Object Detection & Instance Segmentation

Yinan Zhao
University of Texas at Austin

Image Classification

This image is CC0 public domain

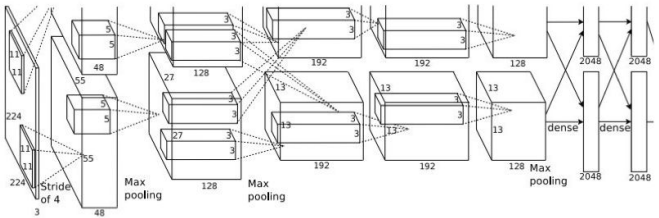


Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

Vector:
4096

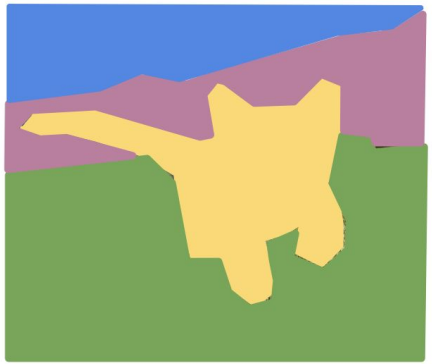
Fully-Connected:
4096 to 1000

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Detection & Segmentation

Semantic Segmentation



GRASS, CAT, TREE, SKY

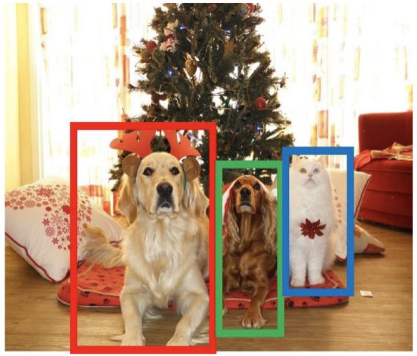
No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection



DOG, DOG, CAT

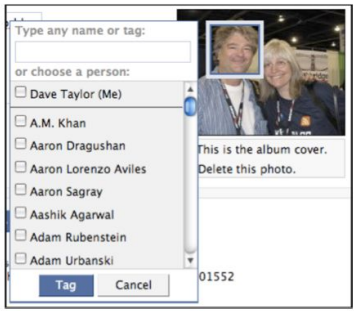
Instance Segmentation

DOG, DOG, CAT

Multiple Object

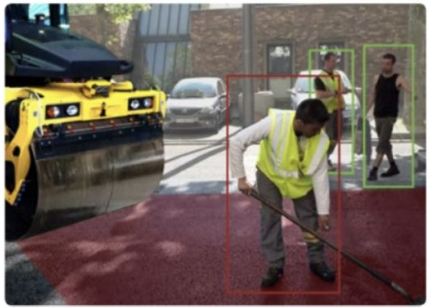
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Application: Object Detection

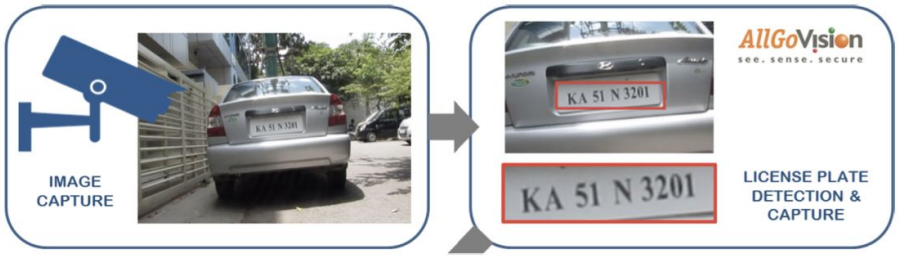


Face detection
(e.g., Facebook)

Mobile check deposit
(e.g., Bank of America)



Pedestrian Detection
(e.g., Blaxtair)
http://media.brintex.com/Occurrence/121/Brochure/3435/brochure.pdf



License Plate Detection (e.g., AllGoVision)

https://www.ischool.utexas.edu/~dannag/Courses/CrowdsourcingForCV/Lectures/07-ObjectDetection.pdf

Application: Object Detection

Counting Fish (e.g., SalmonSoft)
http://www.wecountfish.com/?page_id=143

Business Traffic Analytics

Counting

https://www.ischool.utexas.edu/~dannag/Courses/CrowdsourcingForCV/Lectures/07-ObjectDetection.pdf

Application: Segmentation

Materials Database

[Bell et al; SIGGRAPH; 2013]

Retexturing

(a) Target photo (b) Retextured

Image Search

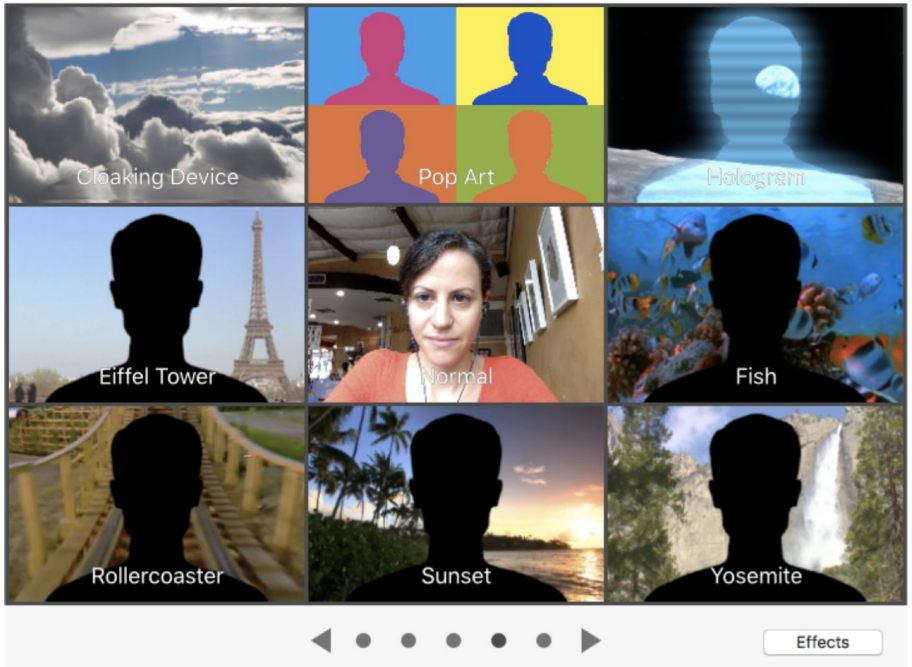
Query

Results: wood floors in kitchens, sorted by diffuse color

https://www.ischool.utexas.edu/~dannag/Courses/CrowdsourcingForCV/Lectures/06-Segmentation.pdf

Application: Segmentation

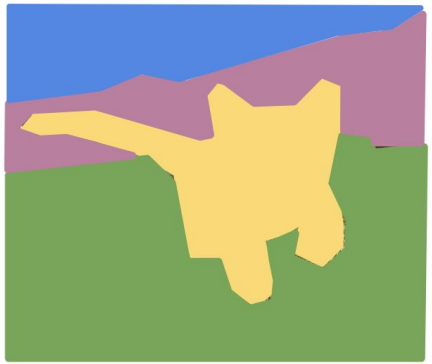
Face Changers (e.g., Photo Booth, phone apps)



https://www.ischool.utexas.edu/~dannag/Courses/CrowdsourcingForCV/Lectures/06-Segmentation.pdf

Detection & Segmentation

Semantic Segmentation



GRASS, CAT, TREE, SKY

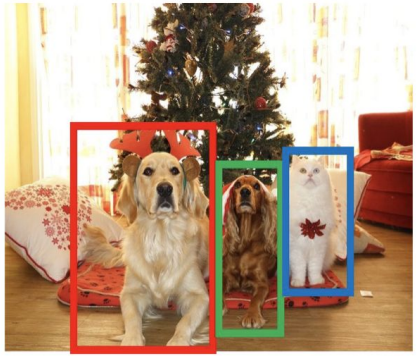
No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection



DOG, DOG, CAT

Instance Segmentation

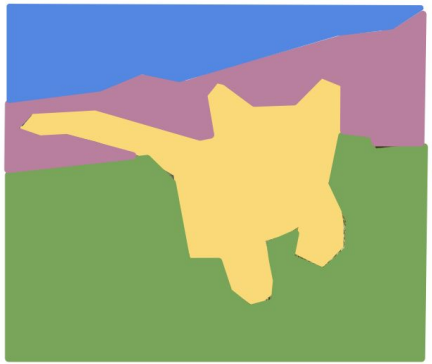
DOG, DOG, CAT

Multiple Object

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Detection & Segmentation

Semantic Segmentation



GRASS, CAT, TREE, SKY

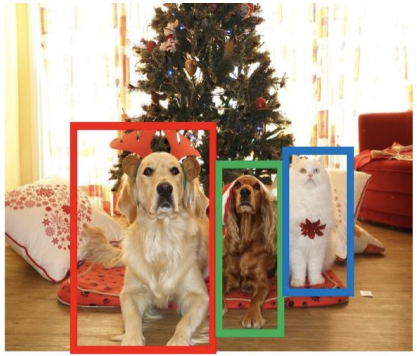
No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection



DOG, DOG, CAT

Instance Segmentation

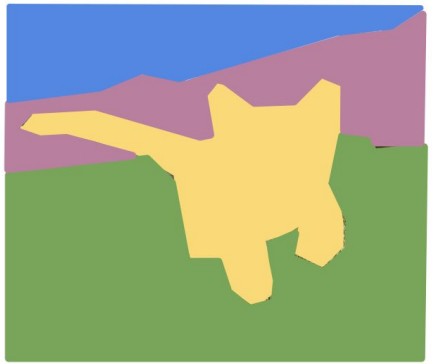
DOG, DOG, CAT

Multiple Object

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Detection & Segmentation

Semantic Segmentation



GRASS, CAT, TREE, SKY

No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection

DOG, DOG, CAT

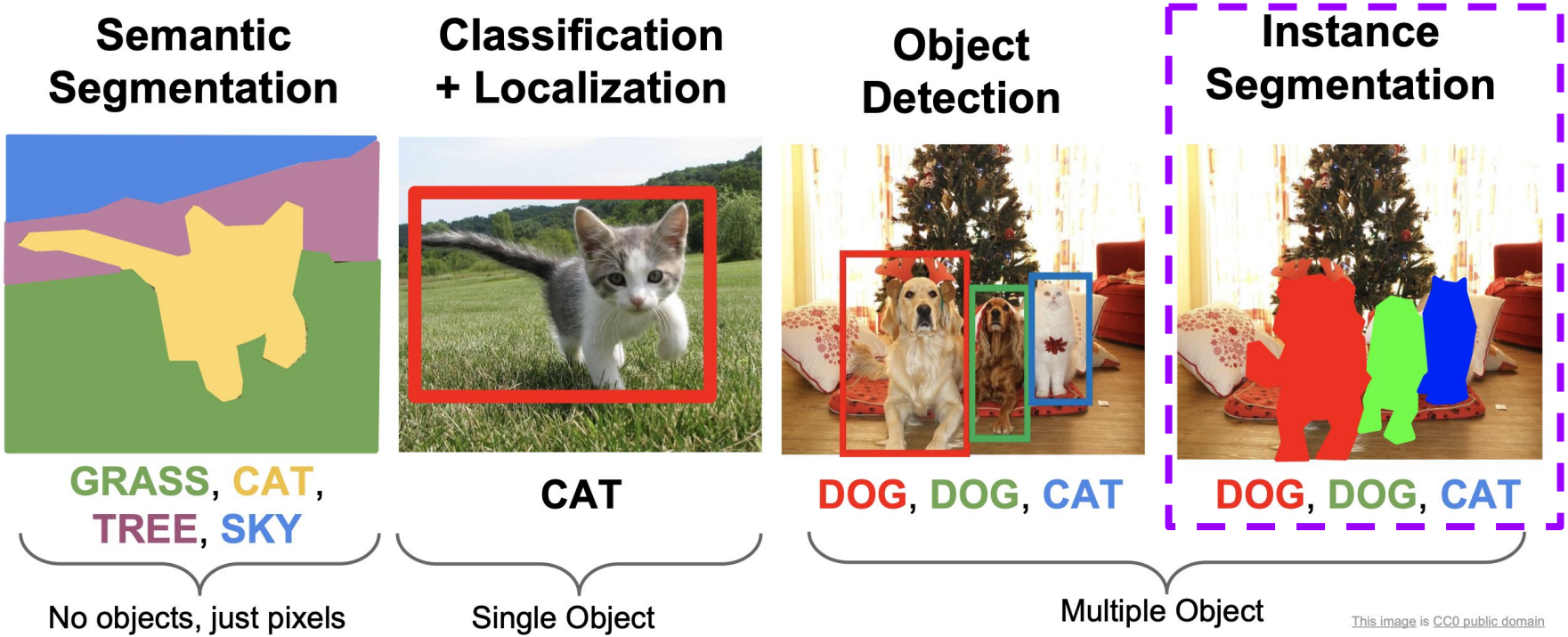
Instance Segmentation

DOG, DOG, CAT

Multiple Object

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

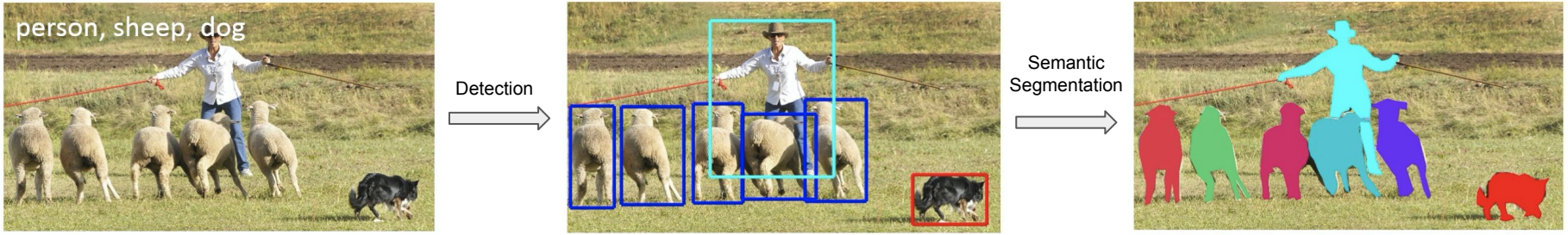
Detection & Segmentation



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Instance Segmentation

Detection first

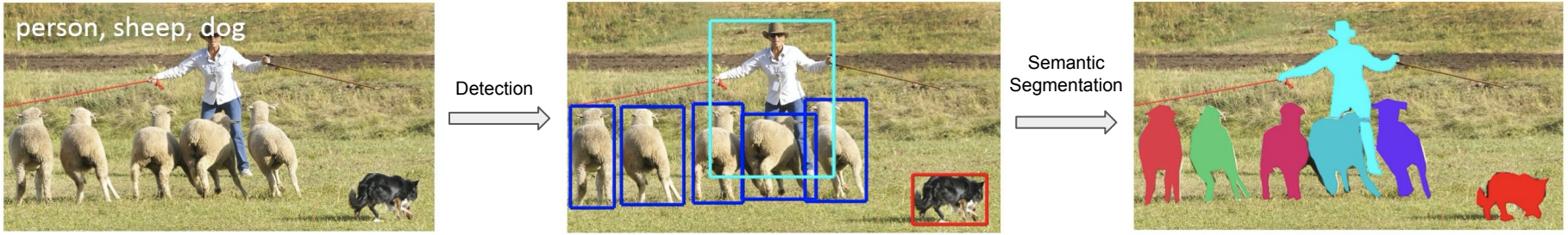


Segmentation first

Instance Segmentation

Detection first

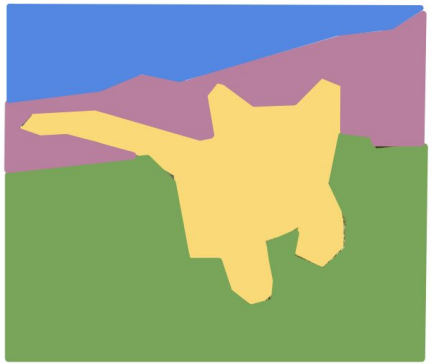
Better performance in practice



Segmentation first

Detection & Segmentation

Semantic Segmentation



GRASS, CAT, TREE, SKY

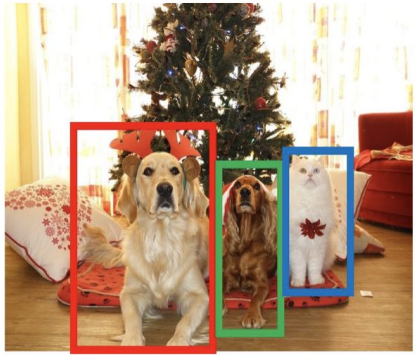
No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection



DOG, DOG, CAT

Instance Segmentation

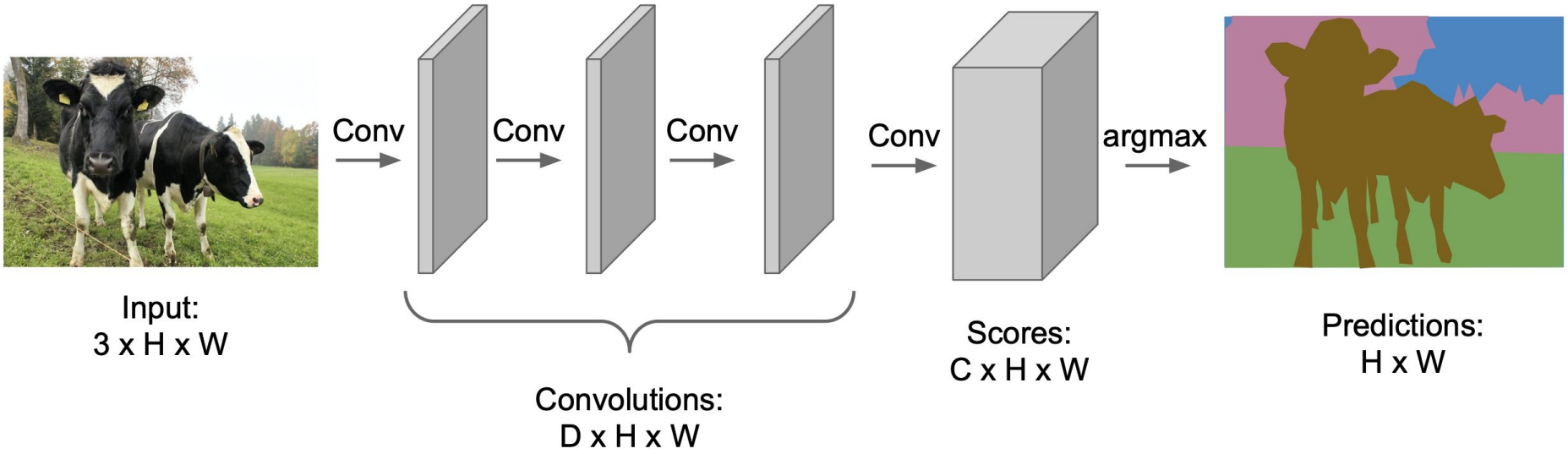
DOG, DOG, CAT

Multiple Object

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Semantic Segmentation: Fully Convolutional Network

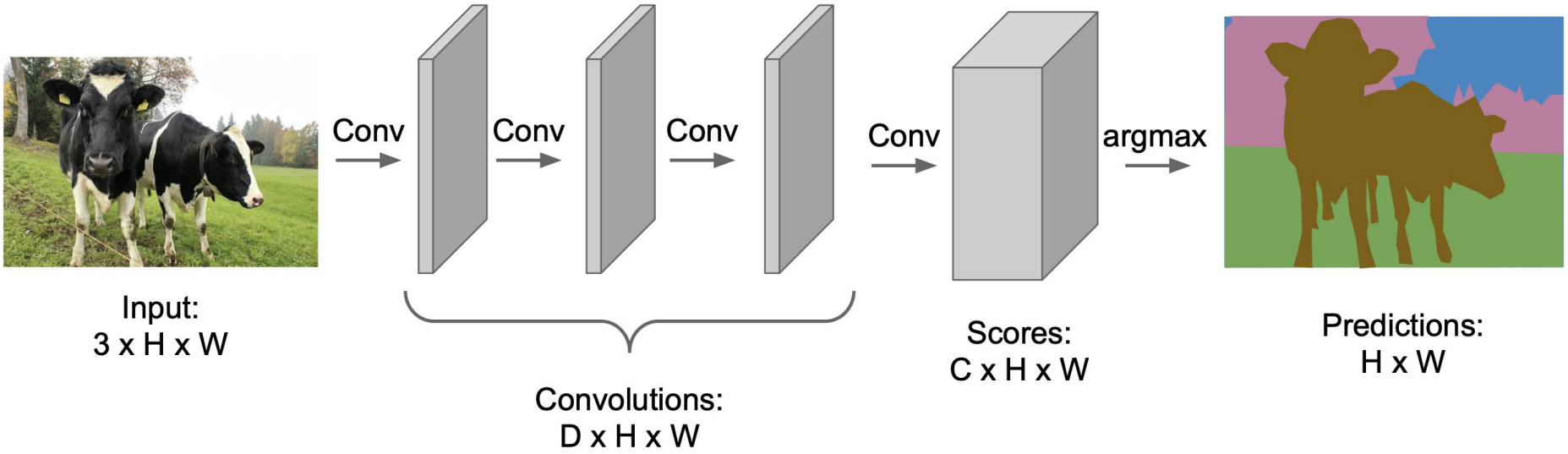
Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Semantic Segmentation: Fully Convolutional Network

Design a network as a bunch of convolutional layers
to make predictions for pixels all at once!

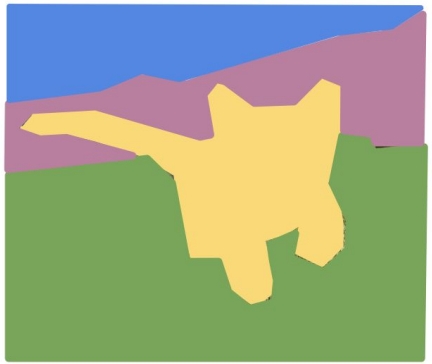


We will also try to tackle object detection by fully convolutional network

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Detection & Segmentation

Semantic Segmentation



GRASS, CAT, TREE, SKY

No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection

DOG, DOG, CAT

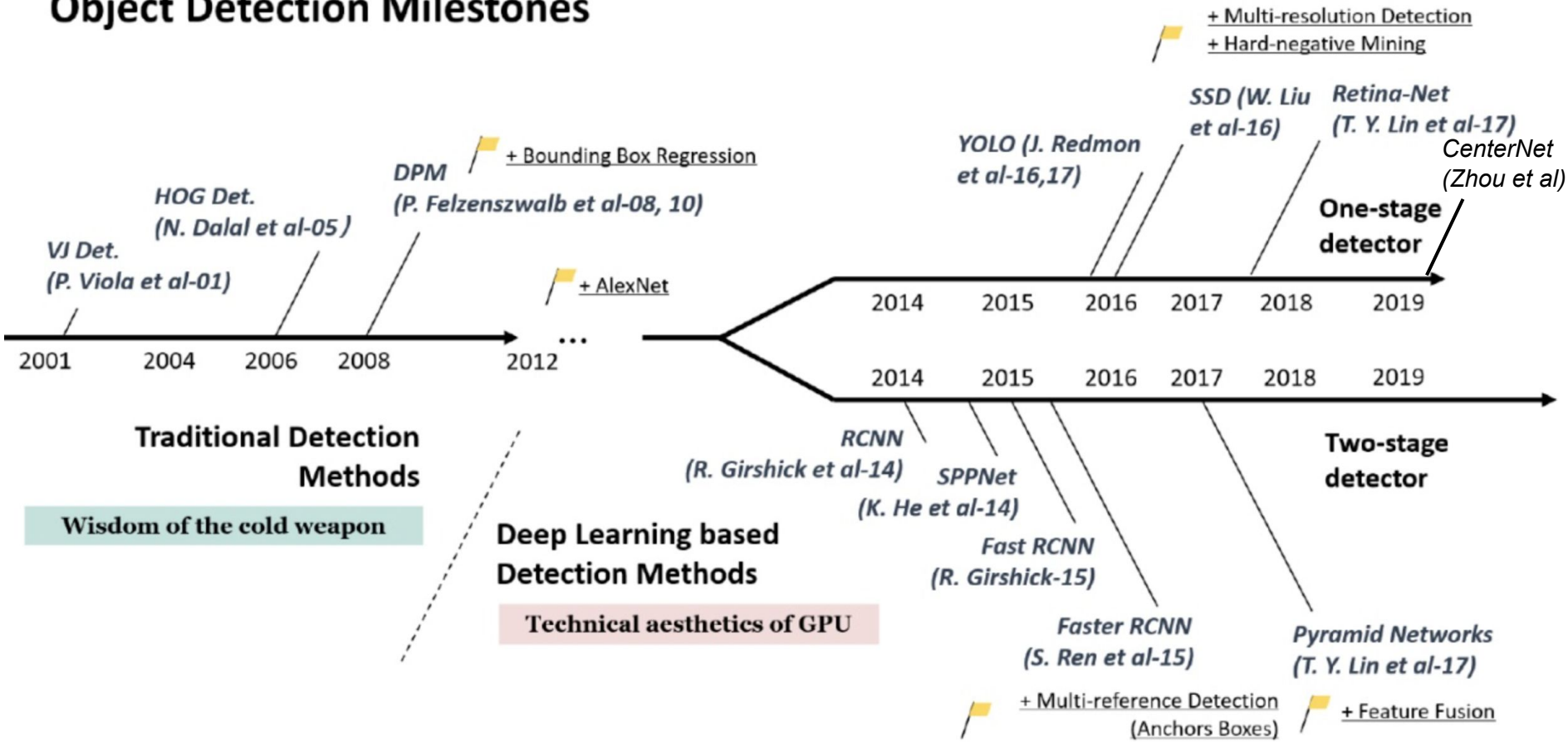
Instance Segmentation

DOG, DOG, CAT

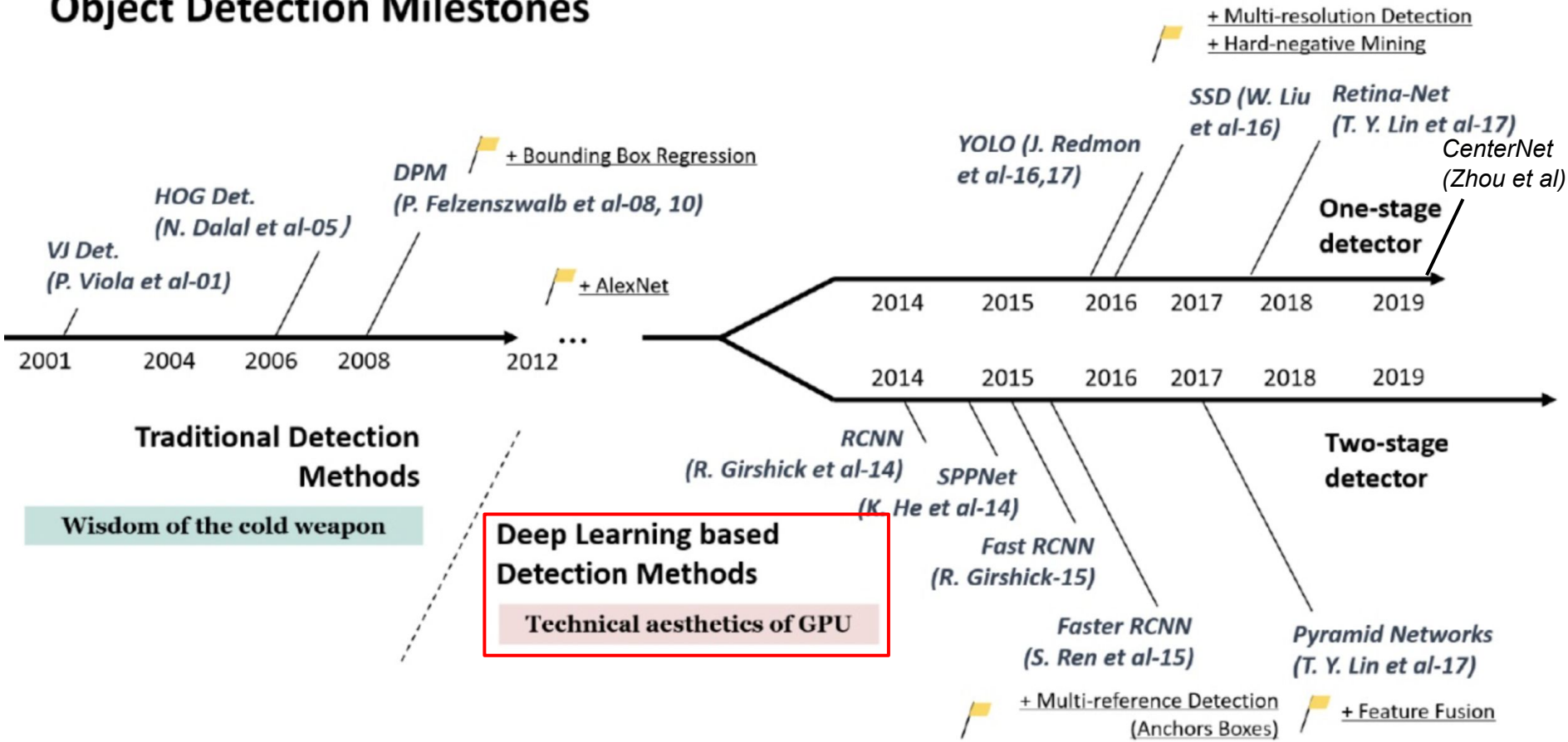
Multiple Object

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

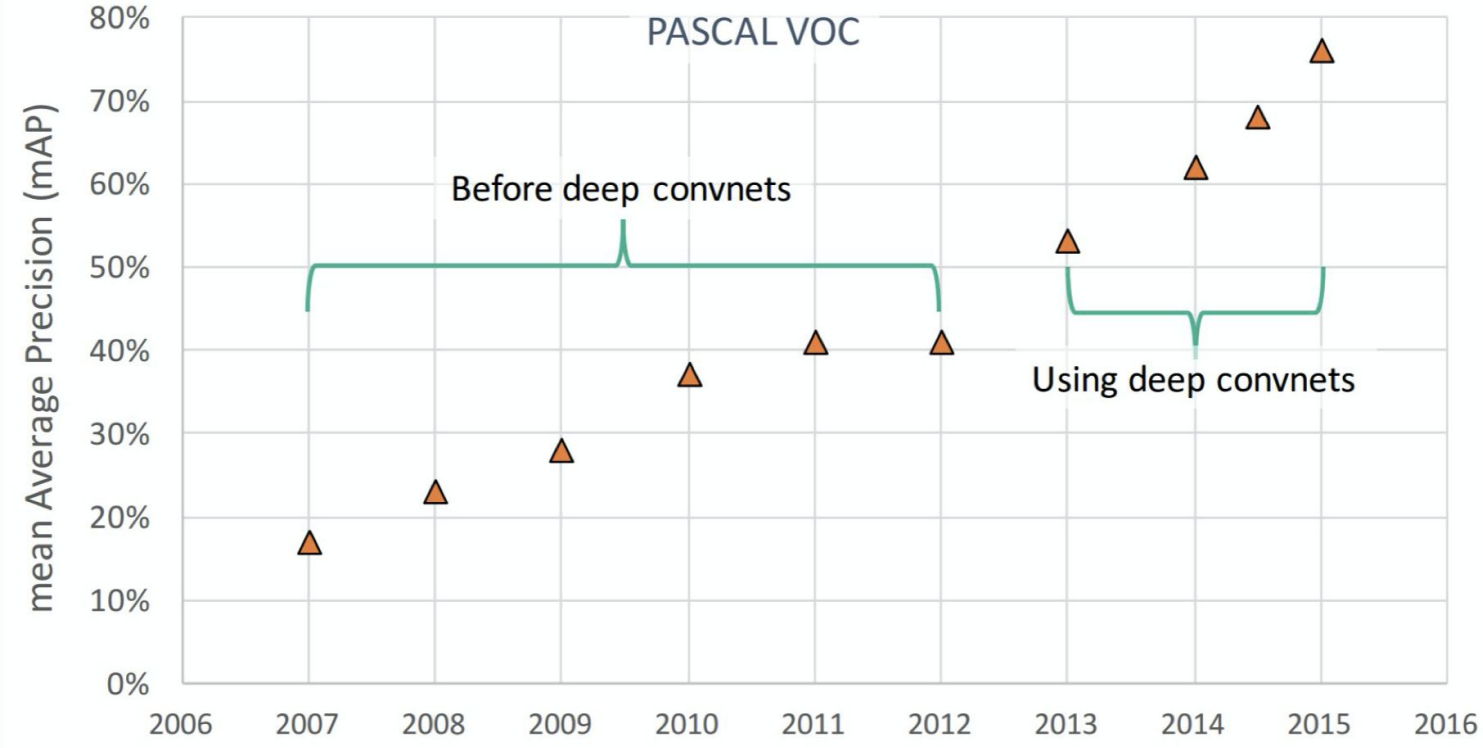
Object Detection Milestones



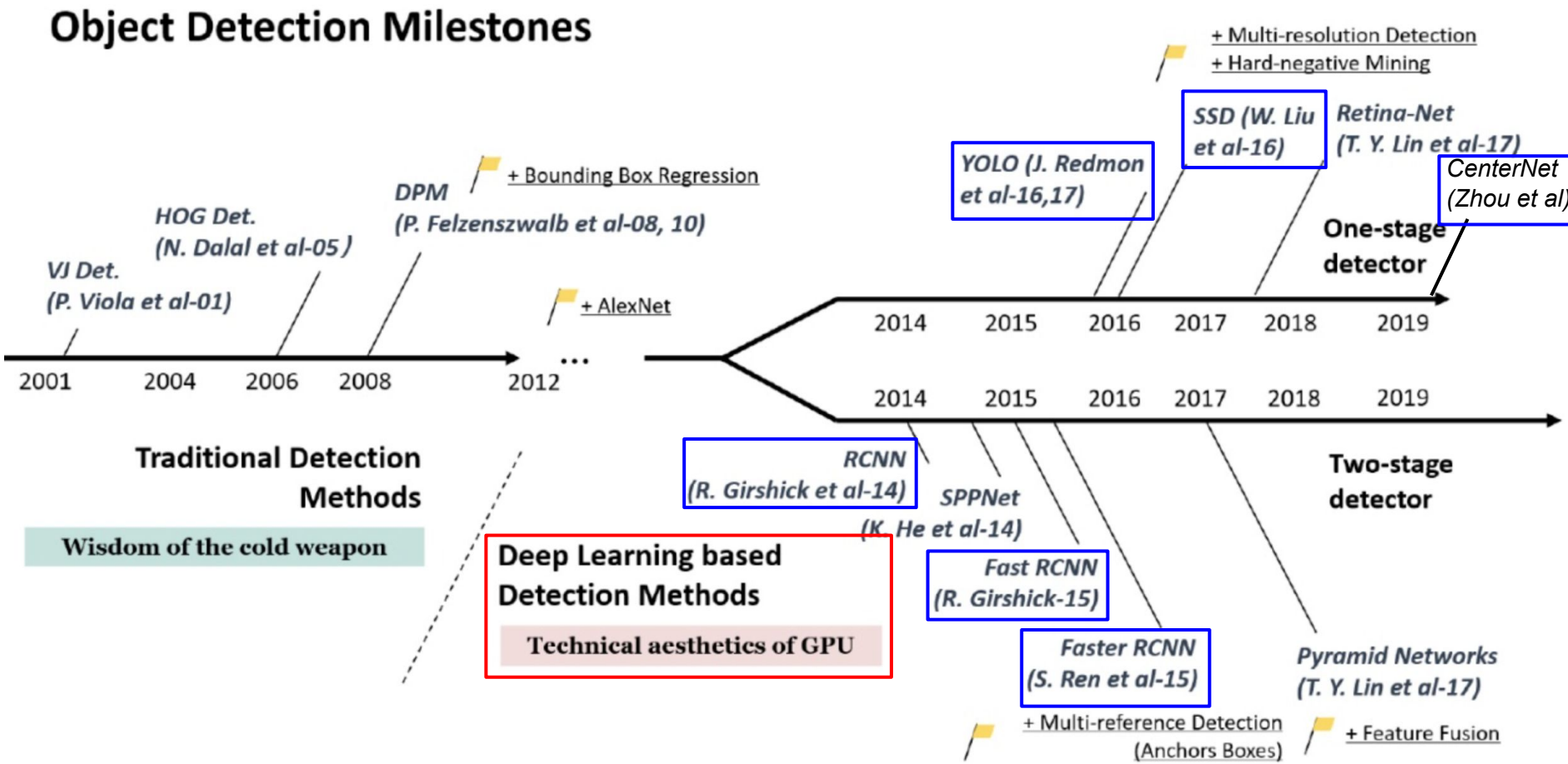
Object Detection Milestones



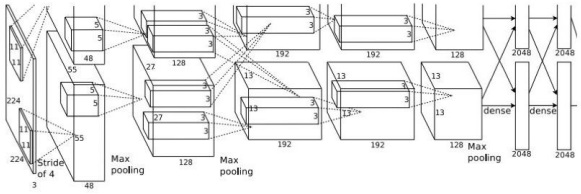
Object Detection: Deep Learning



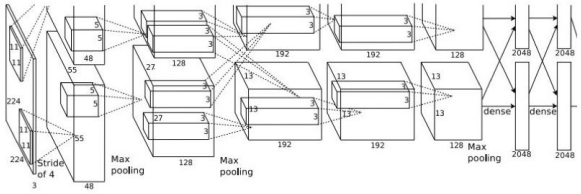
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf


Object Detection Milestones
+ Multi-resolution Detection
+ Hard-negative Mining
SSD (W. Liu et al-16)
Retina-Net (T. Y. Lin et al-17)
YOLO (J. Redmon et al-16,17)
CenterNet (Zhou et al)
+ Bounding Box Regression
DPM (P. Felzenszwalb et al-08, 10)
HOG Det. (N. Dalal et al-05)
VJ Det. (P. Viola et al-01)
+ AlexNet
One-stage detector
2001
2004
2006
2008
2012
2014
2015
2016
2017
2018
2019
Traditional Detection Methods
Wisdom of the cold weapon
RCNN (R. Girshick et al-14)
SPPNet (K. He et al-14)
Fast RCNN (R. Girshick-15)
Faster RCNN (S. Ren et al-15)
Pyramid Networks (T. Y. Lin et al-17)
Two-stage detector
Deep Learning based Detection Methods
Technical aesthetics of GPU
+ Multi-reference Detection (Anchors Boxes)
+ Feature Fusion

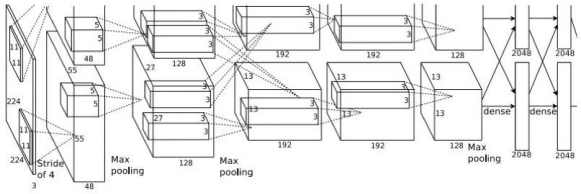
Object Detection: Representation



CAT: (x, y, w, h)



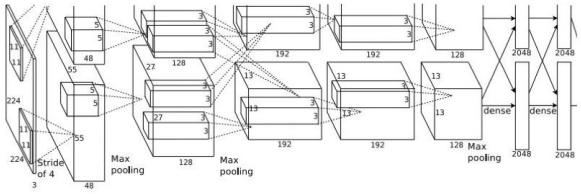
DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)



DUCK: (x, y, w, h)
DUCK: (x, y, w, h)
….

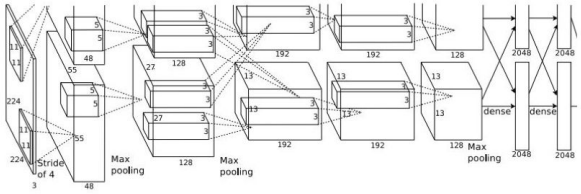
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Object Detection: Representation

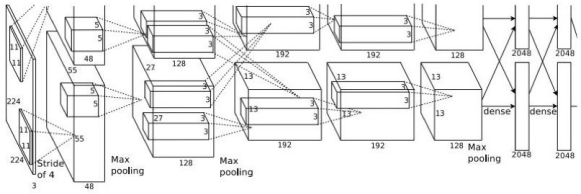


CAT: (x, y, w, h)

Regression



DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)

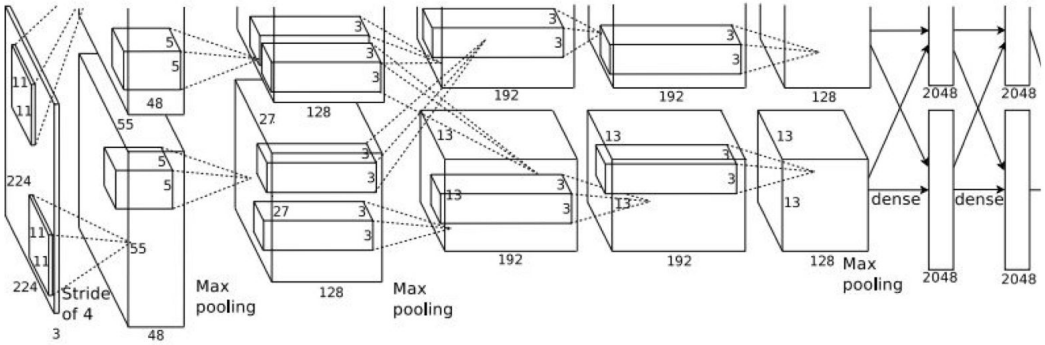


DUCK: (x, y, w, h)
DUCK: (x, y, w, h)
….

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Object Detection: Sliding Window

Apply a CNN to classify many crops of the image

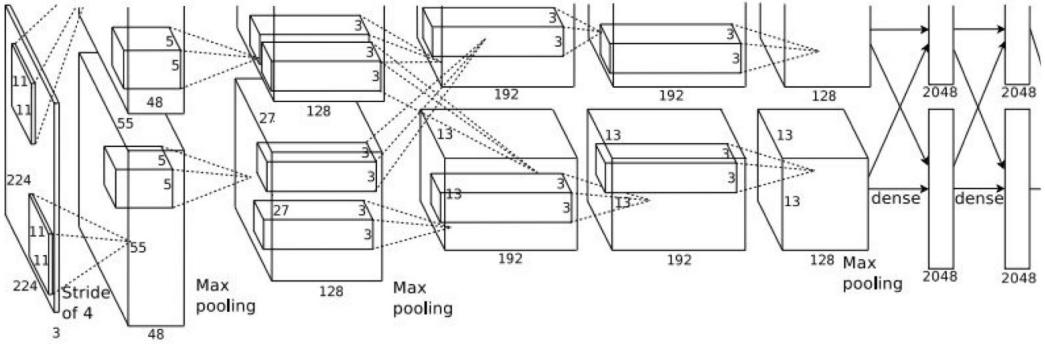


Dog? No
Cat? No
Background? Yes

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Object Detection: Sliding Window

Apply a CNN to classify many crops of the image

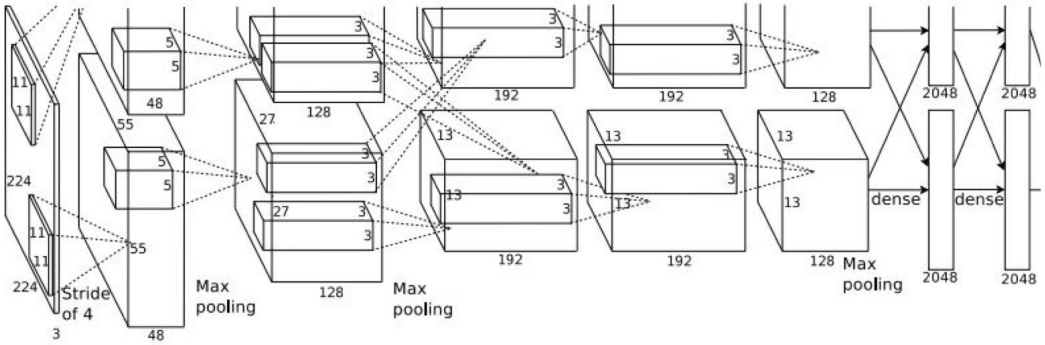


Dog? No
Cat? No
Background? Yes

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Object Detection: Sliding Window

Apply a CNN to classify many crops of the image

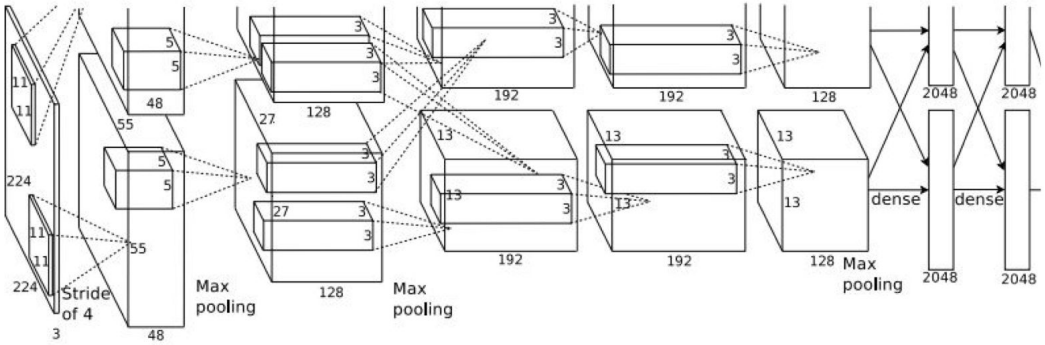


Dog? Yes
Cat? No
Background? No

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Object Detection: Sliding Window

Apply a CNN to classify many crops of the image

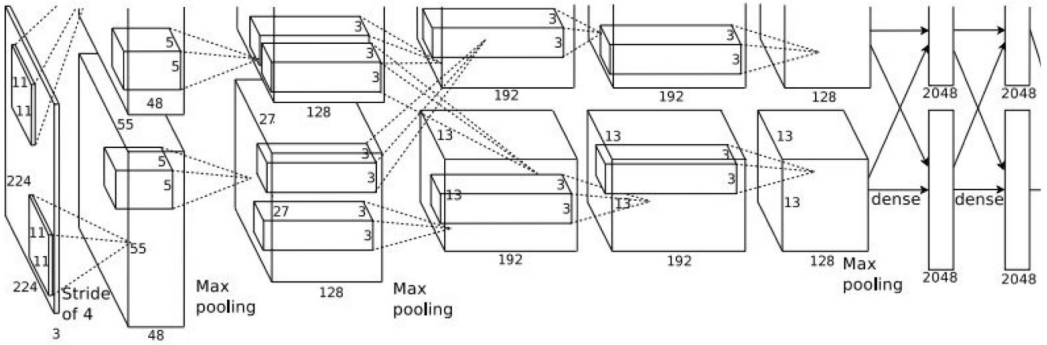


Dog? No
Cat? Yes
Background? No

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Object Detection: Sliding Window

Apply a CNN to classify many crops of the image



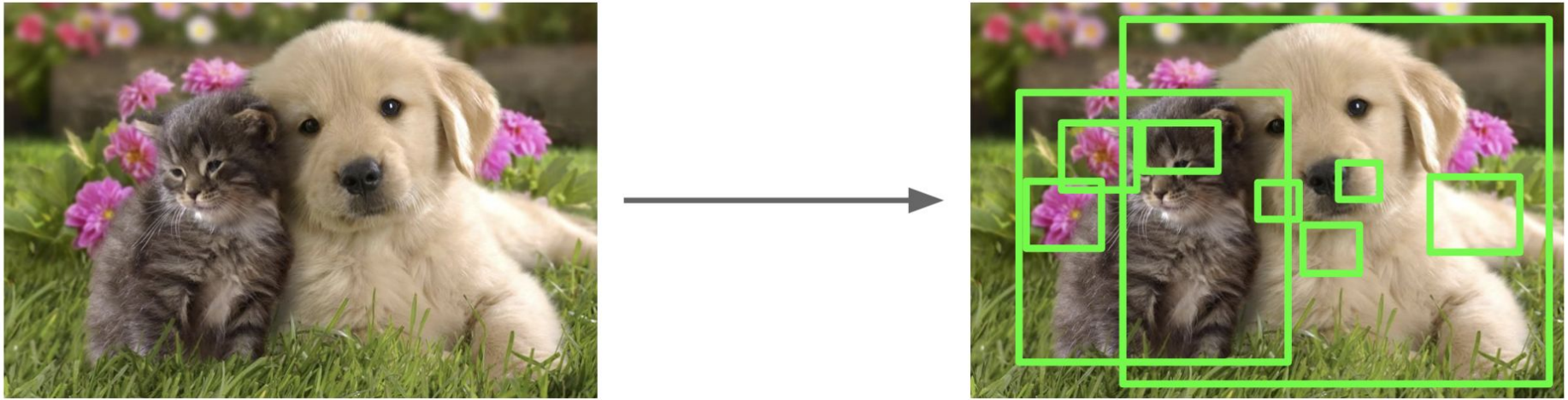
Dog? No
Cat? Yes
Background? No

Expensive: huge number of crops (of different locations and scales)

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

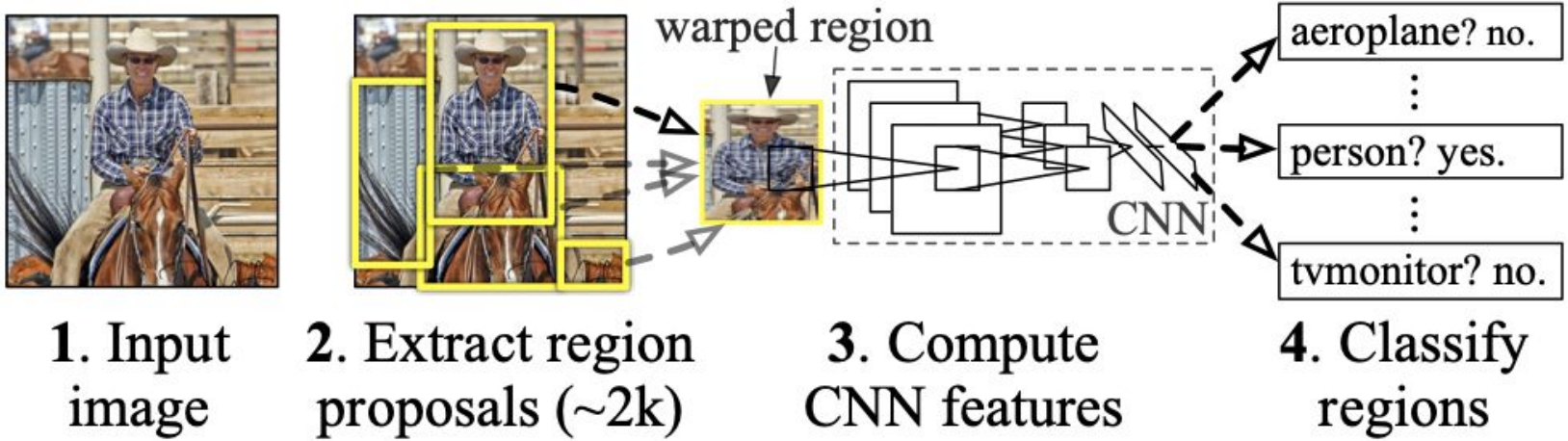
Object Detection: Region Proposal

Quickly find regions which potentially contain objects (Selective Search)



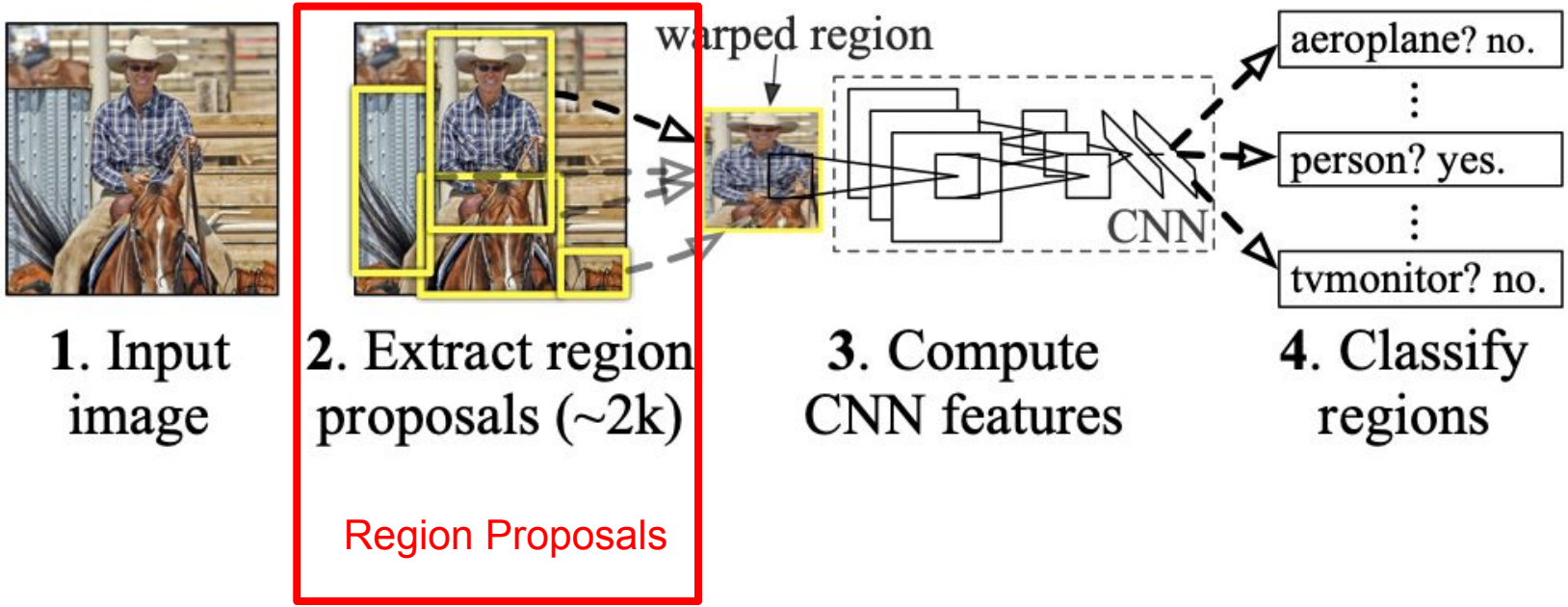
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

R-CNN: *Regions with CNN features*



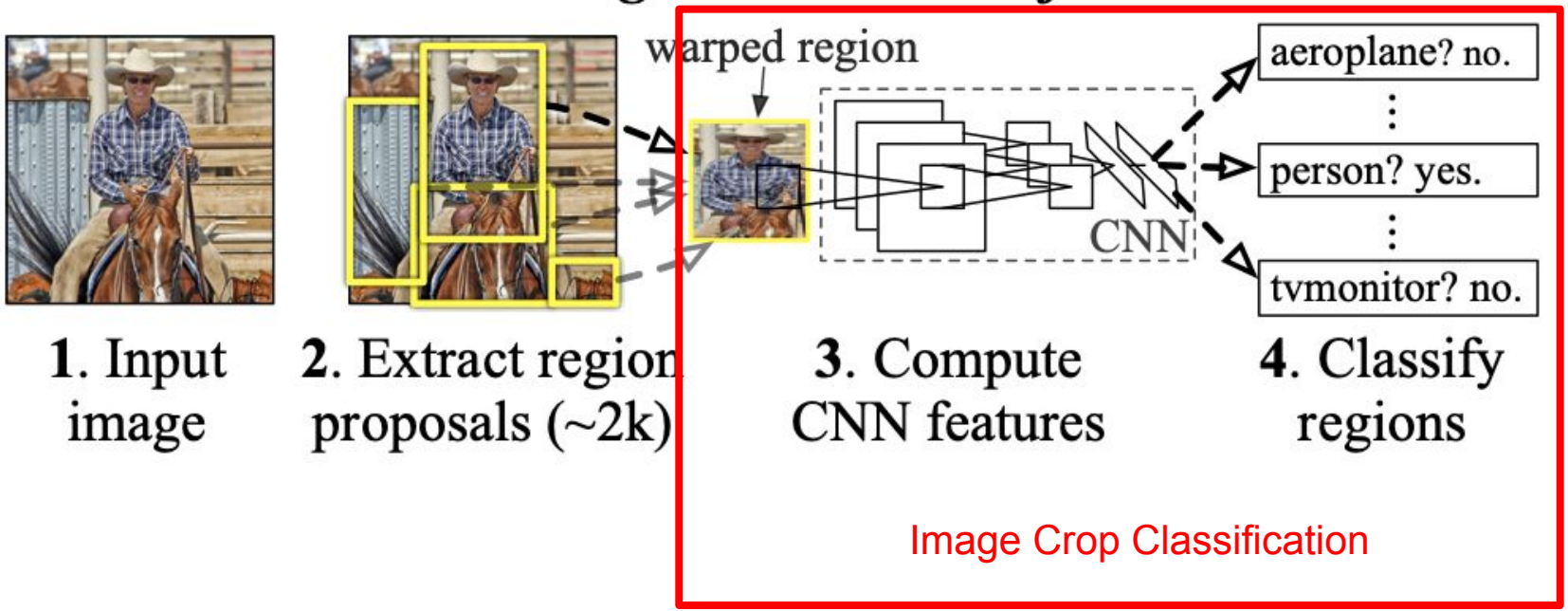
Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

R-CNN: *Regions with CNN features*



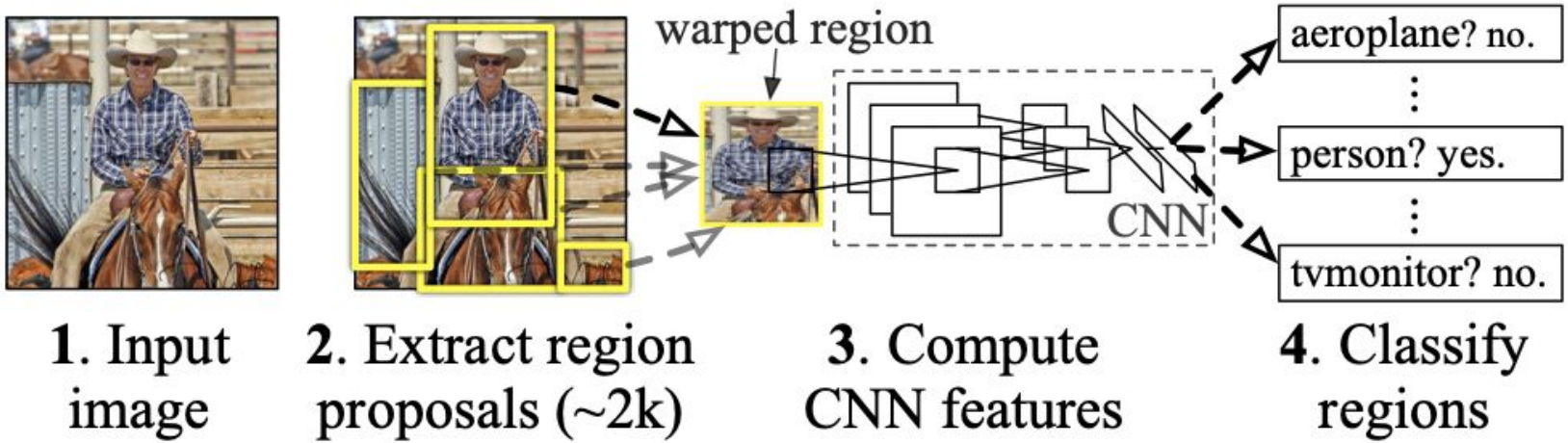
Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

R-CNN: *Regions with CNN features*



Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

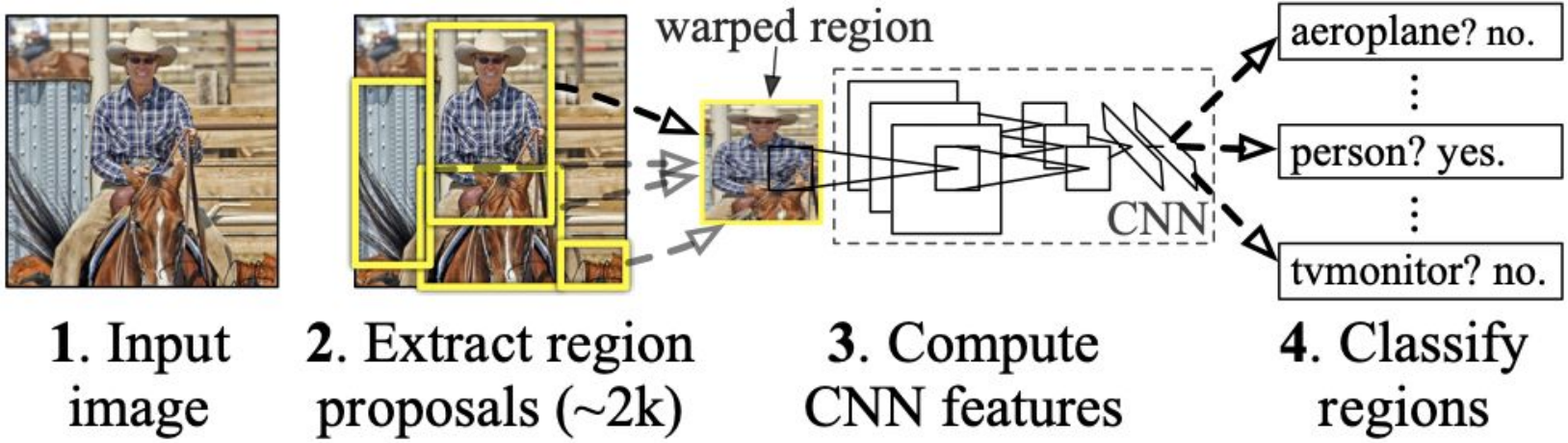
R-CNN: *Regions with CNN features*



- Train post-hoc bounding-box regressions (least squares)
- Training is slow (84h), takes a lot of disk space
- Testing is slow

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

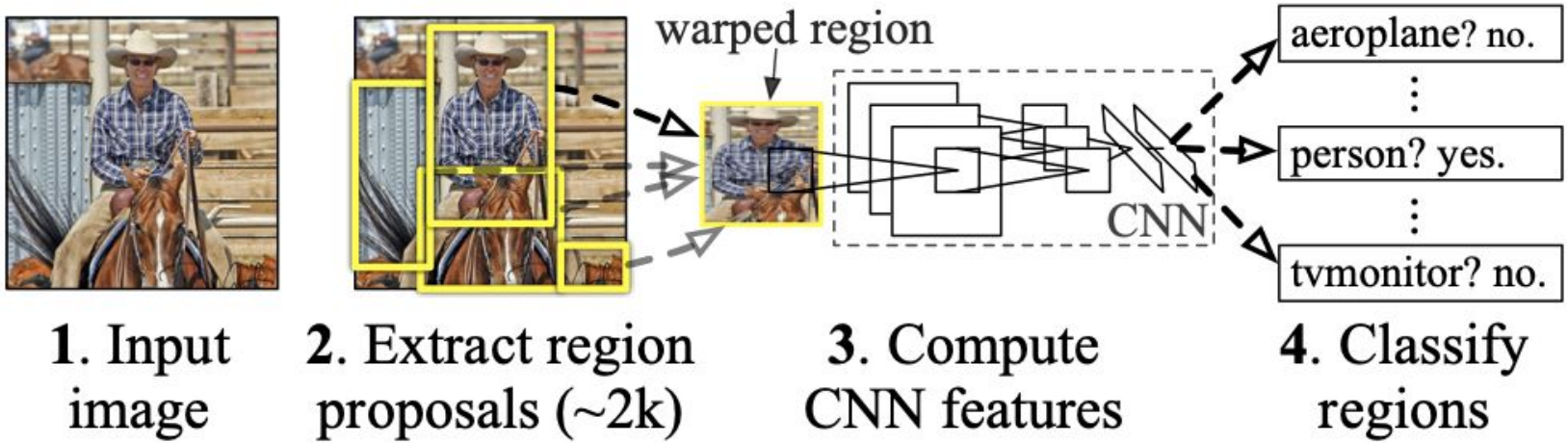
R-CNN: *Regions with CNN features*



Why is RCNN slow at inference time?

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

R-CNN: *Regions with CNN features*

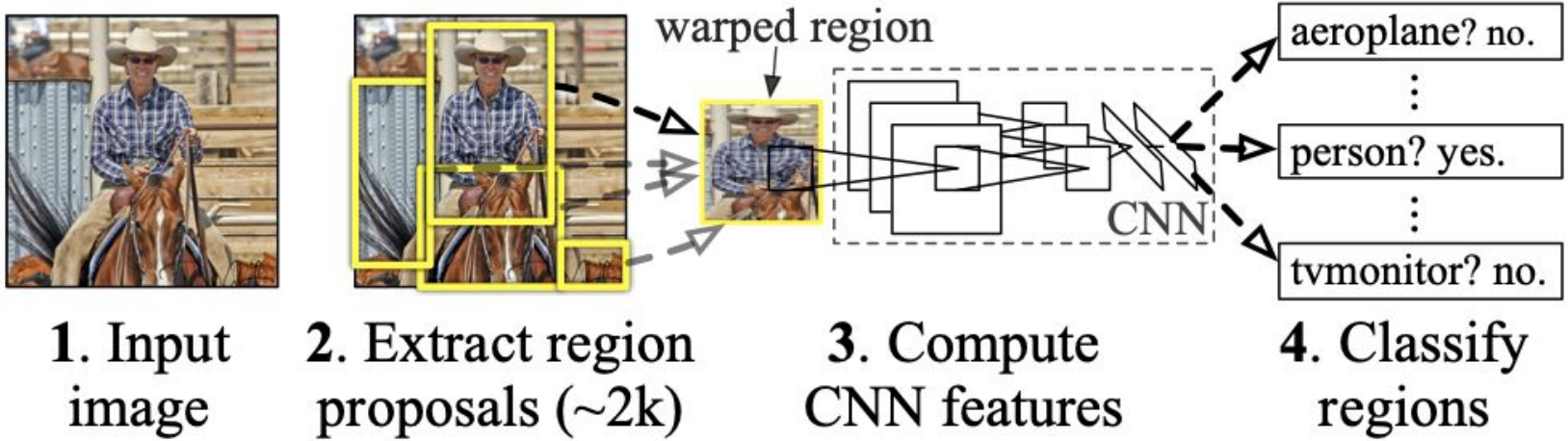


Why is RCNN slow at inference time?

- Every image crop goes through the CNN sequentially

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

R-CNN: *Regions with CNN features*



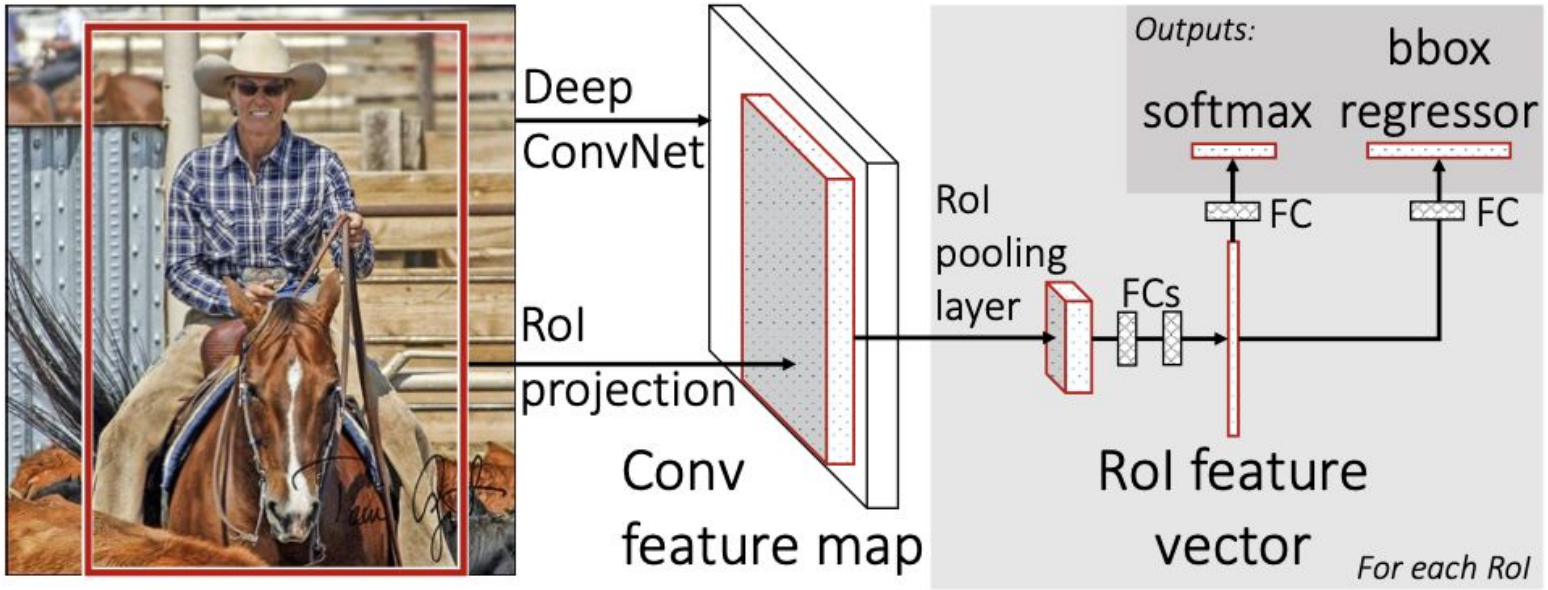
Why is RCNN slow at inference time?

- Every image crop goes through the CNN sequentially

Is it possible to process the image only once and get the features for all image crops?

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

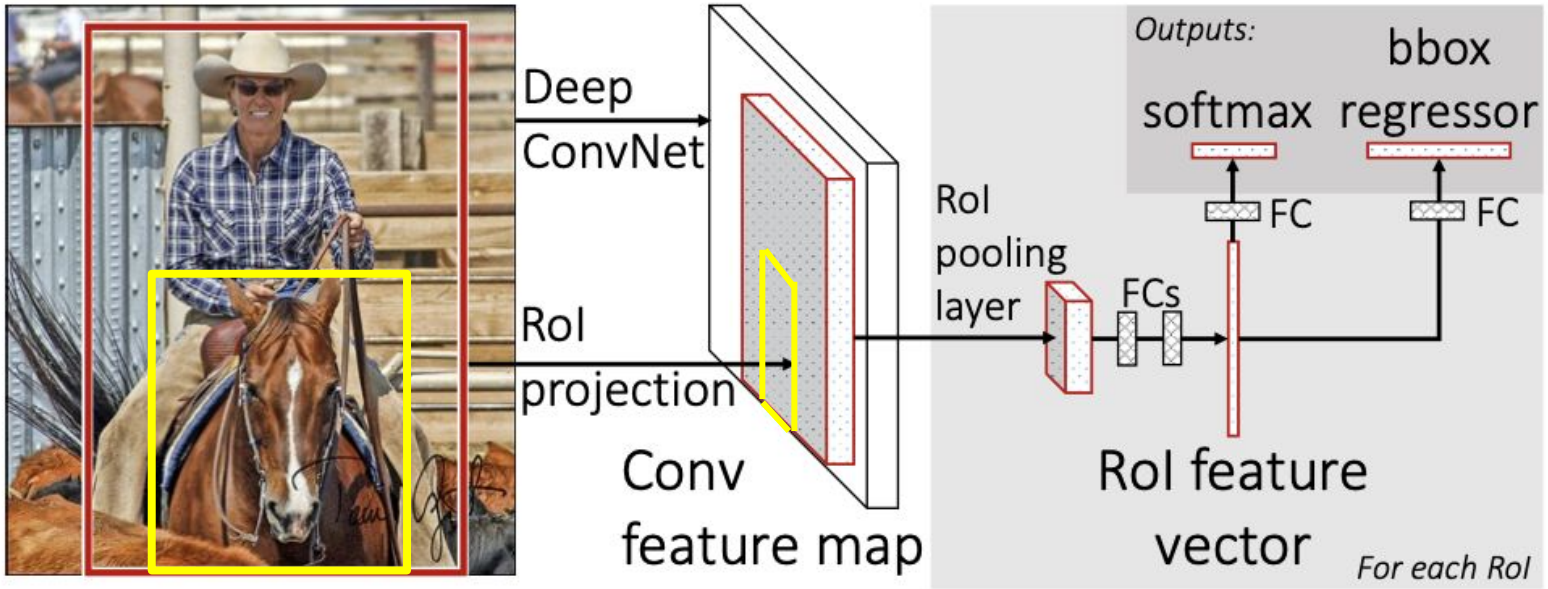
Object Detection: Fast RCNN



Share computation between different RoI (region of interest)!

Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

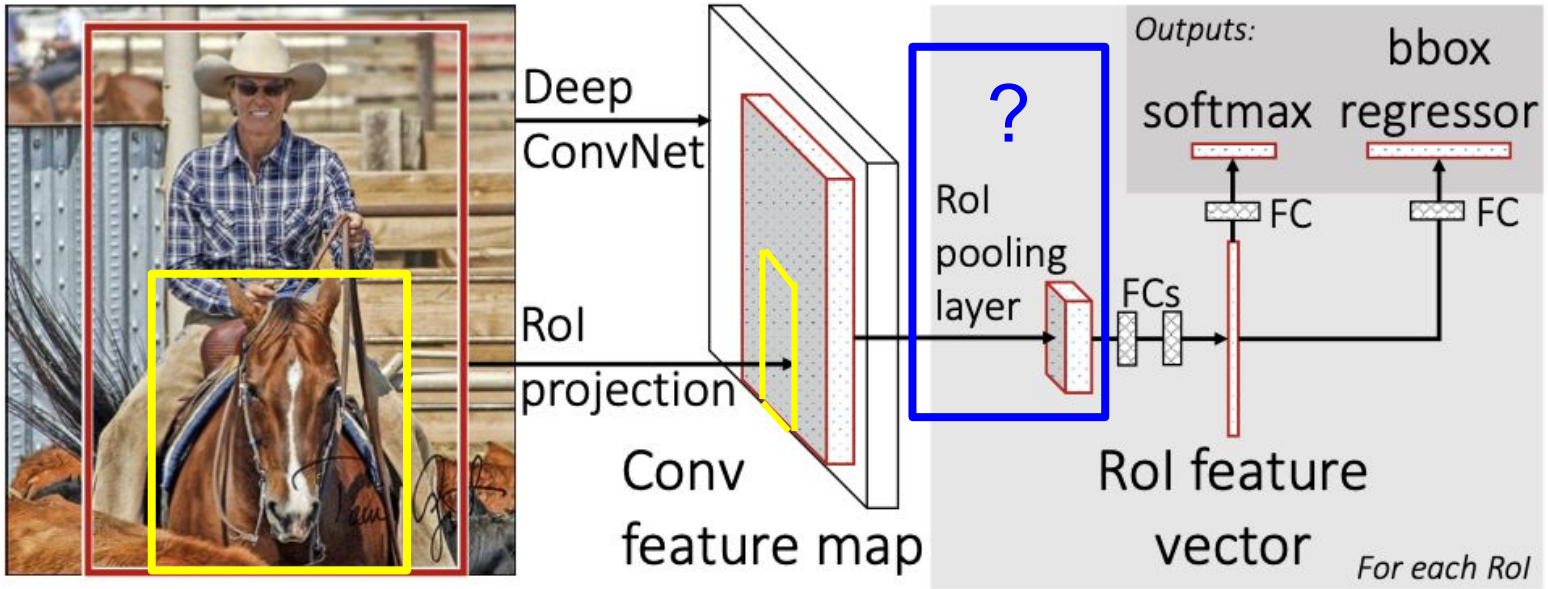
Object Detection: Fast RCNN



Share computation between different RoI (region of interest)!

Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

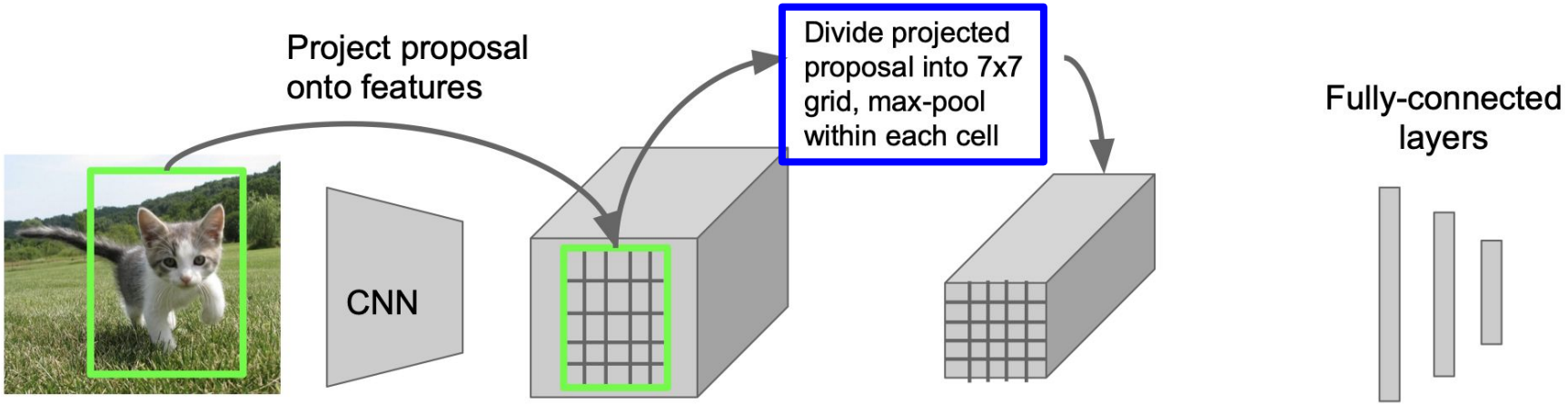
Object Detection: Fast RCNN



Share computation between different RoI (region of interest)!

Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

Faster R-CNN: RoI Pooling



Hi-res input image:
3 x 640 x 480
with region proposal

Hi-res conv features:
512 x 20 x 15;

Projected region proposal is e.g.
512 x 18 x 8
(varies per proposal)

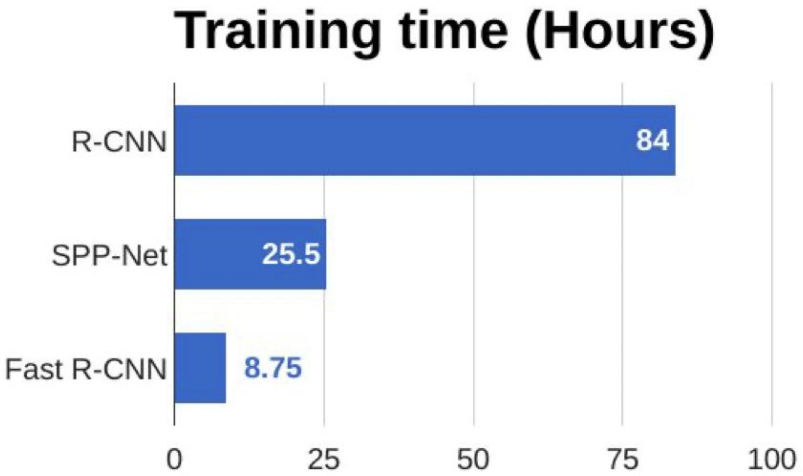
RoI conv features:
512 x 7 x 7
for region proposal

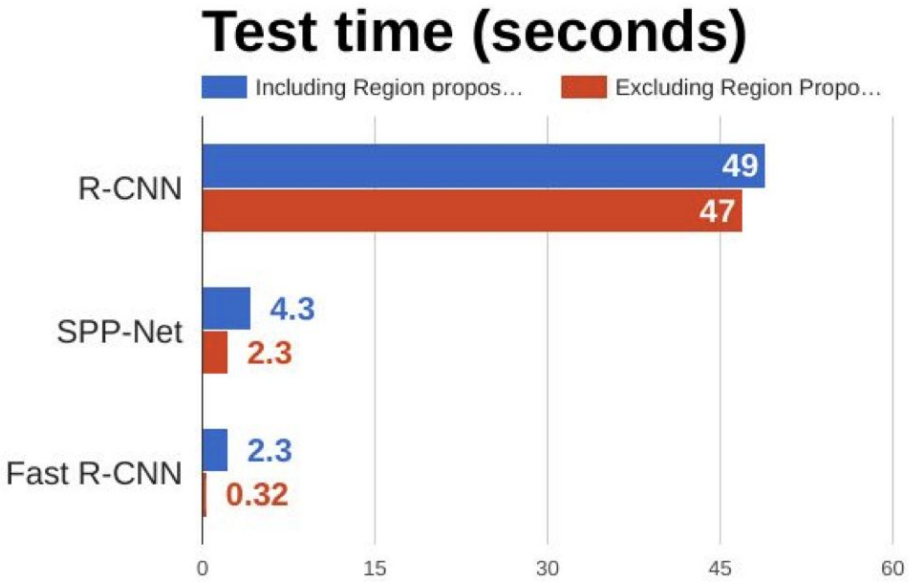
Fully-connected layers expect low-res conv features:
512 x 7 x 7

Girshick, "Fast R-CNN", ICCV 2015.

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

R-CNN vs SPP vs Fast R-CNN

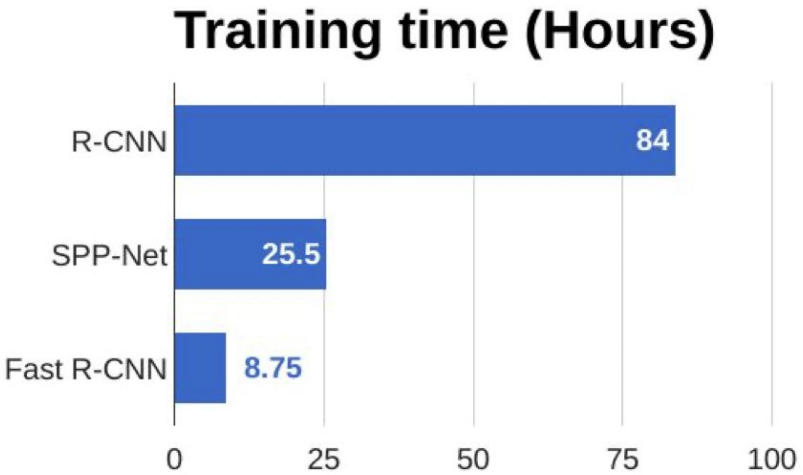


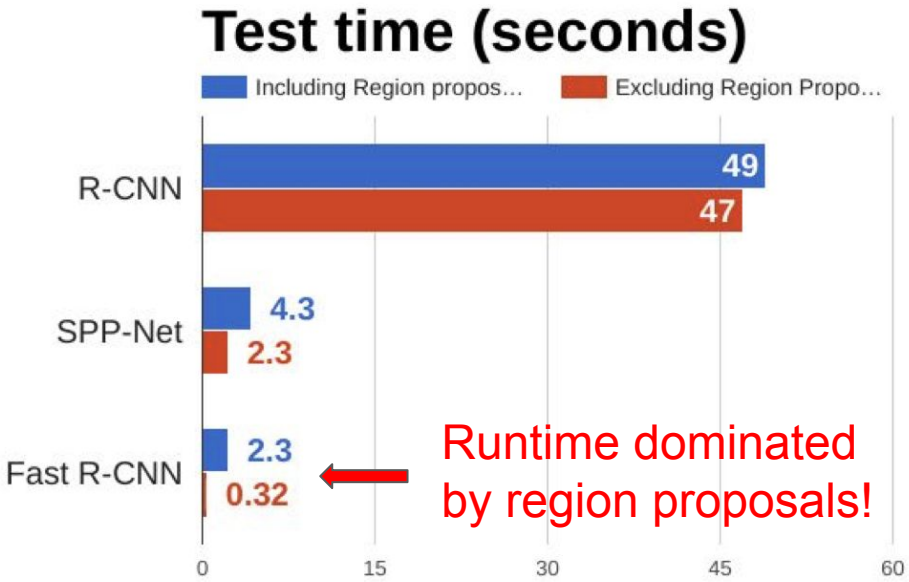


Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
He et al, “Spatial pyramid pooling in deep convolutional networks for visual recognition”, ECCV 2014
Girshick, “Fast R-CNN”, ICCV 2015

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

R-CNN vs SPP vs Fast R-CNN





Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
He et al, "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV 2014
Girshick, "Fast R-CNN", ICCV 2015

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

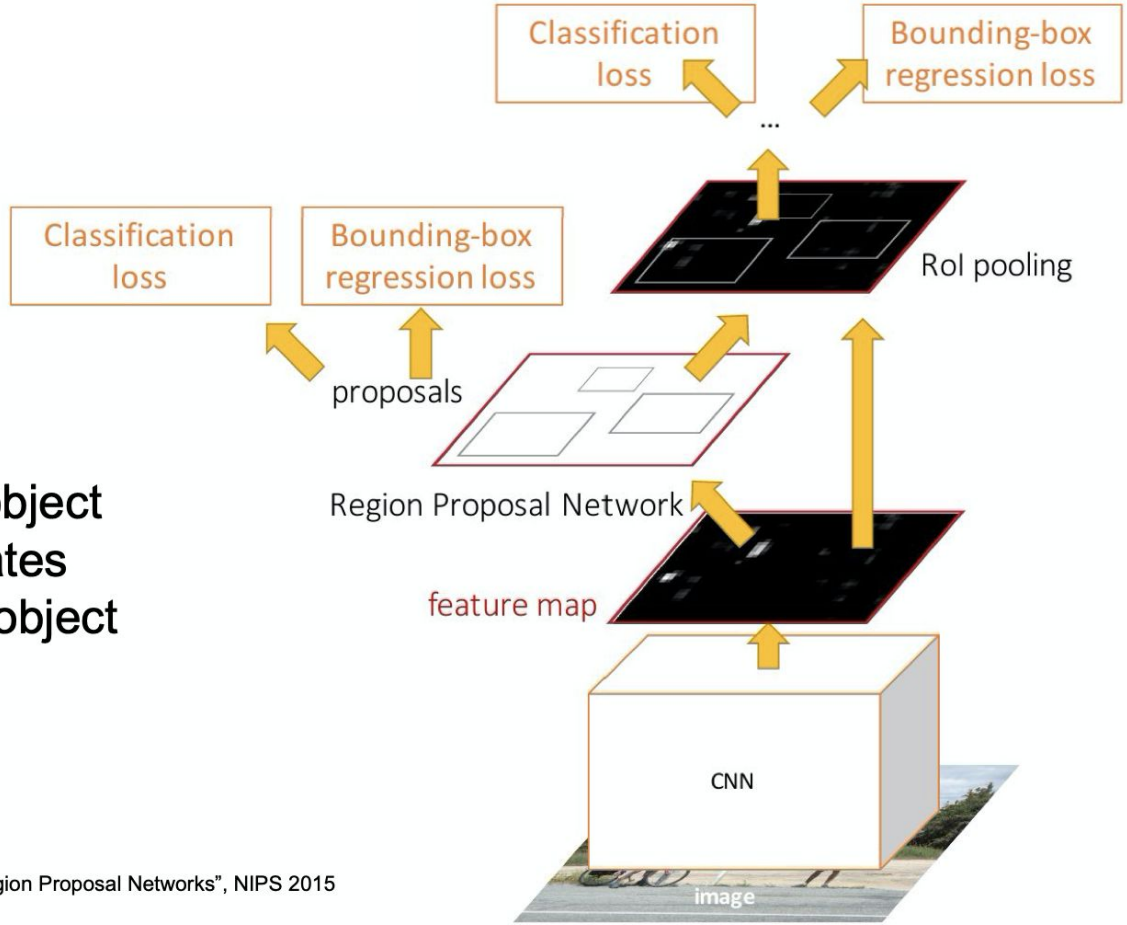
Faster R-CNN:

Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Jointly train with 4 losses:

1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

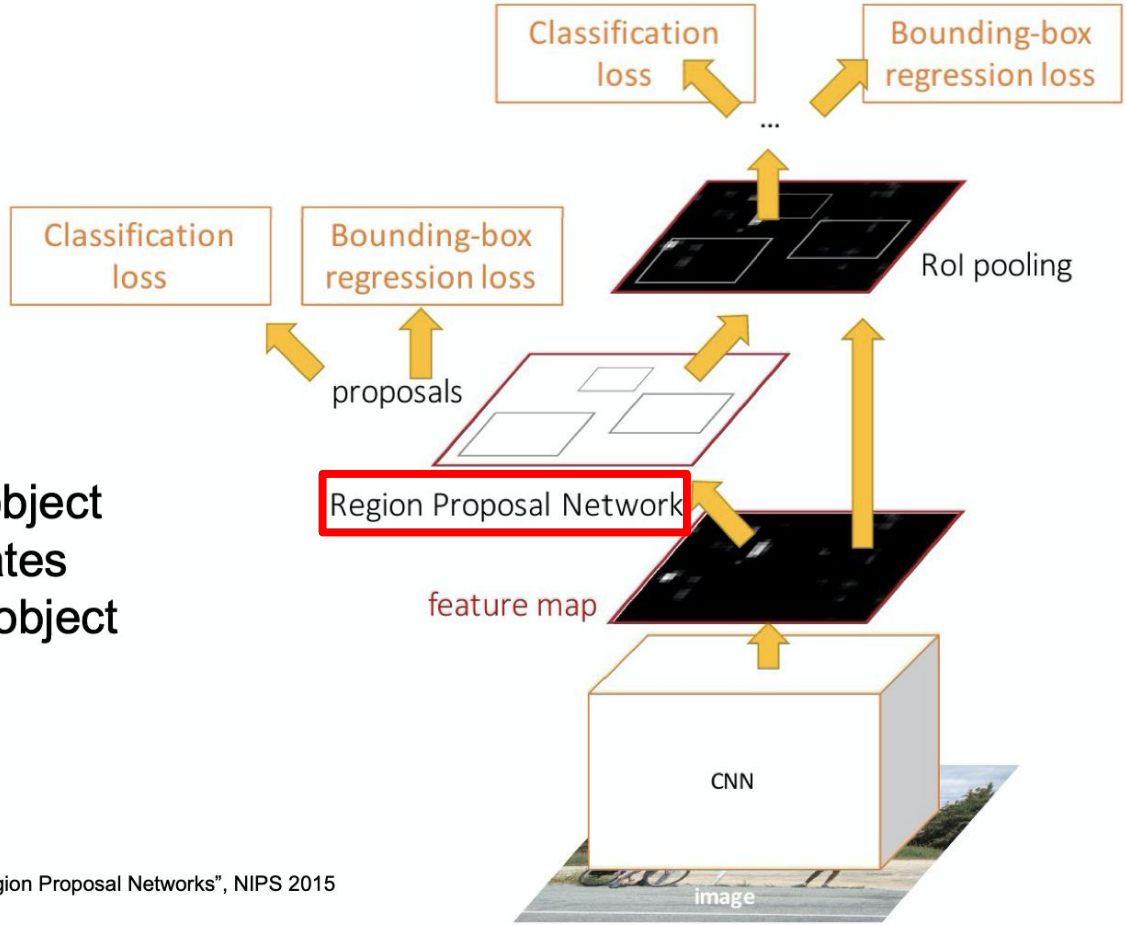
Faster R-CNN:

Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Jointly train with 4 losses:

1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Faster R-CNN:

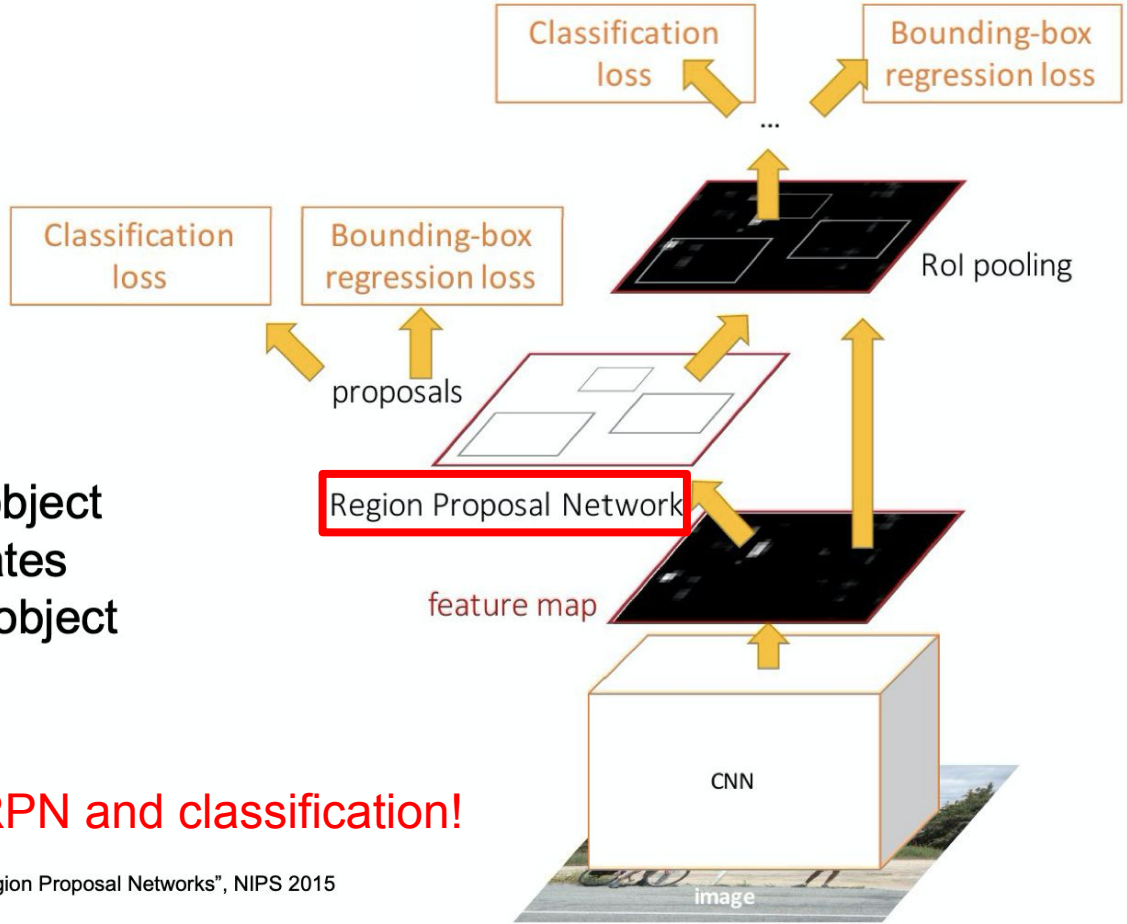
Make CNN do proposals!

Insert **Region Proposal Network (RPN)** to predict proposals from features

Jointly train with 4 losses:

1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates

Share computation between RPN and classification!

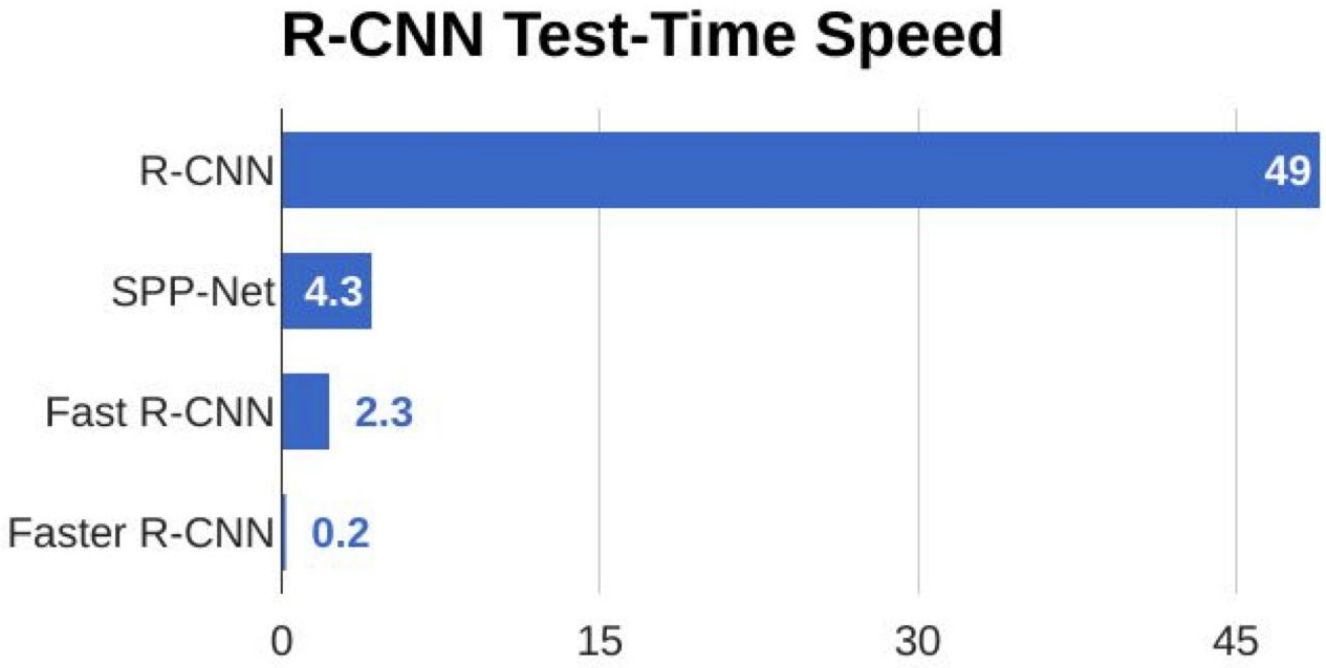


Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Faster R-CNN:

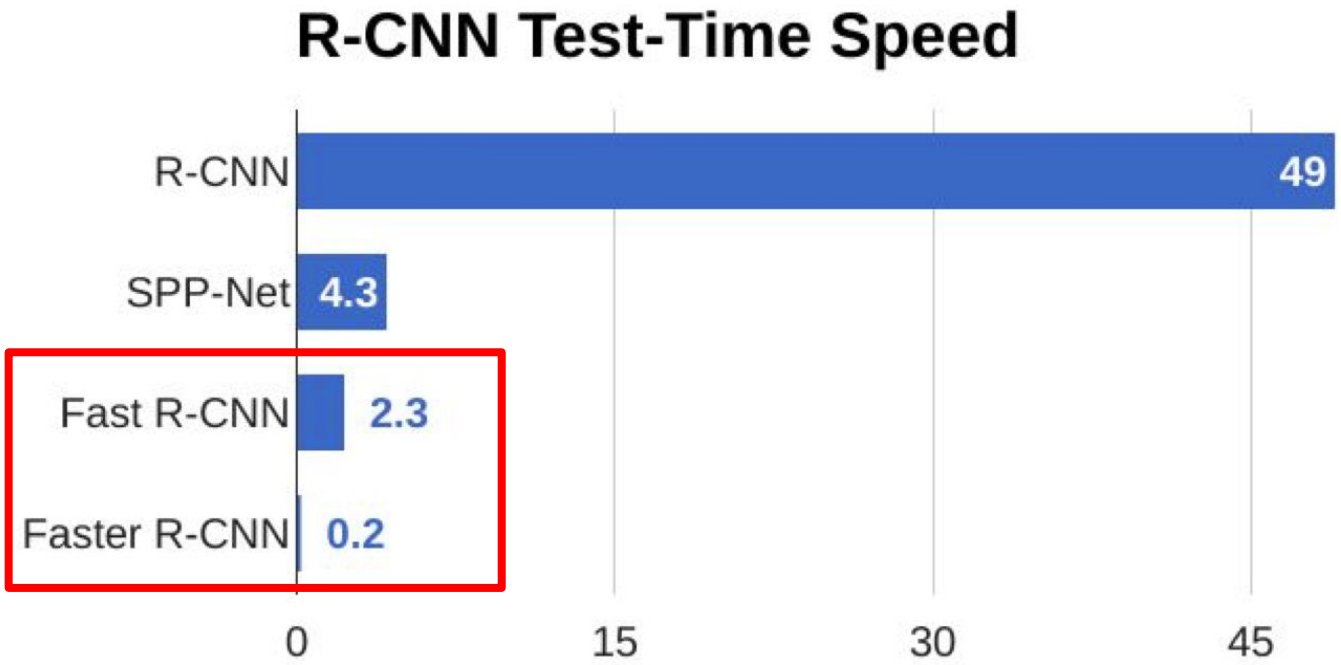
Make CNN do proposals!



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

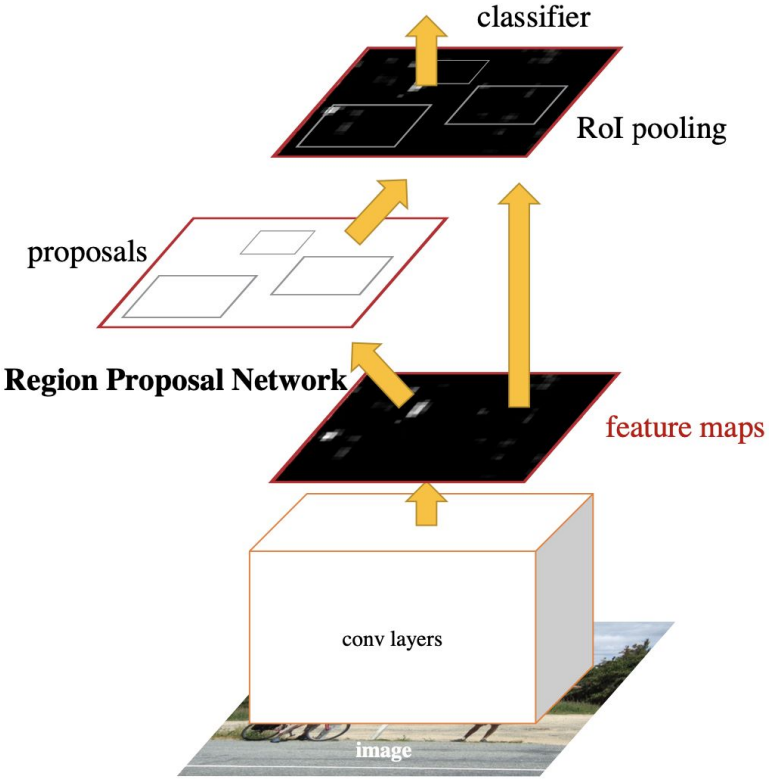
Faster R-CNN:

Make CNN do proposals!



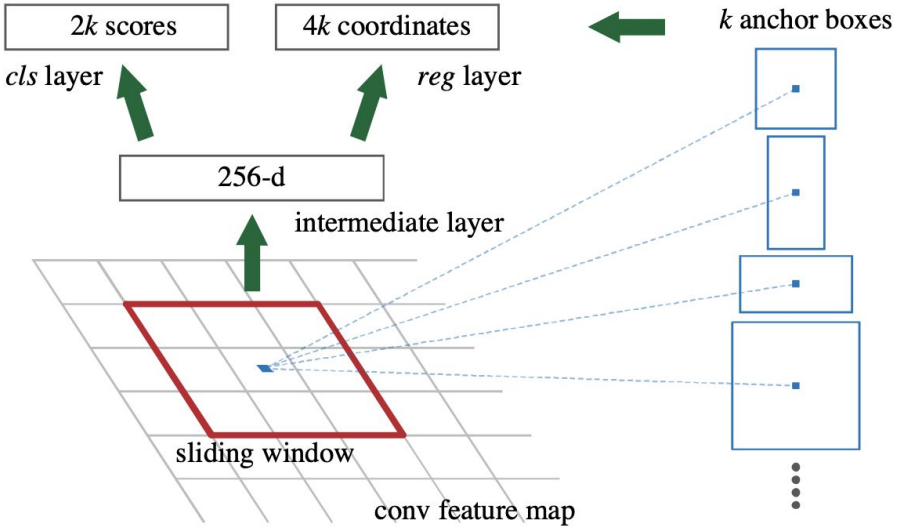
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Faster RCNN (RPN)

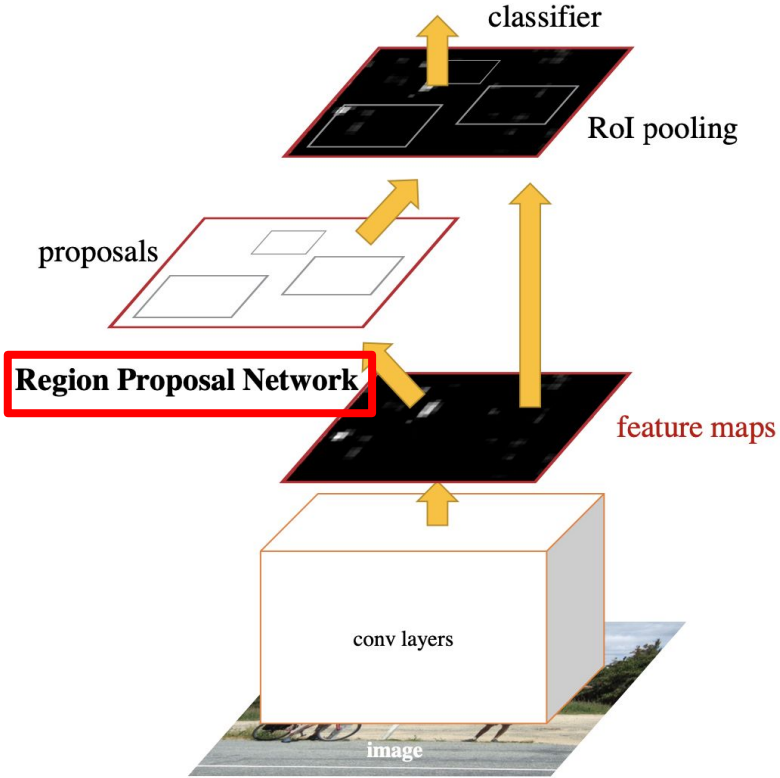


Faster RCNN

Faster RCNN (RPN)

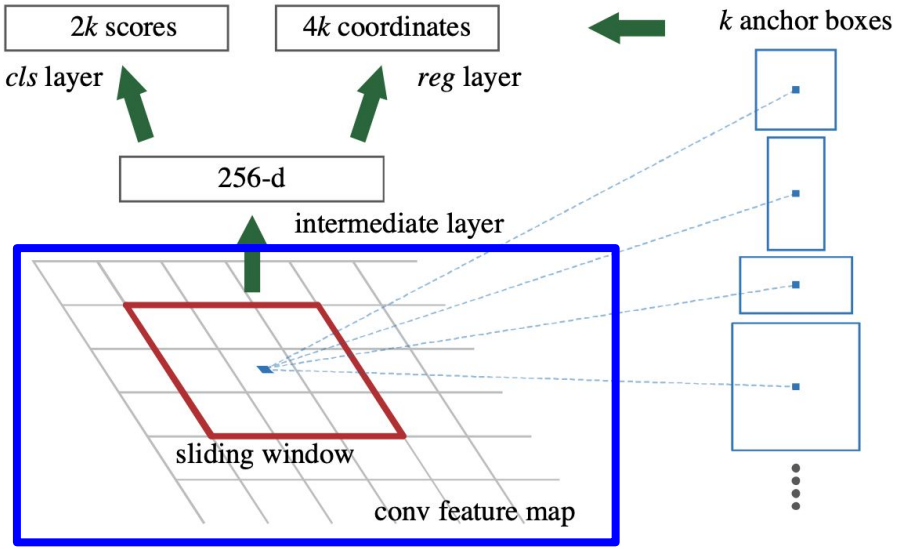


RPN

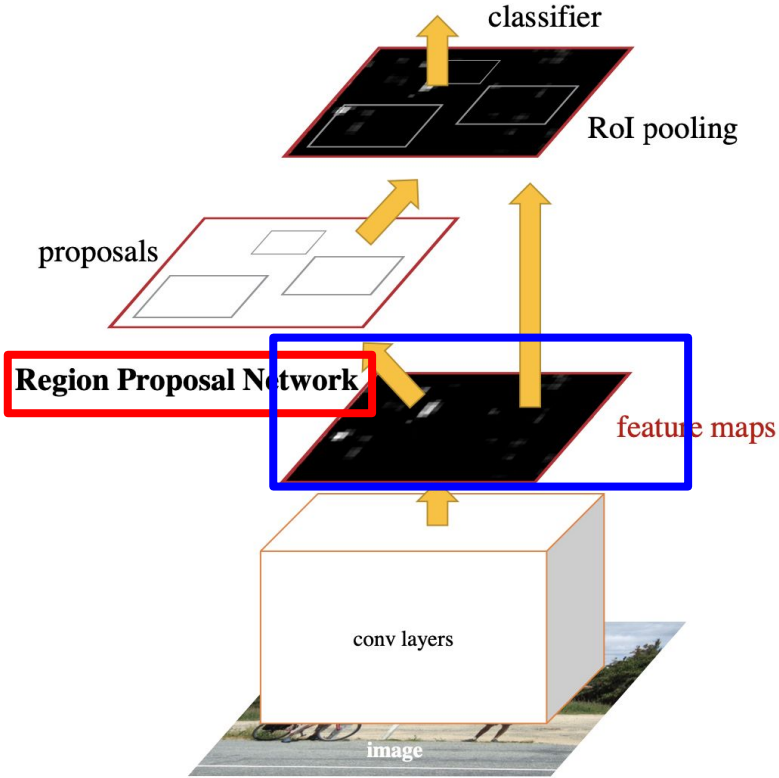


Faster RCNN

Faster RCNN (RPN)

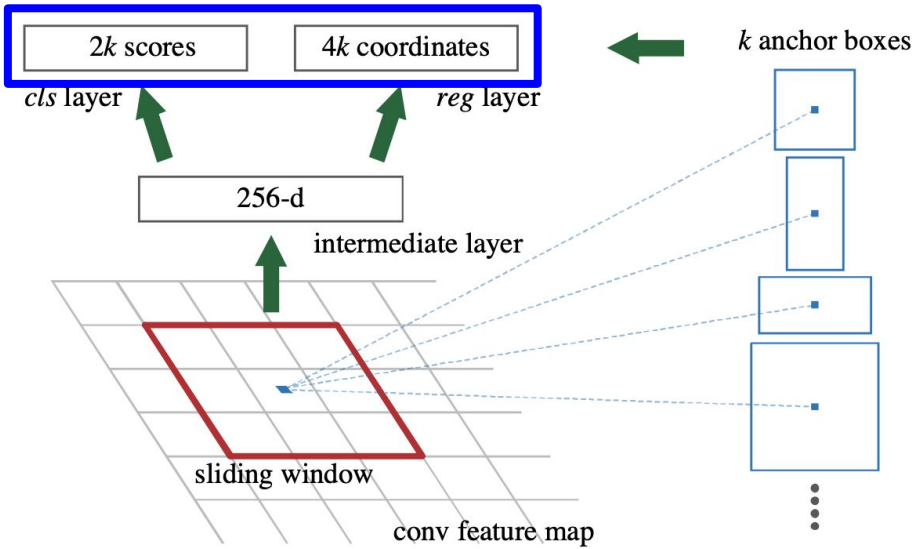


RPN

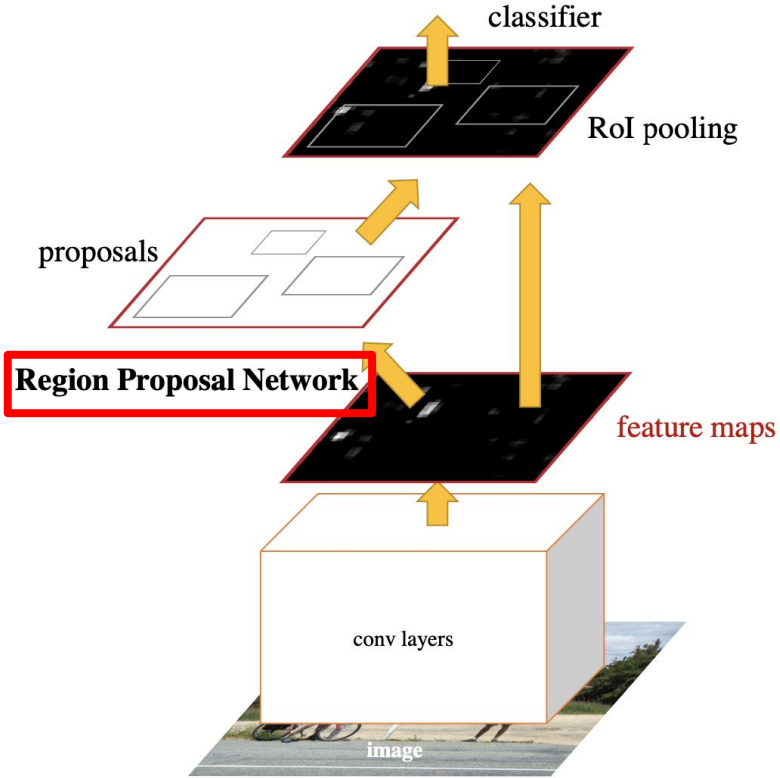


Faster RCNN

Faster RCNN (RPN)

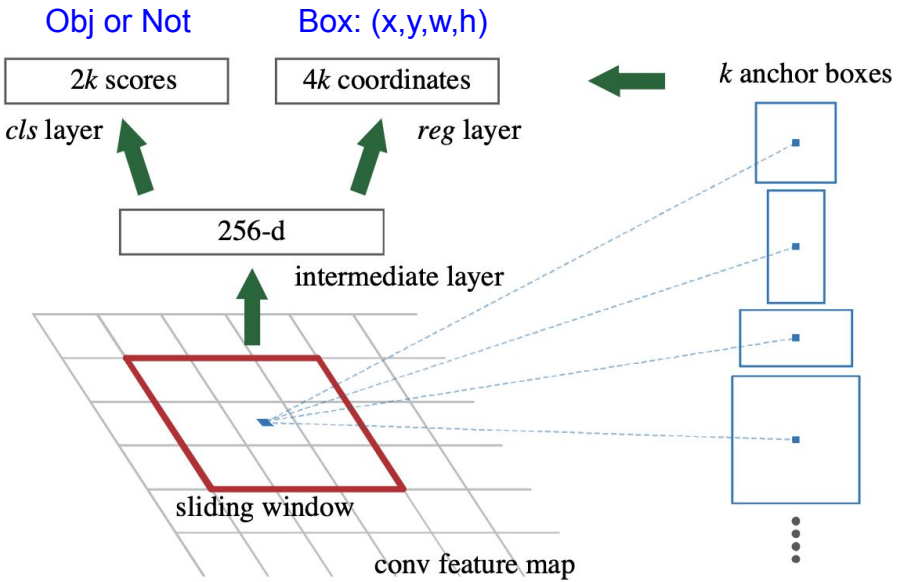


RPN

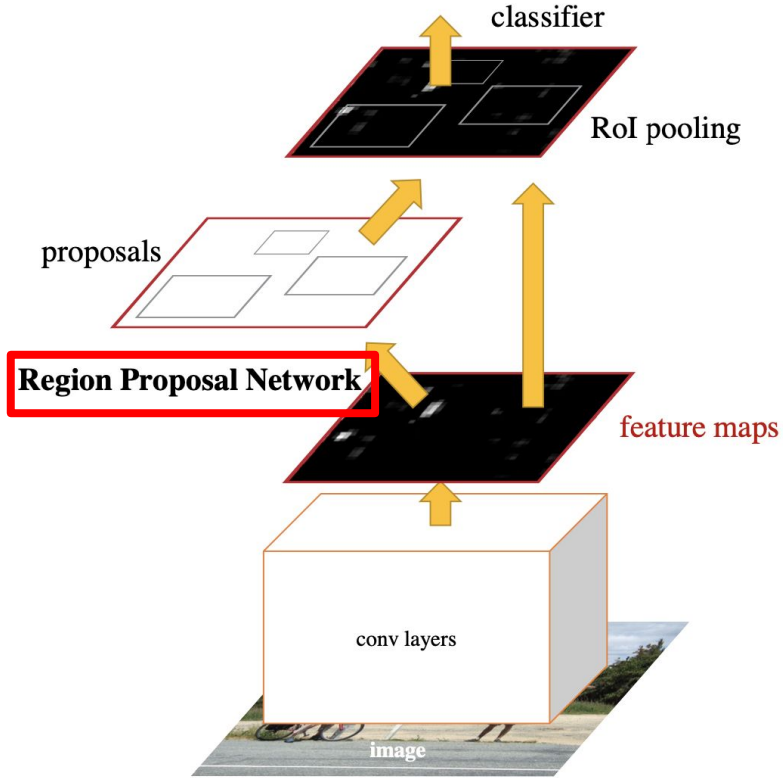


Faster RCNN

Faster RCNN (RPN)

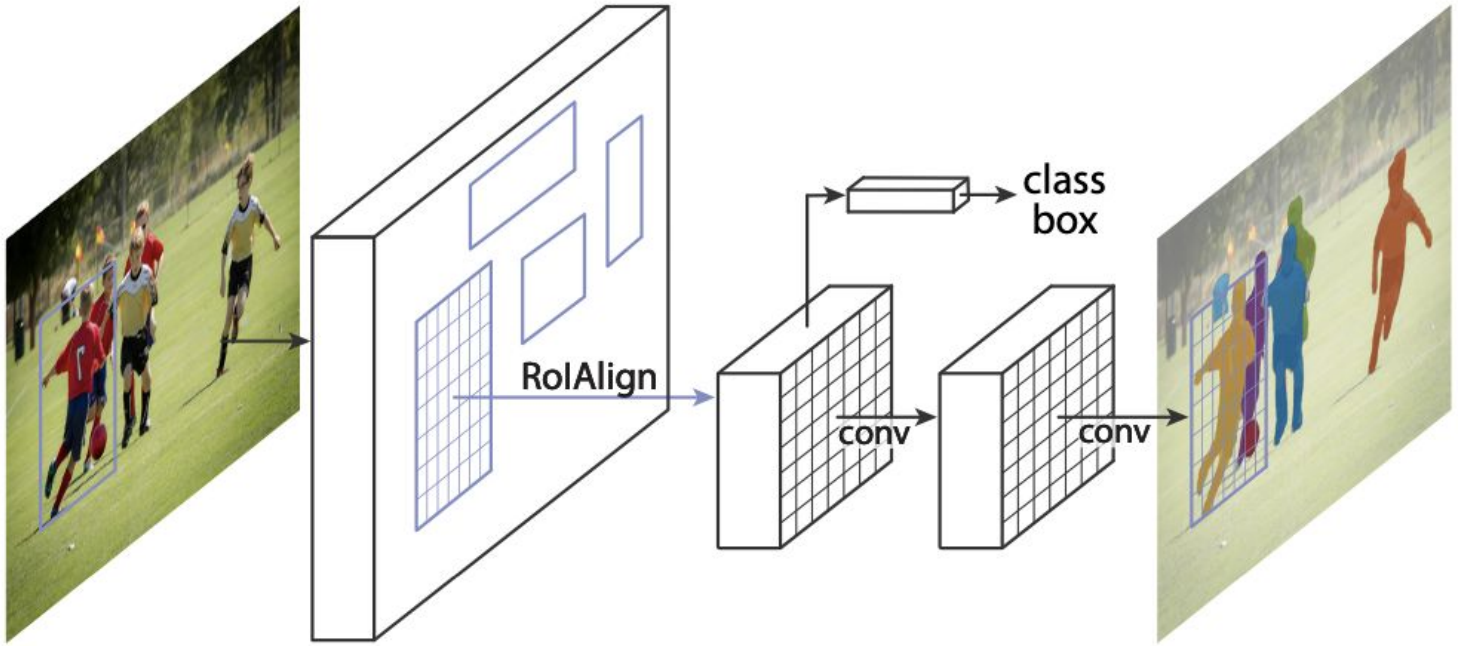


RPN



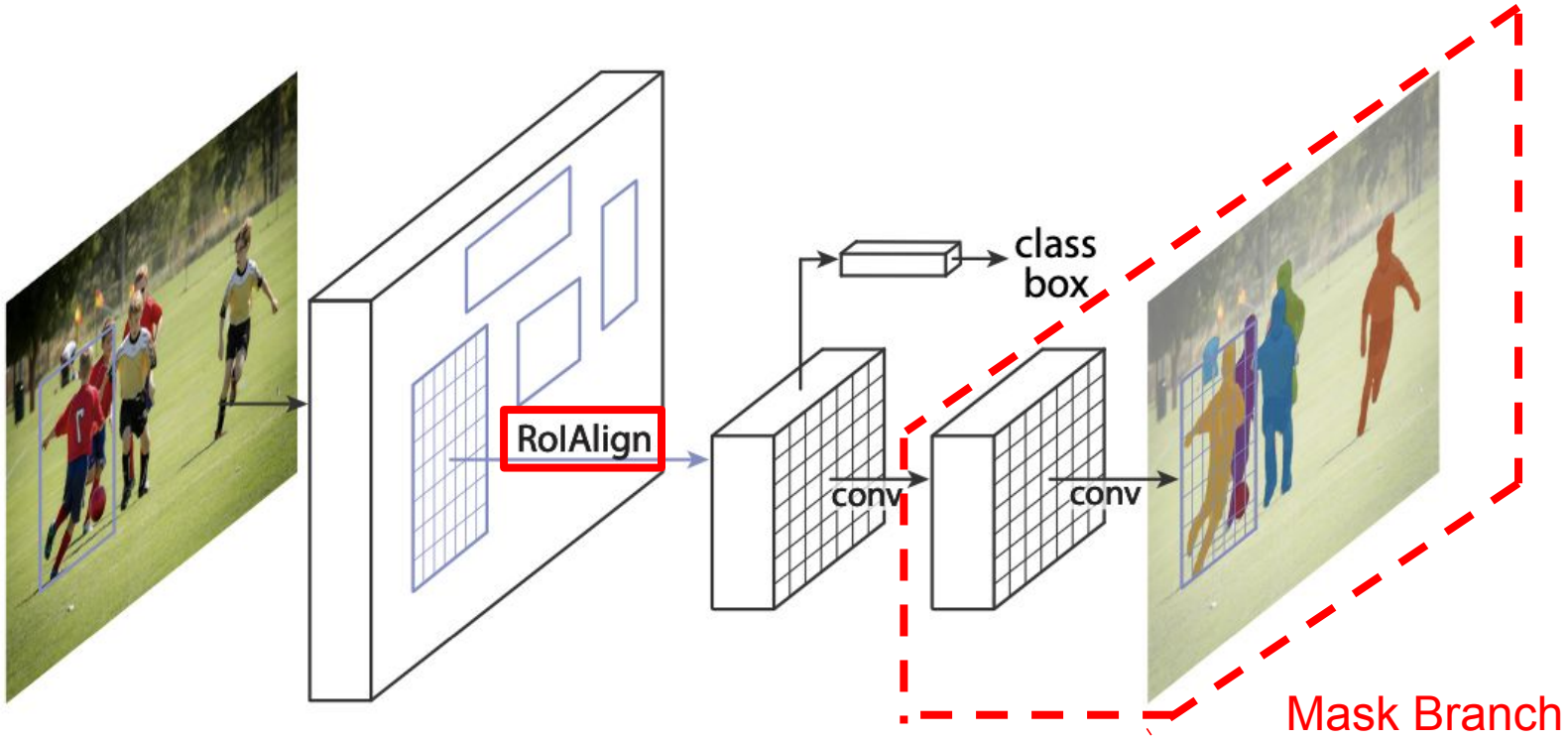
Faster RCNN

Instance Segmentation: Mask RCNN



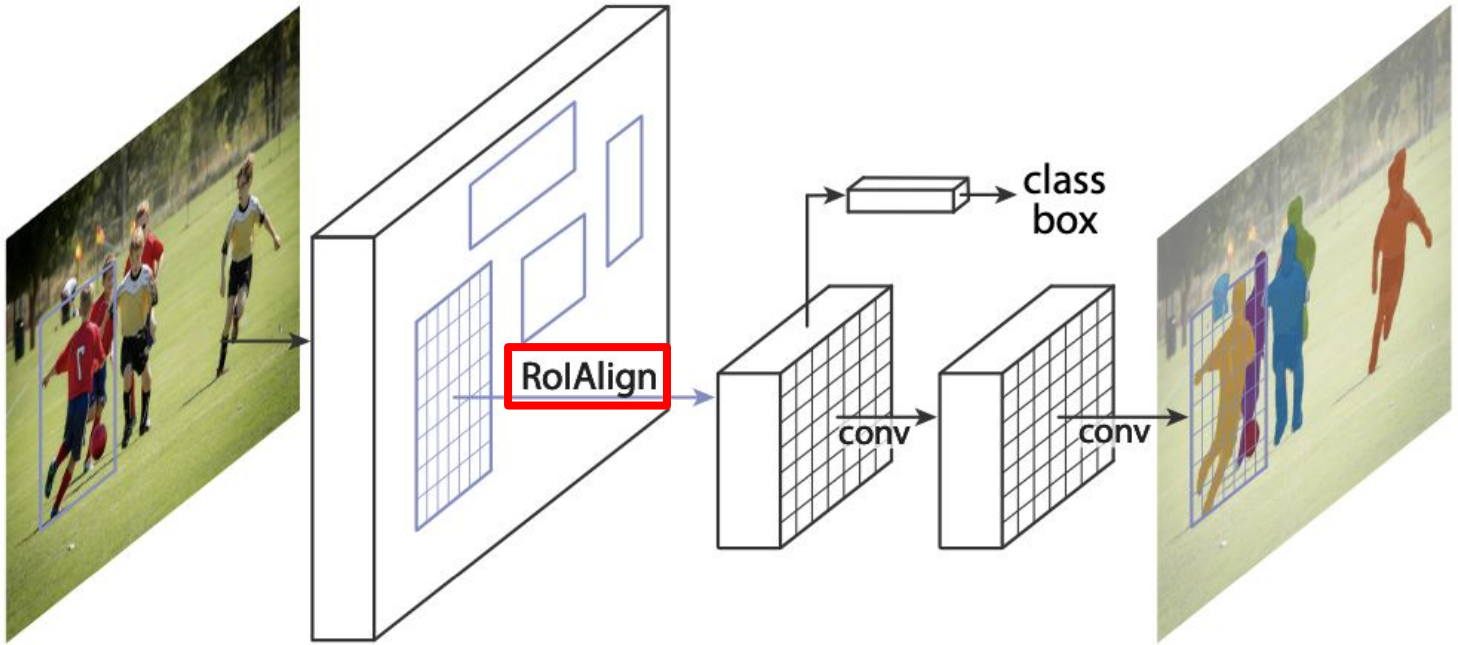
He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Instance Segmentation: Mask RCNN



He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Instance Segmentation: Mask RCNN



Difference between RoIAlign and RoIPool?

He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Mask RCNN (RoIAlign)

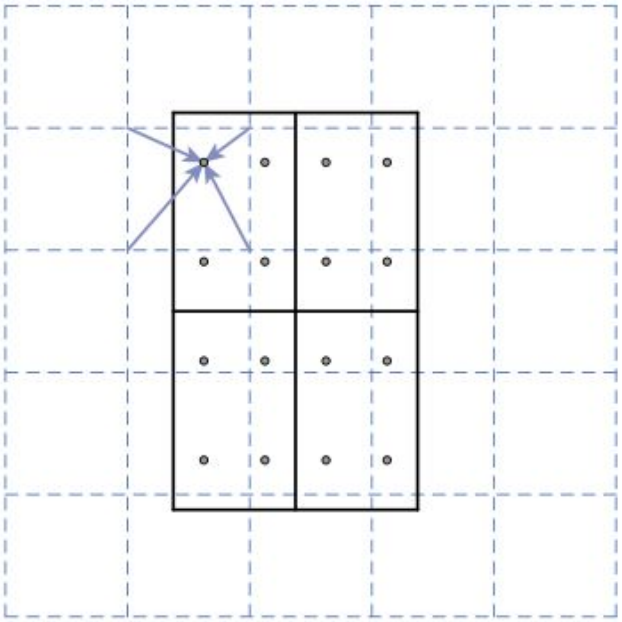


Figure 3. **RoIAlign:** The dashed grid represents a feature map, the solid lines an RoI (with 2×2 bins in this example), and the dots the 4 sampling points in each bin. RoIAlign computes the value of each sampling point by bilinear interpolation from the nearby grid points on the feature map. No quantization is performed on any coordinates involved in the RoI, its bins, or the sampling points.

He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Mask RCNN (RoIAlign)

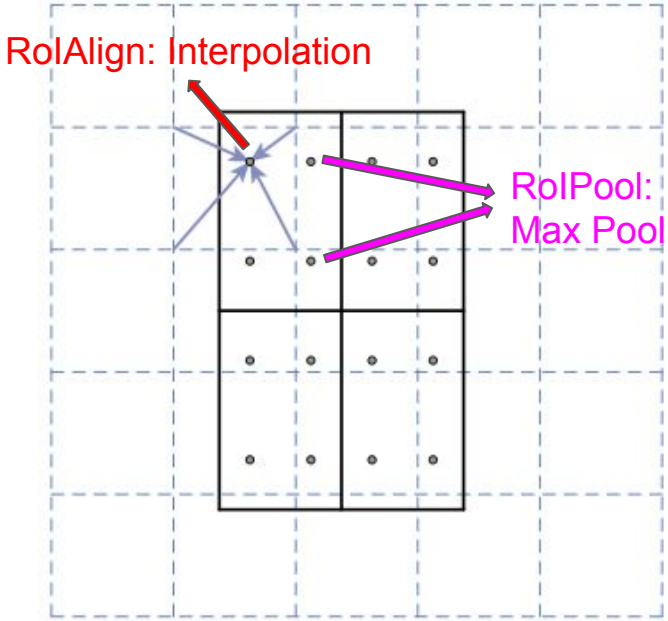
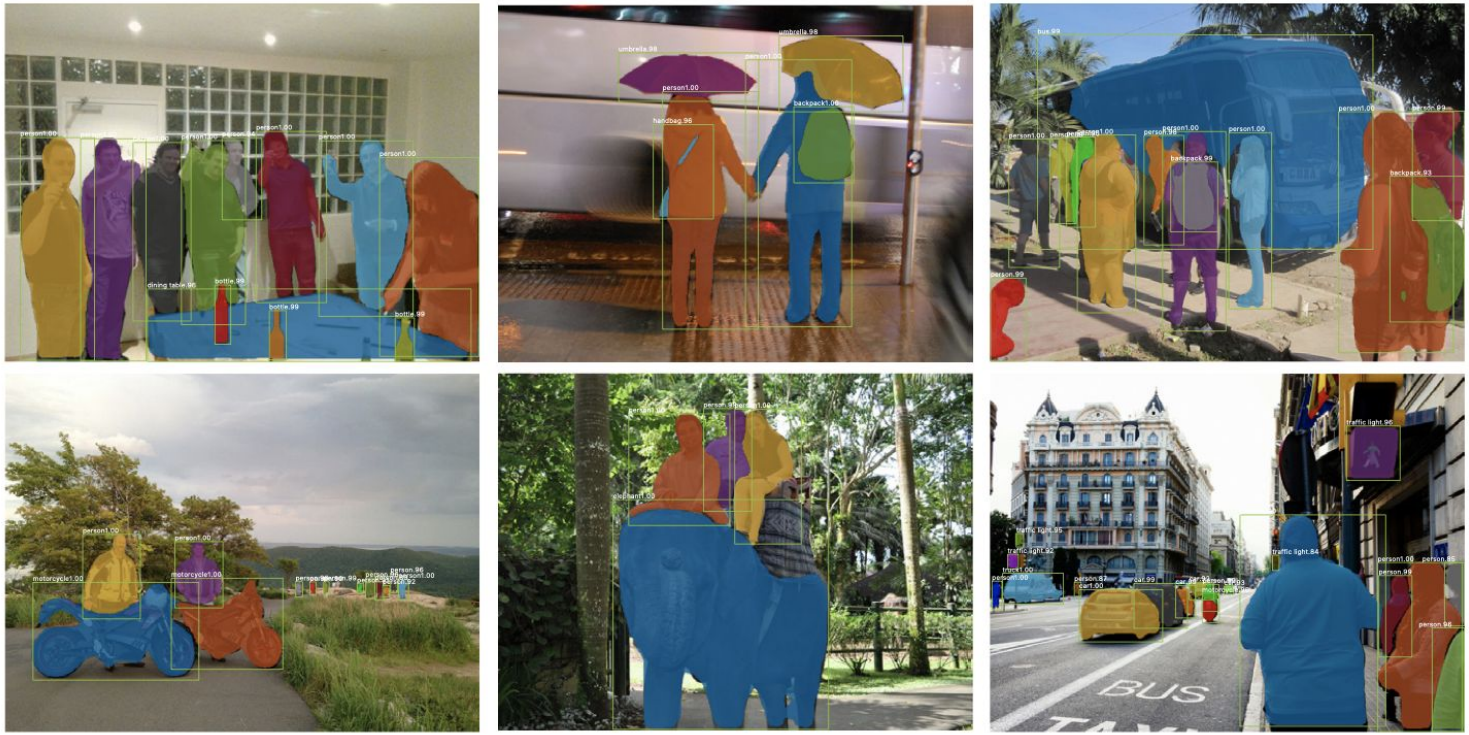


Figure 3. **RoIAlign:** The dashed grid represents a feature map, the solid lines an RoI (with 2×2 bins in this example), and the dots the 4 sampling points in each bin. RoIAlign computes the value of each sampling point by bilinear interpolation from the nearby grid points on the feature map. No quantization is performed on any coordinates involved in the RoI, its bins, or the sampling points.

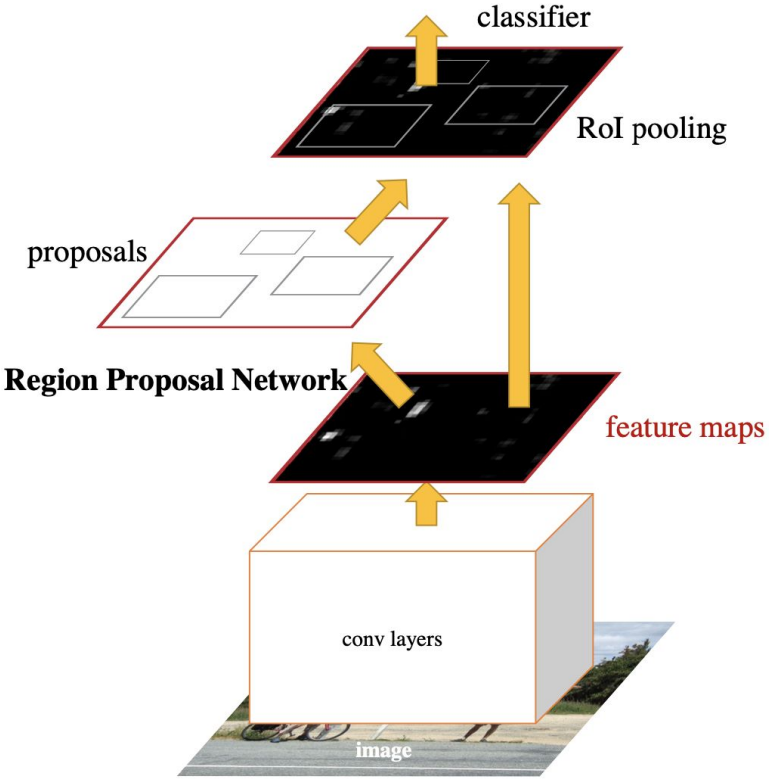
He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Instance Segmentation: Mask RCNN



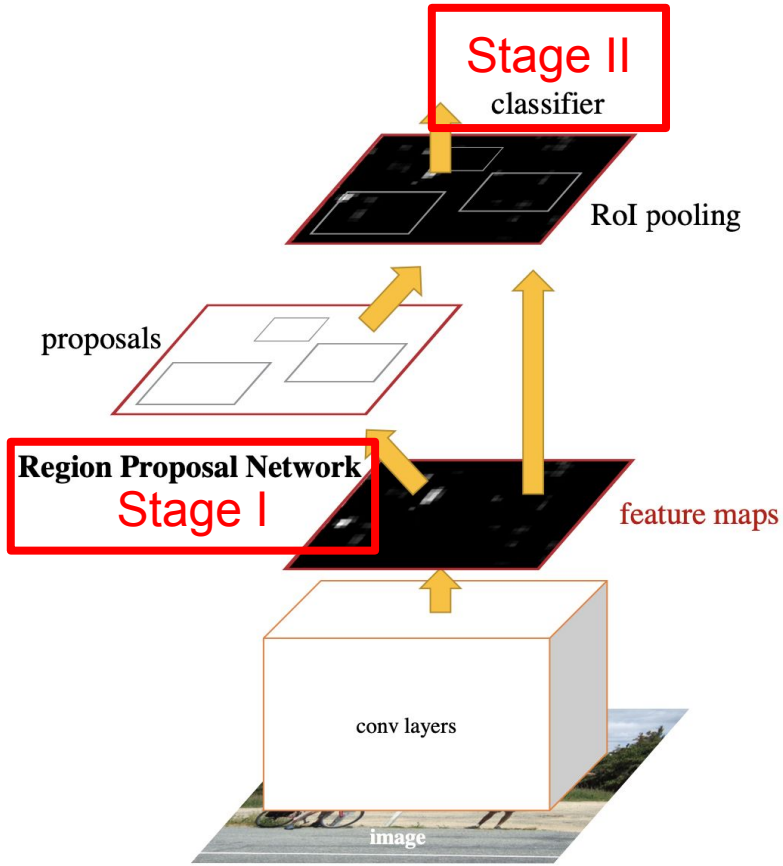
He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

Two Stage Detector



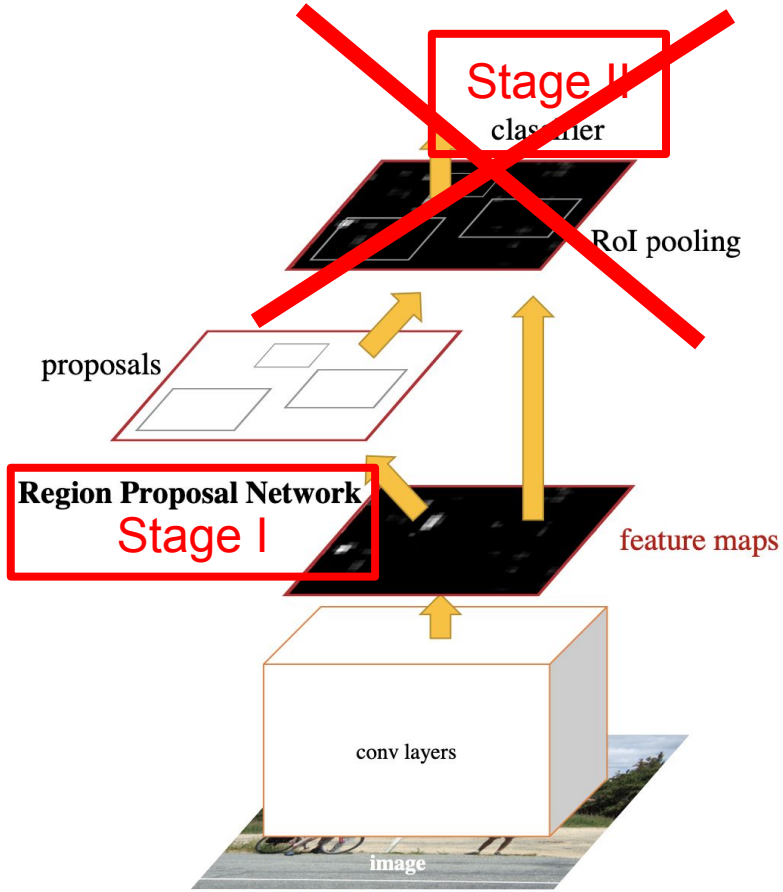
Faster RCNN

Two Stage Detector



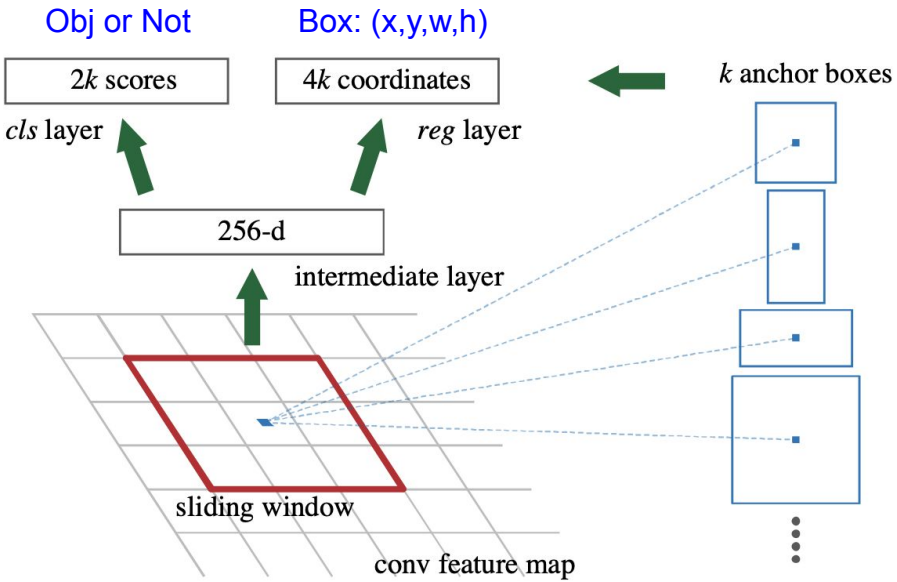
Faster RCNN

Two Stage Detector

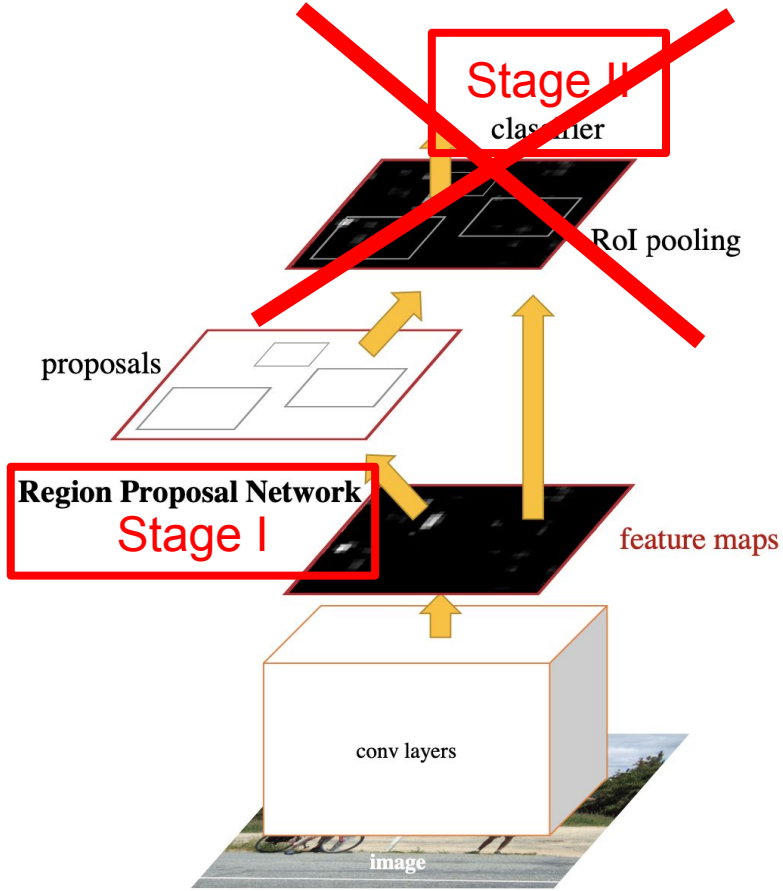


Faster RCNN

Two Stage Detector



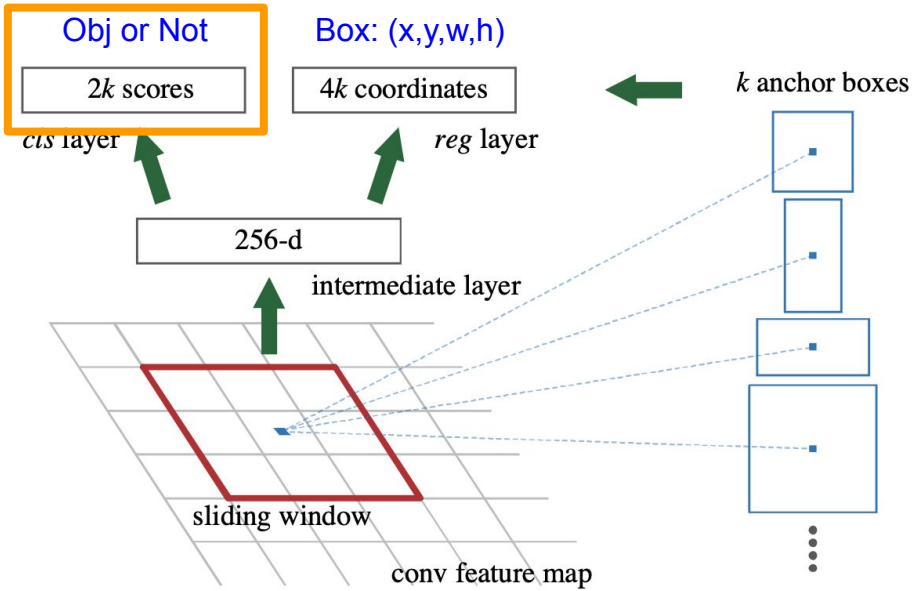
RPN



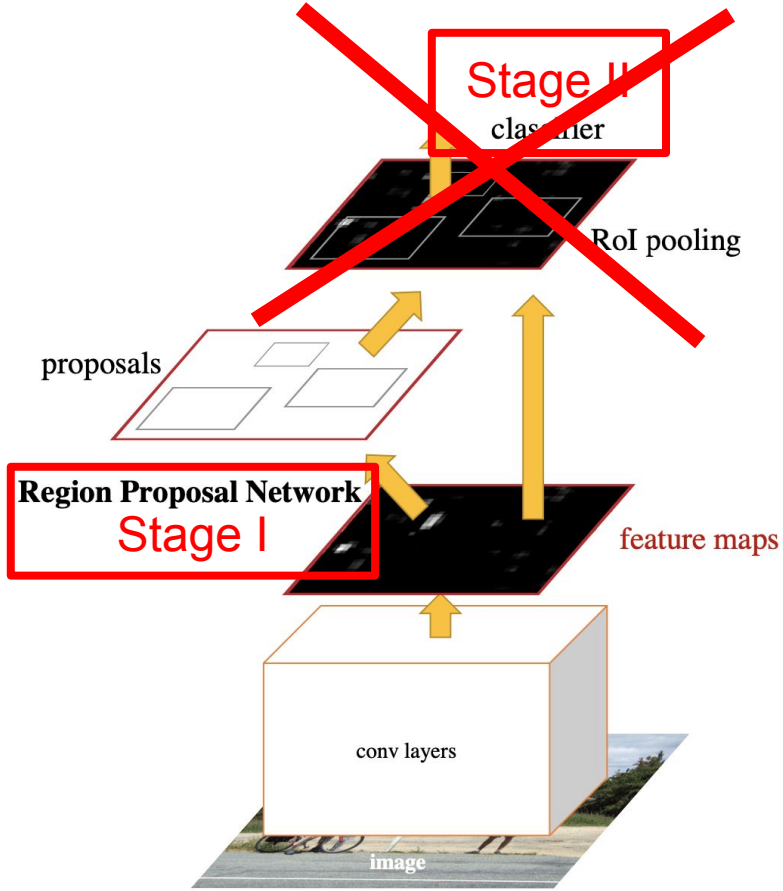
Faster RCNN

Two Stage Detector

Foreground Object Detector

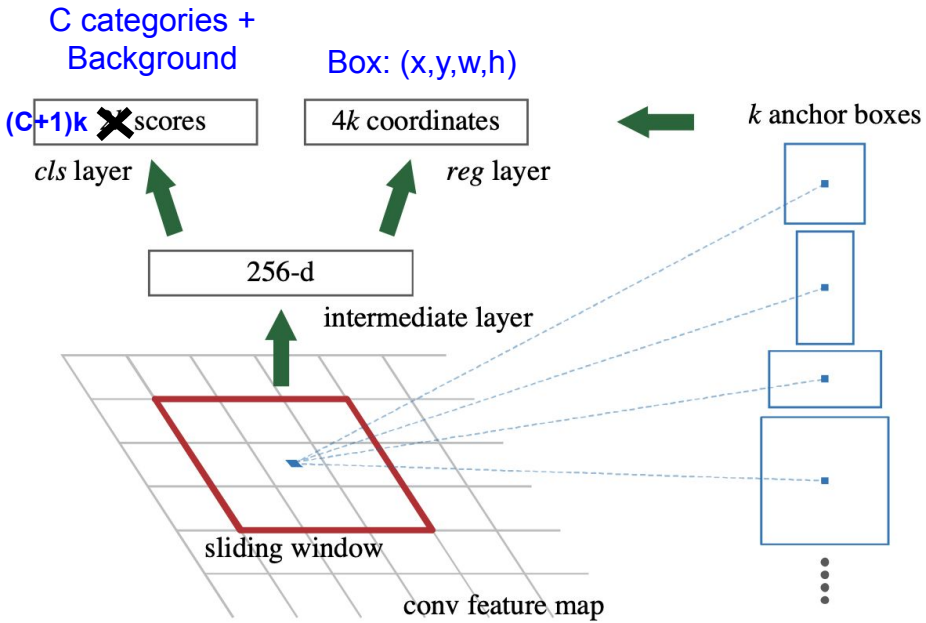


RPN

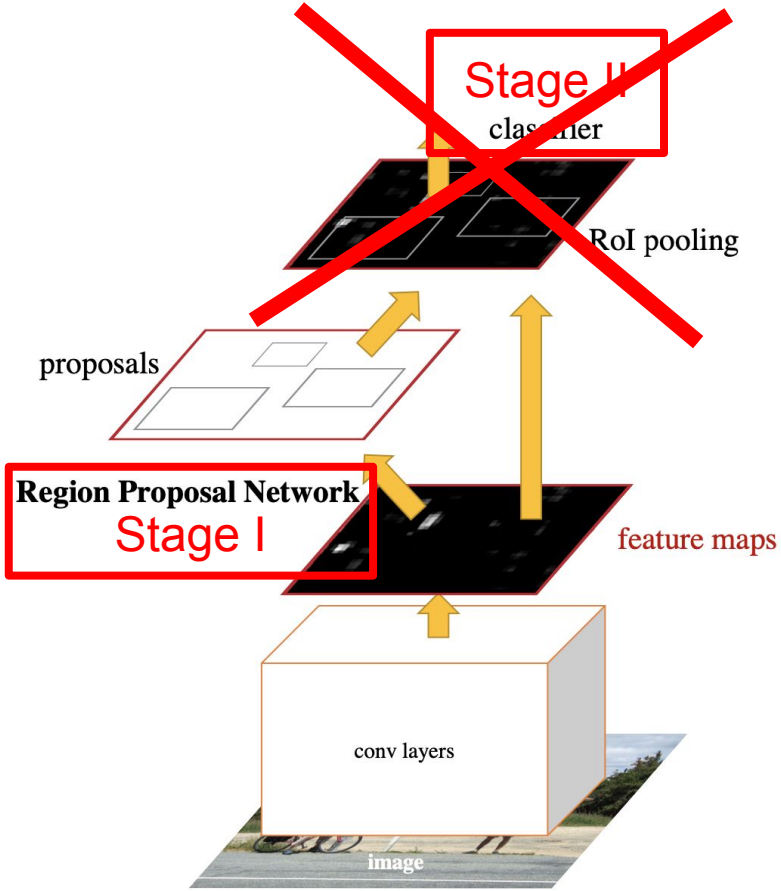


Faster RCNN

Two Stage Detector



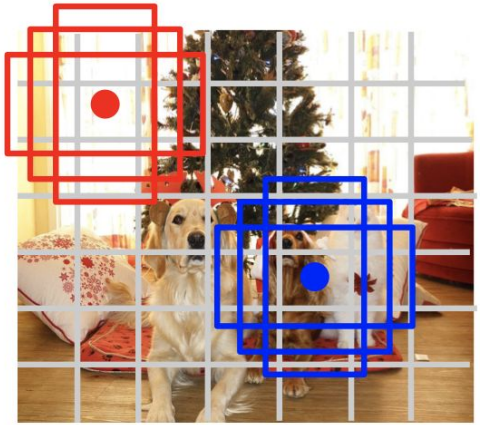
RPN



Faster RCNN

Detection without Proposals: YOLO / SSD

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

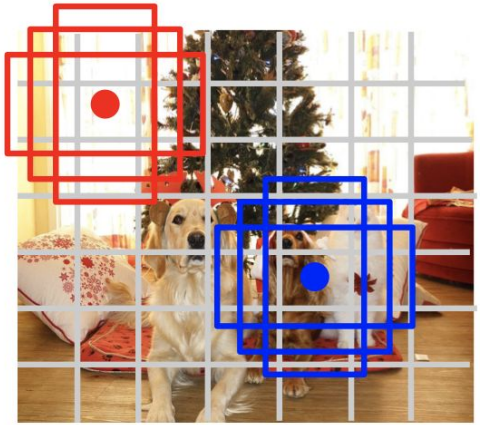
Output:
7 x 7 x (5 * B + C)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Detection without Proposals: YOLO / SSD

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

Output:
7 x 7 x (5 * B + C)

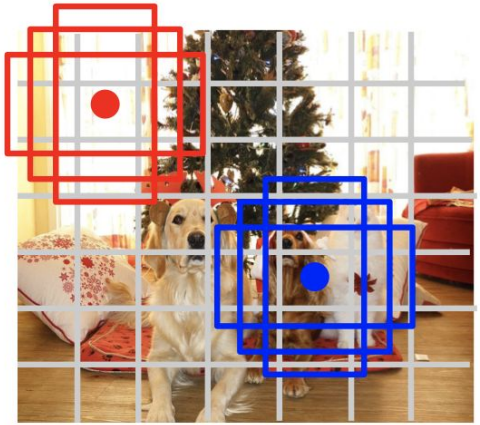
Class-agnostic box

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Detection without Proposals: YOLO / SSD

Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of **base boxes** centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)

Output:
7 x 7 x (5 * B + C)

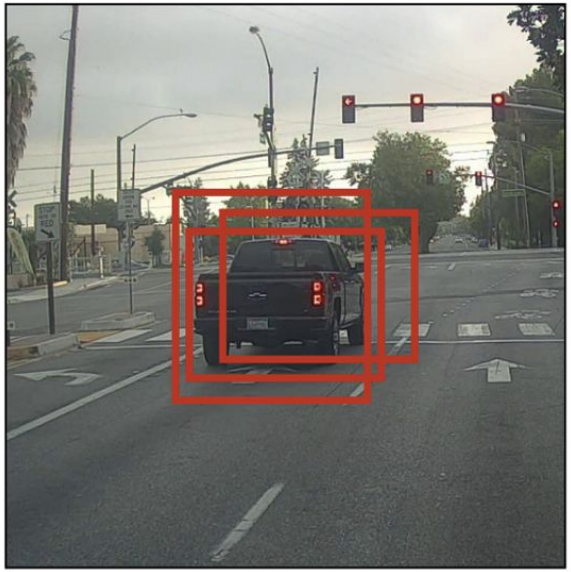
- One Stage!
- Fully Convolutional!

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016

http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Post-processing: Non-maximum Suppression

Before non-max suppression

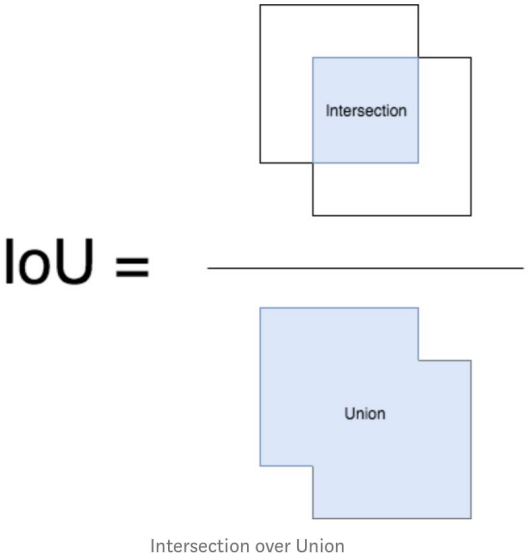


Non-Max Suppression

After non-max suppression

Remove highly overlapping bounding boxes

Post-processing: Non-maximum Suppression



Remove highly overlapping (IoU) bounding boxes

Two Stage vs. One Stage

Two Stage

- Accurate but slow

One Stage

- Fast, simple but less accurate
- Some recent works are even more accurate than two-stage methods

A Recent One Stage Detector: CenterNet

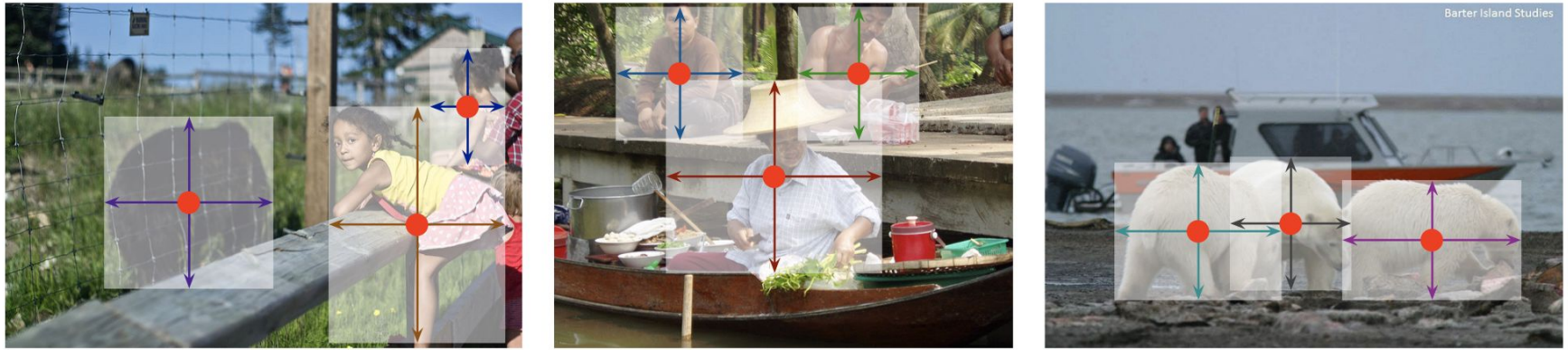


Figure 2: We model an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center. Best viewed in color.

Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).

A Recent One Stage Detector: CenterNet

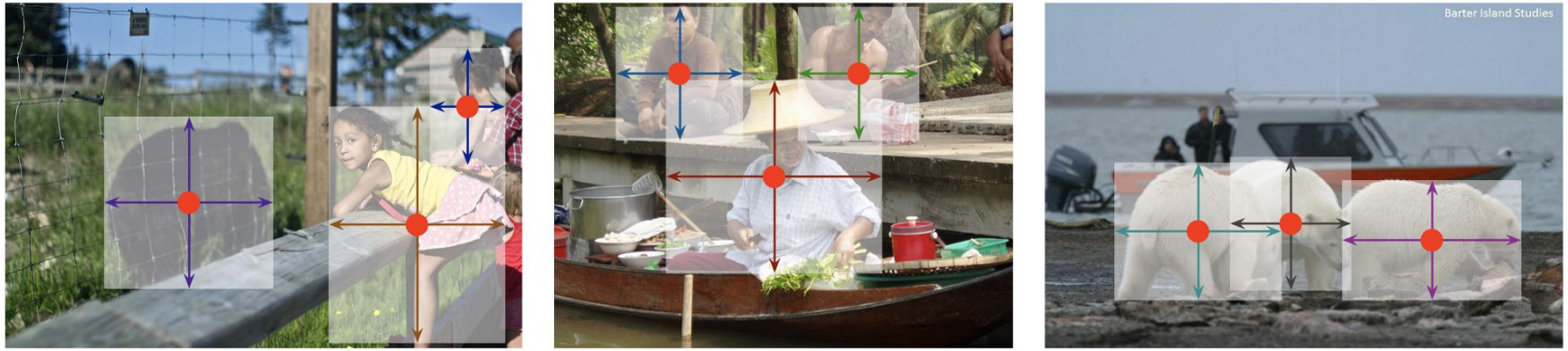


Figure 2: We model an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center. Best viewed in color.

- Anchor free. Object detection as keypoint detection.
- Produce Centerness heatmap. Harvest all the local maximums.
- Non-maximum suppression is not required.

Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).

A Recent One Stage Detector: CenterNet

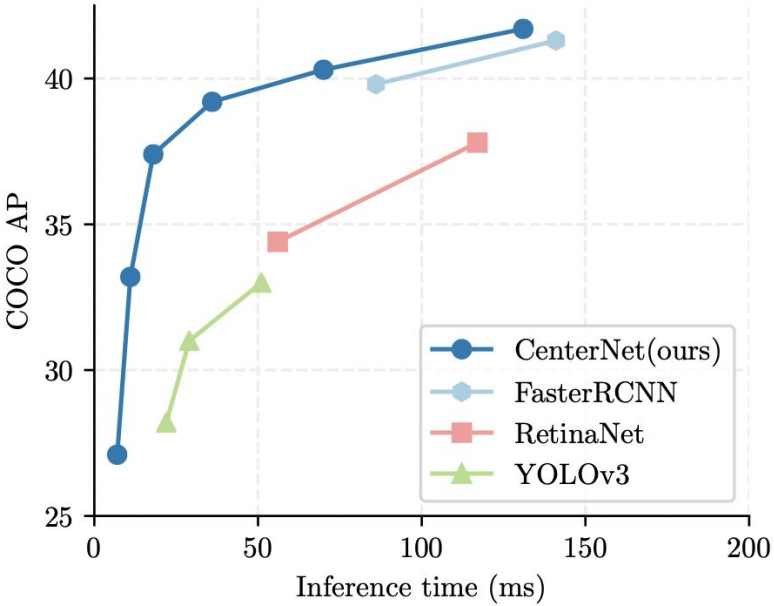


Figure 1: Speed-accuracy trade-off on COCO validation for real-time detectors. The proposed CenterNet outperforms a range of state-of-the-art algorithms.

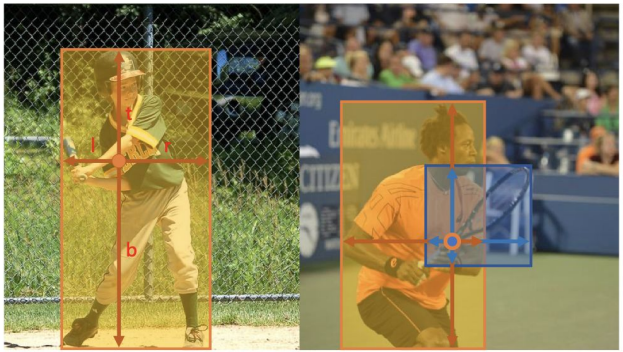
- Fast
- Simple
- Accurate

Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).

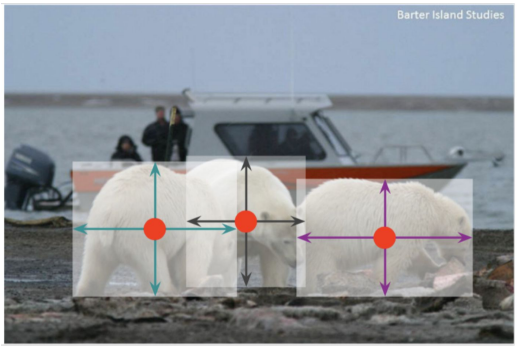
Recent Trend of Object Detection

Anchor-based ➡ Anchor-free, Two-stage ➡ One-stage

CornetNet



FCOS

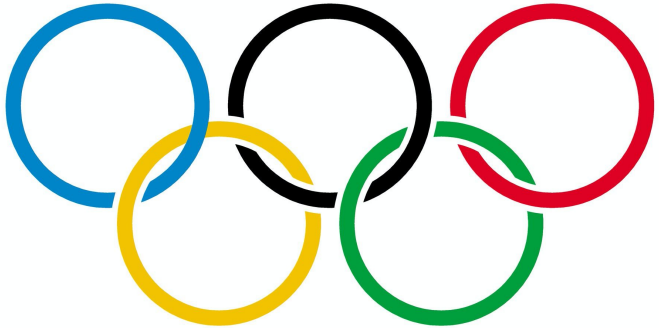


CenterNet

1. Law, Hei, and Jia Deng. "Cornernet: Detecting objects as paired keypoints." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
2 .Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.
3. Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).

Goal

Olympics

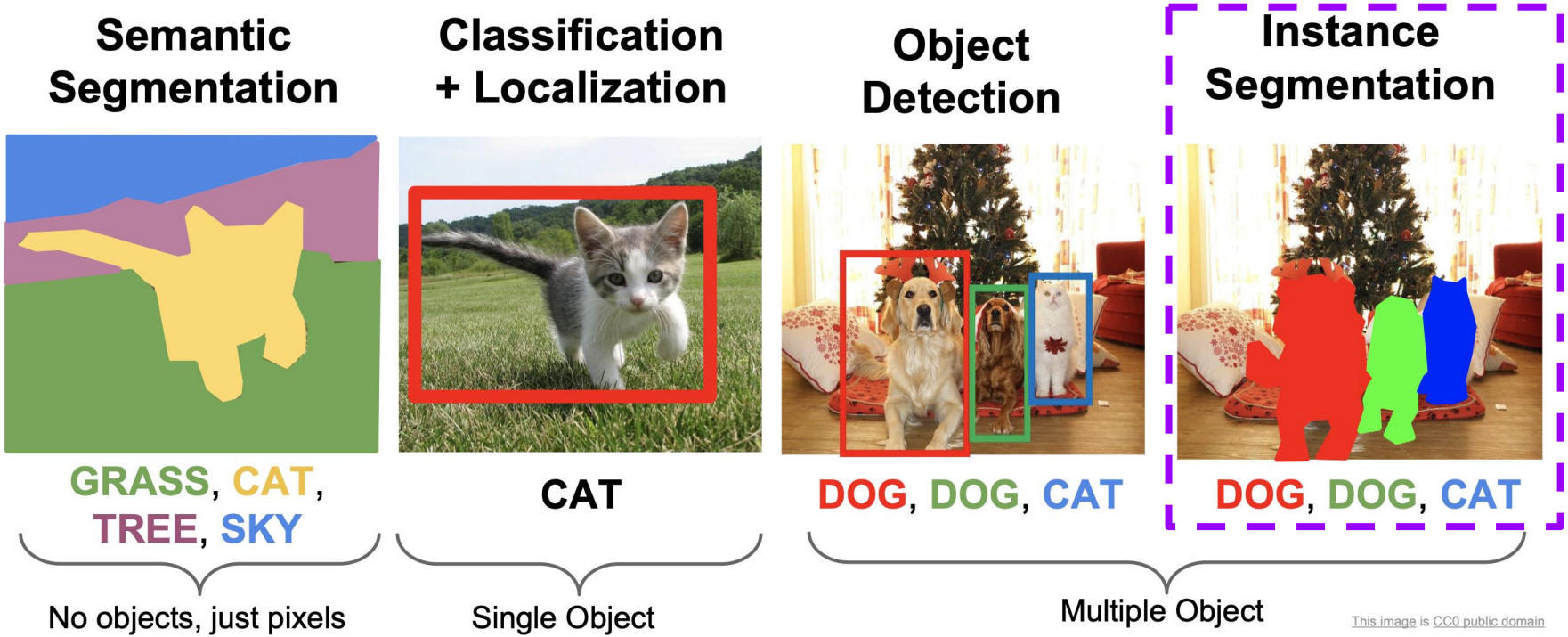


faster, higher, stronger

Object Detection

faster, simpler and more accurate

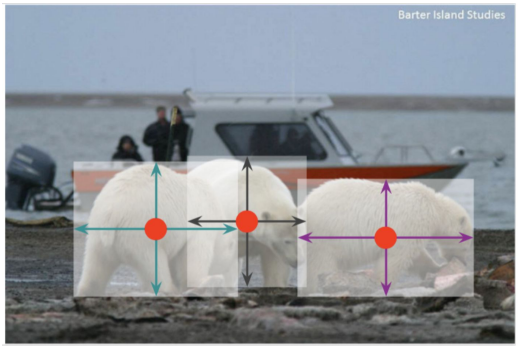
Detection & Segmentation



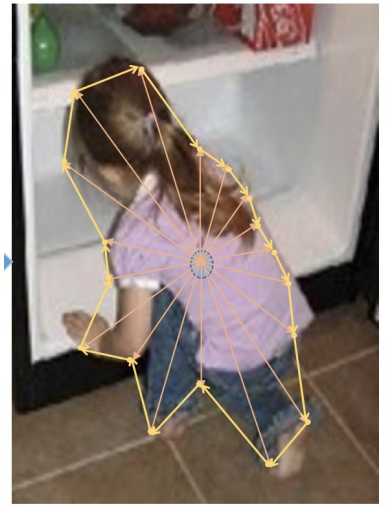
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

One-Stage Instance Segmentation

Parameterized mask representation



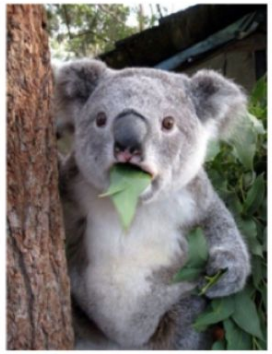
CenterNet



PolarMask

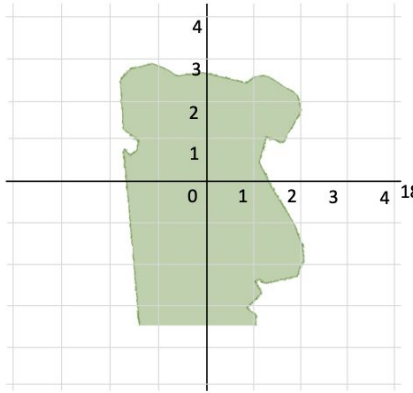
1. Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).
2. Xie, Enze, et al. "Polarmask: Single shot instance segmentation with polar representation." *arXiv preprint arXiv:1909.13226* (2019).

One-Stage Instance Segmentation: PolarMask

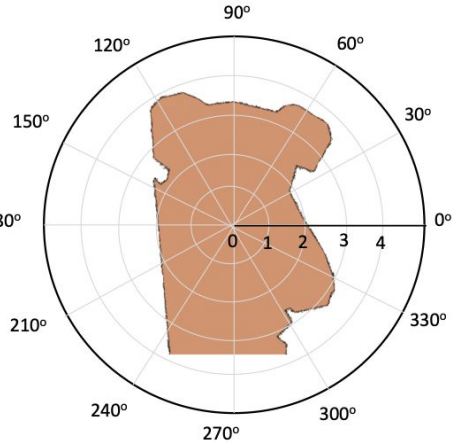


(a) Original image

(b) Pixel-wise Representation



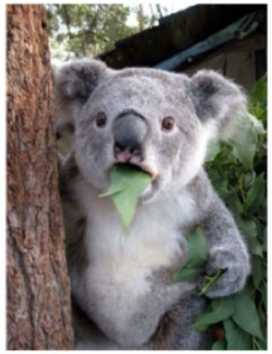
(c) Cartesian Representation



(d) Polar Representation

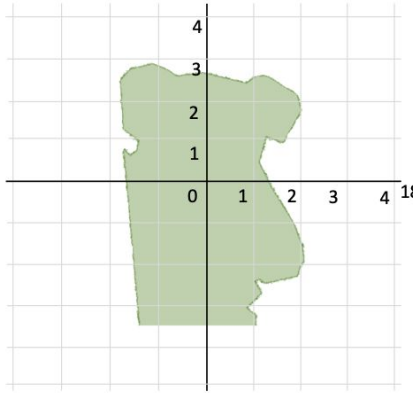
Xie, Enze, et al. "Polarmask: Single shot instance segmentation with polar representation." *arXiv preprint arXiv:1909.13226* (2019).

One-Stage Instance Segmentation: PolarMask

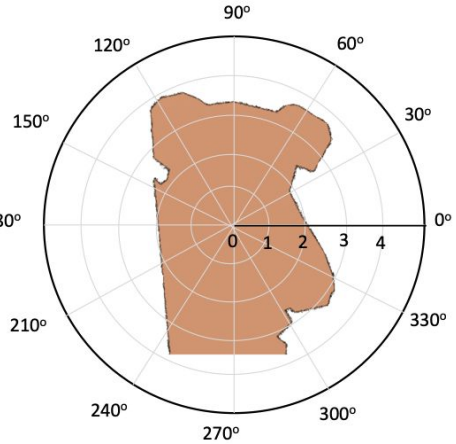


(a) Original image

(b) Pixel-wise Representation



(c) Cartesian Representation

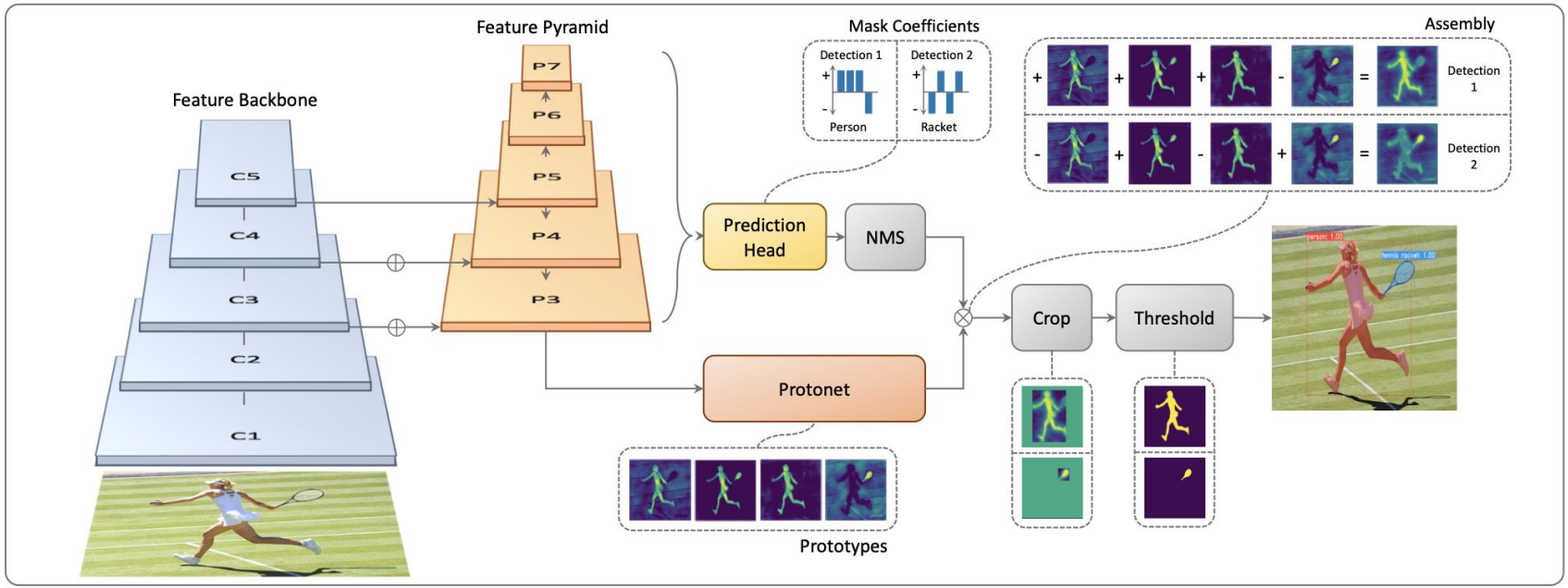


(d) Polar Representation

How to represent a mask is the key!

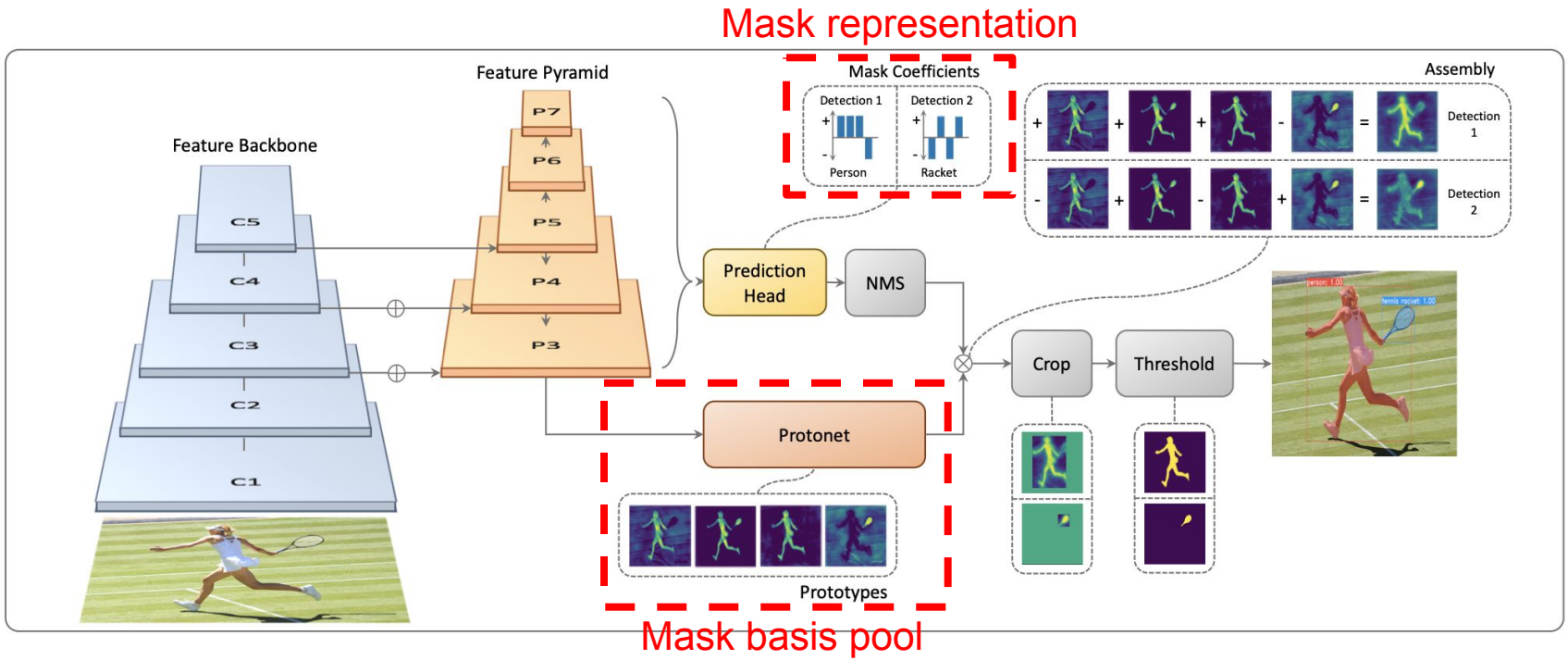
Xie, Enze, et al. "Polarmask: Single shot instance segmentation with polar representation." *arXiv preprint arXiv:1909.13226* (2019).

One-Stage Instance Segmentation: YOLACT



Bolya, Daniel, et al. "YOLACT: real-time instance segmentation." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

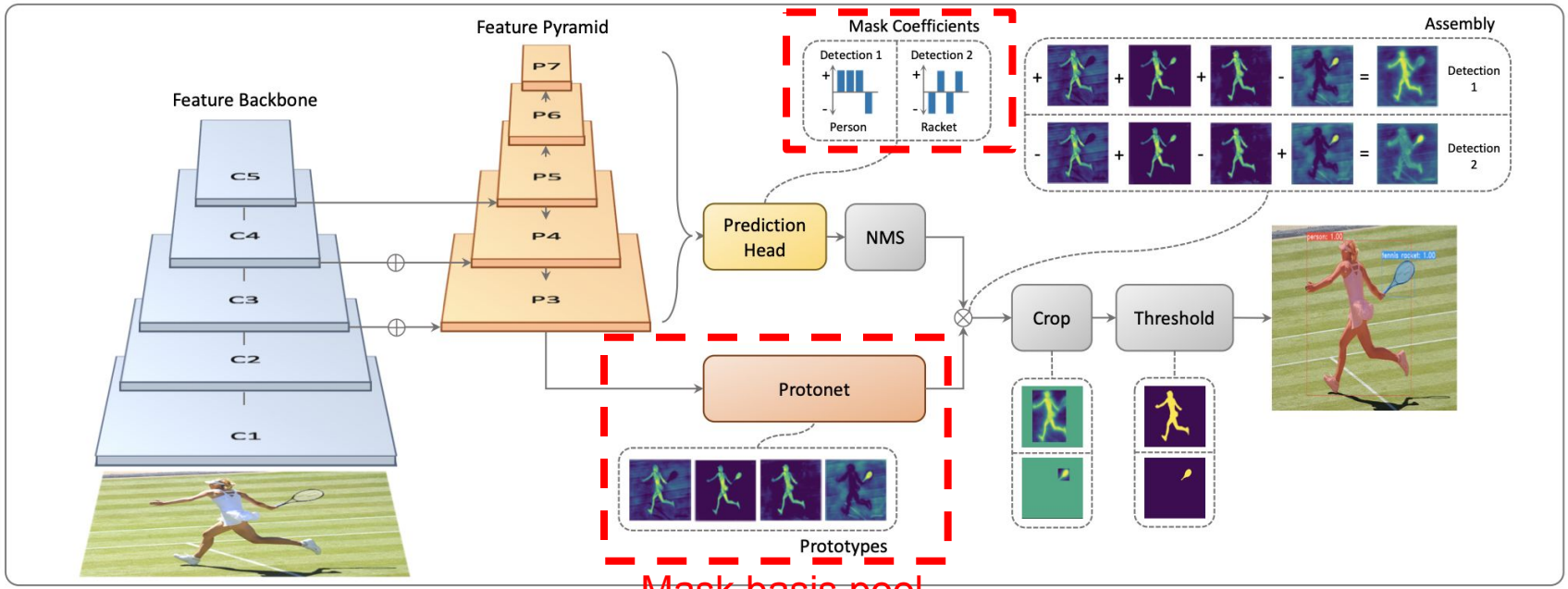
One-Stage Instance Segmentation: YOLACT



Bolya, Daniel, et al. "YOLACT: real-time instance segmentation." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

One-Stage Instance Segmentation: YOLACT

Mask representation

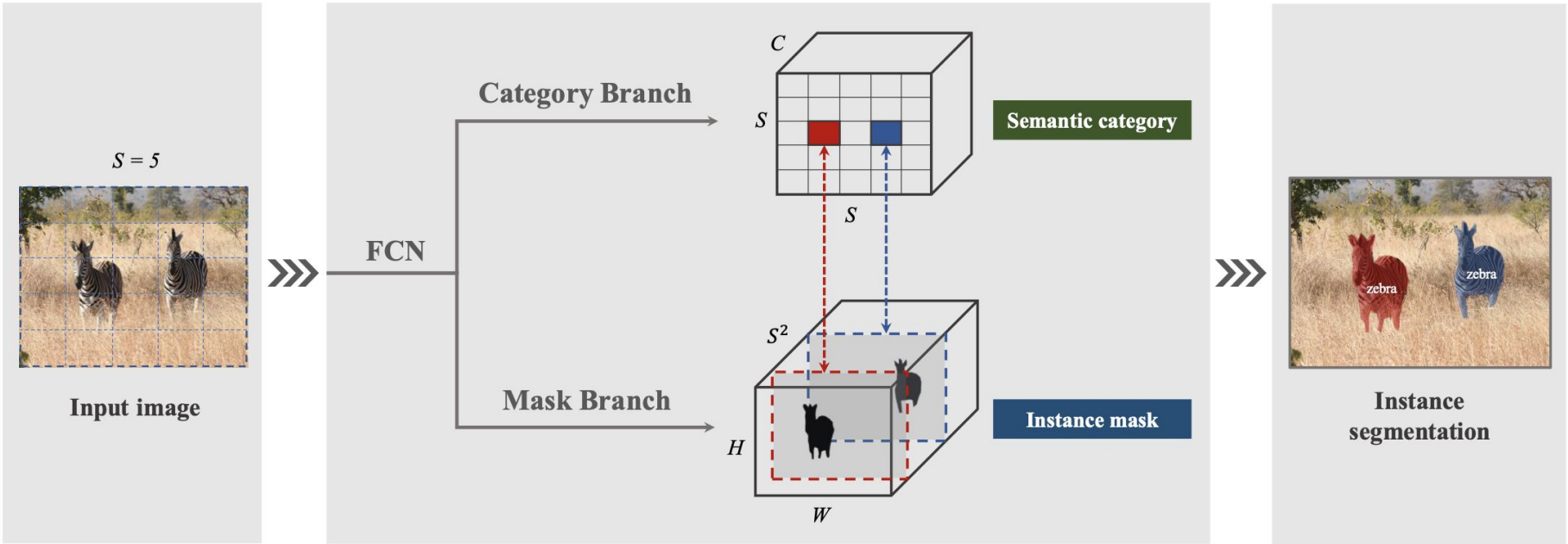


Mask basis pool

Final prediction is a linear combination of several mask basis.

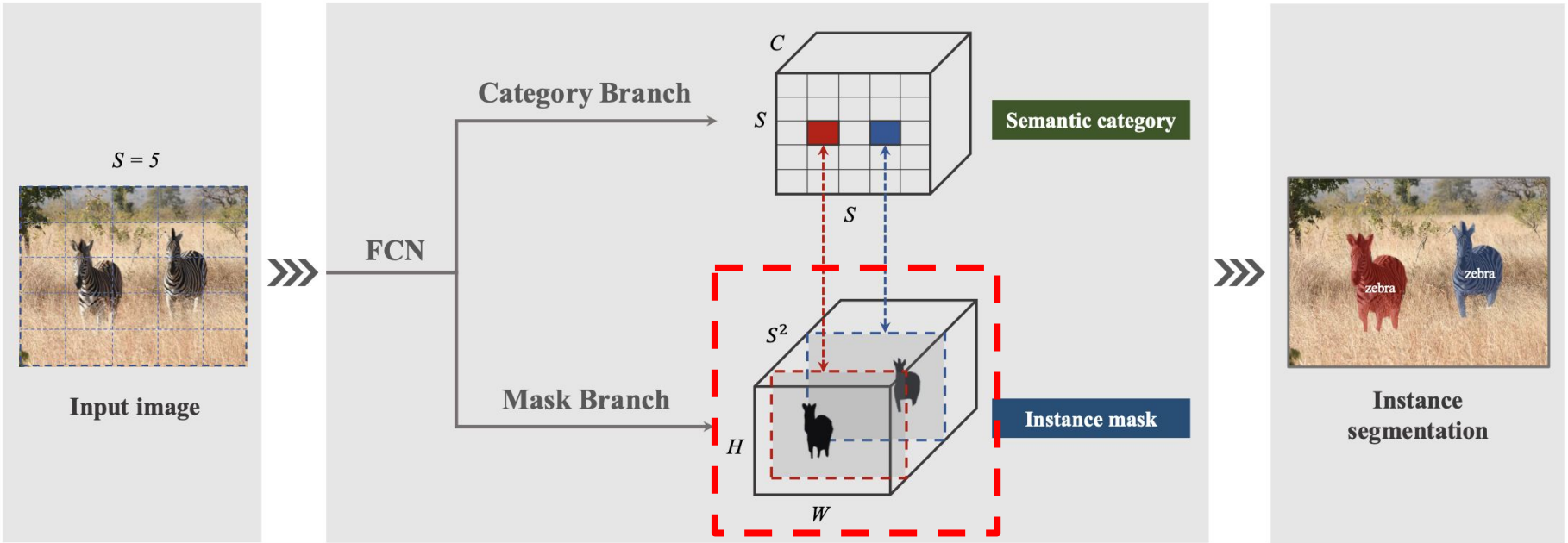
Bolya, Daniel, et al. "YOLACT: real-time instance segmentation." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

One-Stage Instance Segmentation: SOLO



Wang, Xinlong, et al. "SOLO: Segmenting Objects by Locations." *arXiv preprint arXiv:1912.04488* (2019).

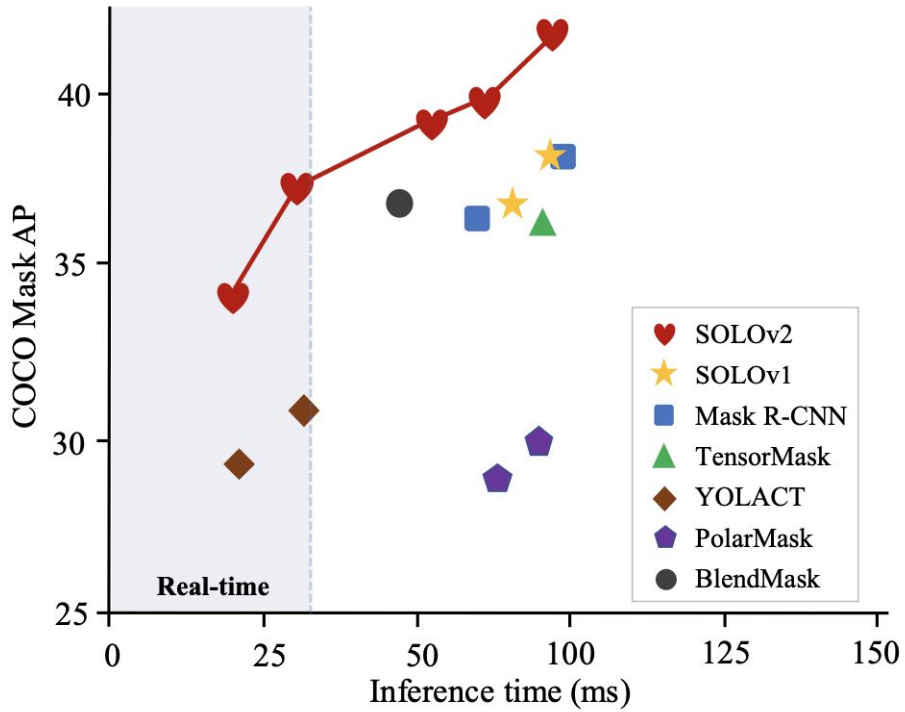
One-Stage Instance Segmentation: SOLO



Each channel represents the mask of an object instance centered at a location

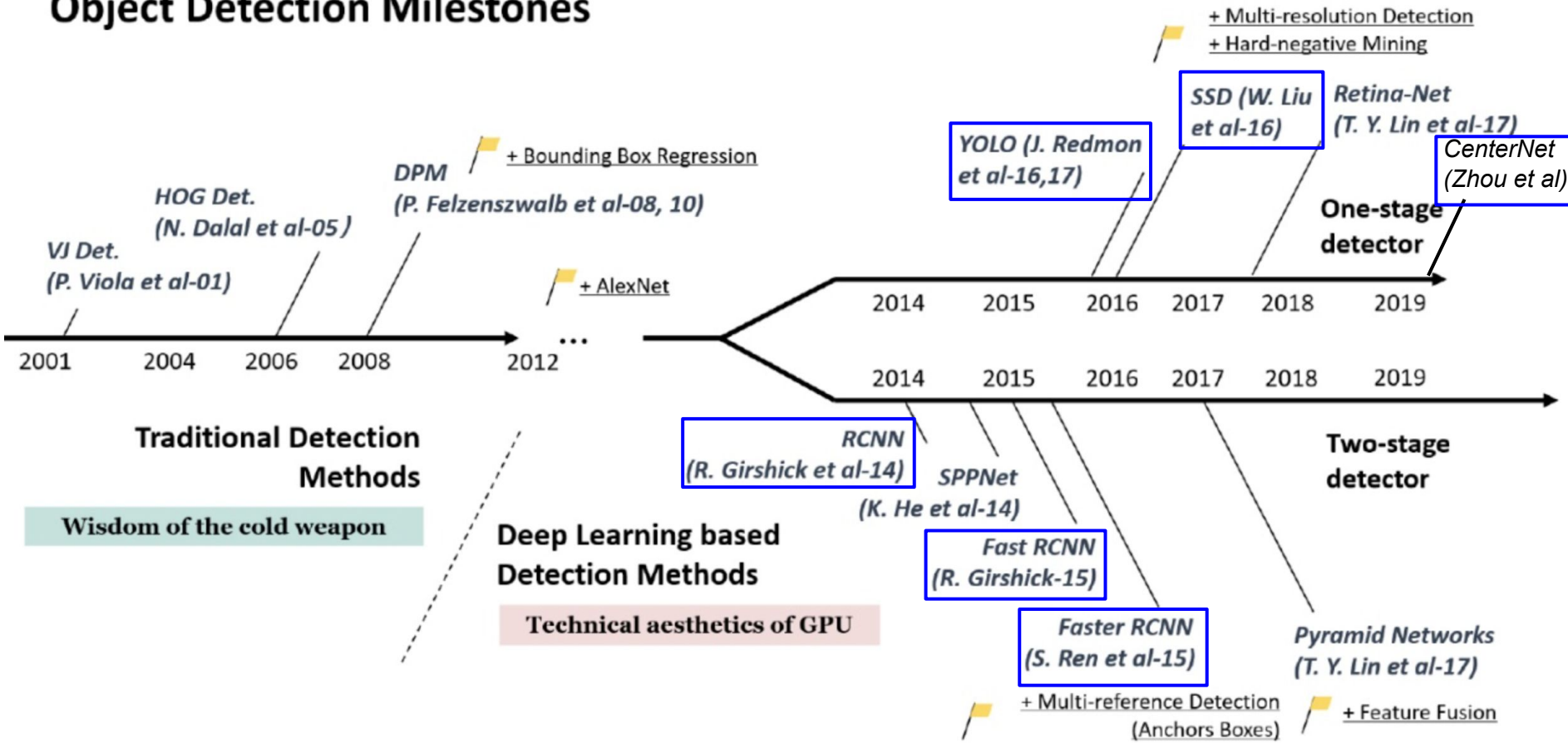
Wang, Xinlong, et al. "SOLO: Segmenting Objects by Locations." *arXiv preprint arXiv:1912.04488* (2019).

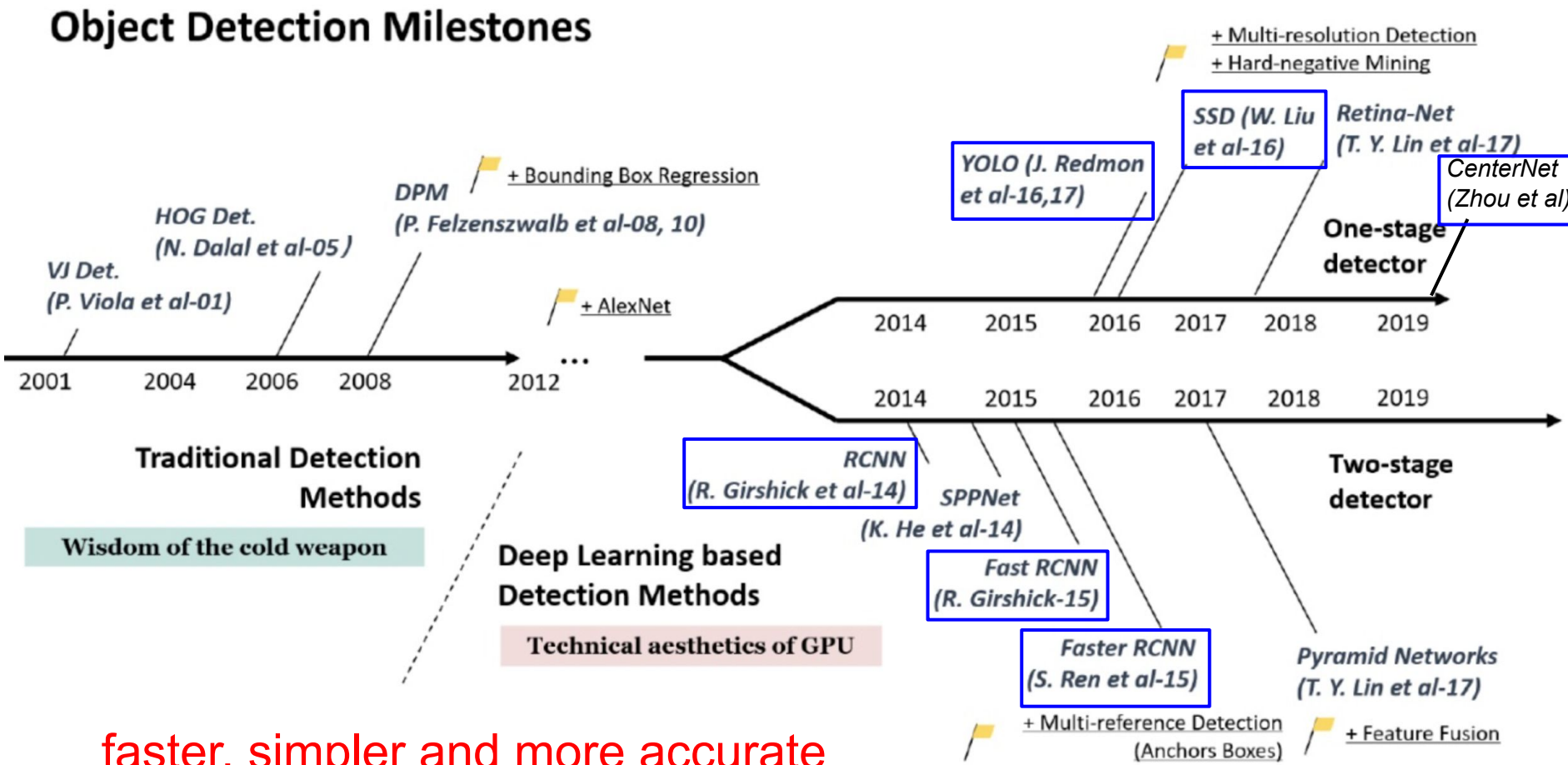
One-Stage Instance Segmentation: SOLO-V2



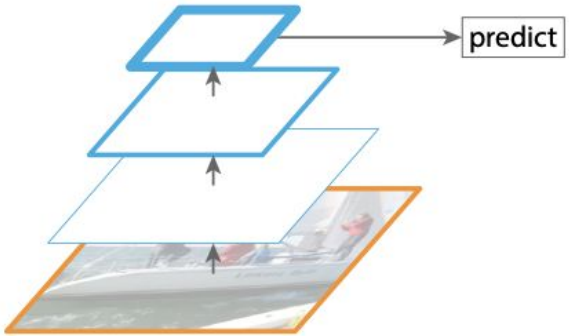
Wang, Xinlong, et al. "SOLOv2: Dynamic, Faster and Stronger." *arXiv preprint arXiv:2003.10152* (2020).

Object Detection Milestones

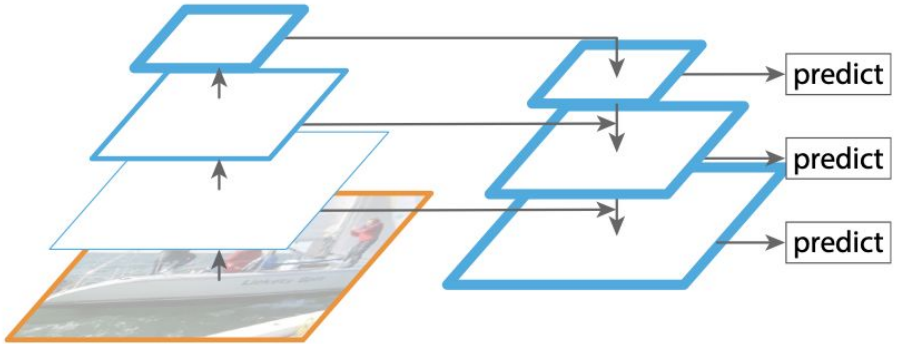



Object Detection Milestones
+ Multi-resolution Detection
+ Hard-negative Mining
SSD (W. Liu et al-16)
Retina-Net (T. Y. Lin et al-17)
YOLO (J. Redmon et al-16,17)
CenterNet (Zhou et al)
One-stage detector
+ Bounding Box Regression
DPM (P. Felzenszwalb et al-08, 10)
HOG Det. (N. Dalal et al-05)
VJ Det. (P. Viola et al-01)
+ AlexNet
2001
2004
2006
2008
2012
2014
2015
2016
2017
2018
2019
Traditional Detection Methods
Wisdom of the cold weapon
RCNN (R. Girshick et al-14)
SPPNet (K. He et al-14)
Fast RCNN (R. Girshick-15)
Faster RCNN (S. Ren et al-15)
Pyramid Networks (T. Y. Lin et al-17)
Two-stage detector
Deep Learning based Detection Methods
Technical aesthetics of GPU
+ Multi-reference Detection (Anchors Boxes)
+ Feature Fusion
faster, simpler and more accurate

Multi-Scale Representation: Feature Pyramid Network



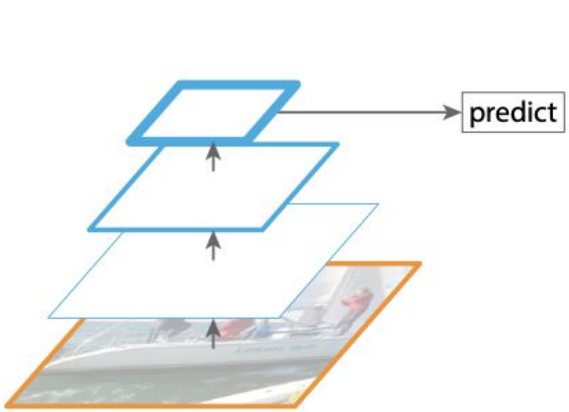
Single Feature Map



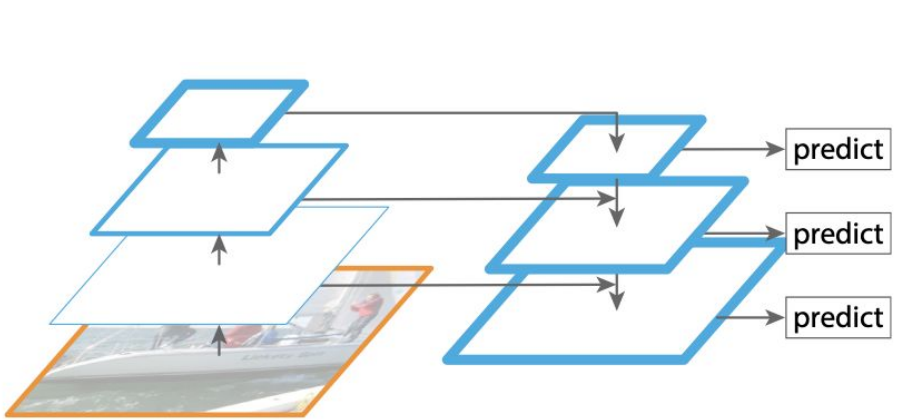
Feature Pyramid Network

Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Multi-Scale Representation: Feature Pyramid Network



Single Feature Map



Feature Pyramid Network

Each level in the hierarchy represents a specific scale

Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.