Mask R-CNN

Outzen Berild, Martin
&
Martinussen, Jakob Gerhard

Overview

- Framework name: **Mask R-CNN**
- Task type: **Object instance segmentation**
- Submission date: **March, 2017**
- Authors: **Facebook AI Research (FAIR)**
- Accolades: **Won COCO Stuff Challenge 2017**

facebook Artificial Intelligence

Problem Complexity

Instance Segmentation

=

Object Detection

+

Semantic Segmentation

=

Complexity ?

“[...] one might expect a complex method is required to achieve good results. However, we show that a surprisingly simple, flexible, and fast system can surpass prior state-of-the-art instance segmentation results.”

Starting Point

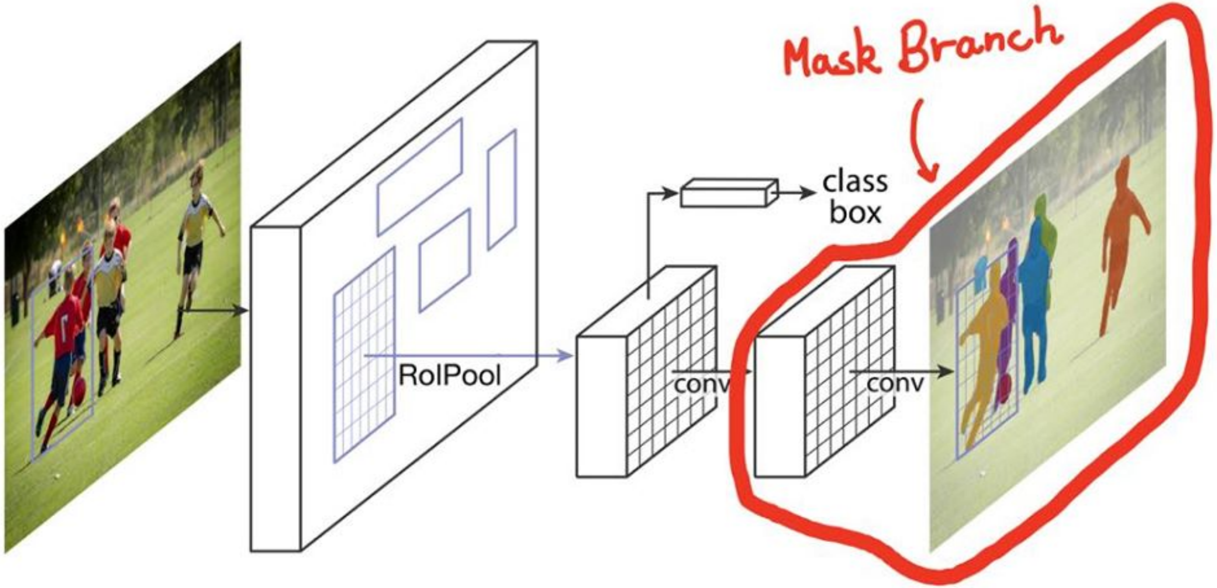
- Extending *Fast R-CNN* in order to predict a segmentation mask

Faster R-CNN = Classification
+
Bounding Box Regression

Mask R-CNN = Faster R-CNN
+
Segmentation Mask

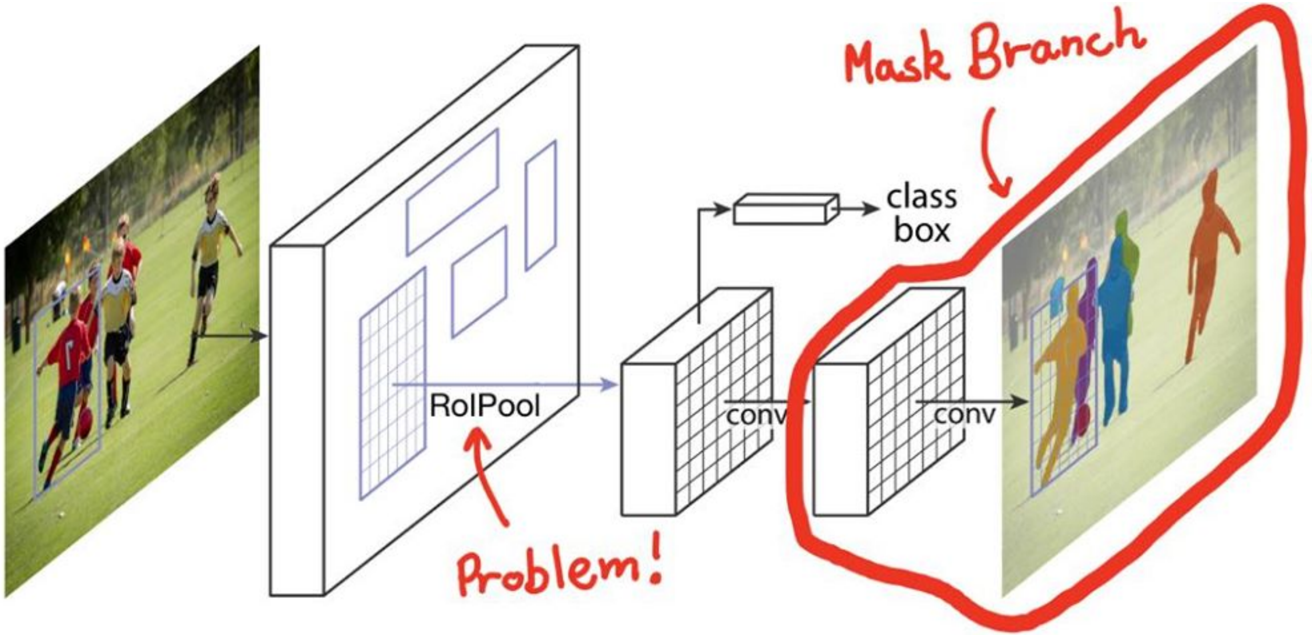
Proposal

Apply a **parallel** *fully convolutional network* (FCN) mask branch to each *region of interest* (RoI)



Problem

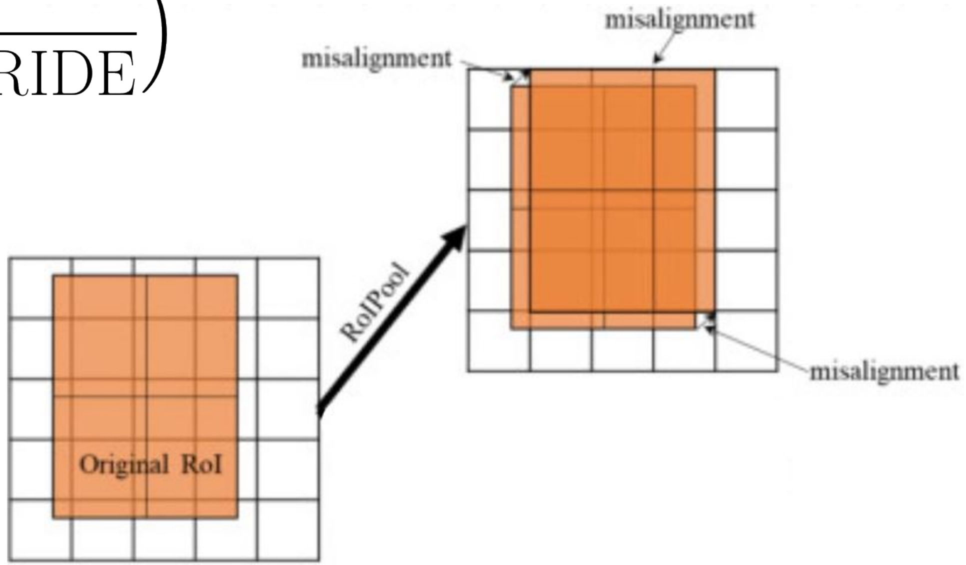
RoIPool introduces a misalignment of input vs. output during quantization into spatial bins.



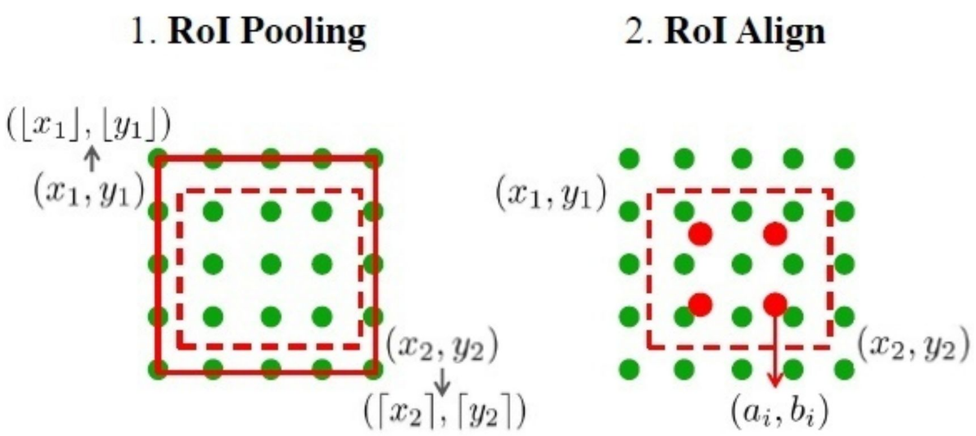
RoIPool Quantization

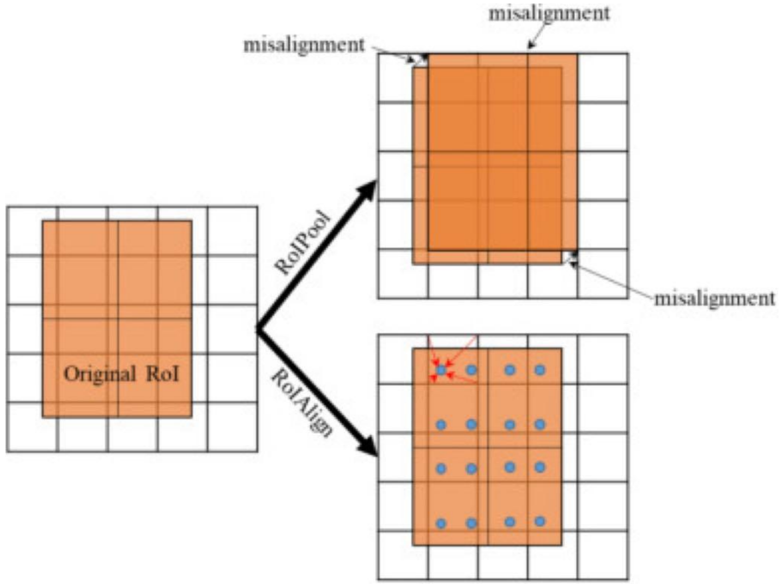
RoIPool quantizes a continuous coordinate *x* by computing:

$$\mathrm{round}\left(\frac{x}{\mathrm{FEATURE\_MAP\_STRIDE}}\right)$$



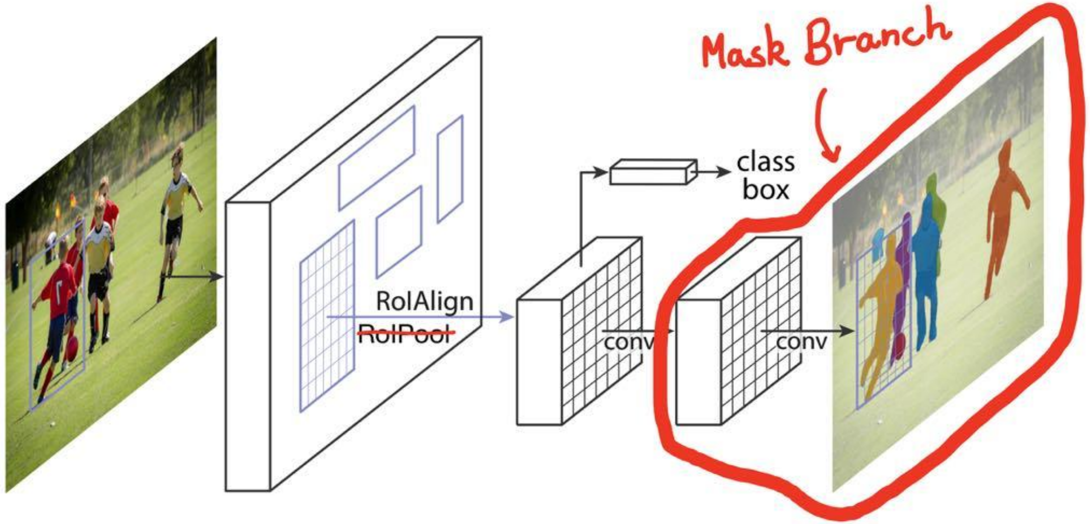
Solution - Bilinear Interpolation





RoIPool → RoIAlign

“[...] RoIAlign has a large impact: it improves mask accuracy by relative 10% to 50%, showing bigger gains under stricter localization metrics.”



	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**

(c) **RoIAlign** (ResNet-50-C4): Mask results with various RoI layers. Our RoIAlign layer improves AP by ~3 points and AP_{75} by ~5 points. Using proper alignment is the only factor that contributes to the large gap between RoI layers.

	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*

(d) **RoIAlign** (ResNet-50-**C5**, *stride 32*): Mask-level and box-level AP using *large-stride* features. Misalignments are more severe than with stride-16 features (Table 2c), resulting in big accuracy gaps.

Decoupling Segmentation and Class Prediction

- **Normal FCN approach** - Per-pixel multi-class categorization
  - Segmentation *precedes* recognition, which is slow and less accurate
  - Loss: Per-pixel *softmax* and a *multinomial* cross-entropy loss
  - *“[...] based on our experiments works poorly for instance segmentation”*

- **Mask R-CNN** - Independent binary mask for each class
  - No class competition
  - RoI classification branch responsible for predicting the class of each pixel
  - *Parallel* prediction of masks and class labels, which is simpler and more flexible
  - Loss: Per-pixel *sigmoid* and a *binary* loss

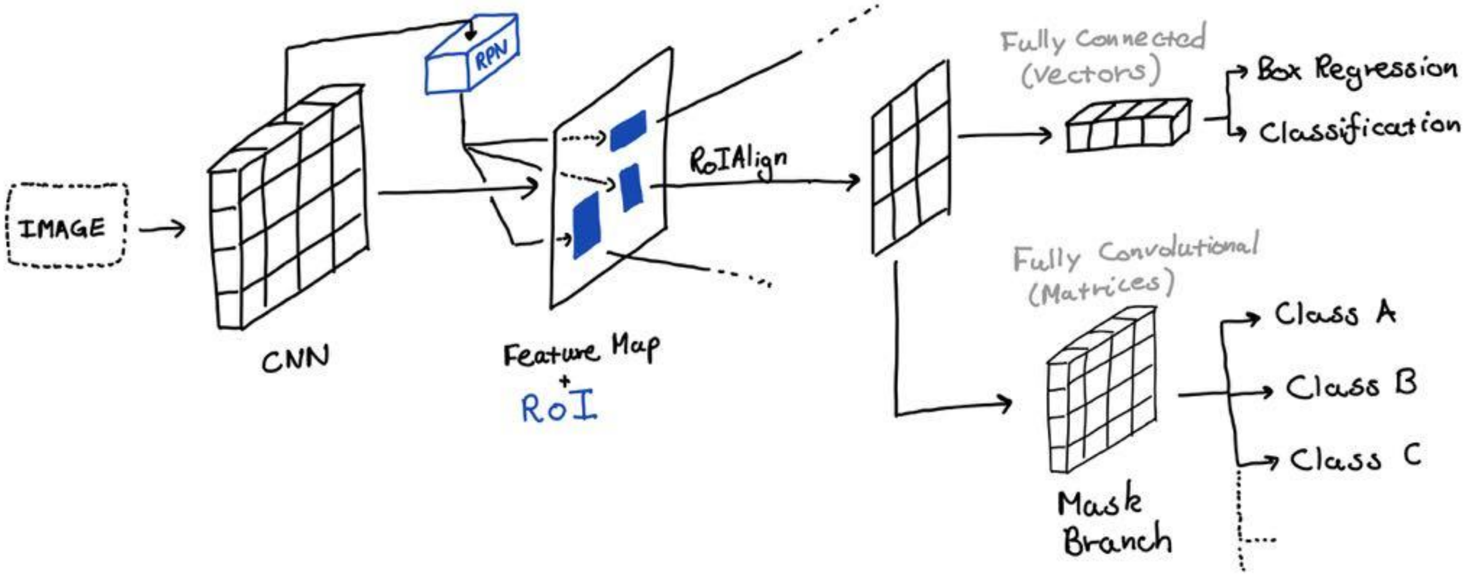


IMAGE
CNN
RPN
Feature Map
+
RoI
RoIAlign
Fully Connected
(vectors)
Box Regression
Classification
Fully Convolutional
(Matrices)
Class A
Class B
Class C
Mask
Branch

Binary Masks

Region of Interest: $m \times m$

Number of classes: K

Mask: Km^2

Loss: $\underbrace{\mathcal{L}_{cls} + \mathcal{L}_{box}}_{\text{Faster R-CNN}} + \underbrace{\mathcal{L}_{mask}}_{\text{sigmoid + BCEL}}$

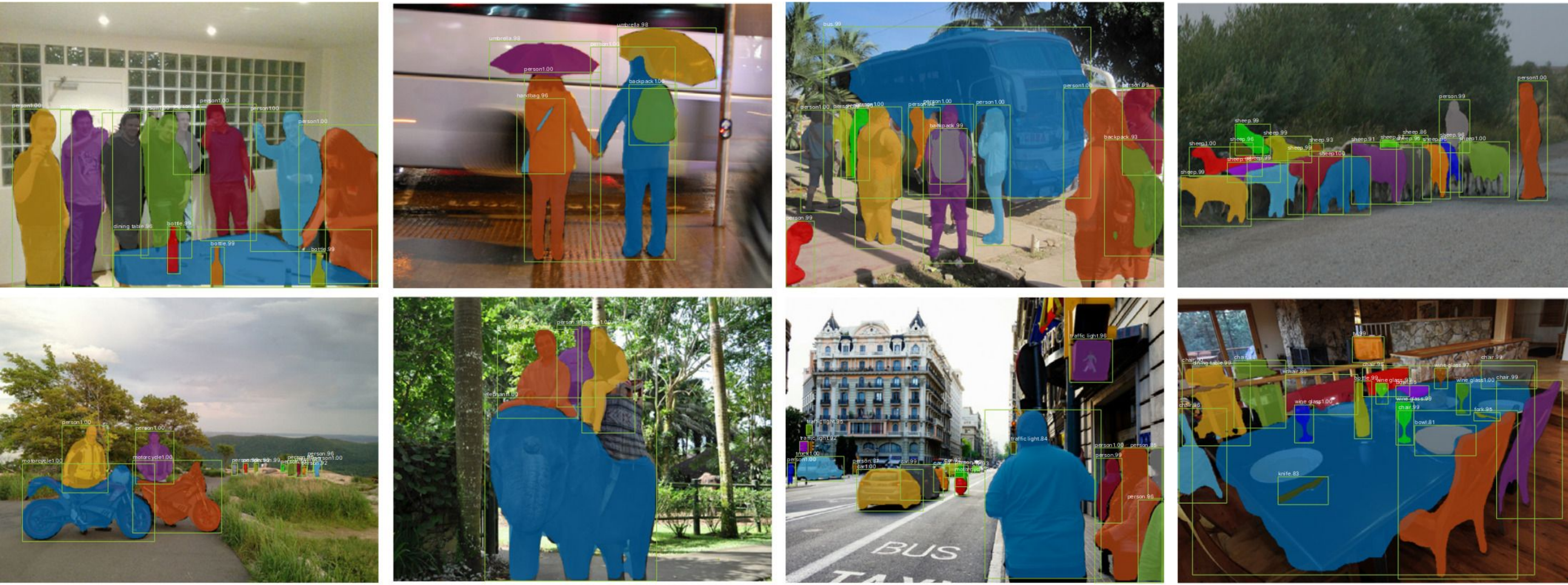
	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(b) **Multinomial *vs*. Independent Masks** (ResNet-50-C4): *Decoupling* via per-class binary masks (sigmoid) gives large gains over multinomial masks (softmax).

	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**

(e) **Mask Branch** (ResNet-50-FPN): Fully convolutional networks (FCN) *vs.* multi-layer perceptrons (MLP, fully-connected) for mask prediction. FCNs improve results as they take advantage of explicitly encoding spatial layout.

Results

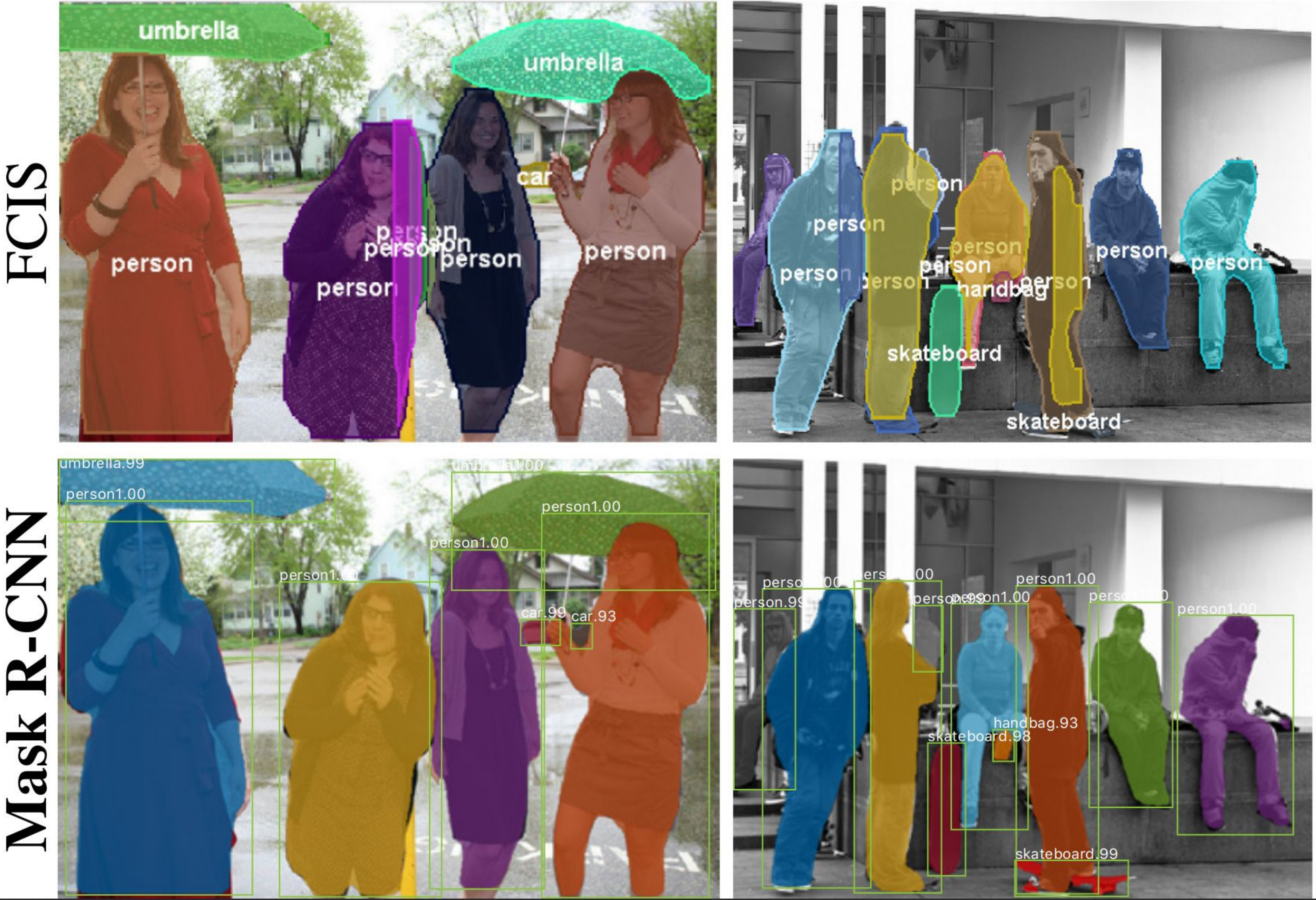


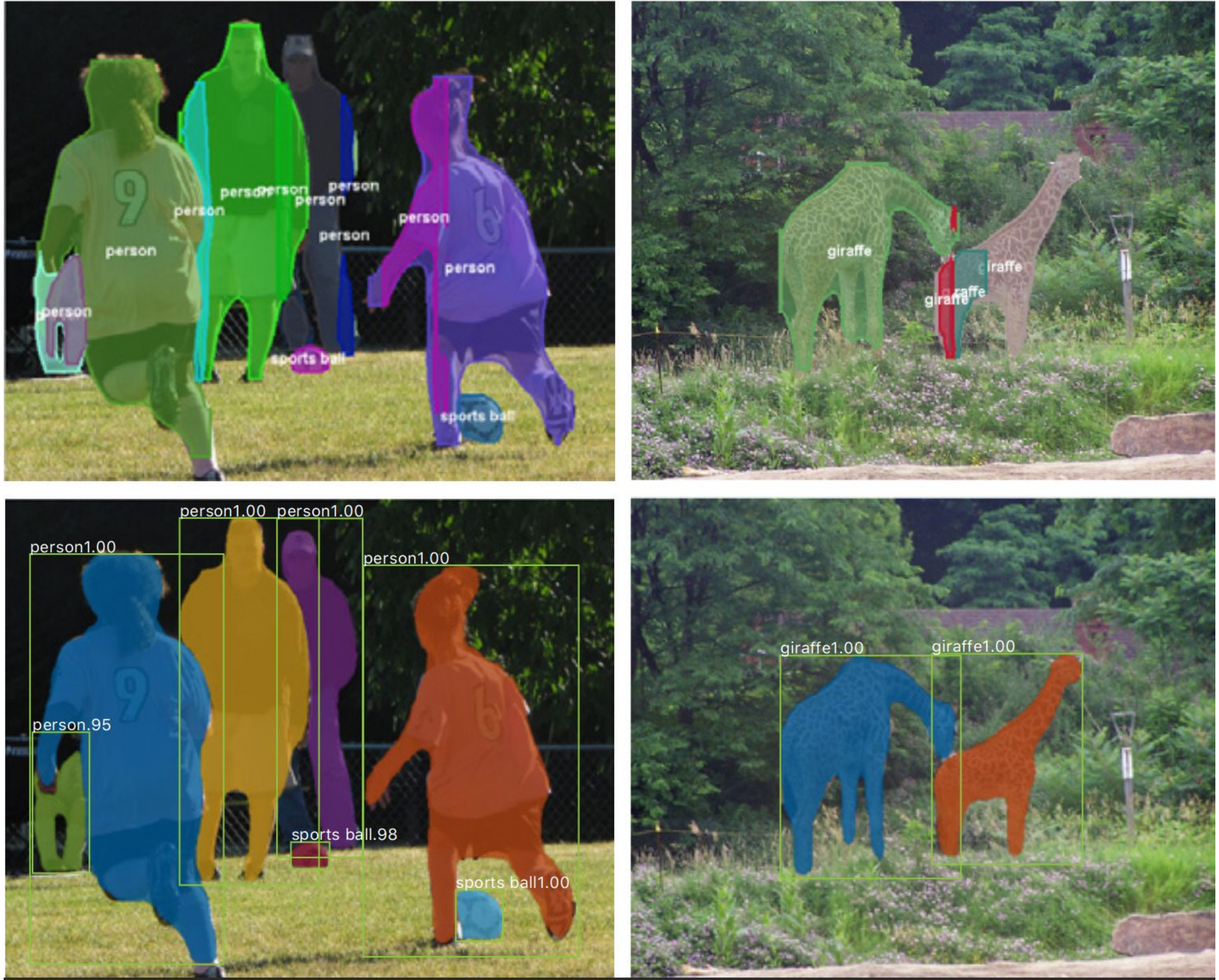
Generalizes to CityScapes

Generalizes to Keypoint Detection

Comparison with FCIS+++

- Position-sensitive output channels by full convolution
- Channels are responsible for *(no decoupling)*:
  - Object classification
  - Bounding box regression
  - Segmentation masks
- Fast but systematic errors on overlapping instances and spurious edges


FCIS
Mask R-CNN
umbrella
umbrella
car
person
handbag
skateboard
skateboard
umbrella.99
person1.00
car.99
car.93
handbag.93
skateboard.98
skateboard.99


person
person
person
person
person
person
person
person
sports ball
sports ball
giraffe
giraffe
person1.00
person1.00
person1.00
person1.00
person.95
sports ball.98
sports ball1.00
giraffe1.00
giraffe1.00


person
person
person
person
tie
tie
train
train
train
train
person1.00
person1.00
tie.95
tie1.00
train.99
train.80
train1.00

Benefits from Deeper Networks

net-depth-features	AP	AP_{50}	AP_{75}
ResNet-50-C4	30.3	51.2	31.5
ResNet-101-C4	32.7	54.2	34.3
ResNet-50-FPN	33.6	55.2	35.3
ResNet-101-FPN	35.4	57.3	37.5
ResNeXt-101-FPN	**36.7**	**59.5**	**38.9**

(a) **Backbone Architecture**: Better backbones bring expected gains: deeper networks do better, FPN outperforms C4 features, and ResNeXt improves on ResNet.

Accolades

- Outperformed every single-model entry on every task in the COCO 2017 Challenge **at time of publication in March**
  - Instance segmentation
  - Bounding-box object detection
  - Person keypoint detection
- FAIR still won *"COCO Stuff"* competition at the end of 2017
- Won *"ICCV 2017 Best Paper Award"*

Key Contributions

RoI alignment

Preservation of pixel alignment in order to predict pixel-accurate segmentation masks.

Independent masks

Decouples mask and class prediction. Mask branch segments independently for each class, while box branch decides on final labels. Negligible computational overhead and less complexity.

Use FCN and not FC layers in mask branch

Using fully convolutional network in mask prediction captures spatial information.

Class-agnostic masks

A single binary mask regardless of class can be nearly as effective provided proper division of labor.