Object Detection

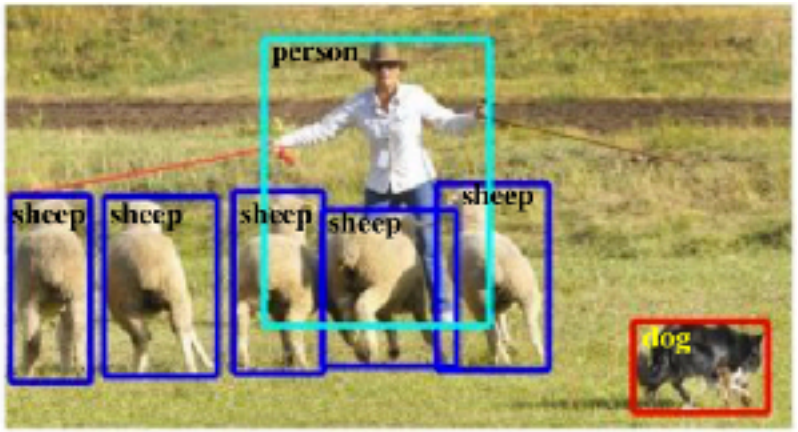
Introduction to RCNN and Yolo

Data Science| Fall 2024@ Knowledge Stream

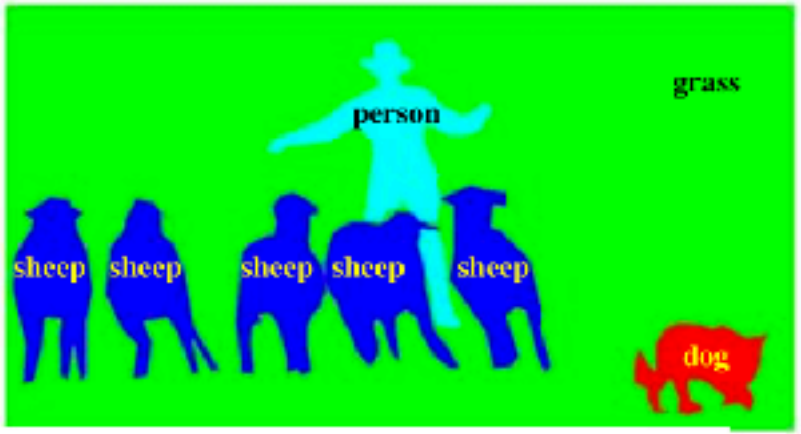
Sana Jabbar

Mask RCNN

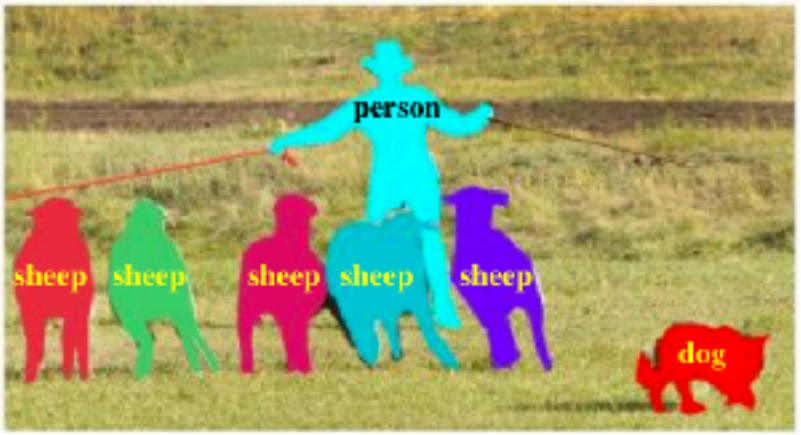
(a) Object Classification



(b) Generic Object Detection (Bounding Box)



(c) Semantic Segmentation



(d) Object Instance Segmetation

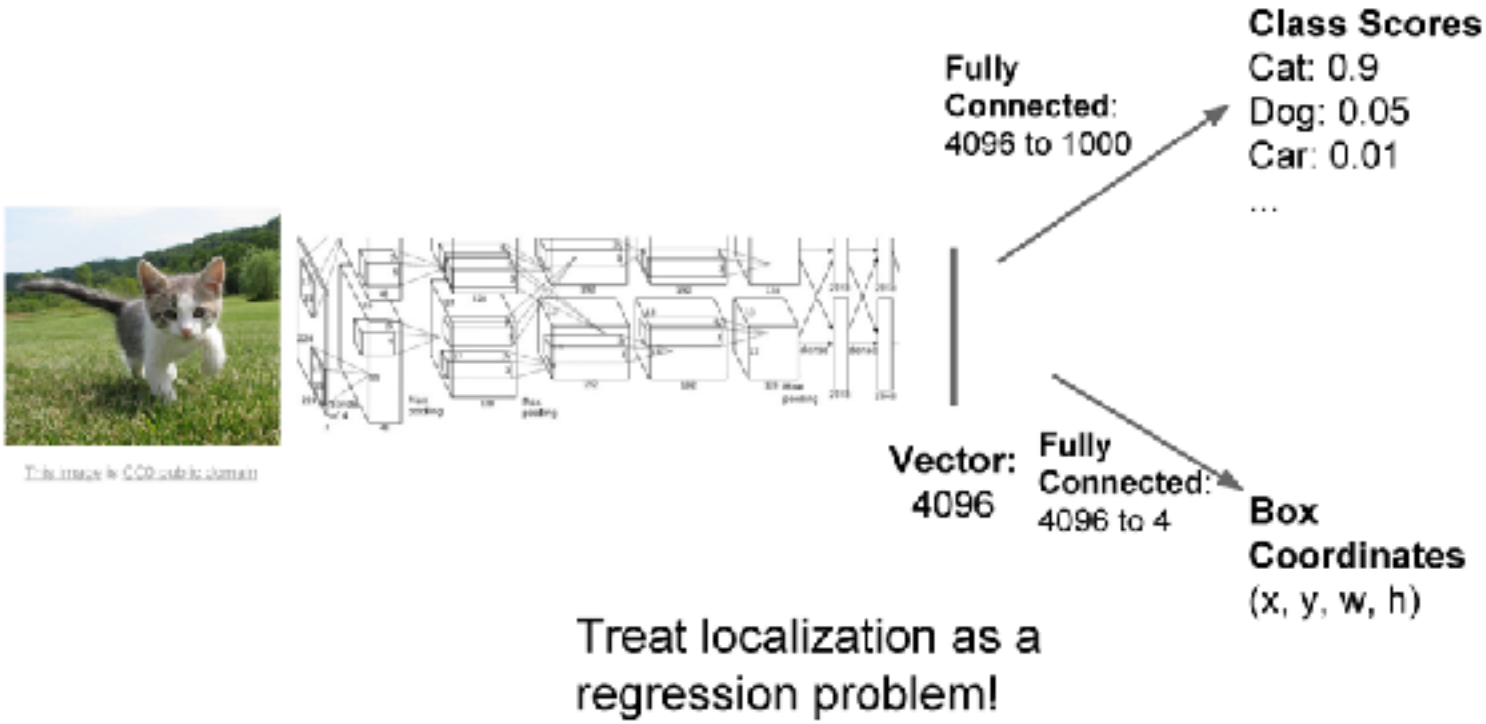
Li Liu et al. Deep Learning for Generic Object Detection: A Survey 2018

Survey of Object Detection

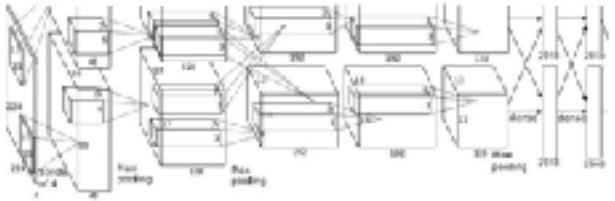
- Deep Learning for Generic Object Detection: A Survey
  Li Liu et al. 2018
  - Region Proposals
  - Region CNN
  - Fast RCNN
  - Faster RCNN
  - Yolo
  - SSD

- 3D Object Detection

Classification + Localization



Classification + Localization



Vector: 4096

Fully Connected: 4096 to 1000

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

Softmax Loss

Correct label: Cat

Fully Connected: 4096 to 4

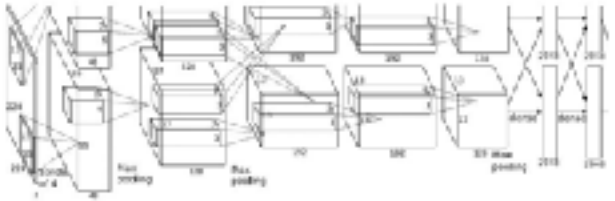
Box Coordinates (x, y, w, h)

L2 Loss

Correct box: (x', y', w', h')

Treat localization as a regression problem!

Classification + Localization



Correct label:
Cat

Fully Connected:
4096 to 1000

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

Softmax Loss

Multitask Loss

+

Loss

Vector:
4096

Fully Connected:
4096 to 4

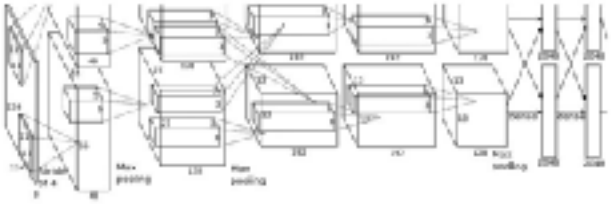
Box Coordinates
(x, y, w, h)

L2 Loss

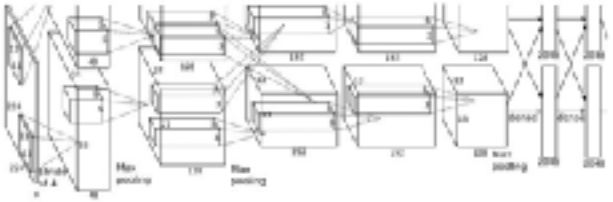
Correct box:
(x', y', w', h')

Treat localization as a regression problem!

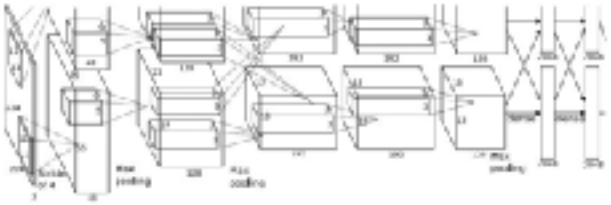
Object Detection as Regression



CAT: (x, y, w, h)



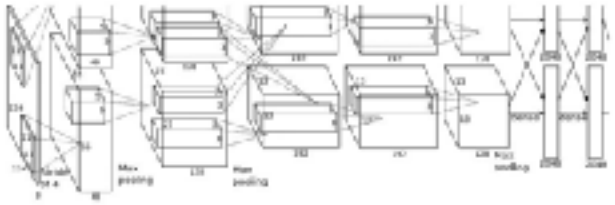
DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)



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

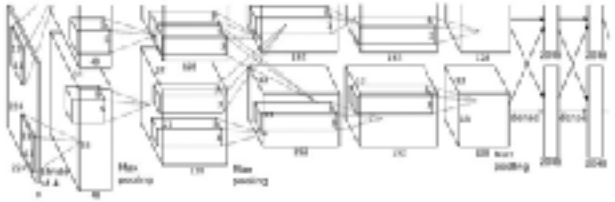
Object Detection as Regression

Each image needs a different number of outputs!



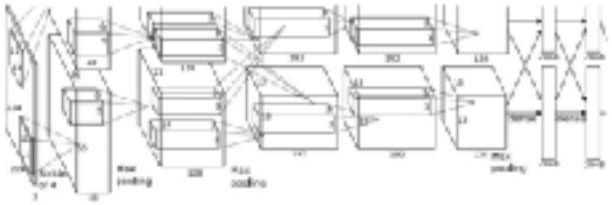
CAT: (x, y, w, h)

4 numbers



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

16 numbers



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

Many numbers!

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

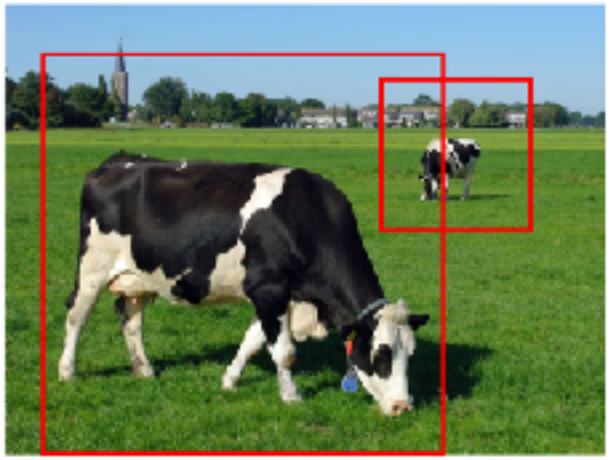
Parsing at fixed scale

Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

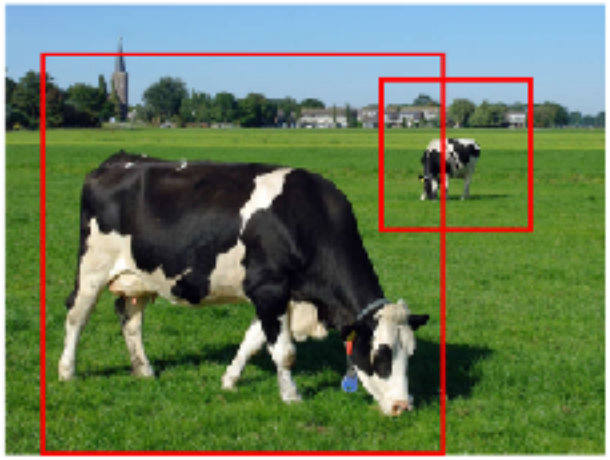


Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

Parsing at fixed scale

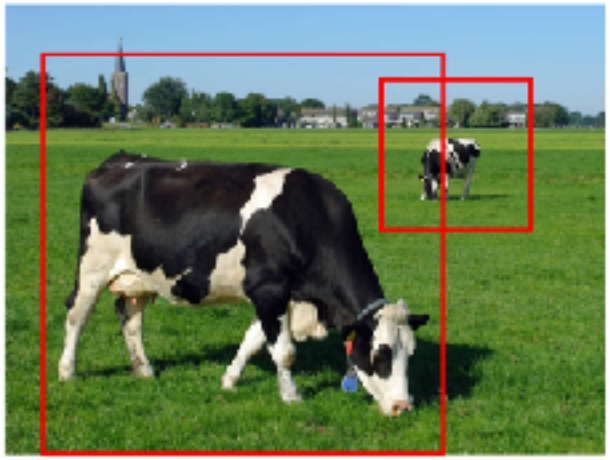


Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

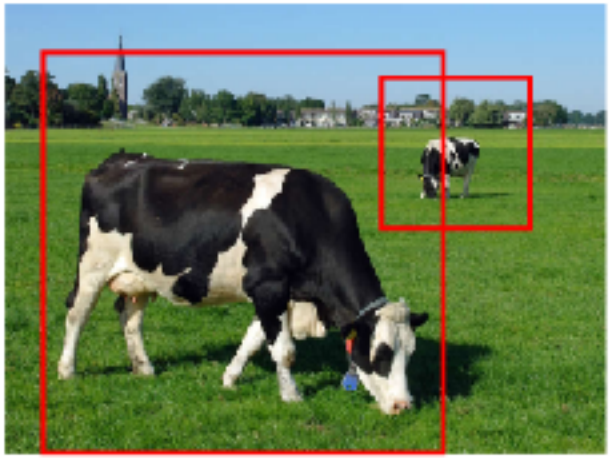
Parsing at fixed scale



Final list of detections

2D Object Detection: Sliding Window

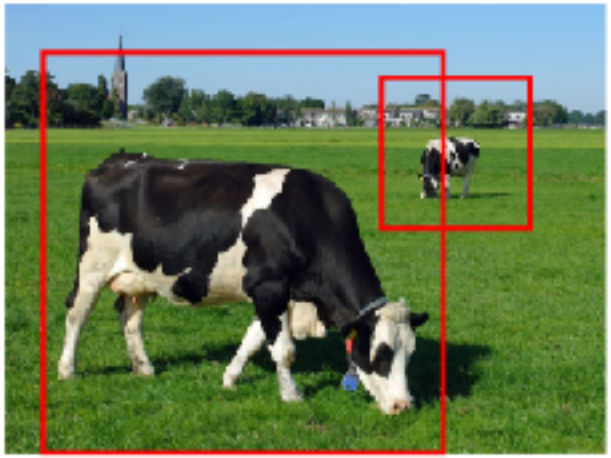
The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.



Final list of detections

2D Object Detection: Sliding Window

The simplest strategy to move from image classification to object detection is to classify local regions, at multiple scales and locations.

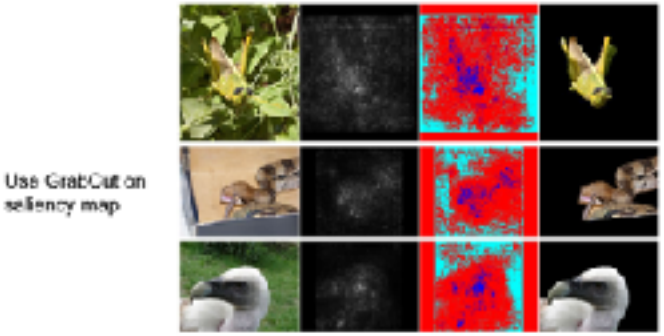


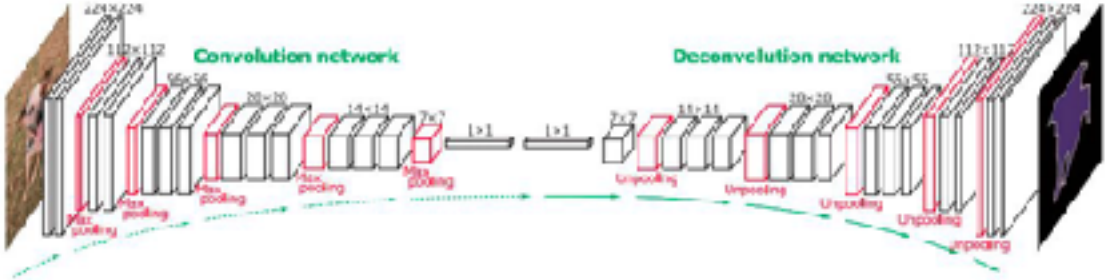
Final list of detections

This "sliding window" approach evaluates a classifier multiple times, and its computational cost increases with the prediction accuracy.

Region Proposals

- Sliding window is computationally expensive
- Can we use segmentation to generate candidate regions?



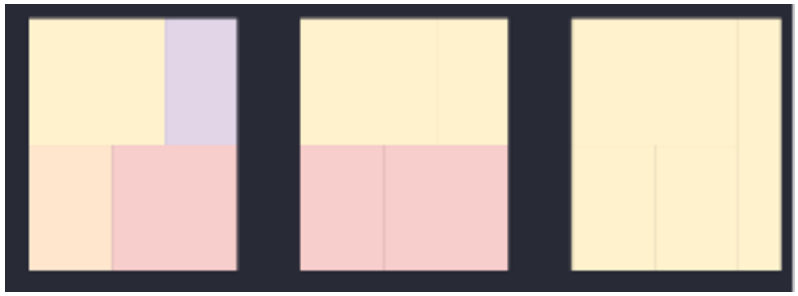


Region Proposals: Selective Search

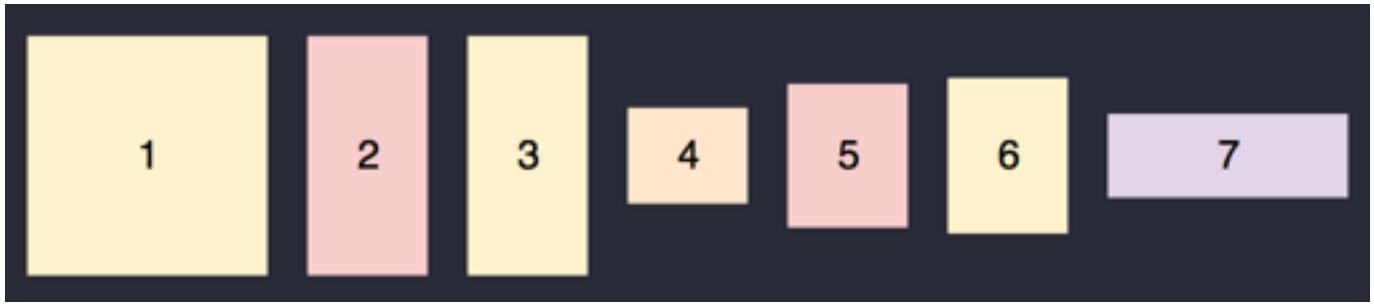
- 1. Generate Initial Sub-segmentation

efficient graph-based segmentation

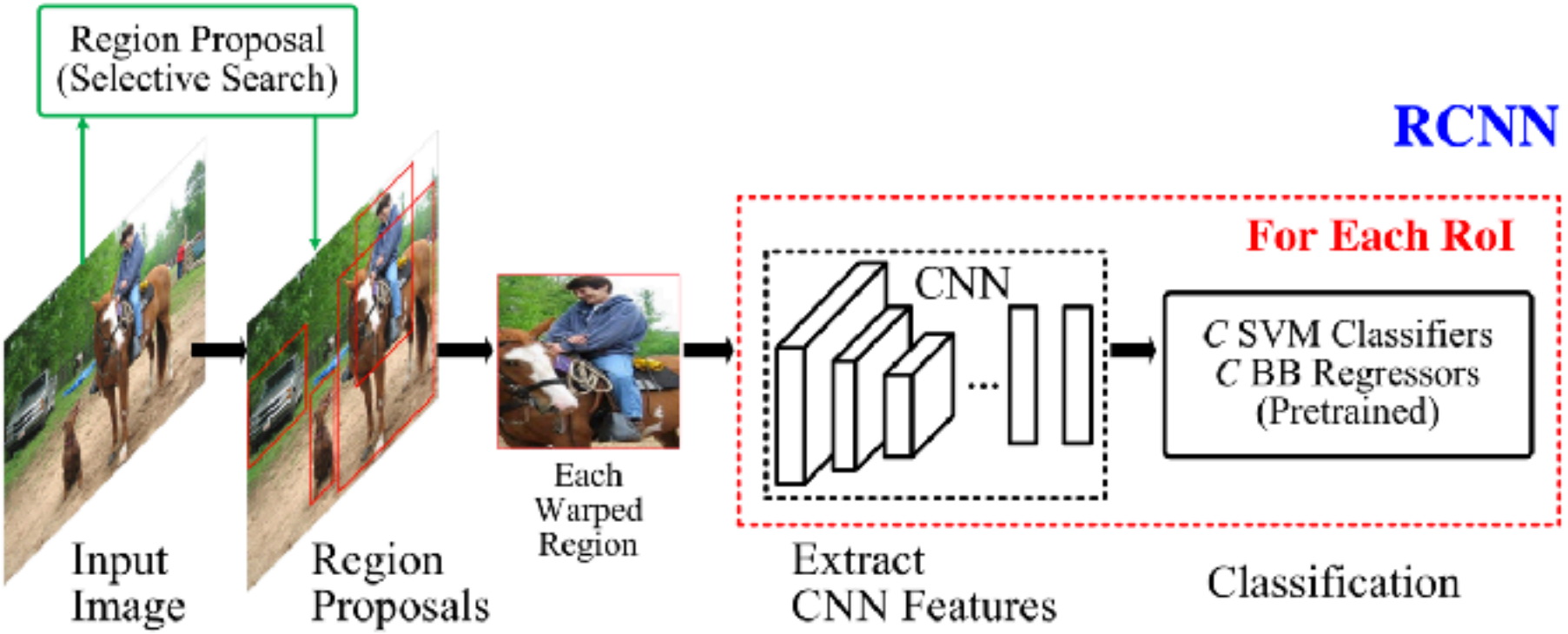
- 2. Combine similar regions into larger ones (color, texture, size)



- 3. Use the generated regions to produce the final candidate region proposals



Region CNN



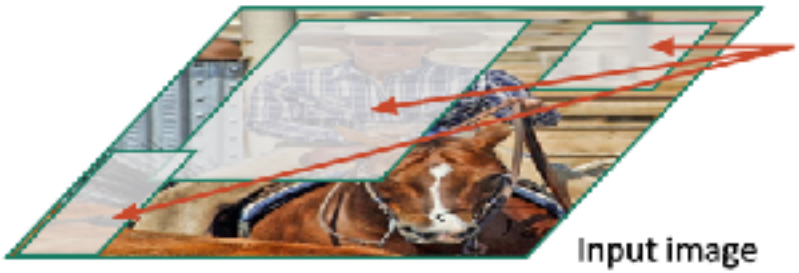
Li Liu et al. Deep Learning for Generic Object Detection: A Survey 2018

Region CNN

Input image

Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Region CNN

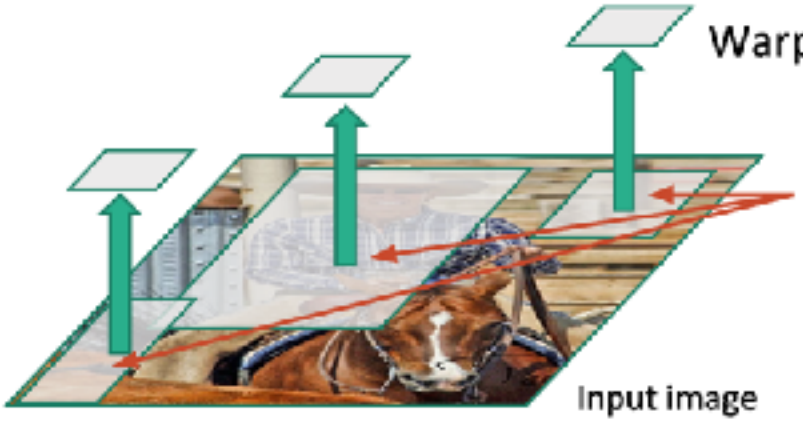


Input image

Regions of Interest (RoI) from a proposal method (~2k)

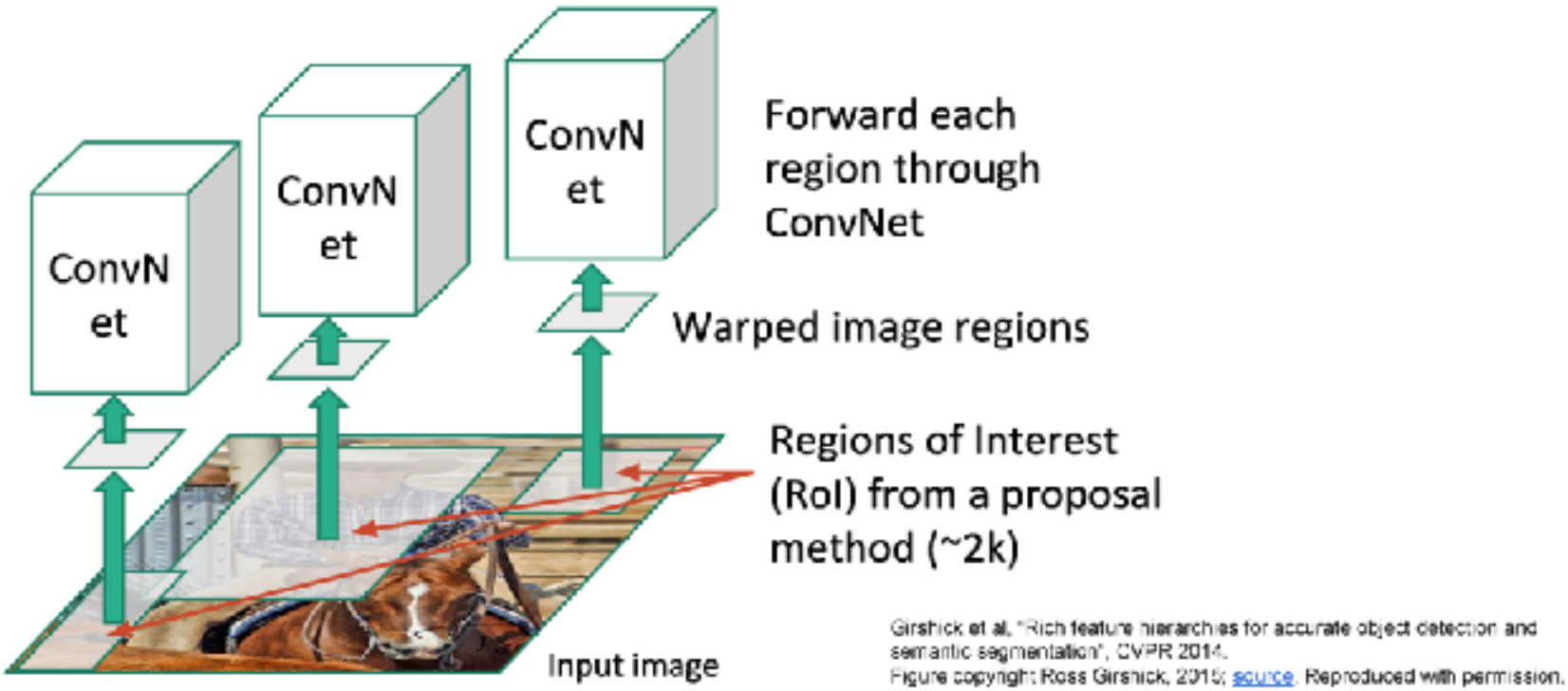
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Region CNN



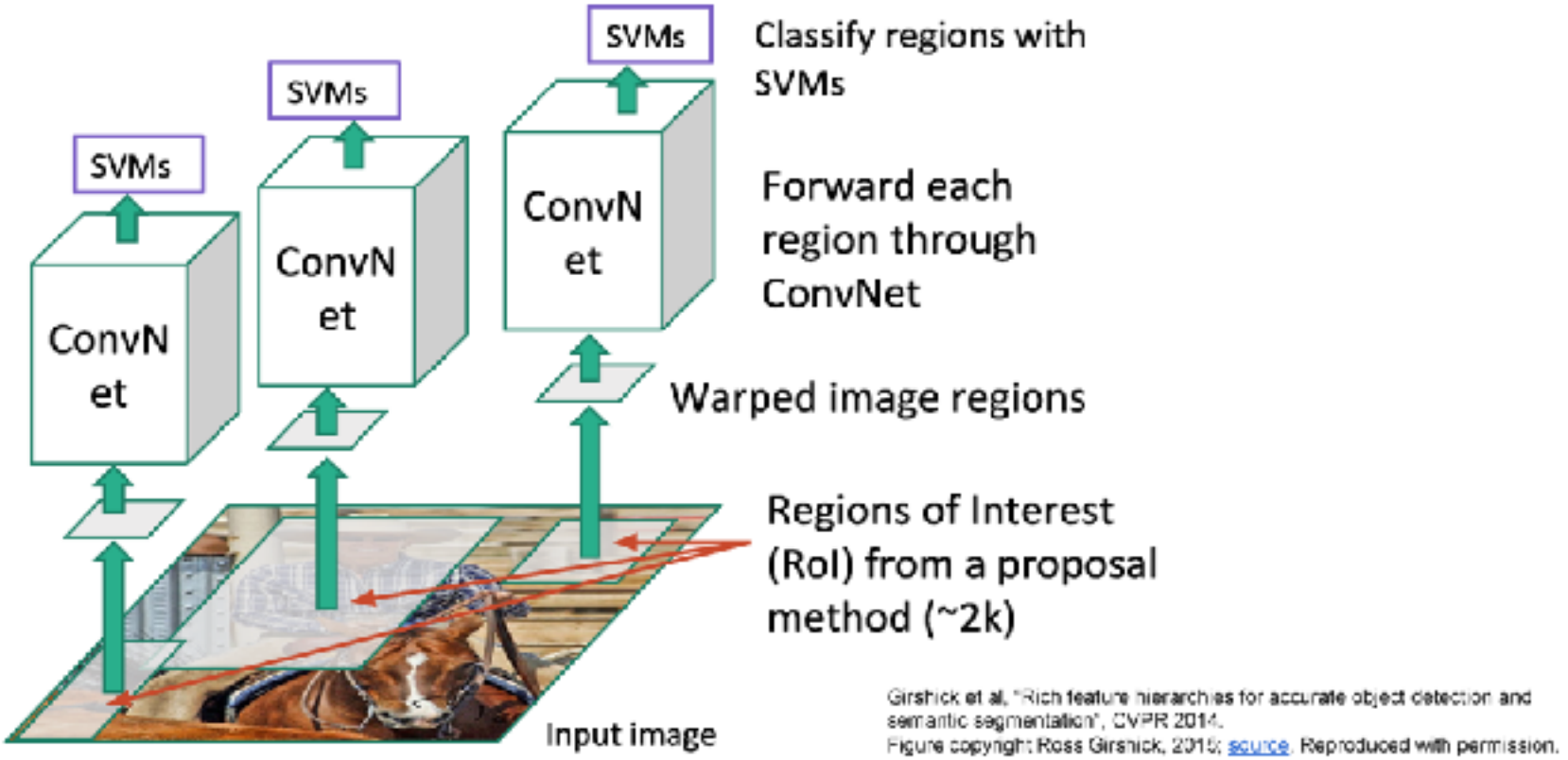
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Region CNN



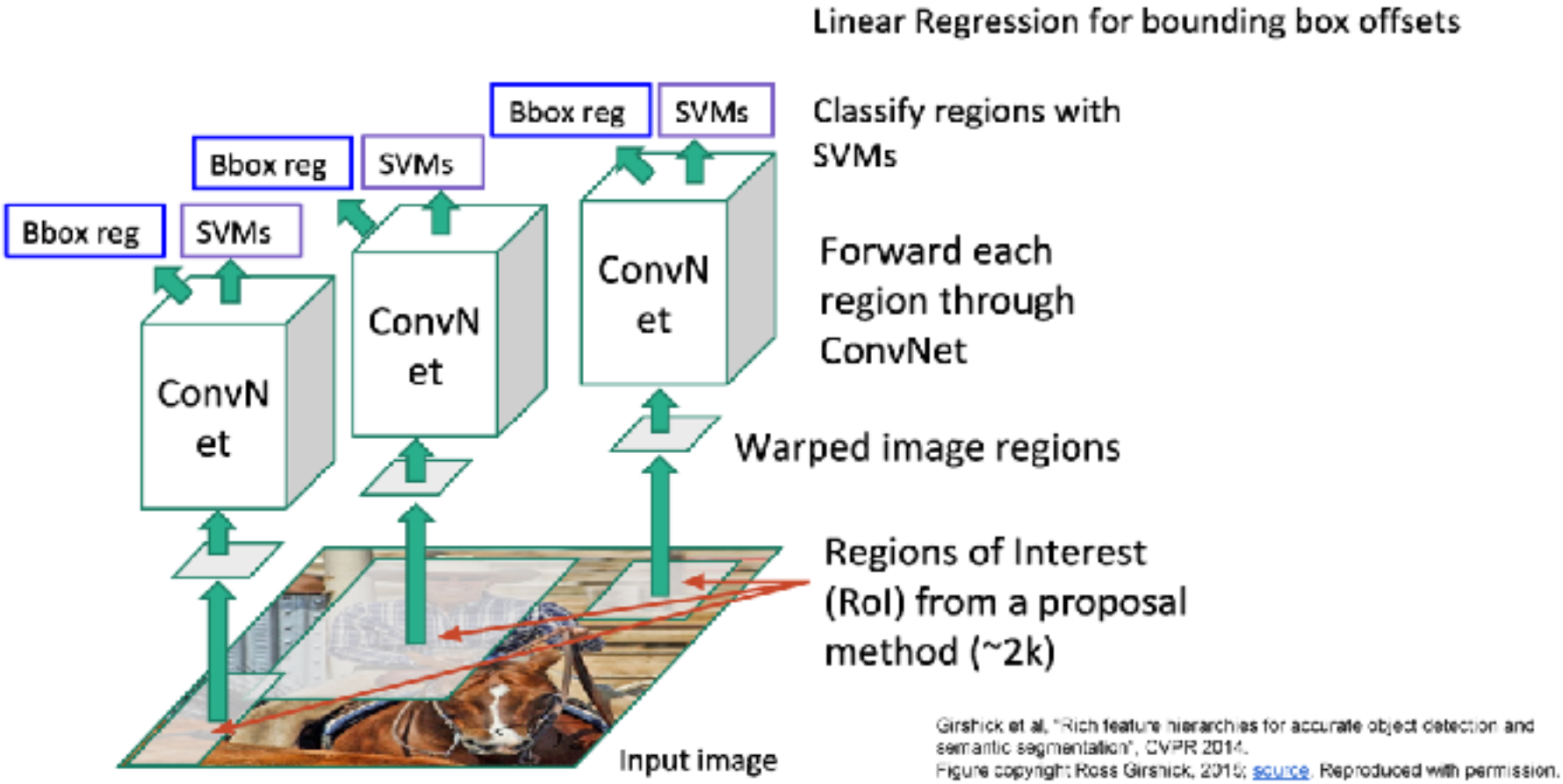
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Region CNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

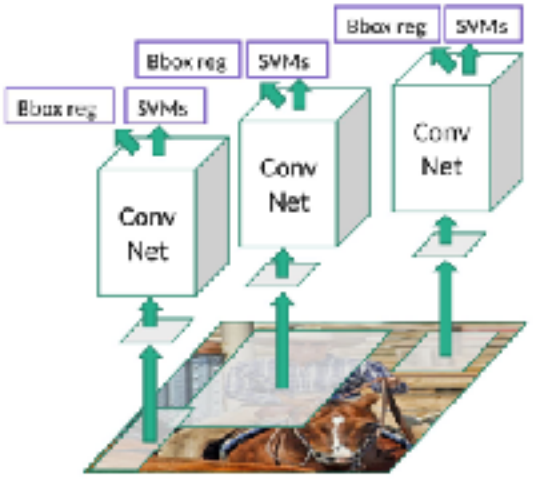
Region CNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Region CNN: Problems

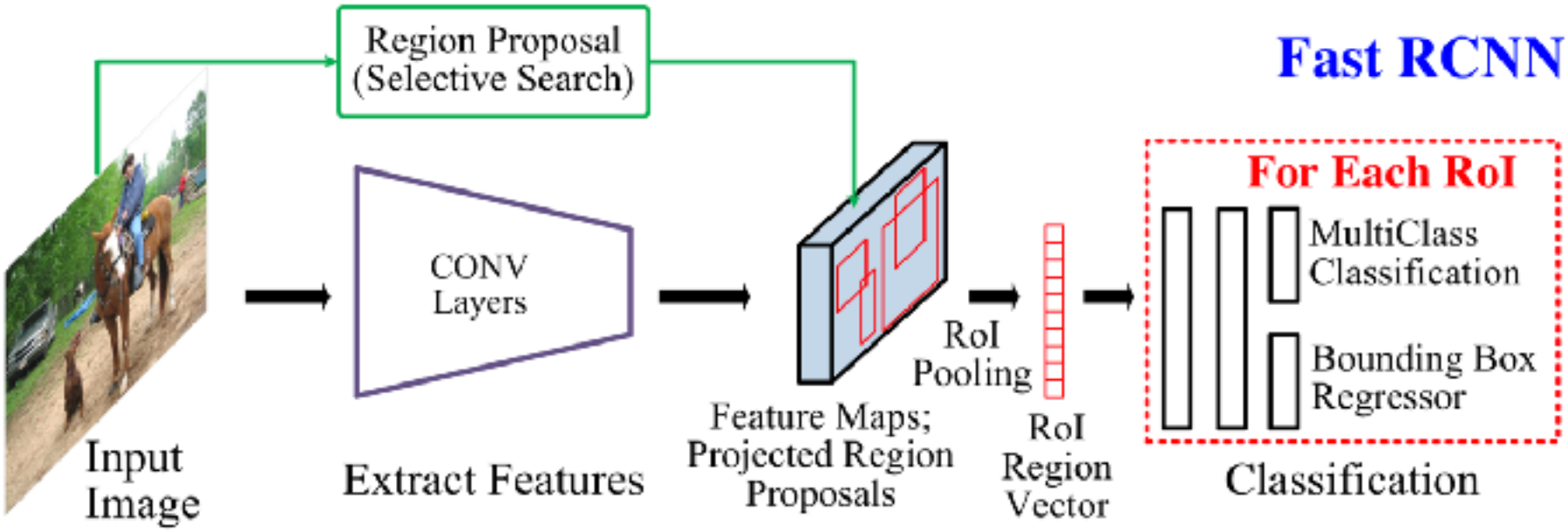
- Ad hoc training objectives
  - Fine-tune network with softmax classifier (log loss)
  - Train post-hoc linear SVMs (hinge loss)
  - Train post-hoc bounding-box regressions (least squares)
- Training is slow (84h), takes a lot of disk space
- Inference (detection) is slow
  - 47s / image with VGG16 [Simonyan & Zisserman. ICLR15]



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Slide copyright Ross Girshick, 2015; source. Reproduced with permission.

- Next: Fast RCNN, Faster RCNN

Fast RCNN



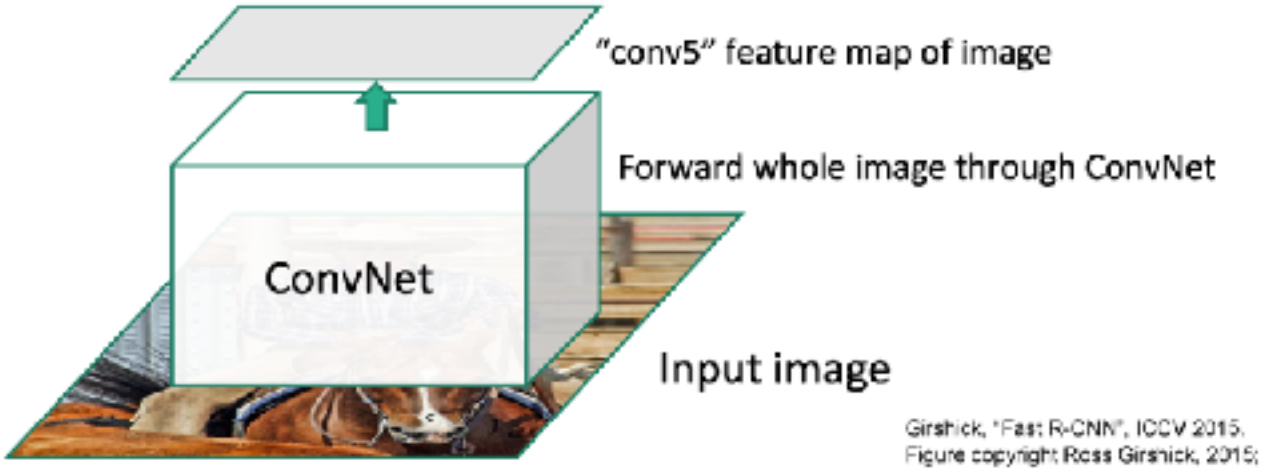
Li Liu et al. Deep Learning for Generic Object Detection: A Survey 2018

Fast RCNN

Input image

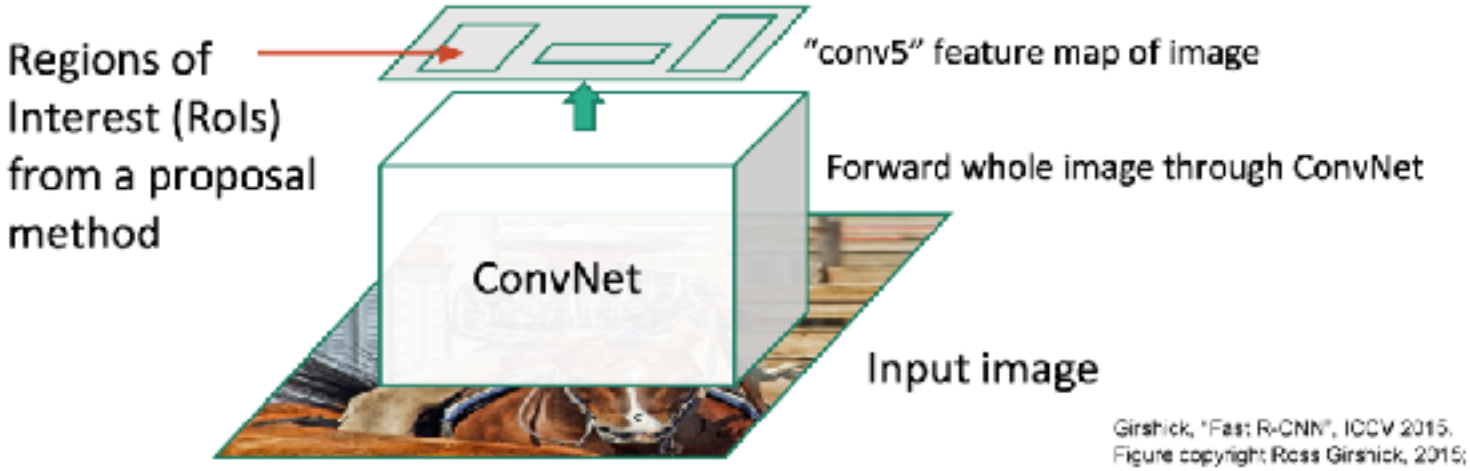
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast RCNN



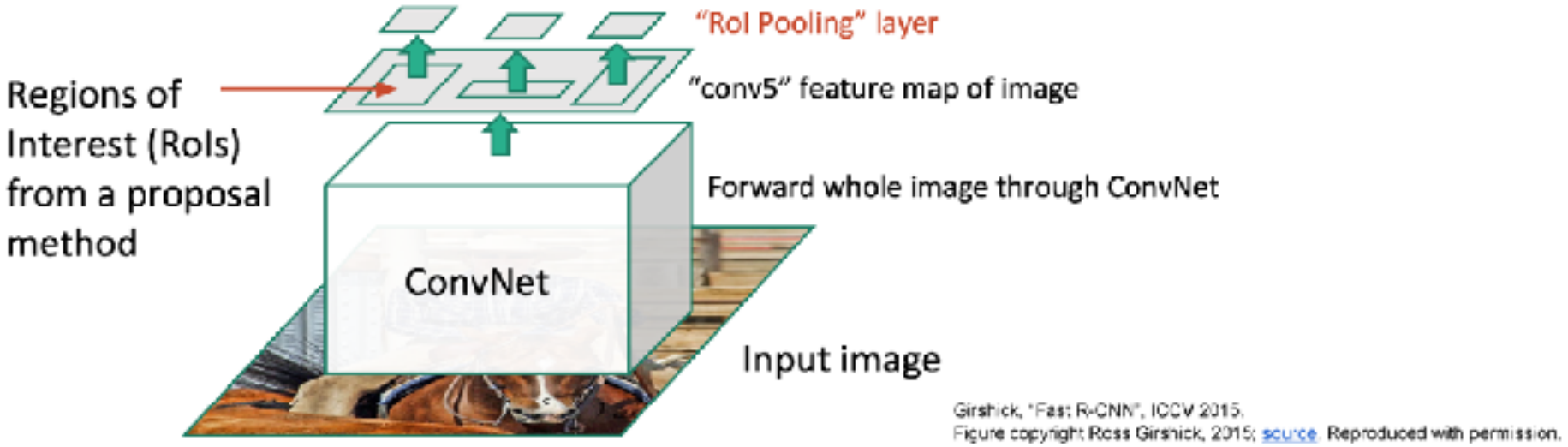
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast RCNN



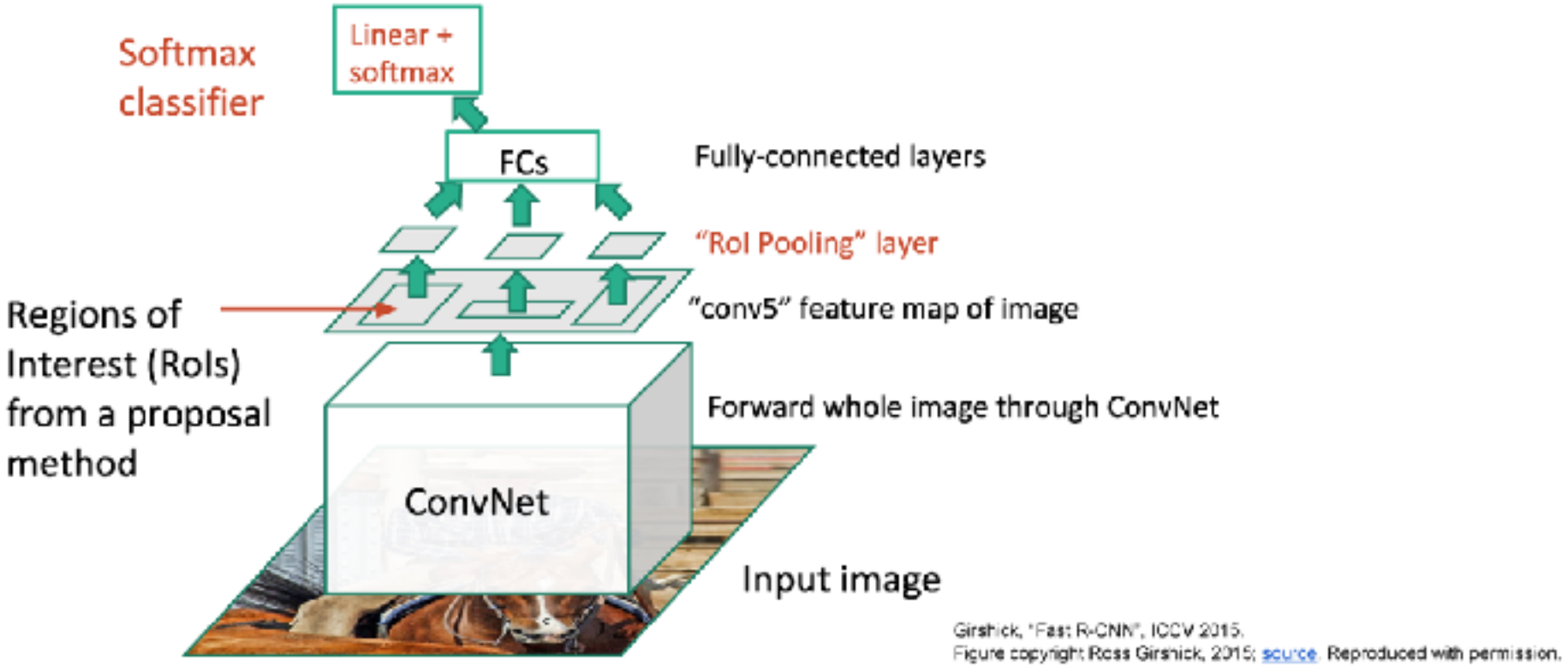
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast RCNN



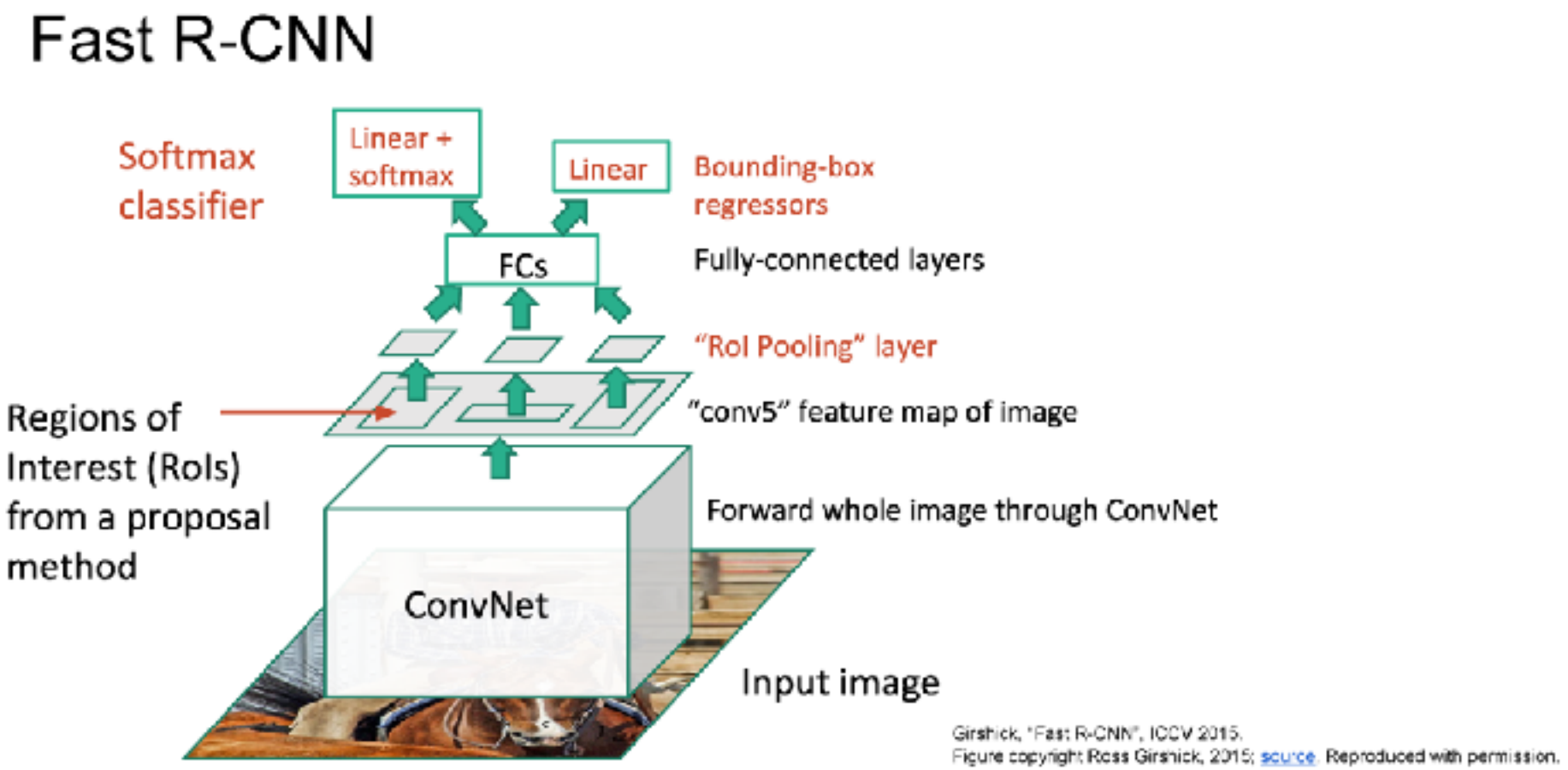
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast RCNN



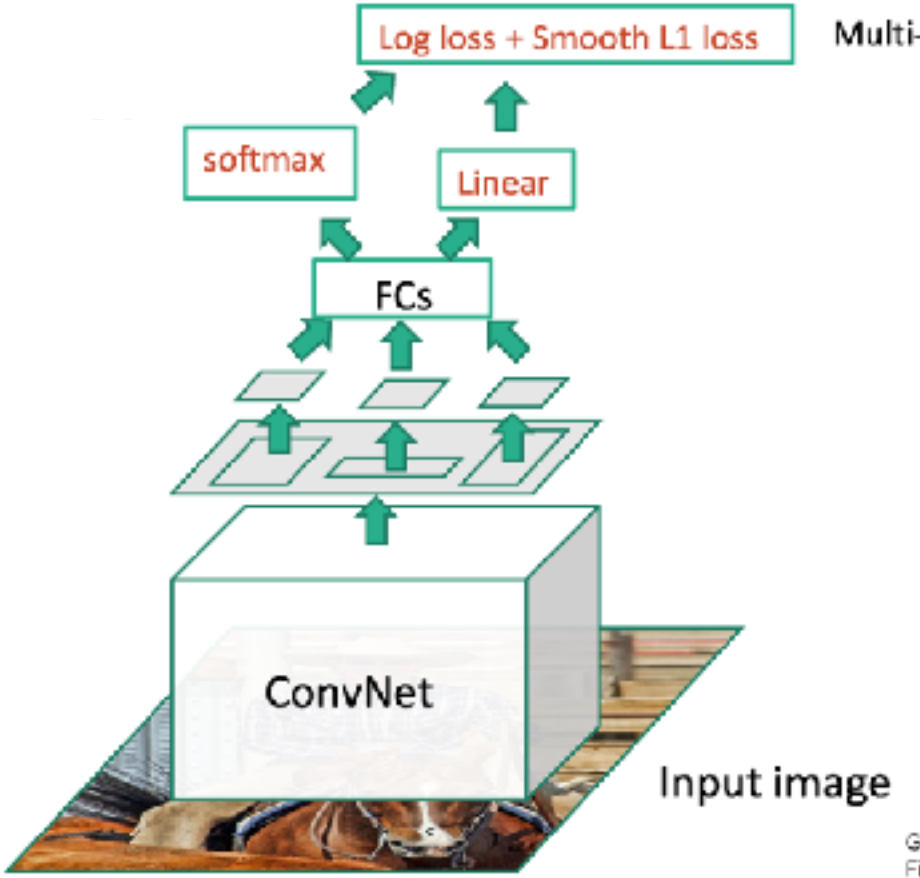
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast RCNN



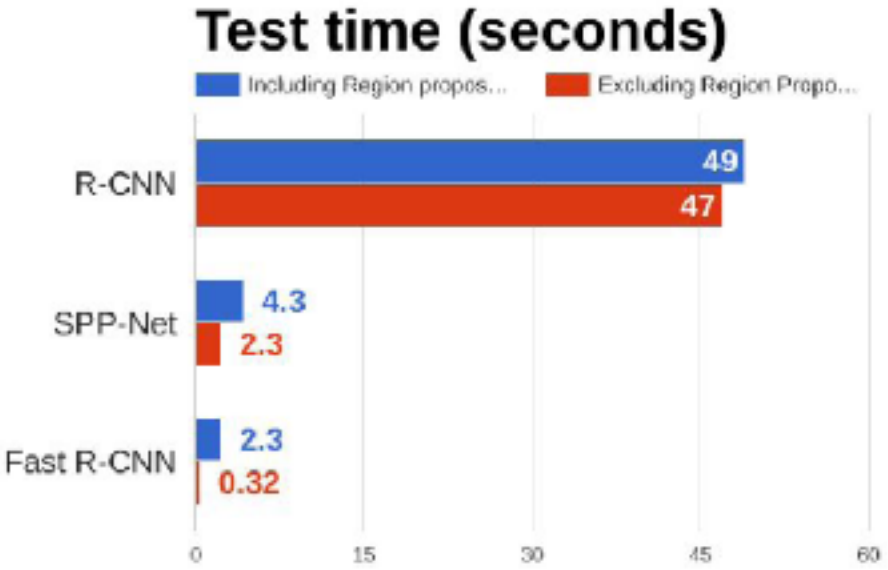
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast RCNN: Training



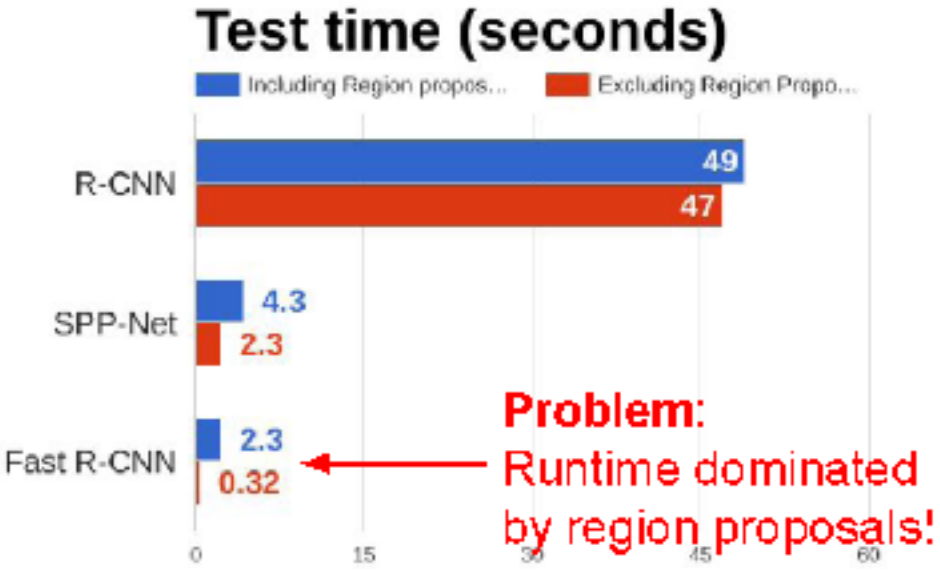
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

RCNN vs. Fast RCNN



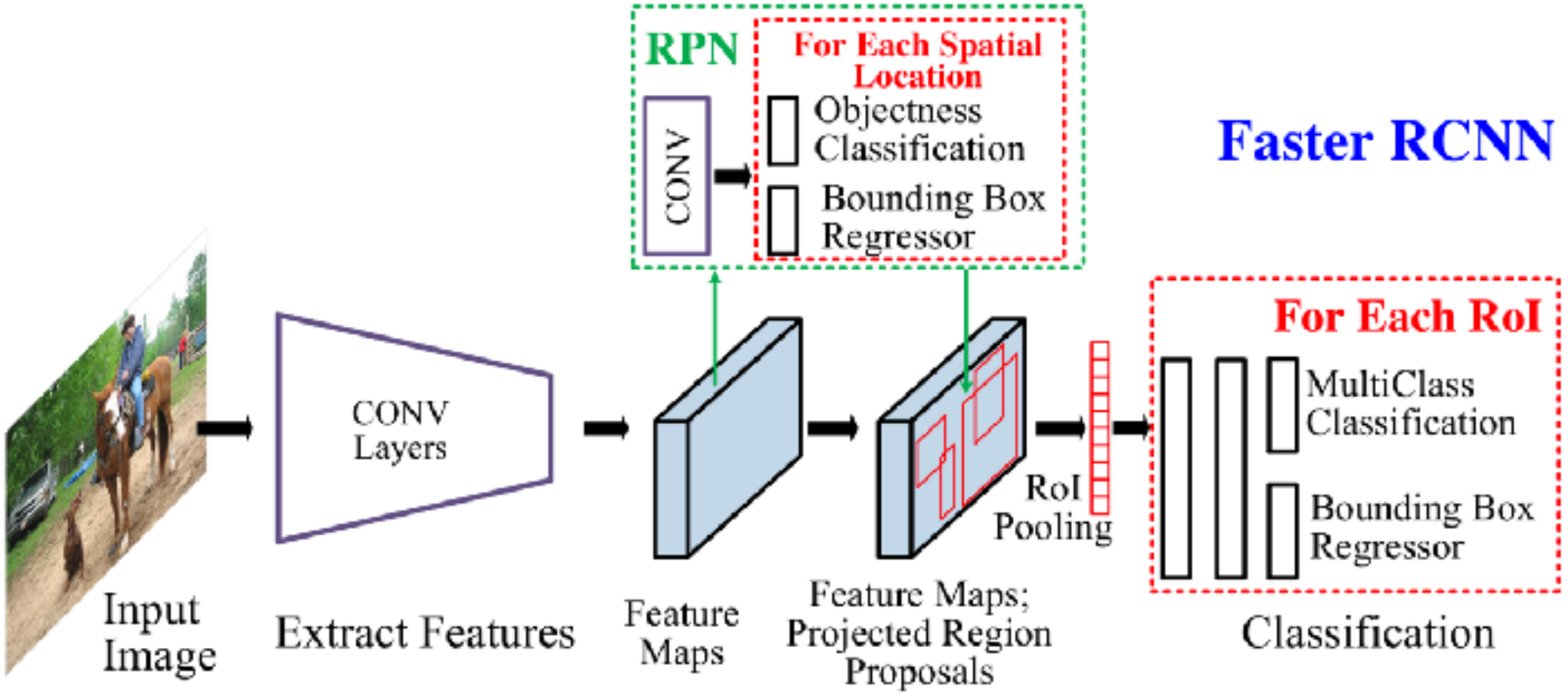
Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014
He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV 2014
Girshick, "Fast R-CNN", ICCV 2015

RCNN vs. Fast RCNN



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

Faster RCNN



Li Liu et al. Deep Learning for Generic Object Detection: A Survey 2018

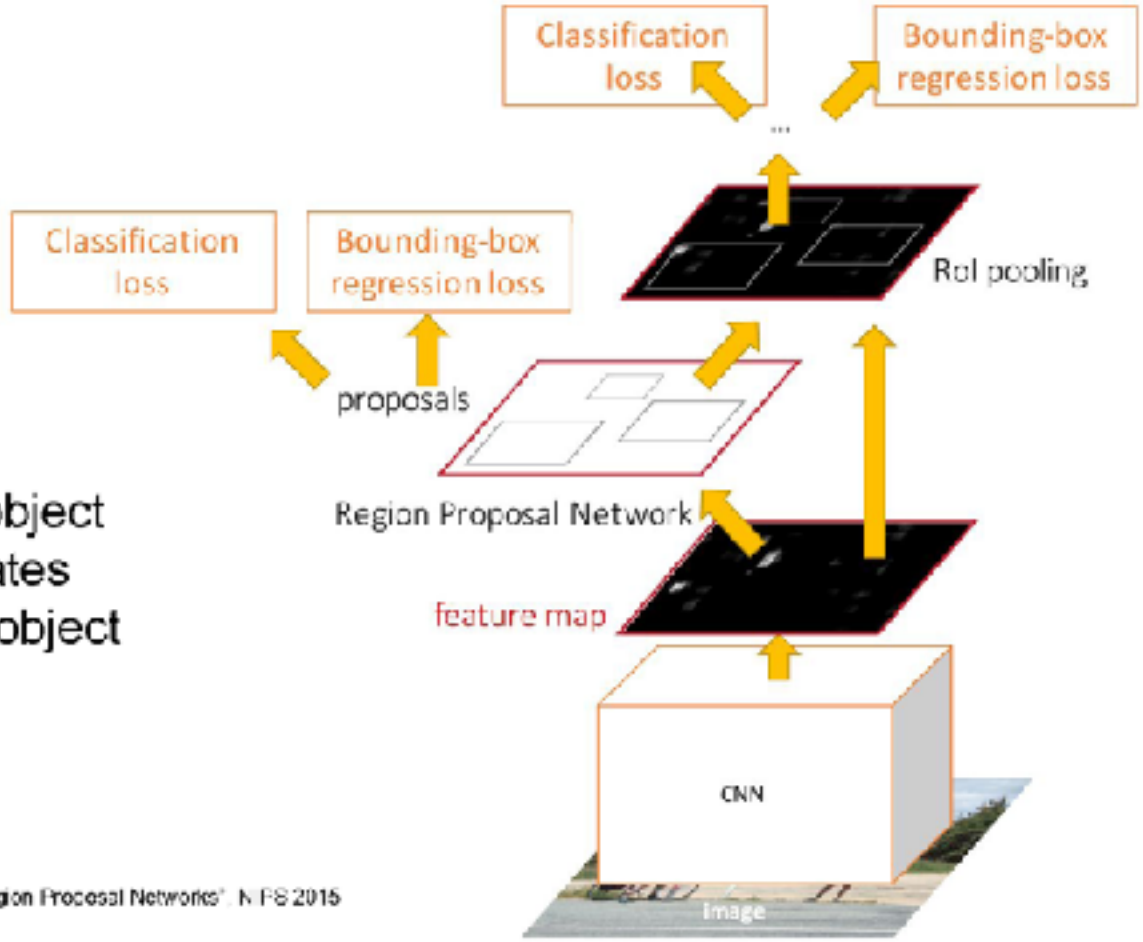
Faster RCNN

Make CNN do proposals!

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

Jointly train with 4 losses:

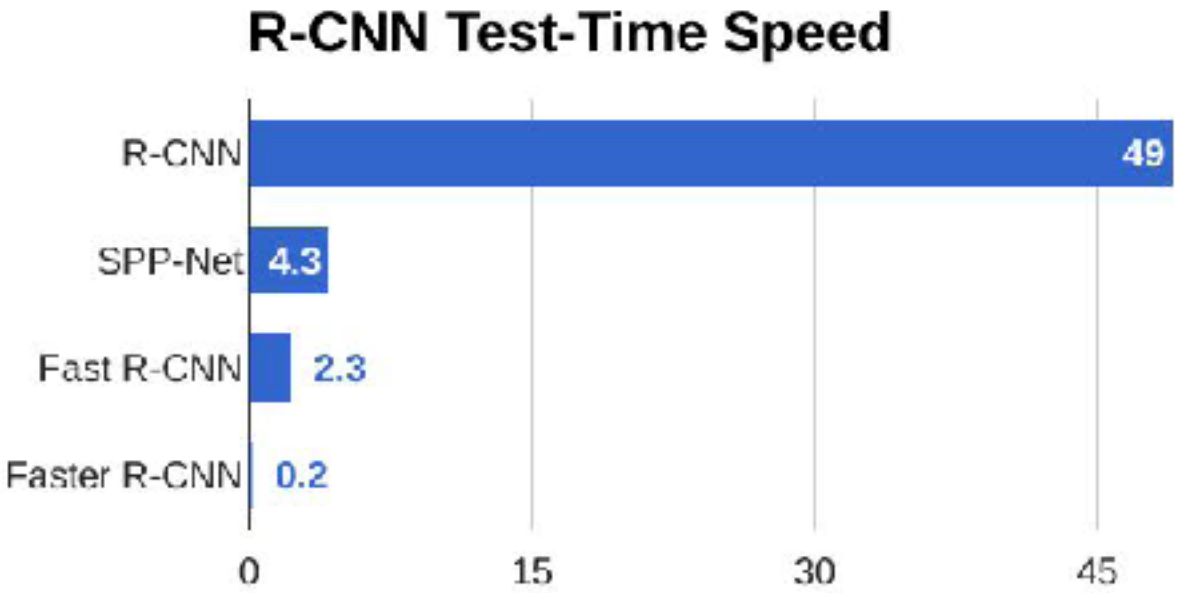
1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



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

Faster RCNN

Make CNN do proposals!



Pre Deep Era, RCNN, Fast RCNN, Faster RCNN

	Candidate Box Selection	**Feature Extraction**	**Classification**
Pre Deep Era	Exhaustive	Hand-crafted (*e.g.* HOG)	Linear
RCNN			
Fast RCNN Faster RCNN			

Pre Deep Era, RCNN, Fast RCNN, Faster RCNN

	Candidate Box Selection	Feature Extraction	Classification
Pre Deep Era	Exhaustive	Hand-crafted (*e.g.* HOG)	Linear
RCNN	Region Proposal	Deep	Linear
Fast RCNN			
Faster RCNN			

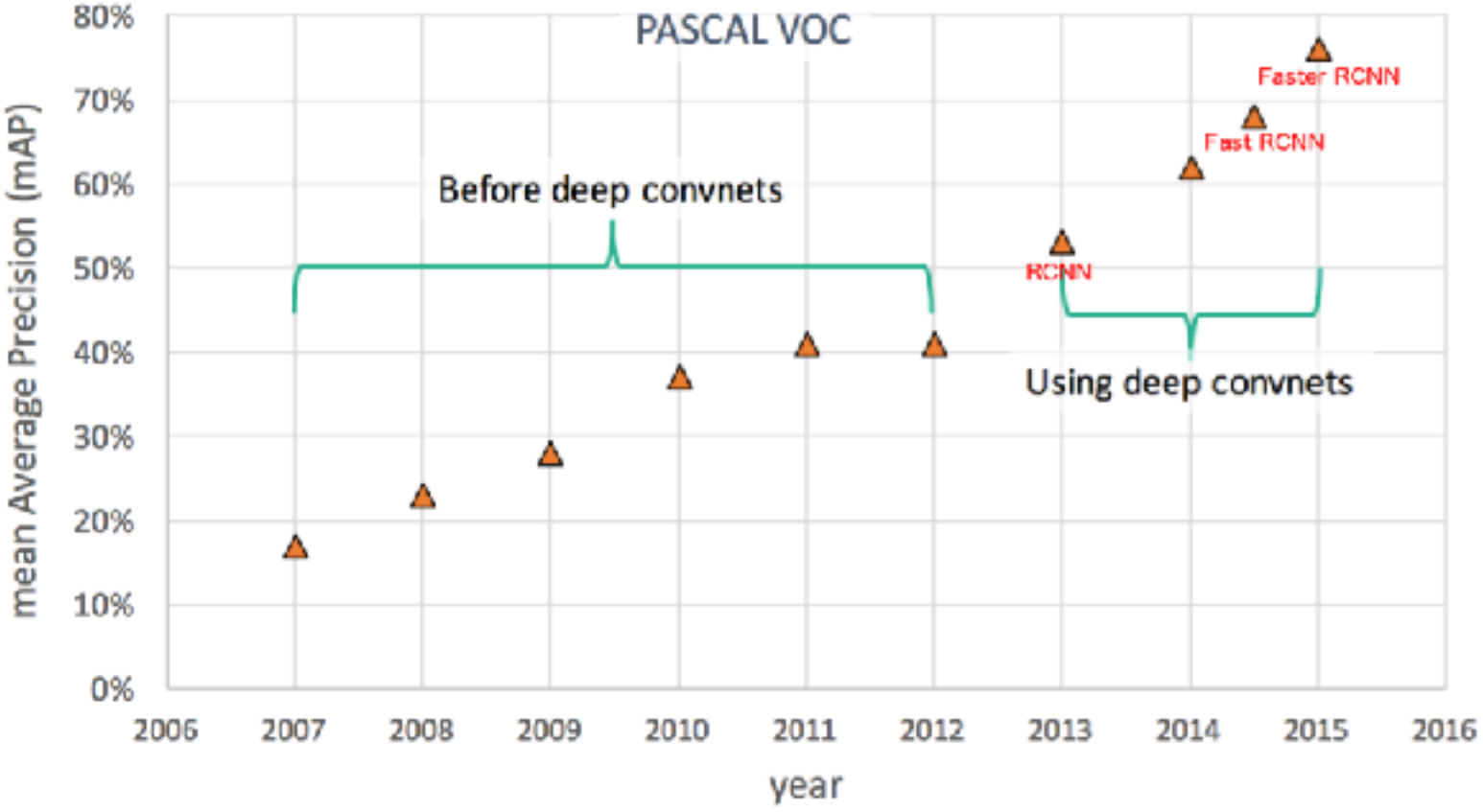
Pre Deep Era, RCNN, Fast RCNN, Faster RCNN

<table>
<tr><th></th><th>Candidate Box Selection</th><th>Feature Extraction</th><th>Classification</th></tr>
<tr><td>Pre Deep Era</td><td>Exhaustive</td><td>Hand-crafted (e.g. HOG)</td><td>Linear</td></tr>
<tr><td>RCNN</td><td>Region Proposal</td><td>Deep</td><td>Linear</td></tr>
<tr><td>Fast RCNN</td><td>Region Proposal</td><td colspan="2">Deep</td></tr>
<tr><td>Faster RCNN</td><td></td><td></td><td></td></tr>
</table>

Pre Deep Era, RCNN, Fast RCNN, Faster RCNN

	Candidate Box Selection	Feature Extraction	Classification
Pre Deep Era	Exhaustive	Hand-crafted (*e.g.* HOG)	Linear
RCNN	Region Proposal	Deep	Linear
Fast RCNN	Region Proposal	Deep	
Faster RCNN	Deep	Deep	

Improvements using Deep ConvNets



Source: *Ross Girshick*