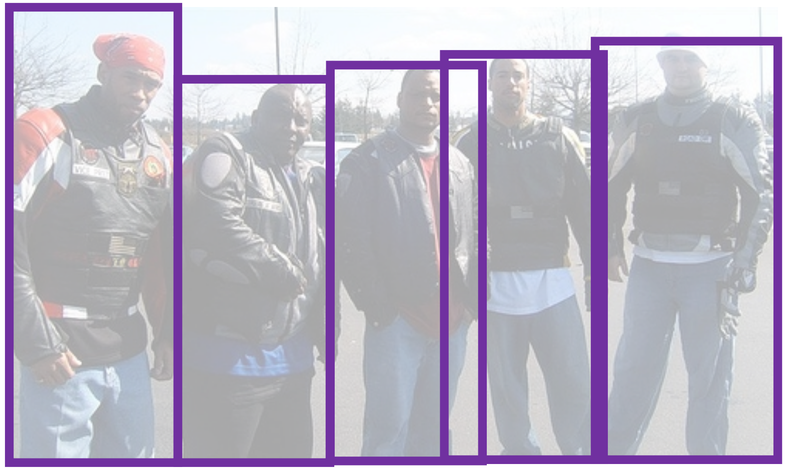
Mask R-CNN

ICCV 2017, Venice, Italy

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick

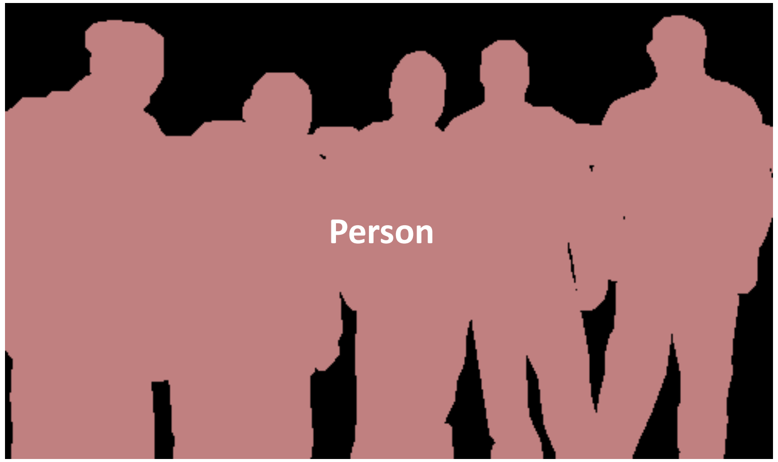
Facebook AI Research (FAIR)

Visual Perception



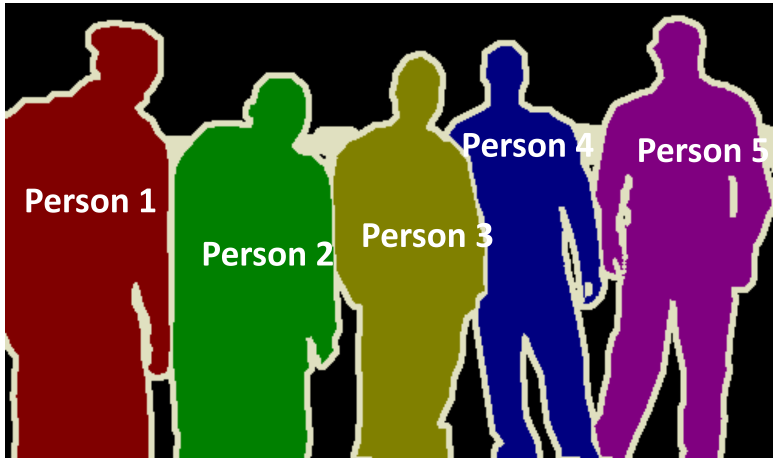
Object Detection

✓



Semantic Segmentation

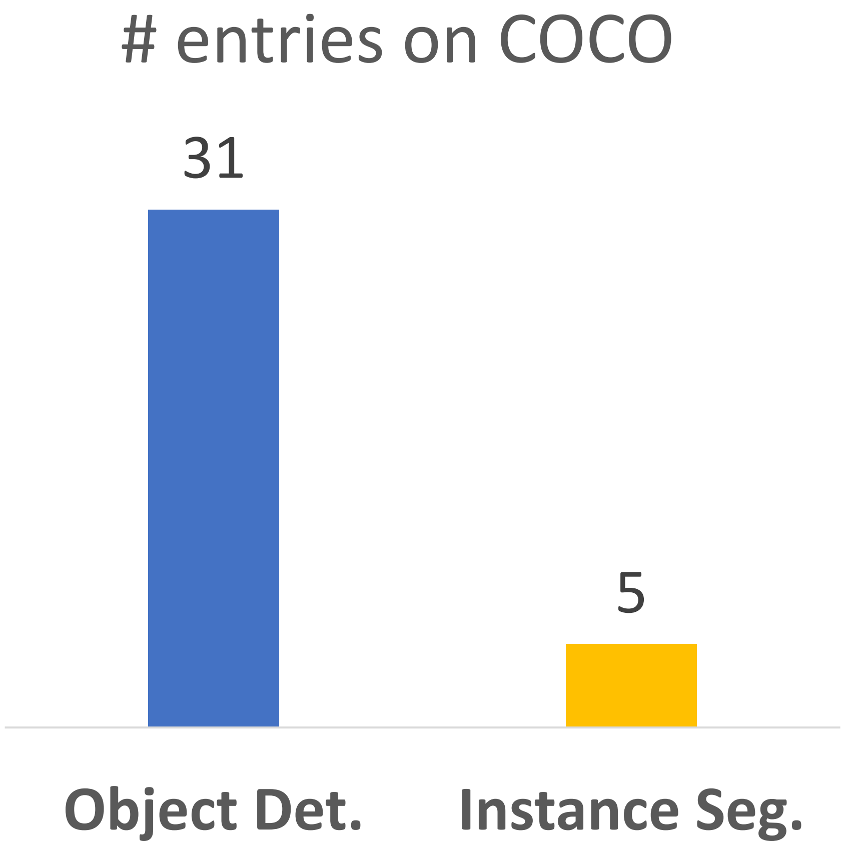
✓

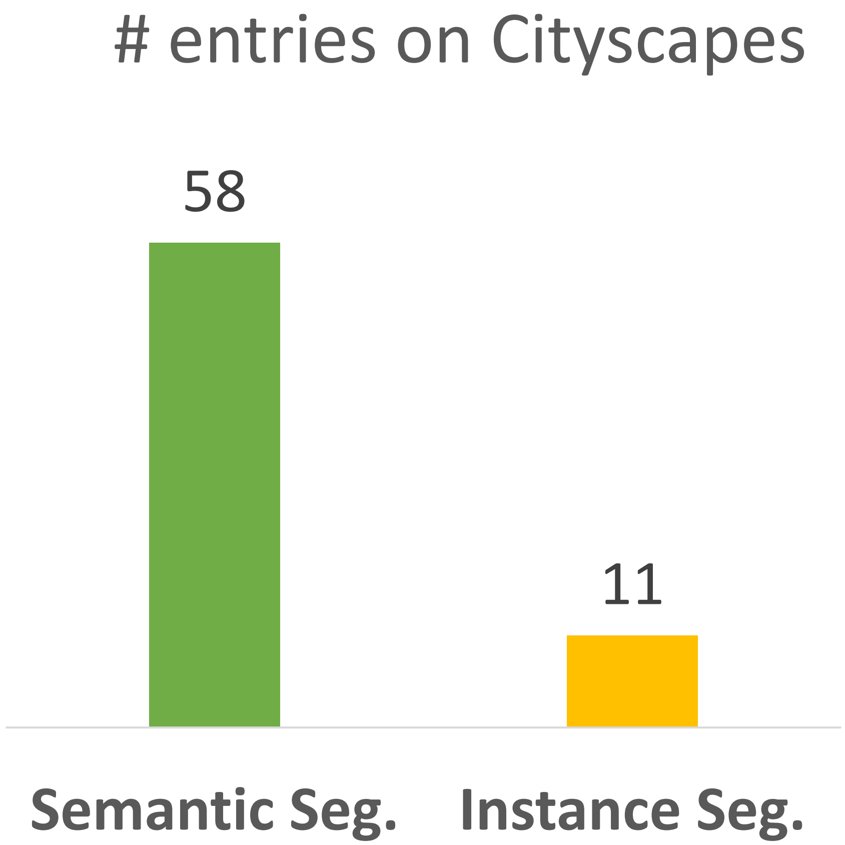


Instance Segmentation

?

A Challenging Problem...

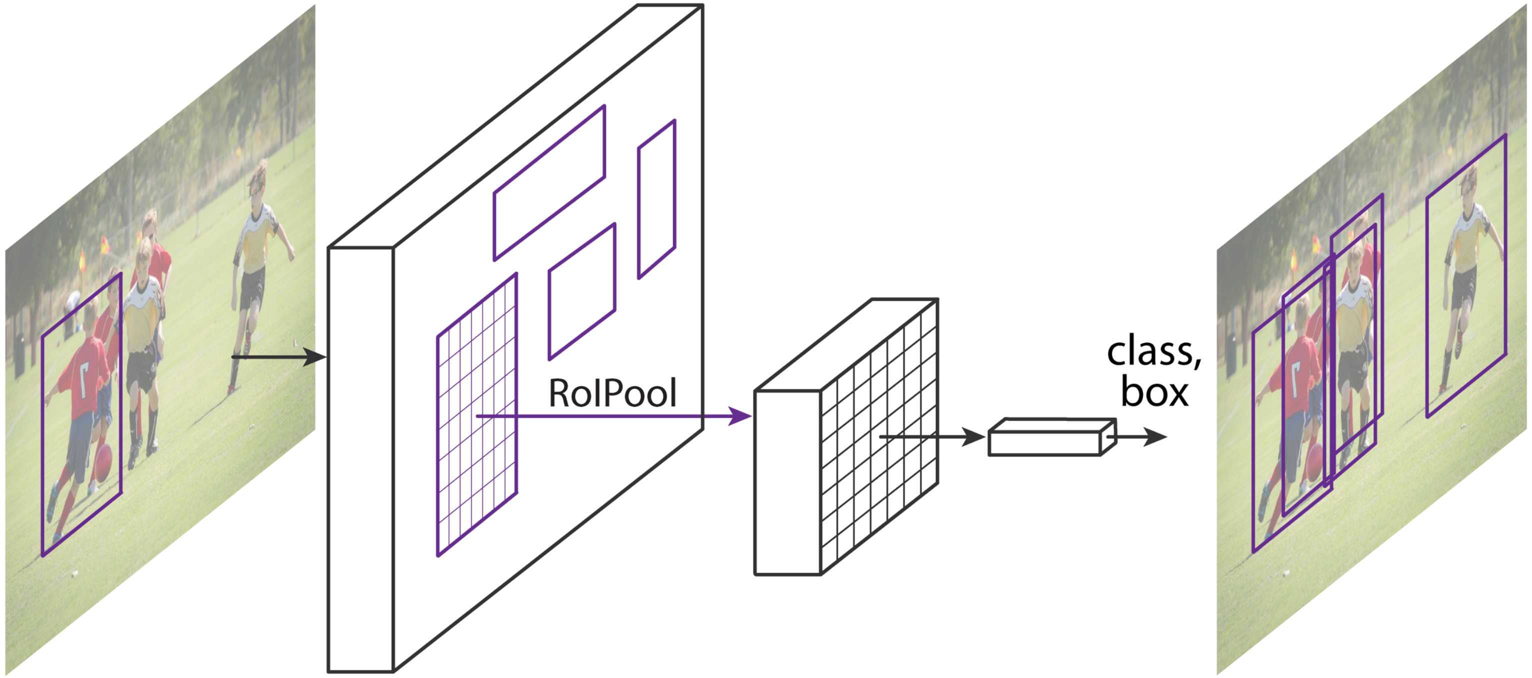




*on the leaderboards

Object Detection

- Fast/Faster R-CNN
  - ✓**Meta-algorithm**
  - ✓Good speed
  - ✓Good accuracy
  - ✓Intuitive
  - ✓Easy to use



Ross Girshick. “Fast R-CNN”. ICCV 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.

Semantic Segmentation

- Fully Convolutional Net (FCN)
  - ✓**Meta-algorithm**
  - ✓Good speed
  - ✓Good accuracy
  - ✓Intuitive
  - ✓Easy to use

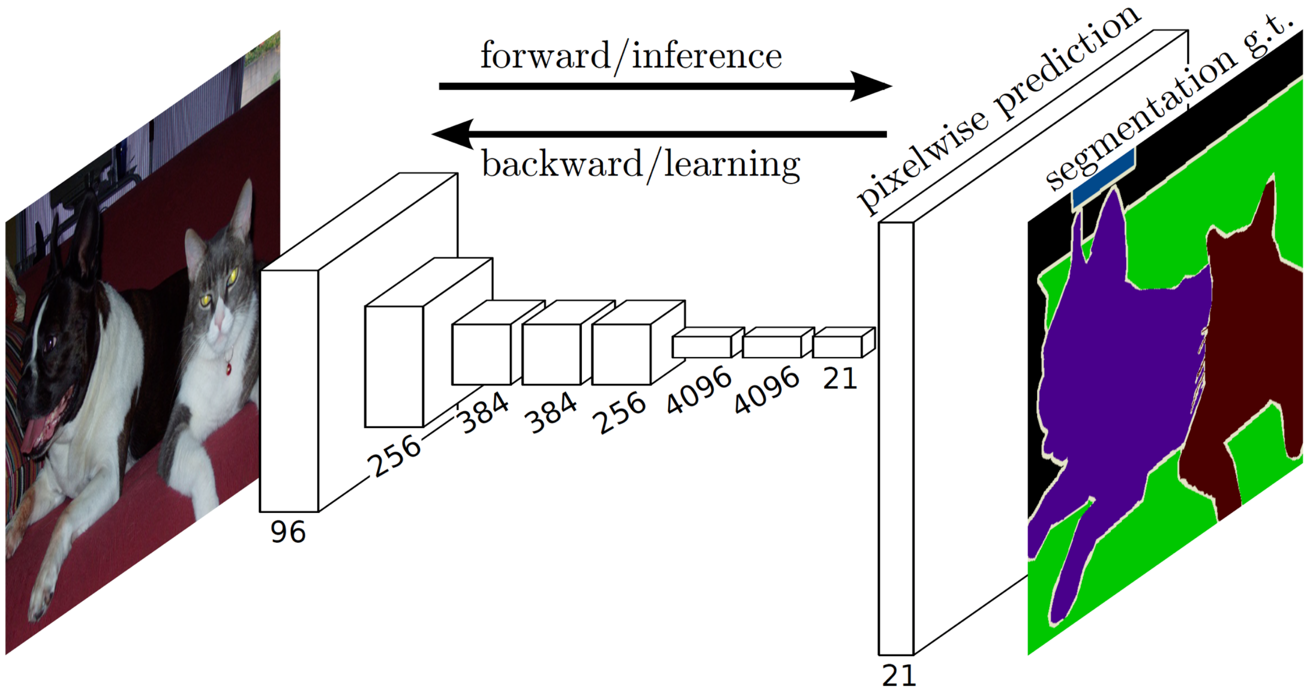
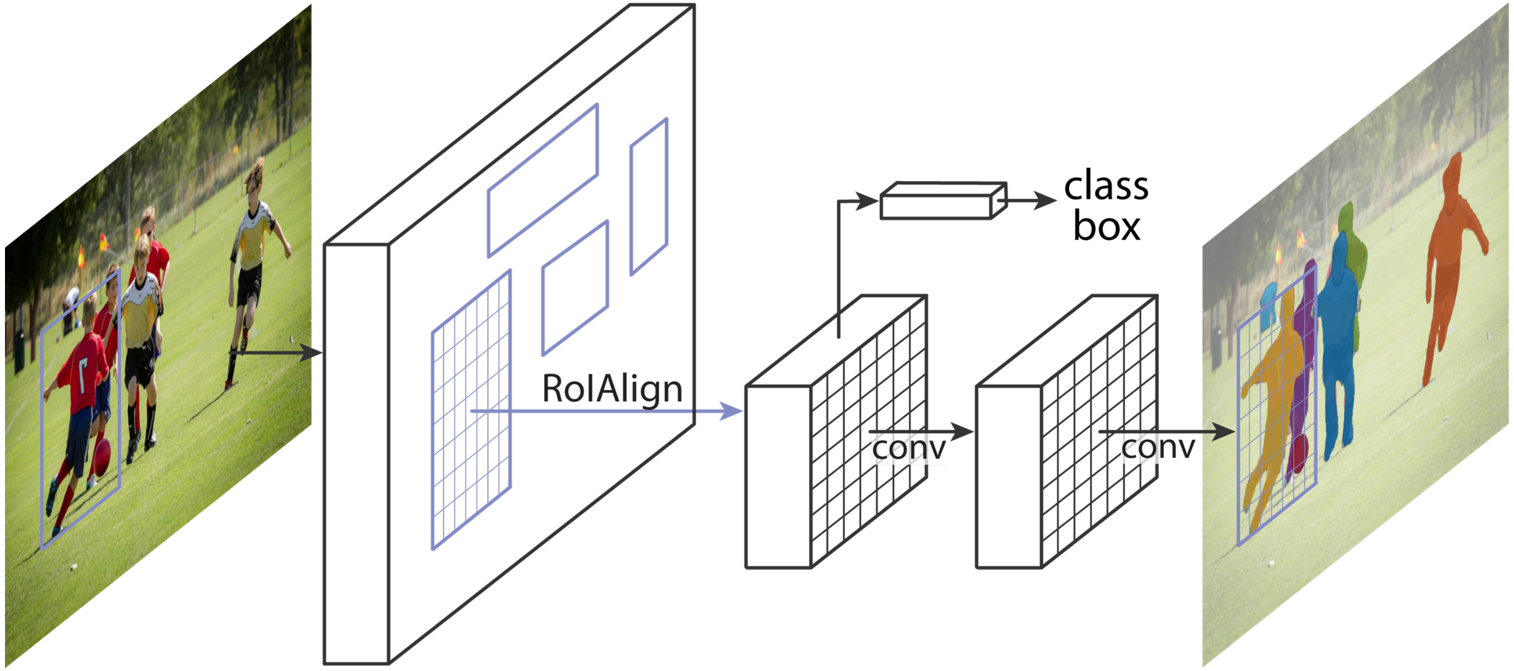


Figure credit: Long et al

Jonathan Long, Evan Shelhamer, & Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation”. CVPR 2015.

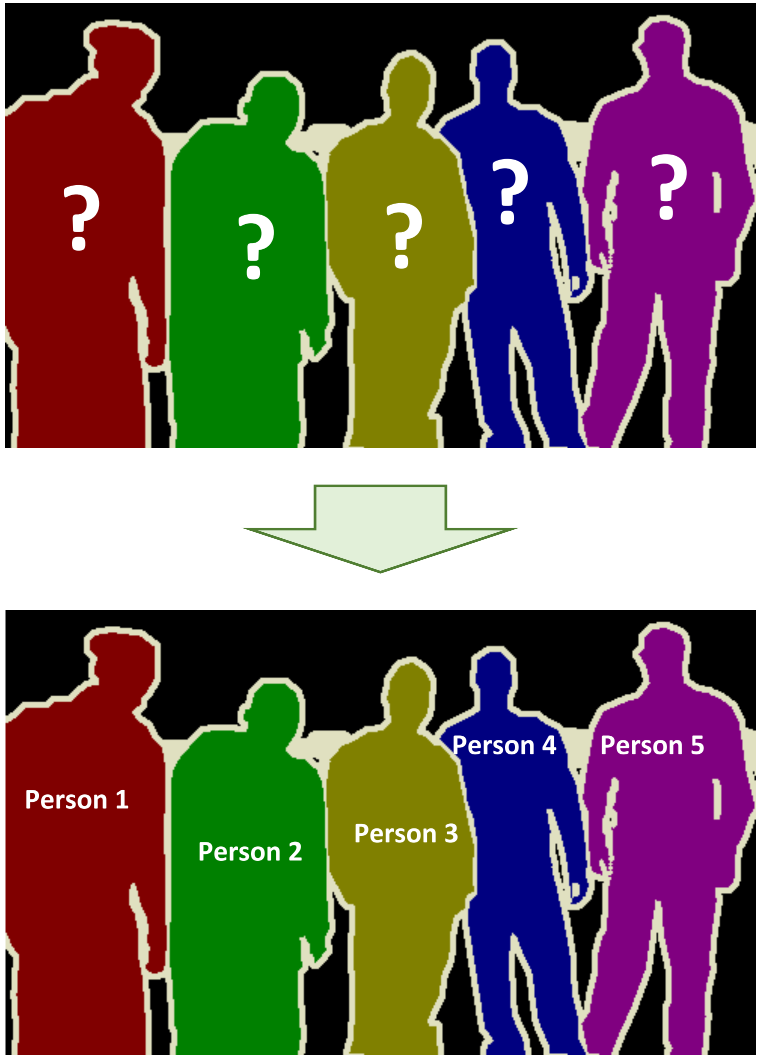
Instance Segmentation

- **Goals** of Mask R-CNN
  - **Meta-algorithm**
  - Good speed
  - Good accuracy
  - Intuitive
  - Easy to use

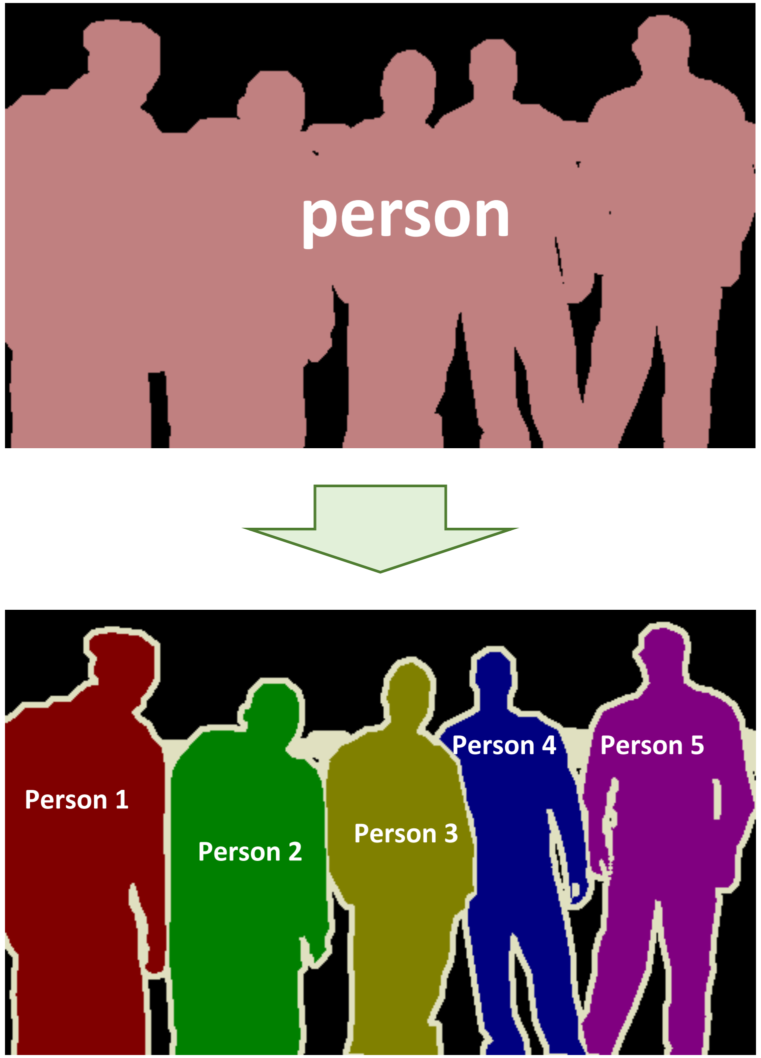


Instance Segmentation Methods

R-CNN driven



FCN driven



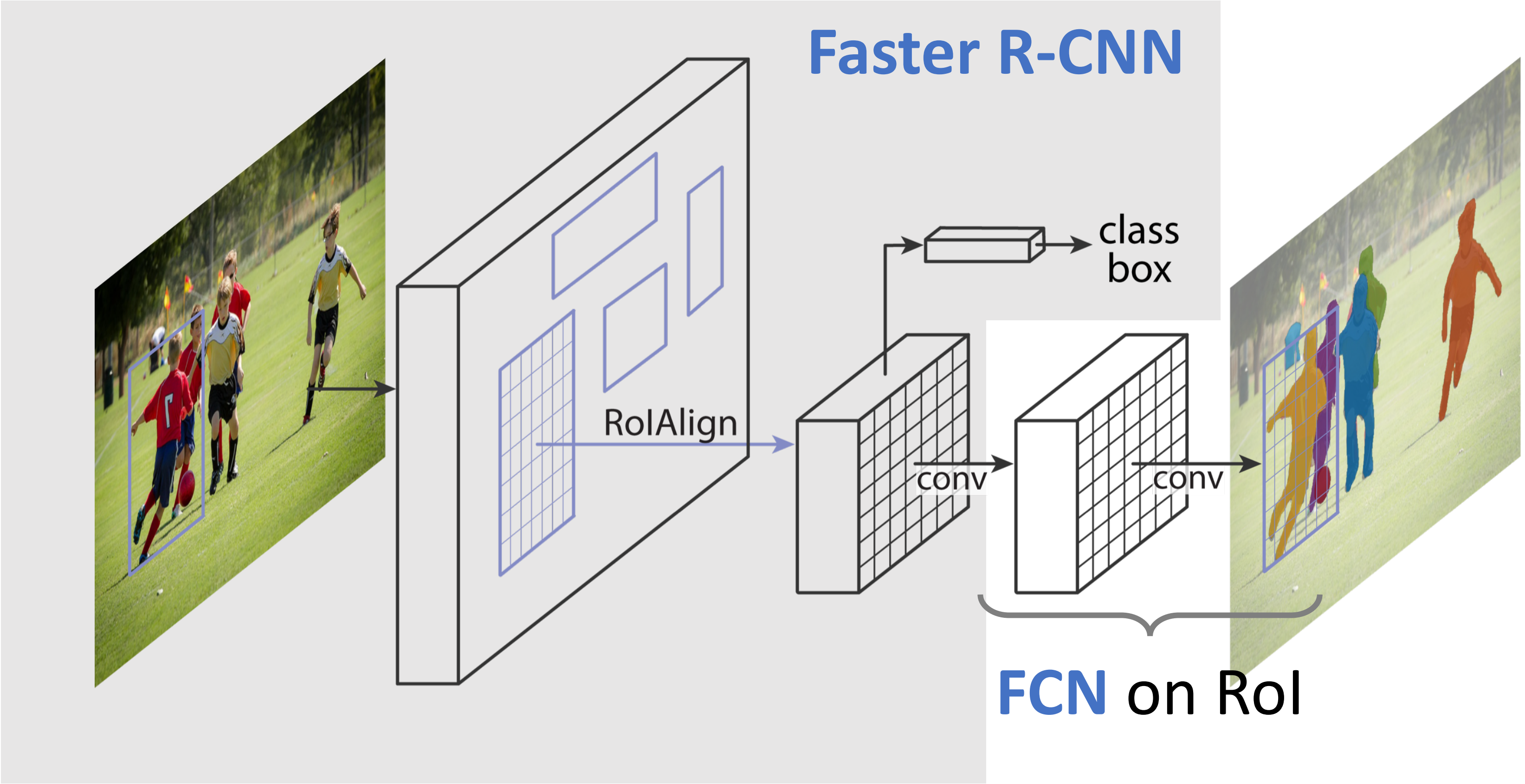
[Hariharan et al, ECCV'14], [Hariharan et al, CVPR'15],
[Dai et al, CVPR'15], [Dai et al, CVPR'16], …

[Li et al, CVPR'17],
[Arnab & Torr, CVPR'17], …

[Liang et al, arXiv'15], [Kirillov et al, CVPR'17],
[Bai & Urtasun, CVPR'17], …

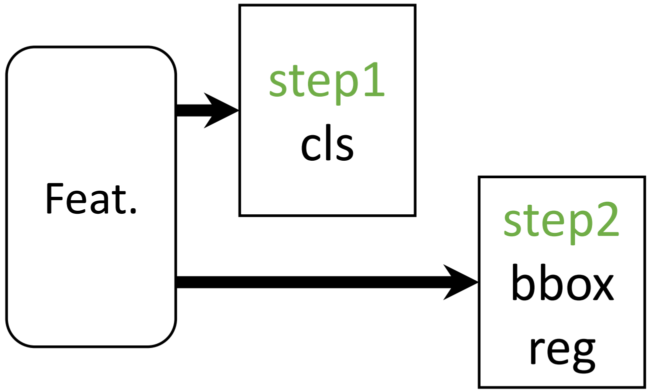
What is Mask R-CNN

- Mask R-CNN = **Faster R-CNN** with **FCN** on RoIs

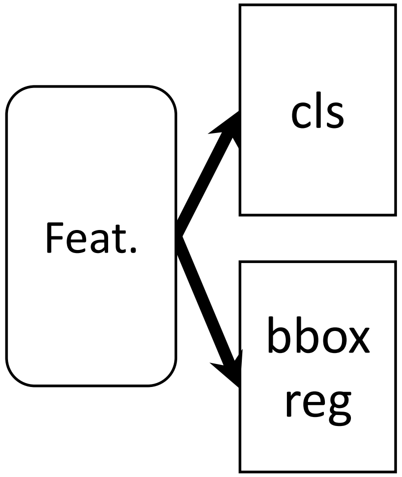


What is Mask R-CNN: Parallel Heads

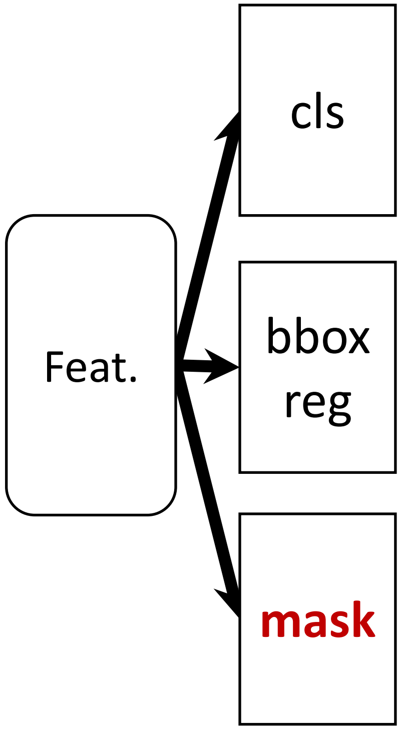
- Easy, fast to implement and use



(slow) R-CNN



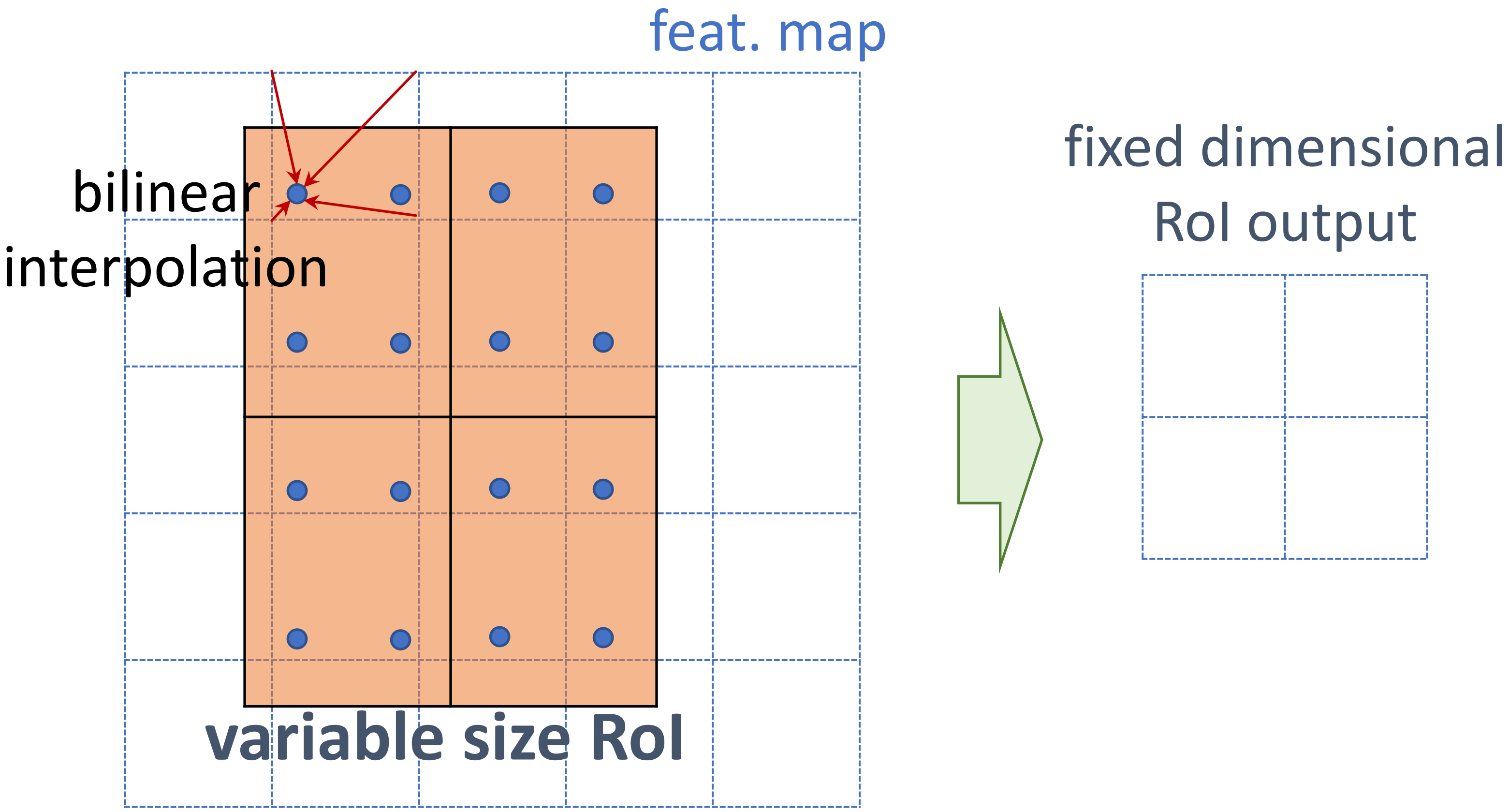
Fast/er R-CNN



Mask R-CNN

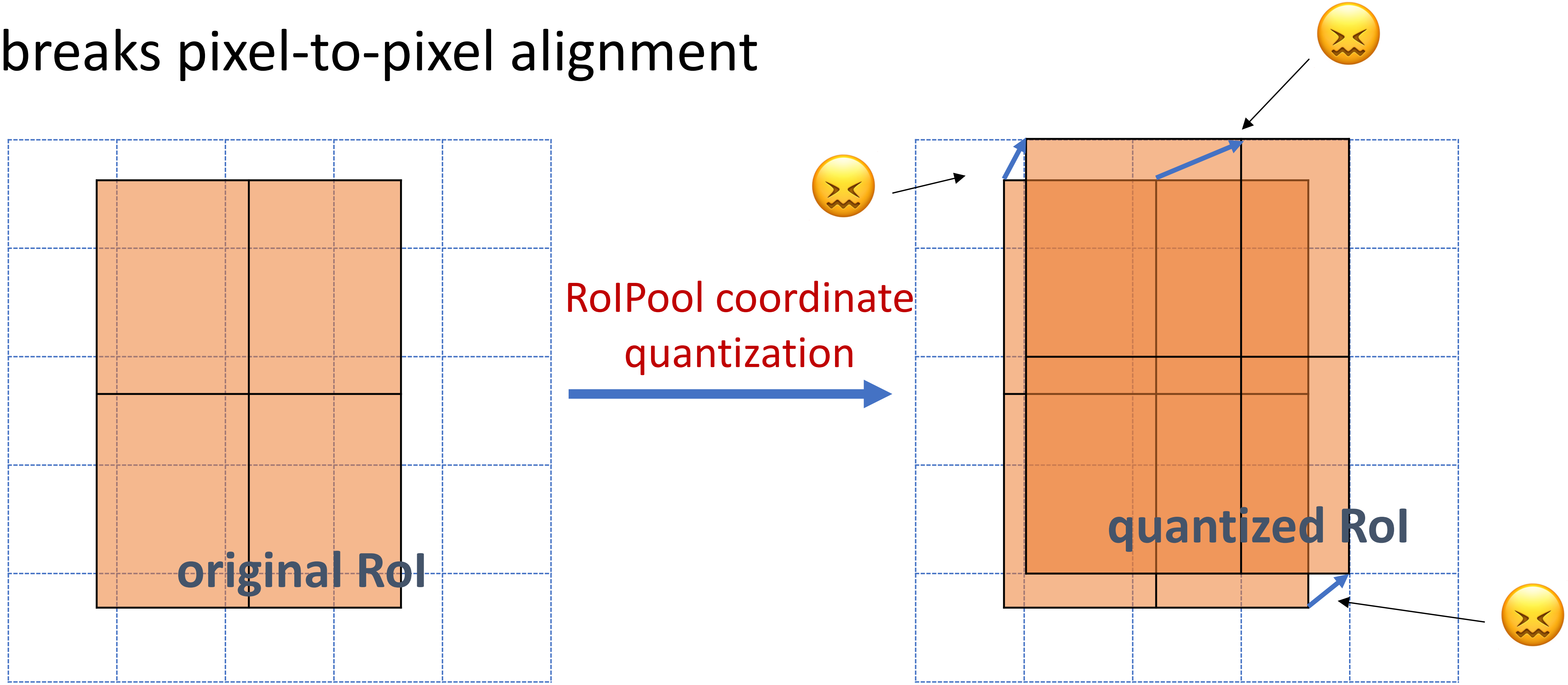
What is Mask R-CNN: RoIAlign

- No quantization



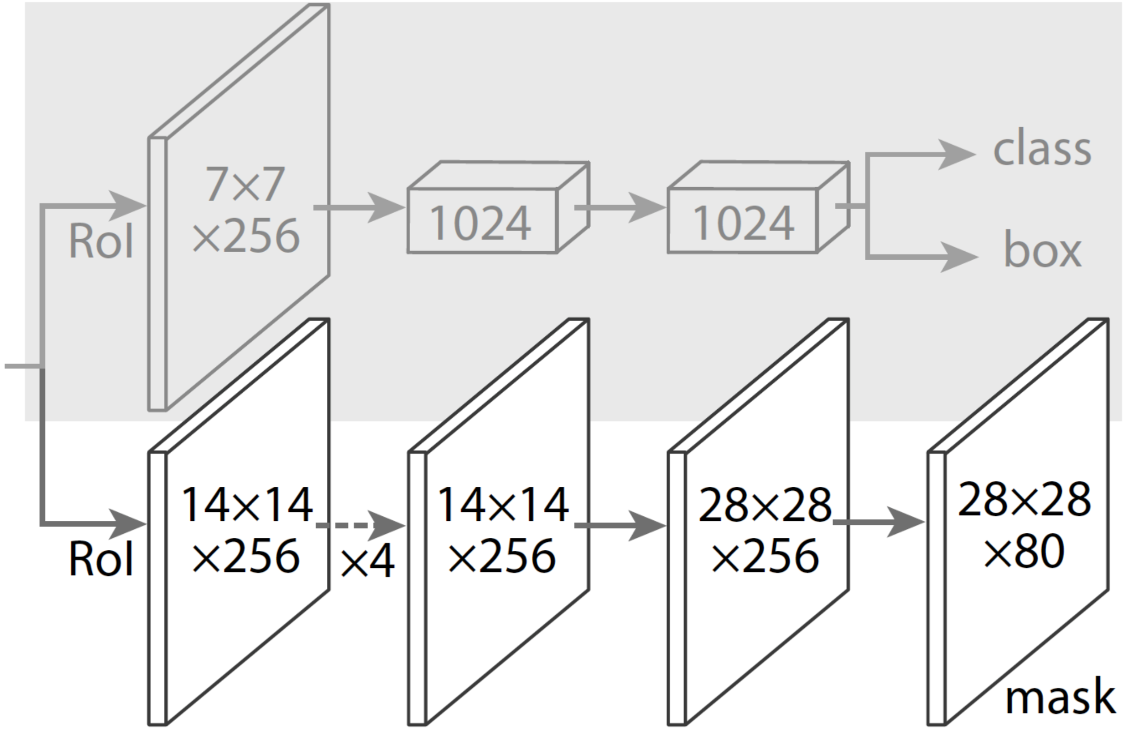
vs. RoIPool

- was not for segmentation
- breaks pixel-to-pixel alignment



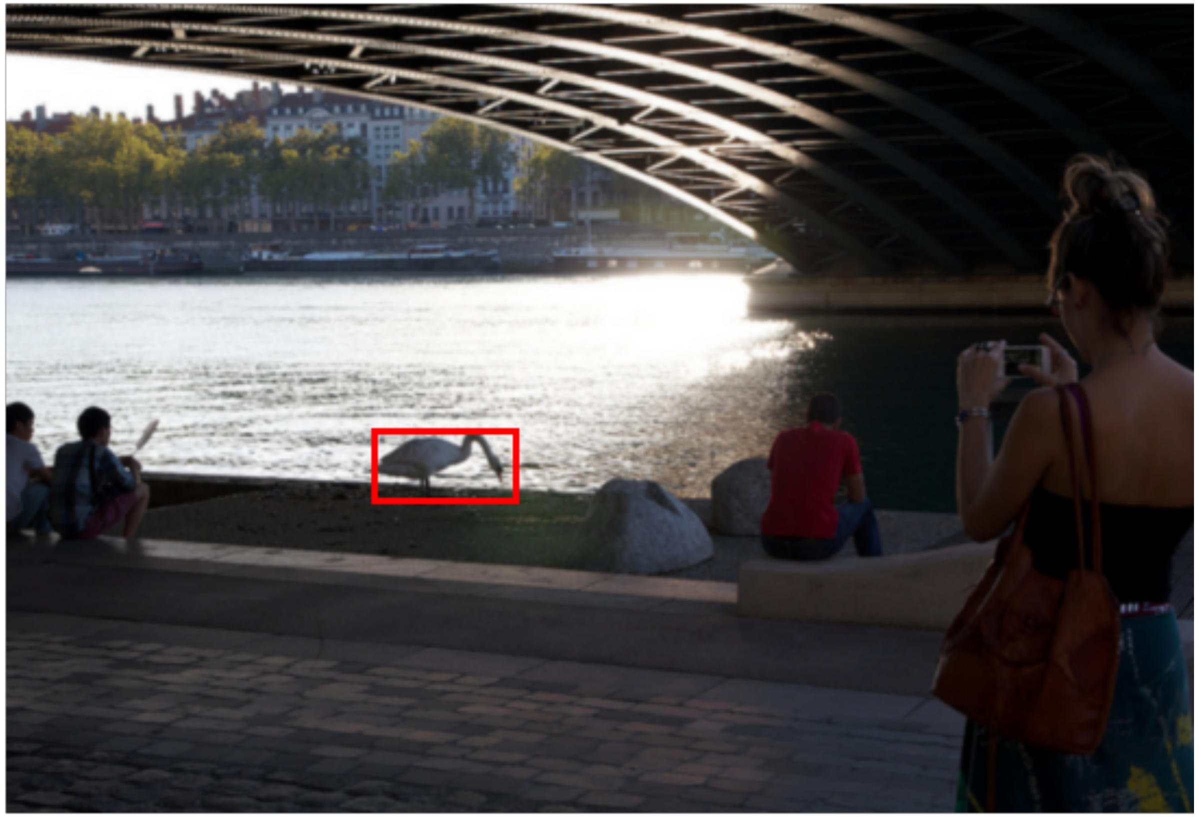
What is Mask R-CNN: FCN Mask Head

- Pixel-to-pixel aligned

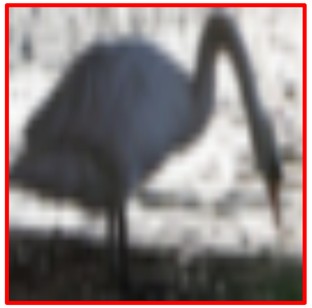


What is Mask R-CNN: FCN Mask Head

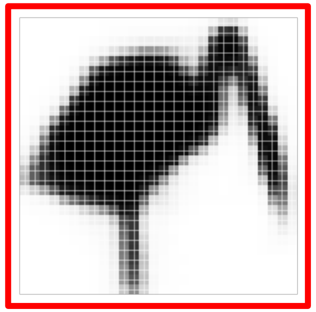
- Pixel-to-pixel aligned



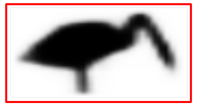
RoI



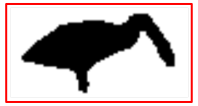
28x28 FCN prediction



resized soft prediction

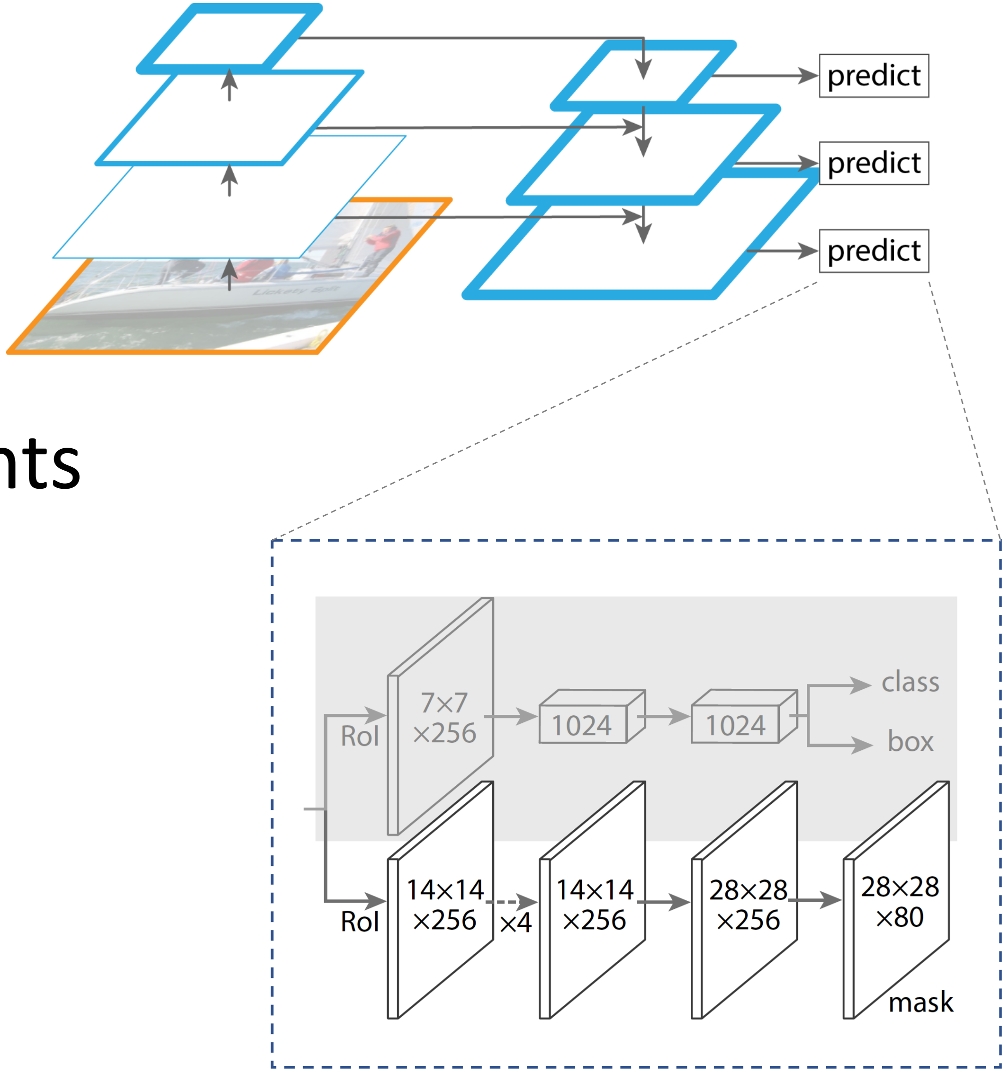


final mask



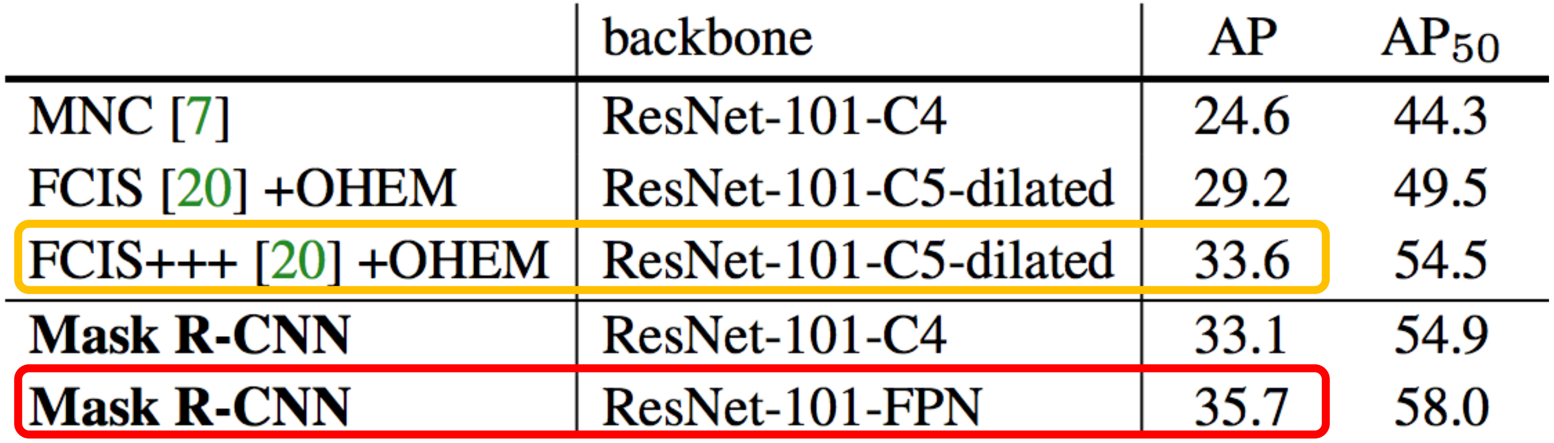
Implementation

- Mask R-CNN is a **meta-algorithm**
- Compatible with other improvements
- We used:
  - ResNet/ResNeXt [Xie et al, CVPR'17]
  - Feature Pyramid Net [Lin et al, CVPR'17]



Results

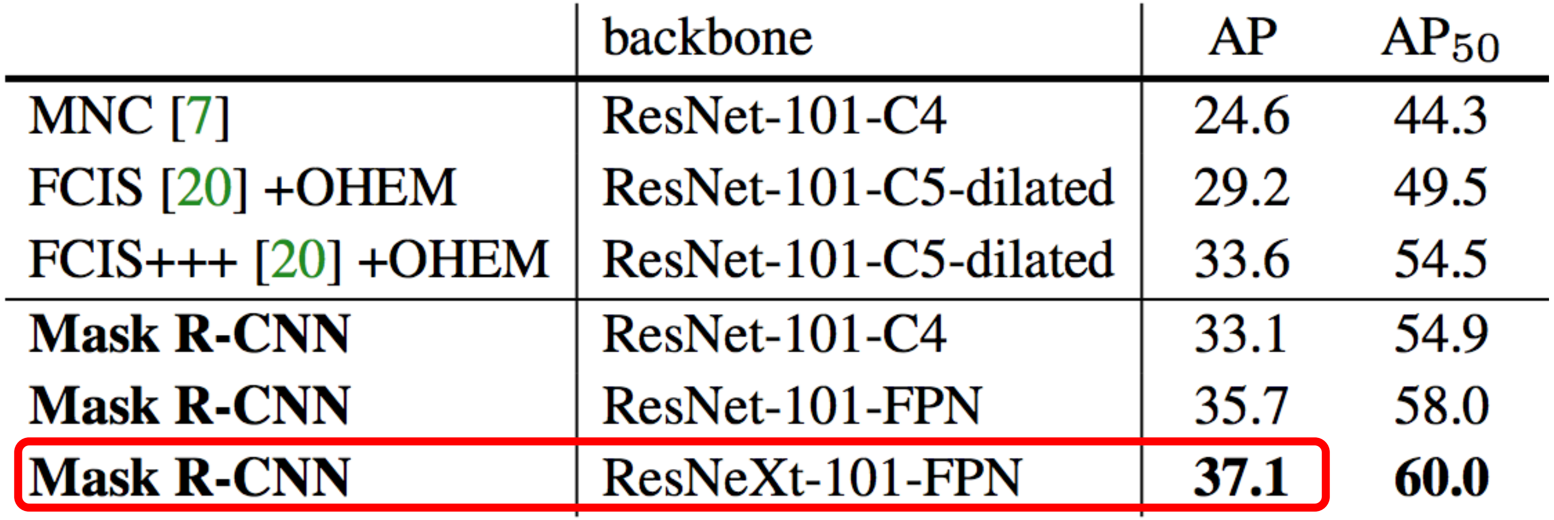
Instance Segmentation Results on COCO



	backbone	AP	AP_{50}
MNC [7]	ResNet-101-C4	24.6	44.3
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5
FCIS+++ [20] +OHEM	ResNet-101-C5-dilated	33.6	54.5
Mask R-CNN	ResNet-101-C4	33.1	54.9
Mask R-CNN	ResNet-101-FPN	35.7	58.0

- without bells and whistles, **2 AP better** than 2016 winner
- **200ms / img**

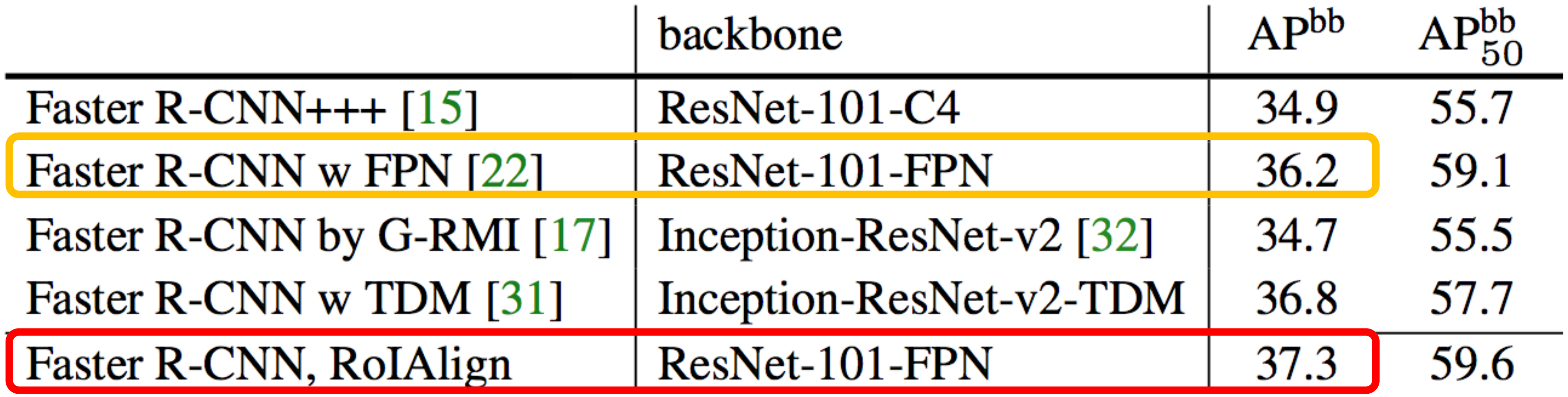
Instance Segmentation Results on COCO



	backbone	AP	AP_{50}
MNC [7]	ResNet-101-C4	24.6	44.3
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5
FCIS+++ [20] +OHEM	ResNet-101-C5-dilated	33.6	54.5
Mask R-CNN	ResNet-101-C4	33.1	54.9
Mask R-CNN	ResNet-101-FPN	35.7	58.0
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**

- Better features: ResNeXt [Xie et al, CVPR’17]

Object Detection Results on COCO

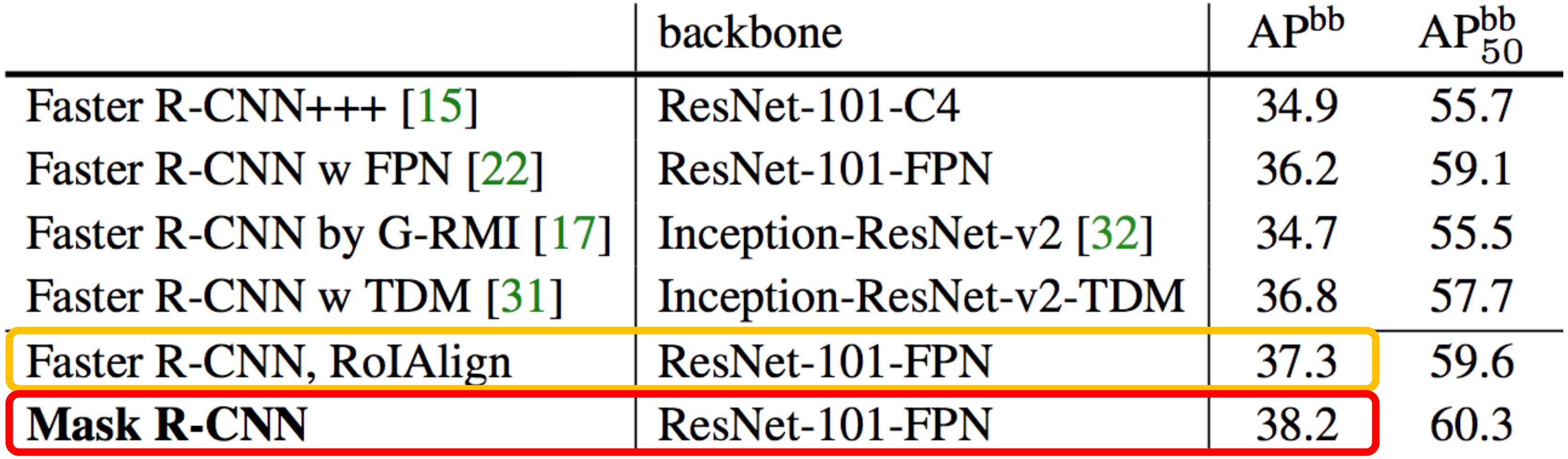


	backbone	$\mathrm{AP}^{\mathrm{bb}}$	$\mathrm{AP}^{\mathrm{bb}}_{50}$
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6

bbox improved by:

- RoIAlign

Object Detection Results on COCO

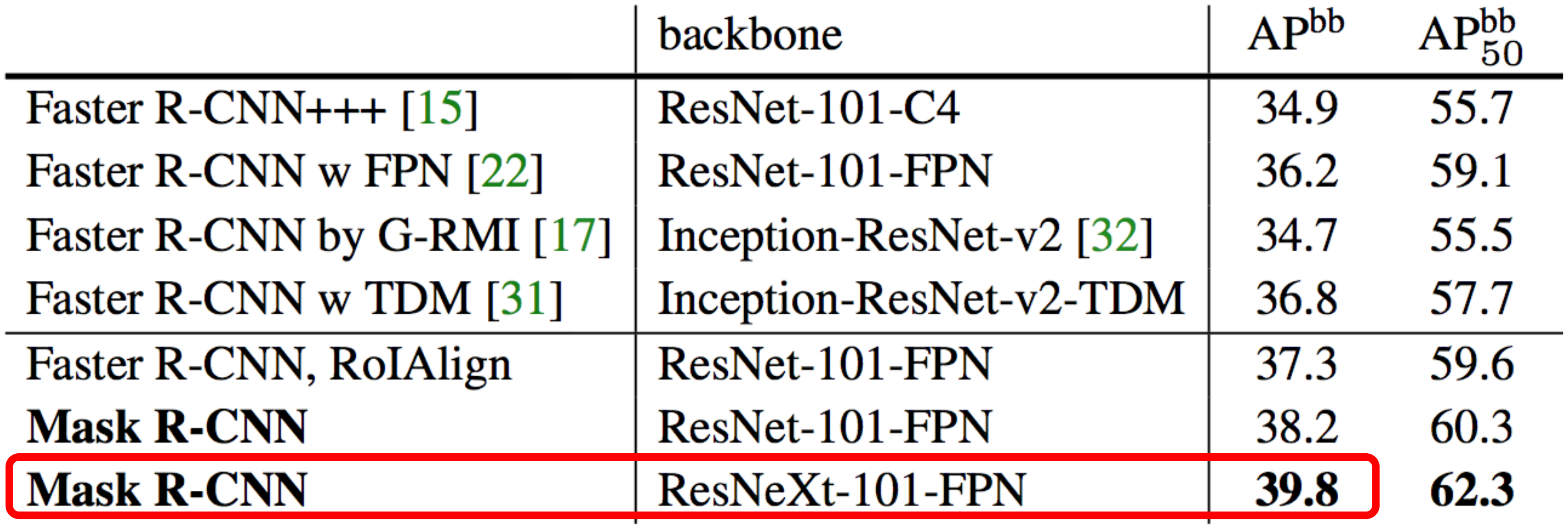


	backbone	AP^{bb}	$\text{AP}^{\text{bb}}_{50}$
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6
Mask R-CNN	ResNet-101-FPN	38.2	60.3

bbox improved by:

- RoIAlign
- Multi-task training w/ mask

Object Detection Results on COCO



	backbone	AP^{bb}	$\text{AP}^{\text{bb}}_{50}$
Faster R-CNN+++ [15]	ResNet-101-C4	34.9	55.7
Faster R-CNN w FPN [22]	ResNet-101-FPN	36.2	59.1
Faster R-CNN by G-RMI [17]	Inception-ResNet-v2 [32]	34.7	55.5
Faster R-CNN w TDM [31]	Inception-ResNet-v2-TDM	36.8	57.7
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6
Mask R-CNN	ResNet-101-FPN	38.2	60.3
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**

bbox improved by:

- RoIAlign
- Multi-task training w/ mask

COCO Competition 2017

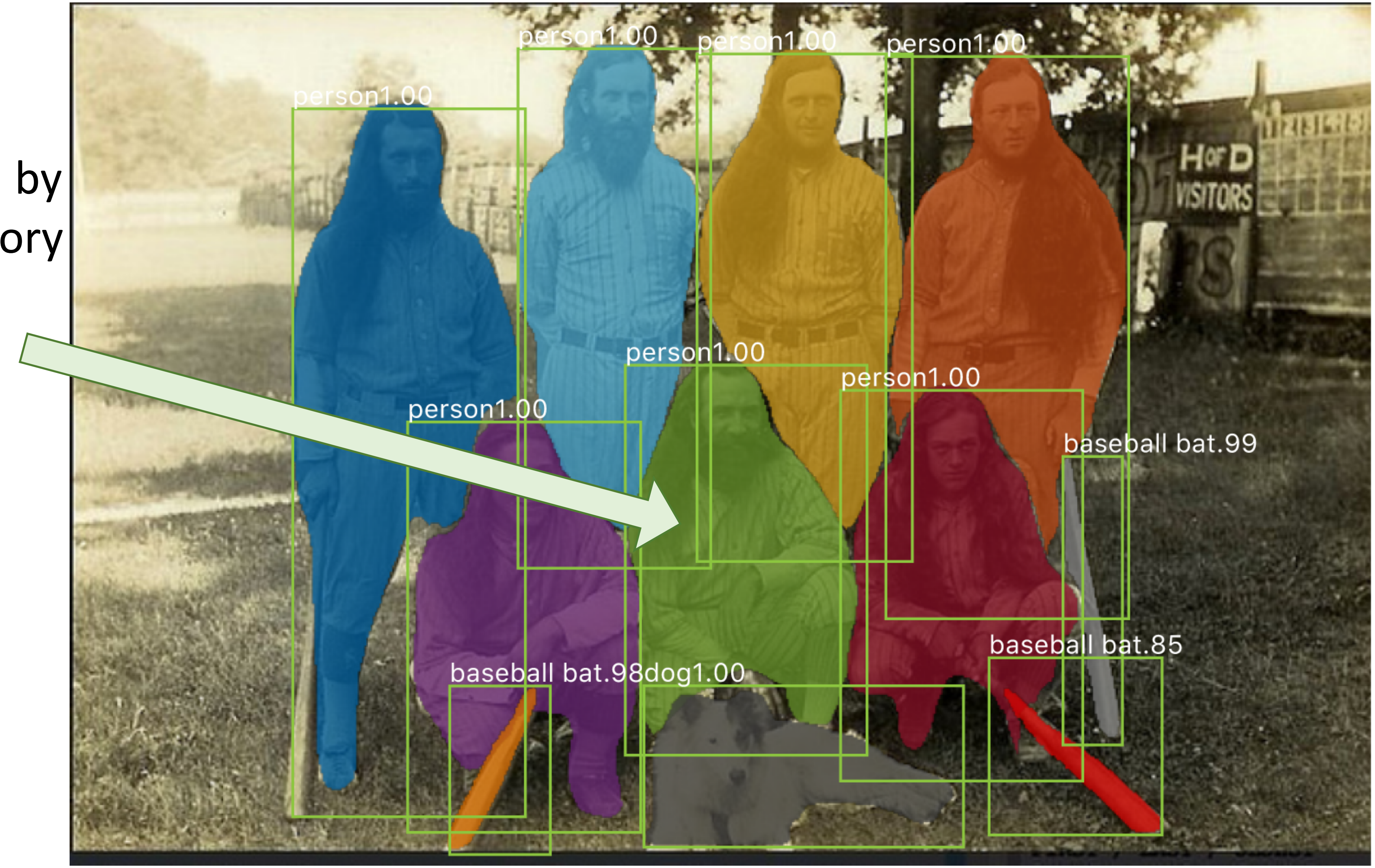
- Mask R-CNN is used by leading teams

- Our Mask R-CNN achieves a *single-model* result of
    - 47.9 bbox AP, 42.6 mask AP

- More in our talk in COCO workshop (10/29, Sunday)

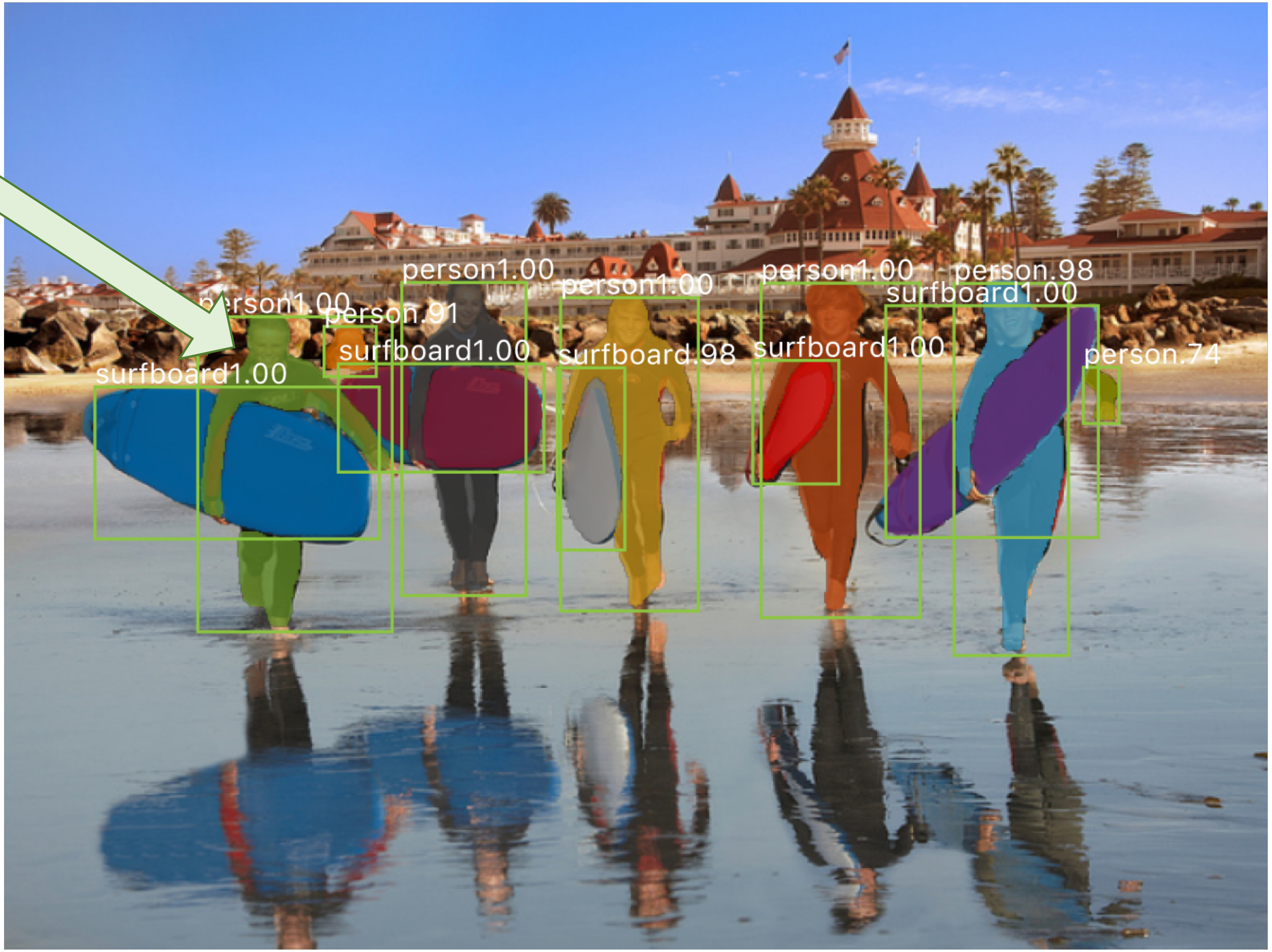
Examples

surrounded by same-category objects



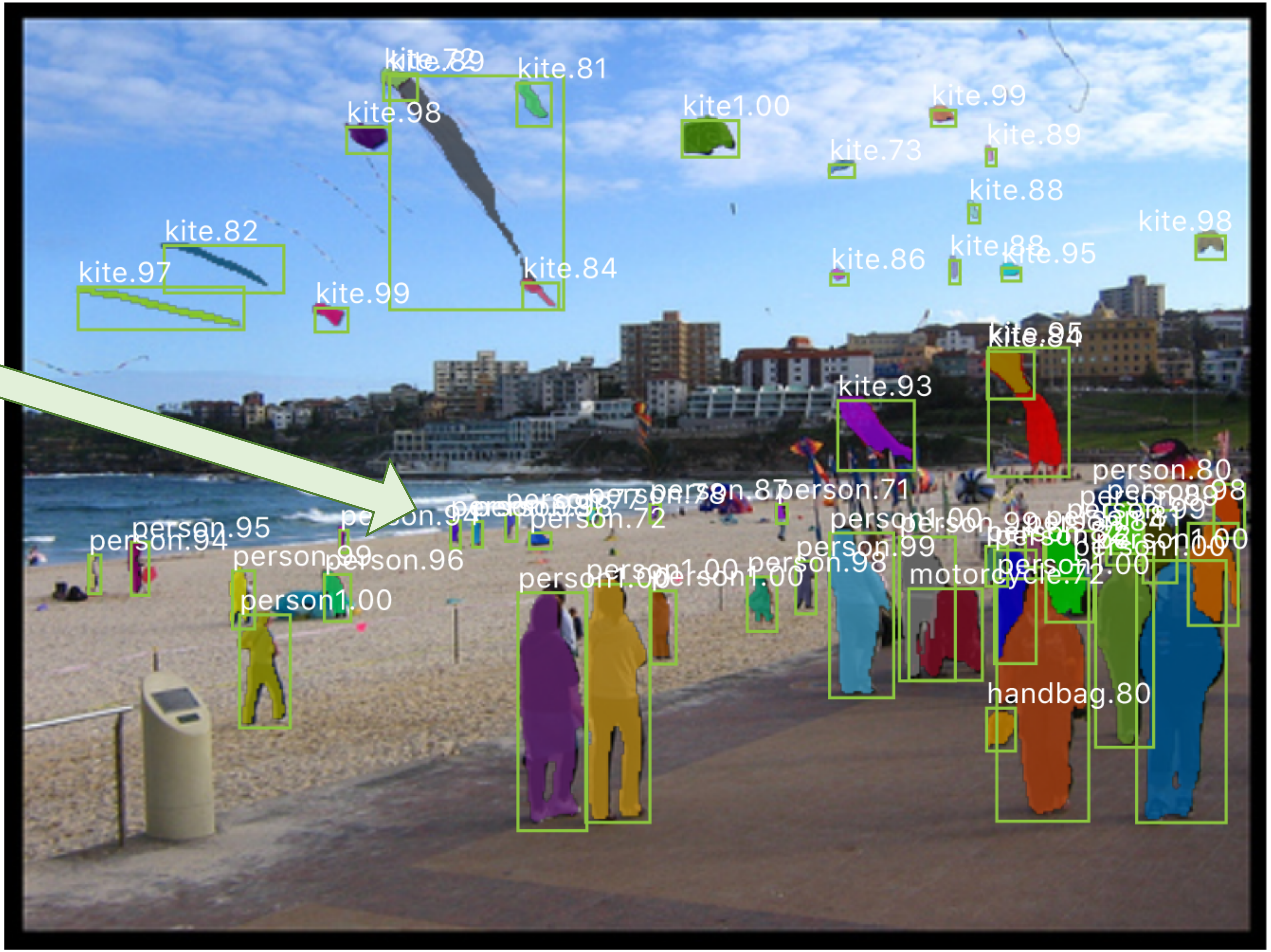
Mask R-CNN results on COCO

disconnected objects



Mask R-CNN results on COCO

small objects

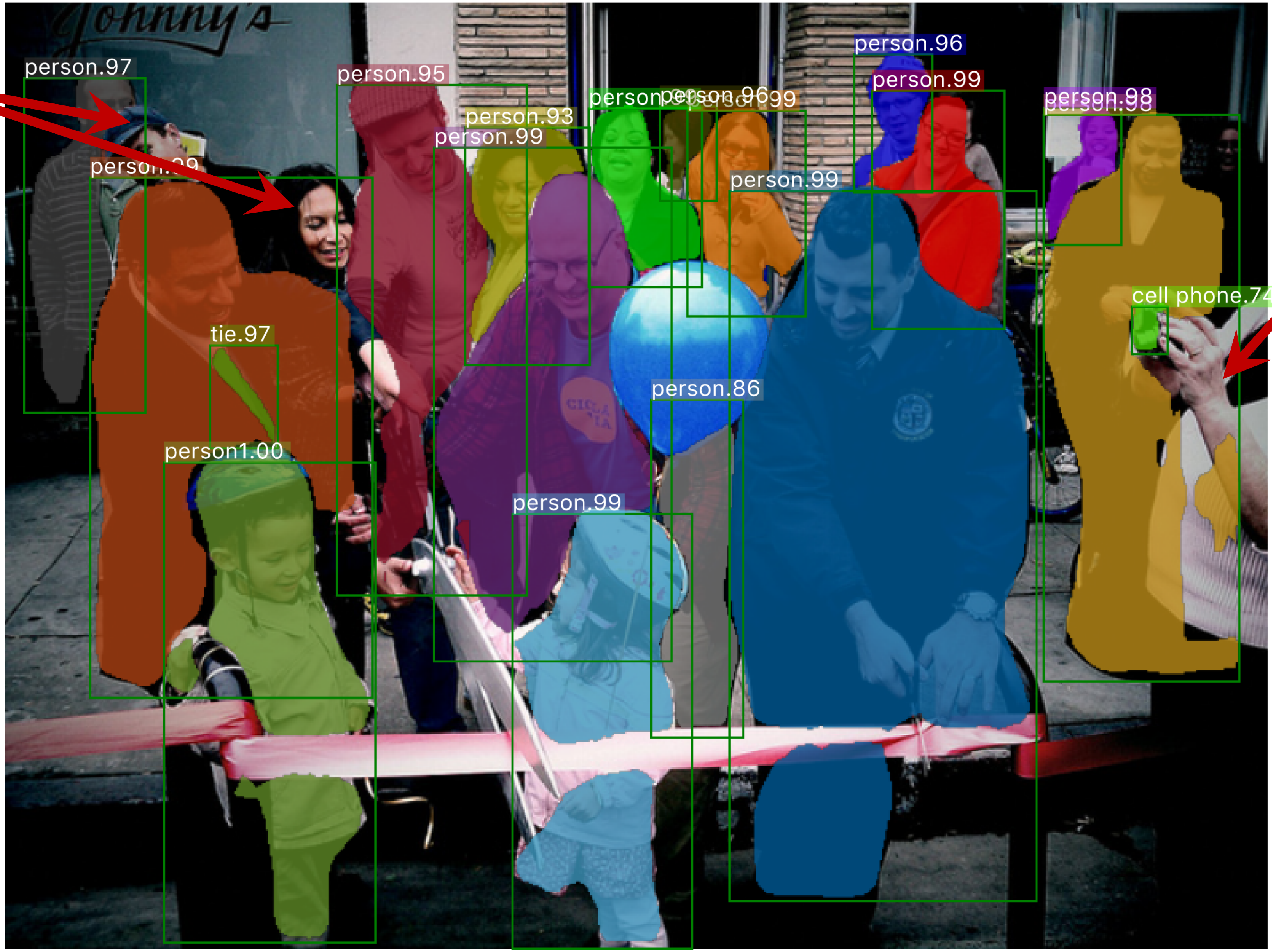


Mask R-CNN results on COCO

Failure: detection/segmentation

missing

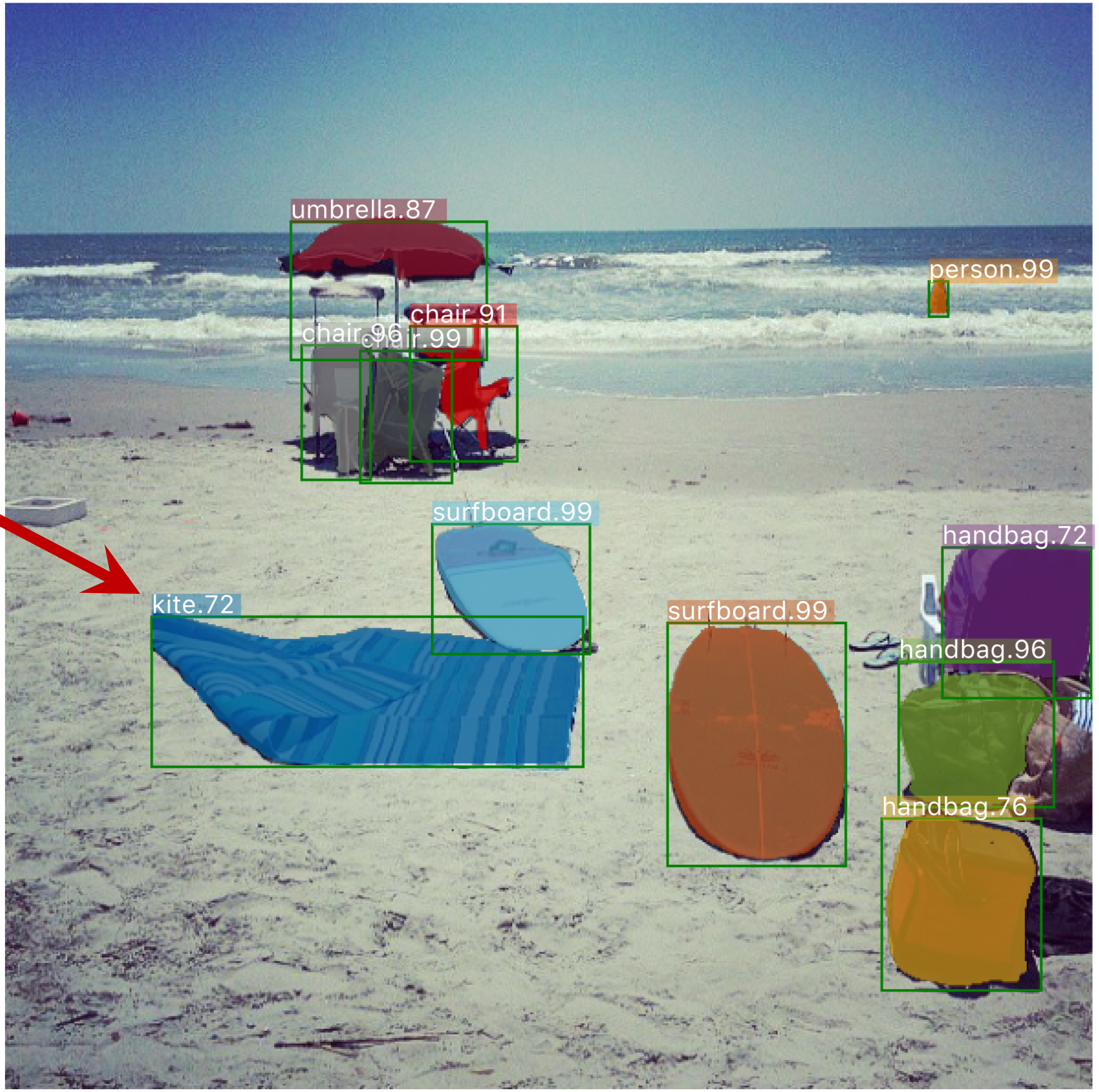
missing, false mask



Mask R-CNN results on COCO

Failure: recognition

not a kite

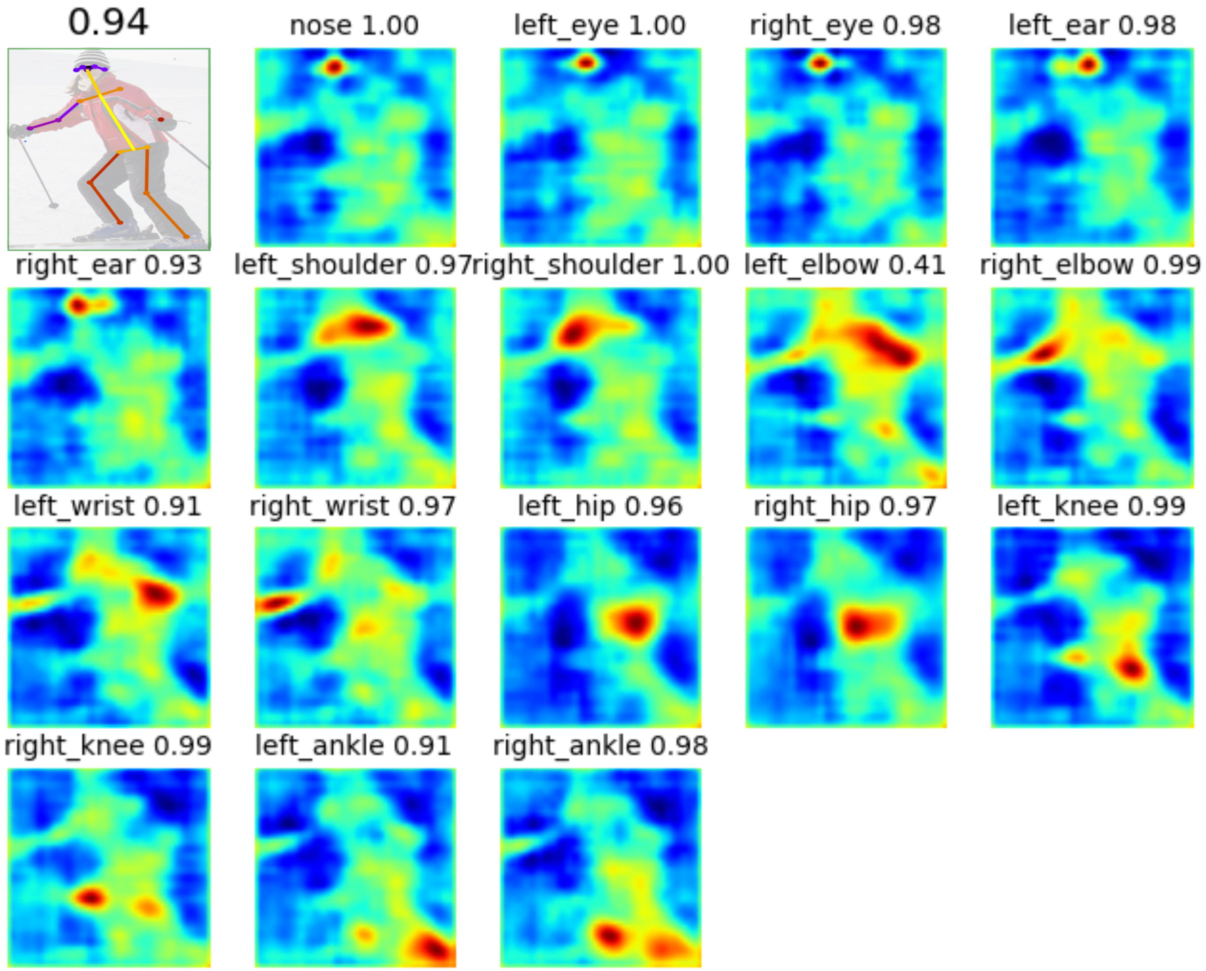


Mask R-CNN results on COCO

For Human Keypoint Detection

- keypoint = 1-hot mask
- human pose = 17 masks

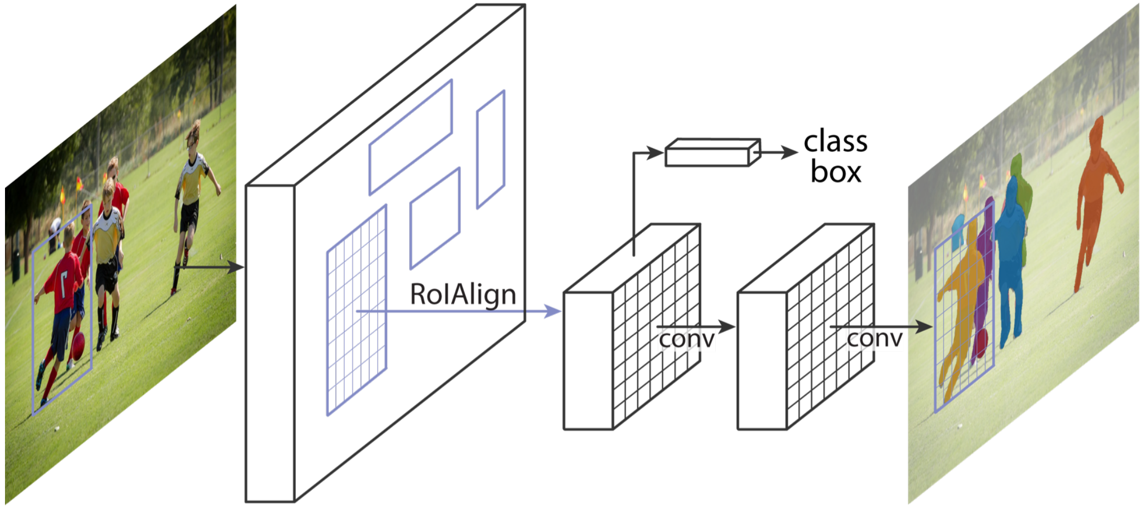
- One framework for
  - ✓bbox
  - ✓mask
  - ✓keypoint



Conclusion

Code will be open-sourced as
Facebook AI Research's **Detectron** platform

- Mask R-CNN
  - ✓**Meta-algorithm**
  - ✓Good speed
  - ✓Good accuracy
  - ✓Intuitive
  - ✓Easy to use


class
box
RoIAlign
conv
conv