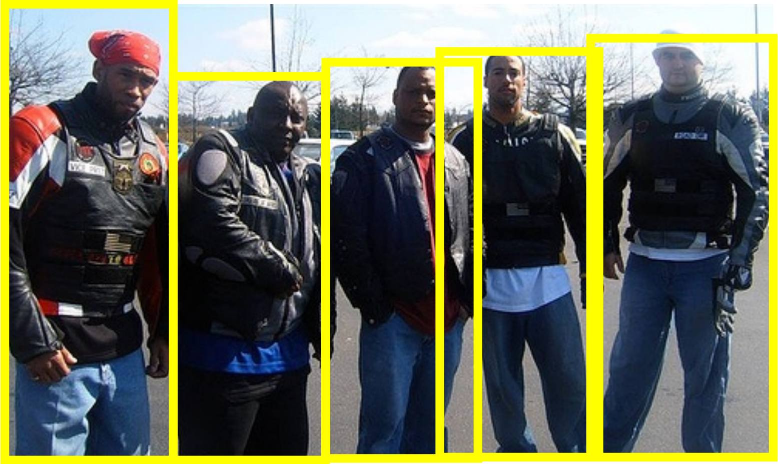
Mask R-CNN:
A Perspective on Equivariance

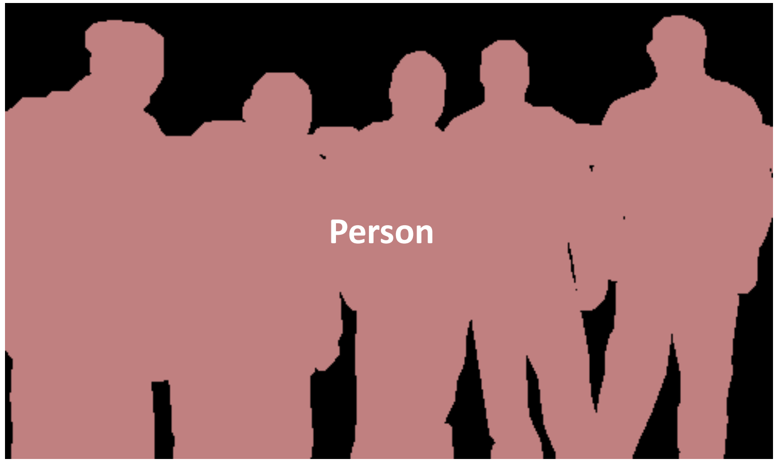
ICCV 2017 Tutorial, Venice, Italy

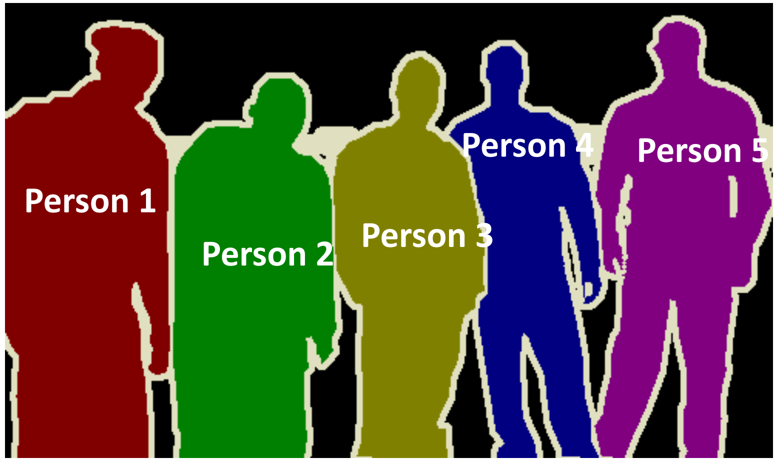
Kaiming He
in collaboration with: Georgia Gkioxari, Piotr Dollár, and Ross Girshick
Facebook AI Research (FAIR)

Introduction

Visual Perception Problems





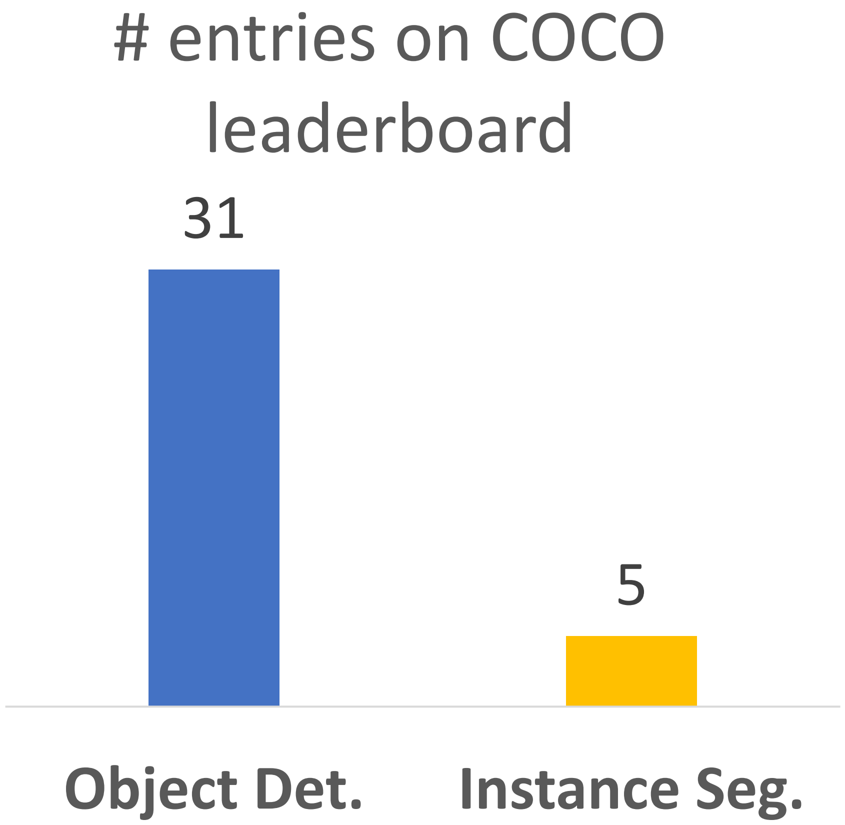


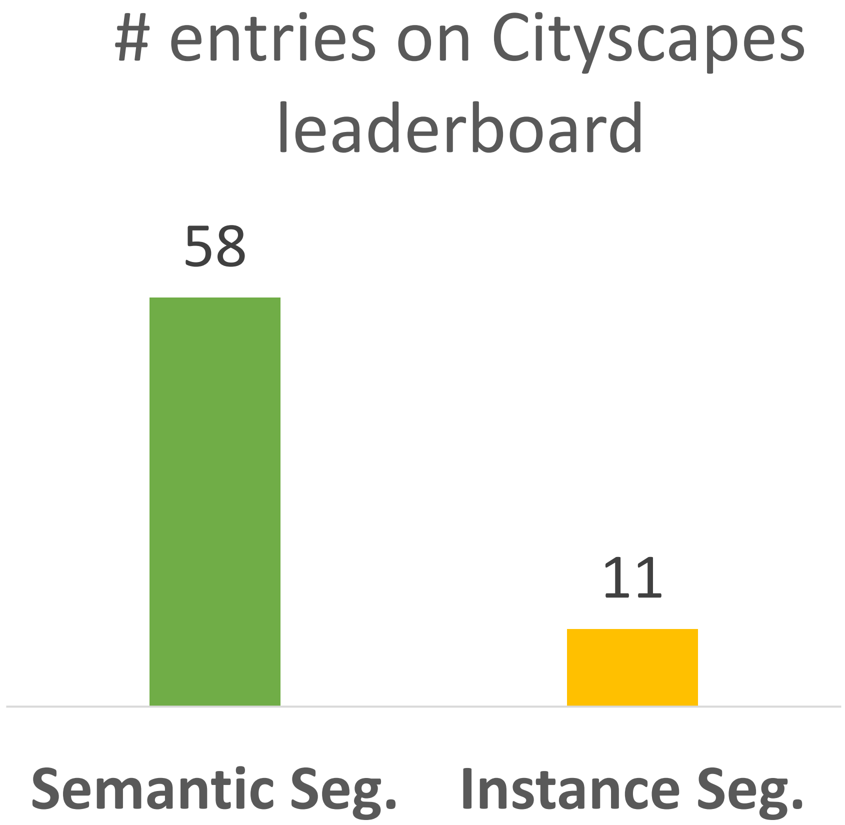
Object Detection ✓

Semantic Segmentation ✓

Instance Segmentation ?

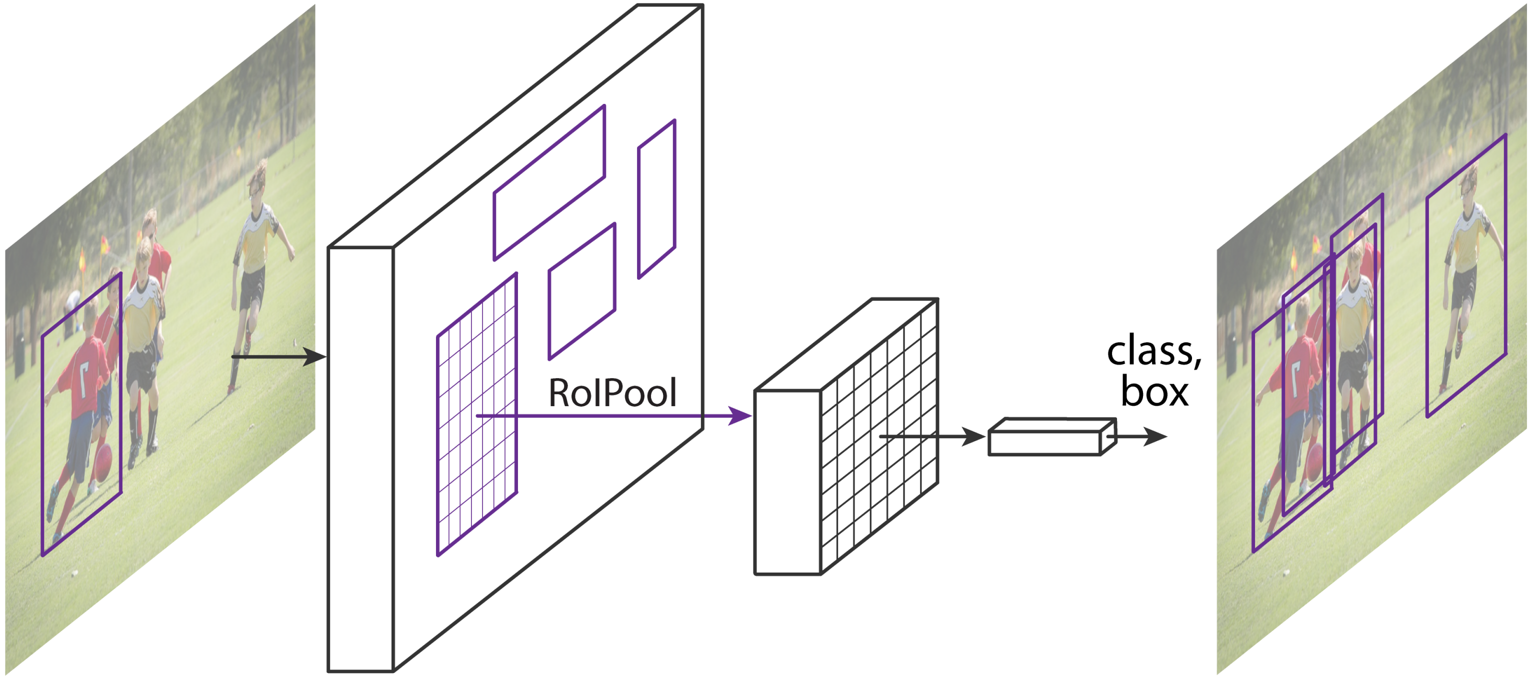
A Challenging Problem...





Object Detection

- Fast/Faster R-CNN
  - ✓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)
  - ✓Good speed
  - ✓Good accuracy
  - ✓Intuitive
  - ✓Easy to use

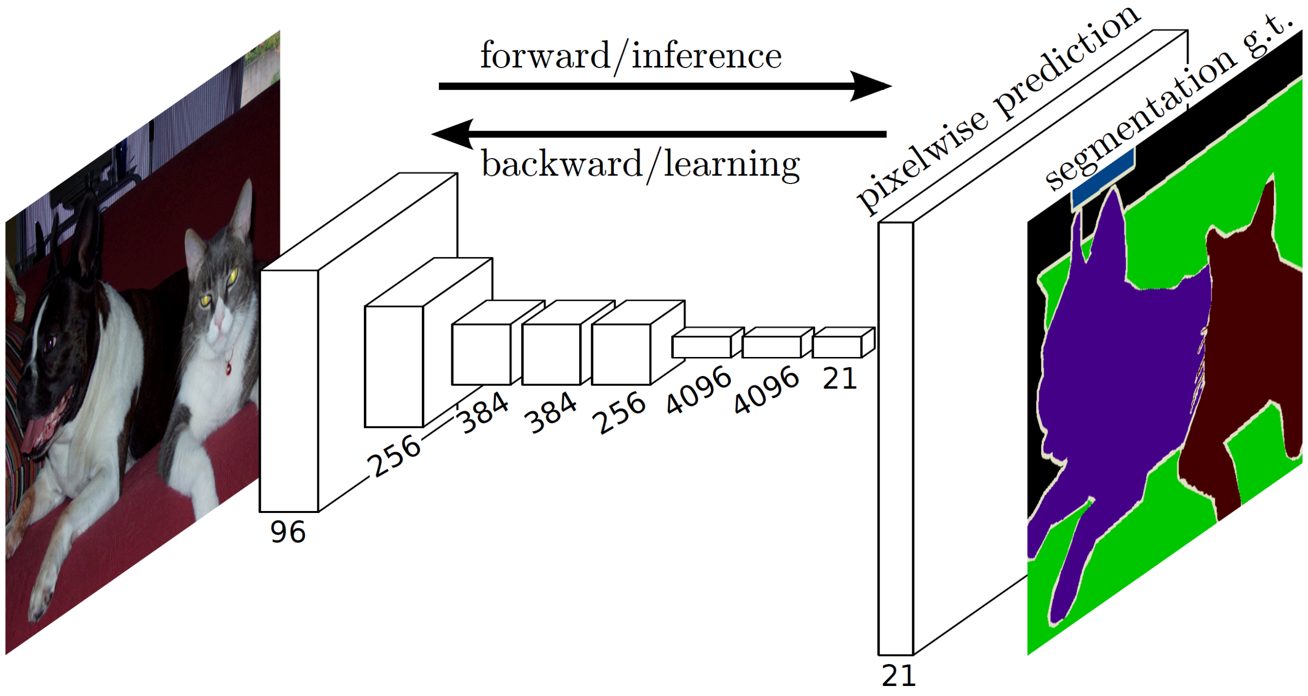
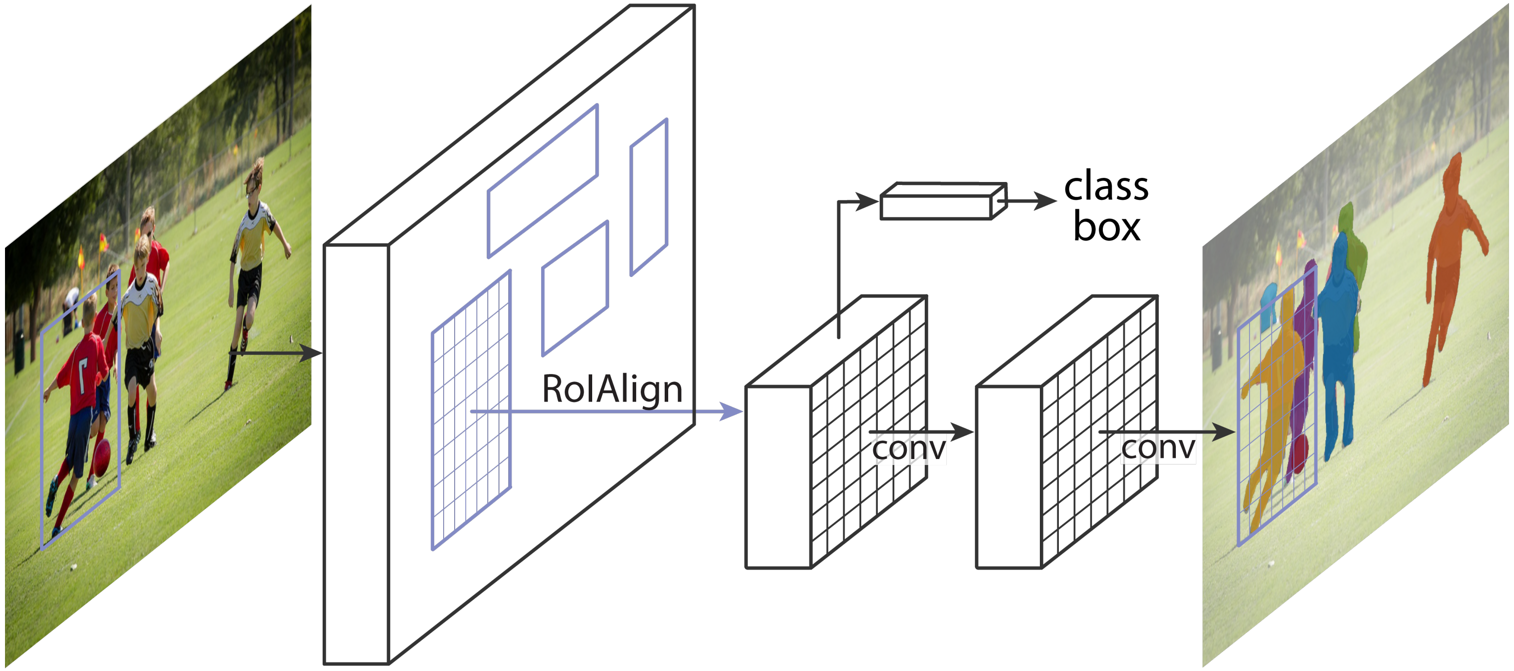


Figure credit: Long et al

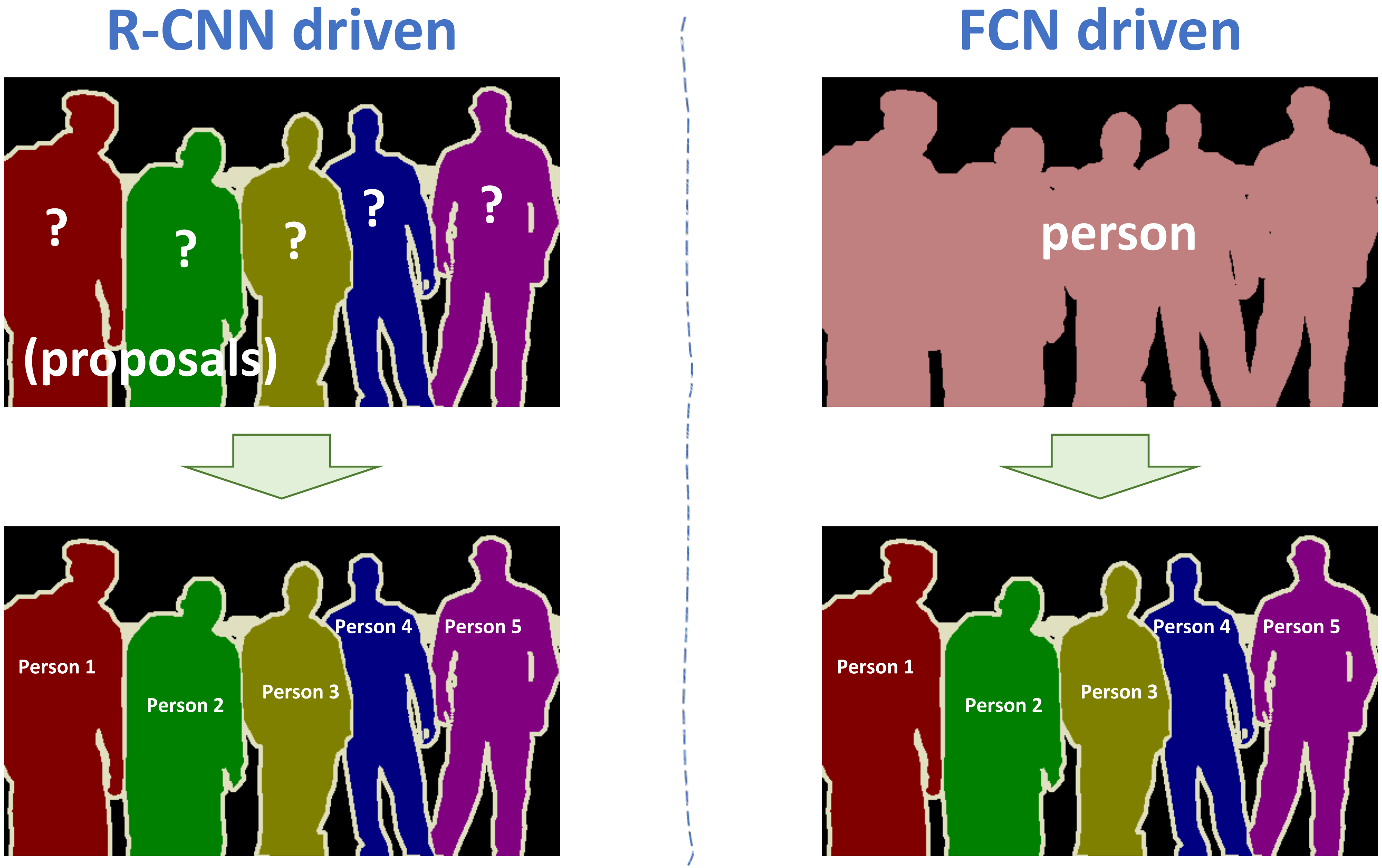
Jonathan Long, Evan Shelhamer, & Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". CVPR 2015.

Instance Segmentation

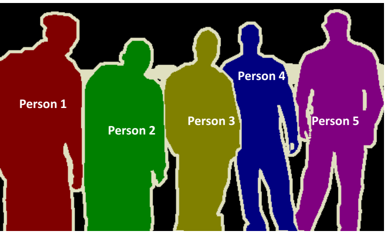
- **Goals** of Mask R-CNN
  - ✓Good speed
  - ✓Good accuracy
  - ✓Intuitive
  - ✓Easy to use



Instance Segmentation Methods

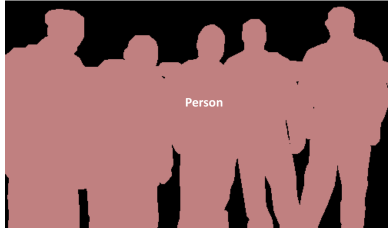


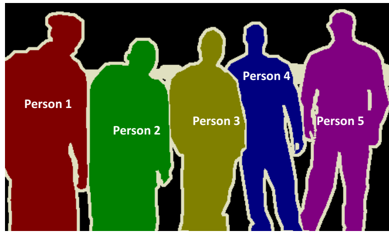
Instance Segmentation Methods



RCNN-driven

- SDS [Hariharan et al, ECCV'14]
- HyperCol [Hariharan et al, CVPR'15]
- CFM [Dai et al, CVPR'15]
- MNC [Dai et al, CVPR'16]





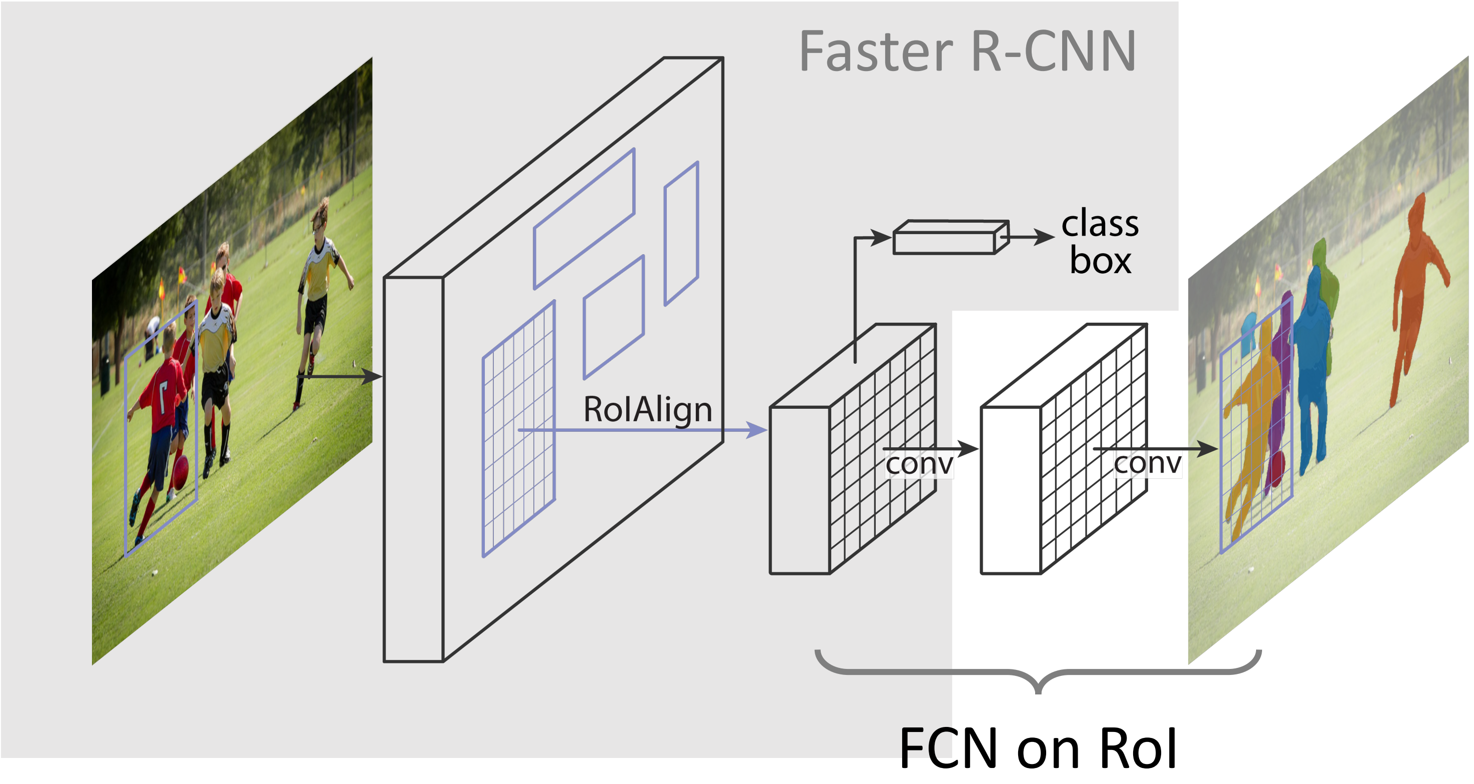
FCN-driven

- PFN [Liang et al, arXiv'15]
- InstanceCut [Kirillov et al, CVPR'17]
- Watershed [Bai & Urtasun, CVPR'17]

- FCIS [Li et al, CVPR'17]
- DIN [Arnab & Torr, CVPR'17]

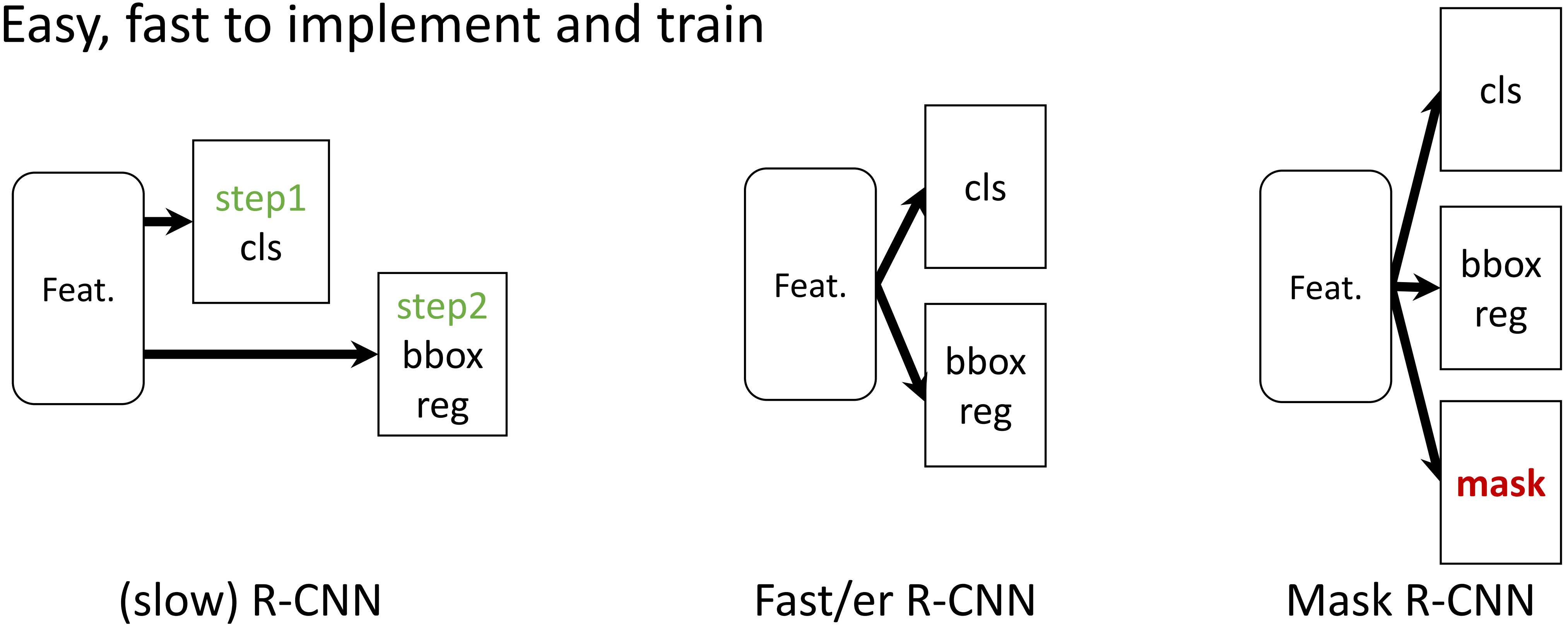
Mask R-CNN

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



Parallel Heads

- Easy, fast to implement and train



A Perspective on Equivariance

How can we draw the *Apple* logo?

How can we draw the *Apple* logo?

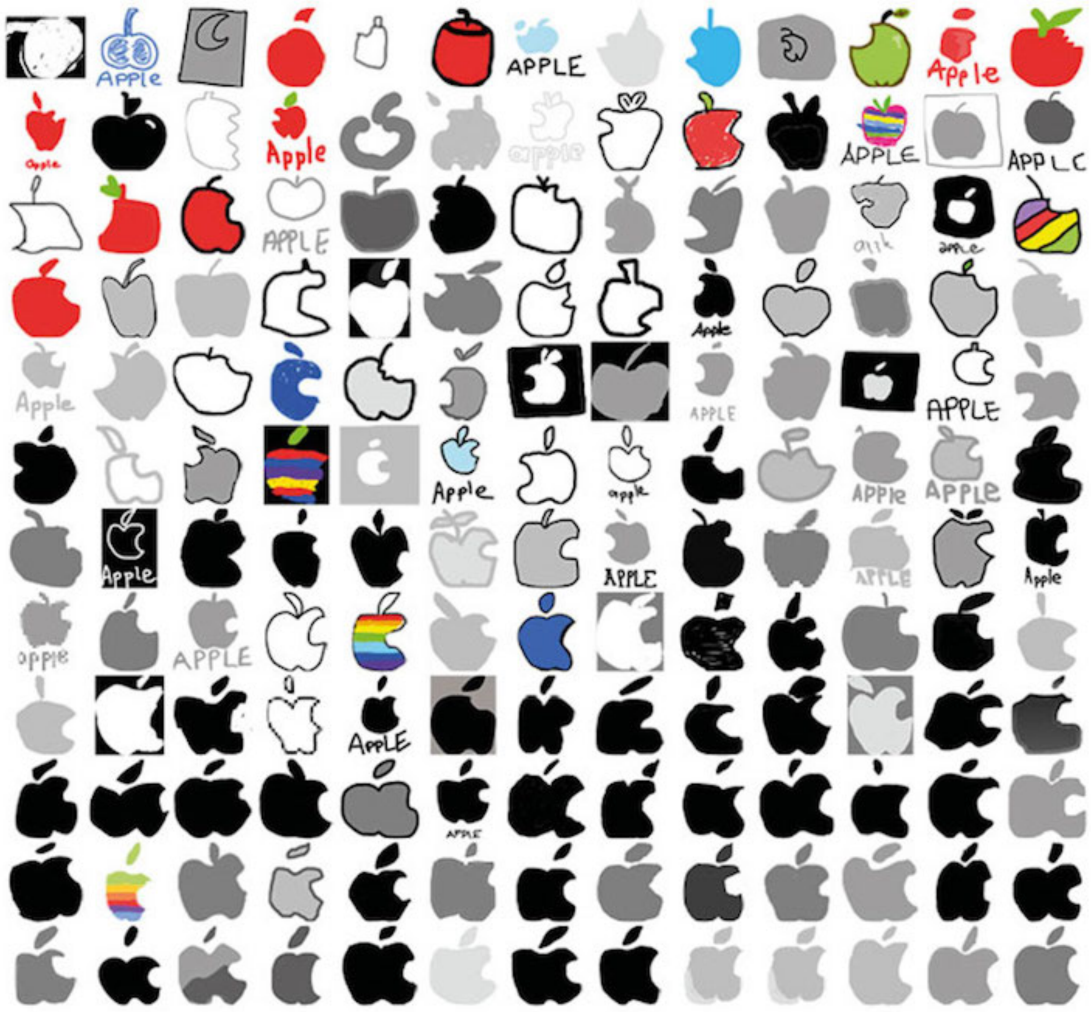


figure source: http://mymodernmet.com/famous-company-logos-memory/

How can we draw the *Apple* logo?

ground truth

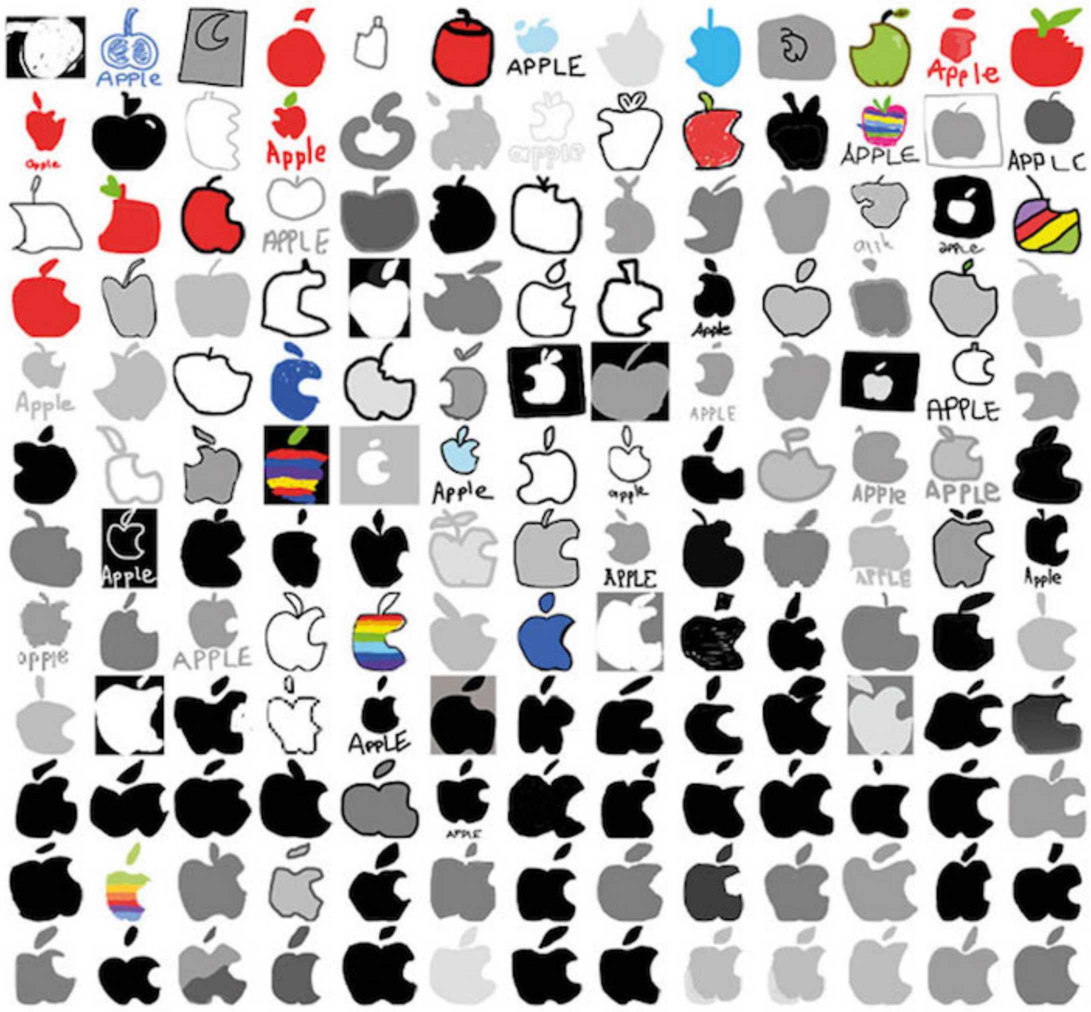
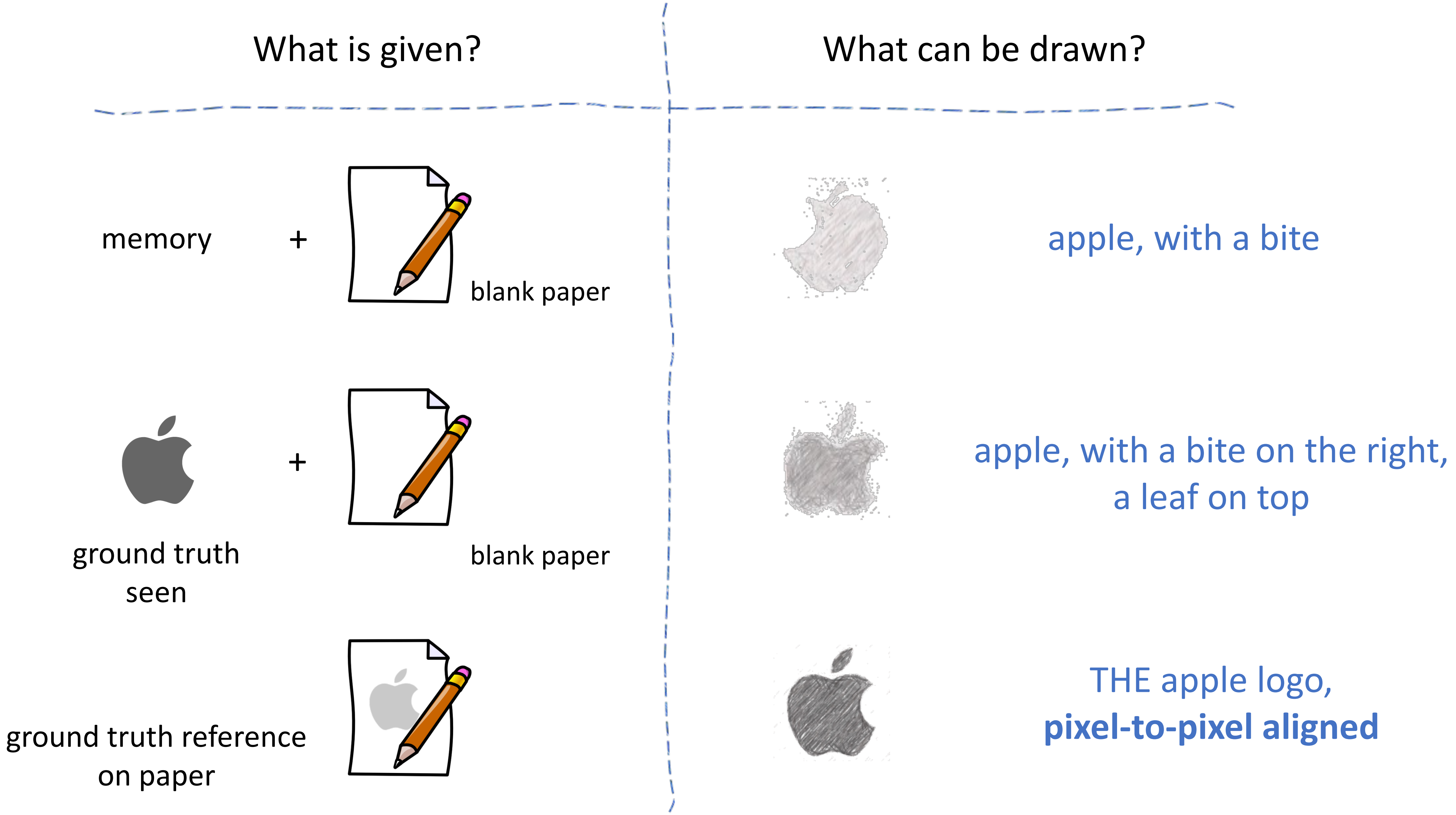
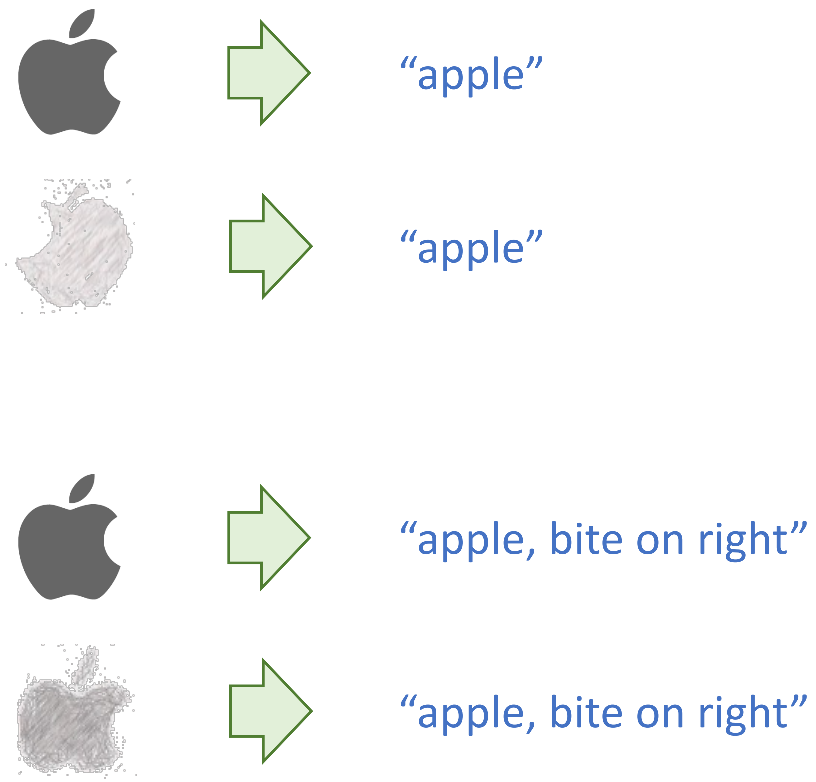


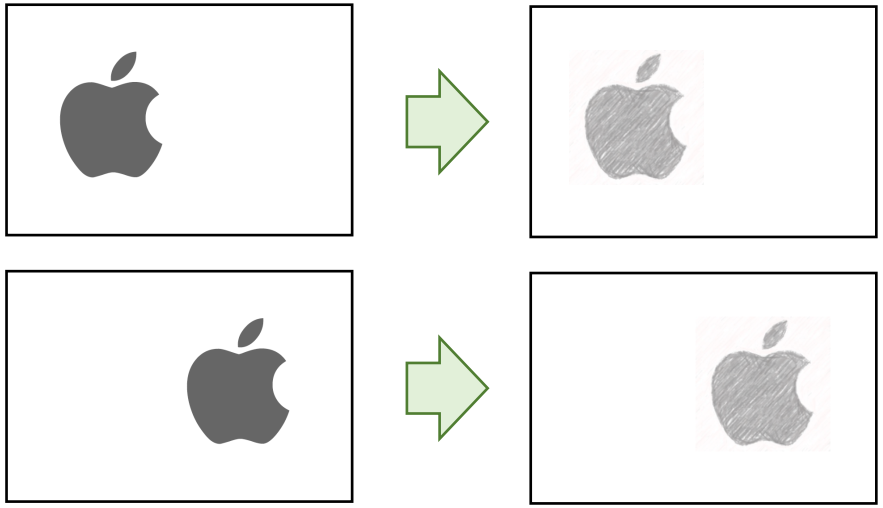
figure source: http://mymodernmet.com/famous-company-logos-memory/



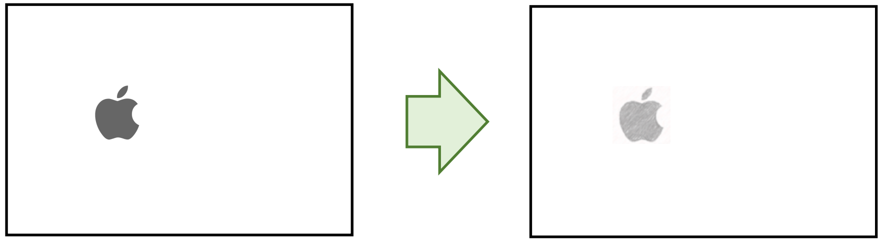
What is given?
What can be drawn?
memory
+
blank paper
apple, with a bite
ground truth
seen
+
blank paper
apple, with a bite on the right,
a leaf on top
ground truth reference
on paper
THE apple logo,
pixel-to-pixel aligned

Invariance vs. Equivariance





translation-equivariant



scale-equivariant

see also “What is wrong with convolutional neural nets?”, Geoffrey Hinton, 2017

Invariance vs. Equivariance

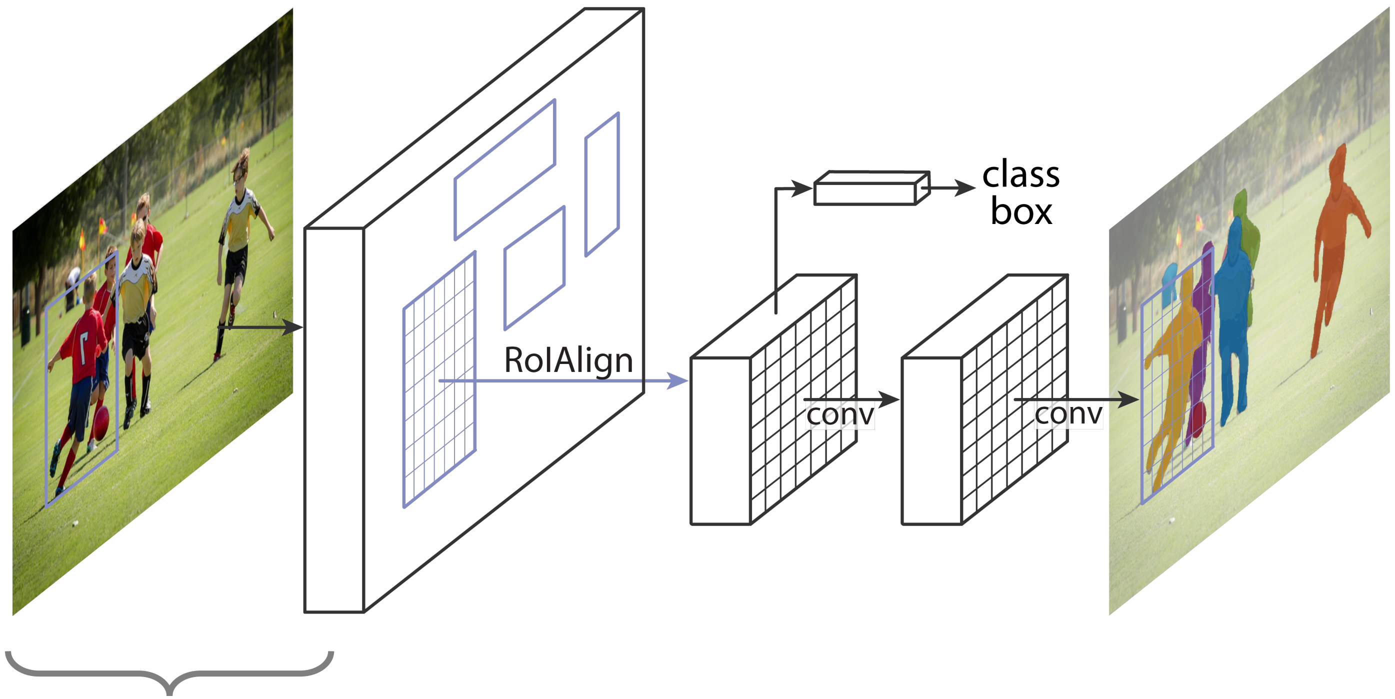
- **Equivariance**: changes in input lead to corresponding changes in output

- *Classification* desires *invariant* representations: output a label
- *Instance Seg.* desires *equivariant* representations:
  - Translated object => translated mask
  - Scaled object => scaled mask
  - *Big and small* objects are equally important (due to AP metric)
    - unlike semantic seg. (counting pixels)

Invariance vs. Equivariance

- Convolutions are translation-equivariant
- *Fully*-ConvNet (FCN) is translation-equivariant
- ConvNet becomes translation-invariant due to fully-connected or global pool layers

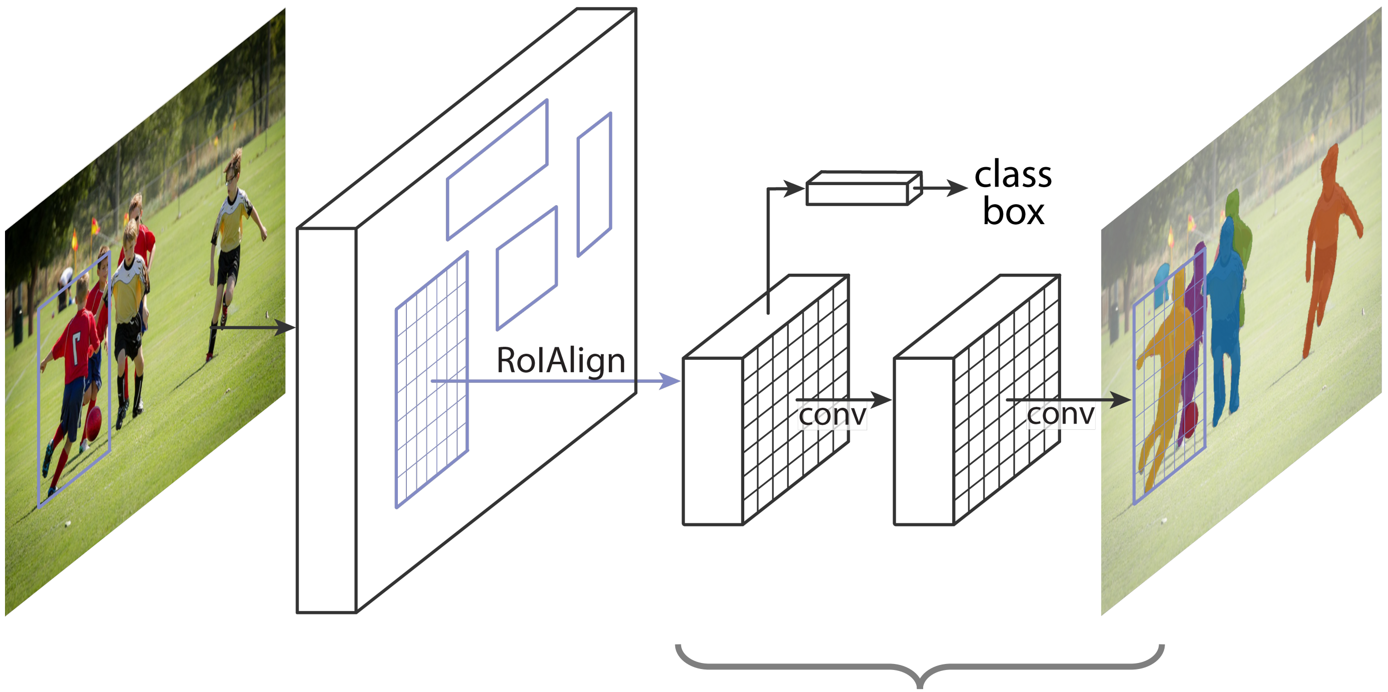
Equivariance in Mask R-CNN



1. Fully-Conv Features:

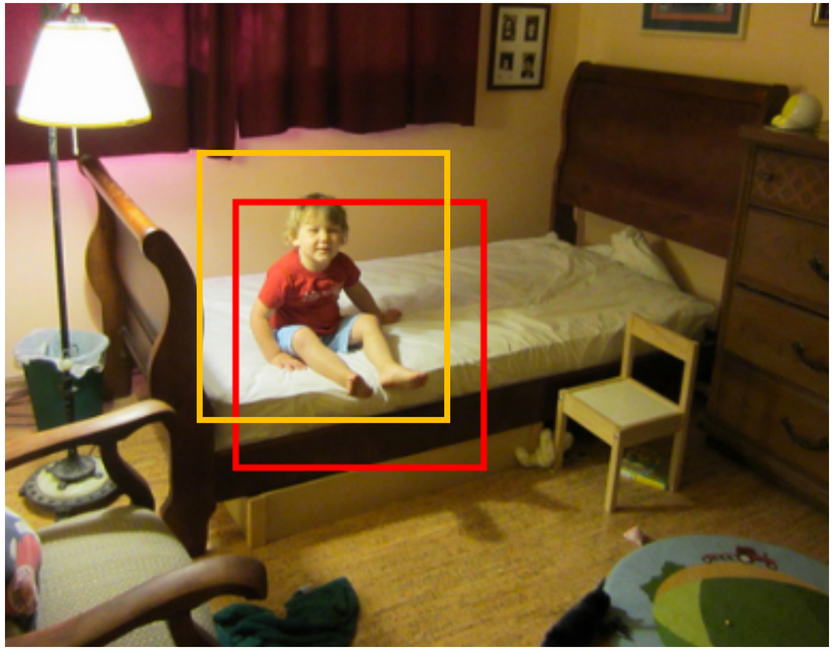
equivariant to global (image) translation

Equivariance in Mask R-CNN

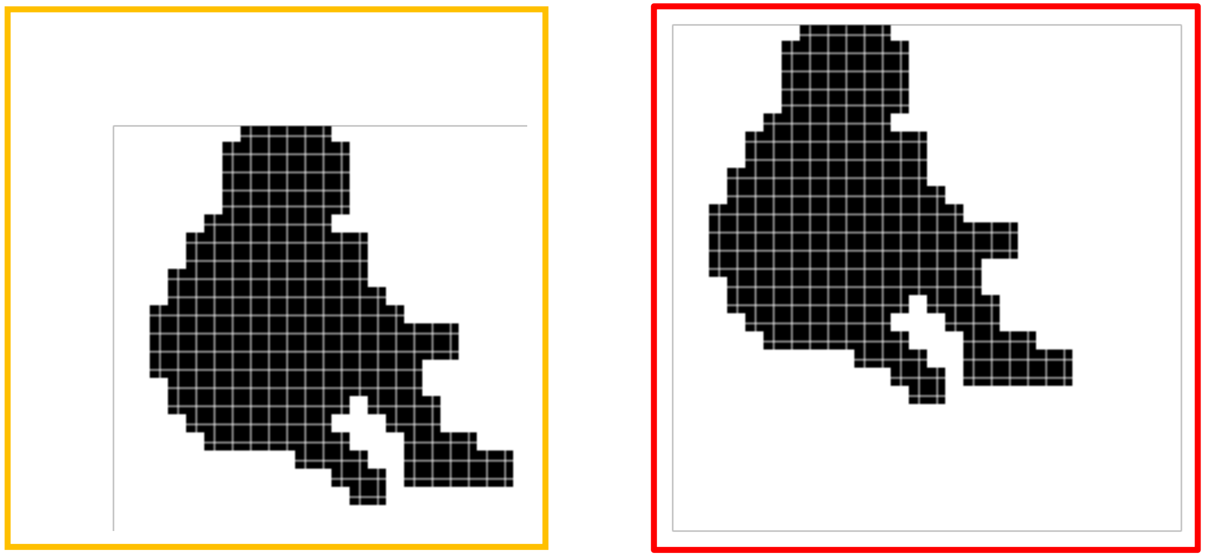


2. Fully-Conv on RoI:
equivariant to translation within RoI

Fully-Conv on RoI



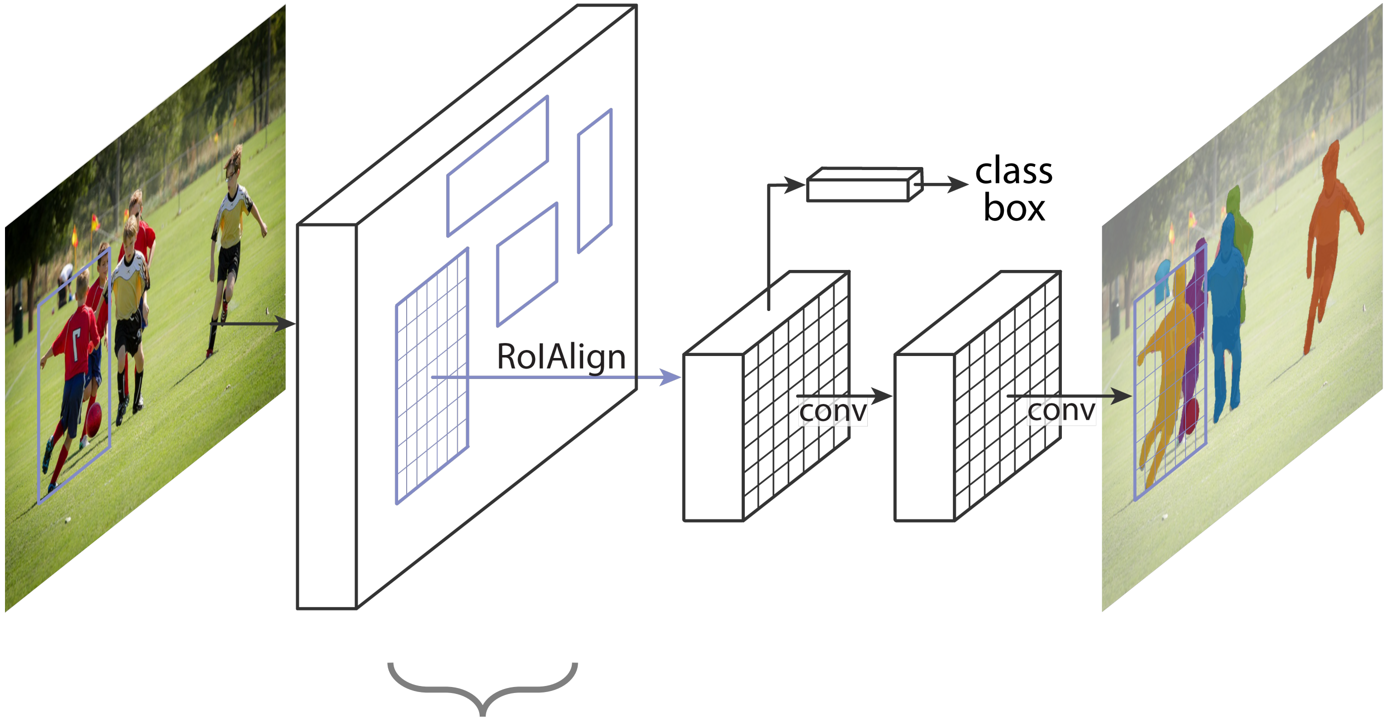
target masks on RoIs



Translation of object in RoI => Same translation of mask in RoI

- Equivariant to small translation of RoIs
- More robust to RoI's localization imperfection

Equivariance in Mask R-CNN



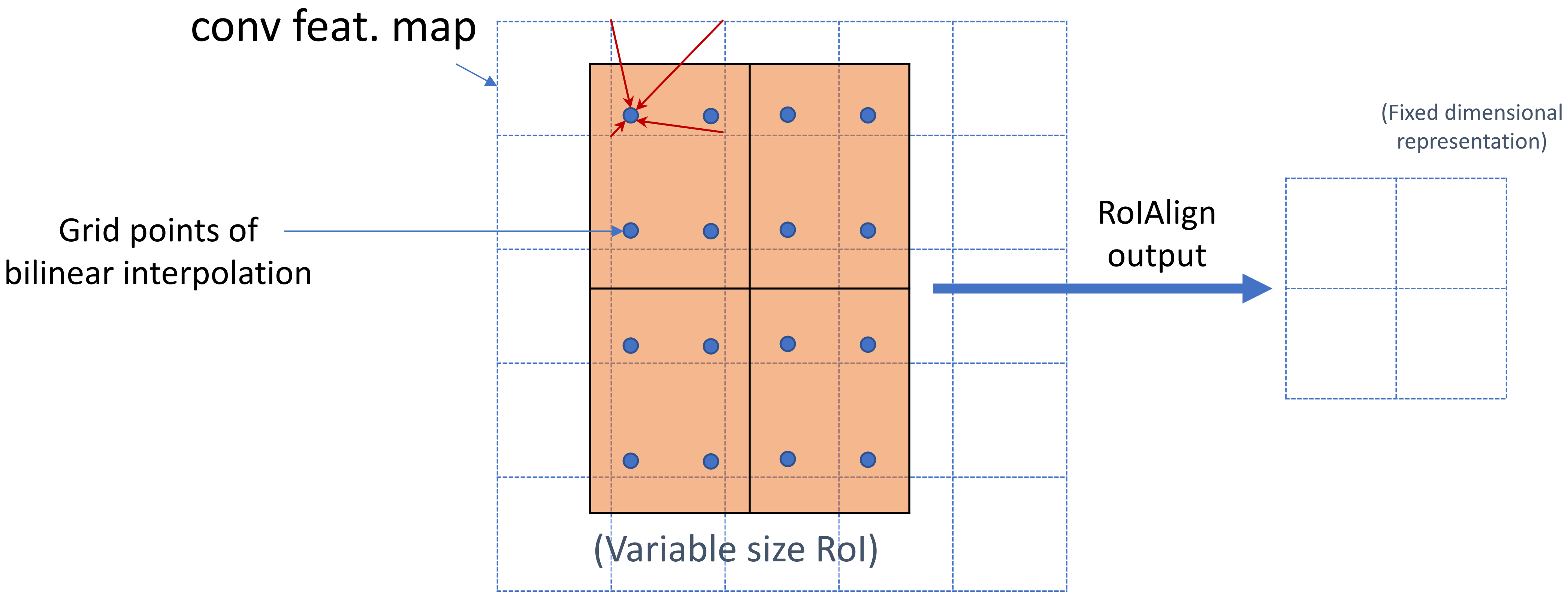
3. RoIAlign:

3a. maintain translation-equivariance before/after RoI

RoIAlign

FAQs: how to sample grid points within a cell?

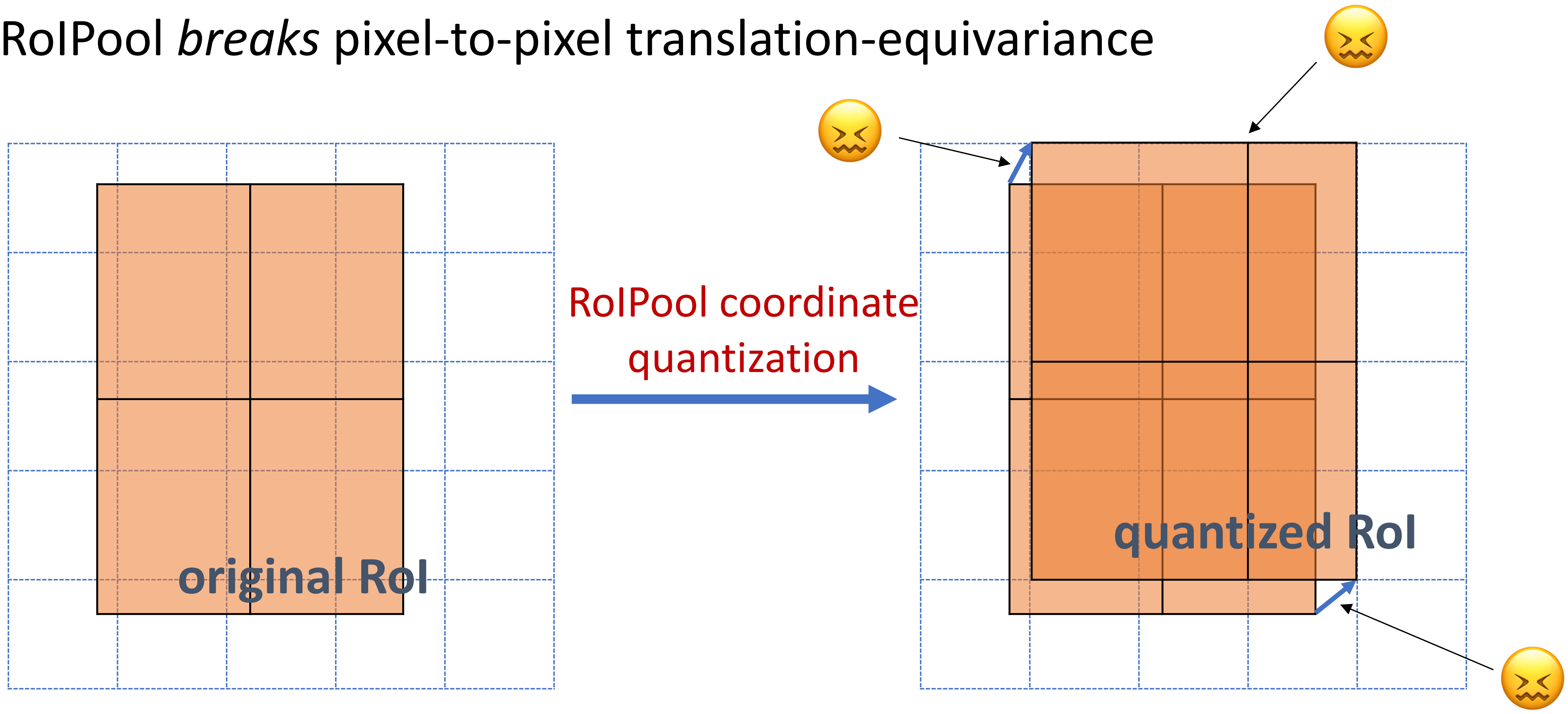
- 4 regular points in 2x2 sub-cells
- other implementation could work



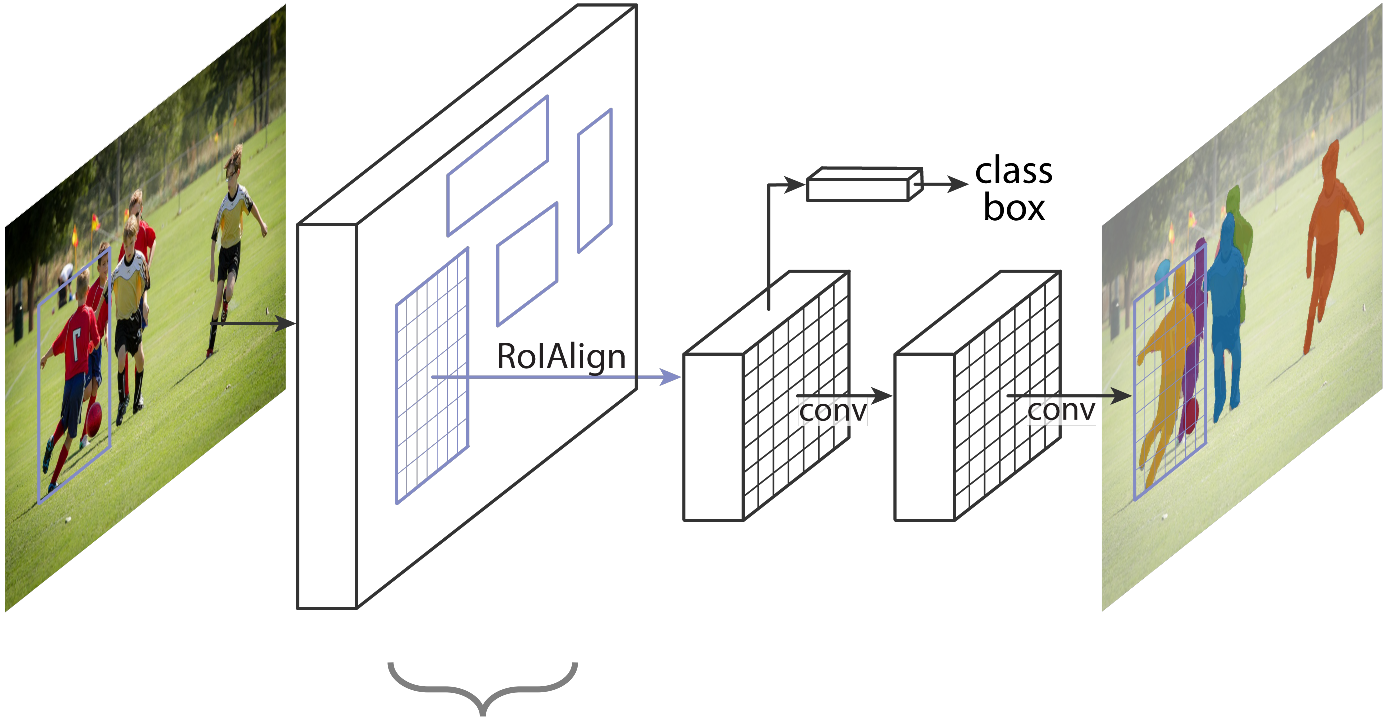
see also “What is wrong with convolutional neural nets?”, Geoffrey Hinton, 2017

RoIAlign vs. RoIPool

- RoIPool *breaks* pixel-to-pixel translation-equivariance



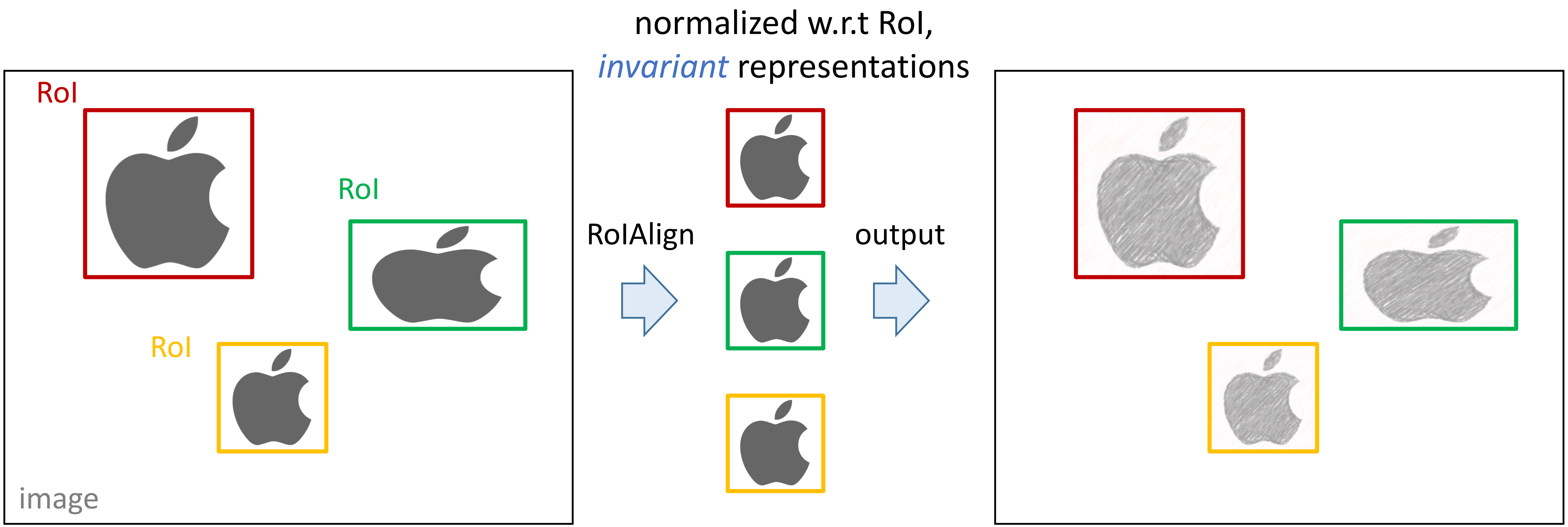
Equivariance in Mask R-CNN



3. RoIAlign:

3b. Scale-equivariant (and aspect-ratio-equivariant)

RoIAlign: Scale-Equivariance

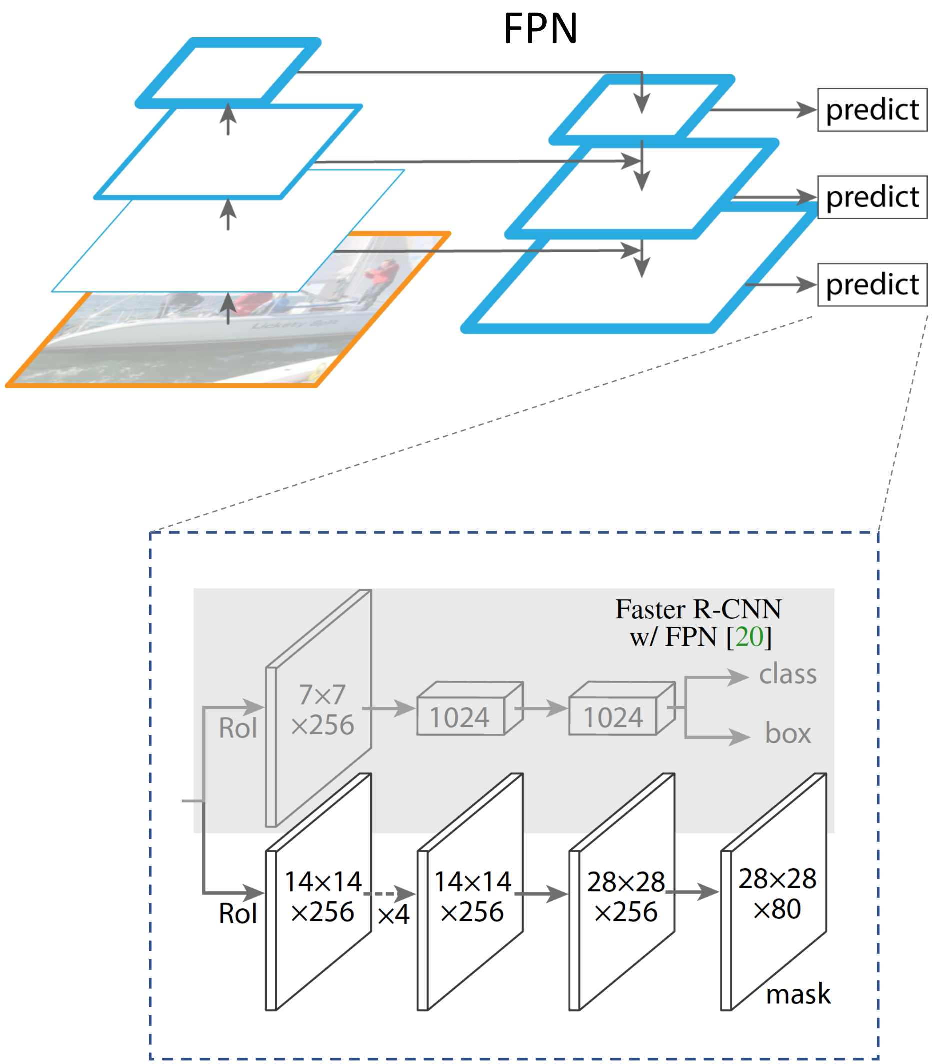


- RoIAlign creates *scale-invariant* representations
- RoIAlign + “output pasted back” provides *scale-equivariance*

More about Scale-Equivariance: FPN

- RoIAlign is scale-invariant if on raw pixels:
  - = (slow) R-CNN: crops and warps RoIs
- RoIAlign is scale-invariant if on scale-invariant feature maps

- Feature Pyramid Network (FPN) [Lin et al. CVPR'17] creates approx. scale-invariant features



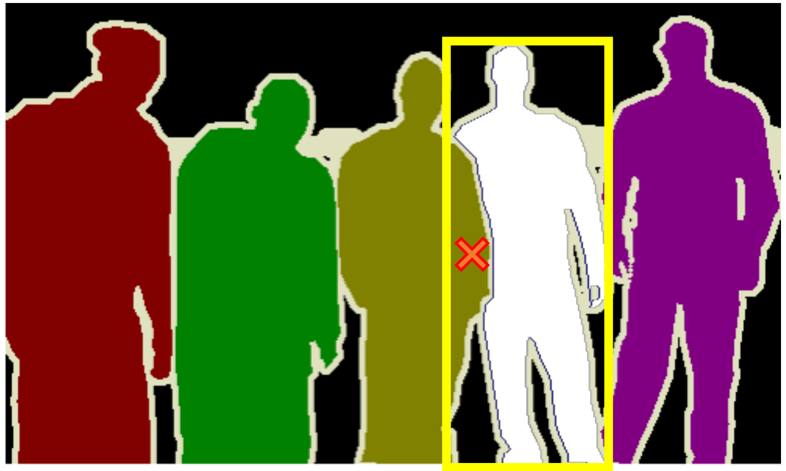
Equivariance in Mask R-CNN: Summary

- Translation-equivariant
  - FCN features
  - FCN mask head
  - RoIAlign (pixel-to-pixel behavior)

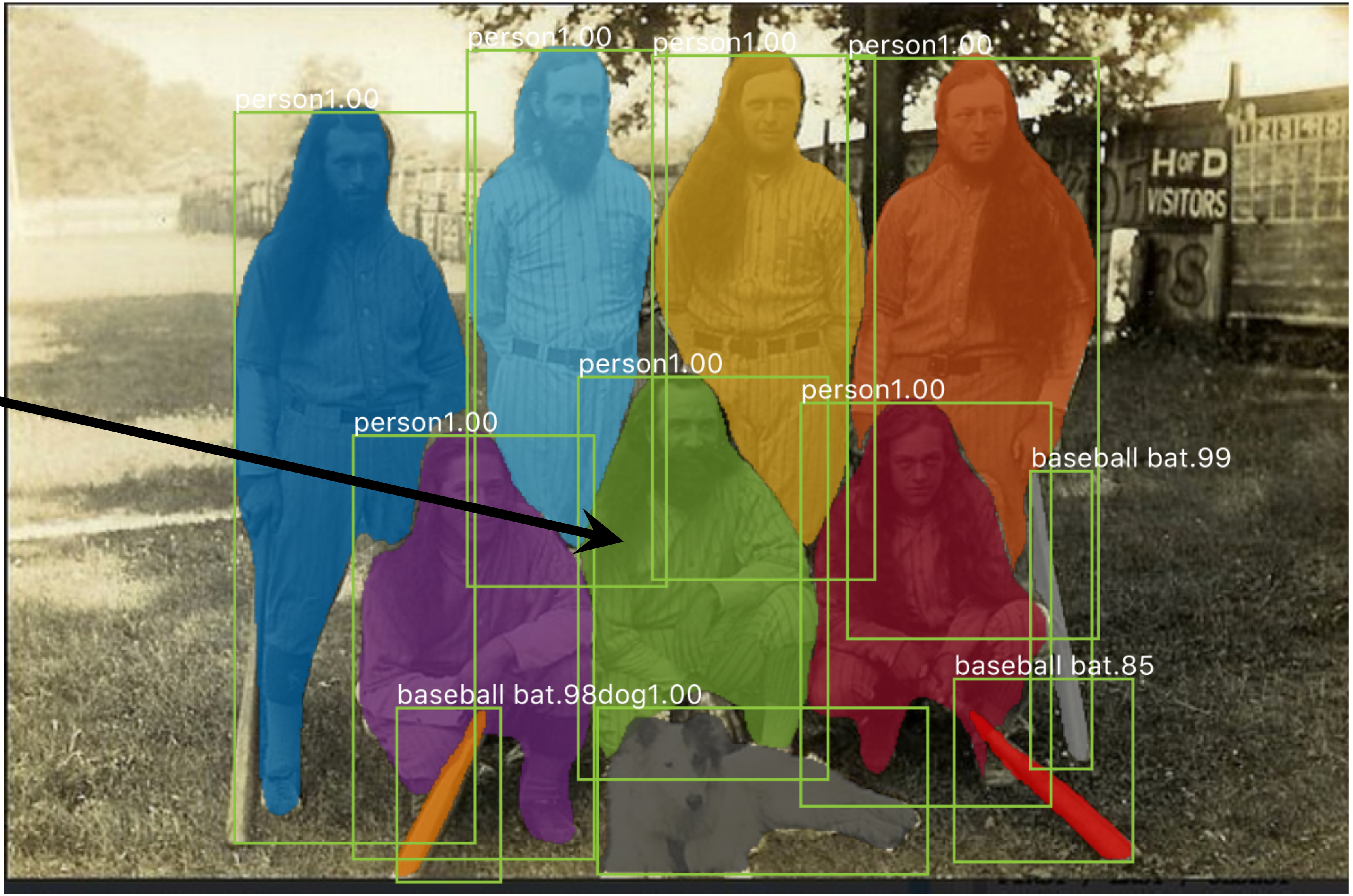
- Scale-equivariant (and aspect-ratio-equivariant)
  - RoIAlign (warping and normalization behavior) + paste-back
  - FPN features

Instance Seg: When we don't want equivariance?

- A pixel x could have a different label w.r.t. different RoIs
  - zero-padding in RoI boundary breaks equivariance
  - outside objects are suppressed
  - only equivariant to small changes of RoIs (which is desired)



object
surrounded by
same-category
objects



Mask R-CNN results on COCO

Result Analysis

Ablation: RoIPool vs. RoIAlign

baseline: ResNet-50-Conv5 backbone, **stride=32**

	mask AP			box AP		
	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*

- huge gain at high IoU,
  in case of big stride (32)

Ablation: RoIPool vs. RoIAlign

baseline: ResNet-50-Conv5 backbone, **stride=32**

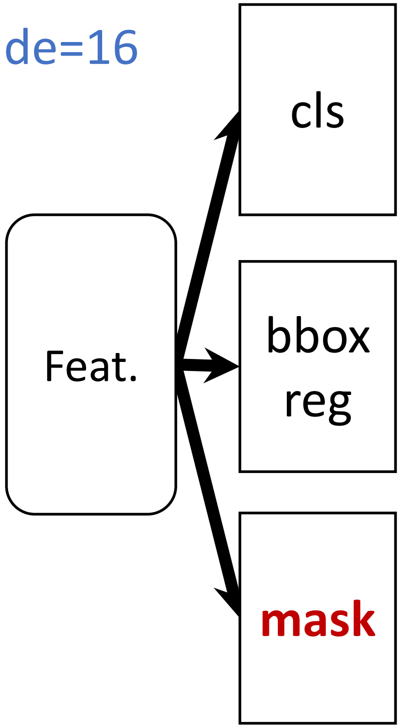
	mask AP			box AP		
	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
RoIPool	23.6	46.5	21.6	28.2	52.7	26.9
RoIAlign	**30.9**	**51.8**	**32.1**	**34.0**	**55.3**	**36.4**
	+*7.3*	+ *5.3*	+*10.5*	+*5.8*	+*2.6*	+*9.5*

- nice box AP without dilation/upsampling

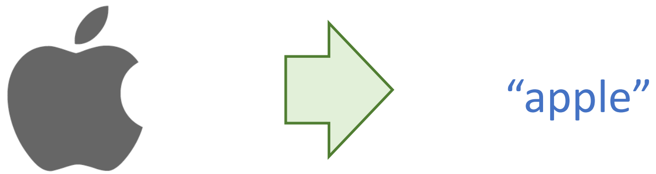
Ablation: Multinomial vs. Binary Masks

baseline: ResNet-50-Conv4 backbone, stride=16

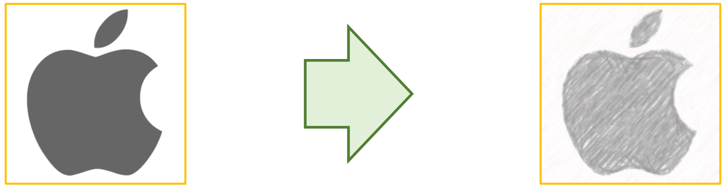
	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*



- cls head: did recognition



- mask head: no need to recognize again



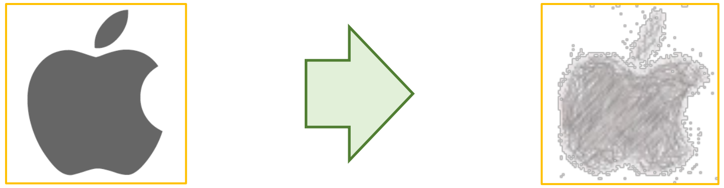
Ablation: MLP vs. FCN mask

baseline: ResNet-50-FPN backbone

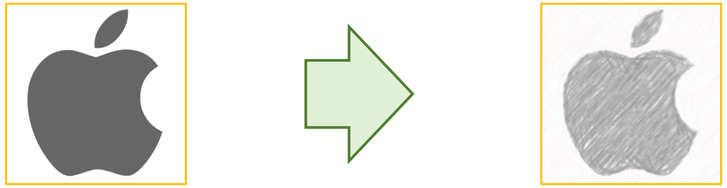
	mask branch	AP	AP_{50}	AP_{75}
MLP	fc: $1024\rightarrow1024\rightarrow80\cdot28^2$	31.5	53.7	32.8
MLP	fc: $1024\rightarrow1024\rightarrow1024\rightarrow80\cdot28^2$	31.5	54.0	32.6
FCN	conv: $256\rightarrow256\rightarrow256\rightarrow256\rightarrow256\rightarrow80$	**33.6**	**55.2**	**35.3**

- +2.1 point

- MLP: lose "place-coded" info, too abstract



- FCN: translation-equivariant



Instance Segmentation Results on COCO

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [20] +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**

- **2 AP better** than SOTA w/ R101, without bells and whistles
- **200ms / img**

Instance Segmentation Results on COCO

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [7]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [20] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [20] +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**

- benefit from better features (ResNeXt [Xie et al. CVPR'17])

Object Detection Results on COCO

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
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	$\boxed{36.2}$	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	**52.1**
Faster R-CNN, RoIAlign	ResNet-101-FPN	$\boxed{37.3}$	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

bbox detection improved by:

- RoIAlign

Object Detection Results on COCO

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
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	**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

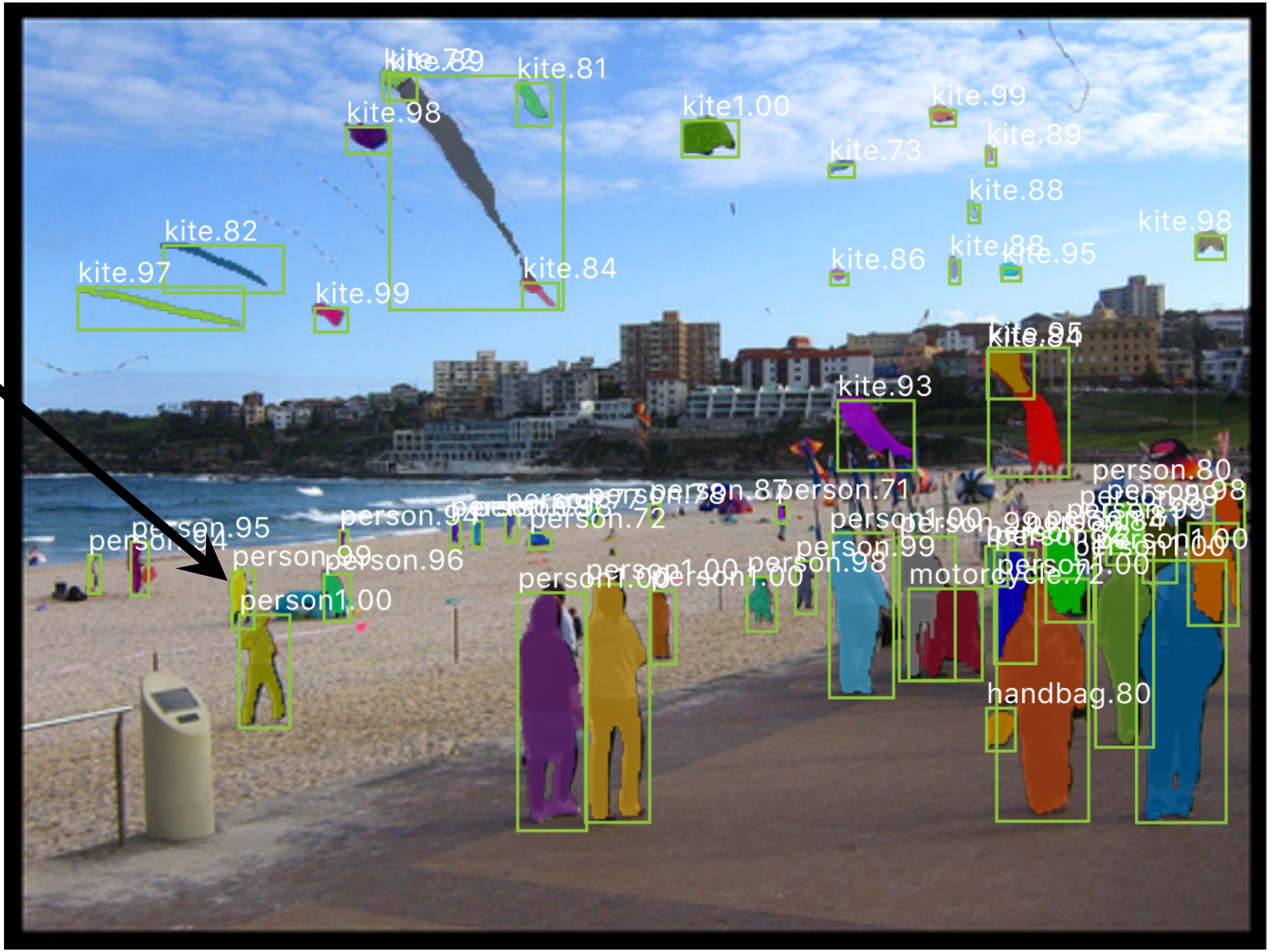
bbox detection improved by:

- RoIAlign
- Multi-task training w/ mask

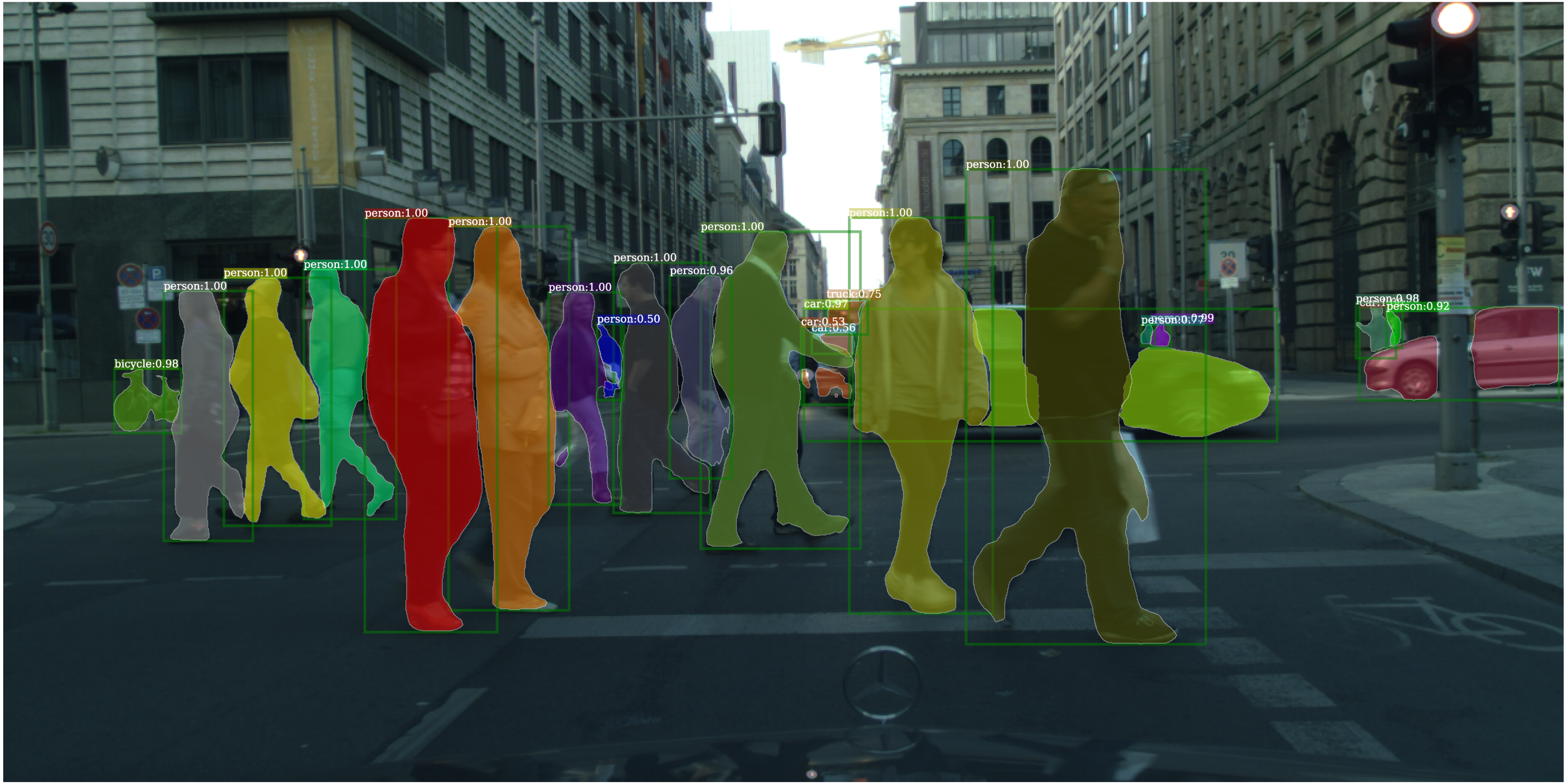
disconnected object

Mask R-CNN results on COCO

small
objects



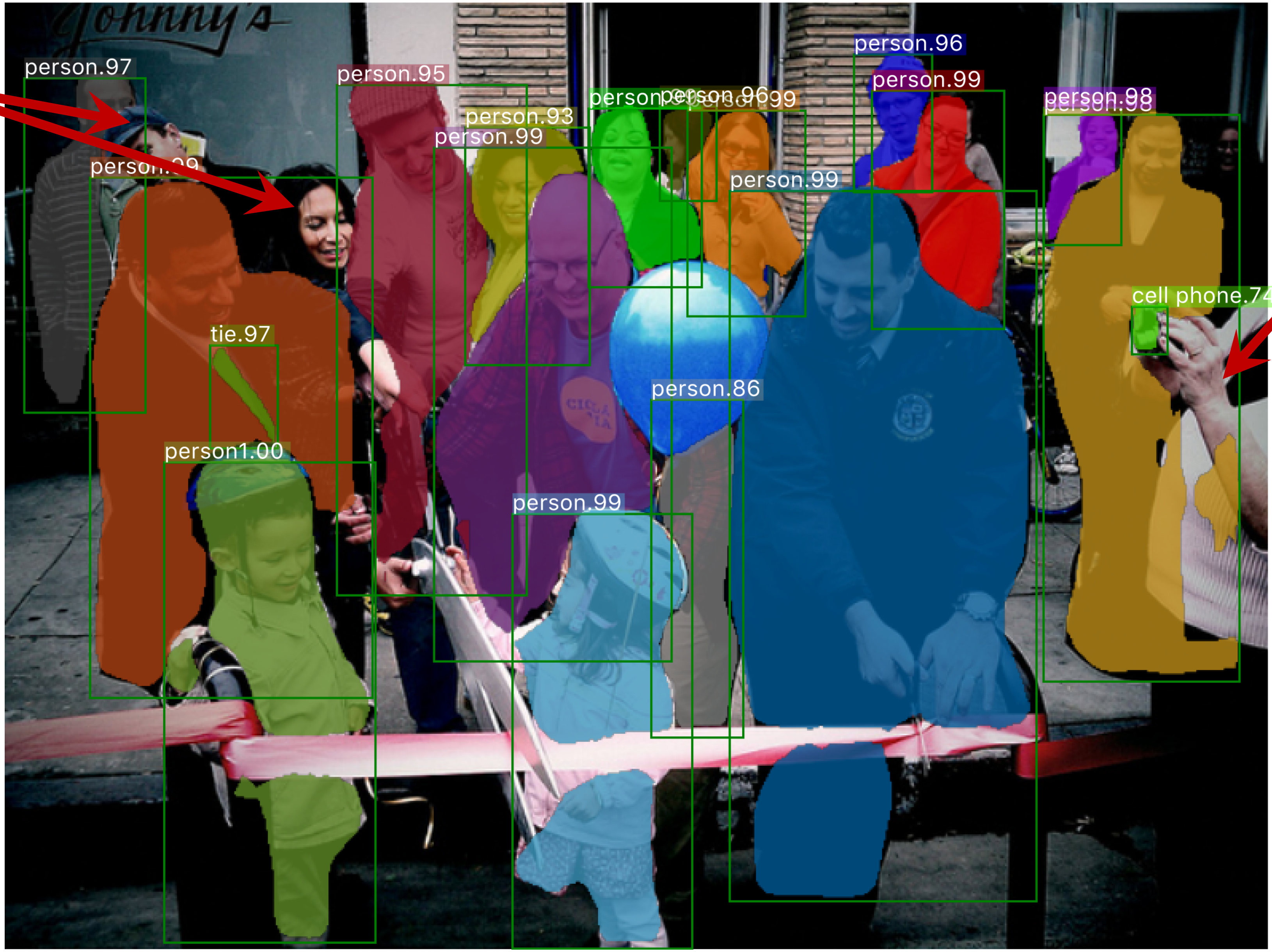
Mask R-CNN results on COCO



Mask R-CNN results on CityScapes

Failure case: detection/segmentation

missing

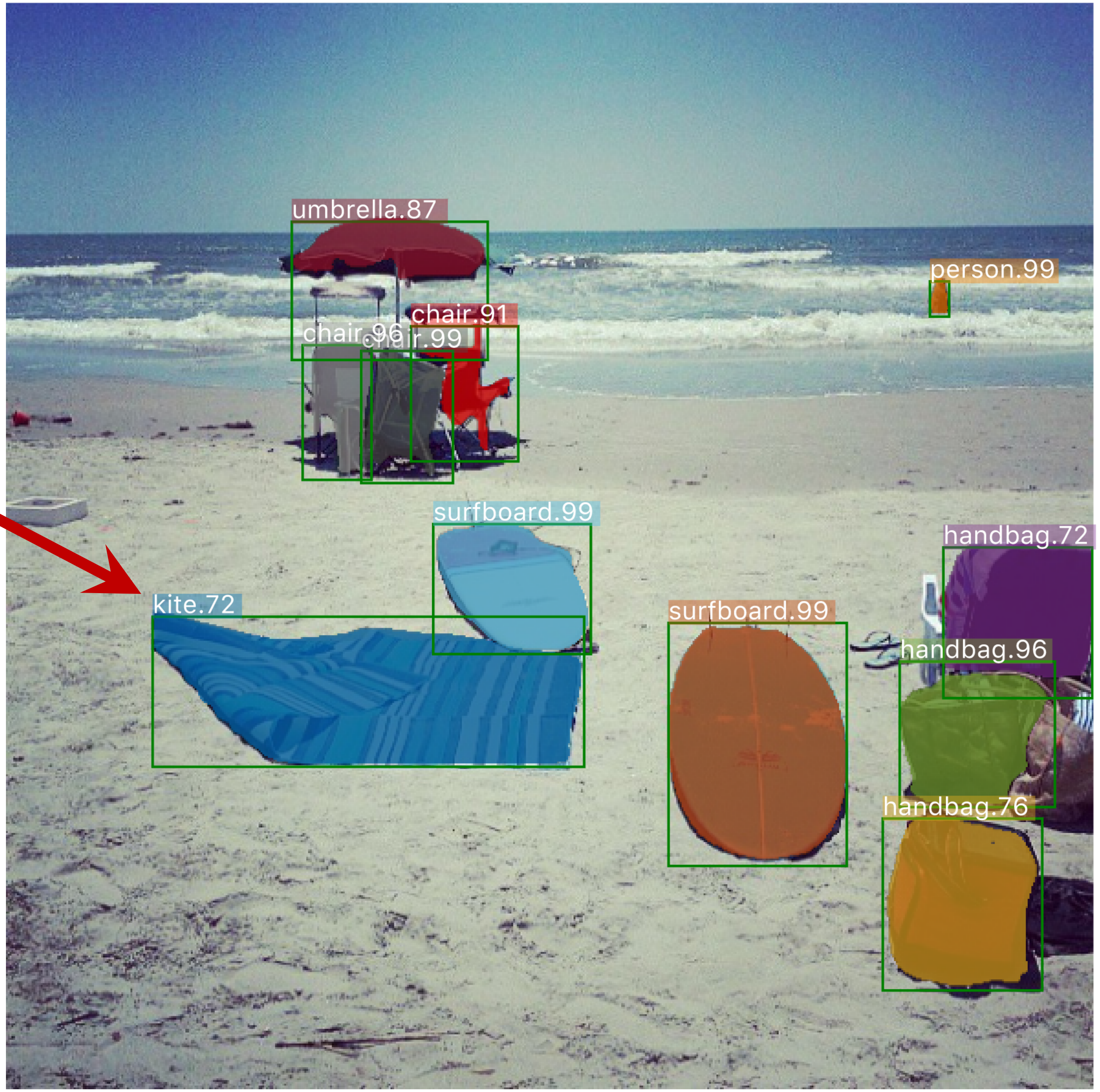


missing,
false mask

Mask R-CNN results on COCO

Failure case: recognition

not a kite

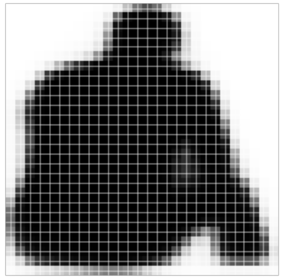


Mask R-CNN results on COCO

Validation image with box detection shown in red

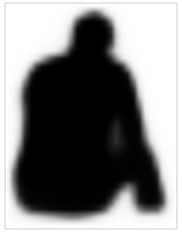
28x28 soft prediction from Mask R-CNN

(enlarged)



Soft prediction resampled to image coordinates

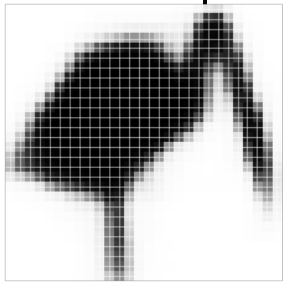
(bilinear and bicubic interpolation work equally well)



Final prediction (threshold at 0.5)

Validation image with box detection shown in red

28x28 soft prediction



Resized Soft prediction

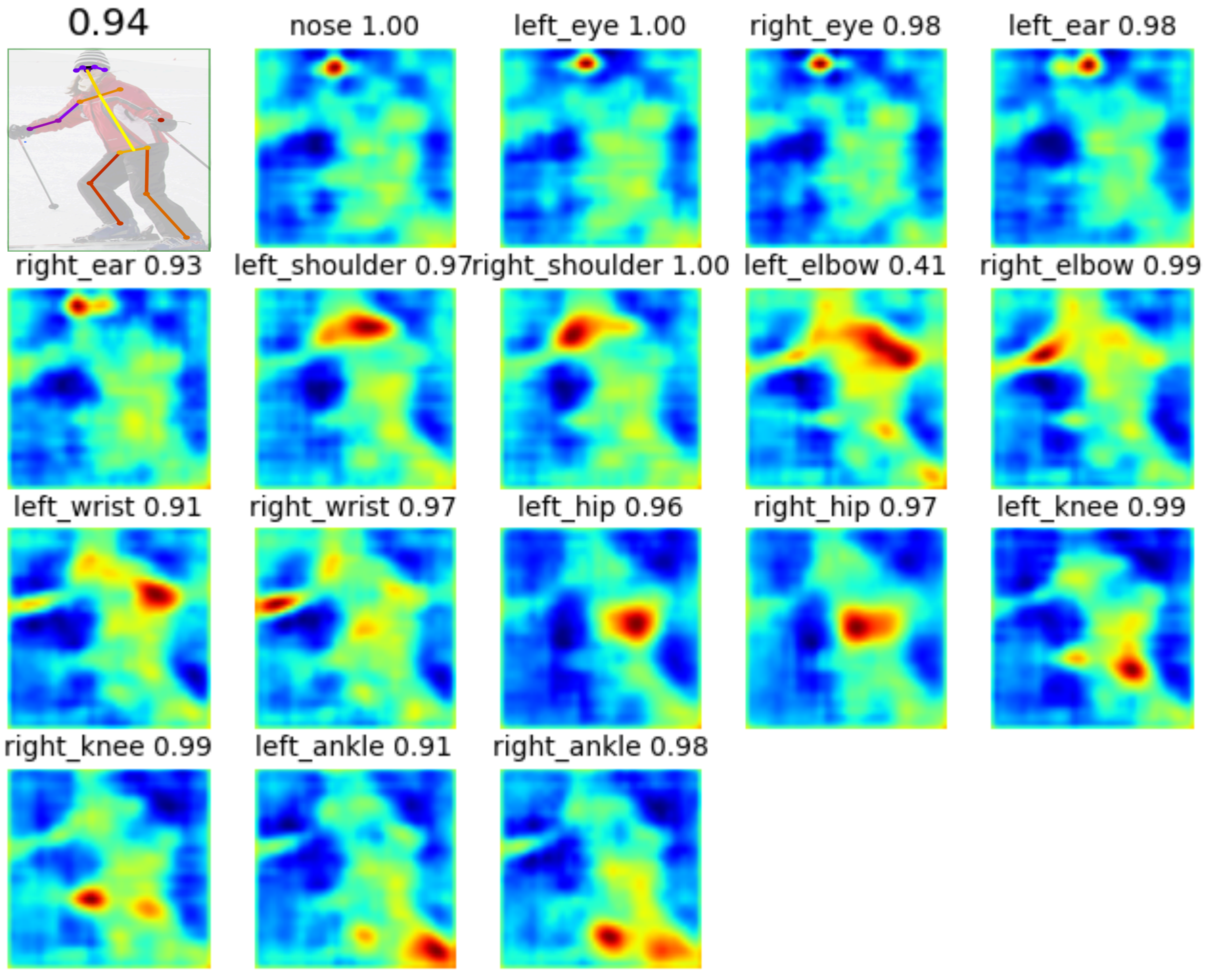
Final mask

Mask R-CNN: for Human Keypoint Detection

- 1 keypoint = 1-hot "mask"
- Human pose = 17 masks

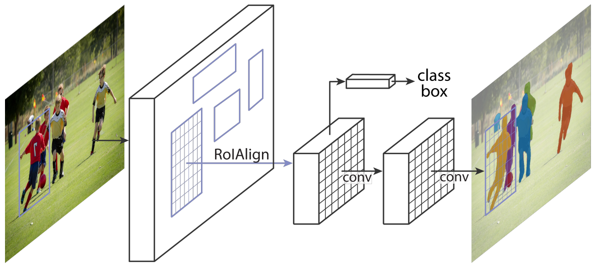
- Softmax over spatial locations
  - e.g. 56^2-way softmax on 56x56

- Desire the same equivariances
  - translation, scale, aspect ratio




www.tfrancis.co.uk

Conclusion



Mask R-CNN

✓Good speed

✓Good accuracy

✓Intuitive

✓Easy to use

✓Equivariance matters

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

More about Mask R-CNN in this ICCV

- **ICCV oral presentation, 10/26, 9am**
- **COCO workshop talk, 10/29, 9am**