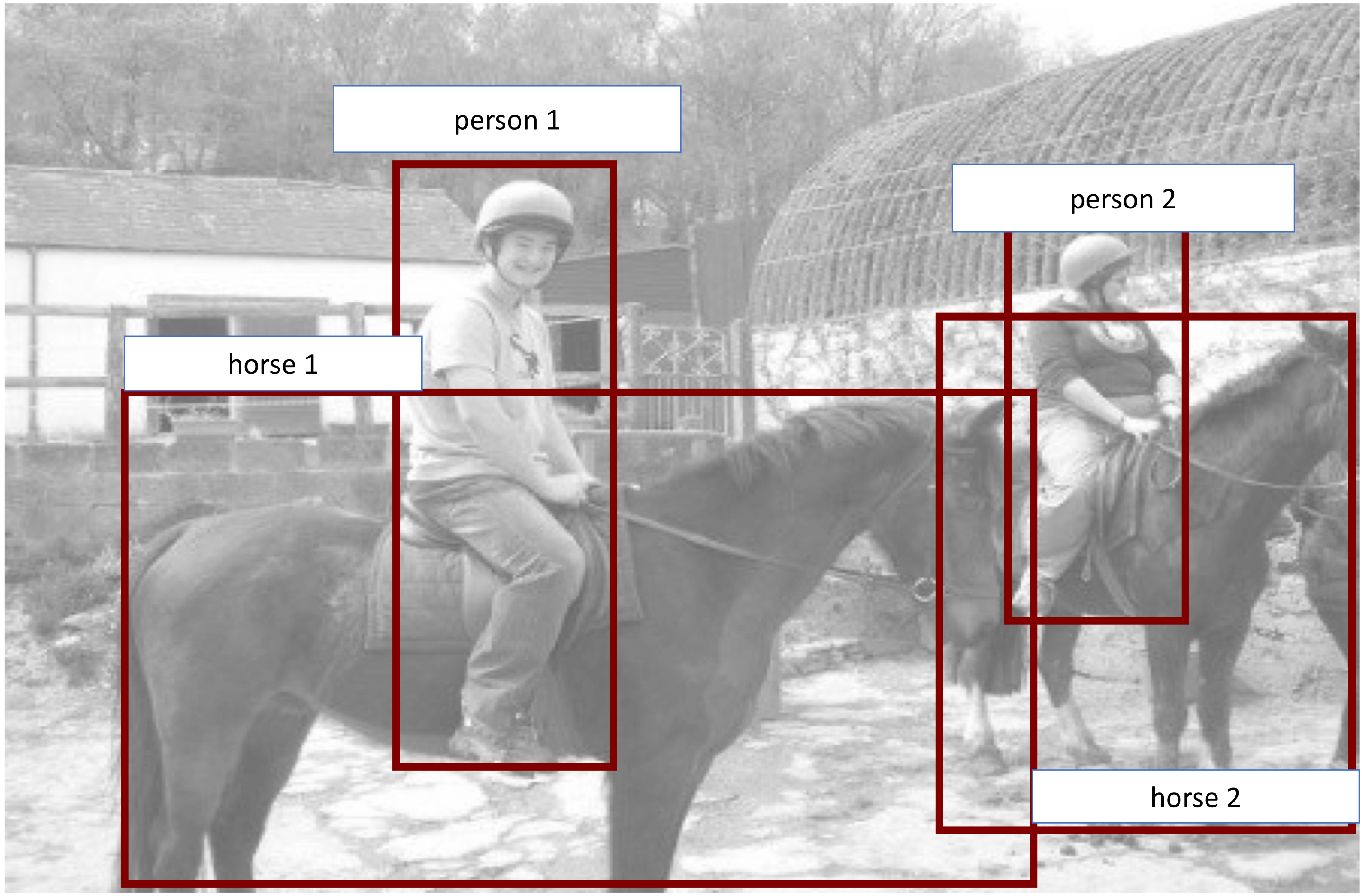
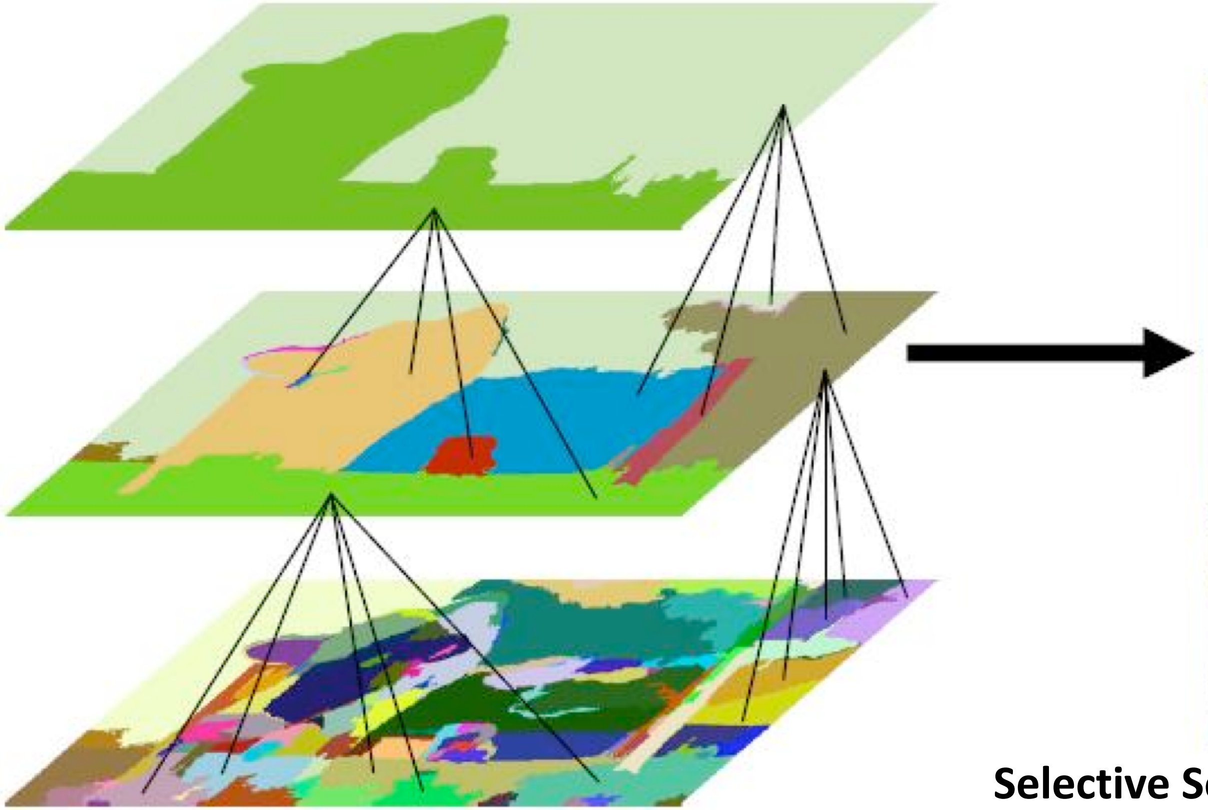
Object detection

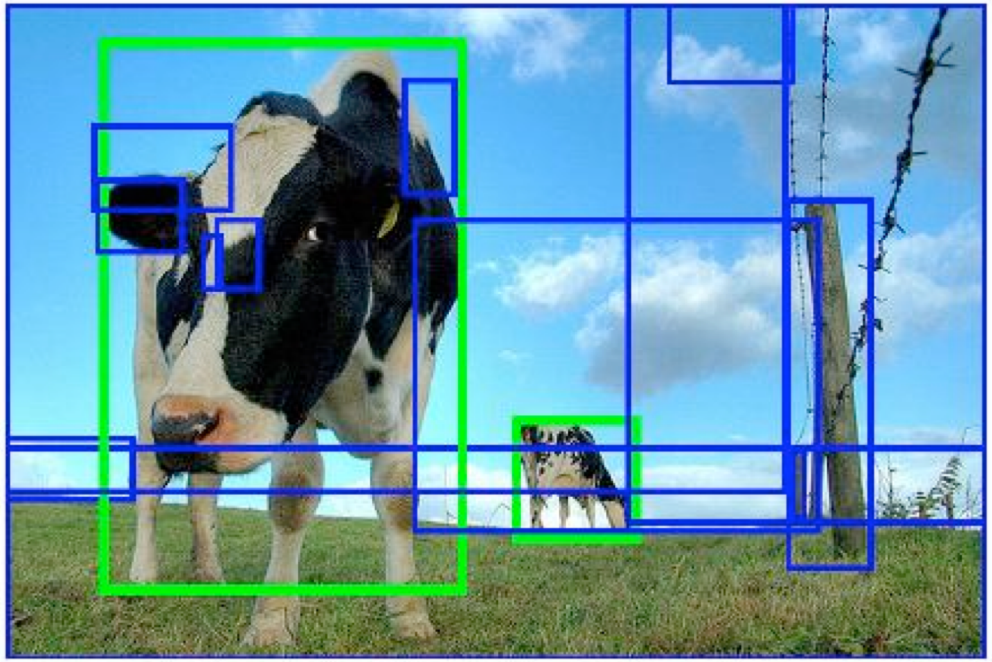
The Task



Step 1: Object proposals

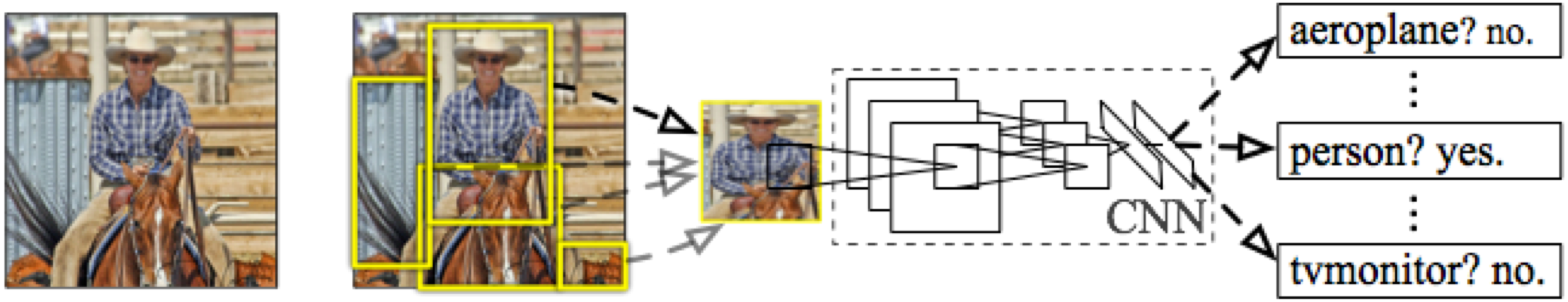
- Use segmentation to produce ~5K candidates





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.

R-CNN: Regions with CNN features



Input image

Extract region proposals (~2k / image)

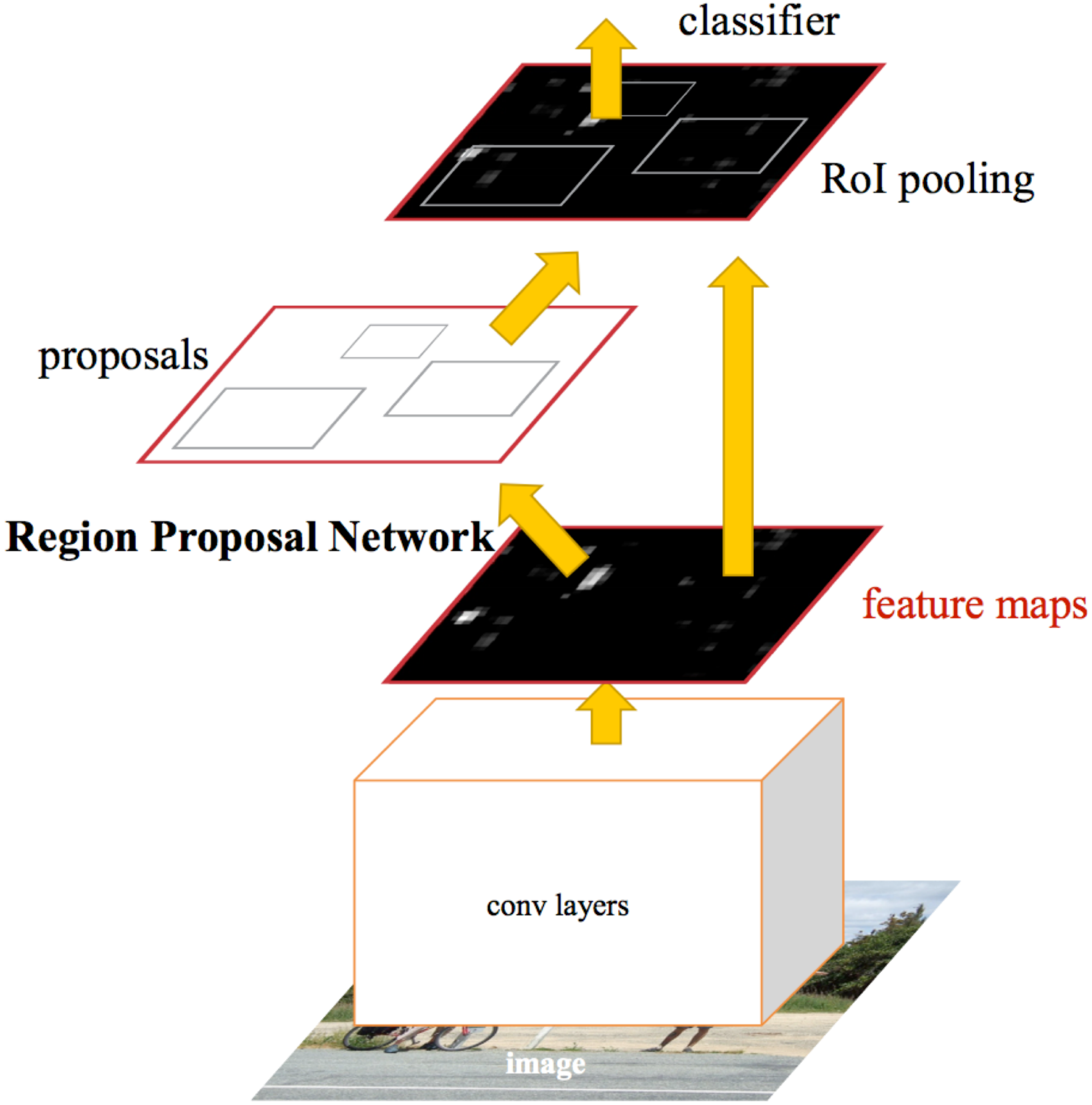
Compute CNN features

Classify regions (linear SVM)

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation
R. Girshick, J. Donahue, T. Darrell, J. Malik
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

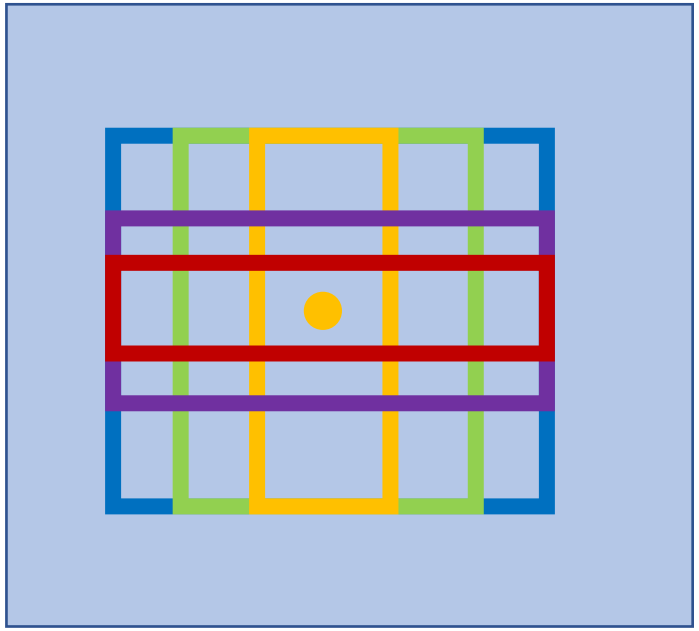
Slide credit : Ross Girshick

Faster R-CNN


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

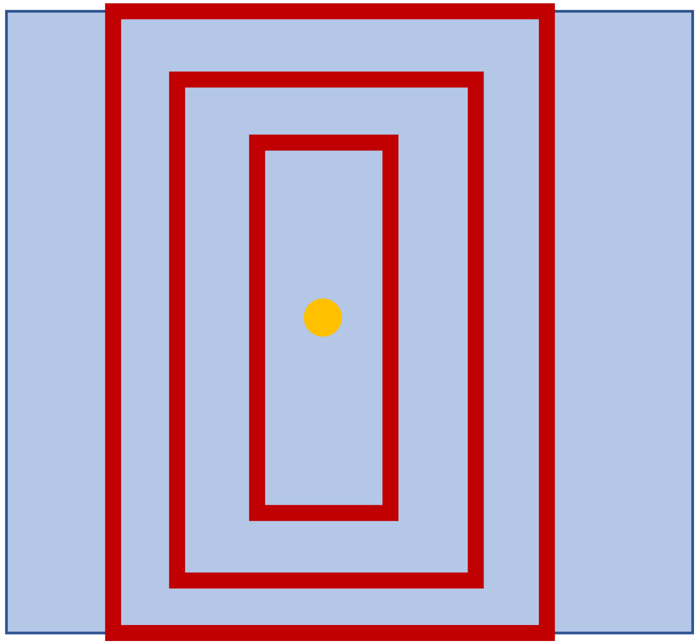
Faster R-CNN

- At each location, consider boxes of many different sizes and aspect ratios



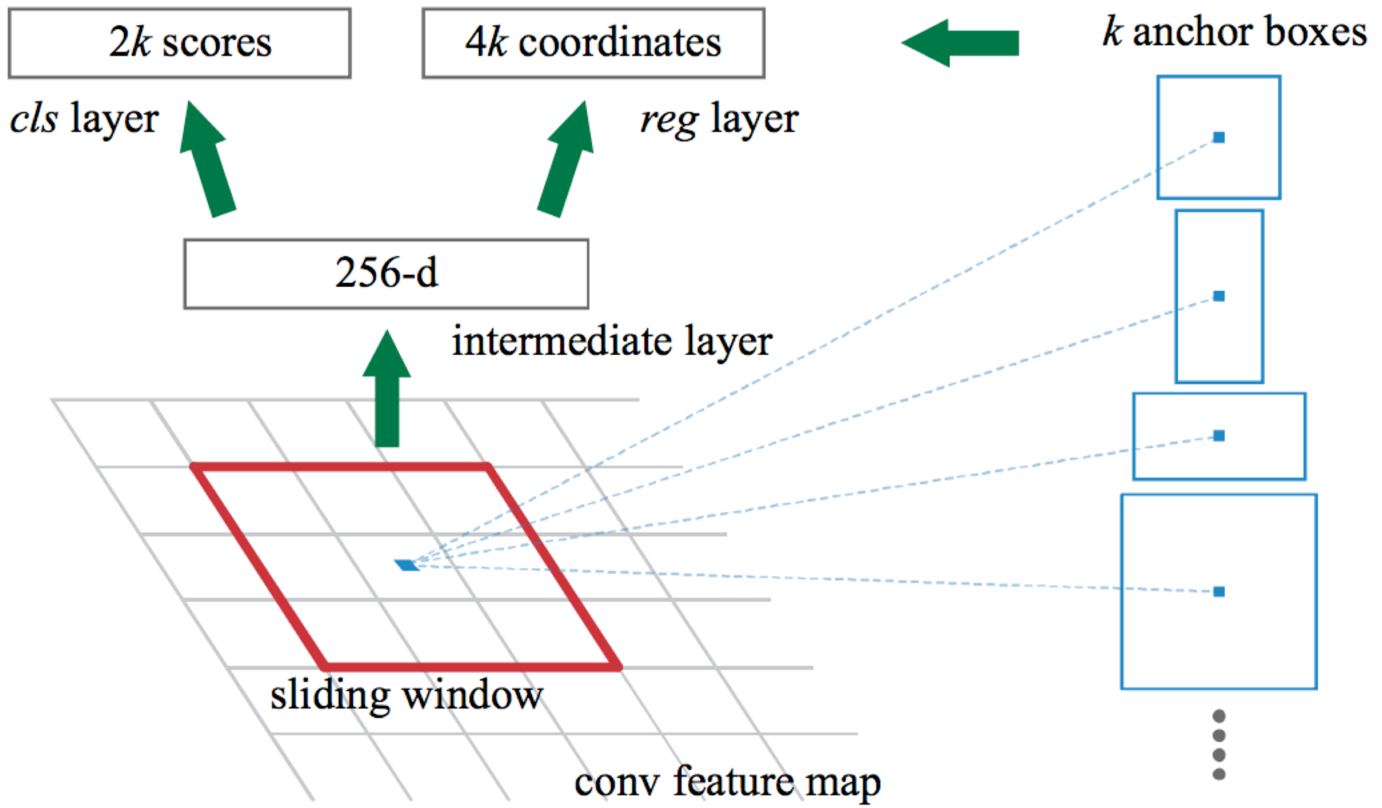
Faster R-CNN

- At each location, consider boxes of many different sizes and aspect ratios



Faster R-CNN

- At each location, consider boxes of many different sizes and aspect ratios



Faster R-CNN

- s scales * a aspect ratios = sa anchor boxes
- Use convolutional layer on top of filter map to produce sa scores
- Pick top few boxes as proposals

Faster R-CNN

Method	mean AP (PASCAL VOC)
Fast R-CNN	65.7
Faster R-CNN	67.0

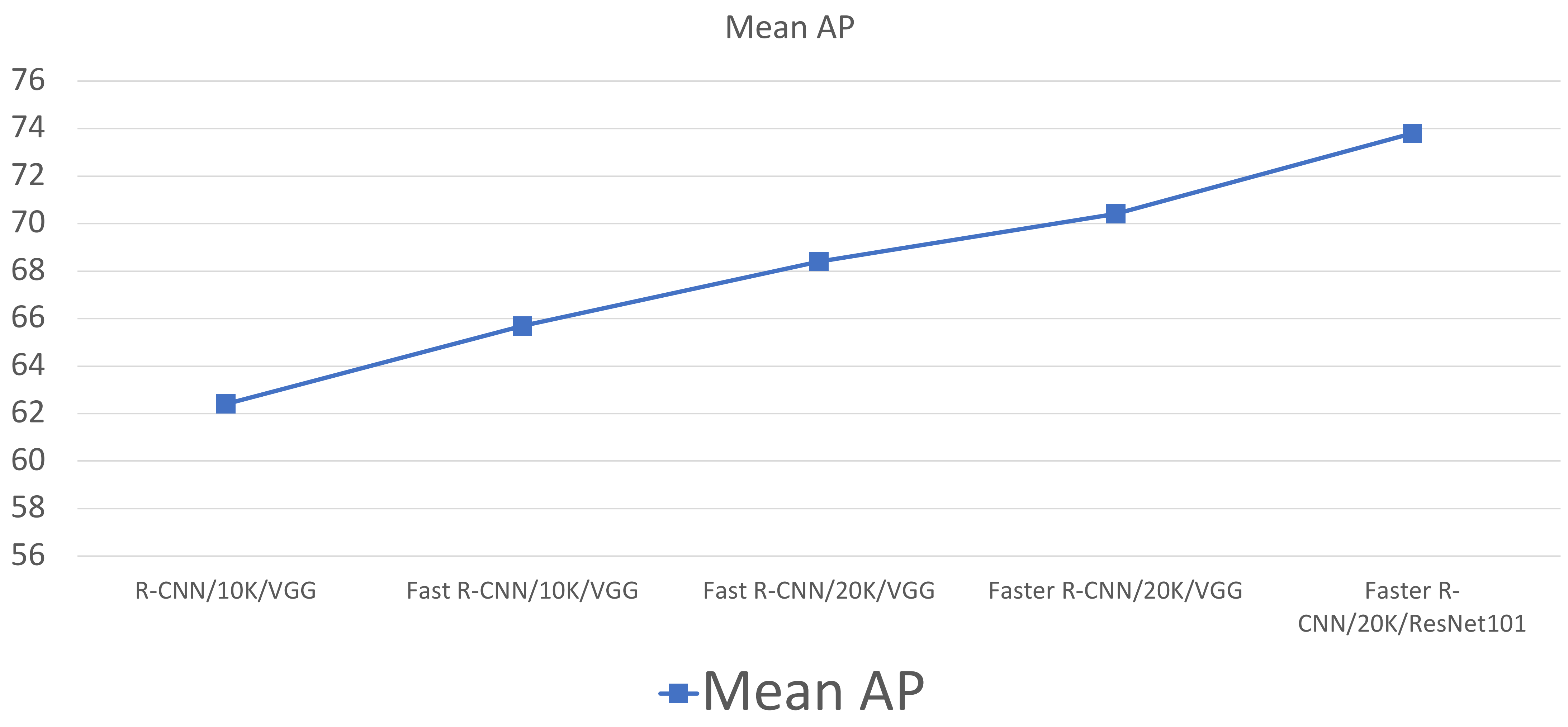
Impact of Feature Extractors

ConvNet	mean AP (PASCAL VOC)
VGG	70.4
ResNet 101	73.8

Impact of Additional Data

Method	Training data	mean AP (PASCAL VOC 2012 Test)
Fast R-CNN	VOC 12 Train (10K)	65.7
Fast R-CNN	VOC07 Trainval + VOC 12 Train	68.4
Faster R-CNN	VOC 12 Train (10K)	67.0
Faster R-CNN	VOC07 Trainval + VOC 12 Train	70.4

The R-CNN family of detectors



SSD (Single Shot Detector)

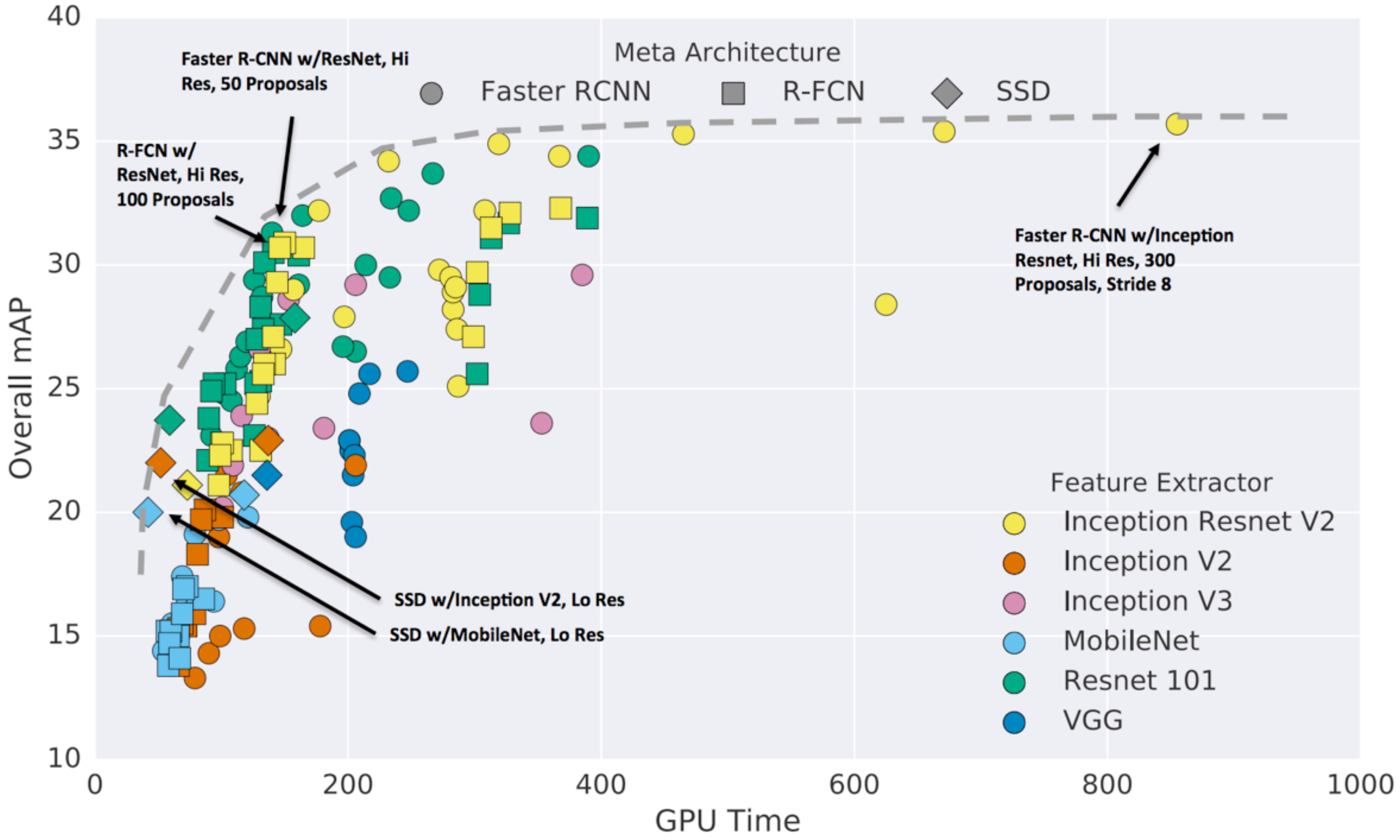
- Why go through separate proposals?
- Directly produce class-specific scores at each location for every scale and aspect ratio
  - s scales * a aspects * c classes = sac scores per location

SSD: Single Shot MultiBox Detector. Wei Liu , Dragomir Anguelov , Dumitru Erhan , Christian Szegedy , Scott Reed , Cheng-Yang Fu , Alexander C. Berg. In ECCV, 2016

SSD

Method	mean AP (PASCAL VOC)
Faster R-CNN	70.4
SSD	74.9

A comprehensive evaluation

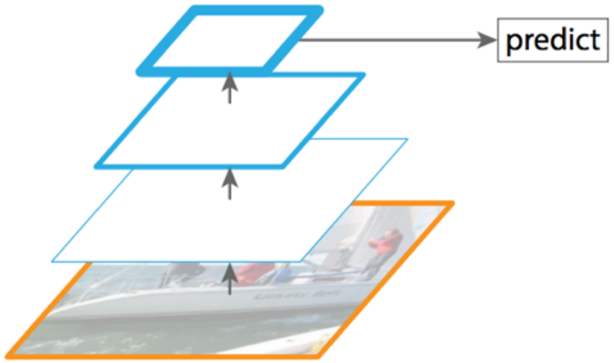


Speed and accuracy trade-offs for modern convolutional object detectors
Alireza Fathi, Anoop Korattikara, Chen Sun, Ian Fischer, Jonathan Huang, Kevin Murphy, Menglong Zhu, Sergio Guadarrama, Vivek Rathod, Yang Song, Zbigniew Wojna
CVPR 2017

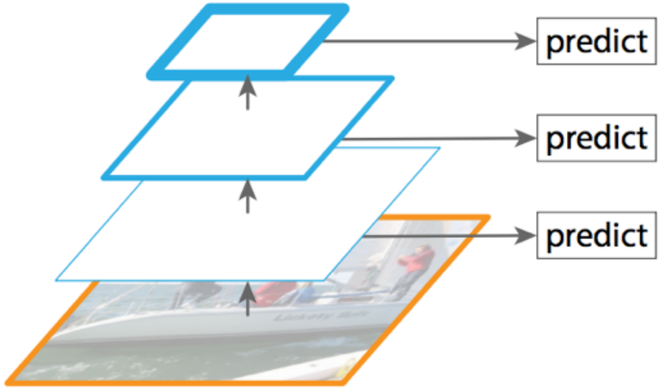
Detecting small objects

- Small objects get low resolution features

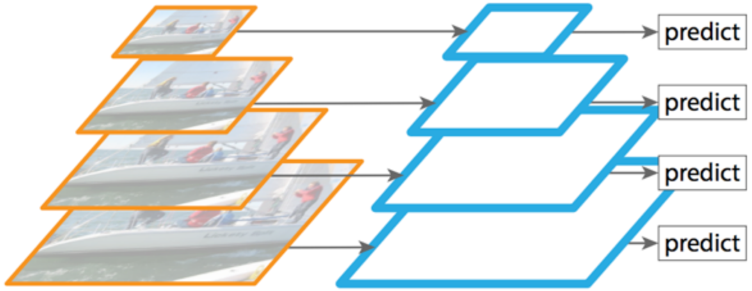
Feature pyramid networks



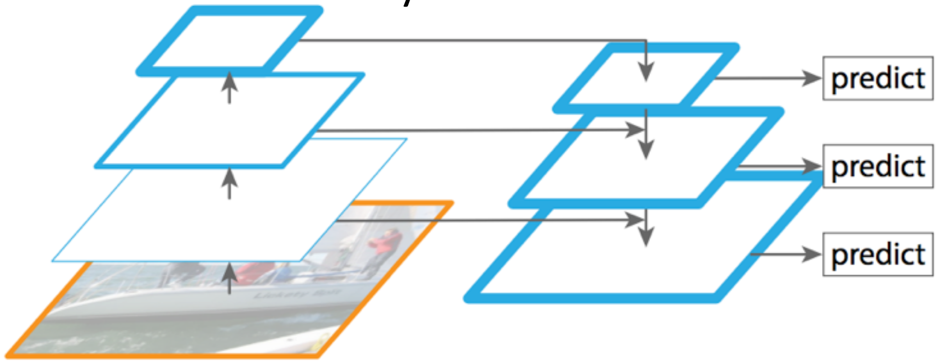
Standard detection



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	*2fc*			51.7	28.0	9.6	31.9	43.1
(c) **FPN**	RPN, $\{P_k\}$	$\{P_k\}$	*2fc*	✓	✓	**56.9**	**33.9**	**17.8**	**37.7**	45.8

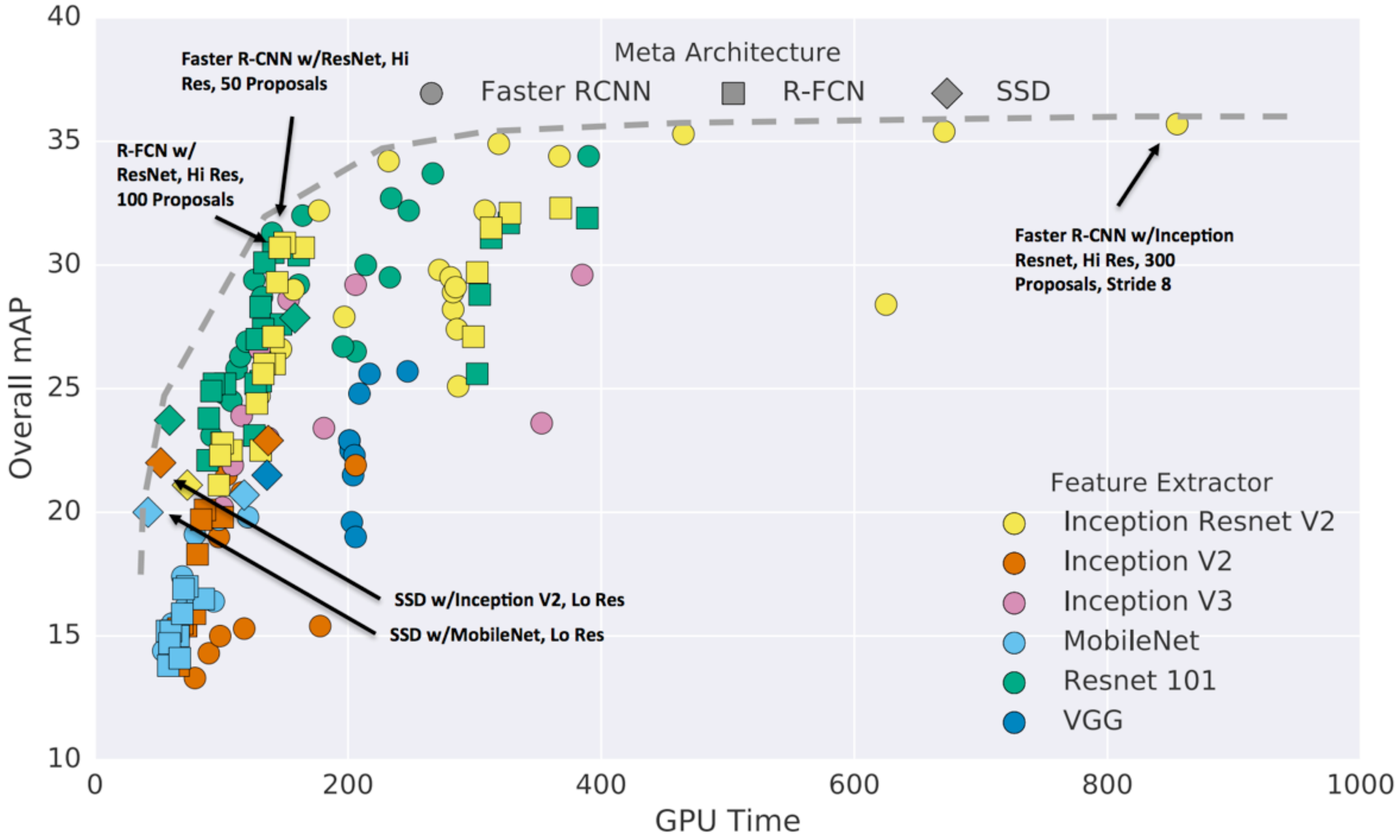
Lin, Tsung-Yi, et al. "Feature Pyramid Networks for Object Detection." *CVPR*. Vol. 1. No. 2. 2017.

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

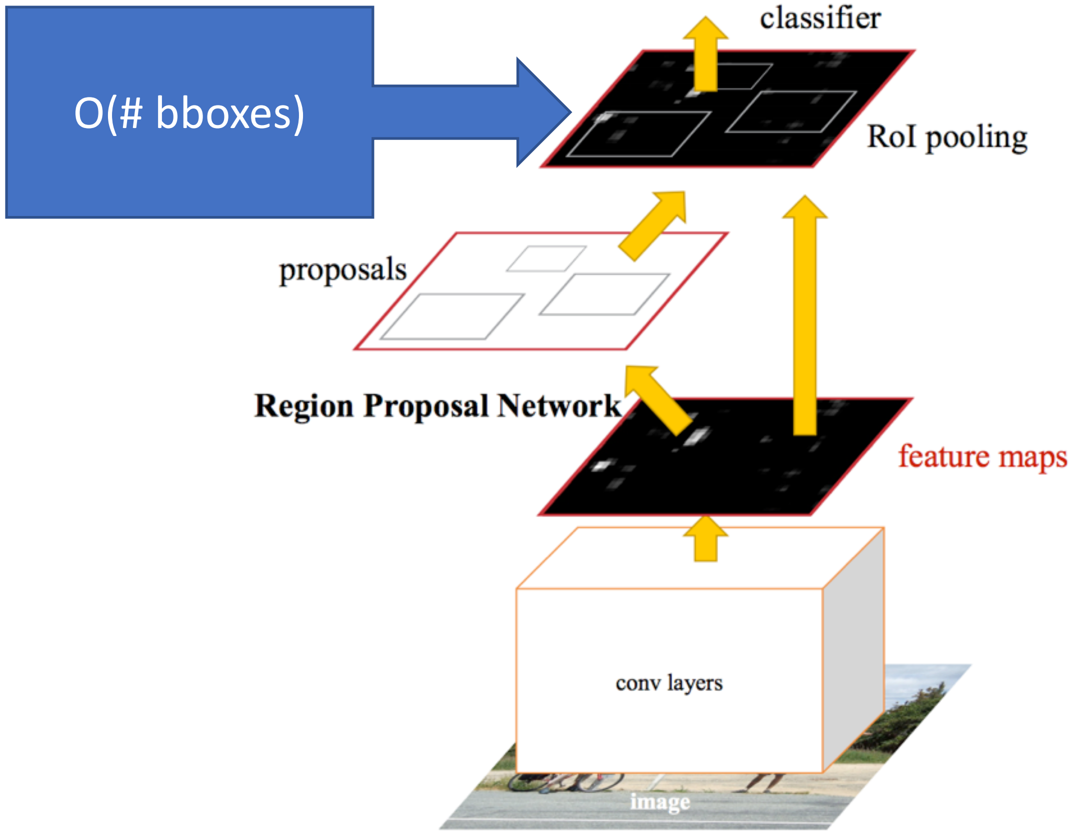
A comprehensive evaluation



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

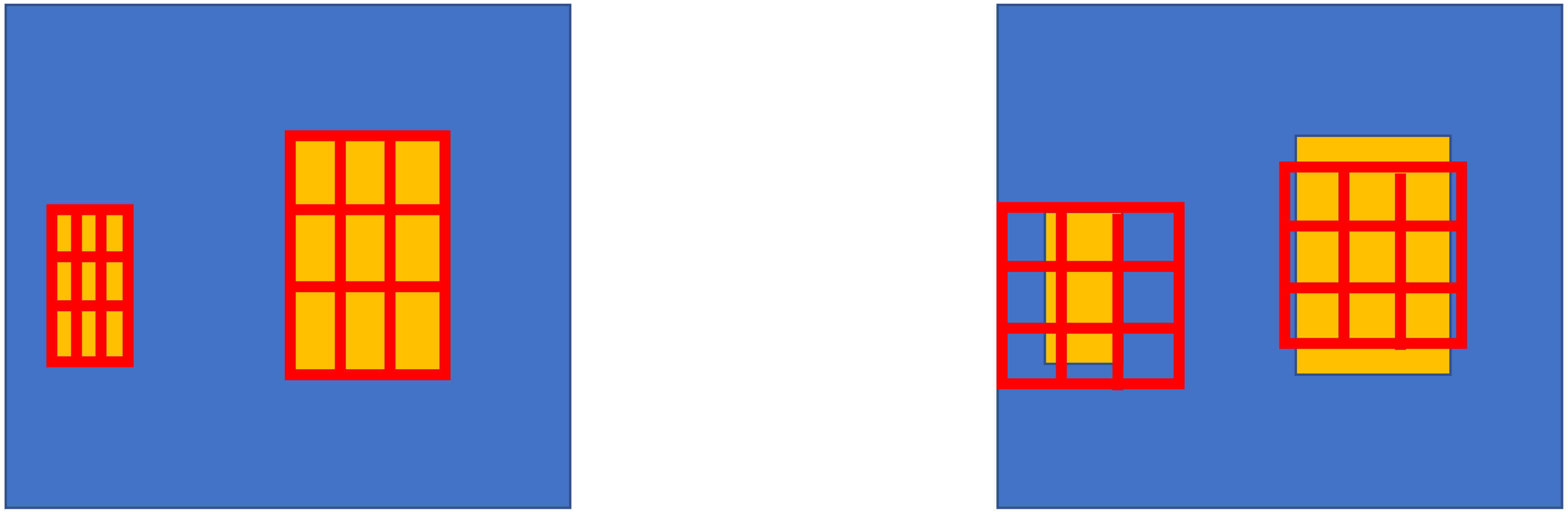
Challenges with RoI pooling

- Still scales with the number of bounding boxes



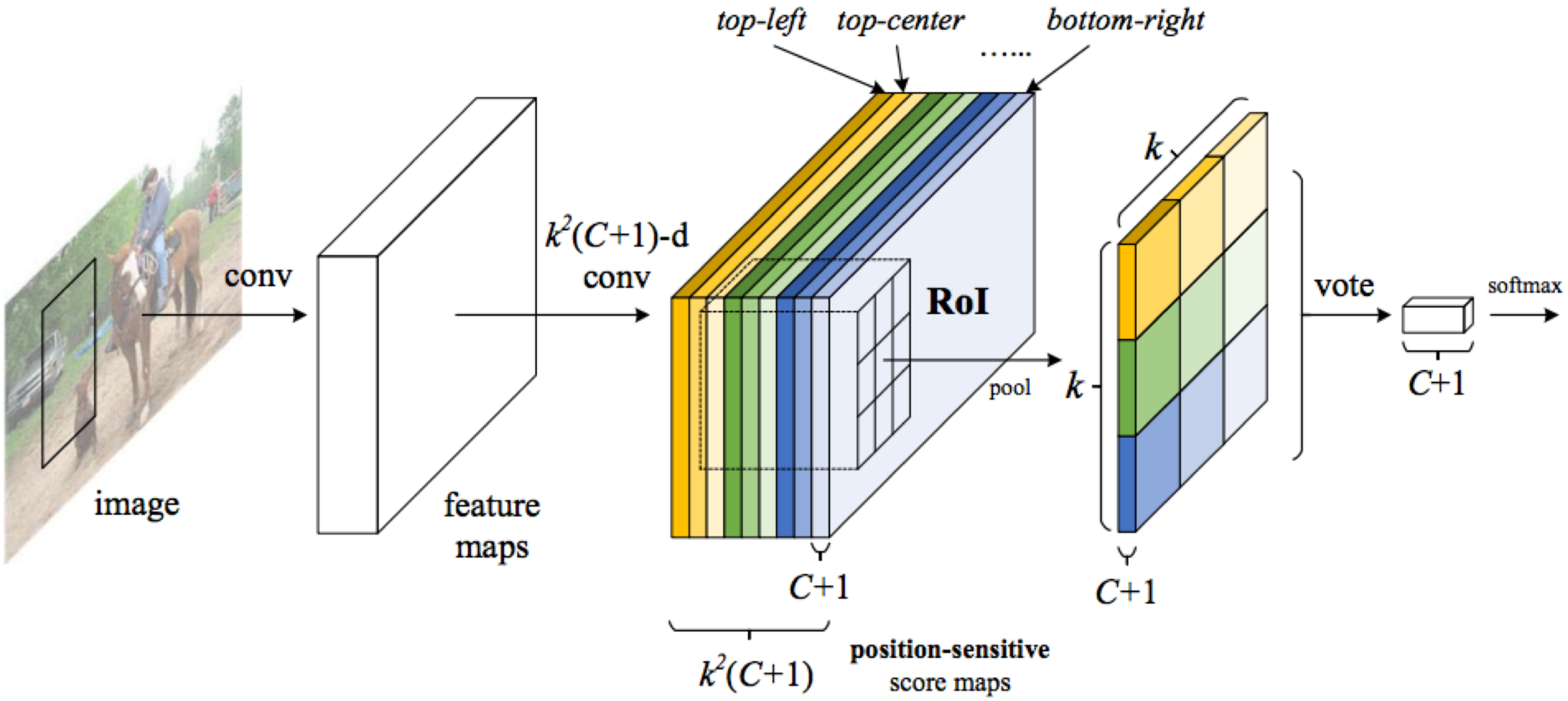
RoI pooling (Faster R-CNN) vs convolution (SSD)

- 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



R-FCN

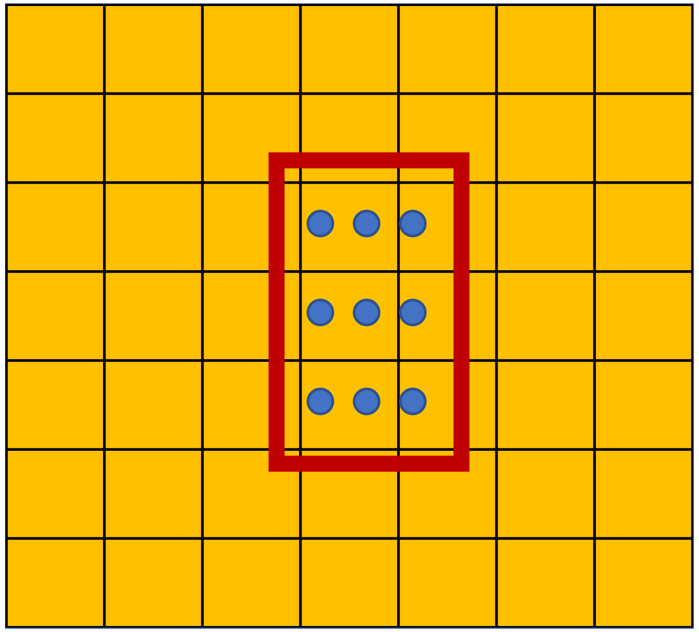
- Can we get benefits of RoI pooling in linear time?



Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." *Advances in neural information processing systems*. 2016.

Other details - ROI Align

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



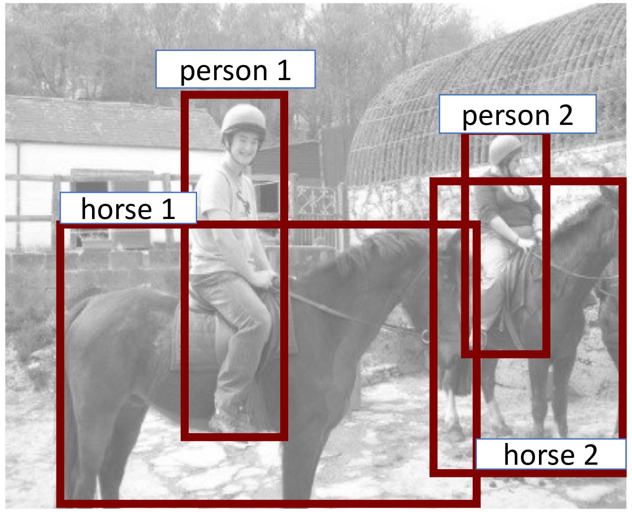
Mask R-CNN. K. He, G. Gkioxari, P. Dollar, R. Girshick. In *ICCV* 2017.

Instance segmentation

Till now

horse, person

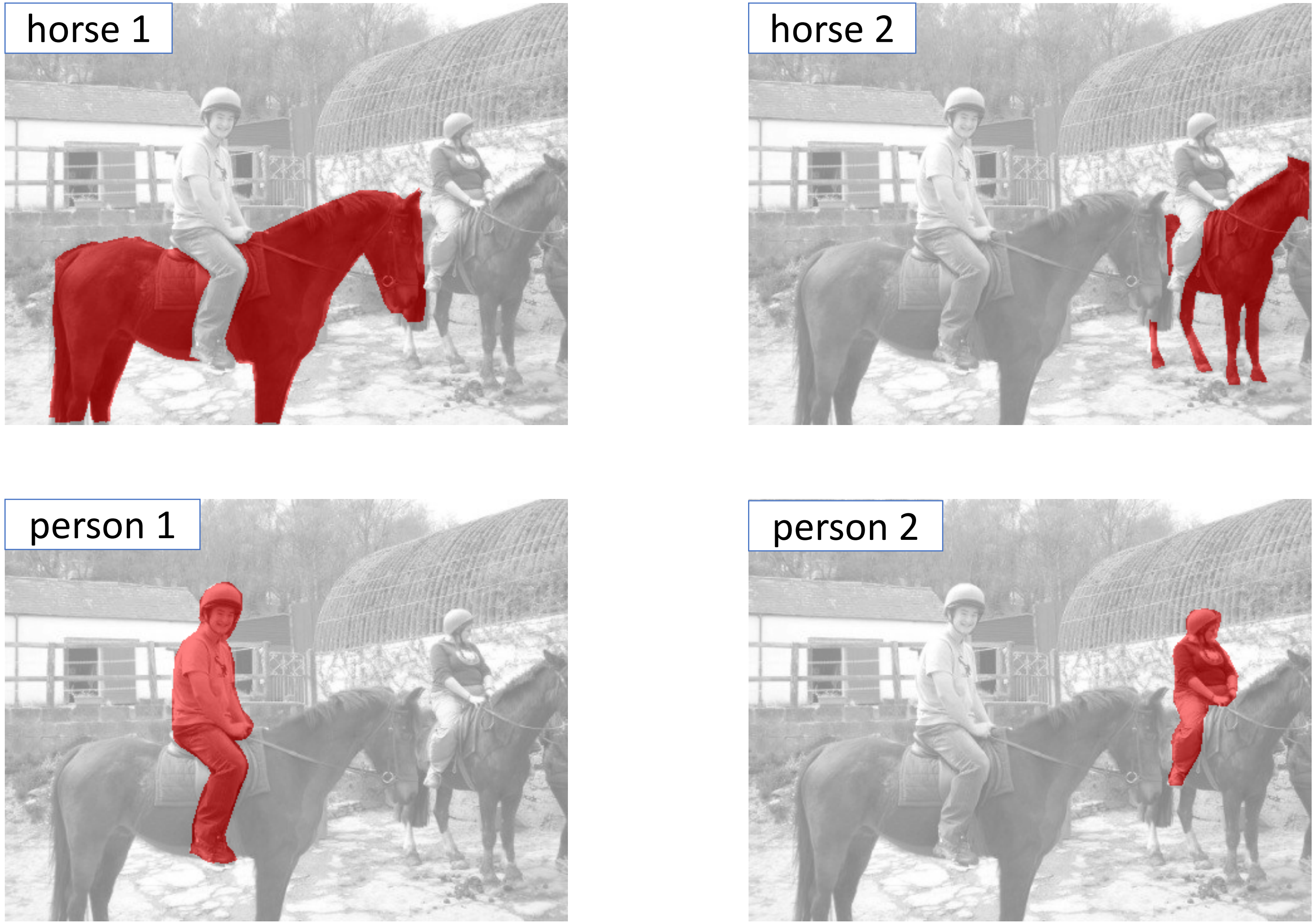
Image Classification



Object Detection

Semantic Segmentation

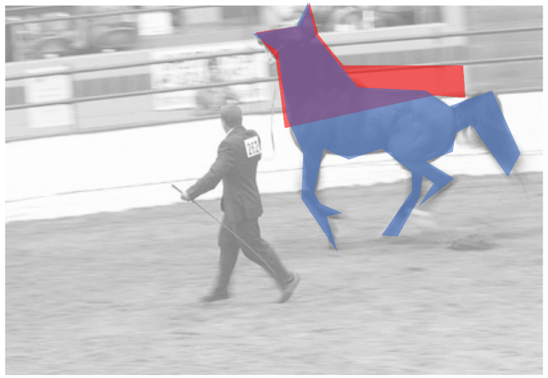
Fine-grained Localization



Instance Segmentation

Evaluation Protocol

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

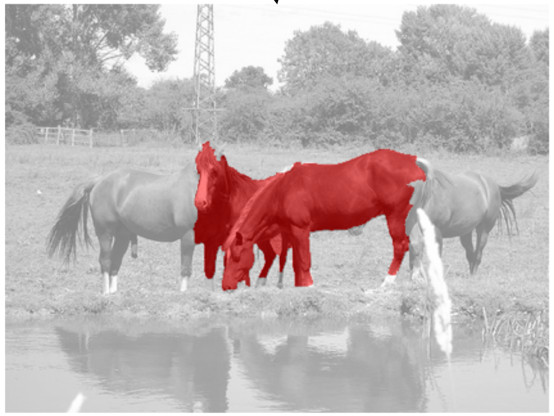


segment overlap = $\frac{\blacksquare \cap \blacksquare}{\blacksquare \cup \blacksquare}$

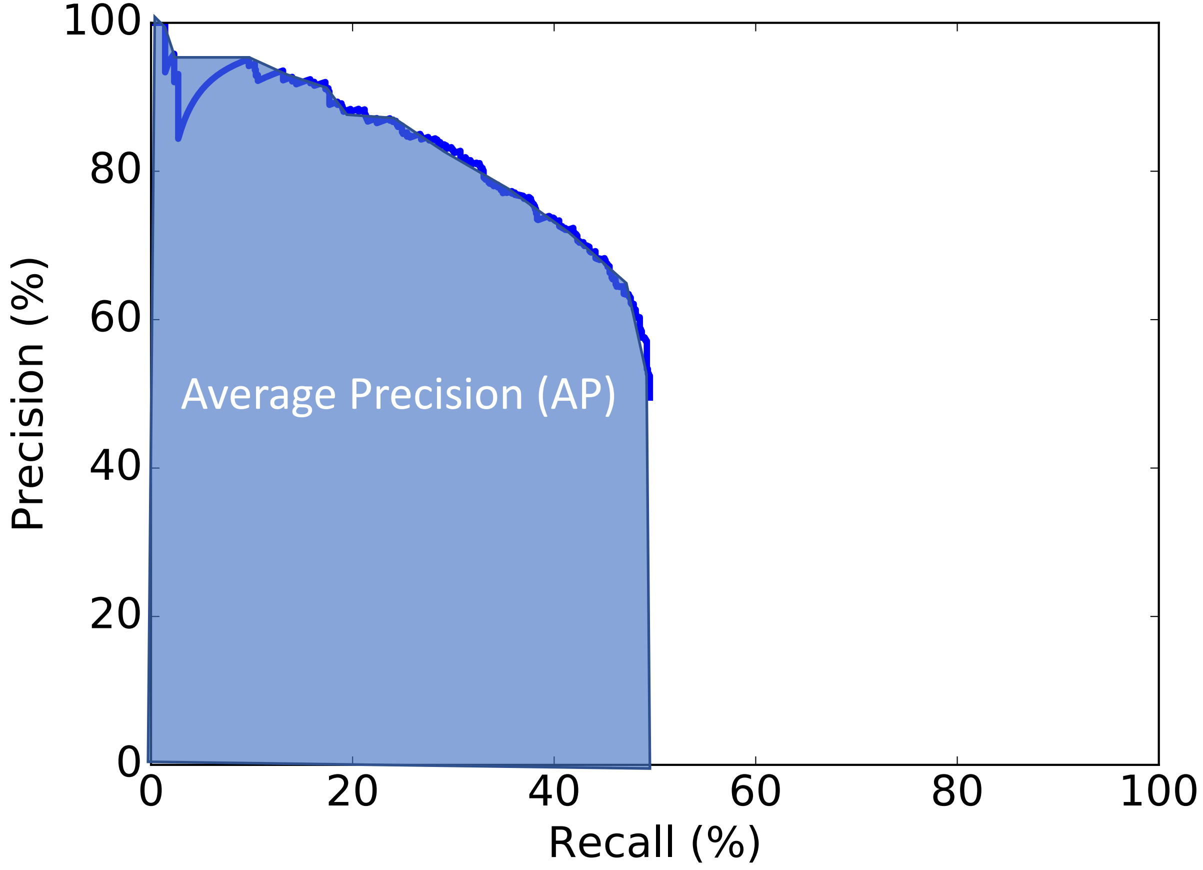
Evaluation Protocol

Labels = [✓ ✓ ✗ ✓ ✗ ✗ ✓ ....]

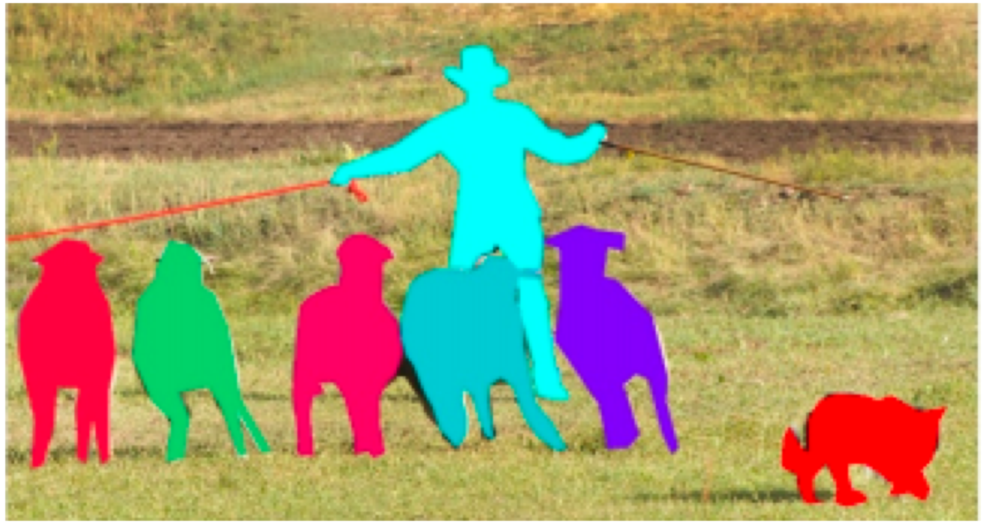
Scores = [0.90 0.87 0.82 0.78 0.70 0.69 0.60 ....]



Evaluation protocol



The COCO Challenge



mscoco.org

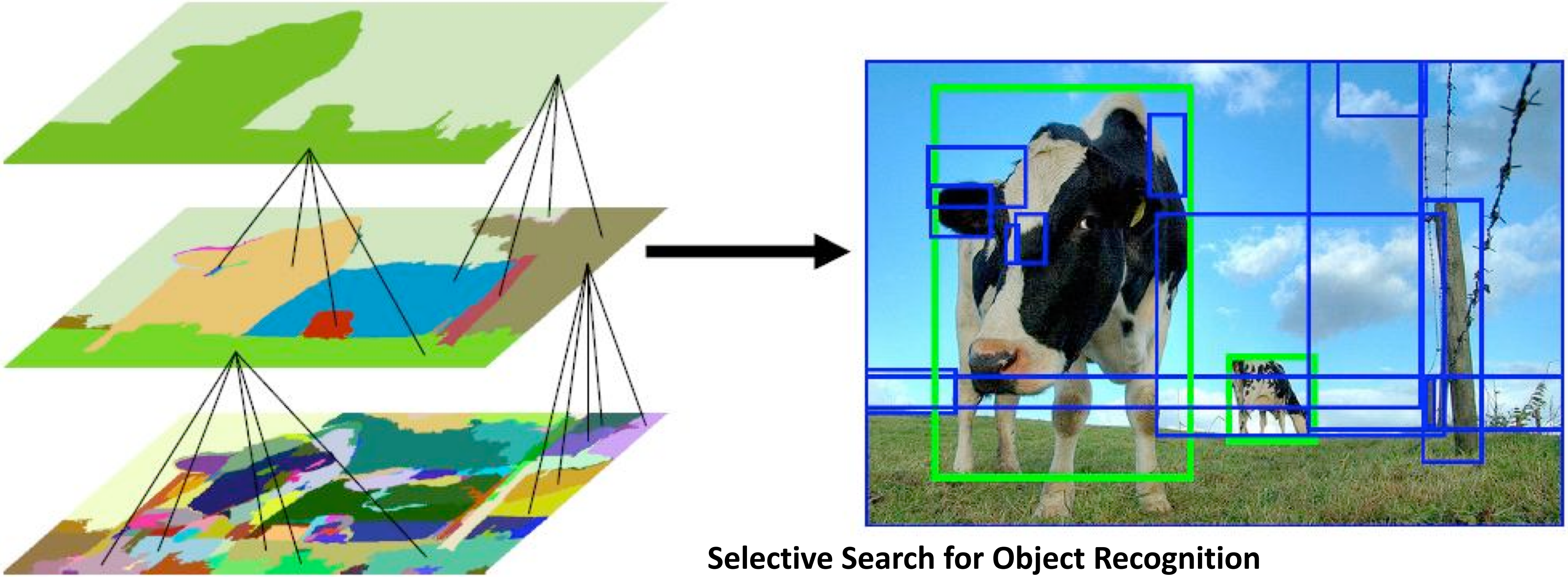
T. Y. Lin et al. **Microsoft COCO: Common Objects in Context**. In ECCV, 2014

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*

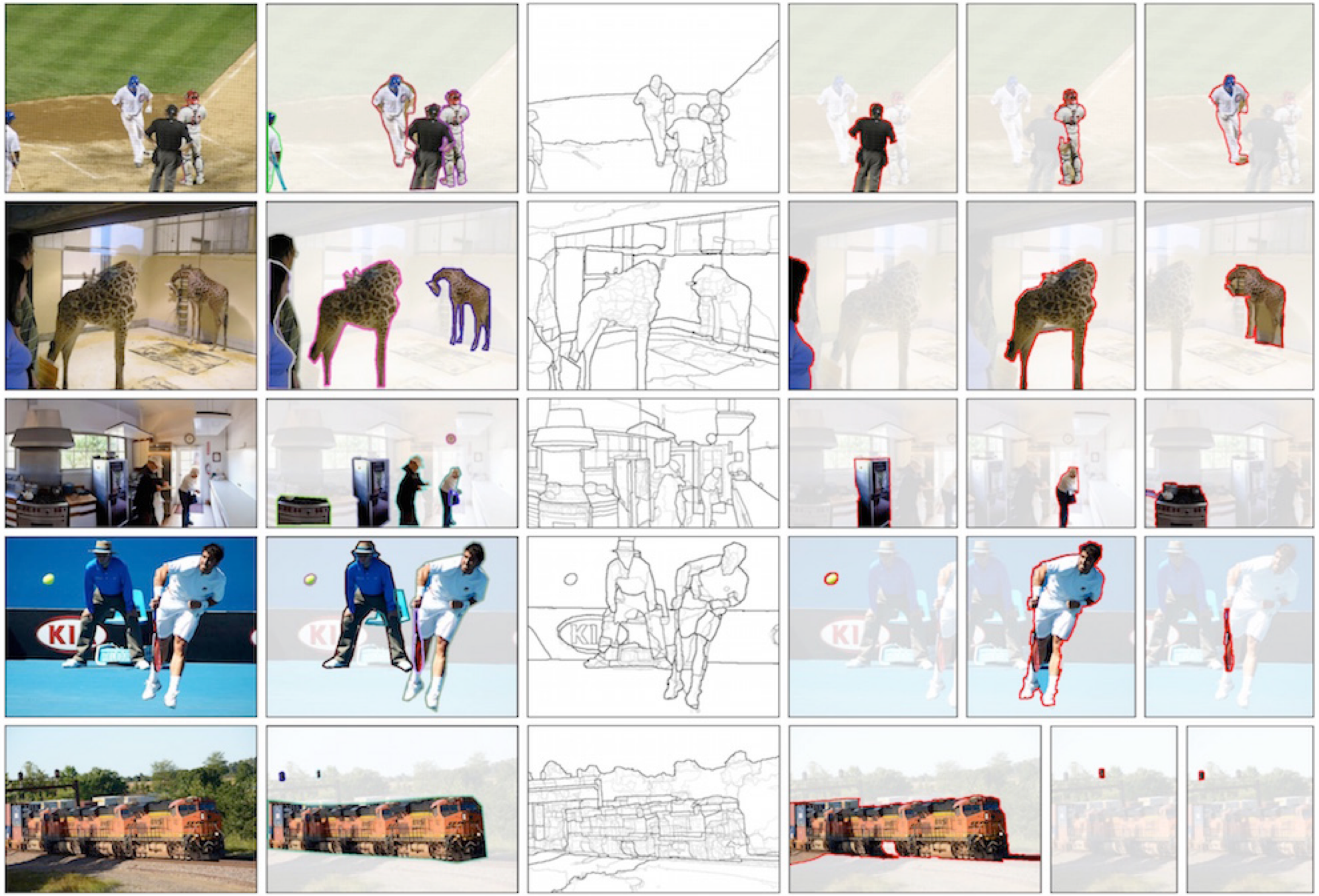
Box proposals

- Use segmentation to produce ~5K candidates



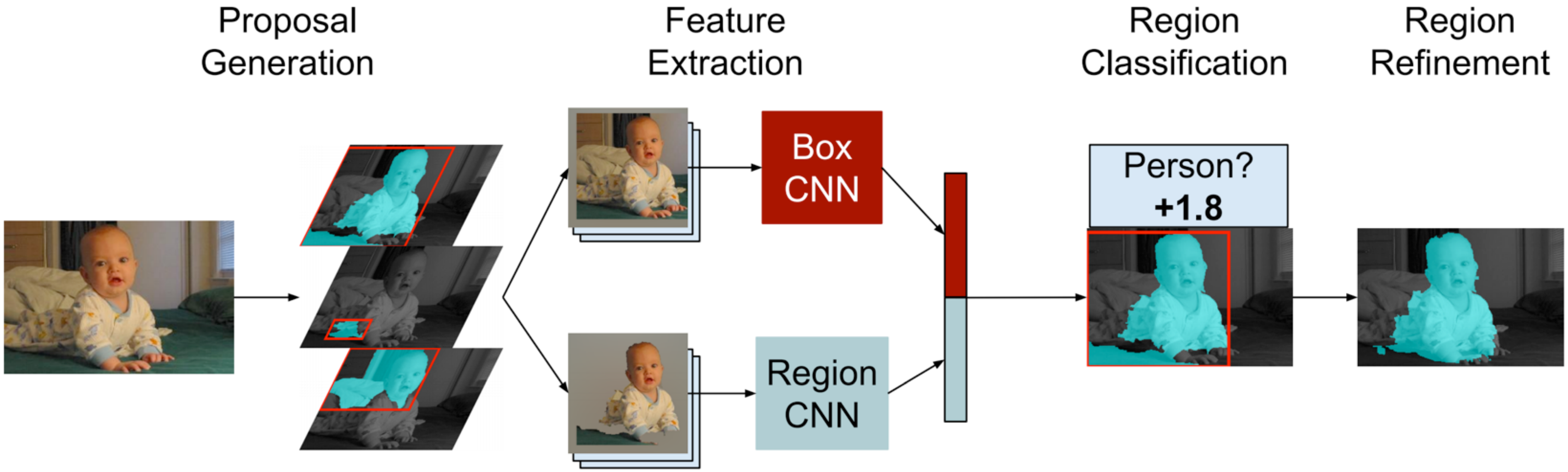
Selective Search for Object Recognition
J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders
In International Journal of Computer Vision 2013.

Segment proposals



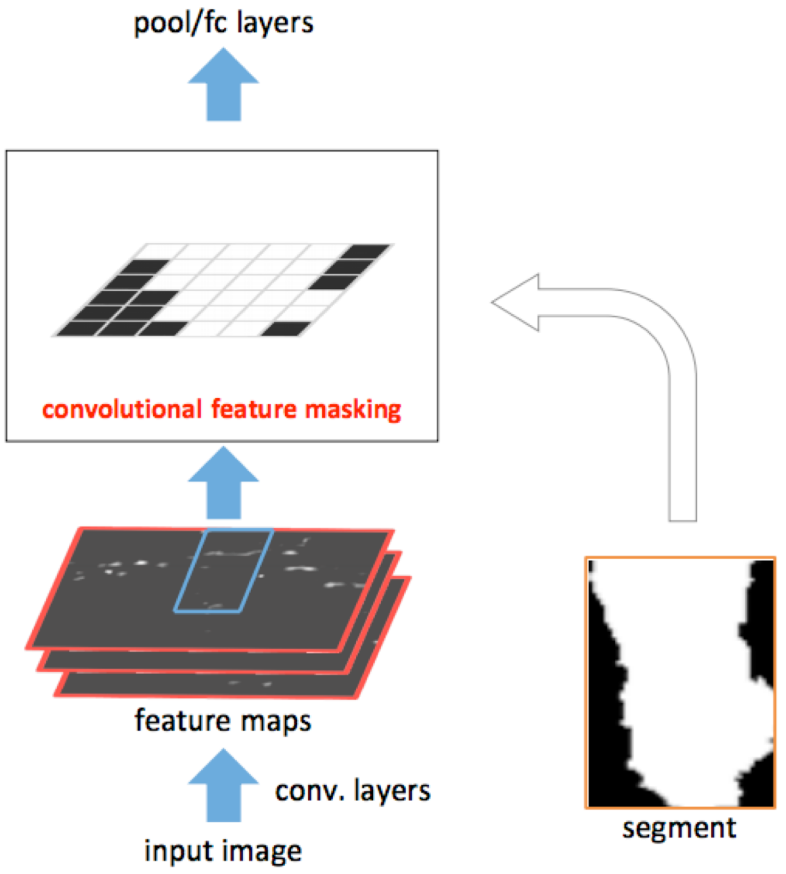
Multi-scale combinatorial grouping. Pablo Arbelaez, Jordi Pont-Tuset, Jonathan Barron, Ferran Marques, Jitendra Malik. In *CVPR,* 2014.

R-CNN for instance segmentation



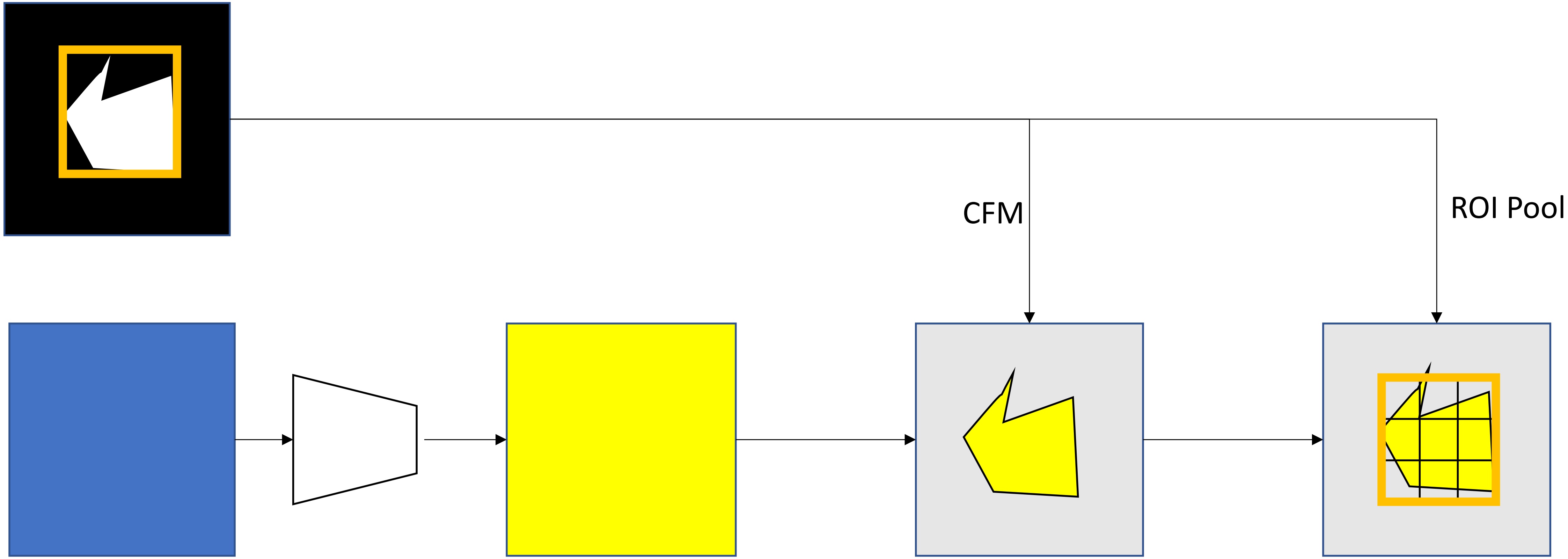
Simultaneous Detection and Segmentation Bharath Hariharan , Pablo Arbelaez , Ross Girshick, Jitendra Malik. In *ECCV* 2014

Fast R-CNN for instance segmentation



Convolutional Feature Masking for Joint Object and Stuff Segmentation. Jifeng Dai, Kaiming He, Jian Sun. In *CVPR* 2015

Fast R-CNN for instance segmentation



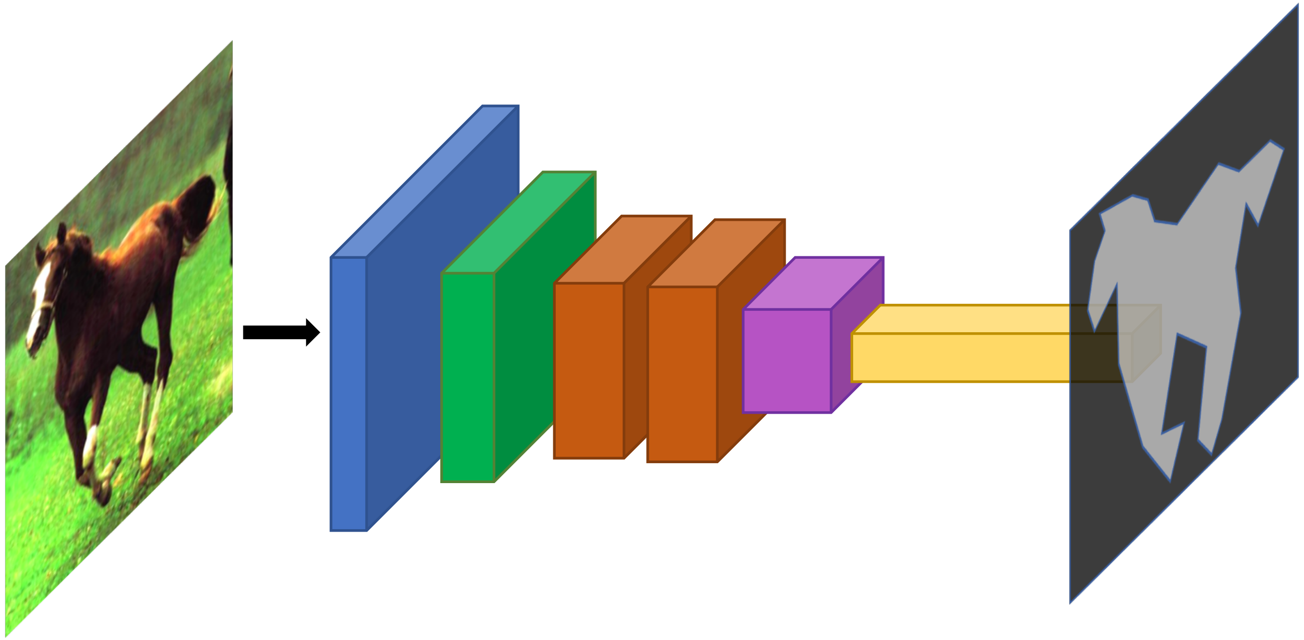
Convolutional Feature Masking for Joint Object and Stuff Segmentation. Jifeng Dai, Kaiming He, Jian Sun. In *CVPR* 2015

Two strategies

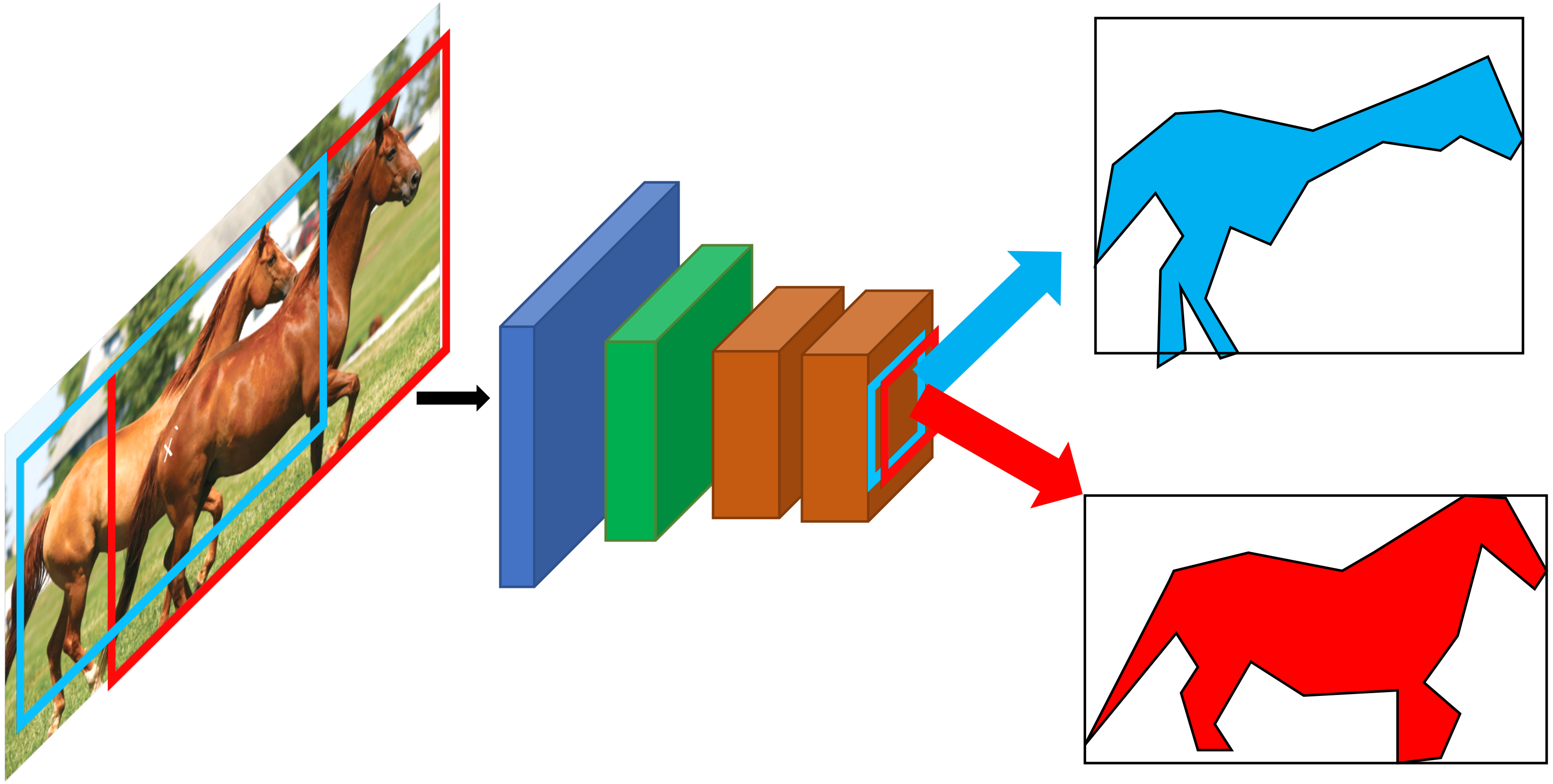
- Segment then classify
  - Use bottom-up techniques to come up with *segment* proposals
  - Classify segment proposals with convnets
  - Segmentation is category agnostic
  - Modification: use convnets to produce segmentation proposals
- Detect then segment
  - Use standard object detection to produce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

Predicting segmentation for bounding boxes

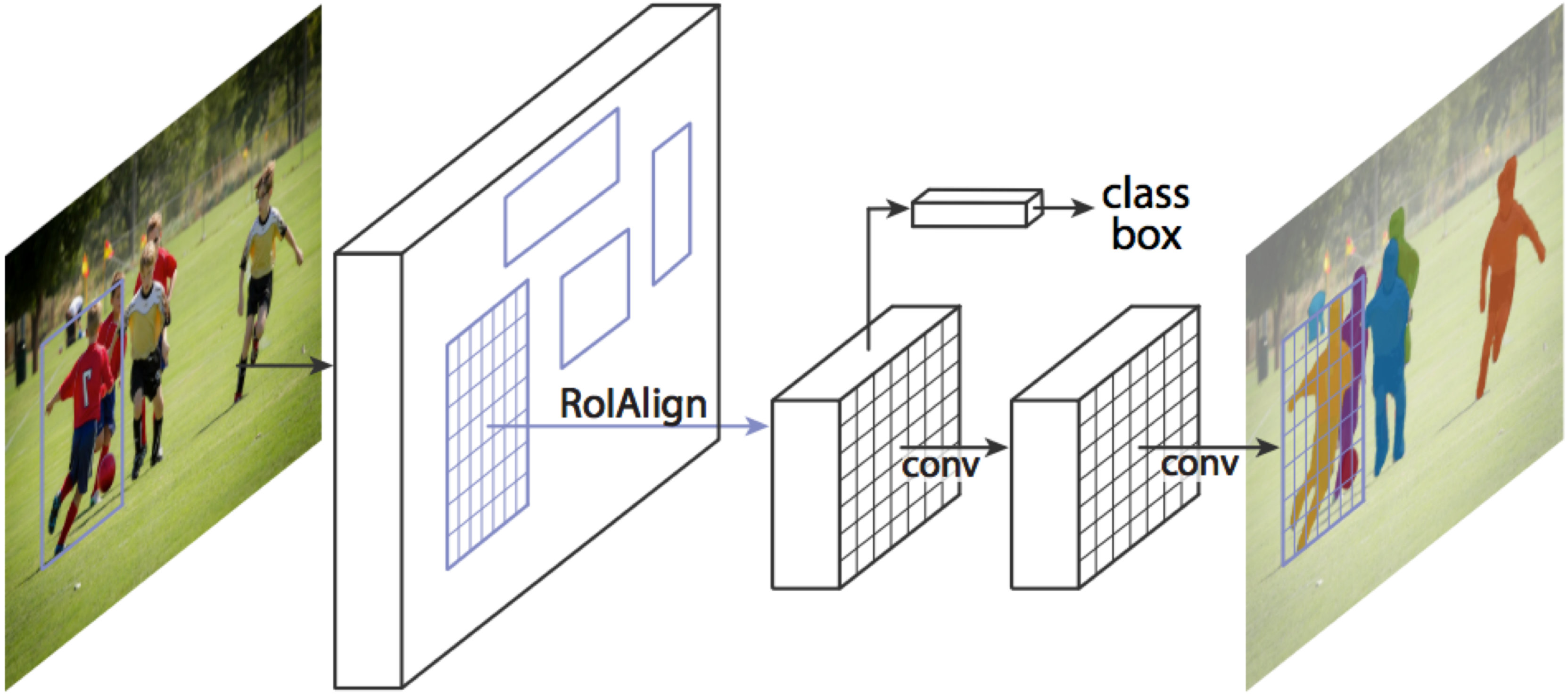
- Given a detection, predict segmentation



Predicting segmentation for bounding boxes

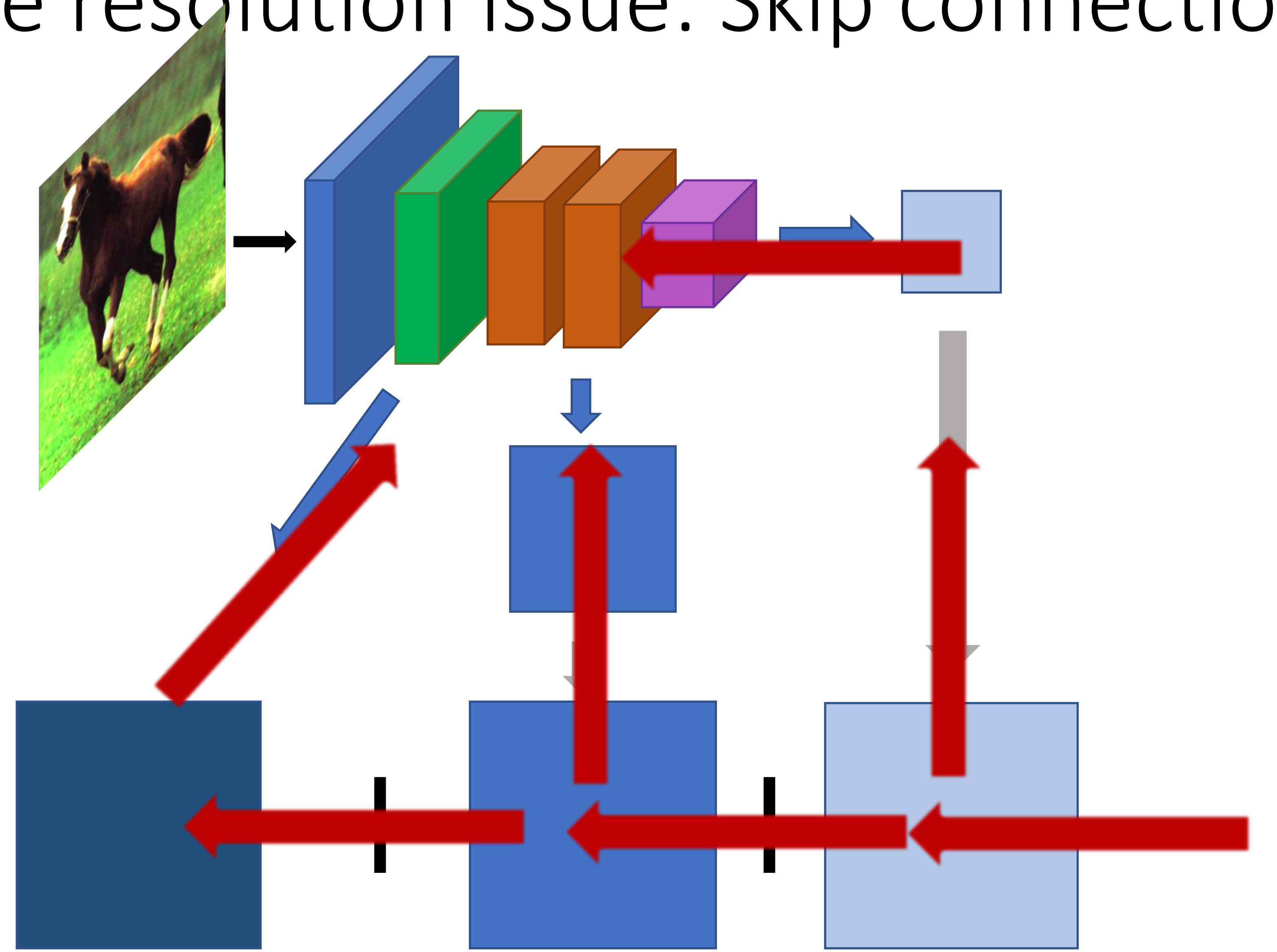


Mask R-CNN

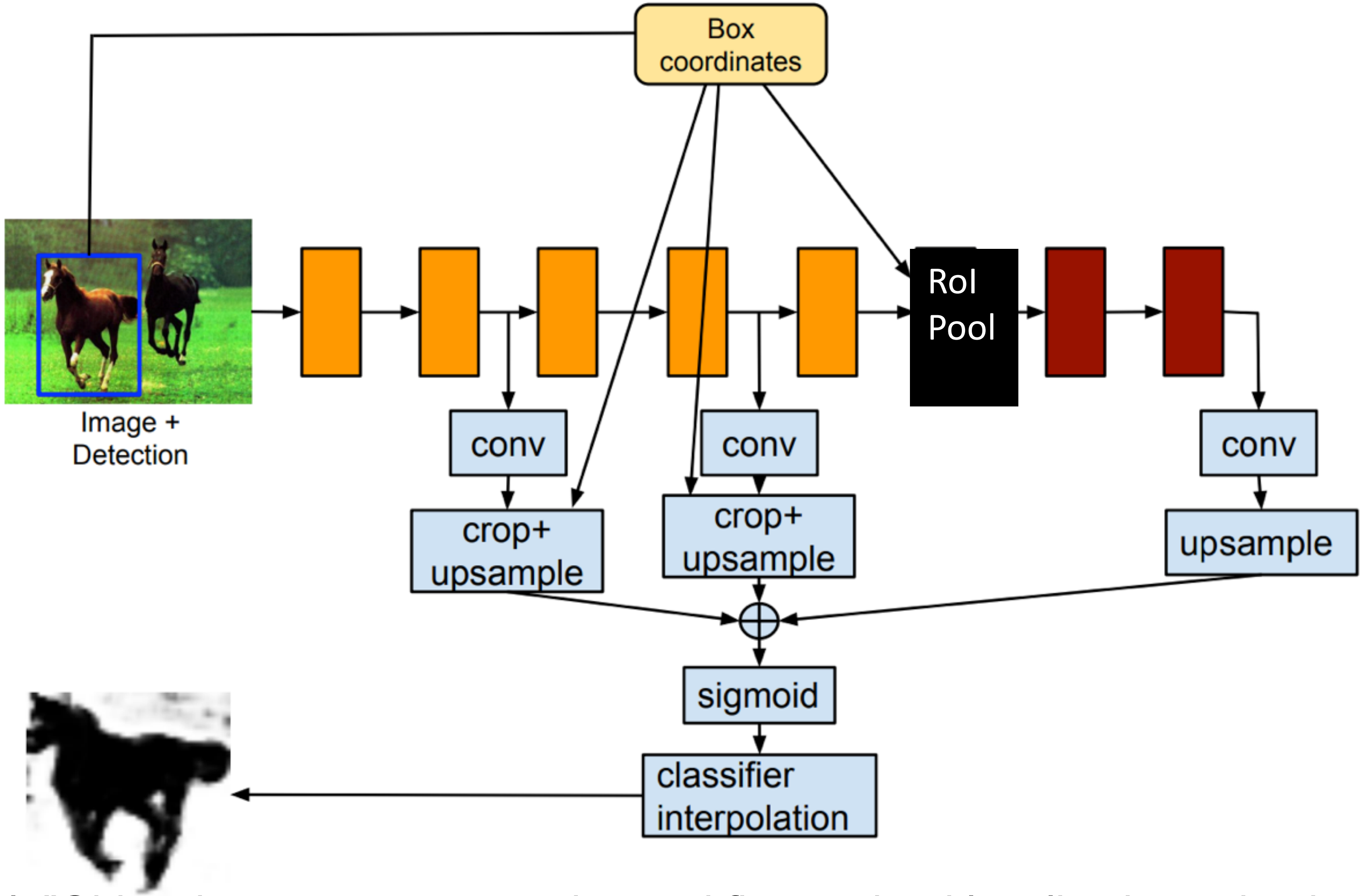


He, Kaiming, et al. "Mask r-cnn." *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017.

The resolution issue: Skip connections

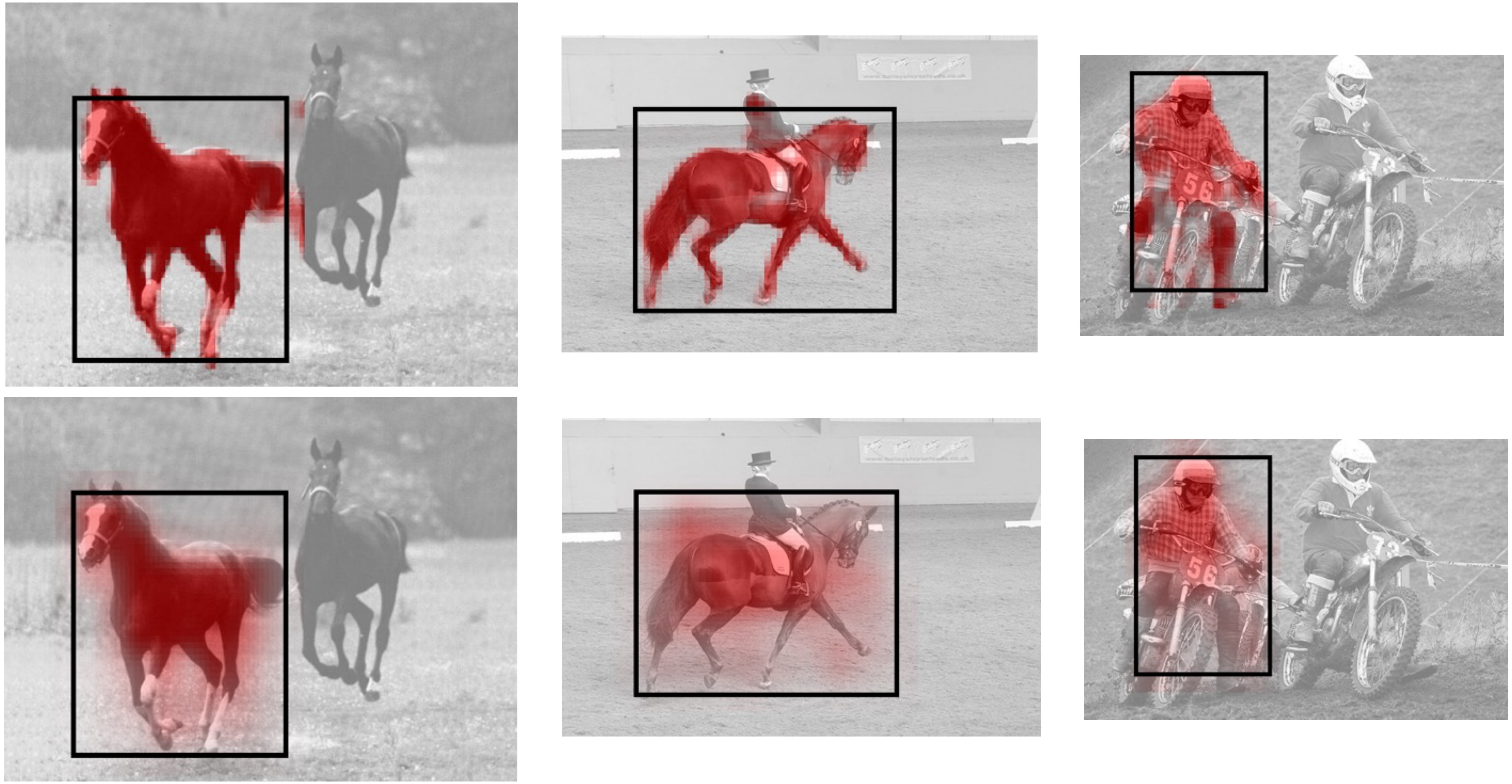


Skip connections with RoI pooling



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

The resolution issue: Skip connections



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

Fully convolutional or not?

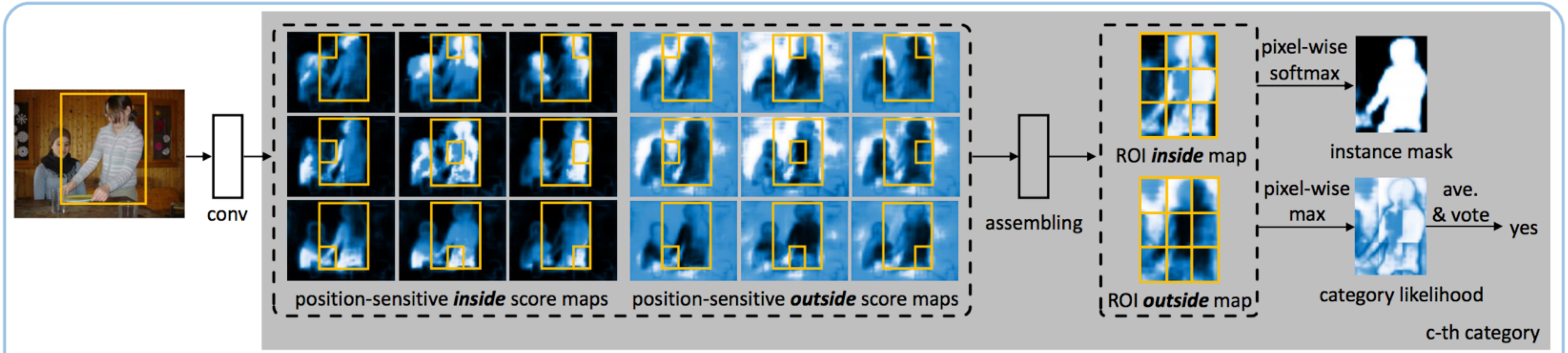
- Convolution is *location invariant*
- Semantic segmentation is *location invariant*
- Figure-ground segmentation of a detected bounding box is *not location invariant*

Fully convolutional or not?

- Convolution is cheaper
- Local connections without parameter sharing?
    - Too many parameters?
    - Predicting 50x50 segmentation: 2500 times more parameters
- Predict lower resolution, then upsample

Fully convolutional:
Mask R-CNN. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. In *ICCV,* 2017.
Non-fully convolutional:
Learning to Segment Object Candidates. Pedro O. Pinheiro, Ronan Collobert, Piotr Dollar. In *ICCV,* 2015.

Instance fully-convolutional networks



(c) Our fully convolutional instance-aware semantic segmentation

Li, Yi, et al. "Fully convolutional instance-aware semantic segmentation." *arXiv preprint arXiv:1611.07709* (2016).

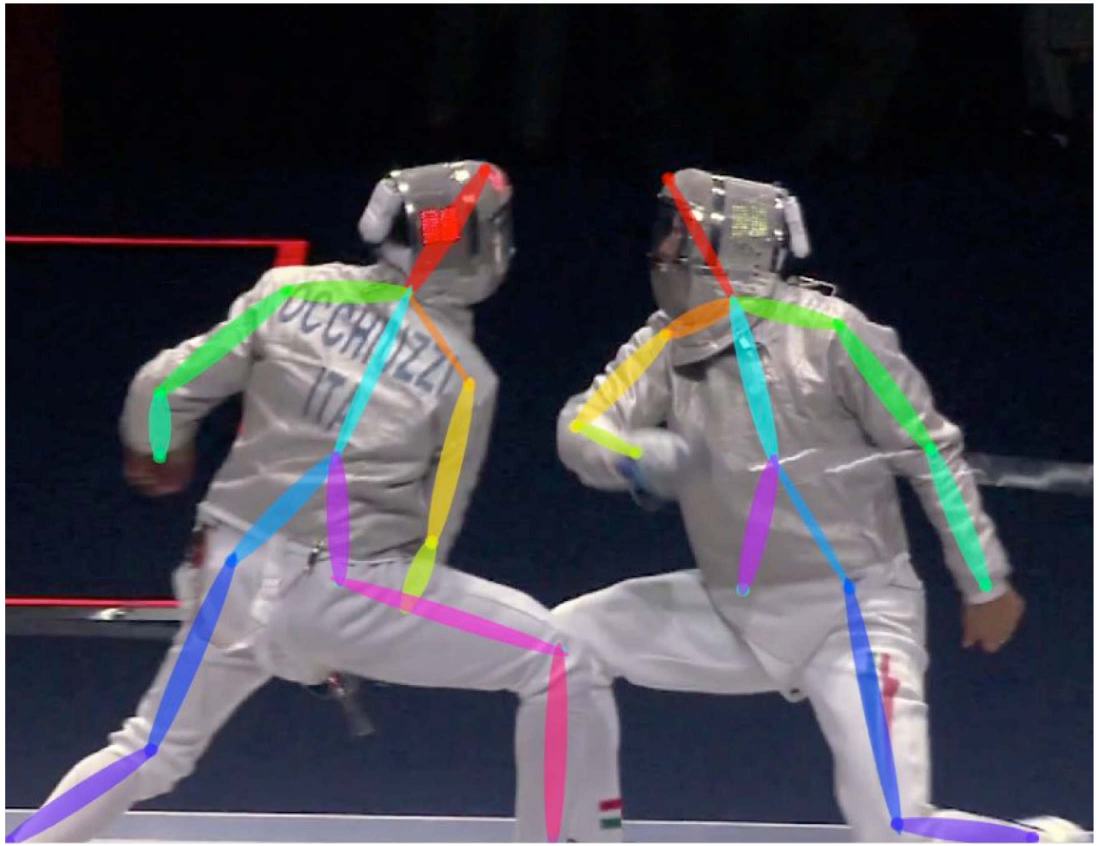
Final results - what works?

- 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

Human pose estimation

The task

- Mark joint locations for person
- Nose
- Right/left shoulder
- Right/left elbow
- Right/left hip
- …



Two versions of task

- Assume people have been detected
- Rough bounding box given
- Key info available:
  - scale
  - only 1 location per joint
- Pros: disentangles detection and pose estimation
- Cons: unrealistic

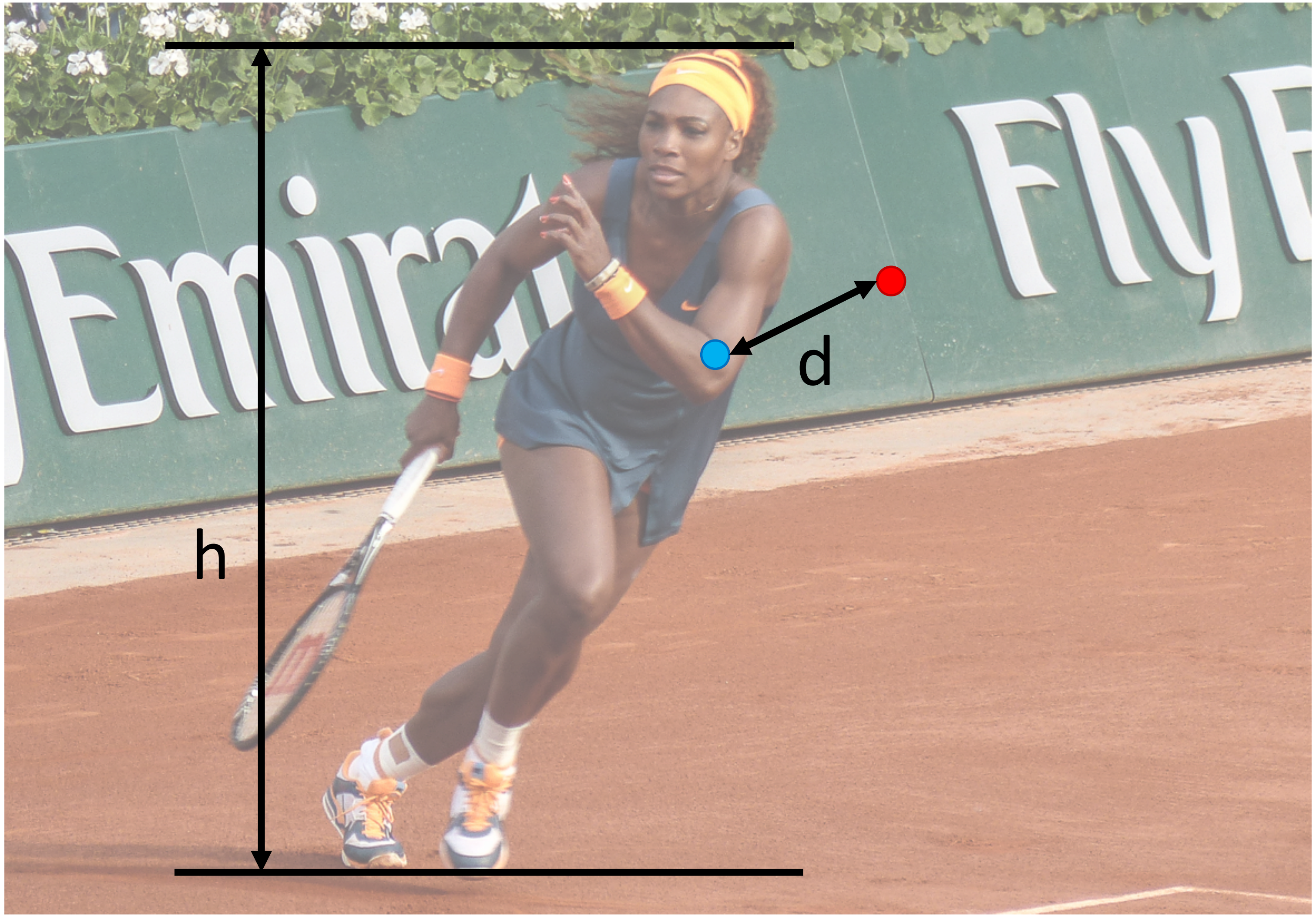
- Tabula rasa without detections
- Challenge: no idea of scale or number
- Possible opportunity: use keypoint estimates to improve detections
- Pros: realistic
- Cons: conflates detection and pose estimation

Pose estimation given detection

Evaluation metric - given detection

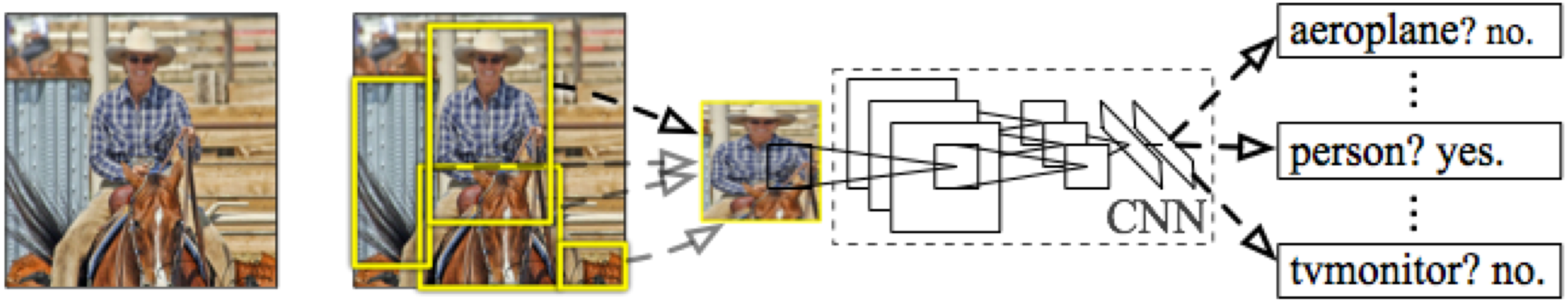
- Evaluate every keypoint separately
- For each person, check if keypoint is correct
- Compure fraction of people for which keypoint is correct: PCK (*Probability of Correct Keypoint*)

Evaluation metric - given detection



$d/h < \alpha$?

R-CNN: Regions with CNN features



Input image

Extract region proposals (~2k / image)

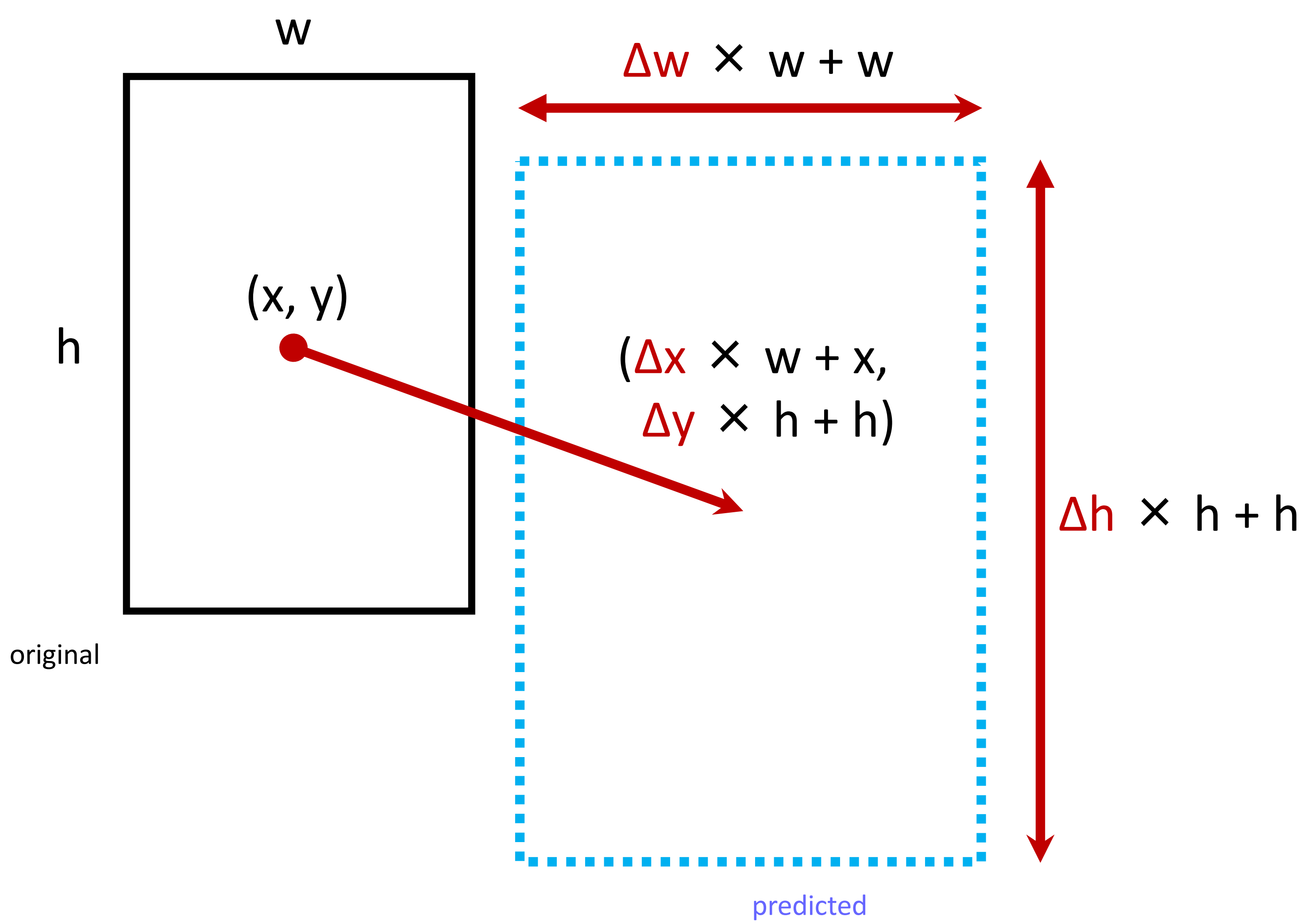
Compute CNN features

Classify regions (linear SVM)

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation
R. Girshick, J. Donahue, T. Darrell, J. Malik
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

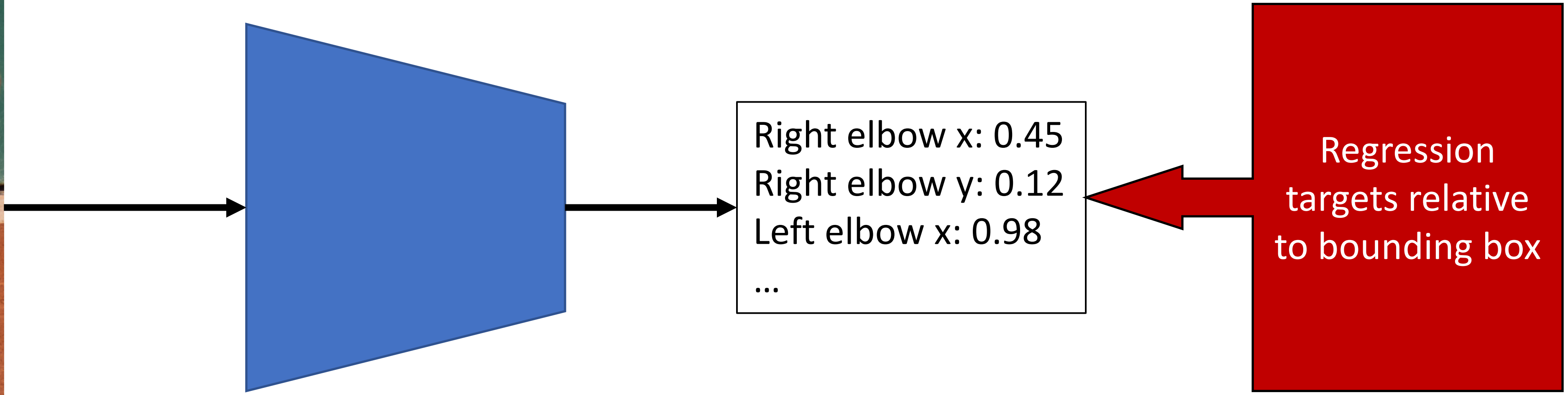
Slide credit : Ross Girshick

Bounding-box regression



Slide credit : Ross Girshick

Strategy 1: Regression



DeepPose: Human Pose Estimation via Deep Neural Networks. Alexander Toshev and Christian Szegedy. In *CVPR,* 2014.

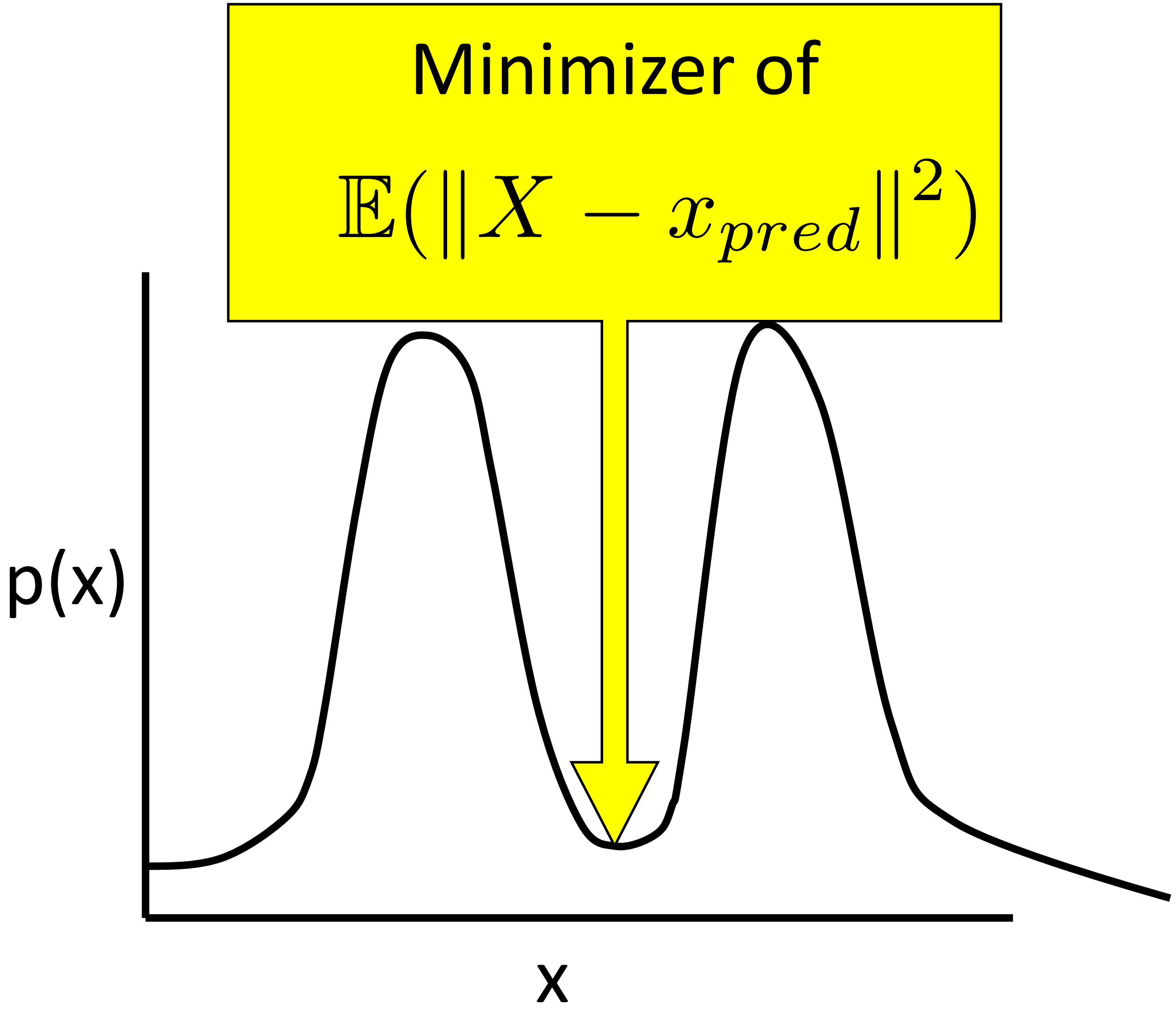
Strategy 1: Regression

- Assumes global object features has enough information for accurate localization
  - Localization info missing due to subsampling?
- Solution: Refinement!

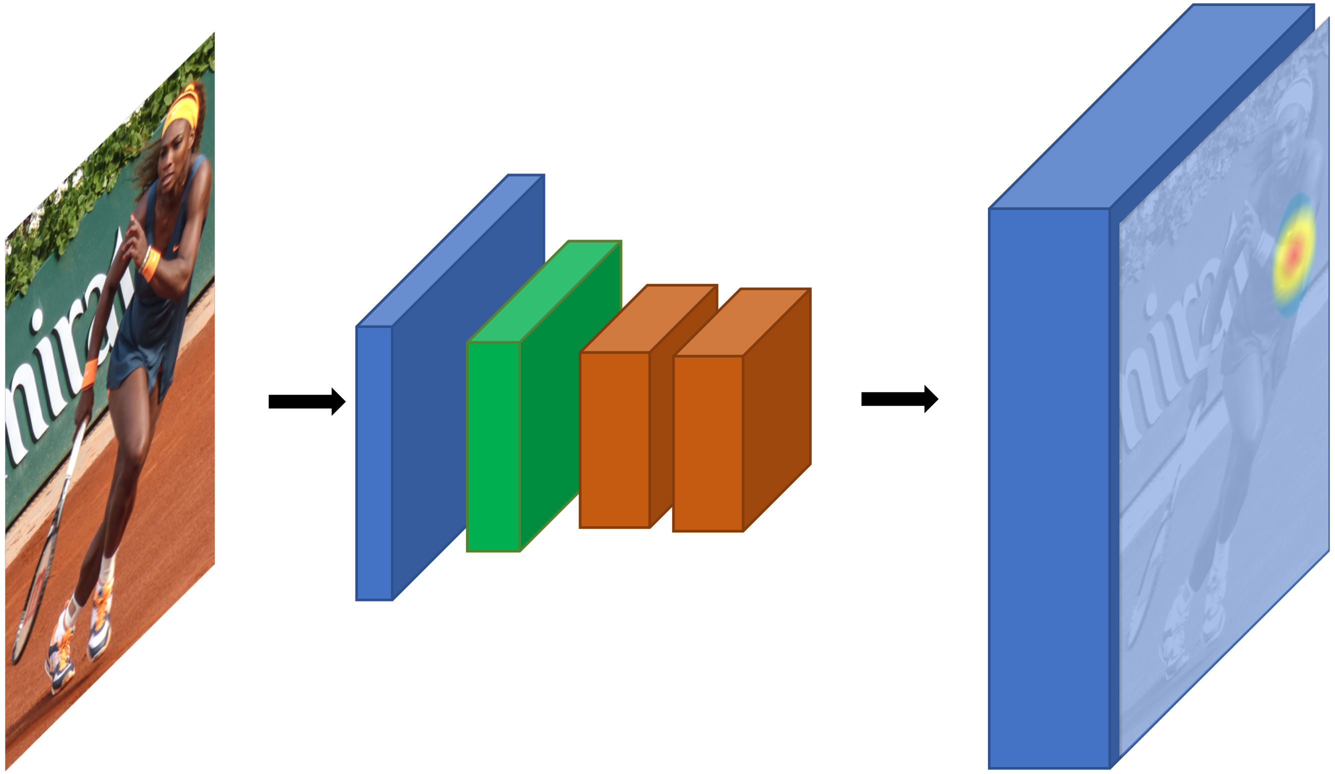
DeepPose: Human Pose Estimation via Deep Neural Networks. Alexander Toshev and Christian Szegedy. In *CVPR,* 2014.

Strategy 1: Regression

- Multimodal distributions?



Strategy 2: Heatmaps



Strategy 2: Heatmaps

- Still have the resolution issue
- Same solutions
  - Dilation?
  - Multiple layers?
  - Multiple image scales?

- abc