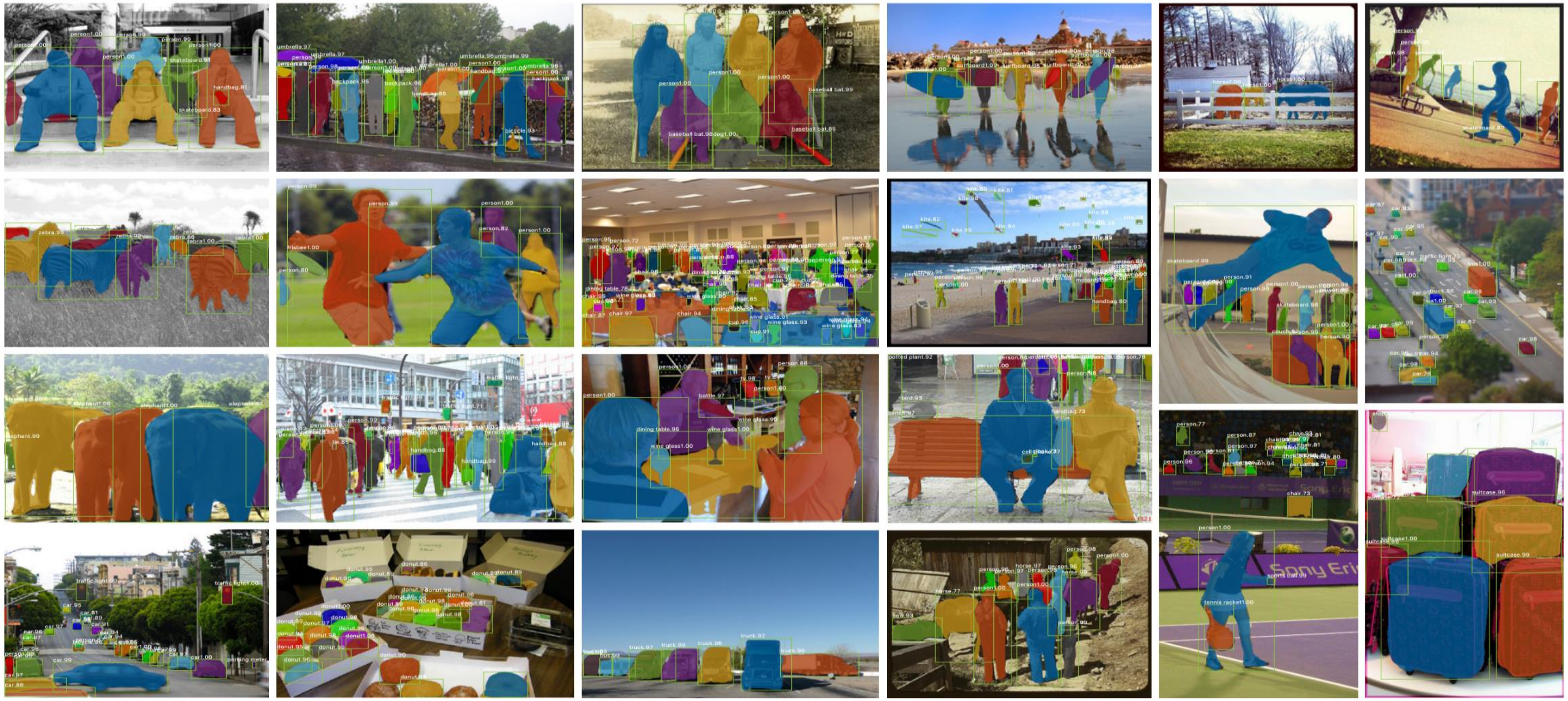


Image Source: https://arxiv.org/pdf/1405.0312.pdf

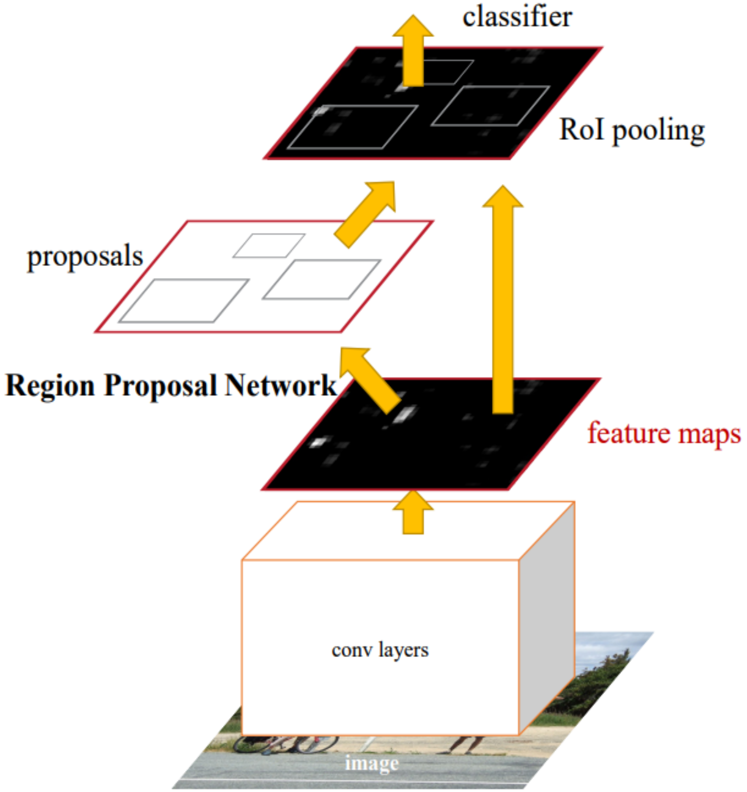
Overview of Mask R-CNN

- Goal: to create a framework for Instance segmentation
- Builds on top of Faster R-CNN by adding a parallel branch
- For each Region of Interest (RoI) predicts segmentation mask using a small FCN
- Changes RoI pooling in Faster R-CNN to a quantization-free layer called RoI Align
- Generate a binary mask for each class independently: decouples segmentation and classification
- Easy to generalize to other tasks: Human pose detection
- Result: performs better than state-of-art models in instance segmentation, bounding box detection and person keypoint detection

Some Results



Background - Faster R-CNN



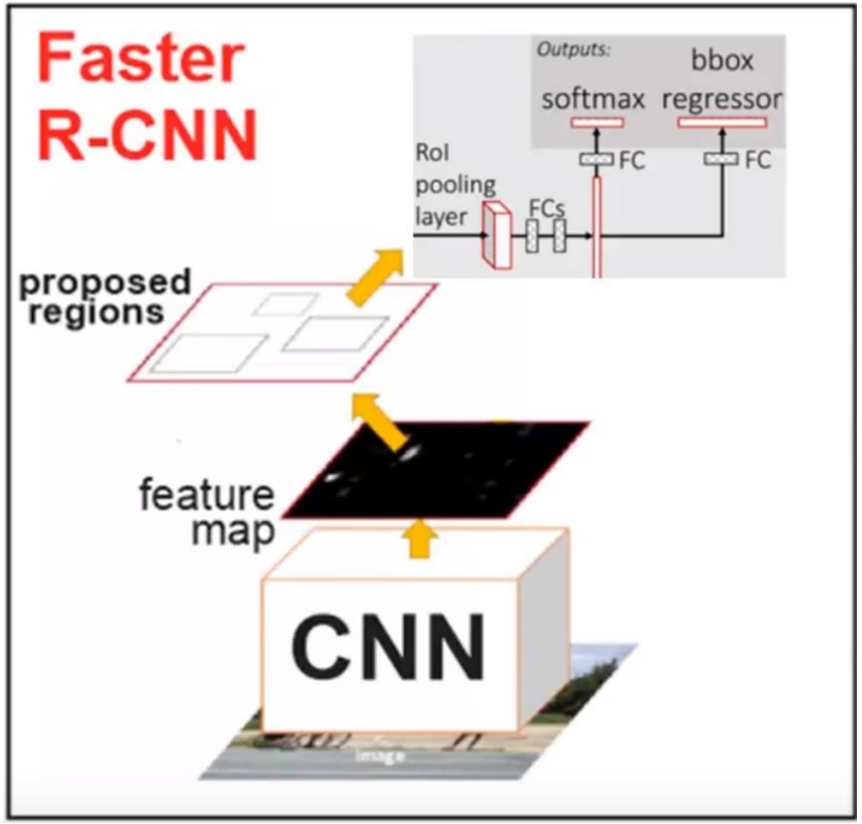


Image Source: https://arxiv.org/pdf/1506.01497.pdf

Image Source:
https://www.youtube.com/watch?v=Ul25zSysk2A&index=1&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

Background - FCN

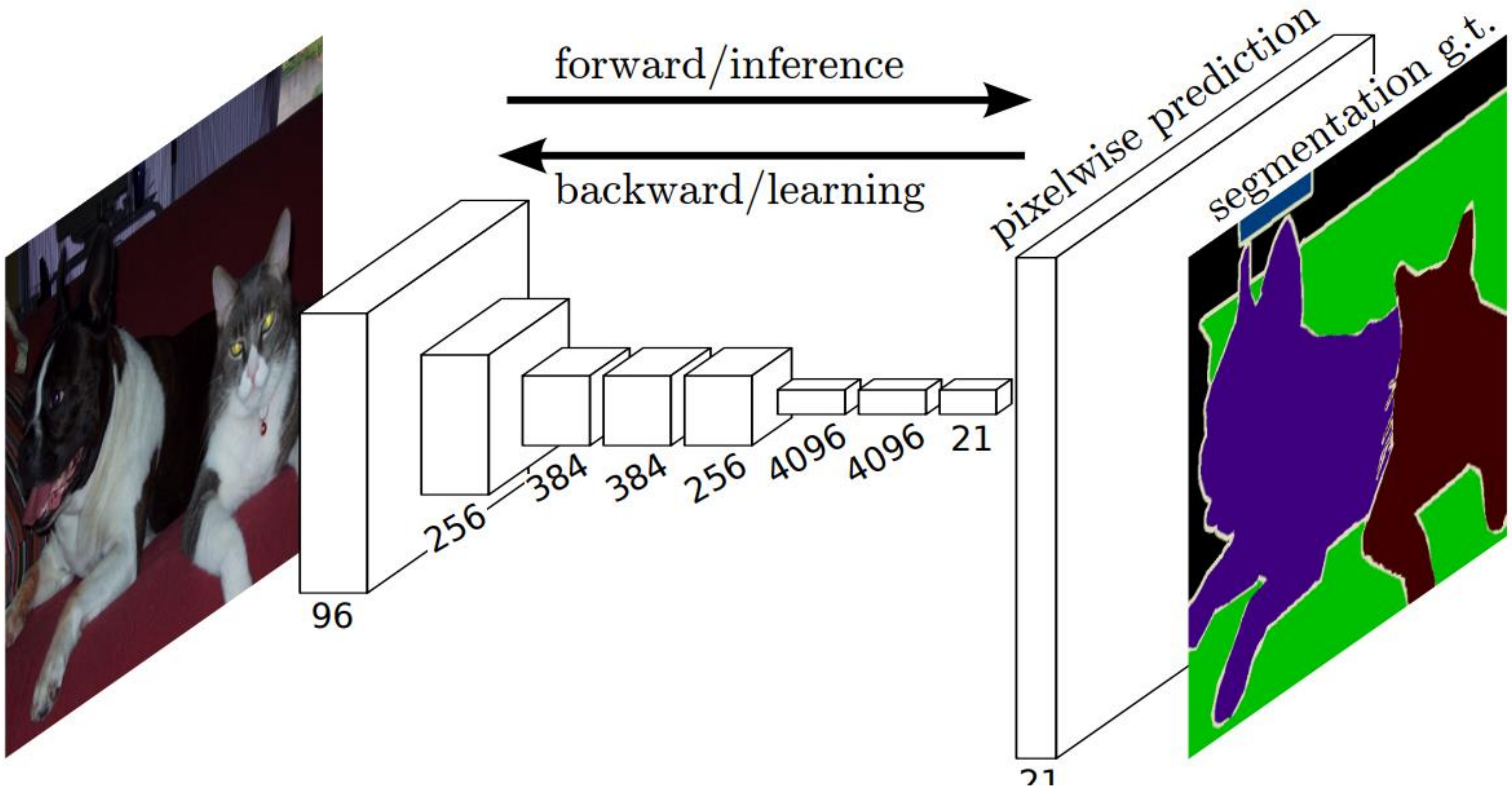


Image Source: https://arxiv.org/pdf/1411.4038.pdf

Related Work

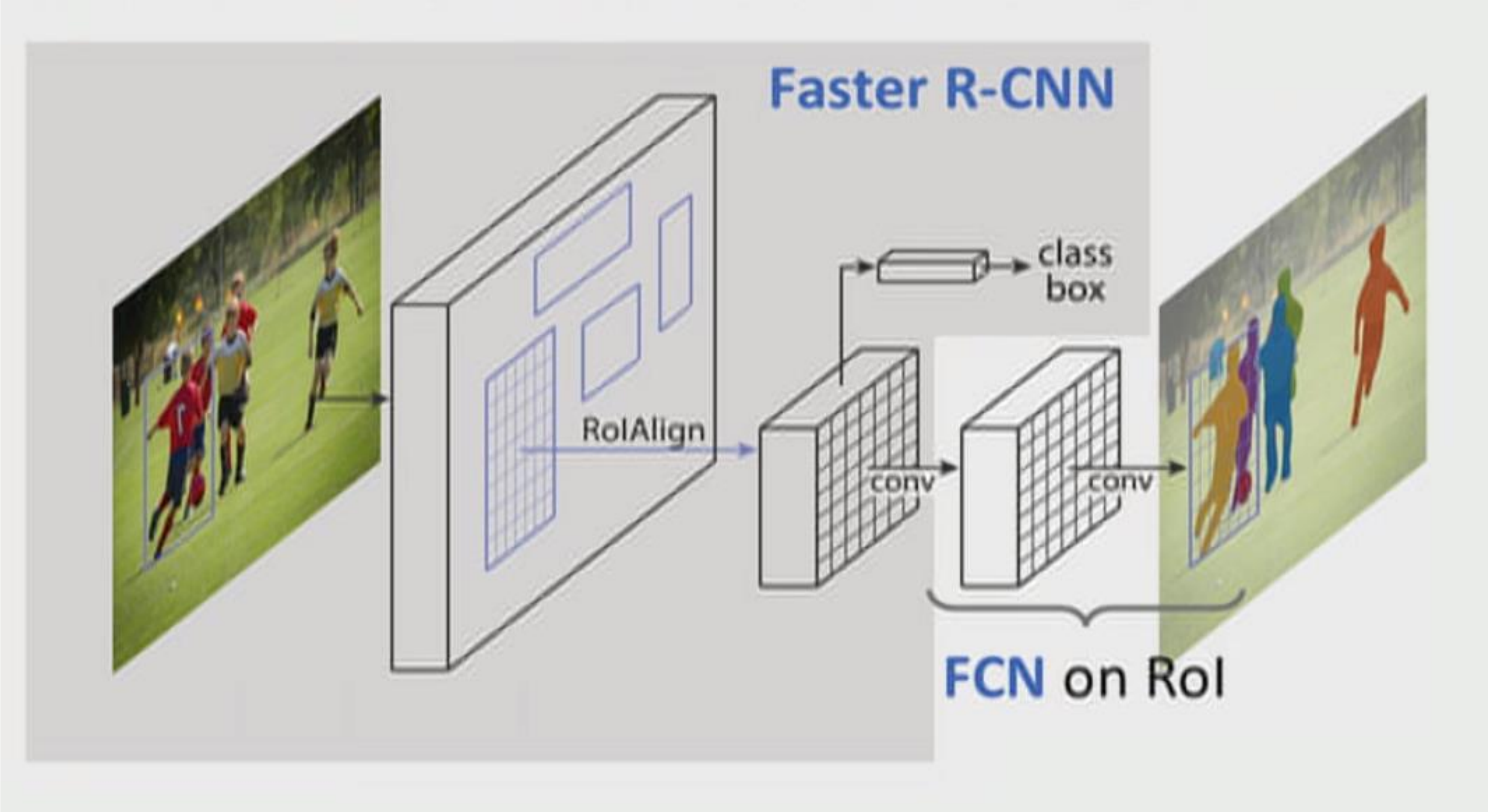


Image Source: https://www.youtube.com/watch?v=g7z4mkfRjI4

Mask R-CNN – Basic Architecture

- Procedure:
  - RPN
  - RoI Align
  - Parallel prediction for the class, box and binary mask for each RoI
- Segmentation is different from most prior systems where classification depends on mask prediction
- Loss function for each sampled RoI

$$L = L_{cls} + L_{box} + L_{mask}$$

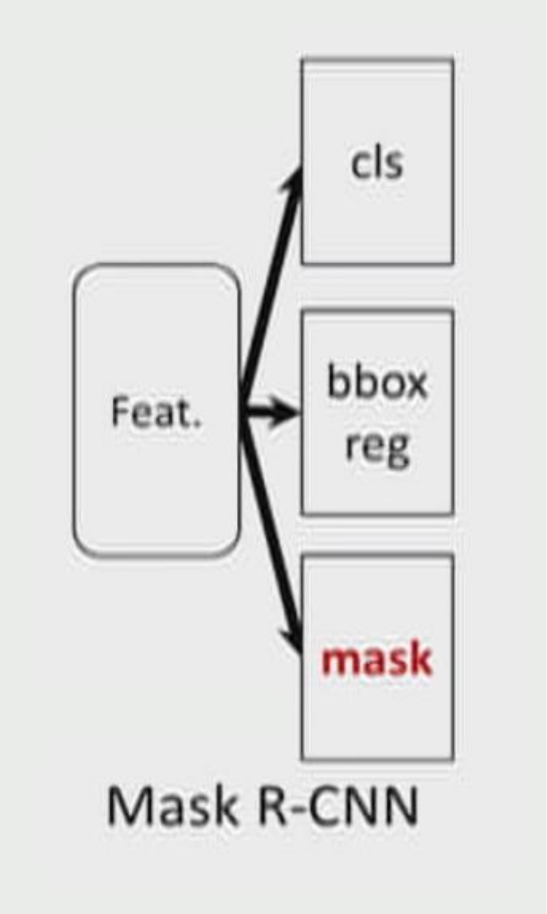
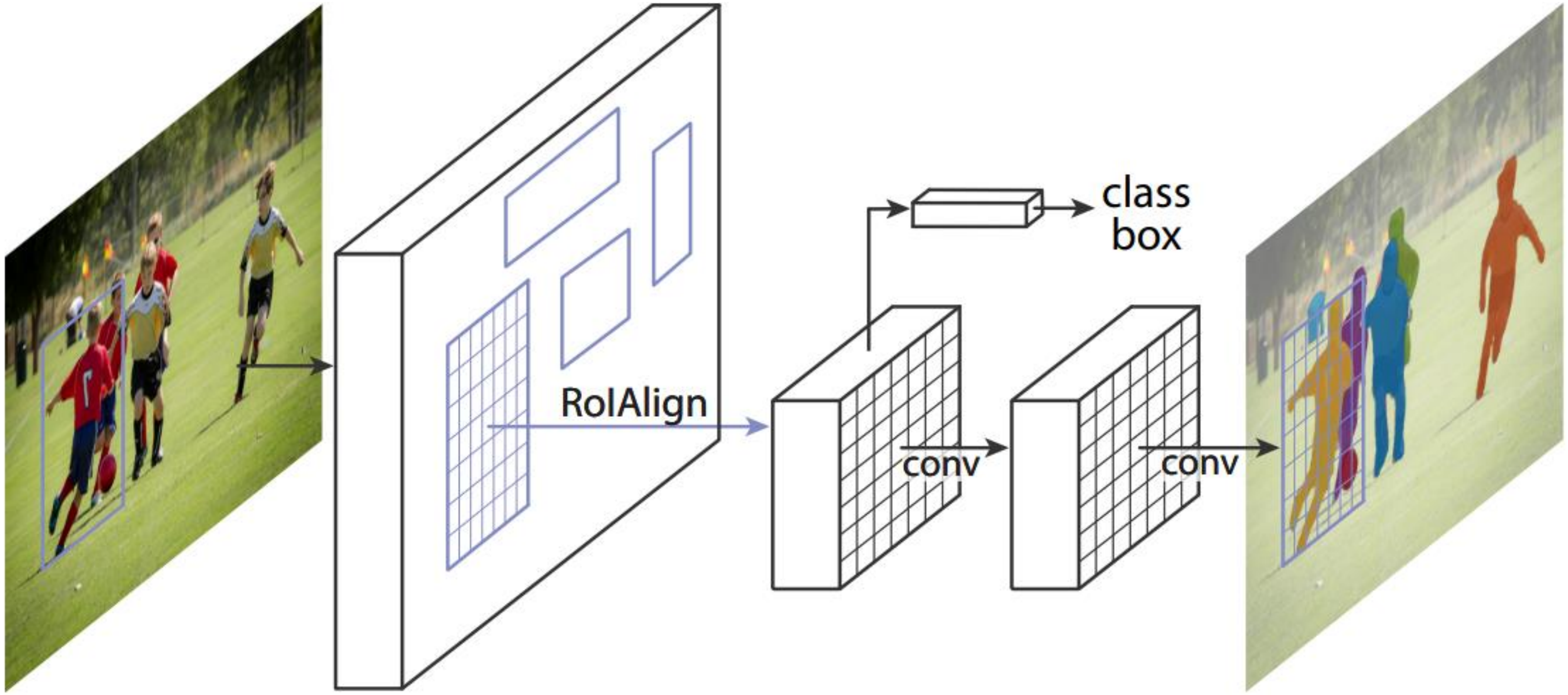


Image Source: https://www.youtube.com/watch?v=g7z4mkfRjI4

Mask R-CNN Framework



RoI Align – Motivation

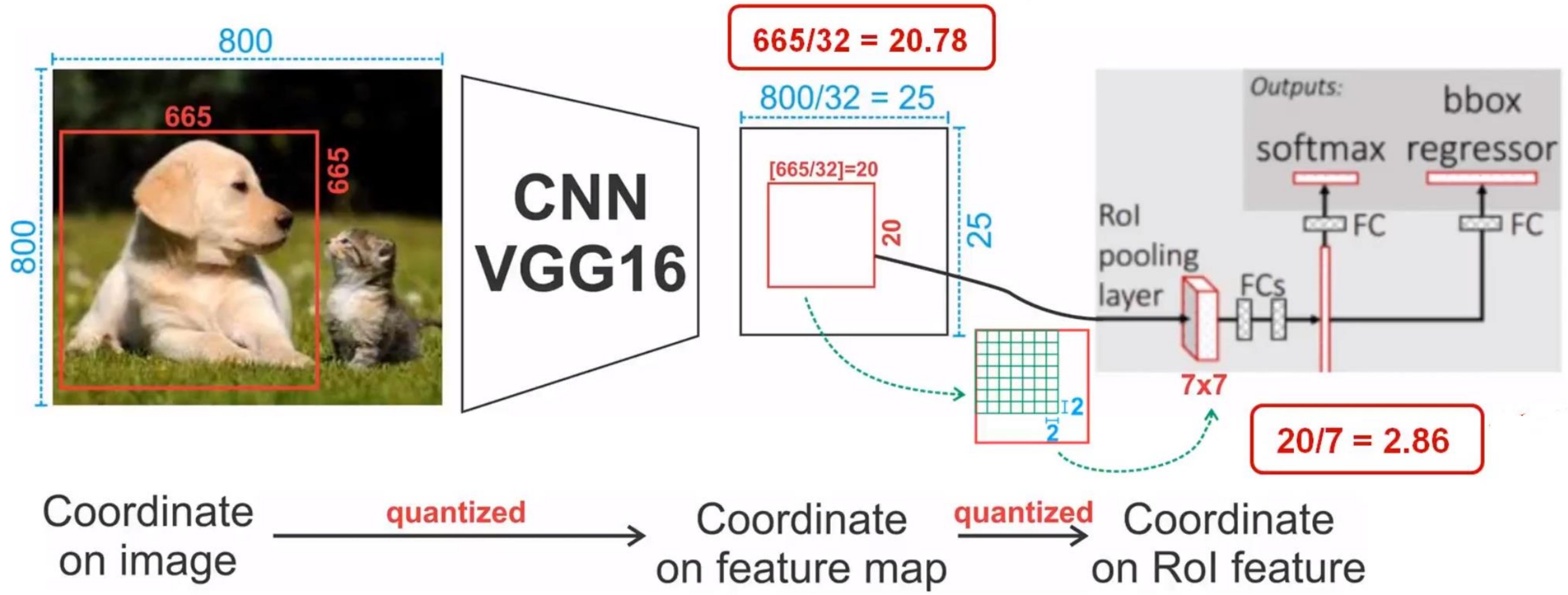
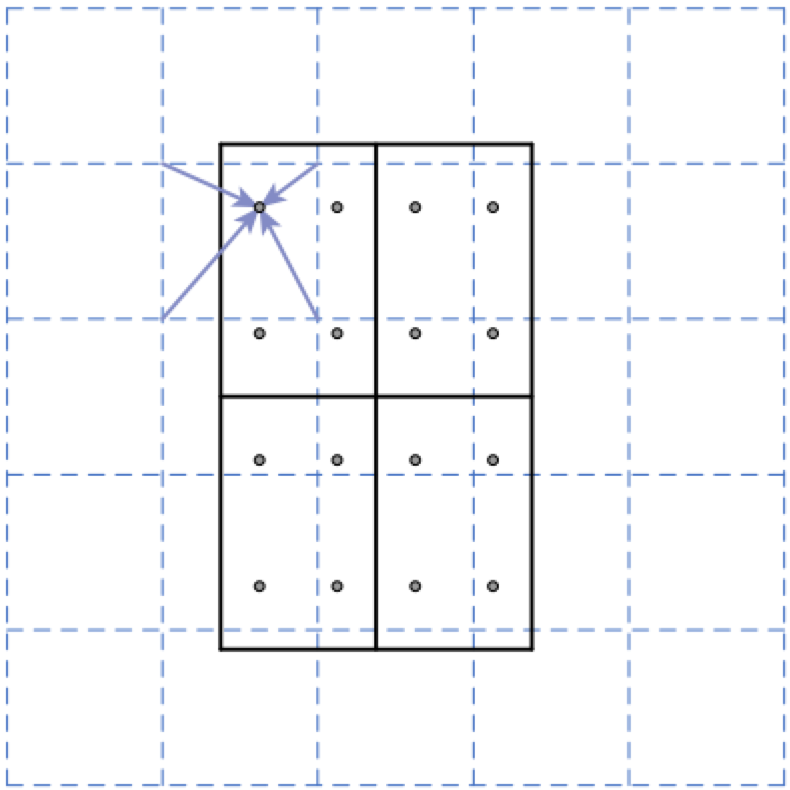


Image Source: https://www.youtube.com/watch?v=Ul25zSysk2A&index=1&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

RoI Align

- Removes this quantization which is causes this misalignment
- For each bin, you regularly sample 4 locations and do bilinear interpolation
- Result are not sensitive to exact sampling location or the number of samples
- Compare results with RoI wrapping: Which basically does bilinear interpolation on feature map only



RoI Align

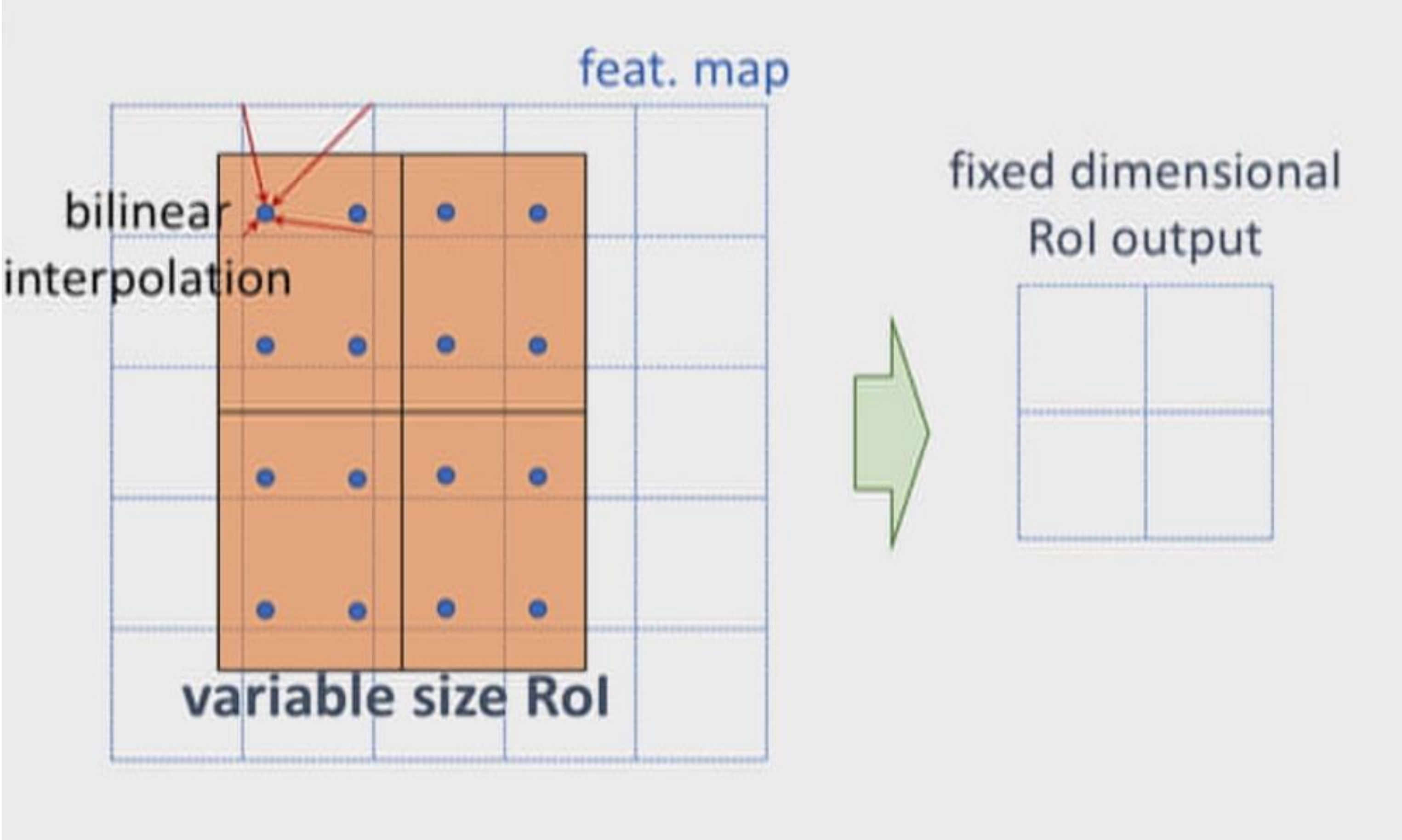


Image Source: https://www.youtube.com/watch?v=g7z4mkfRjI4

RoI Align – Results

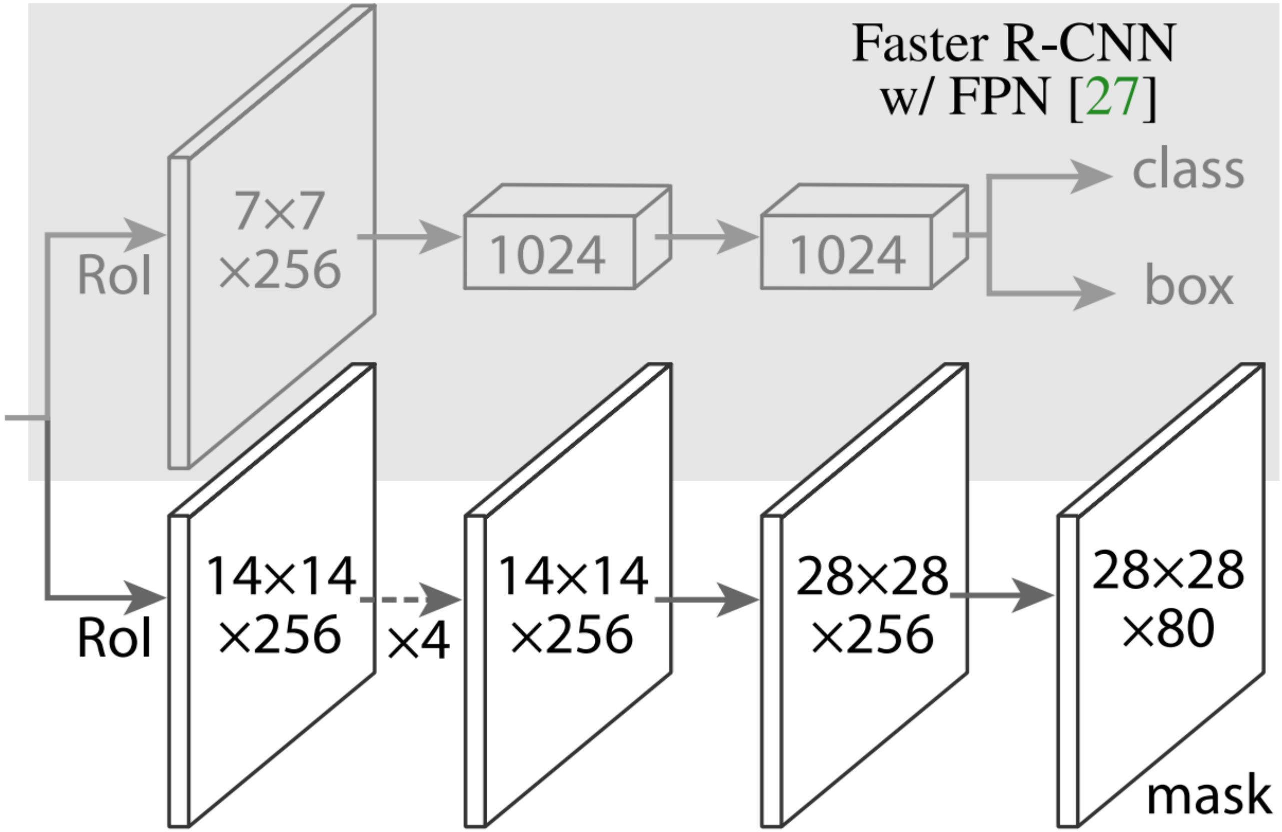
	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

(a) RoIAlign (ResNet-50-C4) comparison

	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
RoIPool	23.6	46.5	21.6	28.2	52.7	26.9
RoIAlign	**30.9**	**51.8**	**32.1**	**34.0**	**55.3**	**36.4**
	+*7.3*	+ *5.3*	+*10.5*	+*5.8*	+*2.6*	+*9.5*

(b) RoIAlign (ResNet-50-C5, stride 32) comparison

FCN Mask Head



Loss Function

$$L = L_{cls} + L_{box} + L_{mask}$$

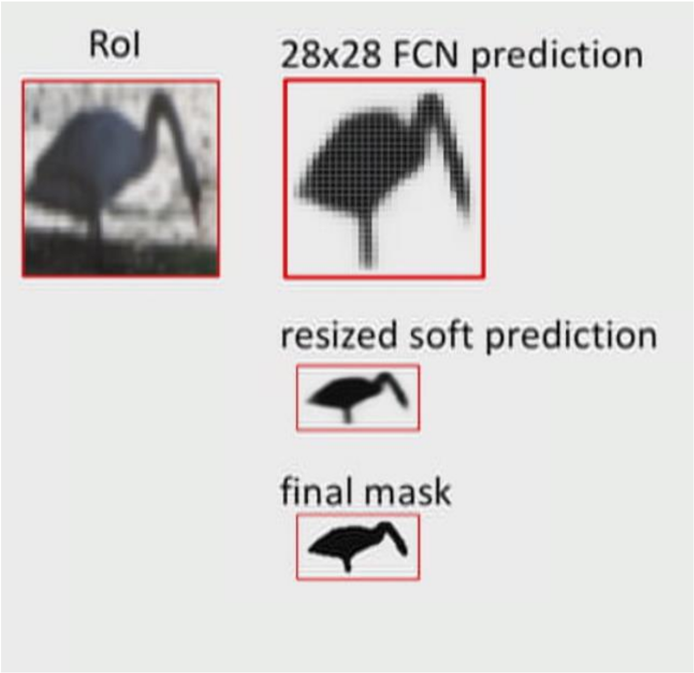
- Loss for classification and box regression is same as Faster R-CNN
- To each map a per-pixel sigmoid is applied
- The map loss is then defined as average binary cross entropy loss
- Mask loss is only defined for the ground truth class
- Decouples class prediction and mask generation
- Empirically better results and model becomes easier to train

Loss Function - Results

	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(a) Multinomial vs. Independent Masks

Mask R-CNN at Test Time



https://www.youtube.com/watch?v=g7z4mkfRjI4

Network Architecture

- Can be divided into two-parts:
  - Backbone architecture : Used for feature extraction
  - Network Head: comprises of object detection and segmentation parts
- Backbone architecture:
  - ResNet
  - ResNeXt: Depth 50 and 101 layers
  - Feature Pyramid Network (FPN)
- Network Head: Use almost the same architecture as Faster R-CNN but add convolution mask prediction branch

Implementation Details

- Same hyper-parameters as Faster R-CNN
- Training:
  - RoI positive if IoU is atleast 0.5; Mask loss is defined only on positive RoIs
  - Each mini-batch has 2 images per GPU and each image has N sampled RoI
  - N is 64 for C4 backbone and 512 for FPN
  - Train on 8 GPUs for 160k iterations
  - Learning rate of 0.02 which is decreased by 10 at 120k iterataions
- Inference:
  - Proposal number 300 for C4 backbone and 1000 for FPN
  - Mask branch is applied to the highest scoring 100 detection boxes; so not done parallel at test time, this speeds up inference and accuracy
  - We also only use the kth-mask where k is the predicted class by the classification branch
  - The m x m mask is resized to the RoI Size

Main Results

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

Main Results

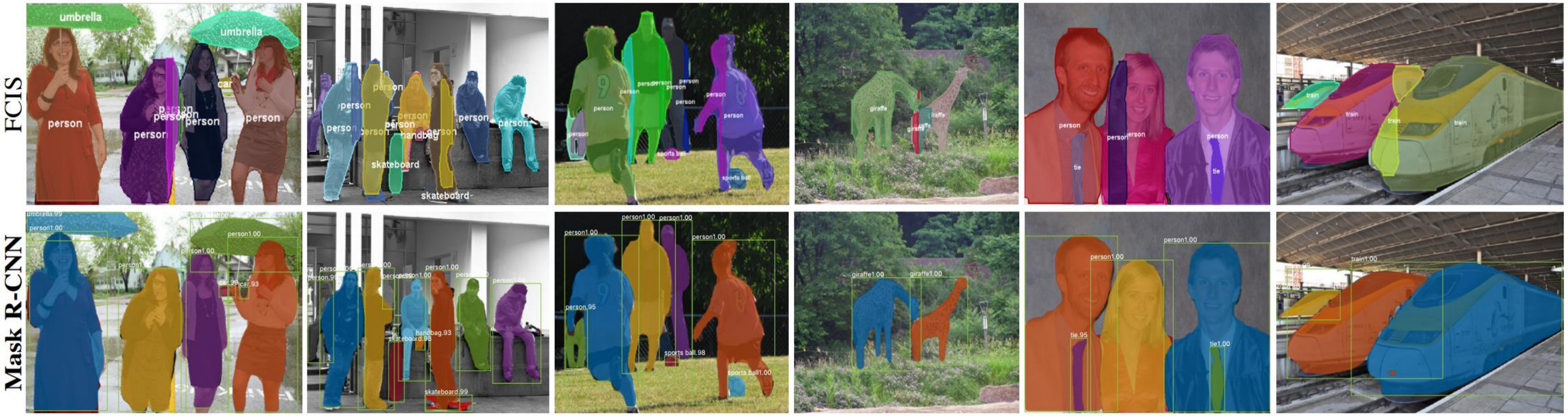


Figure 6. FCIS+++ [26] (top) *vs*. Mask R-CNN (bottom, ResNet-101-FPN). FCIS exhibits systematic artifacts on overlapping objects.

Results: FCN vs MLP

	mask branch	AP	AP_{50}	AP_{75}
MLP	fc: $1024 \rightarrow 1024 \rightarrow 80 \cdot 28^2$	31.5	53.7	32.8
MLP	fc: $1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 80 \cdot 28^2$	31.5	54.0	32.6
FCN	conv: $256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 80$	**33.6**	**55.2**	**35.3**

Main Results – Object Detection

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Faster R-CNN+++ [19]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [27]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [21]	Inception-ResNet-v2 [41]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [39]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3	19.8	40.2	48.8
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**	**22.1**	**43.2**	51.2

Mask R-CNN for Human Pose Estimation

Figure 7. Keypoint detection results on COCO `test` using Mask R-CNN (ResNet-50-FPN), with person segmentation masks predicted from the same model. This model has a keypoint AP of 63.1 and runs at 5 fps.

Mask R-CNN for Human Pose Estimation

- Model keypoint location as a one-hot binary mask
- Generate a mask for each keypoint types
- For each keypoint, during training, the target is a $m \ x \ m$ binary map where only a single pixel is labelled as foreground
- For each visible ground-truth keypoint, we minimize the cross-entropy loss over a m^2-way softmax output

Results for Pose Estimation

	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
CMU-Pose+++ [6]	61.8	84.9	67.5	57.1	68.2
G-RMI [32]†	62.4	84.0	68.5	**59.1**	68.1
Mask R-CNN, keypoint-only	62.7	87.0	68.4	57.4	71.1
Mask R-CNN, keypoint & mask	**63.1**	**87.3**	**68.7**	57.8	**71.4**

(a) Keypoint detection AP on COCO test-dev

	AP^{bb}_{person}	AP^{mask}_{person}	AP^{kp}
Faster R-CNN	52.5	-	-
Mask R-CNN, mask-only	**53.6**	**45.8**	-
Mask R-CNN, keypoint-only	50.7	-	64.2
Mask R-CNN, keypoint & mask	52.0	45.1	**64.7**

(b) Multi-task learning

	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
RoIPool	59.8	86.2	66.7	55.1	67.4
RoIAlign	**64.2**	**86.6**	**69.7**	**58.7**	**73.0**

(c) RoIAlign vs. RoIPool

Experiments on Cityscapes

Experiments on Cityscapes

	training data	AP [val]	AP	AP_{50}	person	rider	car	truck	bus	train	mcycle	bicycle
InstanceCut [23]	fine + coarse	15.8	13.0	27.9	10.0	8.0	23.7	14.0	19.5	15.2	9.3	4.7
DWT [4]	fine	19.8	15.6	30.0	15.1	11.7	32.9	17.1	20.4	15.0	7.9	4.9
SAIS [17]	fine	-	17.4	36.7	14.6	12.9	35.7	16.0	23.2	19.0	10.3	7.8
DIN [3]	fine + coarse	-	20.0	38.8	16.5	16.7	25.7	20.6	30.0	23.4	17.1	10.1
SGN [29]	fine + coarse	29.2	25.0	44.9	21.8	20.1	39.4	24.8	33.2	30.8	17.7	12.4
Mask R-CNN	fine	31.5	26.2	49.9	30.5	23.7	46.9	22.8	32.2	18.6	19.1	16.0
Mask R-CNN	fine + COCO	**36.4**	**32.0**	**58.1**	**34.8**	**27.0**	**49.1**	**30.1**	**40.9**	**30.9**	**24.1**	**18.7**

Latest Results – Instance Segmentation

description	backbone	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
original baseline	X-101-FPN	36.7	59.5	38.9	39.6	61.5	43.2
+ updated baseline	X-101-FPN	37.0	59.7	39.0	40.5	63.0	43.7
+ e2e training	X-101-FPN	37.6	60.4	39.9	41.7	64.1	45.2
+ ImageNet-5k	X-101-FPN	38.6	61.7	40.9	42.7	65.1	46.6
+ train-time augm.	X-101-FPN	39.2	62.5	41.6	43.5	65.9	47.2
+ deeper	X-152-FPN	39.7	63.2	42.2	44.1	66.4	48.4
+ Non-local [43]	X-152-FPN-NL	**40.3**	**64.4**	**42.8**	**45.0**	**67.8**	**48.9**
+ test-time augm.	X-152-FPN-NL	**41.8**	**66.0**	**44.8**	**47.3**	**69.3**	**51.5**

Latest Result – Pose Estimation

description	backbone	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
original baseline	R-50-FPN	64.2	86.6	69.7	58.7	73.0
+ updated baseline	R-50-FPN	65.1	86.6	70.9	59.9	73.6
+ deeper	R-101-FPN	66.1	87.7	71.7	60.5	75.0
+ ResNeXt	X-101-FPN	67.3	88.0	73.3	62.2	75.6
+ data distillation [35]	X-101-FPN	**69.1**	**88.9**	**75.3**	**64.1**	**77.1**
+ test-time augm.	X-101-FPN	**70.4**	**89.3**	**76.8**	**65.8**	**78.1**

Future work

- Interesting direction would be to replace rectangular RoI
- Extend this to segment multiple background (sky, ground)
- Any other ideas?

Conclusion

- A framework to do state-of-art instance segmentation
- Generates high-quality segmentation mask
- Model does Object Detection, Instance Segmentation and can also be extended to human pose estimation!!!!!!
- All of them are done in parallel
- Simple to train and adds a small overhead to Faster R-CNN

Resources

- Official code: https://github.com/facebookresearch/Detectron
- TensorFlow unofficial code: https://github.com/matterport/Mask_RCNN
- ICCV17 video: https://www.youtube.com/watch?v=g7z4mkfRjI4
- Tutorial Videos: https://www.youtube.com/watch?v=Ul25zSysk2A&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

References

- https://arxiv.org/pdf/1703.06870.pdf
- https://arxiv.org/pdf/1405.0312.pdf
- https://arxiv.org/pdf/1411.4038.pdf
- https://arxiv.org/pdf/1506.01497.pdf
- http://cs231n.stanford.edu/
- https://www.youtube.com/watch?v=OOT3UIXZztE
- https://www.youtube.com/watch?v=Ul25zSysk2A&index=1&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

Thank You

Any Questions?