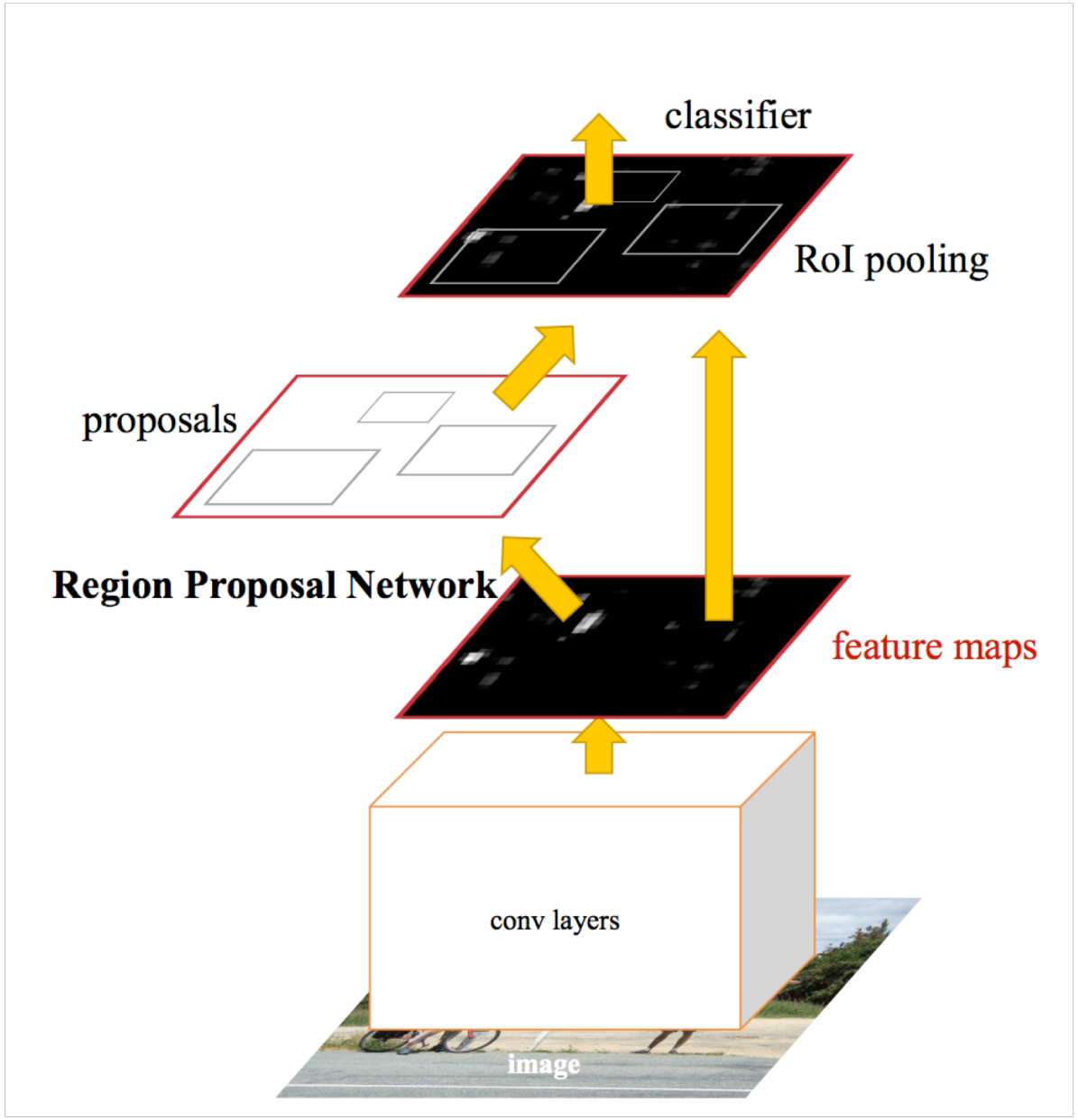
Object detection

Faster R-CNN

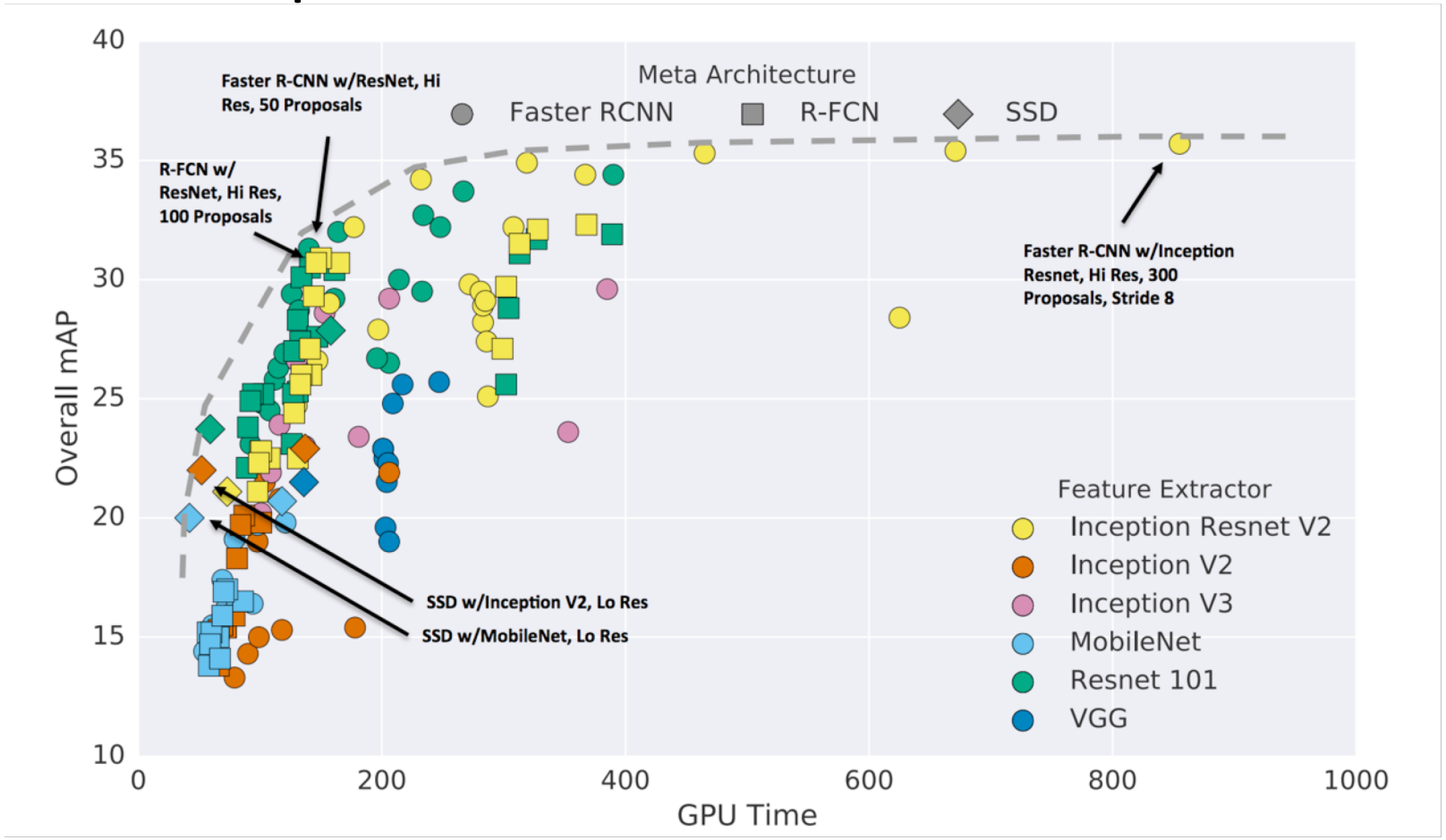

classifier
RoI pooling
proposals
Region Proposal Network
feature maps
conv layers
image

Other details - Non-max suppression

Other details - Non-max suppression

- Go down the list of detections starting from highest scoring
- Eliminate any detection that overlaps highly with a higher scoring detection
- Separate, heuristic step

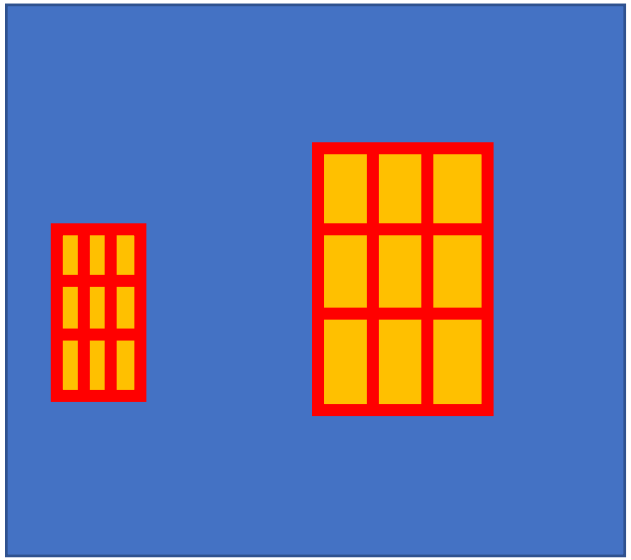
A comprehensive evaluation

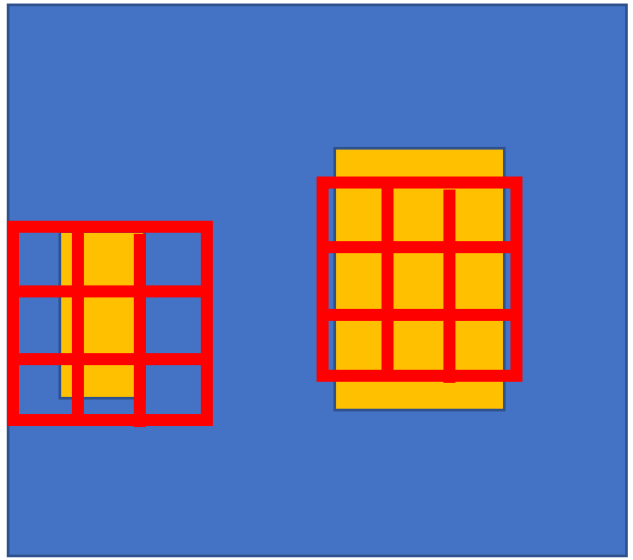


Speed and accuracy trade-offs for modern convolutional object detectors
Alireza Fathi, Anoop Korattikara, Chen Sun, Ian Fischer, Jonathan Huang, Kevin Murphy, Menglong Zhu, Sergio Guadarrama, Vivek Rathod, Yang Song, Zbigniew Wojna
CVPR 2017

RoI pooling (Faster R-CNN) vs convolution (SSD)

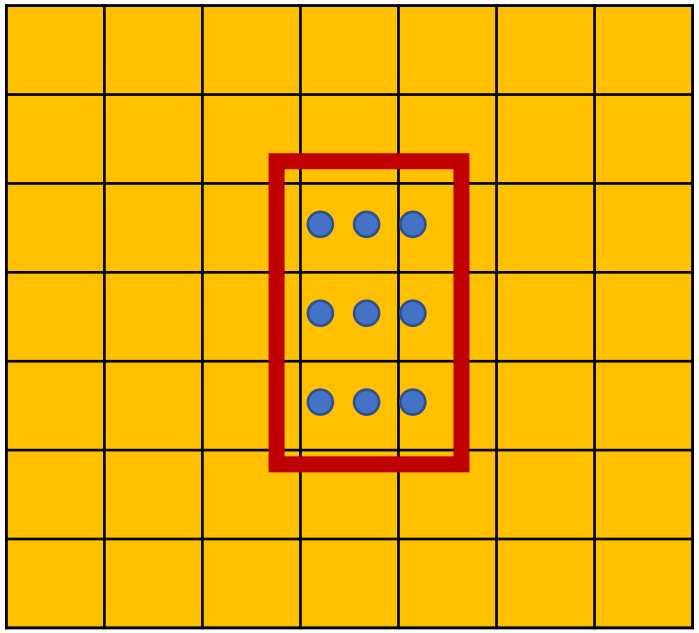
- Why does Faster R-CNN tend to be more accurate?
- RoI pooling takes information from the precise box
- Convolution takes information from just the kernel window





Other details - ROI Align

- Snapping box to grid introduces quantization artifacts
- Instead, use bilinear interpolation

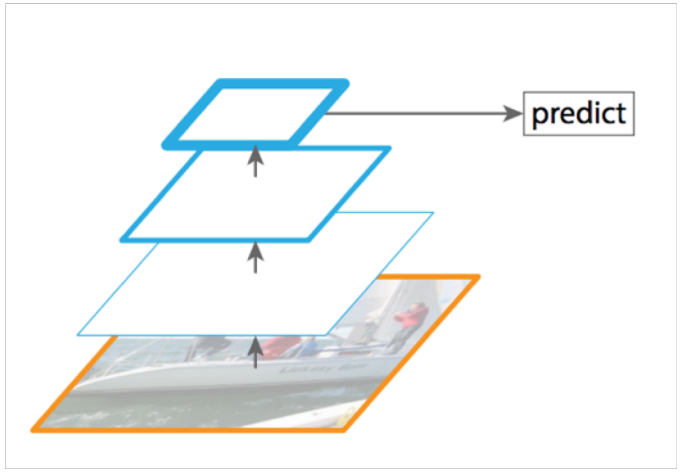


Mask R-CNN. K. He, G. Gkioxari, P. Dollar, R. Girshick. In *ICCV* 2017.

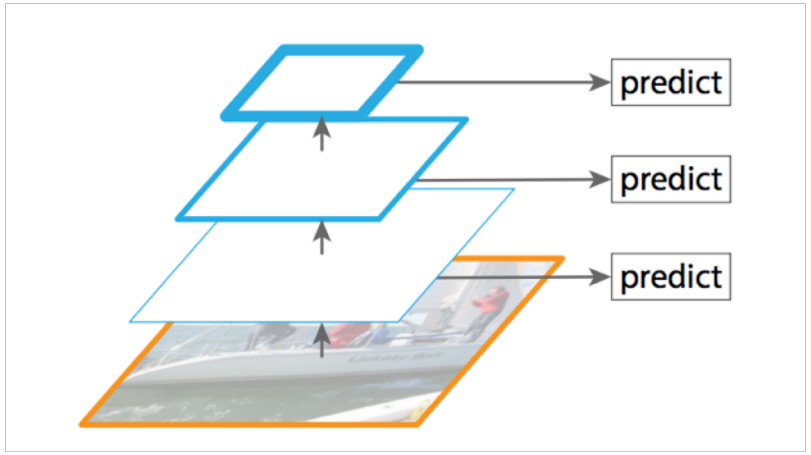
Detecting small objects

- Small objects get low resolution features

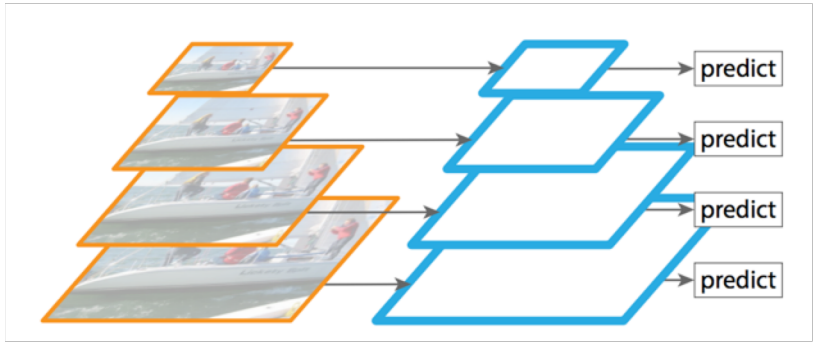
Feature pyramid networks



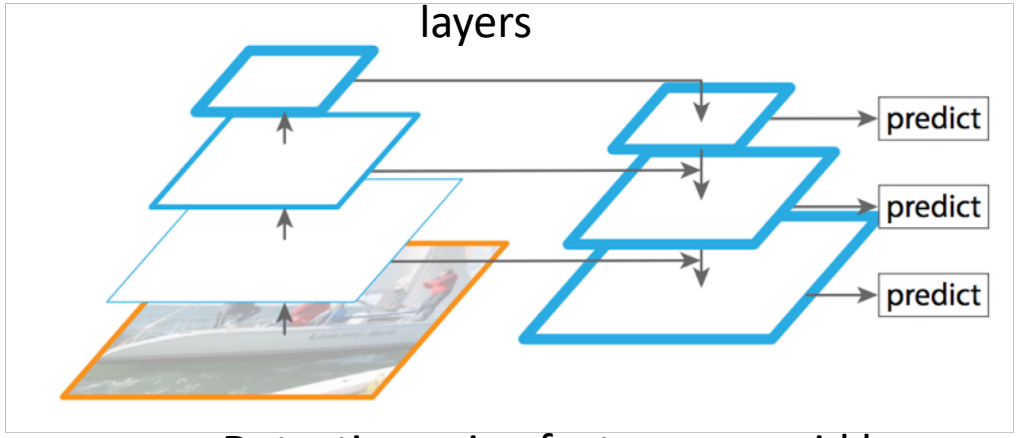
Standard detection



Detection using multiple layers



Detection on image pyramid



Detection using feature pyramid layers

Lin, Tsung-Yi, et al. "Feature Pyramid Networks for Object Detection." *CVPR*. Vol. 1. No. 2. 2017.

Feature pyramid networks

Faster R-CNN	proposals	feature	head	lateral?	top-down?	AP@0.5	AP	AP_s	AP_m	AP_l
(*) baseline from He *et al.* [16]†	RPN, C_4	C_4	conv5			47.3	26.3	-	-	-
(a) baseline on conv4	RPN, C_4	C_4	conv5			53.1	31.6	13.2	35.6	**47.1**
(b) baseline on conv5	RPN, C_5	C_5	2*fc*			51.7	28.0	9.6	31.9	43.1
(c) **FPN**	RPN, $\{P_k\}$	$\{P_k\}$	2*fc*	✓	✓	**56.9**	**33.9**	**17.8**	**37.7**	45.8

Lin, Tsung-Yi, et al. "Feature Pyramid Networks for Object Detection." *CVPR*. Vol. 1. No. 2. 2017.

Deformable conv

- Convolution uses the same kernel size always
- Want to capture more or less object region depending on scale
  - Or more generally properties of pixel
- Idea: *learn what pixels to combine using convolution*

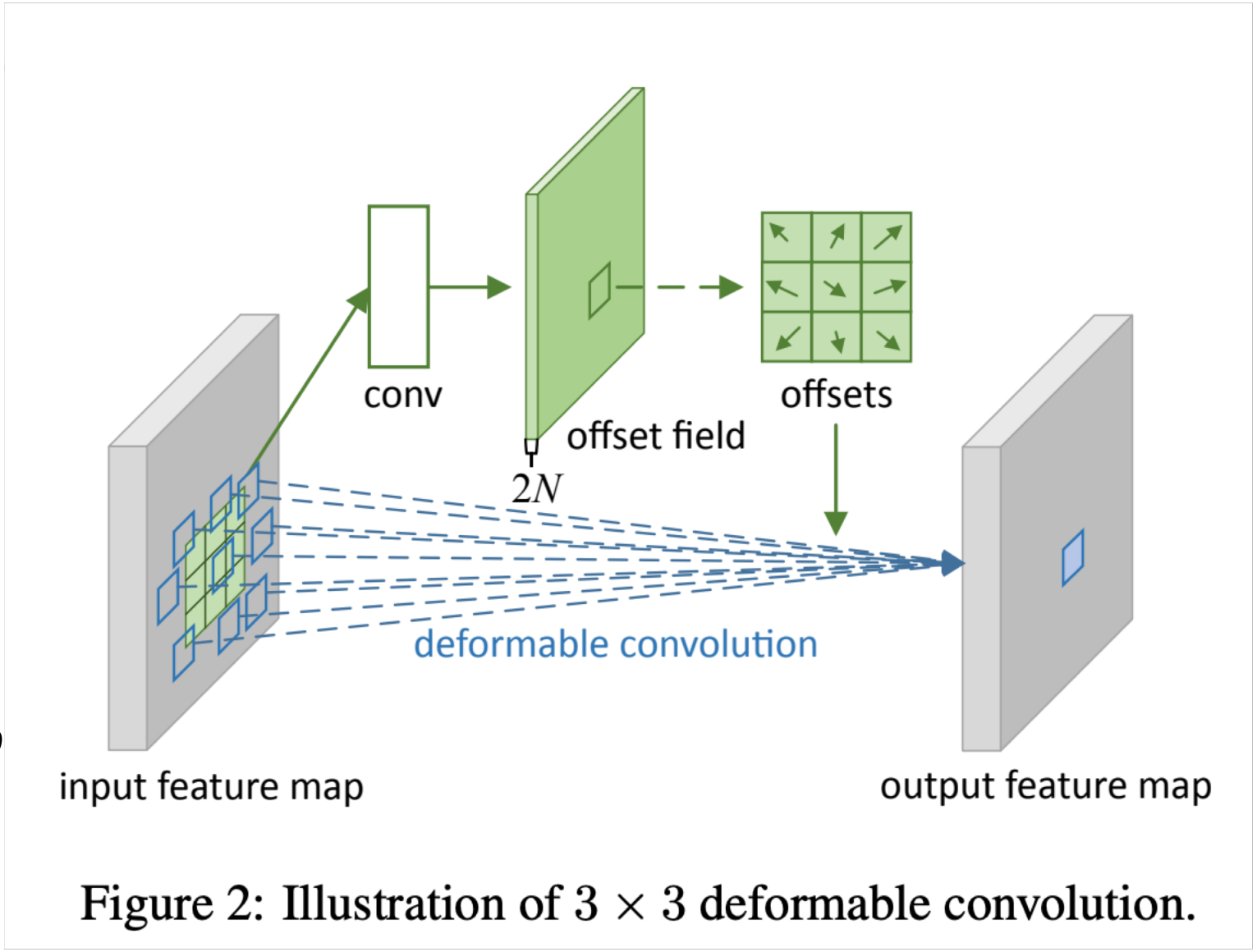


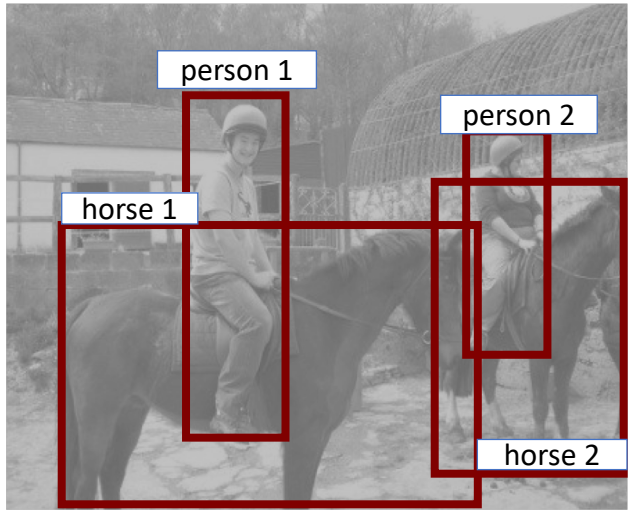
Figure 2: Illustration of 3×3 deformable convolution.

Instance segmentation

Till now

horse, person

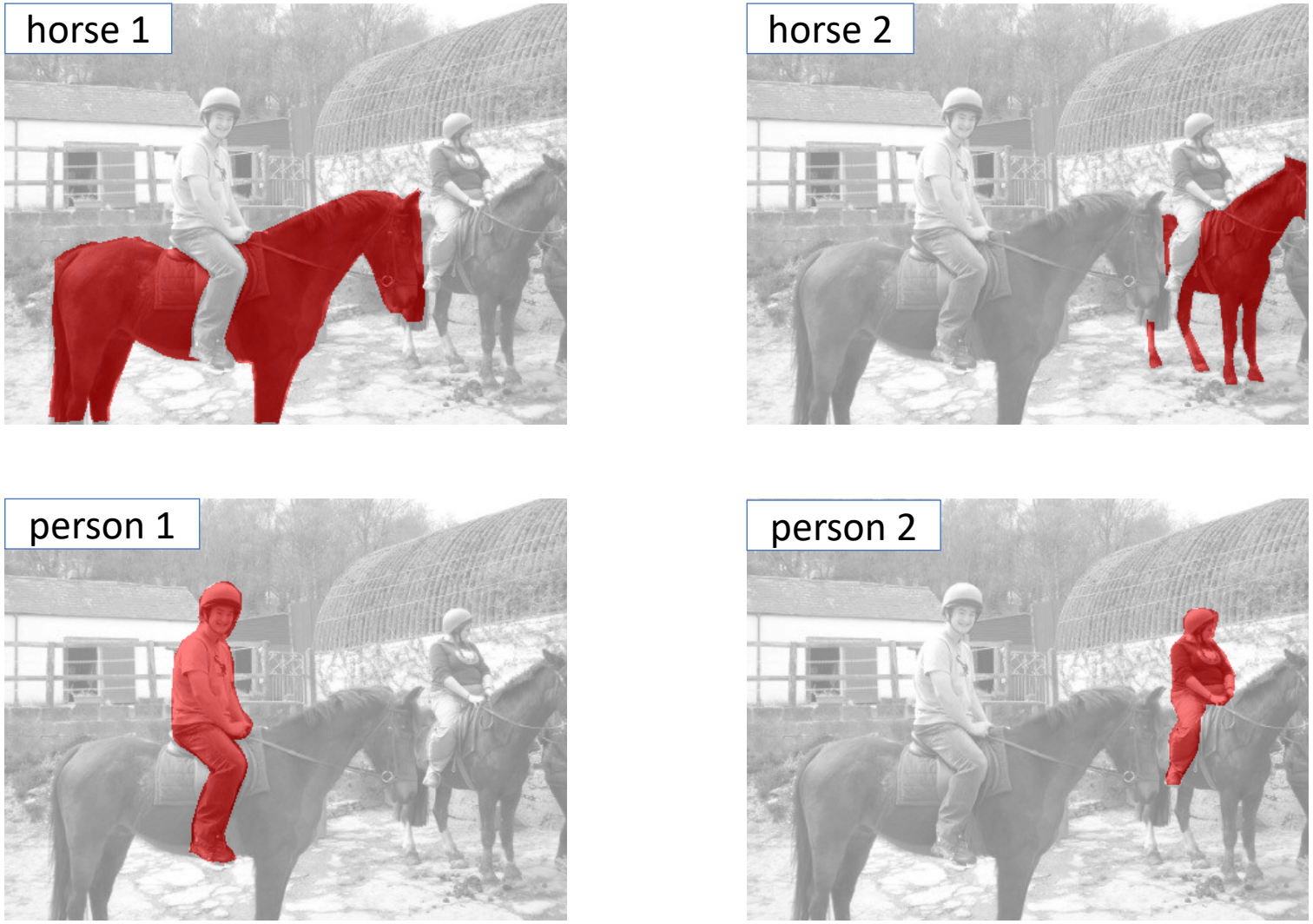
Image Classification



Object Detection

Semantic Segmentation

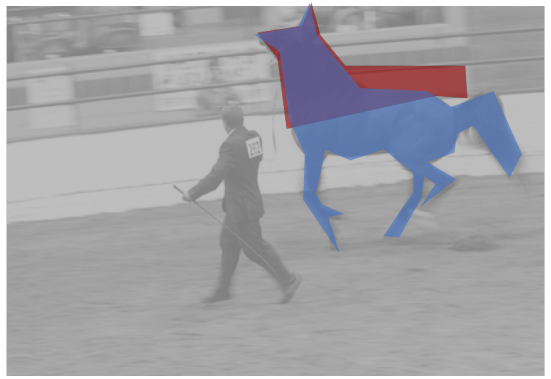
Fine-grained Localization



Instance Segmentation

Evaluation Protocol

- Sort predicted instances by confidence
- Match prediction to closest annotation based on *segment overlap*
  - If segment overlap > threshold, correct

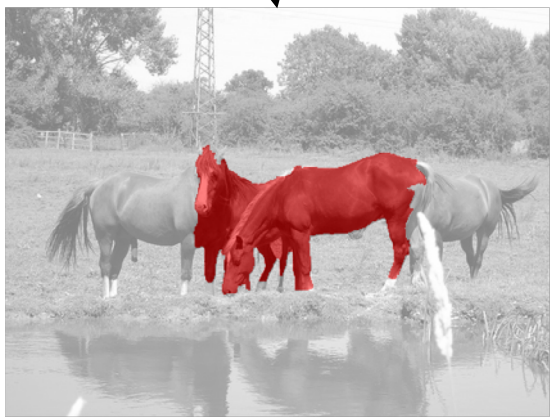


segment overlap = $\frac{\square \cap \square}{\square \cup \square}$

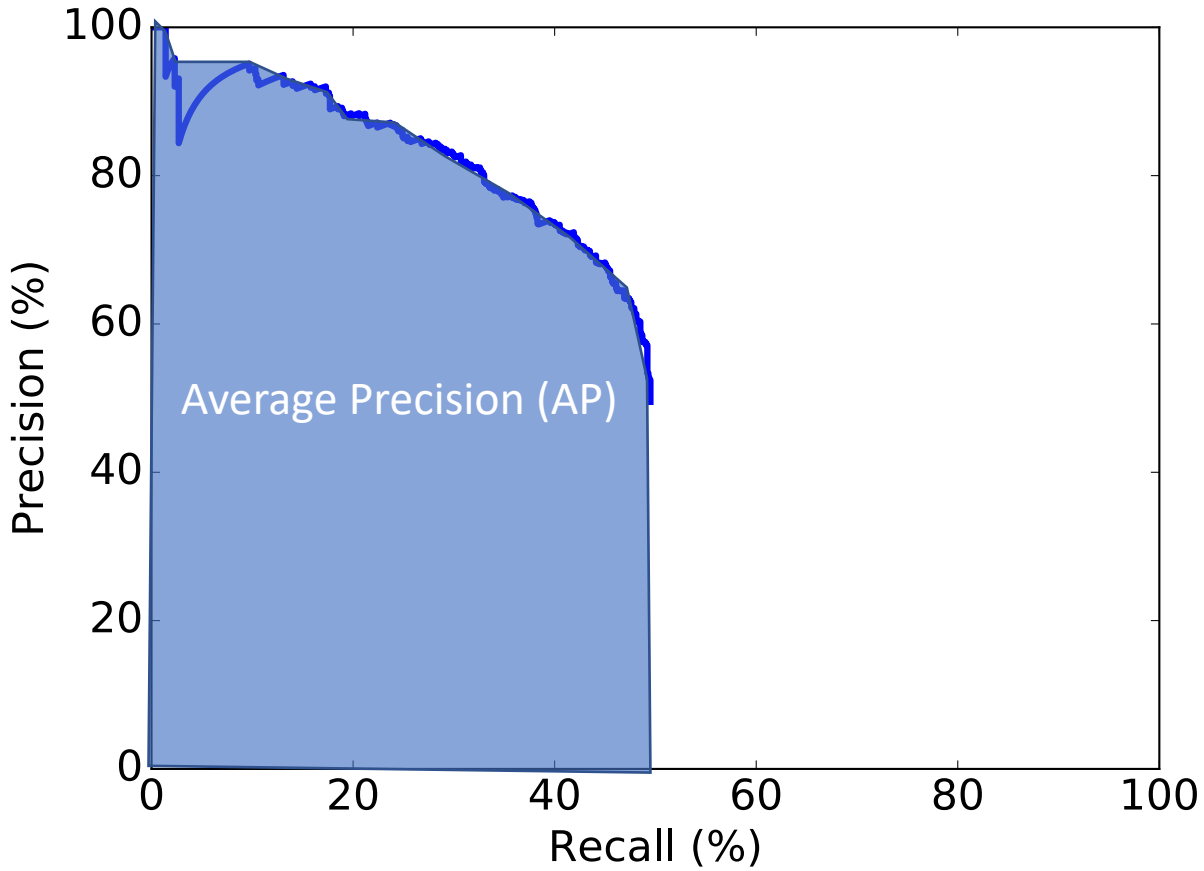
Evaluation Protocol

Labels = [✓ ✓ ✗ ✓ ✗ ✗ ✓ ....]

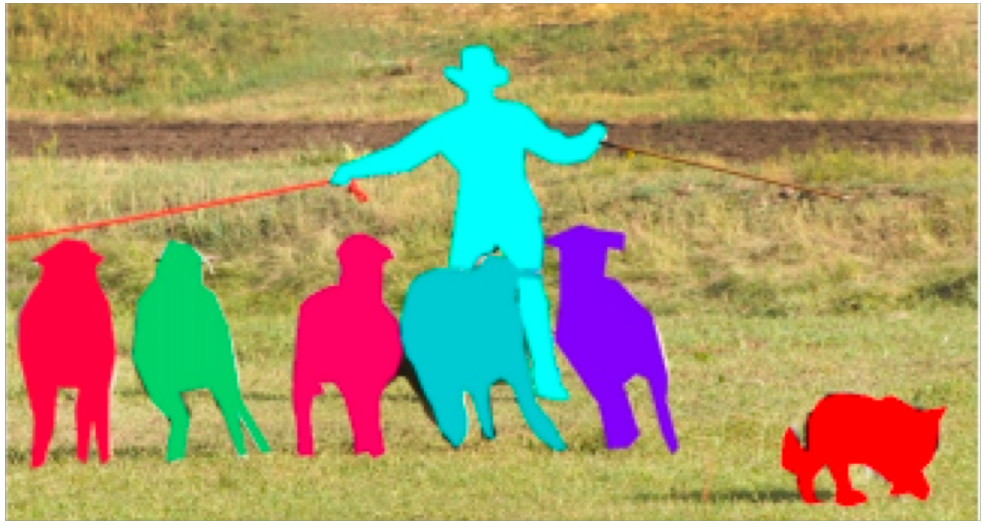
Scores = [0.90 0.87 0.82 0.78 0.70 0.69 0.60 ....]



Evaluation protocol



The COCO Challenge



mscoco.org

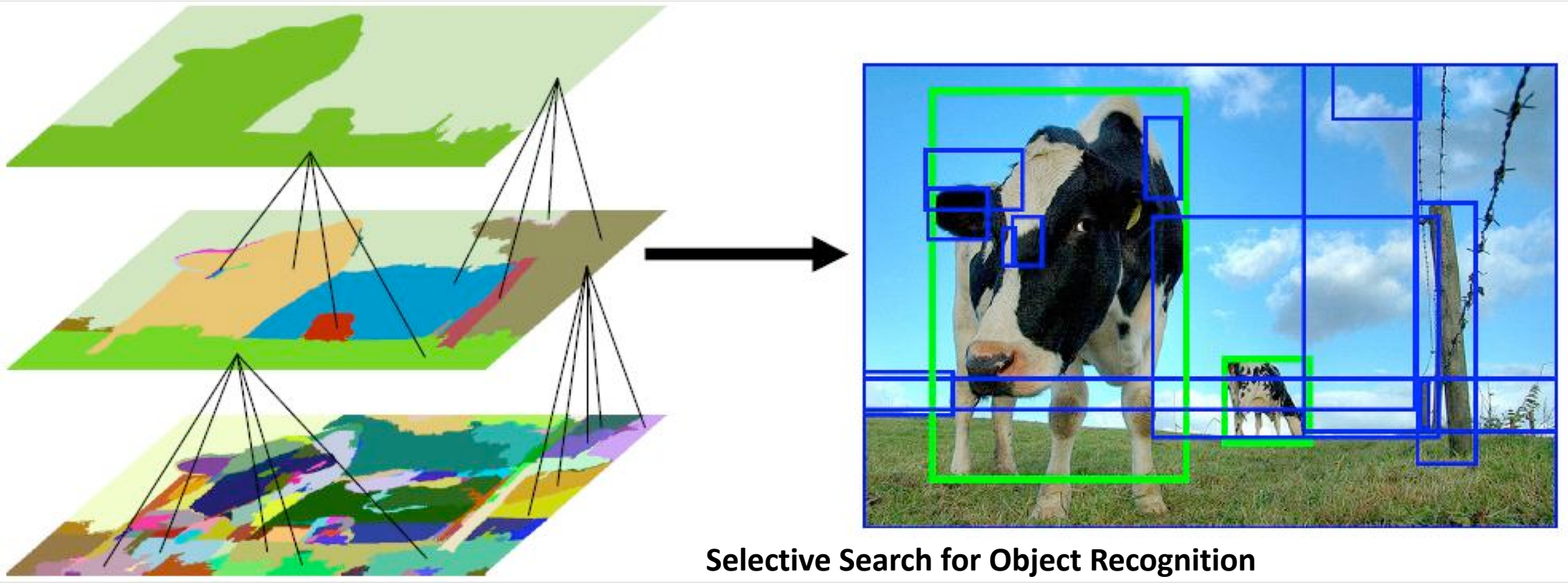
T. Y. Lin et al. **Microsoft COCO: Common Objects in Context**. In ECCV, 2014

Two strategies

- Segment then classify
  - Use bottom-up techniques to come up with *segment* proposals
  - Classify segment proposals with convnets
  - Segmentation is category agnostic
  - Modification: use convnets to produce segmentation proposals
- Detect then segment
  - Use standard object detection to produce boxes
  - Segment boxes
  - Segmentation is *category specific*

Box proposals

- Use segmentation to produce ~5K candidates

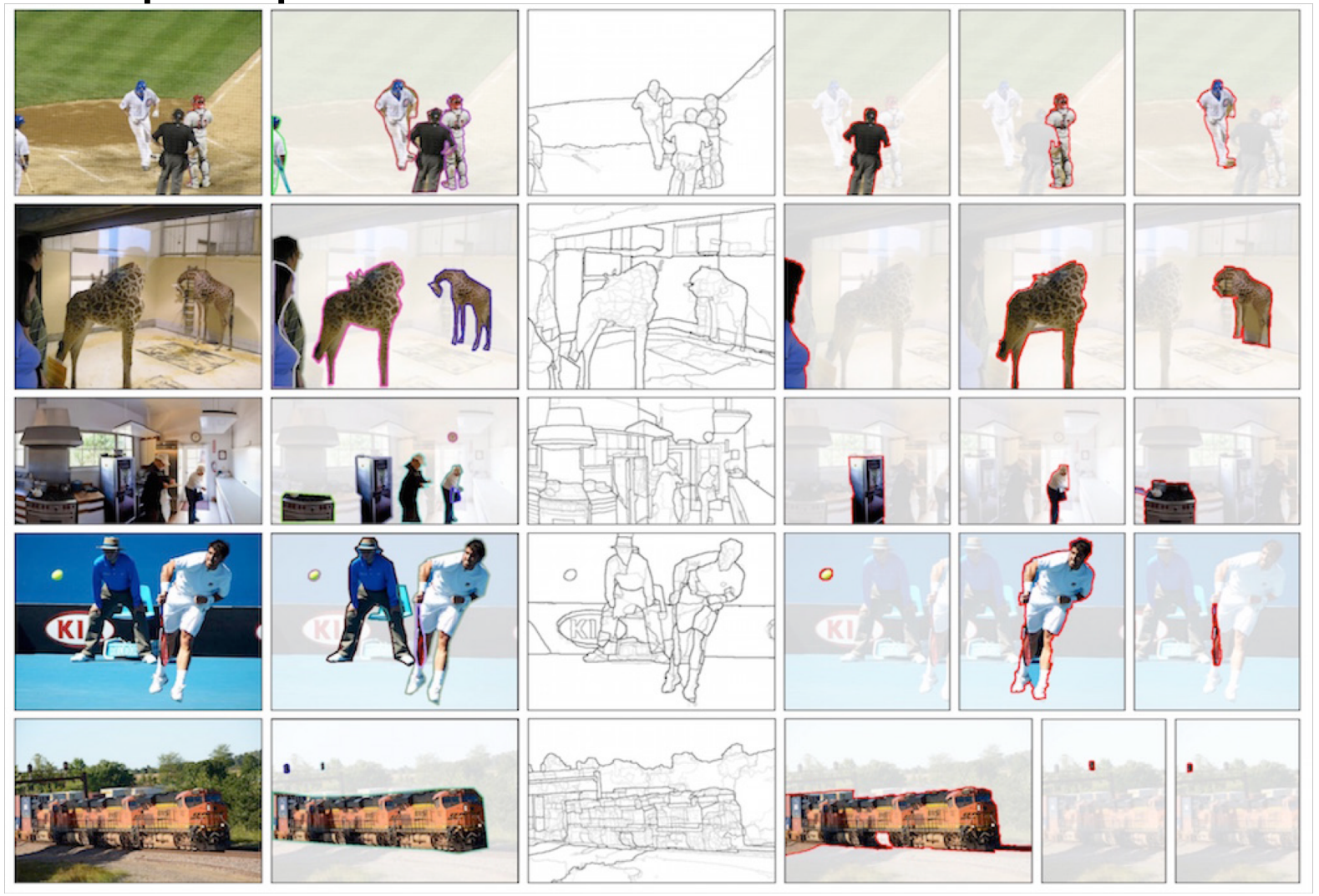


Selective Search for Object Recognition

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders

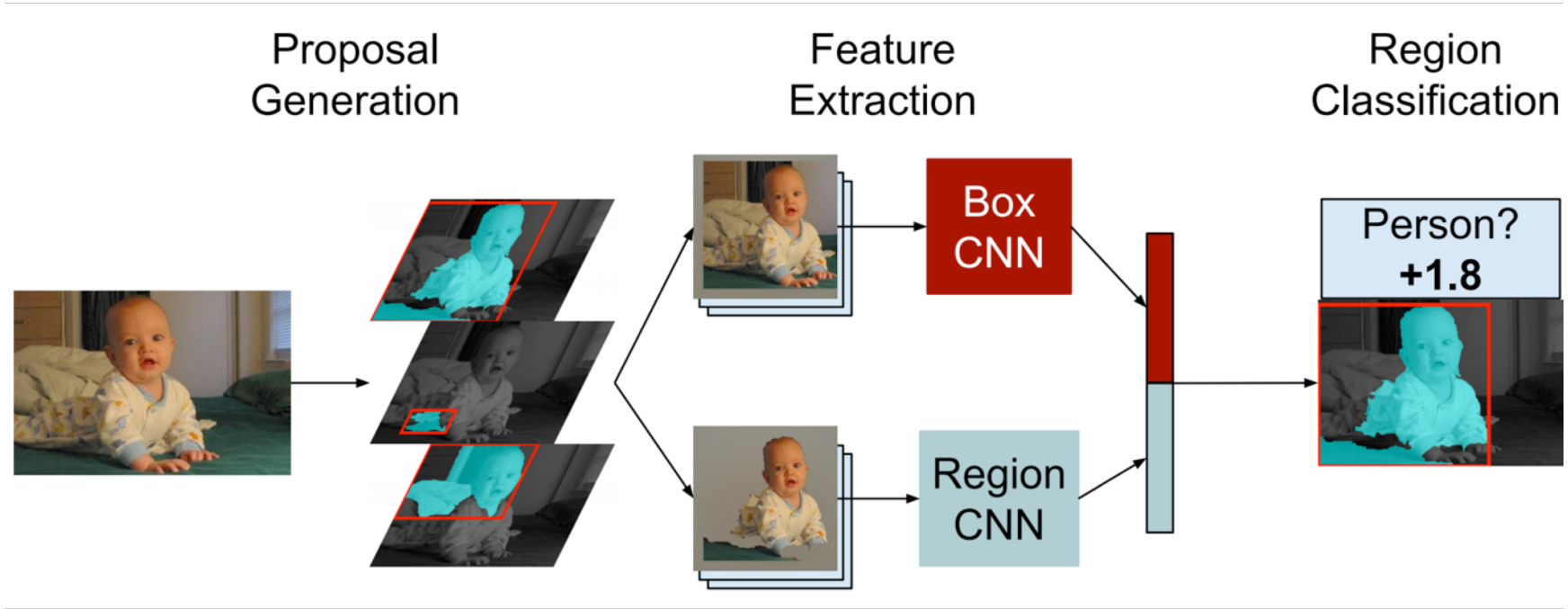
In International Journal of Computer Vision 2013.

Segment proposals



Multi-scale combinatorial grouping. Pablo Arbelaez, Jordi Pont-Tuset, Jonathan Barron, Ferran Marques, Jitendra Malik. In *CVPR,* 2014.

R-CNN for instance segmentation



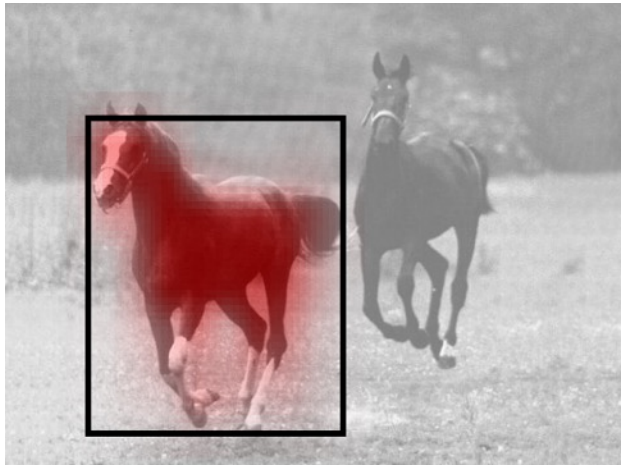
Simultaneous Detection and Segmentation Bharath Hariharan , Pablo Arbelaez , Ross Girshick, Jitendra Malik. In *ECCV* 2014

Two strategies

- Segment then classify
  - Use bottom-up techniques to come up with *segment* proposals
  - Classify segment proposals with convnets
  - Segmentation is category agnostic
  - Modification: use convnets to produce segmentation proposals
- Detect then segment
  - Use standard object detection to produce boxes
  - Segment boxes
  - Segmentation is *category specific*

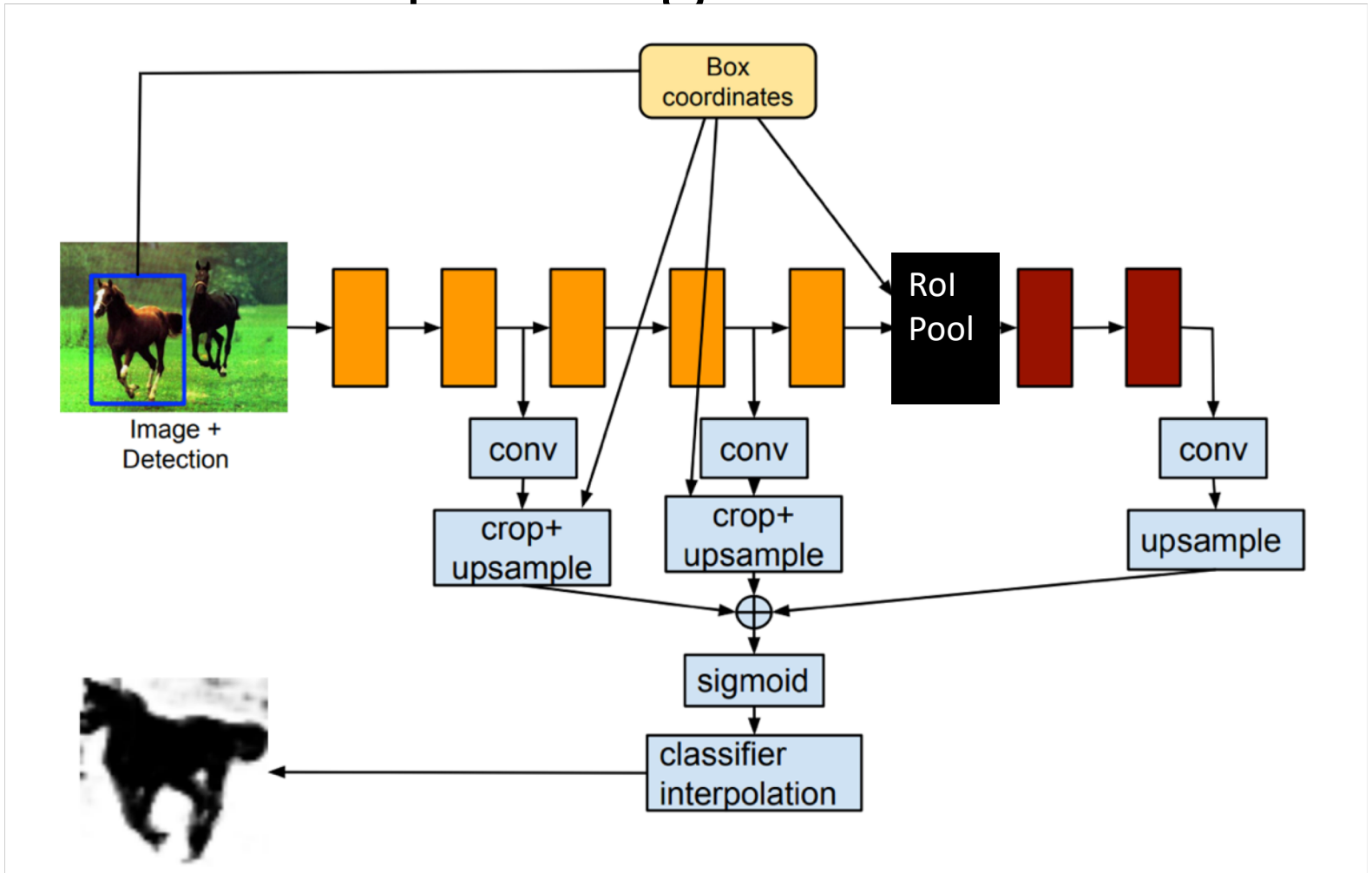
Detect then segment

- How should we segment a detected object?
- We have already computed features using ROI Pooling
- Idea: use features to predict mask!
  - Can either use a simple linear layer
  - Or can use convolution
  - Issue: can be very coarse



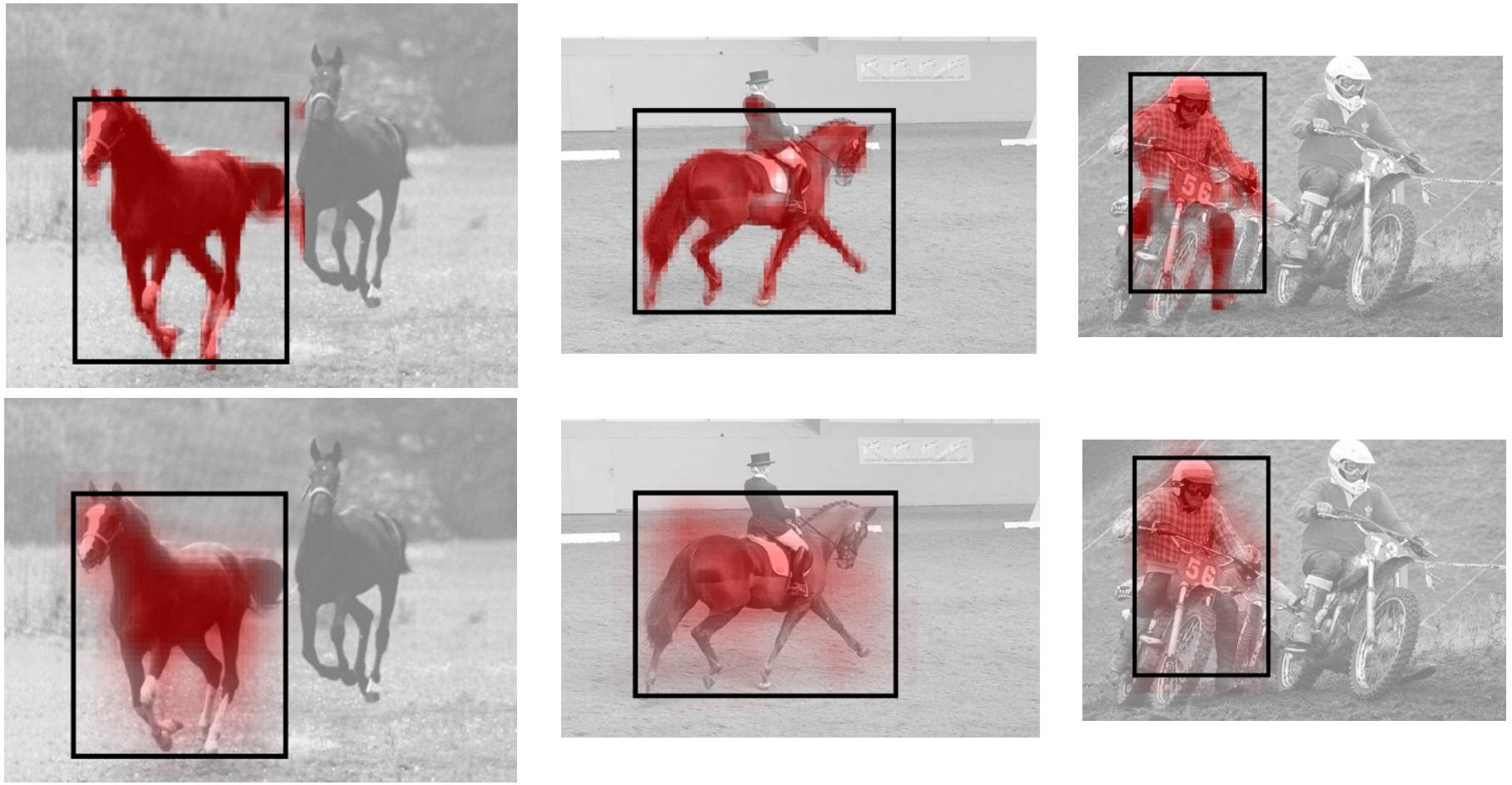
Skip connections with RoI pooling

- Finer-grained segmentation: tap into earlier layers



Hariharan, Bharath, et al. "Object instance segmentation and fine-grained localization using hypercolumns." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 4 (2017): 627-639.

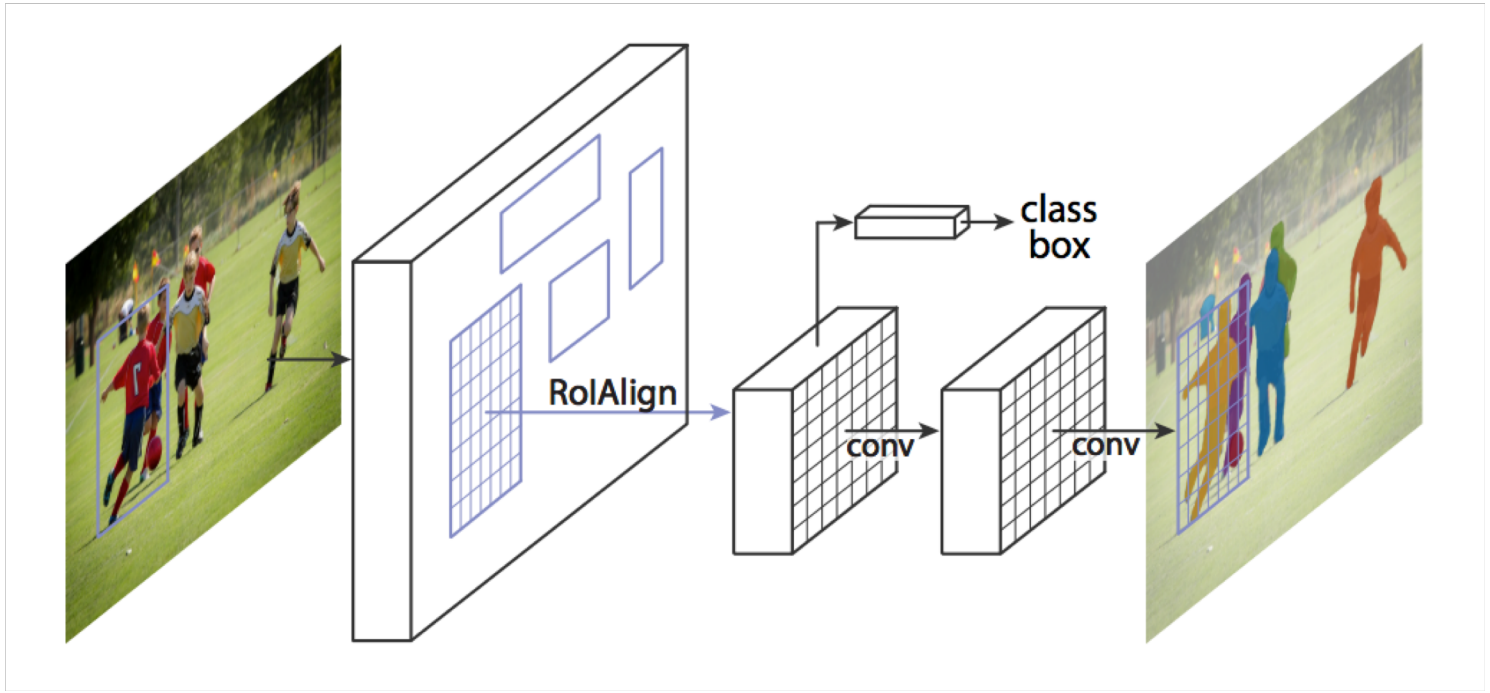
Skip connections for finer-grained details



Hypercolumns for Object Segmentation and Fine-grained Localization. Bharath Hariharan, Pablo Arbeláez, Ross Girshick, Jitendra Malik. CVPR, 2015

Mask R-CNN

- With deeper networks and ROI Align, skip connections not needed (?)



He, Kaiming, et al. "Mask r-cnn." *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017.

Final results - what works?

- First detect, then segment
- Big problem for instance segmentation is object detection
- Mask R-CNN (Faster R-CNN + convolution on RoI Pooled feature to get masks) is good starting point