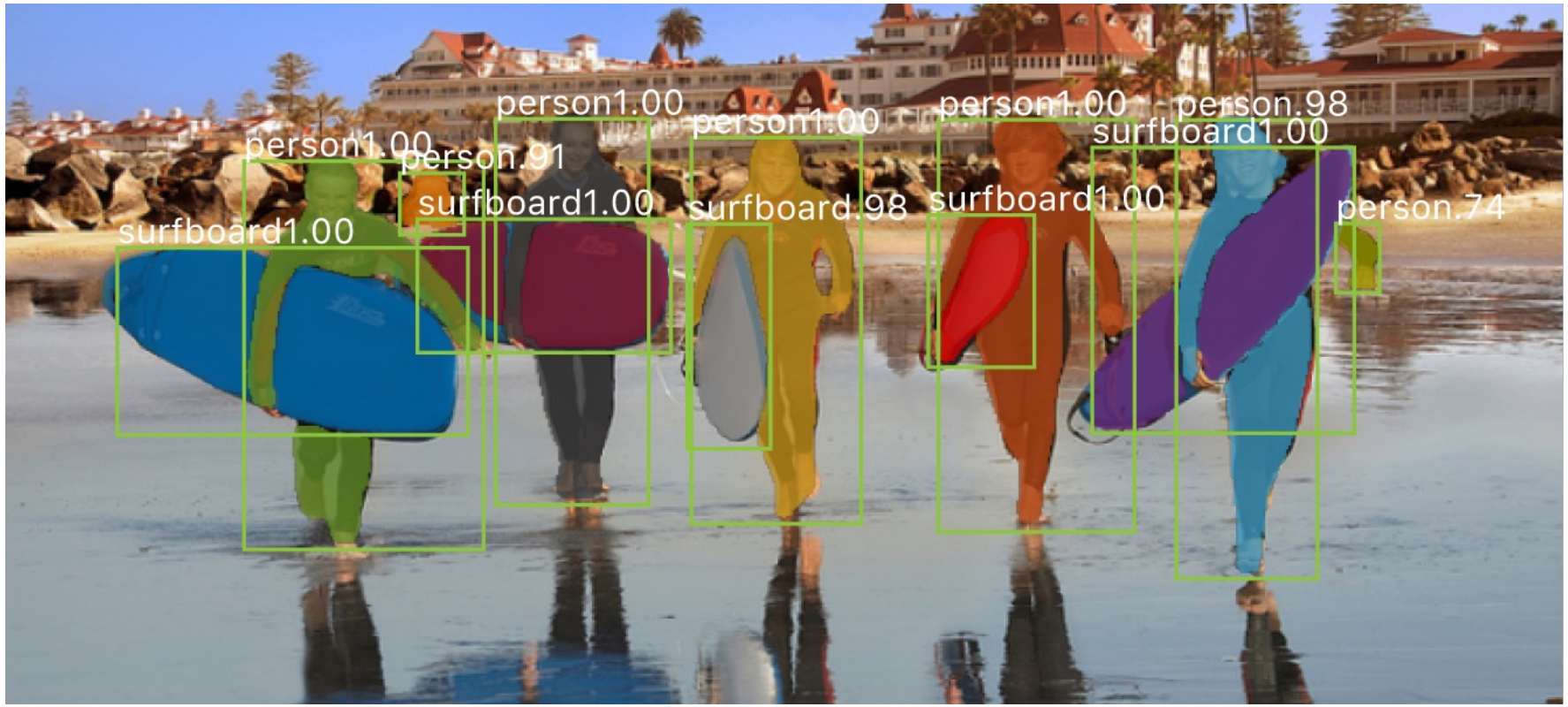
Lecture 13

Object Detection and Image Segmentation



Overview

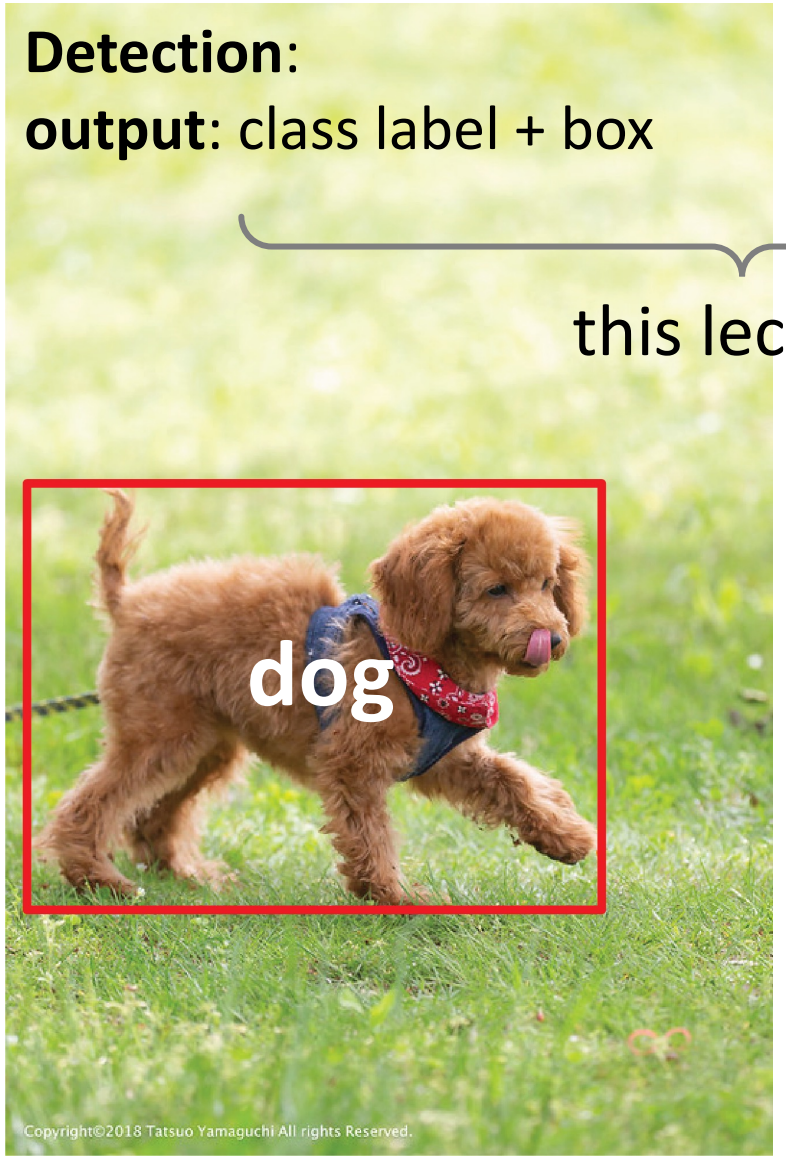
- Visual Recognition Tasks
- Methodology: ConvNets, Fully Conv Nets
- Semantic Segmentation
- Object Detection
- Instance Segmentation

Visual Recognition Tasks

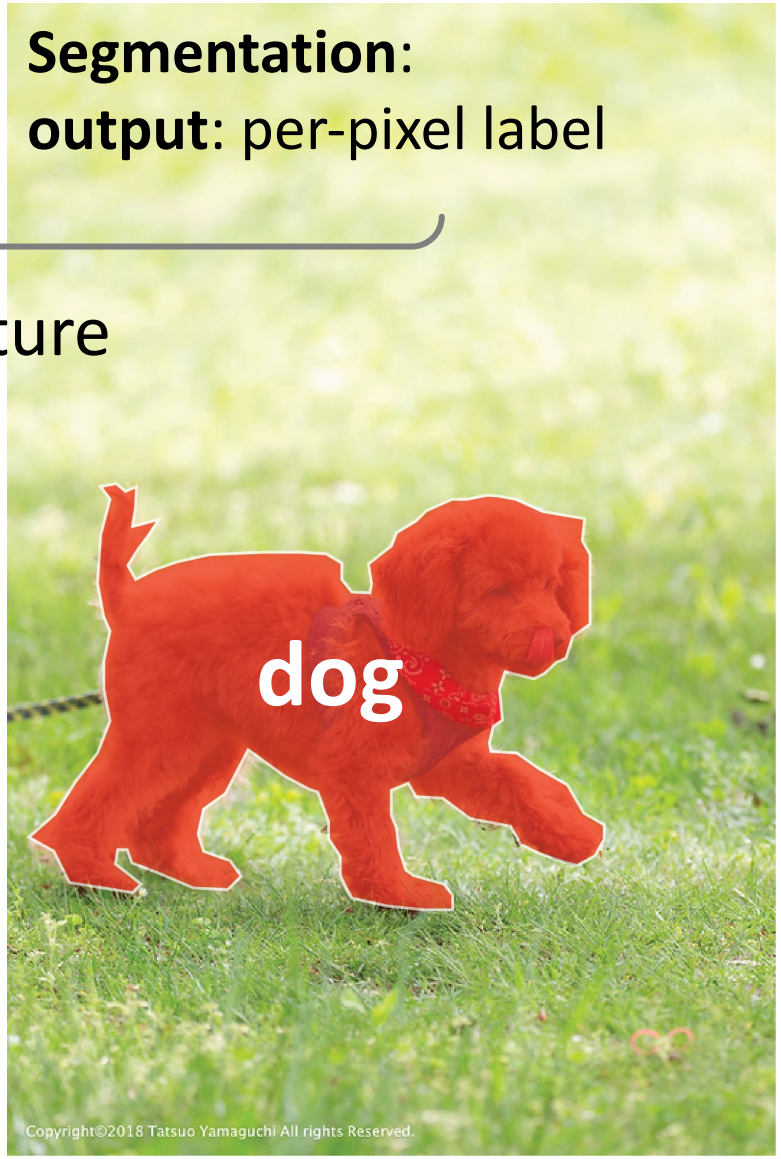
Classification:
output: class label

Detection:
output: class label + box

this lecture



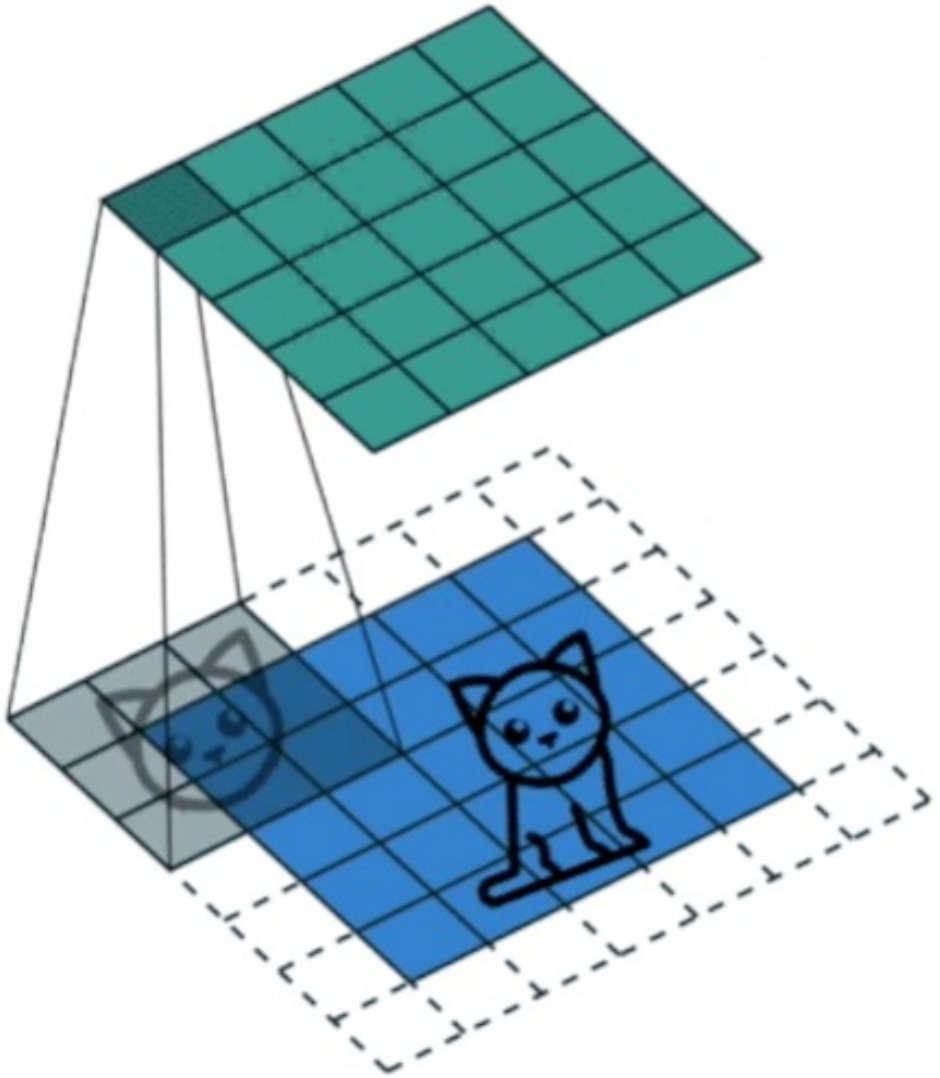
Segmentation:
output: per-pixel label



Methodology:
ConvNets, Fully Conv Nets

Convolution is template matching

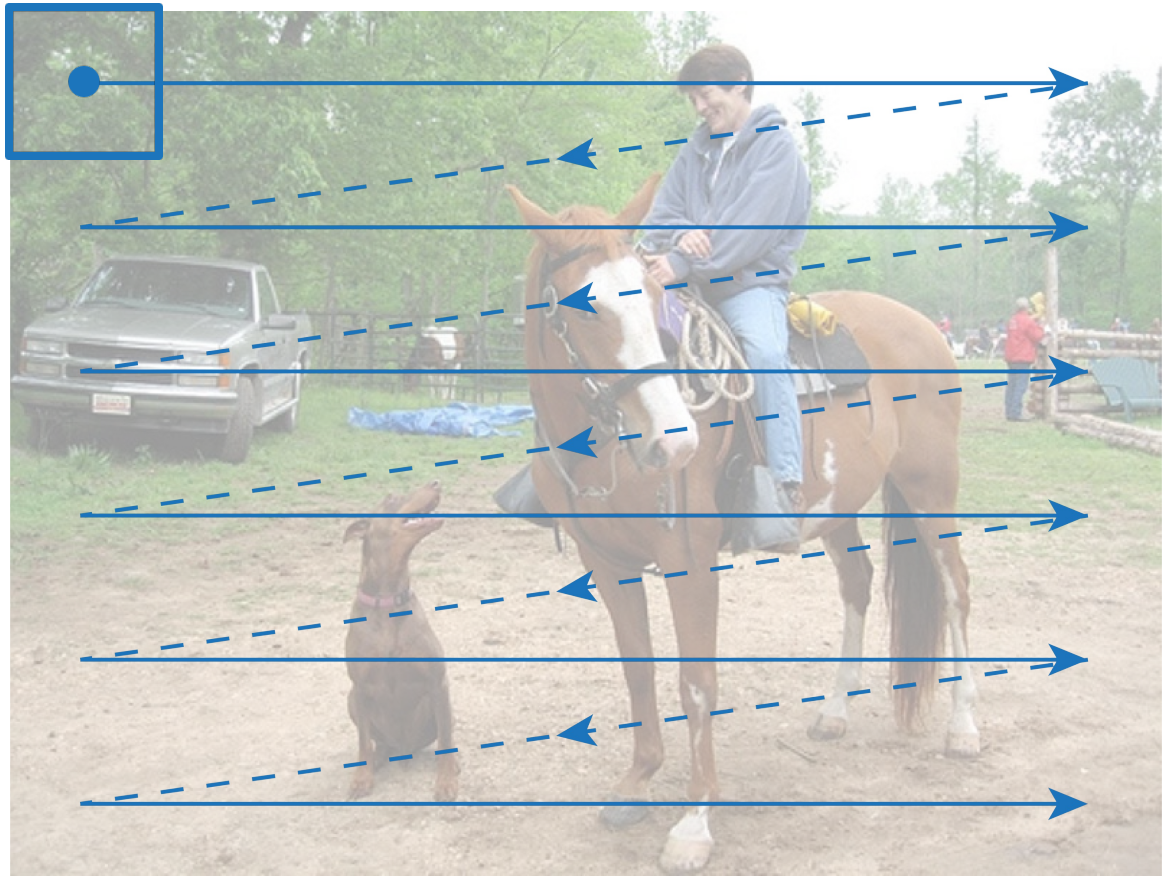
- sliding window
- abstract templates



Animation adapted from: https://lena-voita.github.io/nlp_course/models/convolutional.html

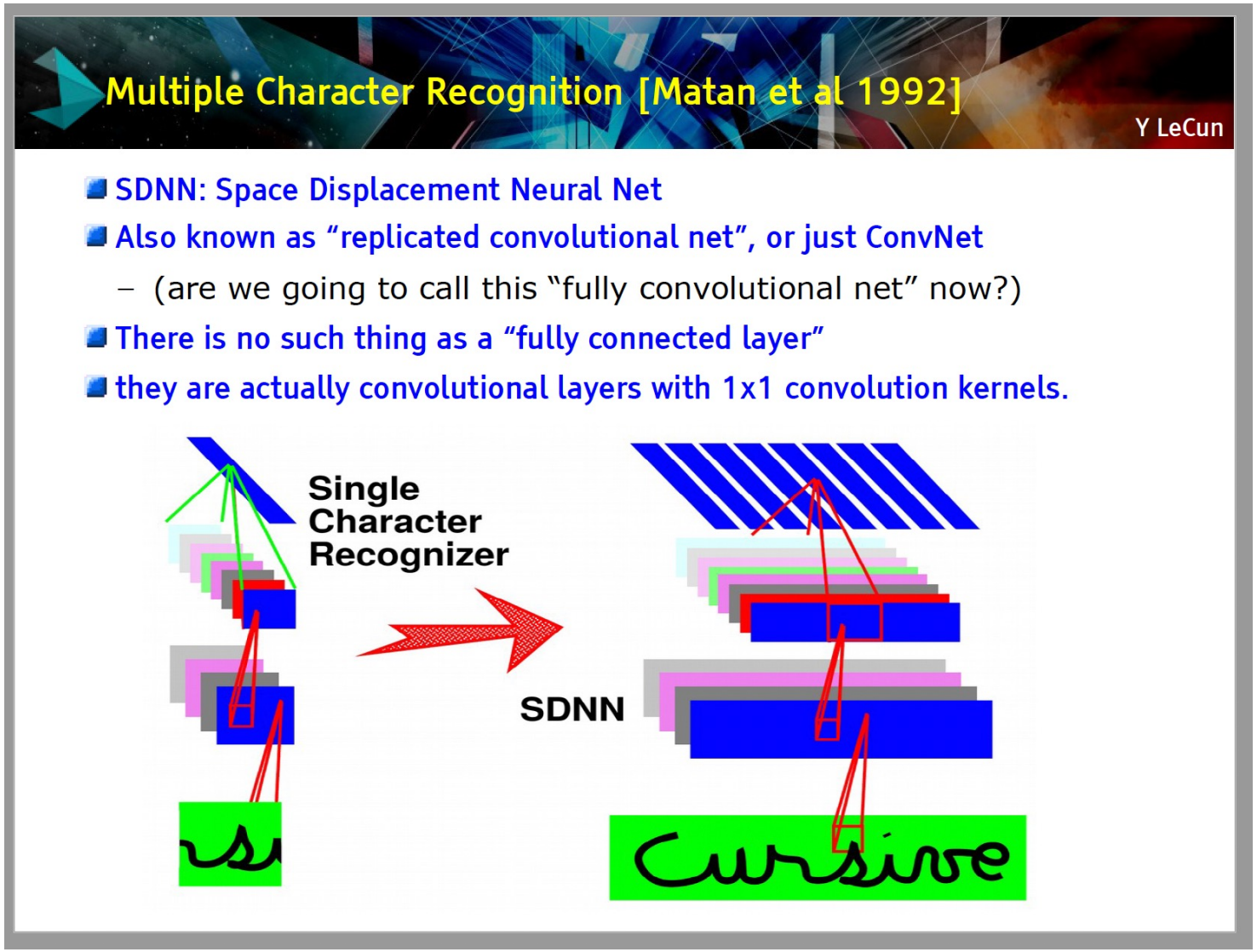
Convolution is template matching

- sliding window
- abstract templates



ConvNets Are Fully Convolutional

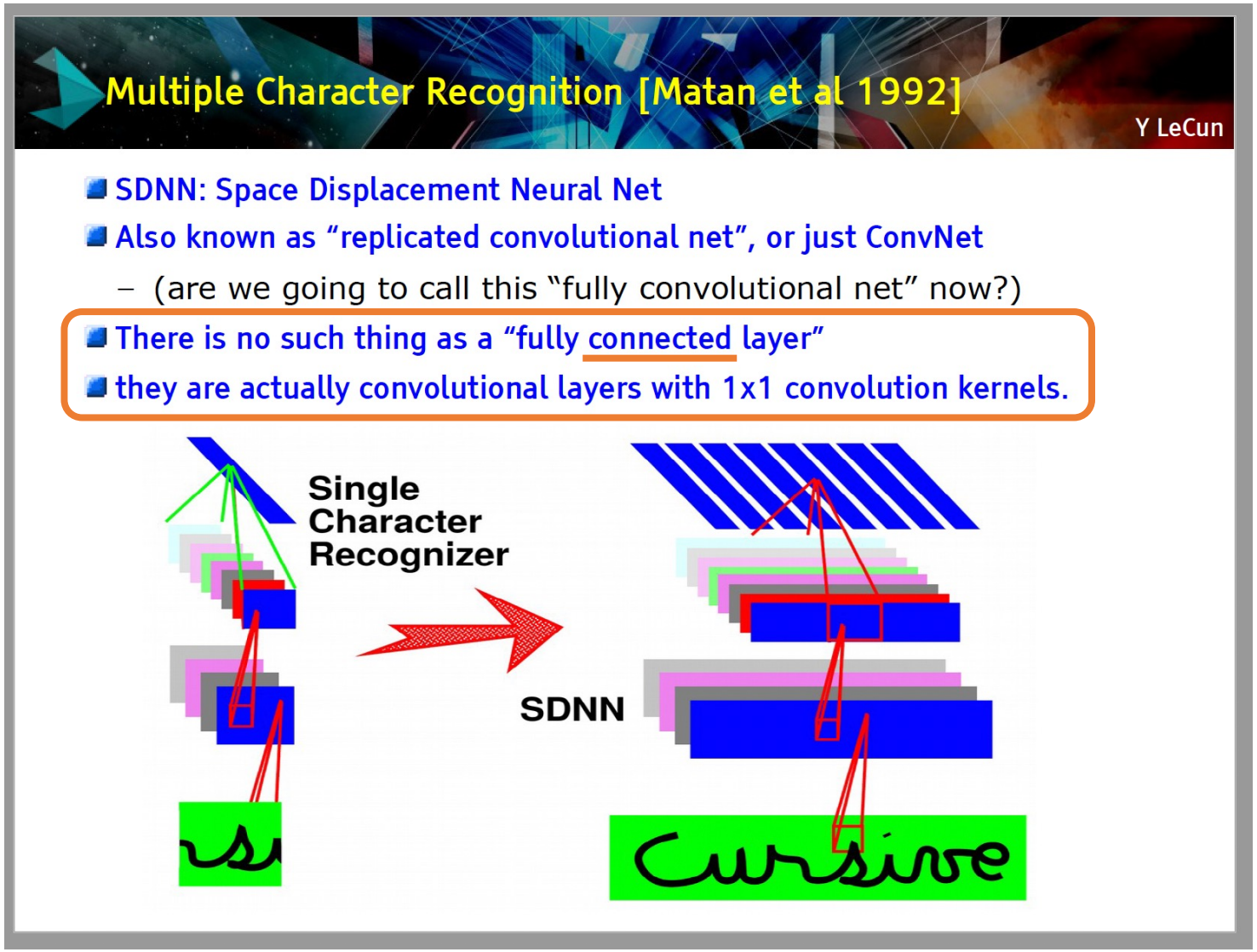
Screenshot from: “What's Wrong With Deep Learning?” Yann LeCun, 2015



https://dl.icdst.org/pdfs/files4/23e2b3914c137d1a3ee1a81ada45ed4b.pdf

ConvNets Are Fully Convolutional

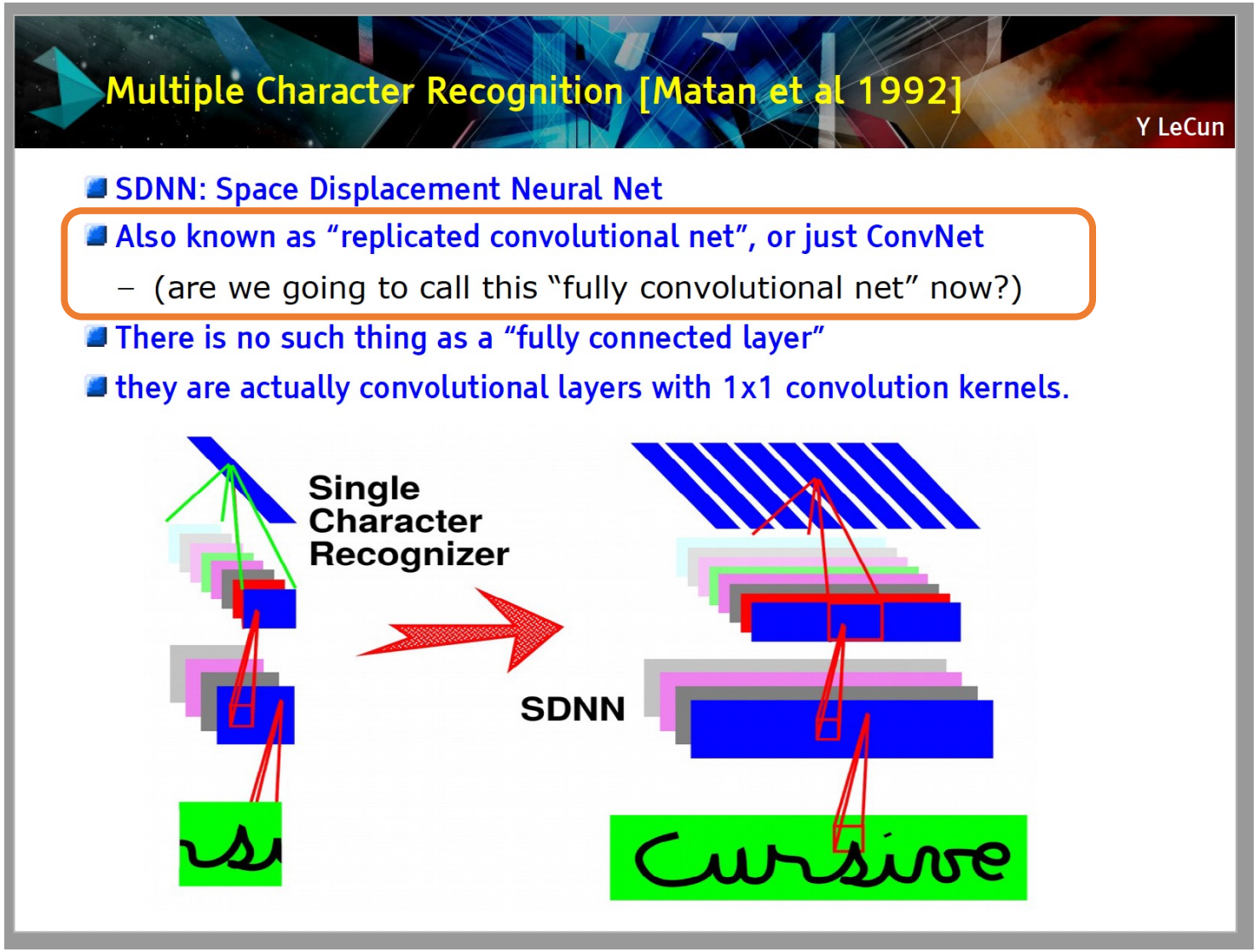
Screenshot from: “What's Wrong With Deep Learning?” Yann LeCun, 2015



https://dl.icdst.org/pdfs/files4/23e2b3914c137d1a3ee1a81ada45ed4b.pdf

ConvNets Are Fully Convolutional

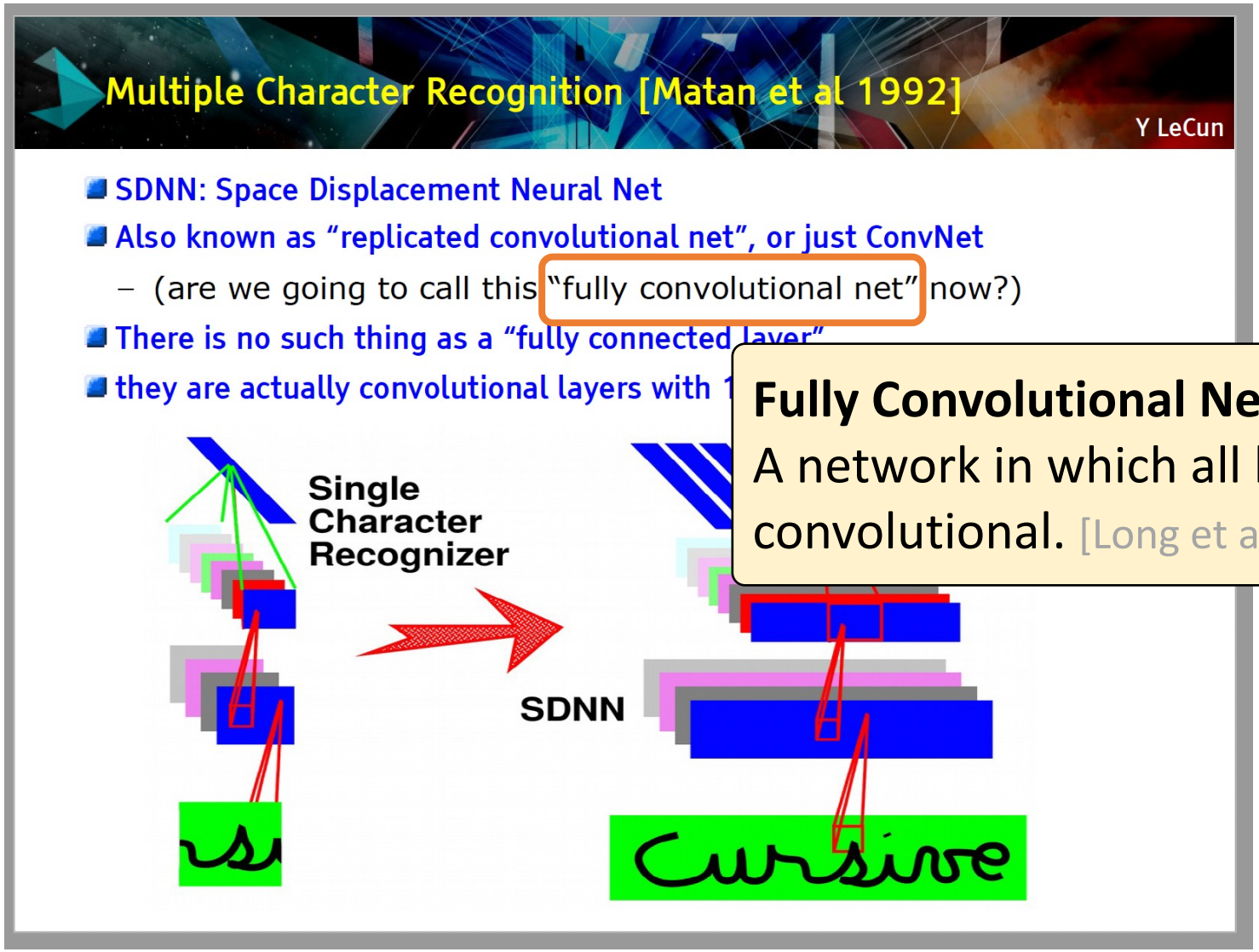
Screenshot from: “What's Wrong With Deep Learning?” Yann LeCun, 2015



https://dl.icdst.org/pdfs/files4/23e2b3914c137d1a3ee1a81ada45ed4b.pdf

ConvNets Are Fully Convolutional

Screenshot from: “What's Wrong With Deep Learning?” Yann LeCun, 2015



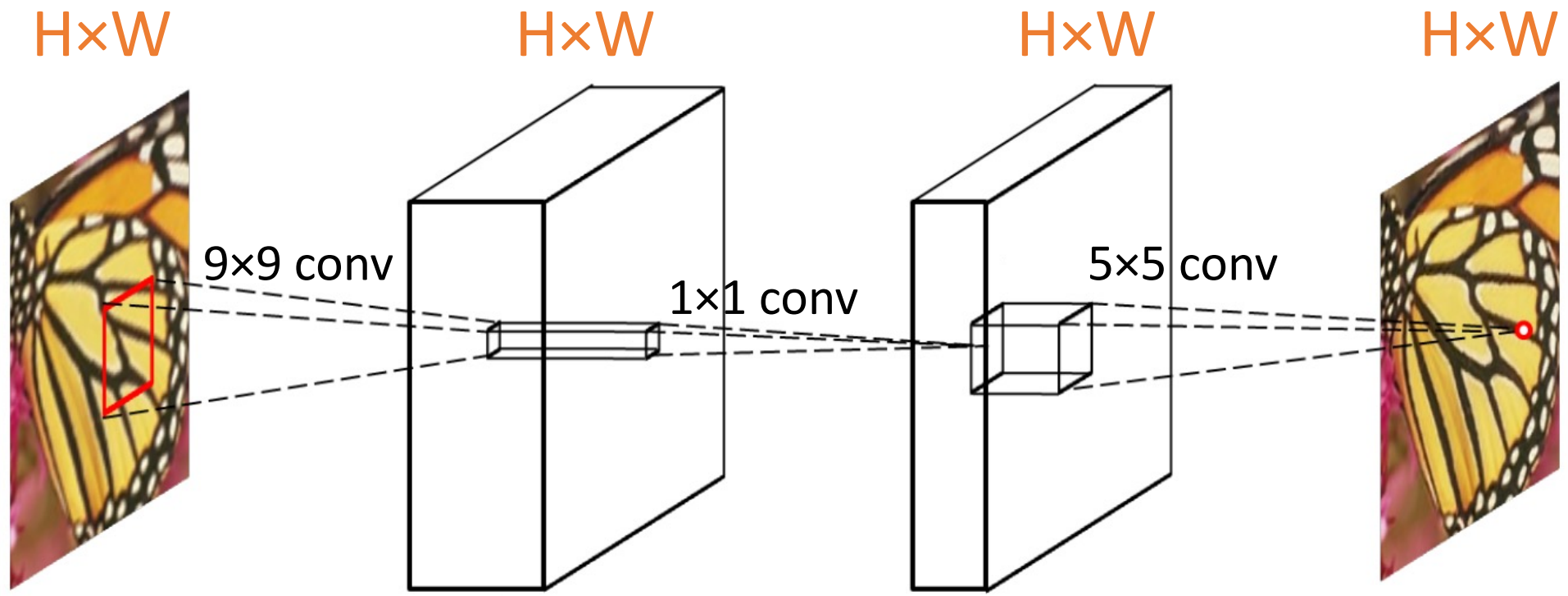
Fully Convolutional Net (FCN): A network in which all layers are convolutional. [Long et al. CVPR 2015]

https://dl.icdst.org/pdfs/files4/23e2b3914c137d1a3ee1a81ada45ed4b.pdf

ConvNets Are Fully Convolutional

Example: **Super-Resolution CNN (SRCNN)**

- arbitrary input size
- proportional output size

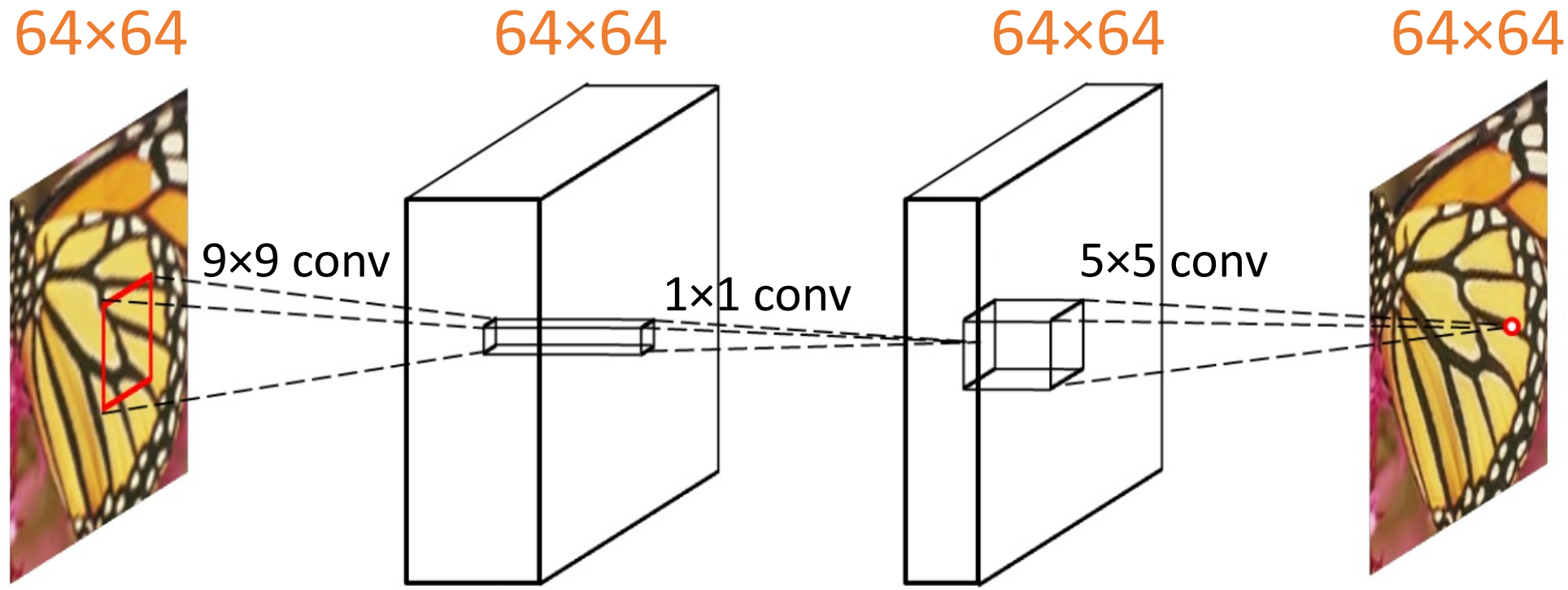


Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, “Learning a Deep Convolutional Network for Image Super-Resolution”, ECCV 2014

ConvNets Are Fully Convolutional

Example: **Super-Resolution CNN (SRCNN)**

- arbitrary input size
- proportional output size

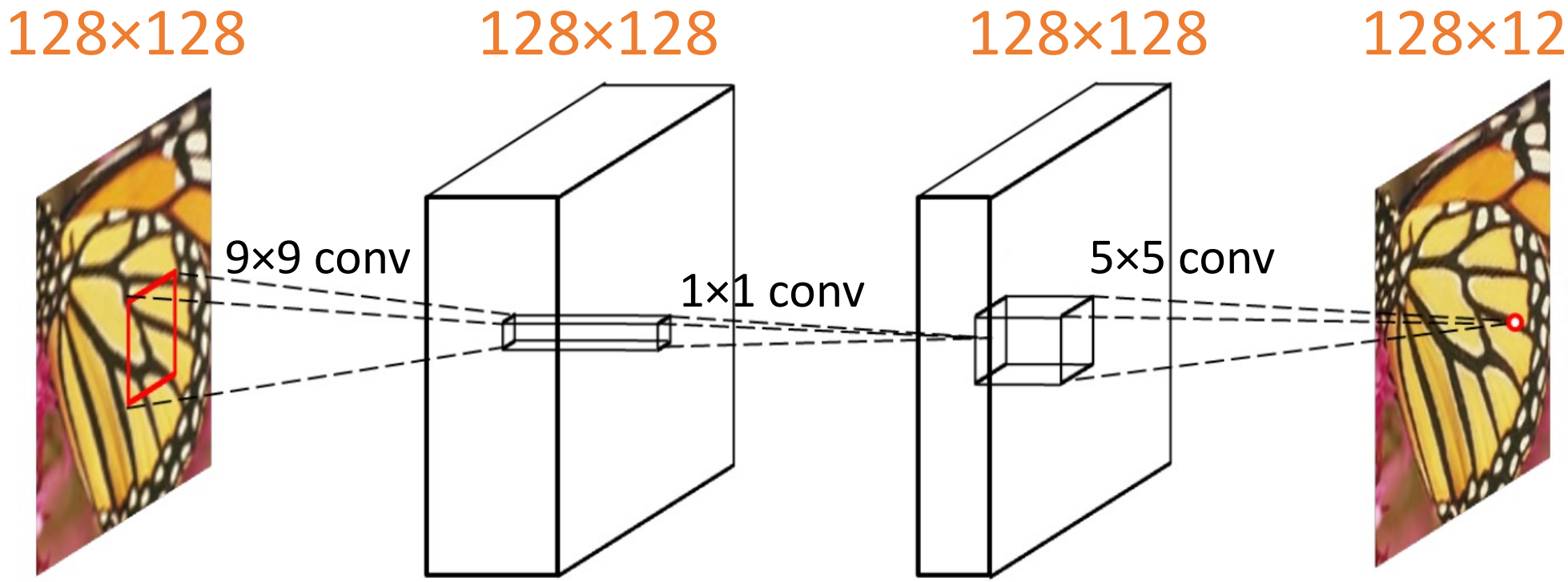


Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, “Learning a Deep Convolutional Network for Image Super-Resolution”, ECCV 2014

ConvNets Are Fully Convolutional

Example: **Super-Resolution CNN (SRCNN)**

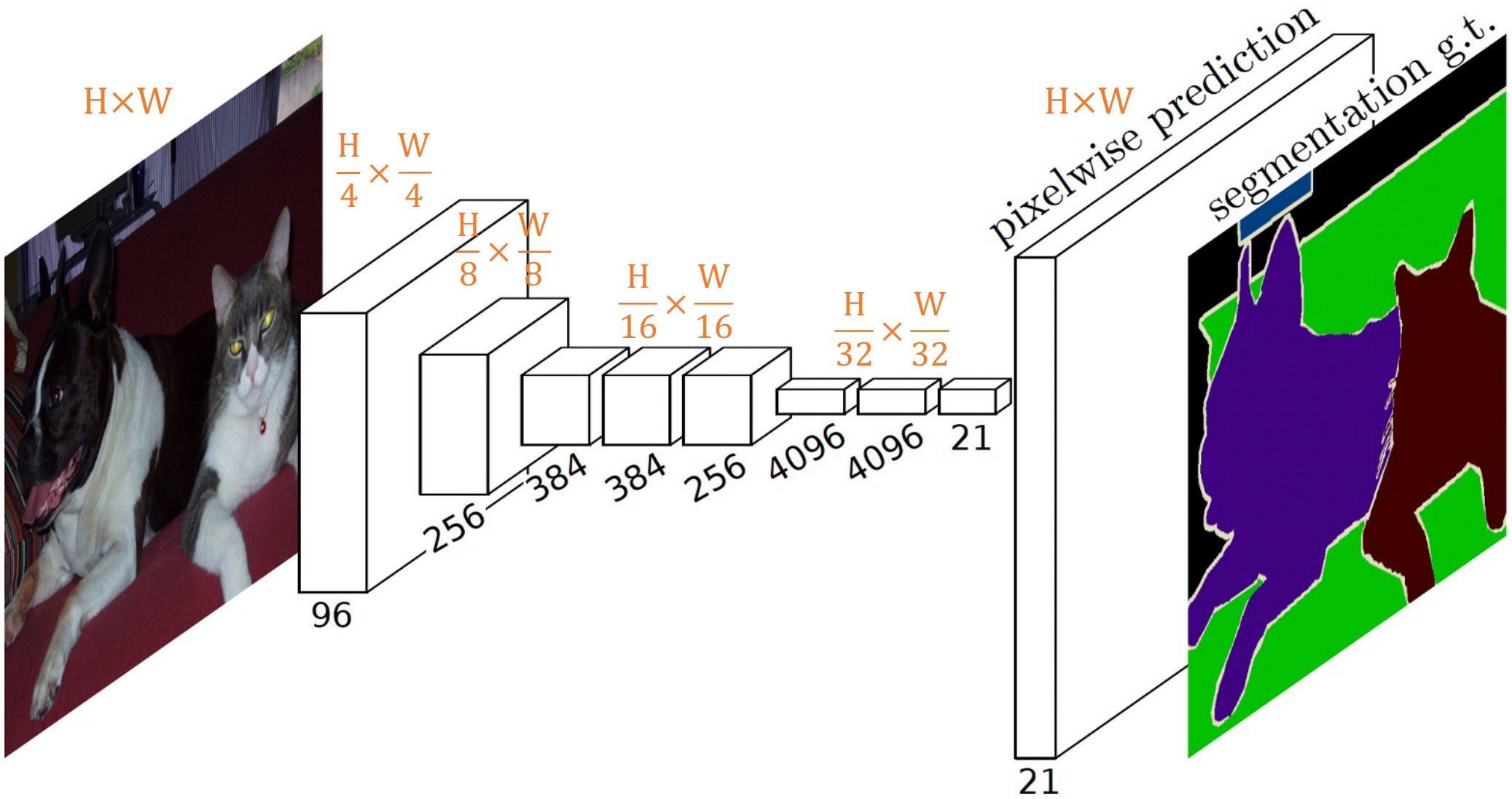
- arbitrary input size
- proportional output size



Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, “Learning a Deep Convolutional Network for Image Super-Resolution”, ECCV 2014

ConvNets Are Fully Convolutional

Example: **Fully Convolutional Network (FCN)**

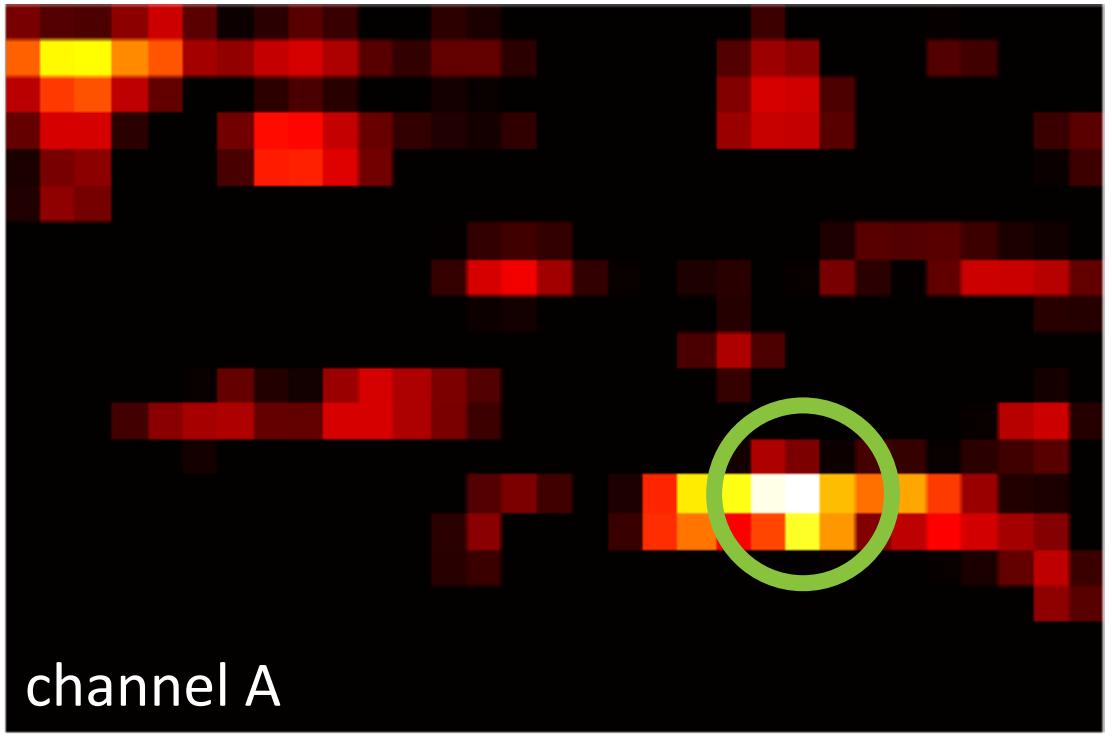


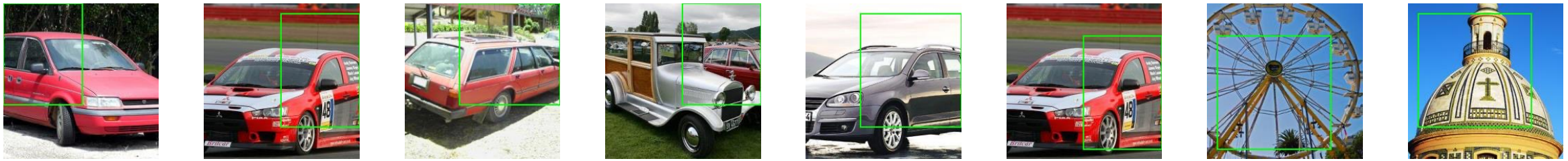
Jonathan Long, Evan Shelhamer, Trevor Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

ConvNets Are Fully Convolutional

input H×W

conv5's output $\frac{H}{32} \times \frac{W}{32}$





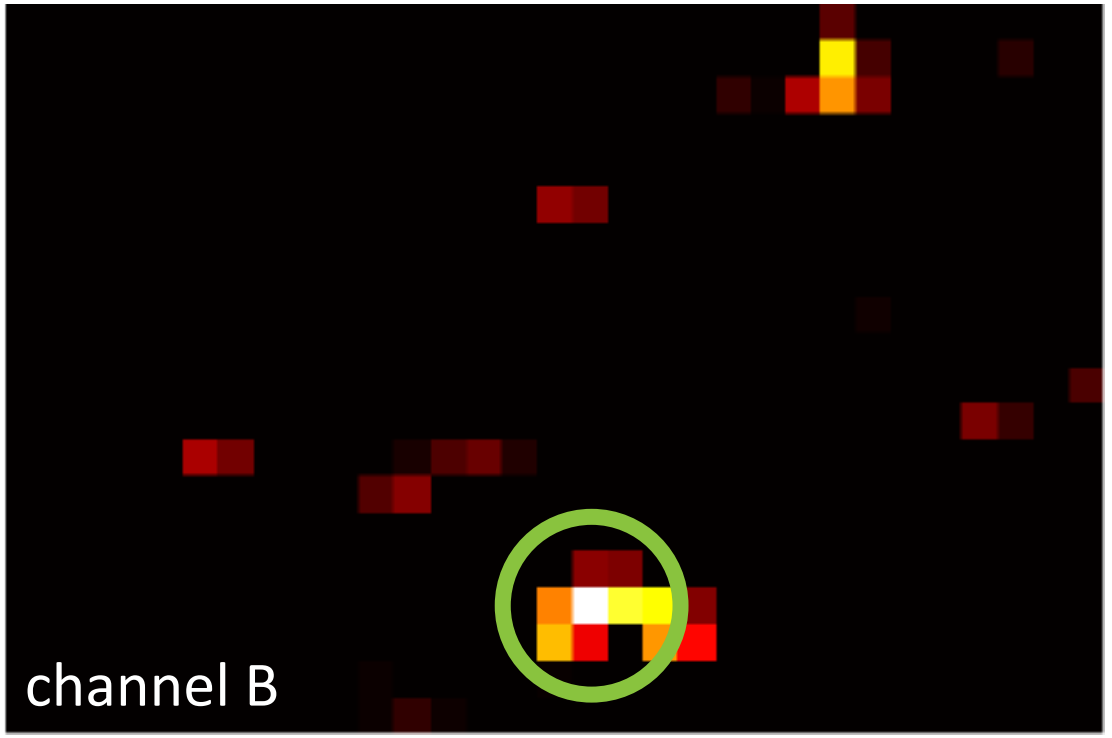
strongest activations of this channel

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", ECCV 2014

ConvNets Are Fully Convolutional

input H×W

conv5's output $\frac{H}{32} \times \frac{W}{32}$

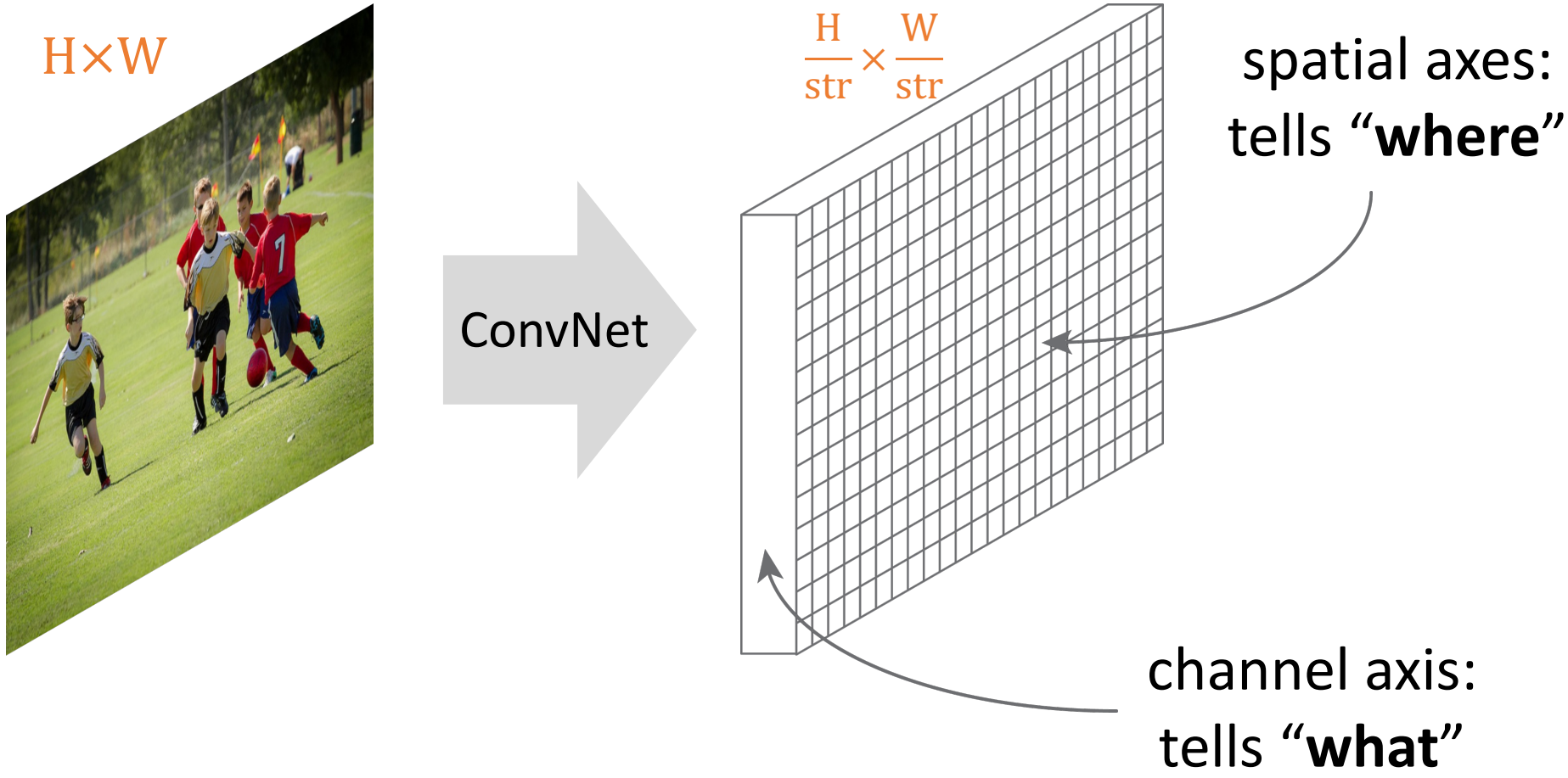


strongest activations of this channel

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", ECCV 2014

ConvNets Are Fully Convolutional

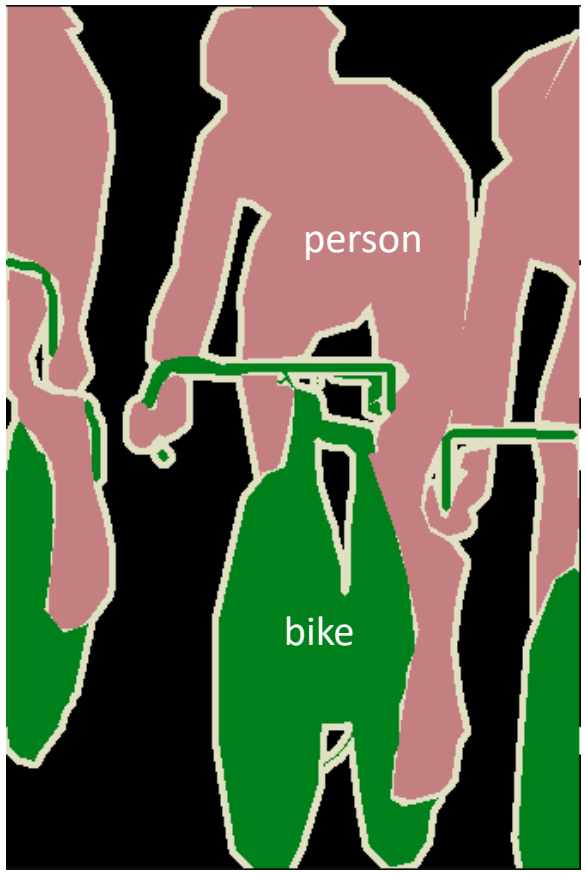
- Conceptually, detection and segmentation are solved in a “fully convolutional” fashion.



Semantic Segmentation

Semantic Segmentation: Problem Definition

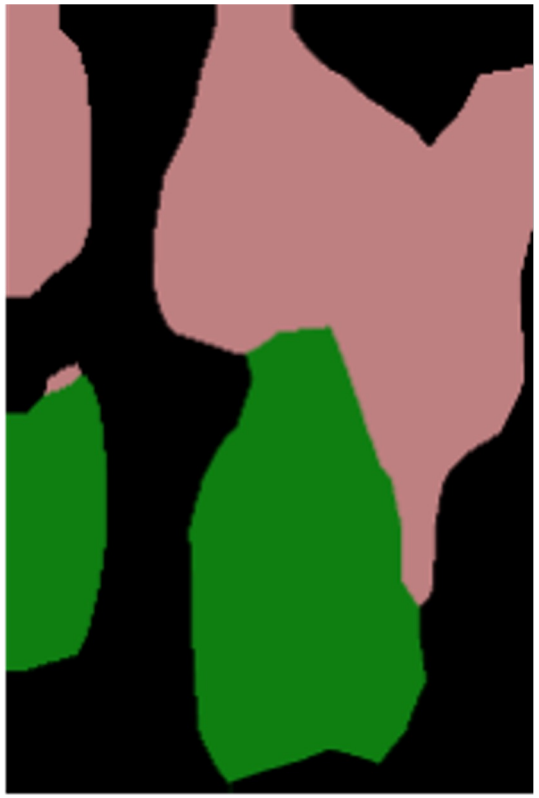
- **Task**: Label every pixel
  - Historically: don't differentiate instances



Semantic Segmentation: Evaluation

- Intersection over Union (IoU), averaged over classes

$$\text{IoU} = \frac{\text{intersection}}{\text{union}}$$

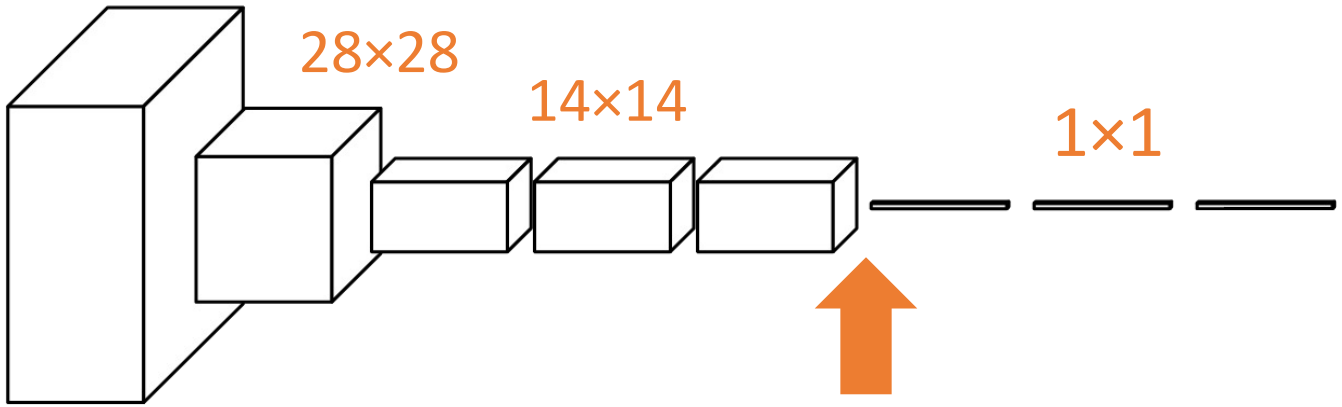


prediction

ground truth

Fully Convolutional Network (FCN)

- Repurpose a classification ConvNet for semantic segmentation



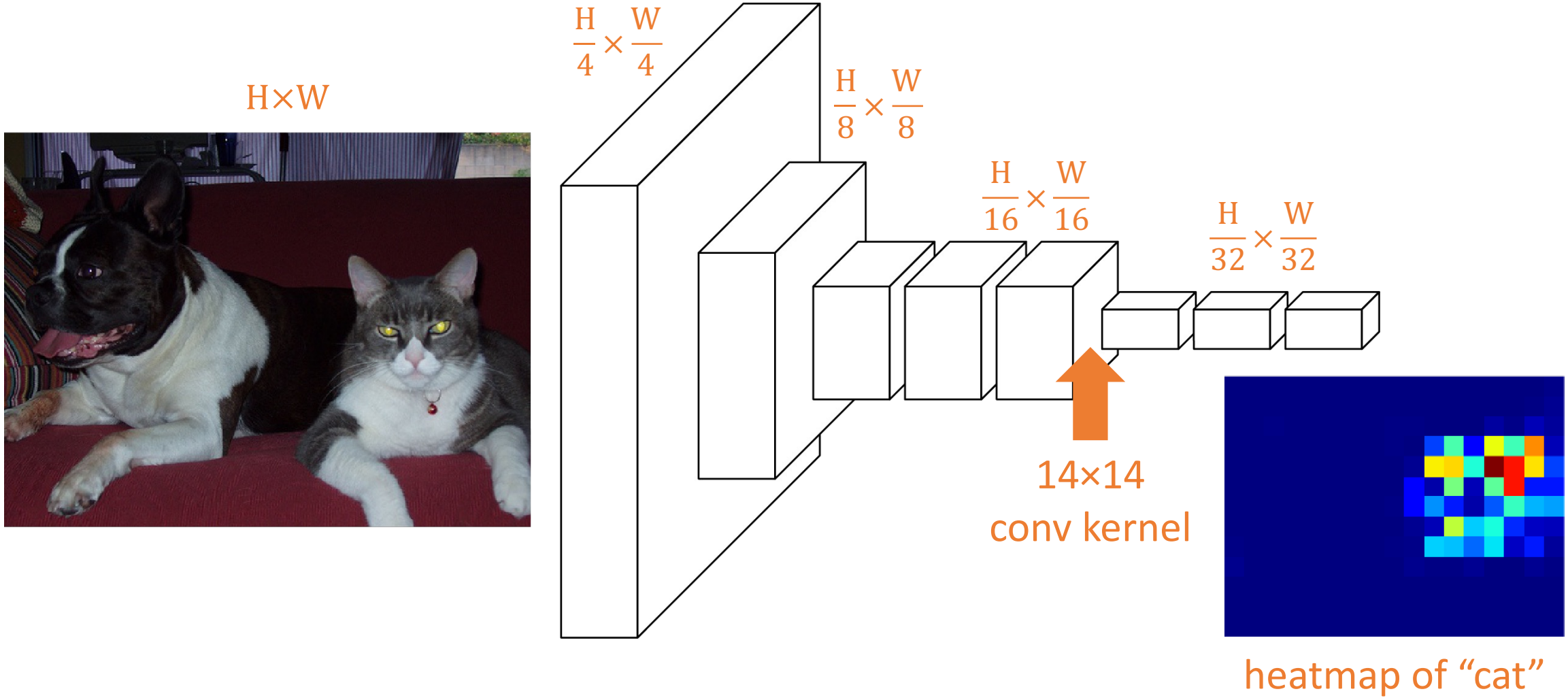
A classification ConvNet has an output spatial size of 1×1.

This fully-connected (fc) layer can be viewed as a 14×14 conv kernel.

Jonathan Long, Evan Shelhamer, Trevor Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

Fully Convolutional Network (FCN)

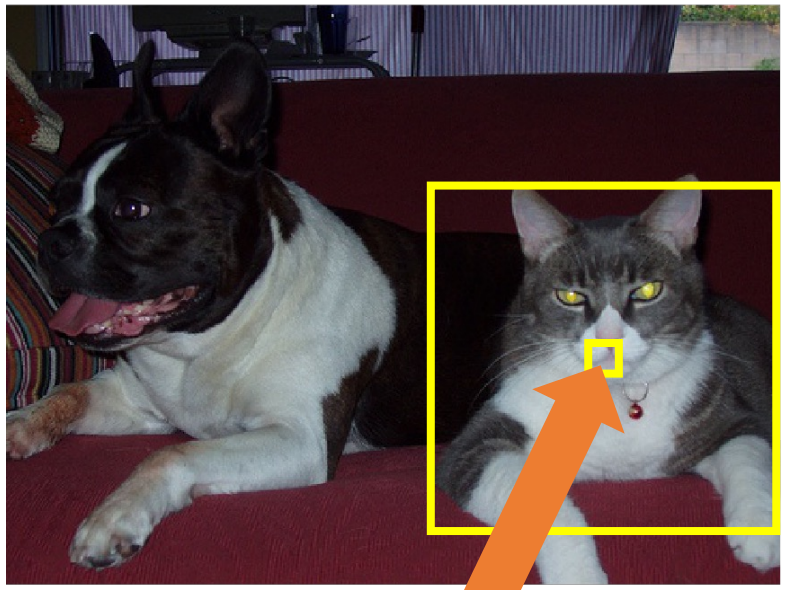
- Repurpose a classification ConvNet for semantic segmentation



Jonathan Long, Evan Shelhamer, Trevor Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

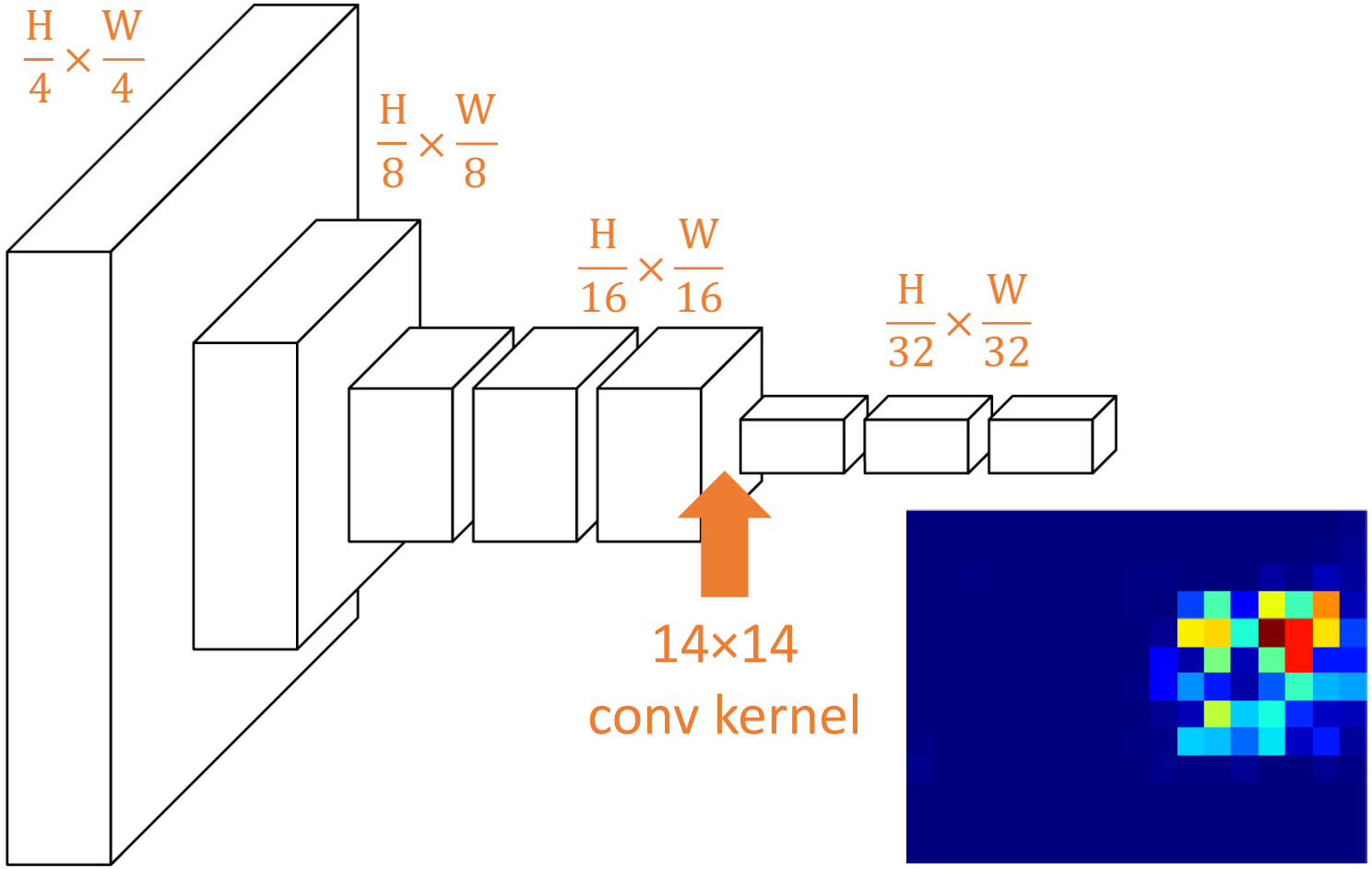
Fully Convolutional Network (FCN)

- Repurpose a classification ConvNet for semantic segmentation



H×W

prediction at this pixel is like running classification on this window

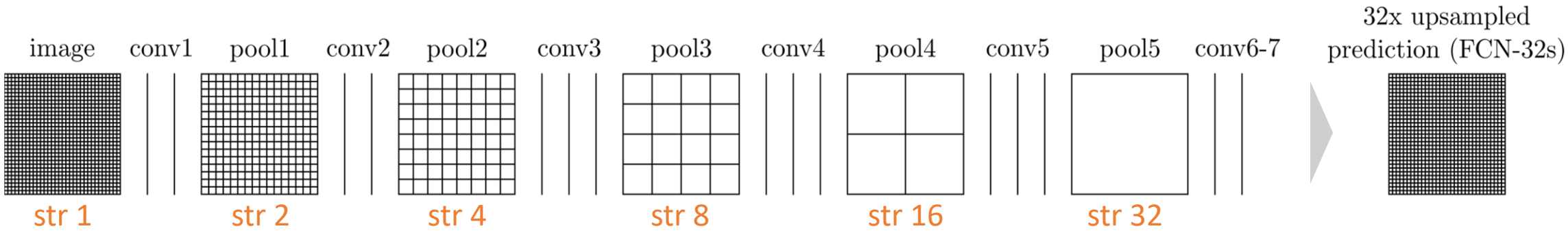


heatmap of “cat”

Jonathan Long, Evan Shelhamer, Trevor Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

Fully Convolutional Network (FCN)

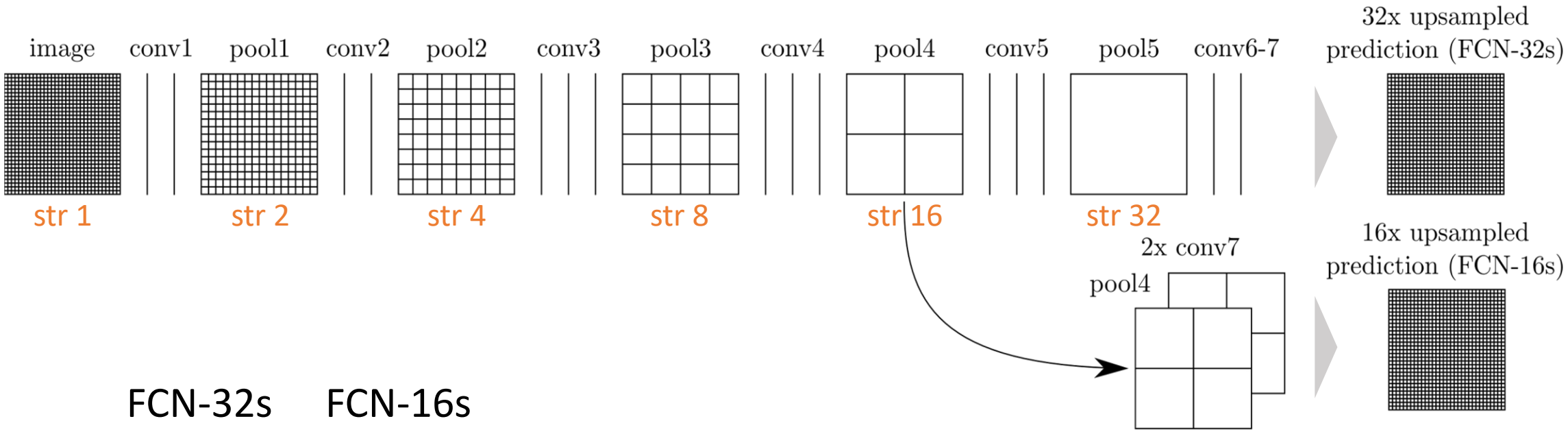
- Upsampling: combining “where” and “what”



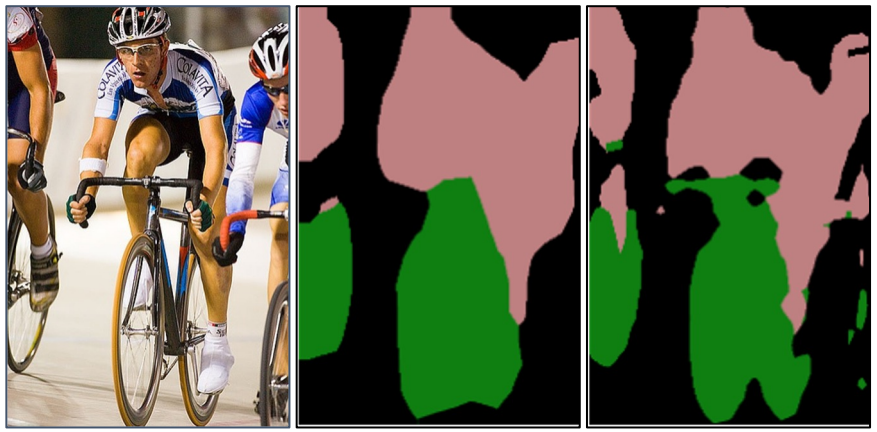
FCN-32s

Fully Convolutional Network (FCN)

- Upsampling: combining “where” and “what”

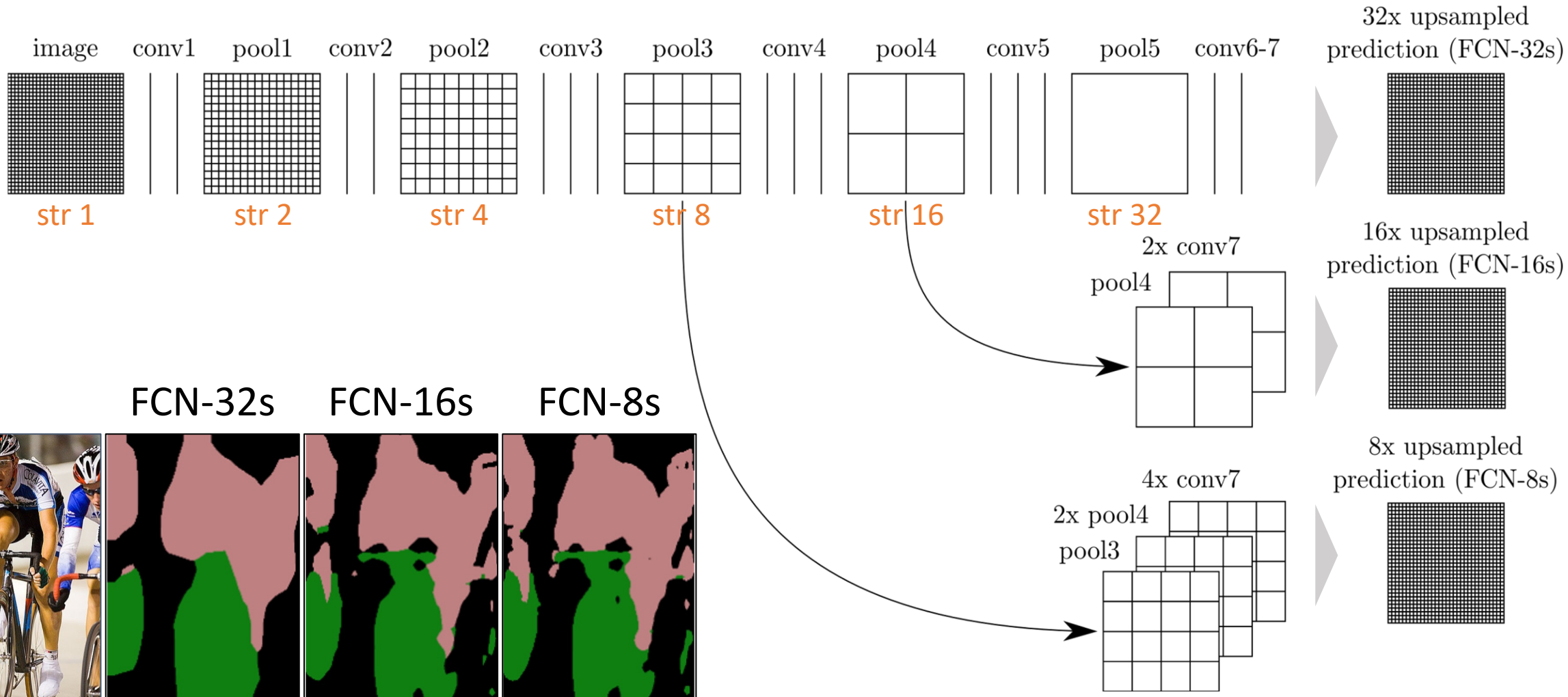


FCN-32s FCN-16s



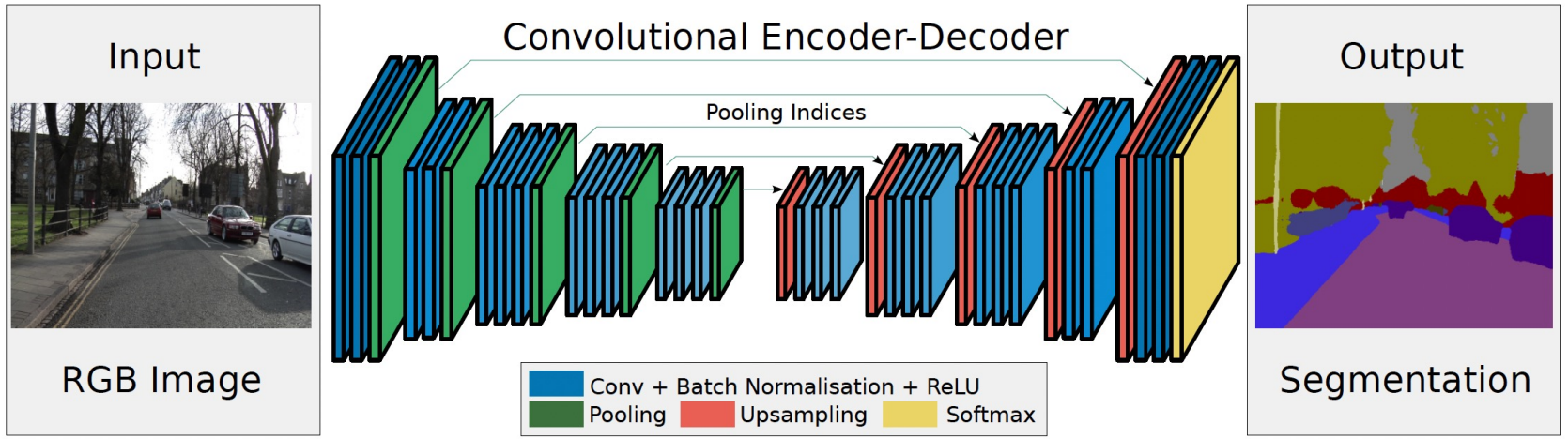
Fully Convolutional Network (FCN)

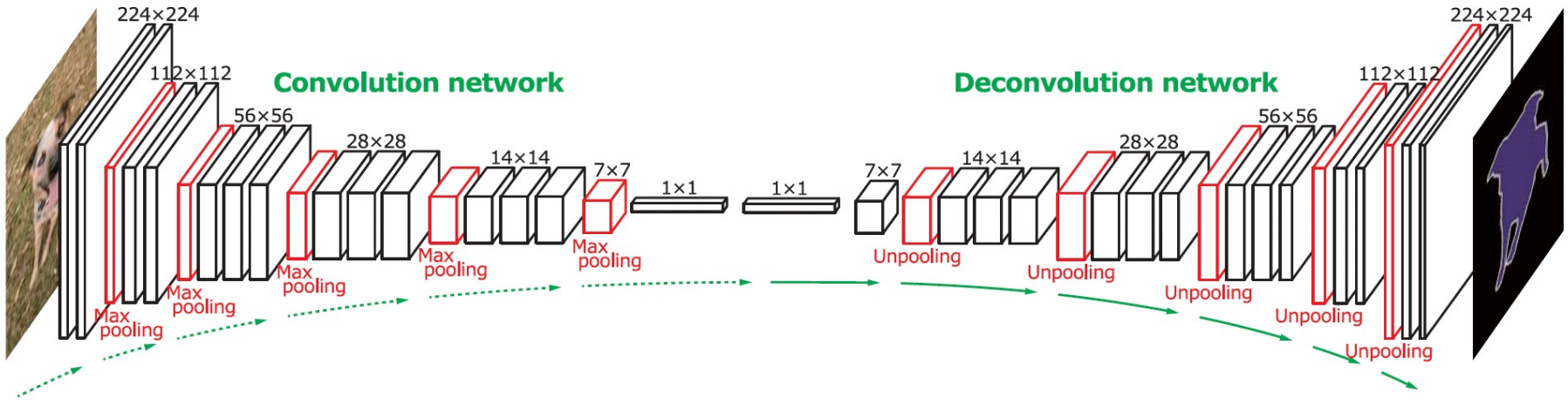
- Upsampling: combining “where” and “what”



Jonathan Long, Evan Shelhamer, Trevor Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

More upsampling strategies

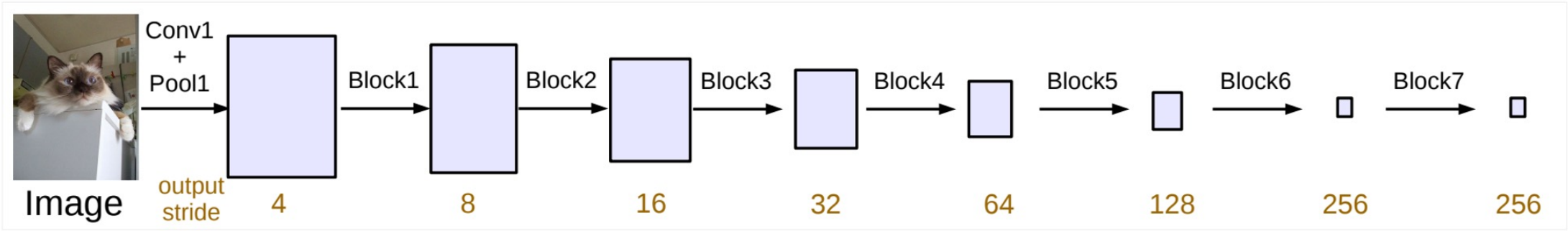


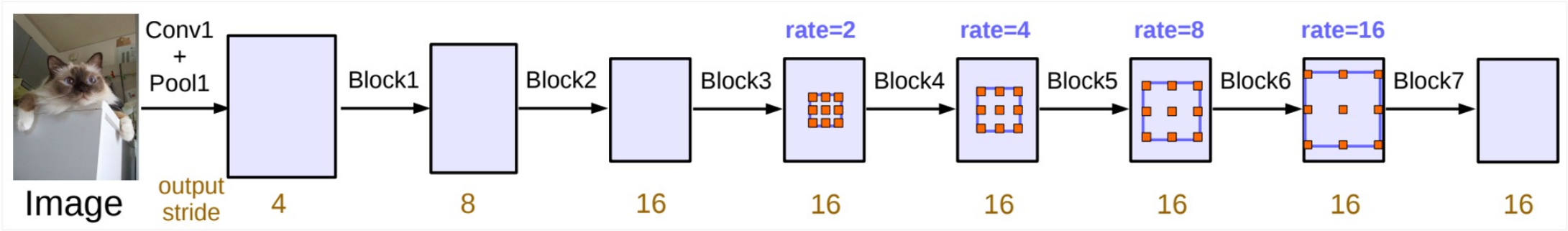


Hyeonwoo Noh, Seunghoon Hong, Bohyung Han, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", arXiv 2015, PAMI 2017

More upsampling strategies





à trous / dilated conv

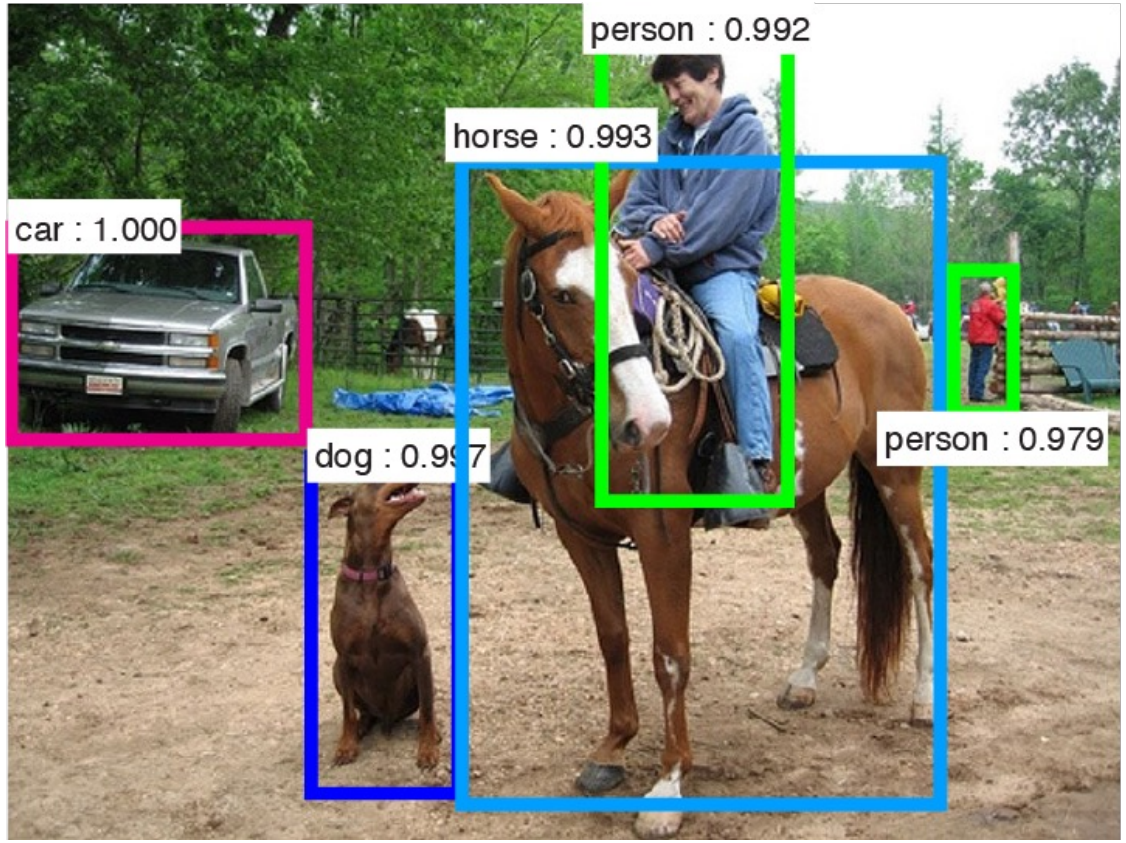
Don’t stride on the feature map; stride on the kernel

Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation”, arXiv 2017

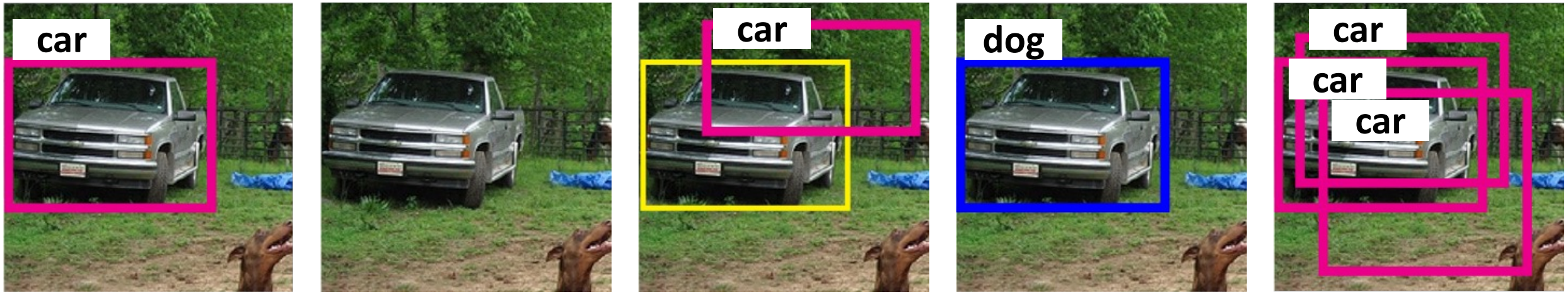
Object Detection

Object Detection: Problem Definition

- **Task**: identify, locate, and recognize objects
- **Identify**:
  - Are there objects presented?
- **Locate**:
  - Where is each object?
  - box: (x, y, h, w)
- **Recognize**:
  - What is each object?
  - class label & confidence

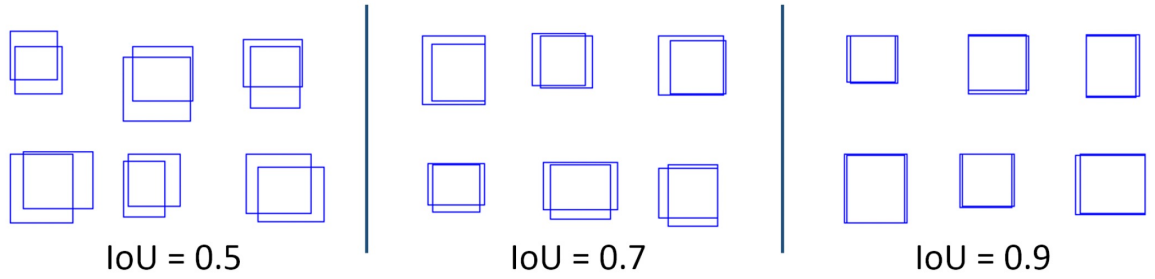


Object Detection: Evaluation



correct (true positive)	missed (false negative)	mislocalized (false positive)	misrecognized (false positive)	duplicated (false positive)

define by an IoU threshold



Object Detection: Evaluation

- **Precision**: What % of detected objects are correct?

$$\textbf{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

- **Recall**: What % of objects are detected?

$$\textbf{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

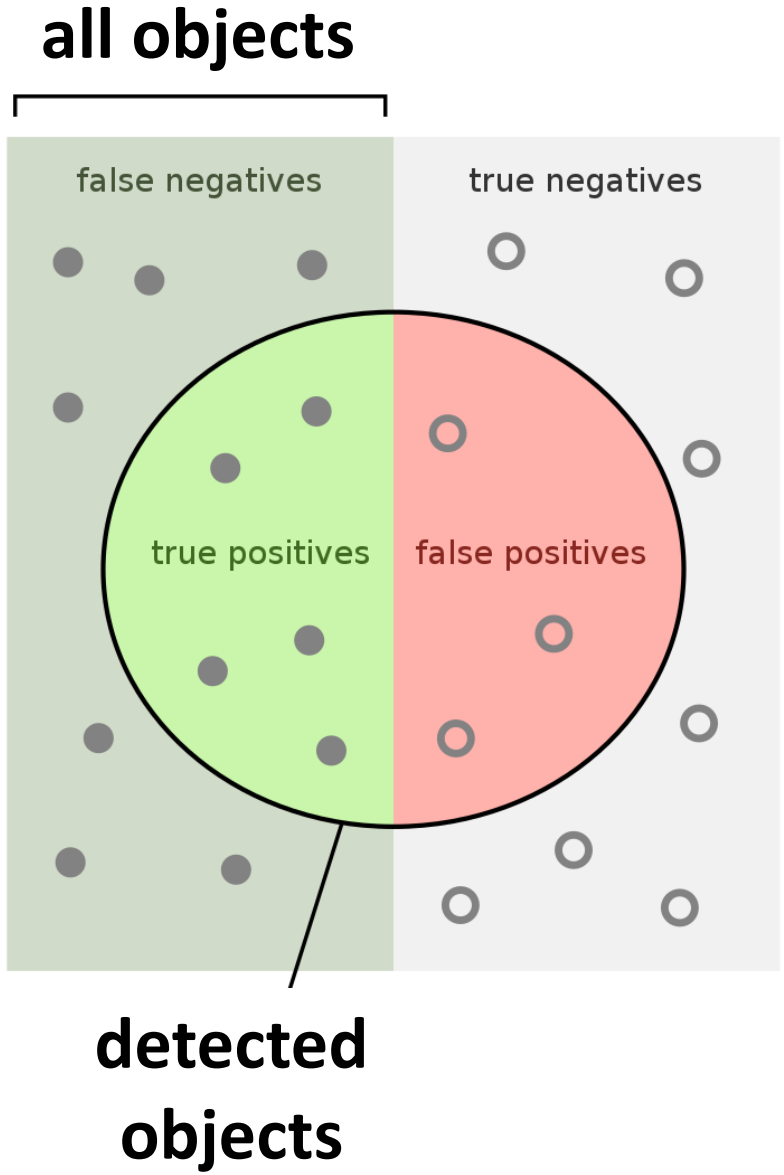
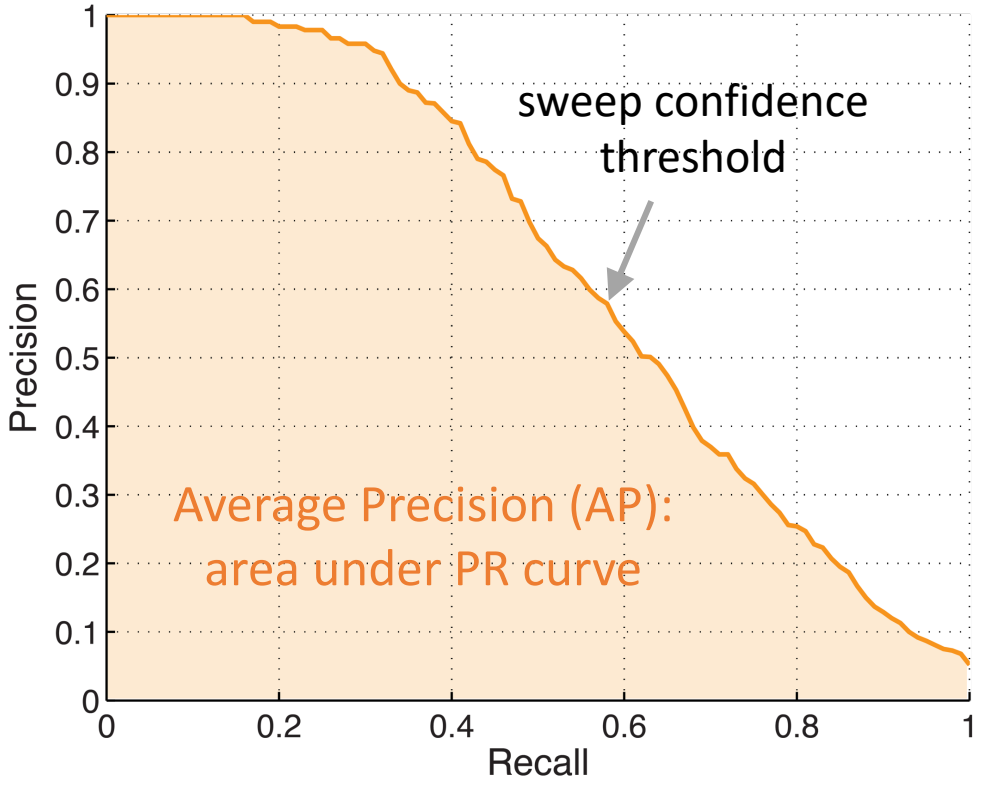


Figure adapted from: https://en.wikipedia.org/wiki/Precision_and_recall

Object Detection: Evaluation

- Precision-Recall Tradeoff

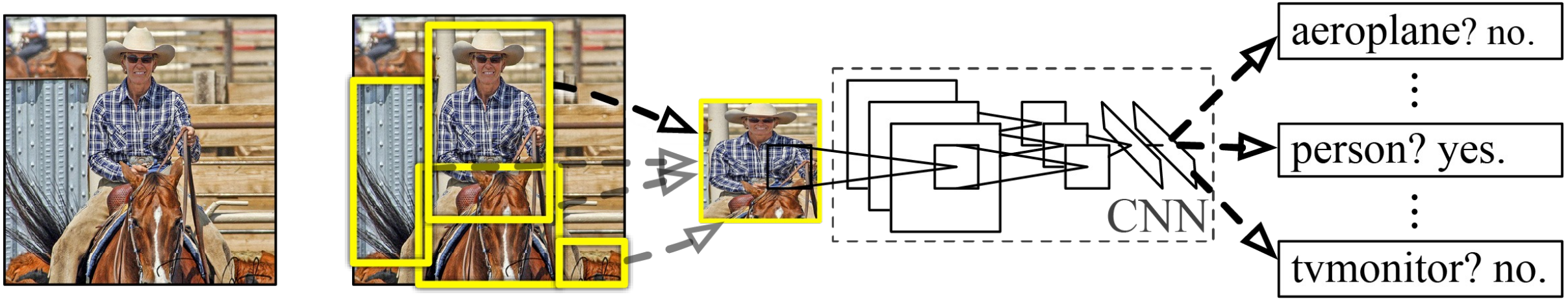


Life is a precision-recall tradeoff.

threshold = **0.4**: Recall ↑, Precision ↓

threshold = **0.7**: Recall ↓, Precision ↑

Region-Based CNN (R-CNN)

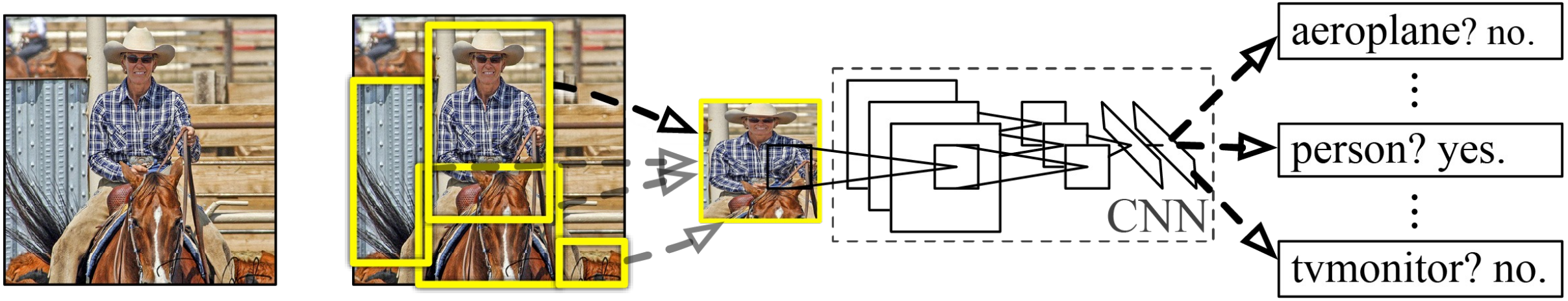


- **The deep learning revolution in the vision community.**

the ILSVRC workshop at ECCV 2012.[1] The central issue of debate can be distilled as: **To what extent do the CNN *classification* results on ImageNet generalize to object *detection* results on the PASCAL VOC Challenge?** In this paper, we study object detection using features computed by a large convolutional neural network, answering this important scientific question.[2]

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", CVPR 2014

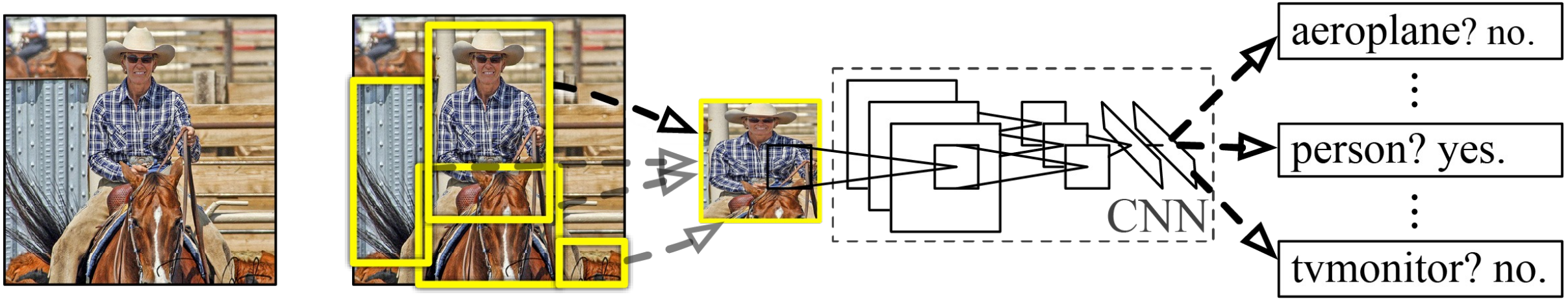
Region-Based CNN (R-CNN)



- **The deep learning revolution in the vision community.**
- **R-CNN**: Regions with CNN features
- **40%** relative improvement over classical methods
- Deep Learning made work beyond image classification

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)



input

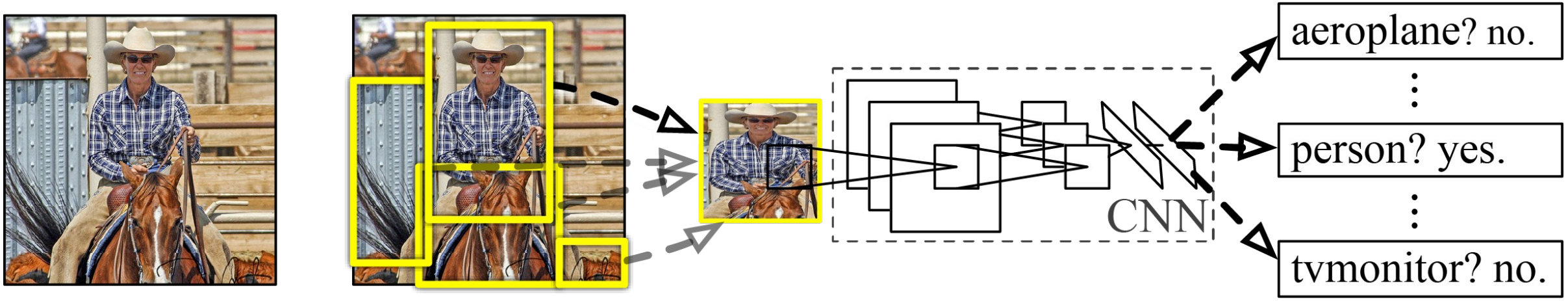
extract
“region proposals”

compute CNN features
for each region

classify regions
& refine boxes

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)

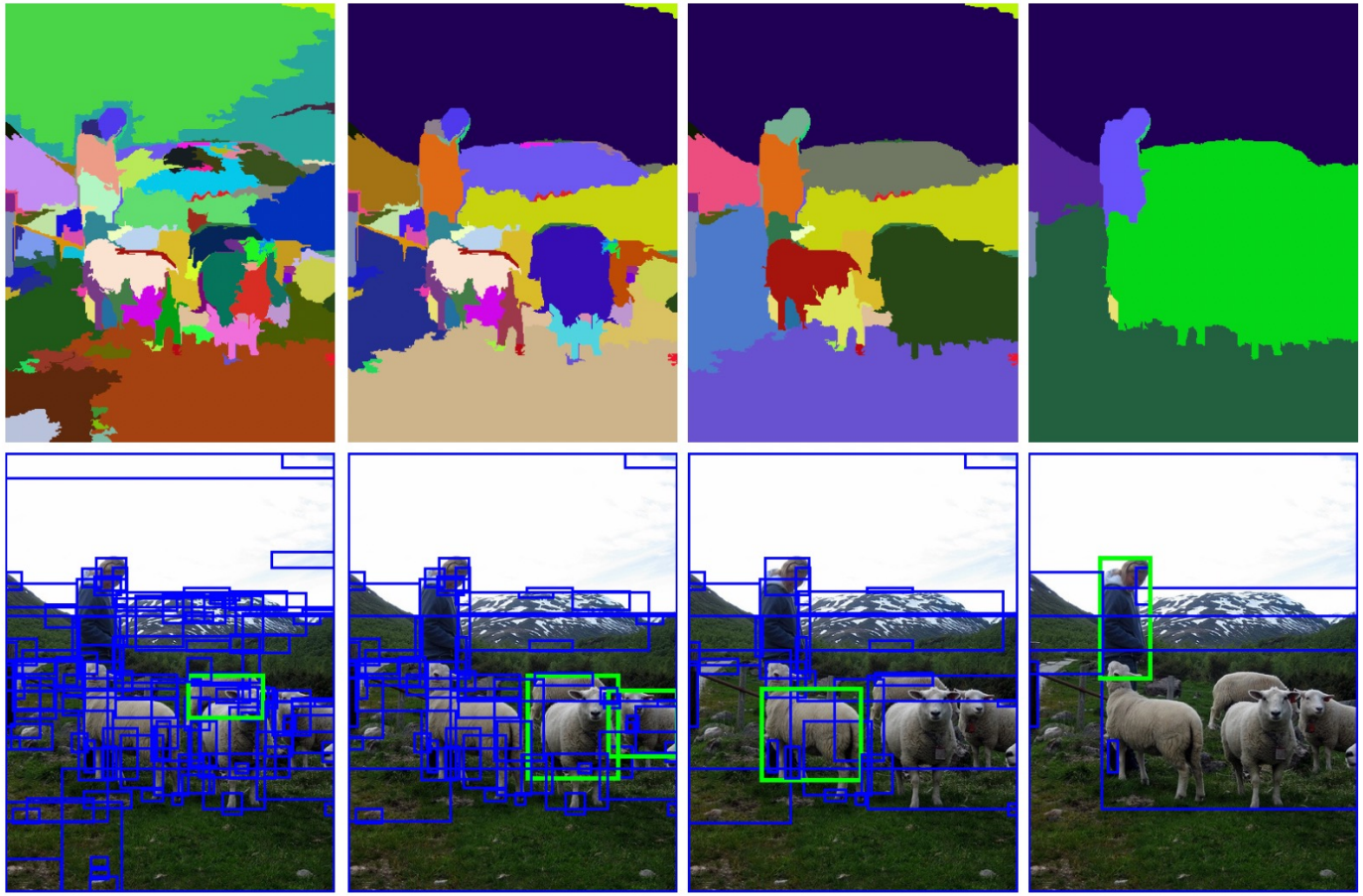


extract
"region proposals"

- **Reduce the search space ("selective search")**
- e.g., ~2000 regions / image
- Recall-driven
  - a proposal may not be an object
  - an object is highly likely in the proposals

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", CVPR 2014

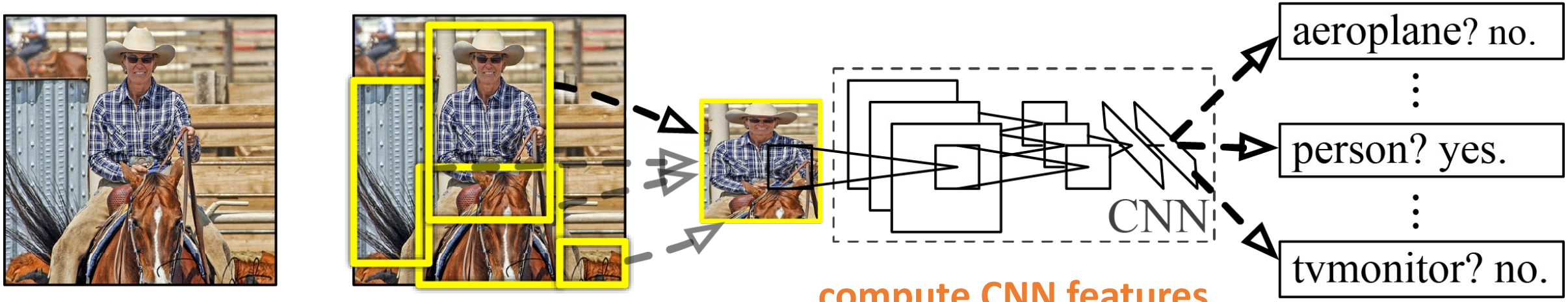
(e.g.) Selective Search for Region Proposal



- Multi-scale over-segmentation
- Hierarchically group similar pixels
- “Hand-engineered” features
- Bounding boxes of segments

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. “Selective search for object recognition”. IJCV 2013

Region-Based CNN (R-CNN)

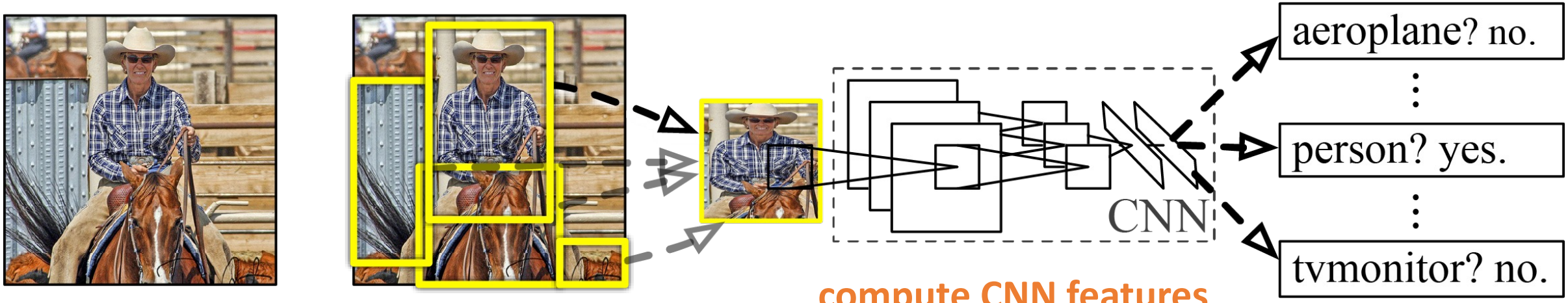


compute CNN features
for each region

one proposal

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)



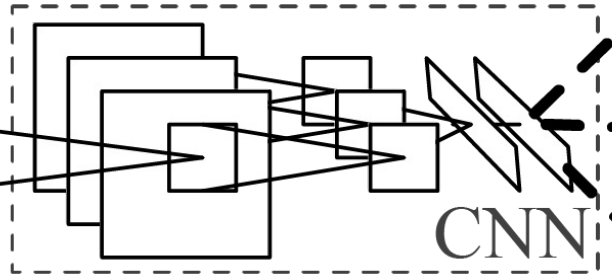
compute CNN features
for each region

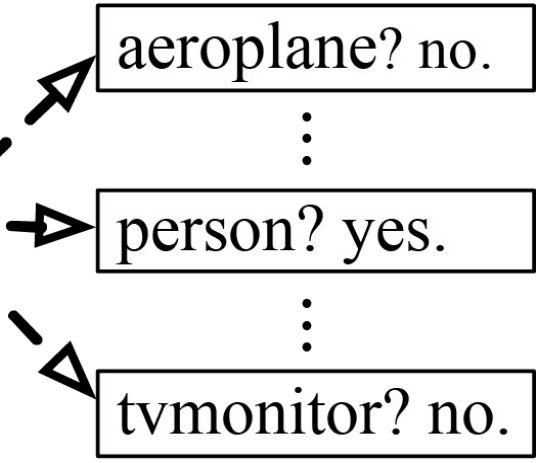
one proposal

crop

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)





compute CNN features for each region

one proposal

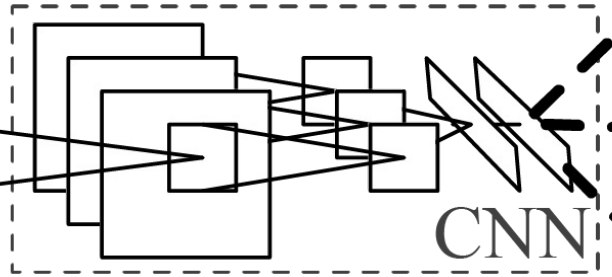
crop

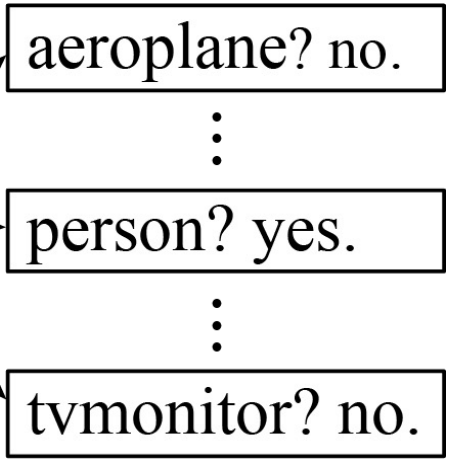
warp

(e.g., 224×224)

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)



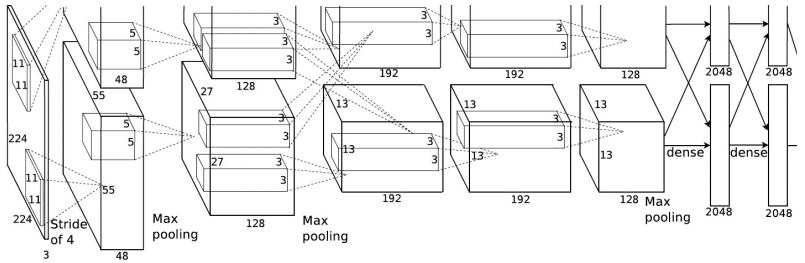


compute CNN features
for each region

one proposal

crop

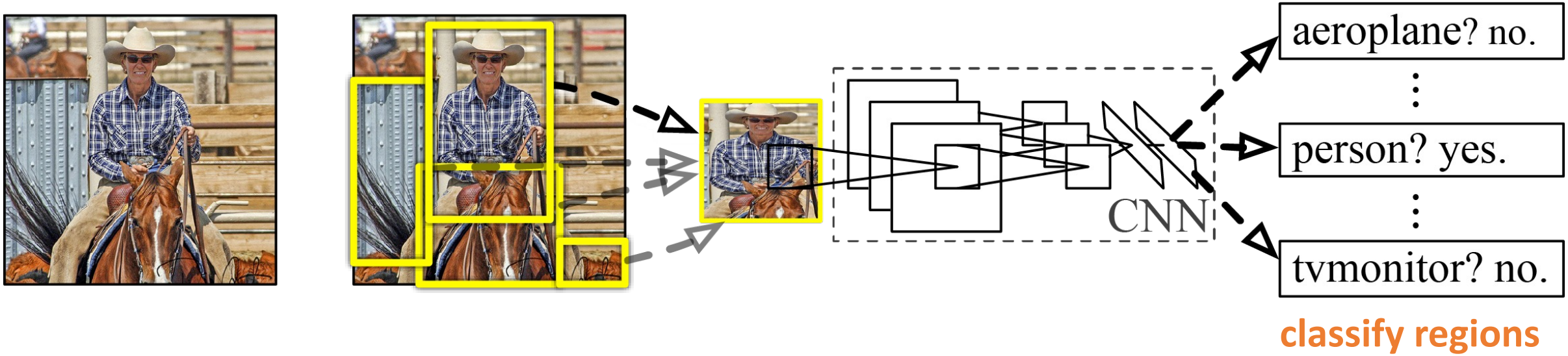
warp
(e.g., 224×224)



apply a classification ConvNet
(e.g., AlexNet)

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

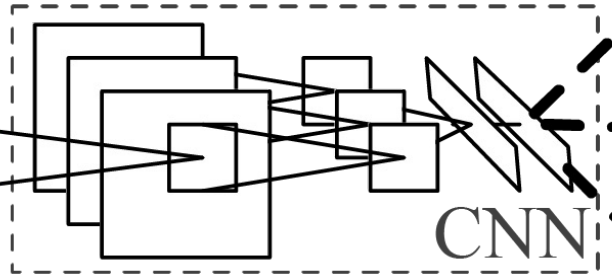
Region-Based CNN (R-CNN)



classify regions

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)

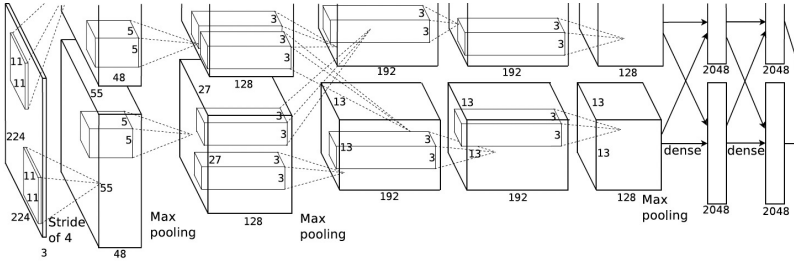


aeroplane? no.

person? yes.

tvmonitor? no.

classify regions



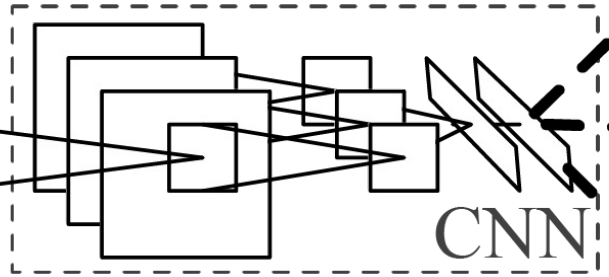
background: 0.0

person: 0.9

horse: 0.1

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)

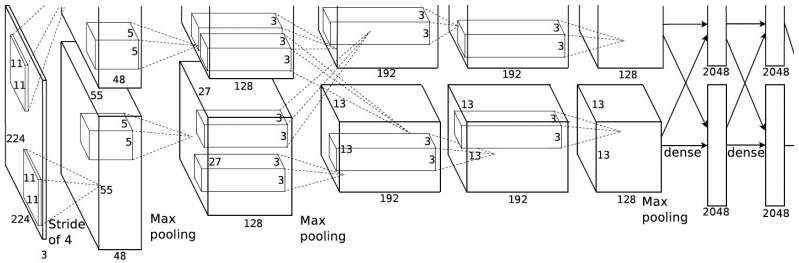


aeroplane? no.

person? yes.

tvmonitor? no.

classify regions



Detection is first of all
a foreground/background
classification problem.

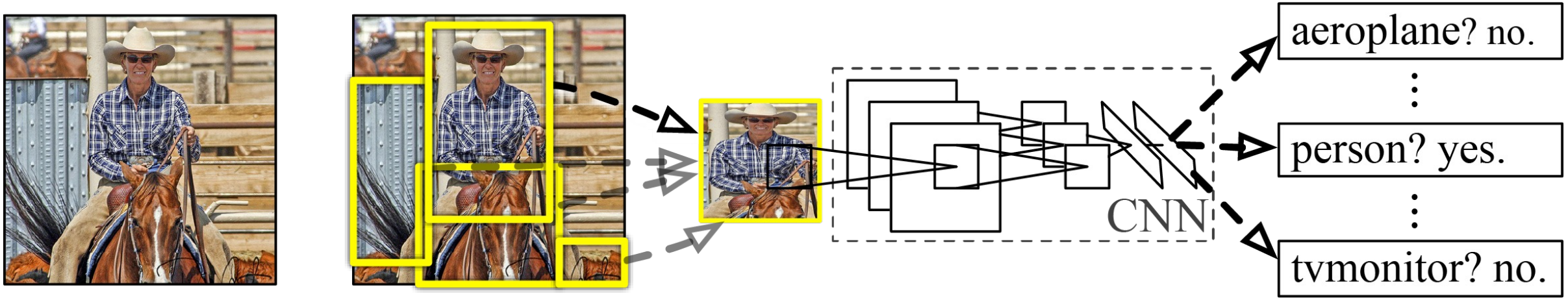
background: 0.9

person: 0.0

horse: 0.1

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Region-Based CNN (R-CNN)

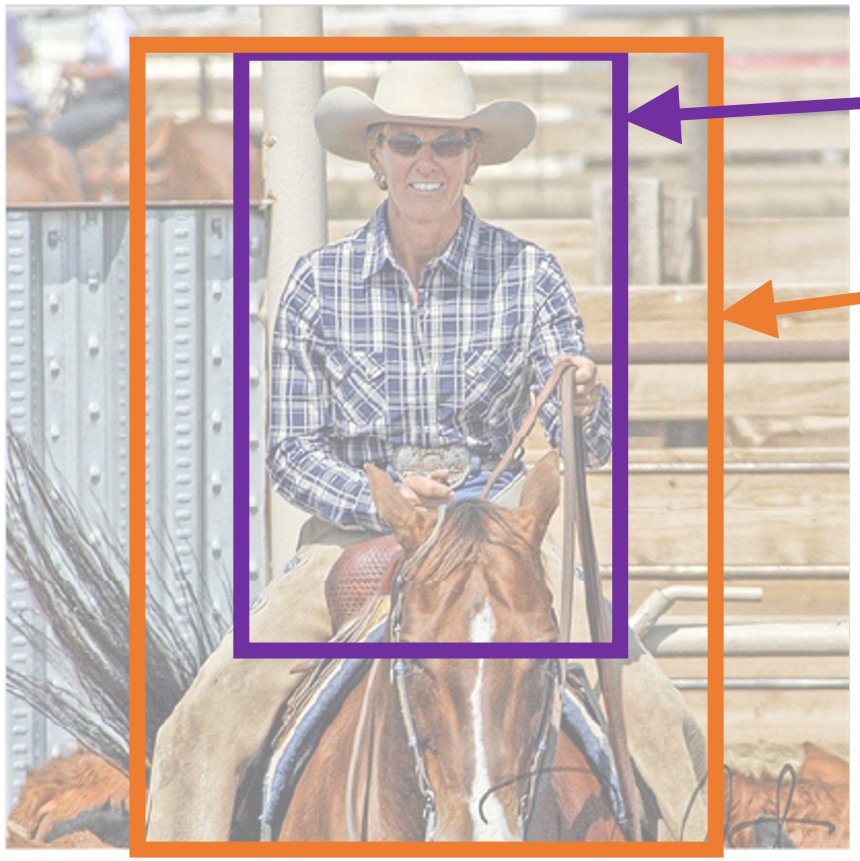


classify regions
& **refine boxes**

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

Bounding-box Regression

- **Geometric prediction** from deep neural networks



reference box: (x, y, w, h)

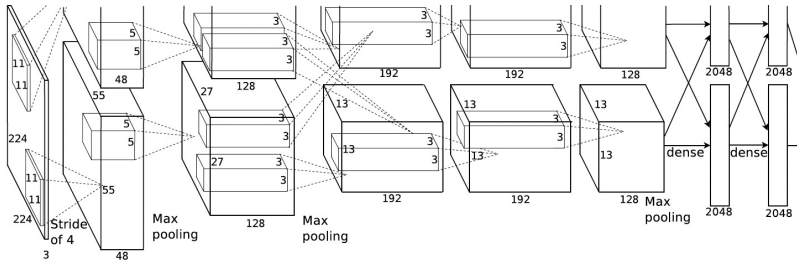
target box: (x', y', w', h')

$$\Delta x = (x' - x)/w$$
$$\Delta y = (y' - y)/h$$
$$\Delta \log w = \log w' - \log w$$
$$\Delta \log h = \log h' - \log h$$

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", CVPR 2014

Bounding-box Regression

- **Geometric prediction** from deep neural networks
- one network, multiple tasks



classification

bbox regression

$\Delta x, \Delta y, \Delta \log w, \Delta \log h$

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

R-CNN

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

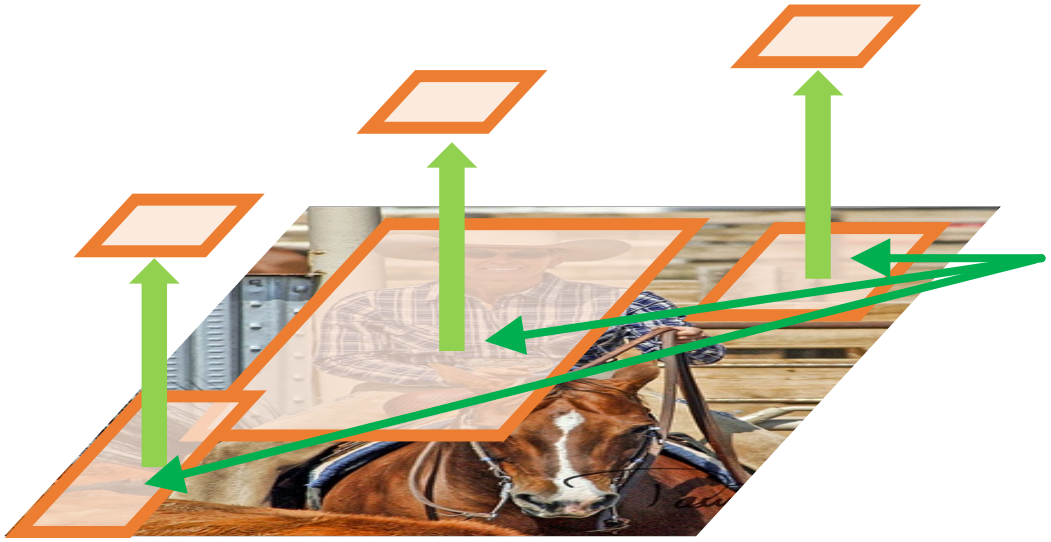
R-CNN

- Propose ~2000 Regions of Interest (RoI)

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

R-CNN

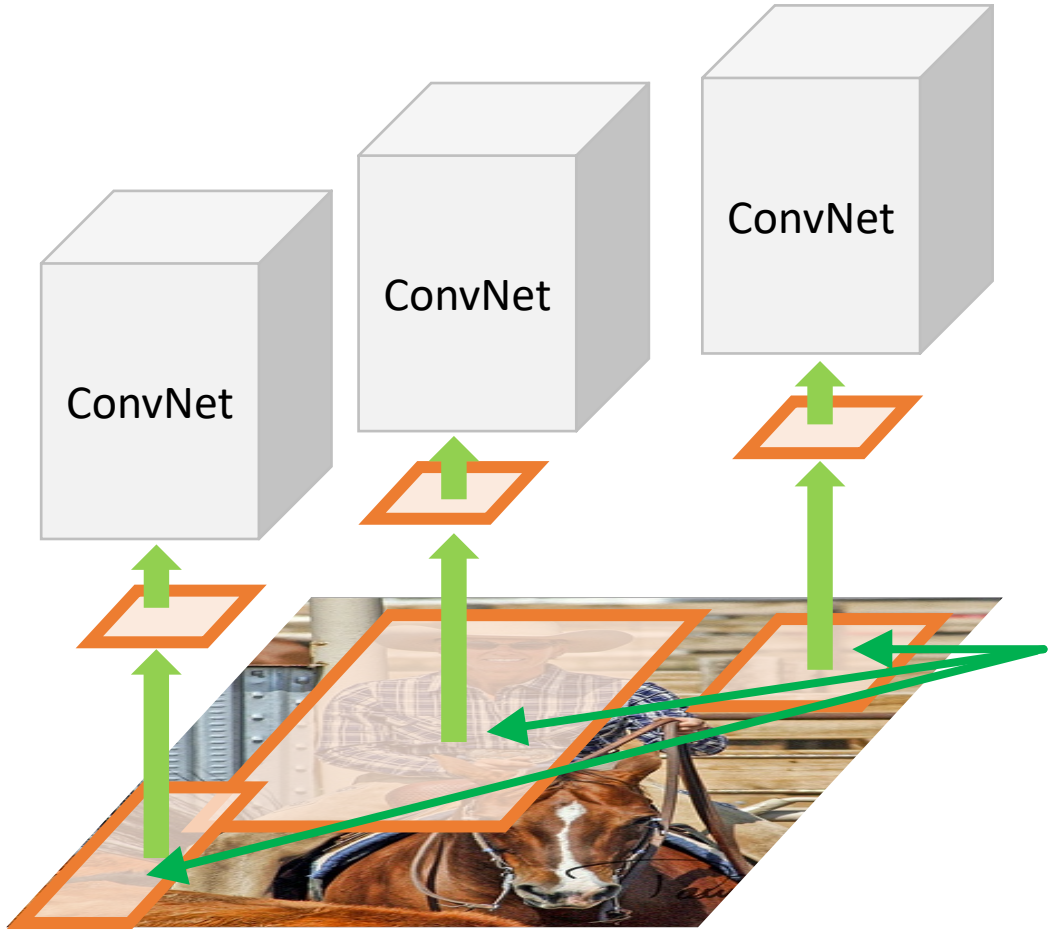


- Warp regions
- Propose ~2000 Regions of Interest (RoI)

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

R-CNN

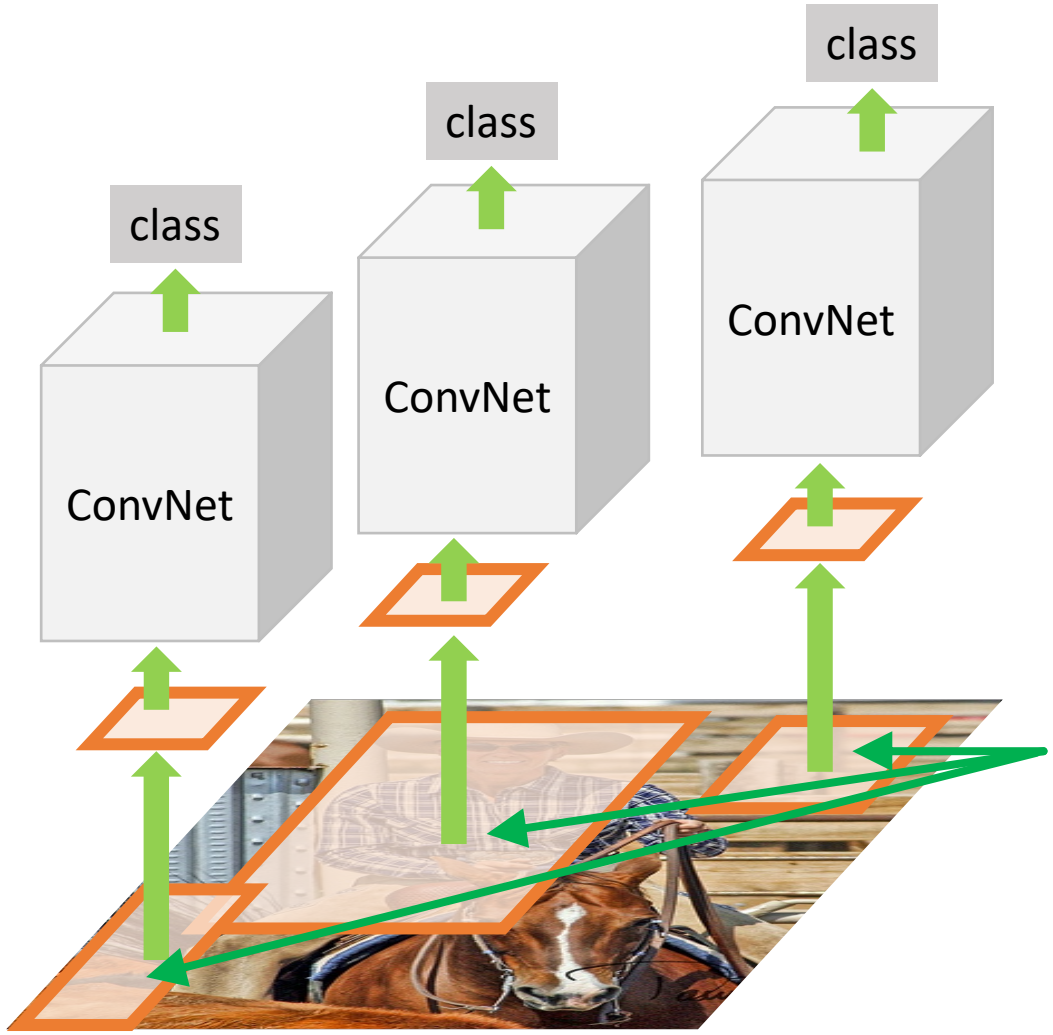


- Apply ConvNets to regions
- Warp regions
- Propose ~2000 Regions of Interest (RoI)

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

R-CNN

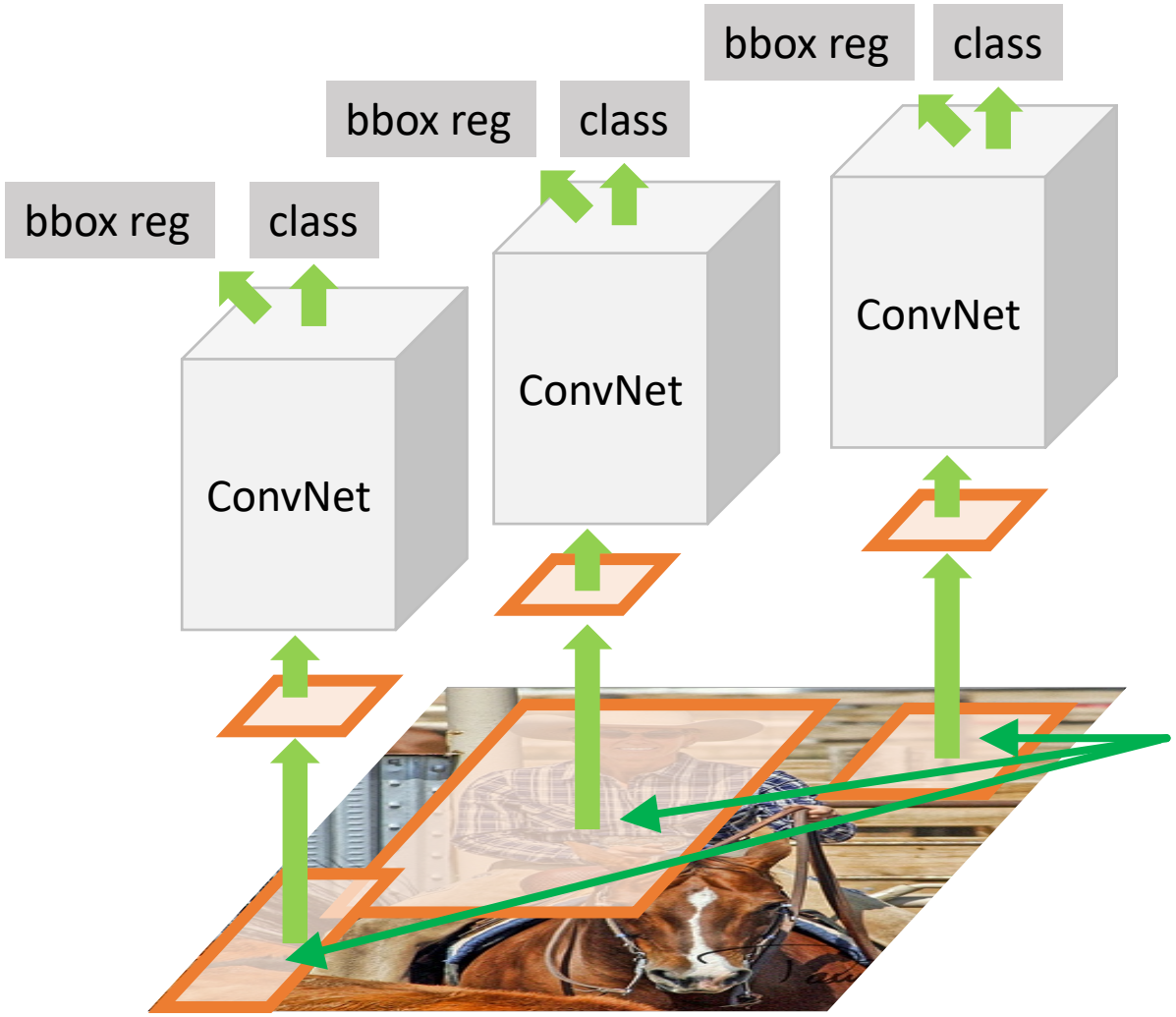


- Classify regions
- Apply ConvNets to regions
- Warp regions
- Propose ~2000 Regions of Interest (RoI)

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", CVPR 2014

R-CNN

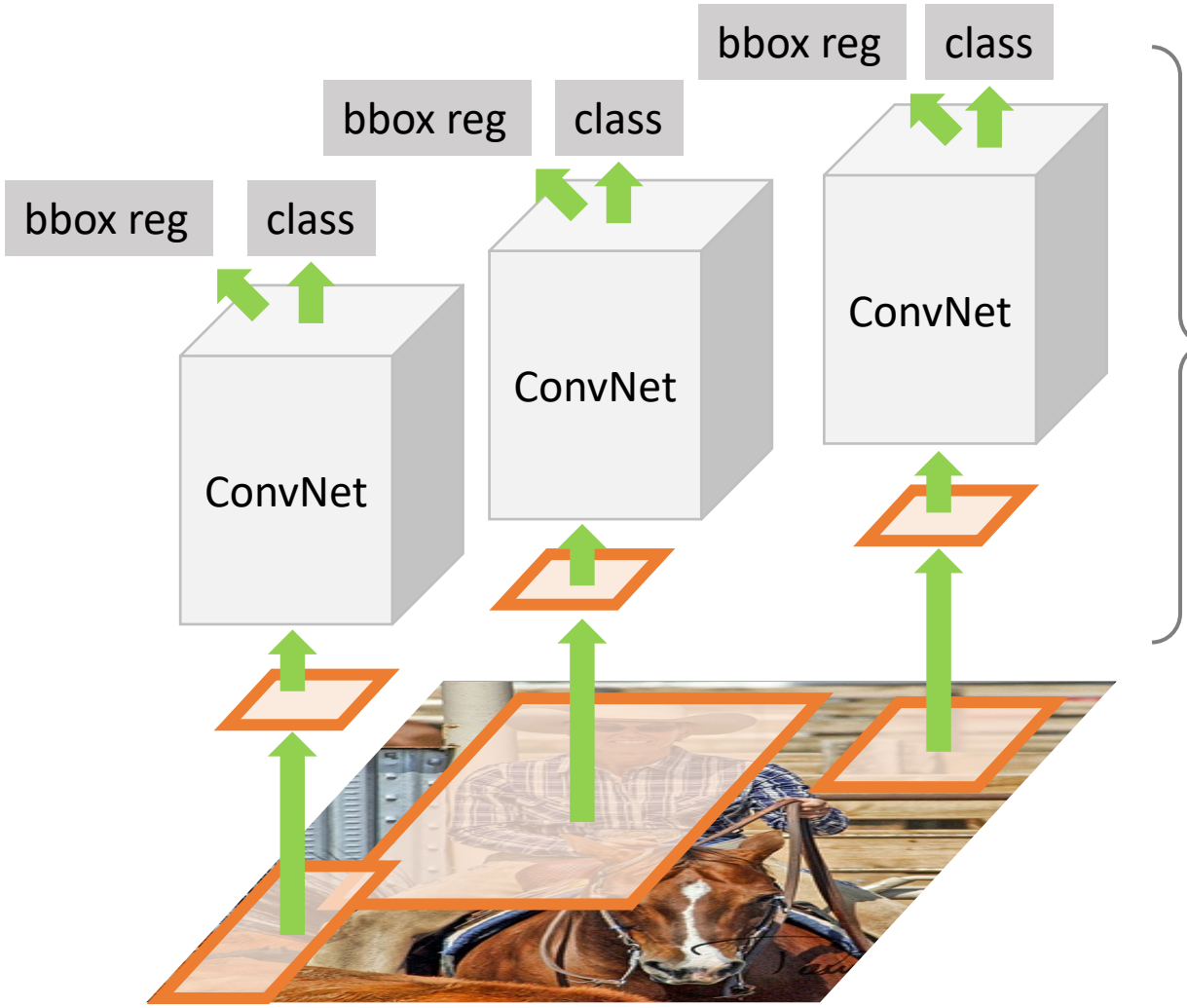


- Regress bounding-boxes
- Classify regions
- Apply ConvNets to regions
- Warp regions
- Propose ~2000 Regions of Interest (RoI)

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”, CVPR 2014

What's wrong with "Slow" R-CNN? (quote [Ross Girshick, ICCV 2015])



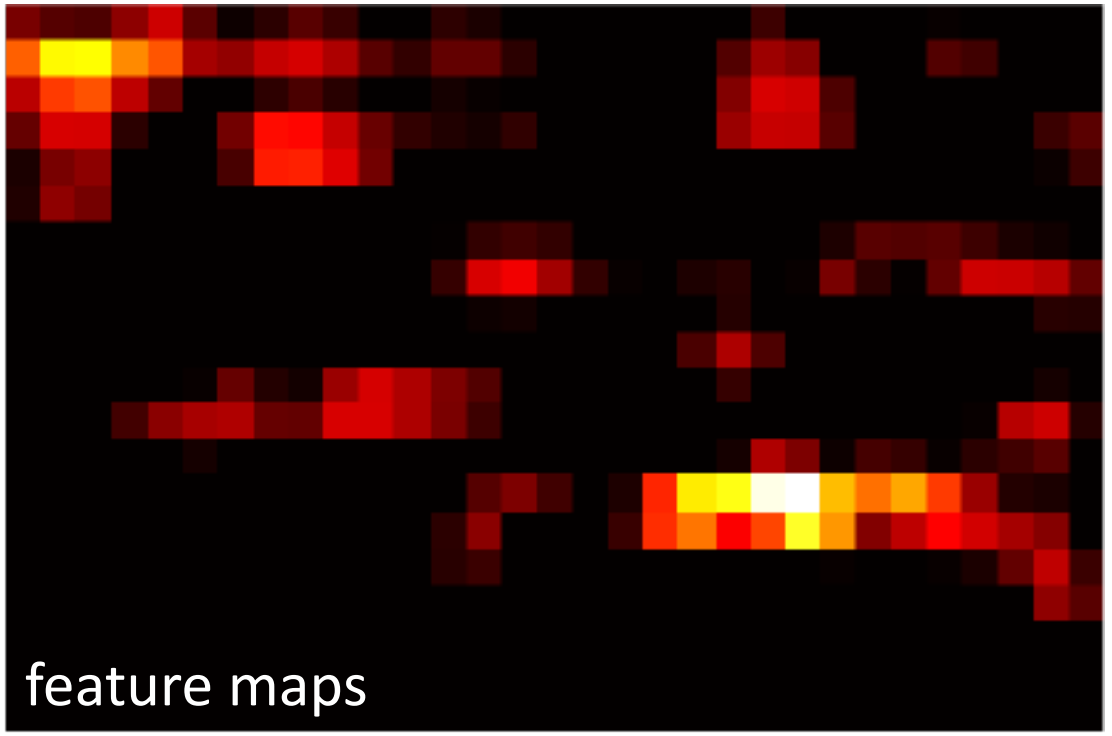
Many (~2000) ConvNets applied on regions!

Slides adapted from: Ross Girshick, ICCV 2015

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", CVPR 2014

Towards Fast R-CNN (SPPnet)

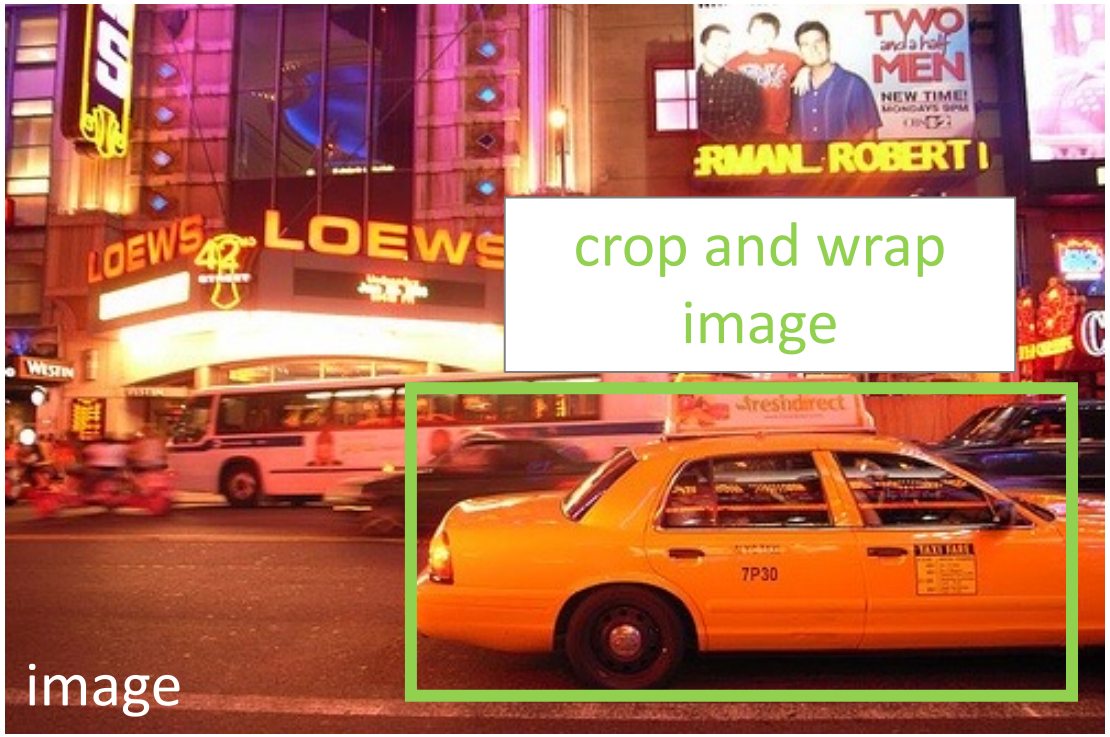
- **ConvNets are Fully Convolutional!**

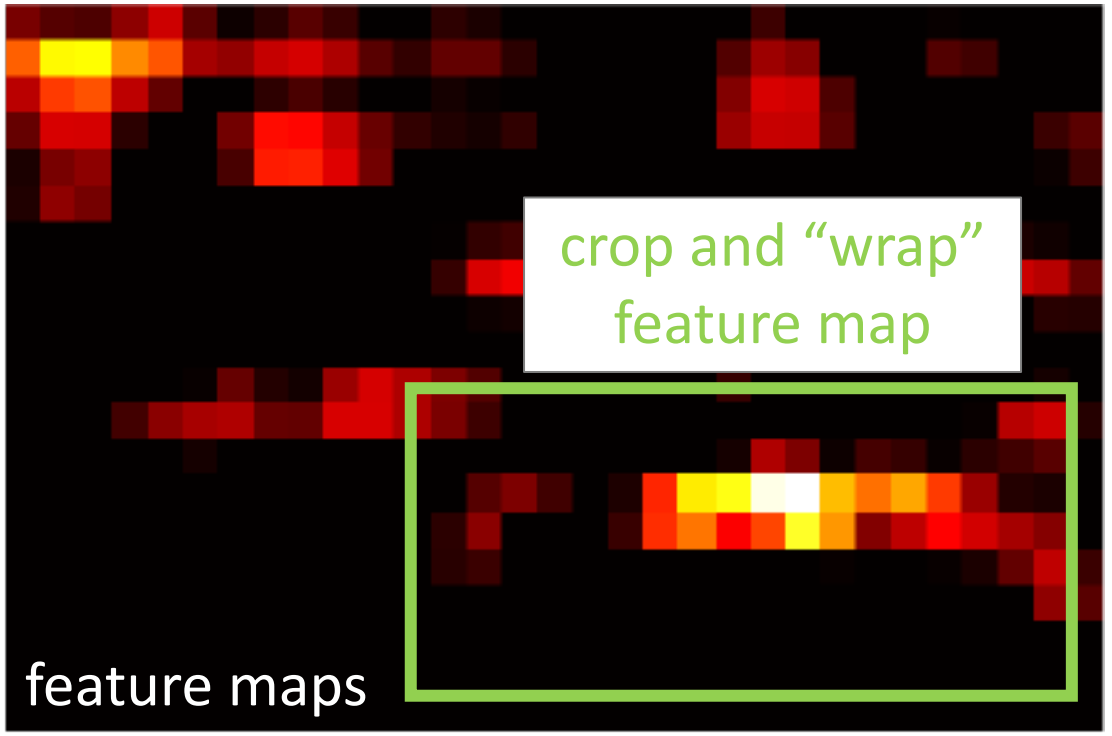


Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, ECCV 2014

Towards Fast R-CNN (SPPnet)

- **ConvNets are Fully Convolutional!**





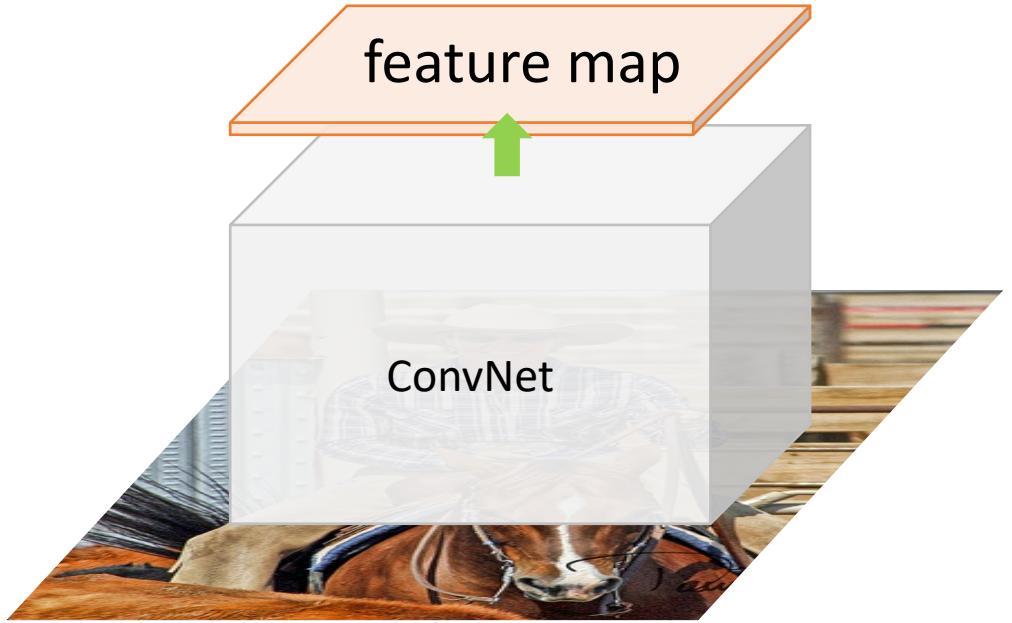
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, ECCV 2014

Towards Fast R-CNN (SPPnet)

Slides adapted from: Ross Girshick, ICCV 2015

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, ECCV 2014

Towards Fast R-CNN (SPPnet)

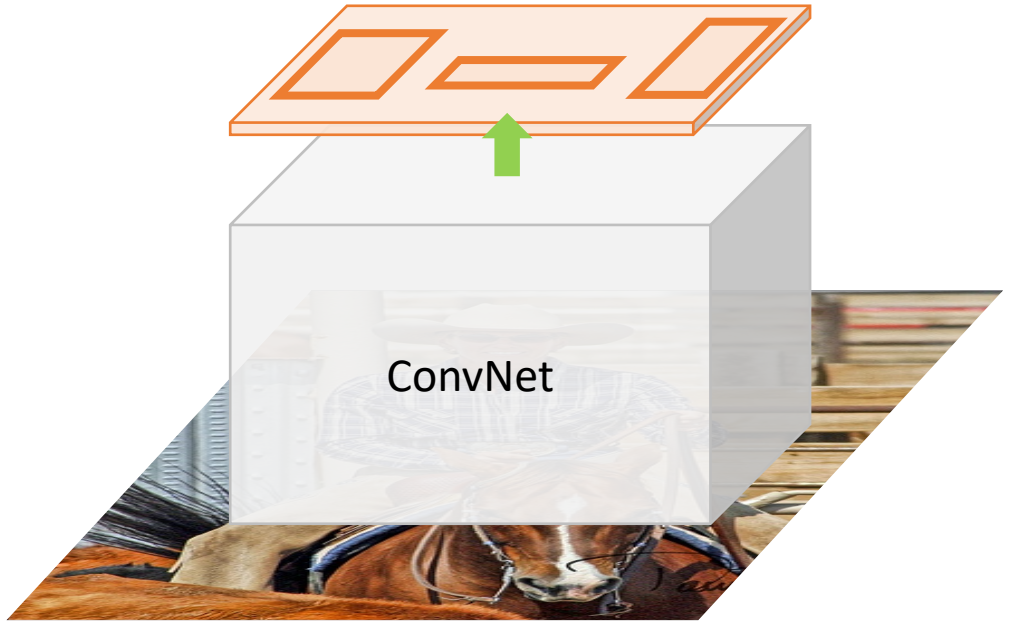


- Apply ConvNet to **whole** image

Slides adapted from: Ross Girshick, ICCV 2015

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, ECCV 2014

Towards Fast R-CNN (SPPnet)

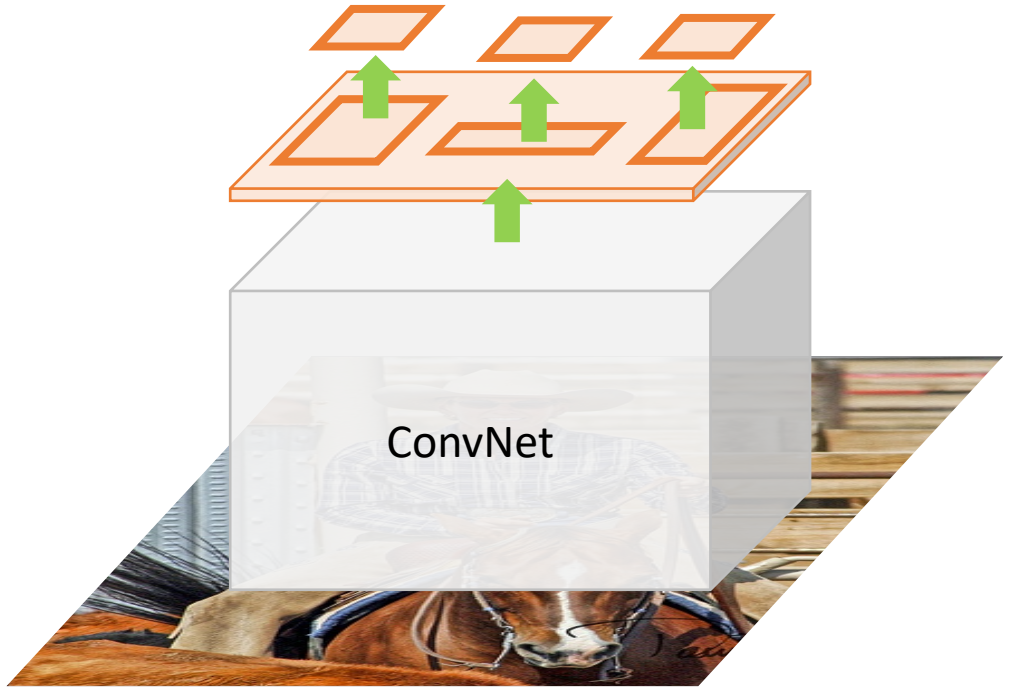


- Map proposed RoIs to feature maps
- Apply ConvNet to **whole** image

Slides adapted from: Ross Girshick, ICCV 2015

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, ECCV 2014

Towards Fast R-CNN (SPPnet)



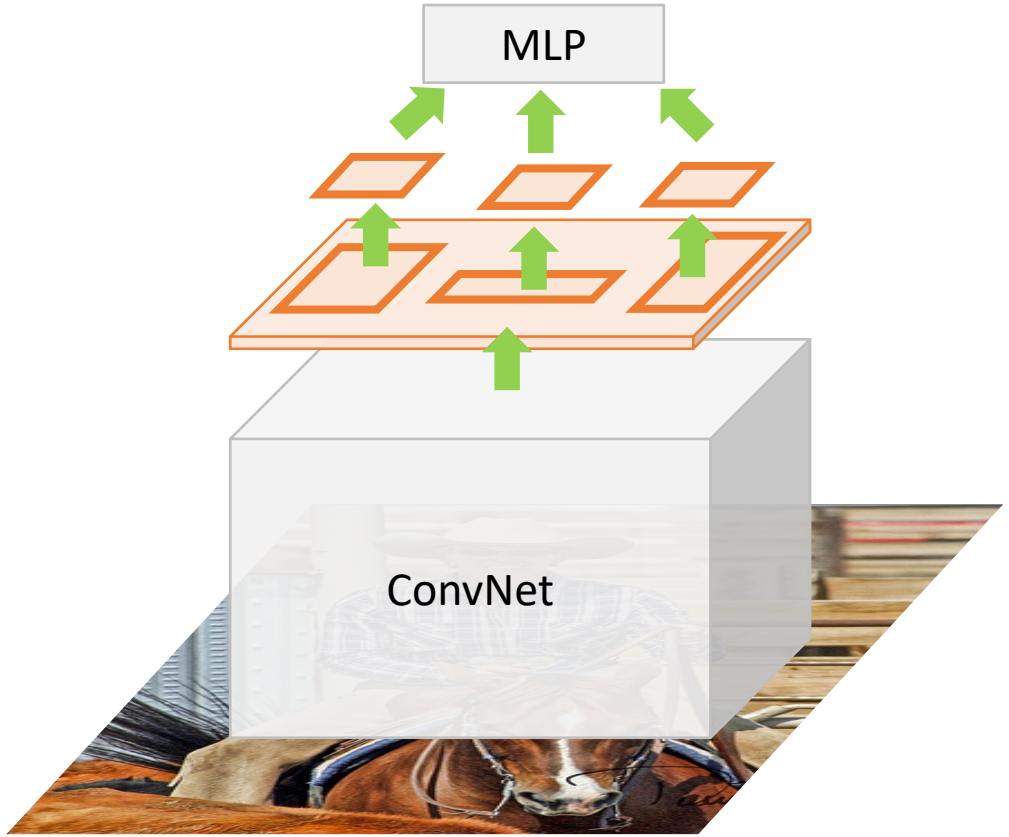
- Crop and **pool** features
- Map proposed RoIs to feature maps
- Apply ConvNet to **whole** image

*In original SPPnet, pooling is done with a fancier variant (spatial pyramid pooling)

Slides adapted from: Ross Girshick, ICCV 2015

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, ECCV 2014

Towards Fast R-CNN (SPPnet)

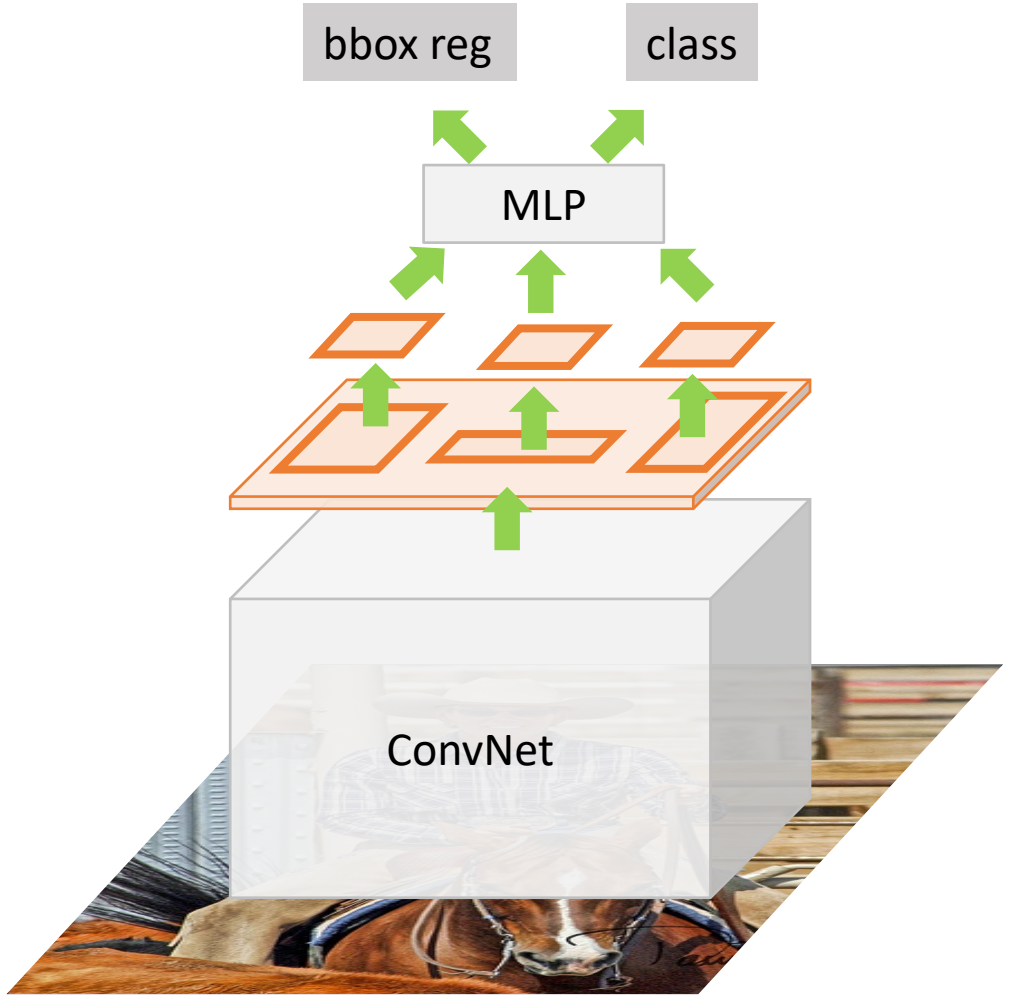


- Apply MLP on regions
- Crop and **pool** features
- Map proposed RoIs to feature maps
- Apply ConvNet to **whole** image

Slides adapted from: Ross Girshick, ICCV 2015

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", ECCV 2014

Towards Fast R-CNN (SPPnet)



- Classify and regress boxes
- Apply MLP on regions
- Crop and **pool** features
- Map proposed RoIs to feature maps
- Apply ConvNet to **whole** image

Slides adapted from: Ross Girshick, ICCV 2015

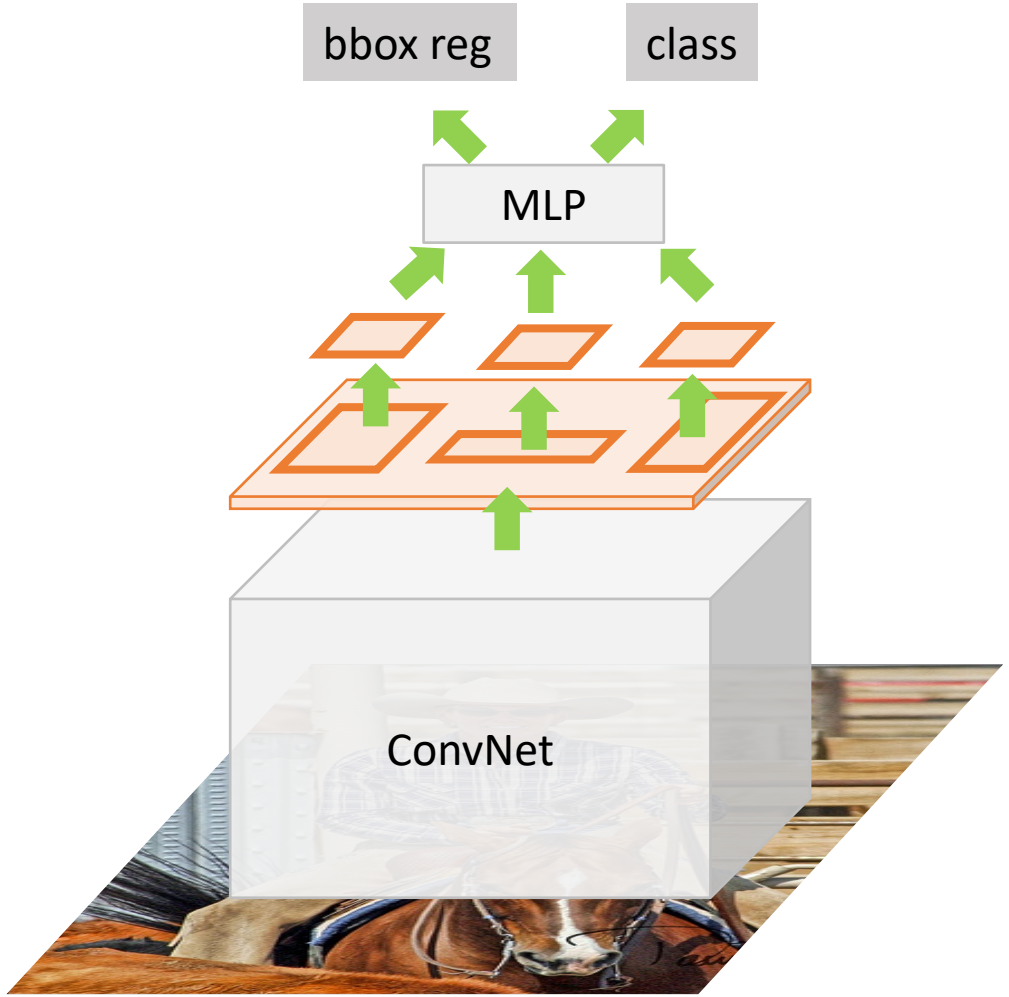
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", ECCV 2014

Towards Fast R-CNN (SPPnet)

- **What's good with "fully convolutional" ConvNet?**

many but **small** MLPs on regions

one "fully conv" ConvNet



- Classify and regress boxes
- Apply MLP on regions
- Crop and **pool** features
- Map proposed RoIs to feature maps
- Apply ConvNet to **whole** image

Slides adapted from: Ross Girshick, ICCV 2015

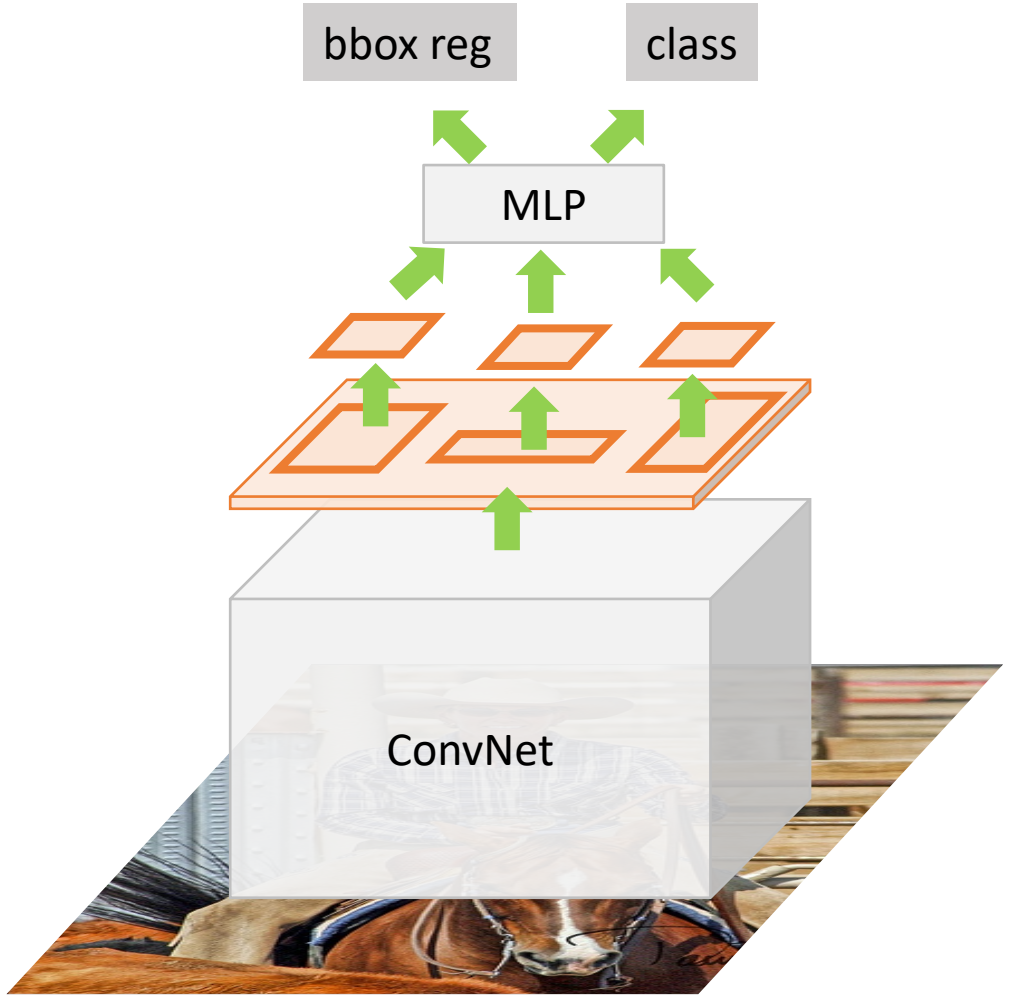
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", ECCV 2014

Towards Fast R-CNN (SPPnet)

- **What's still wrong?** Not yet "Differentiable Programming"

batching is **w.r.t. regions** (not images)

frozen pre-trained ConvNet



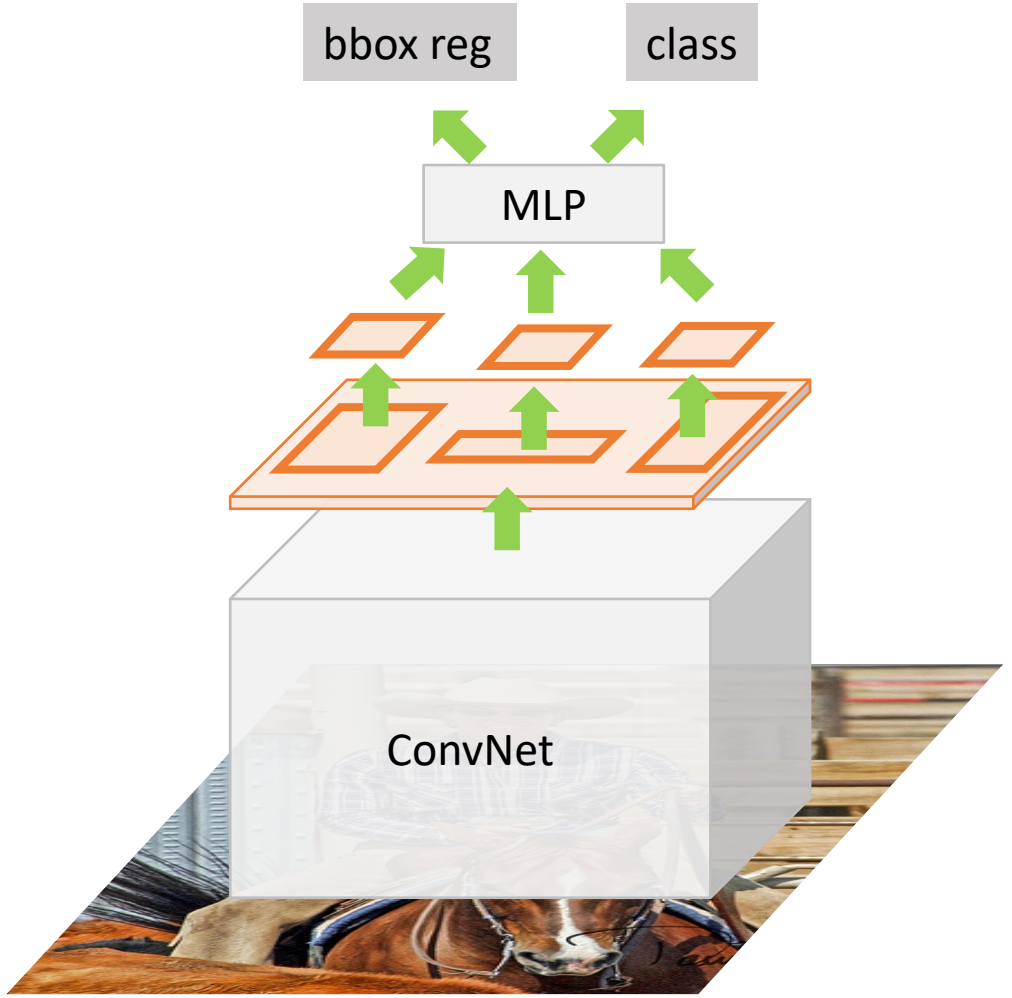
- Classify and regress boxes
- Apply MLP on regions
- Crop and **pool** features
- Map proposed RoIs to feature maps
- Apply ConvNet to **whole** image

Slides adapted from: Ross Girshick, ICCV 2015

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", ECCV 2014

Fast R-CNN

- **Differentiable Programming**

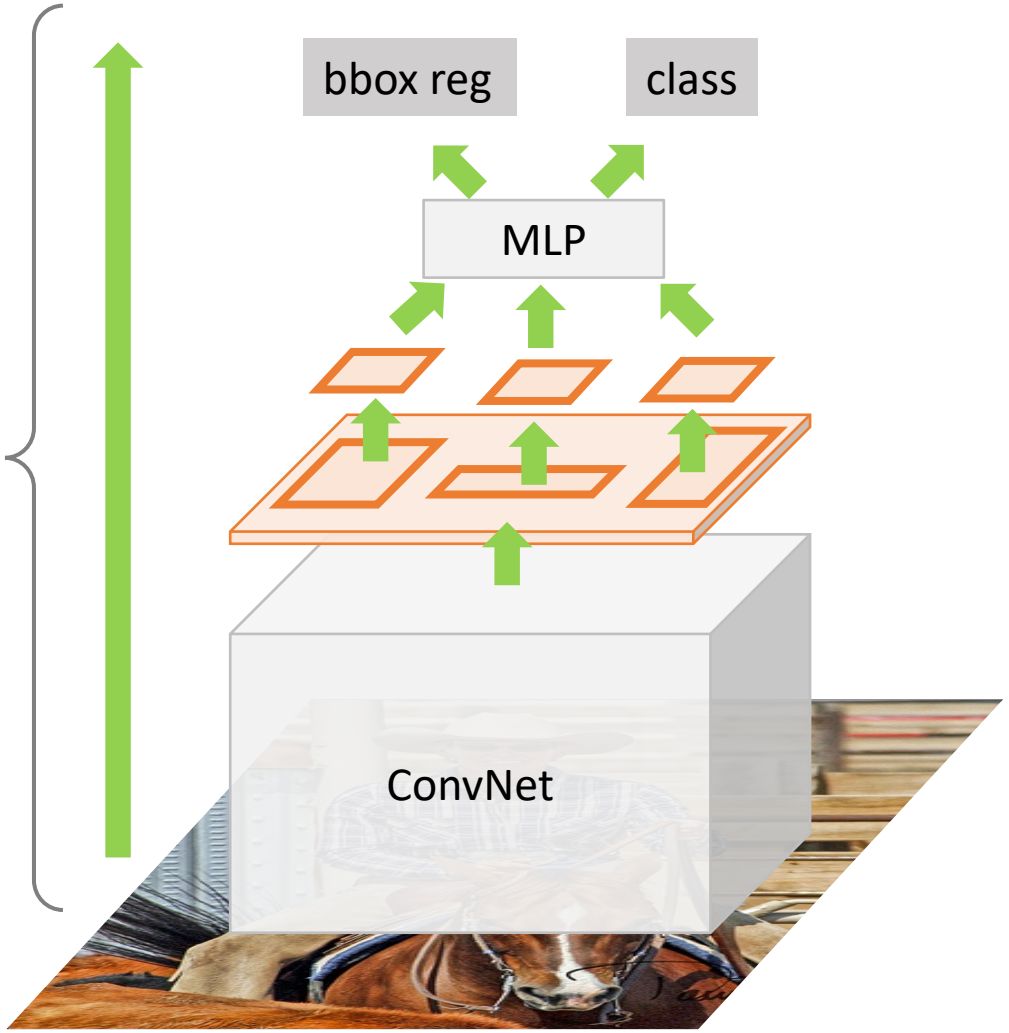


Slides adapted from: Ross Girshick, ICCV 2015
Ross Girshick, “Fast R-CNN”, ICCV 2015

Fast R-CNN

- **Differentiable Programming**

forward path
(a program)

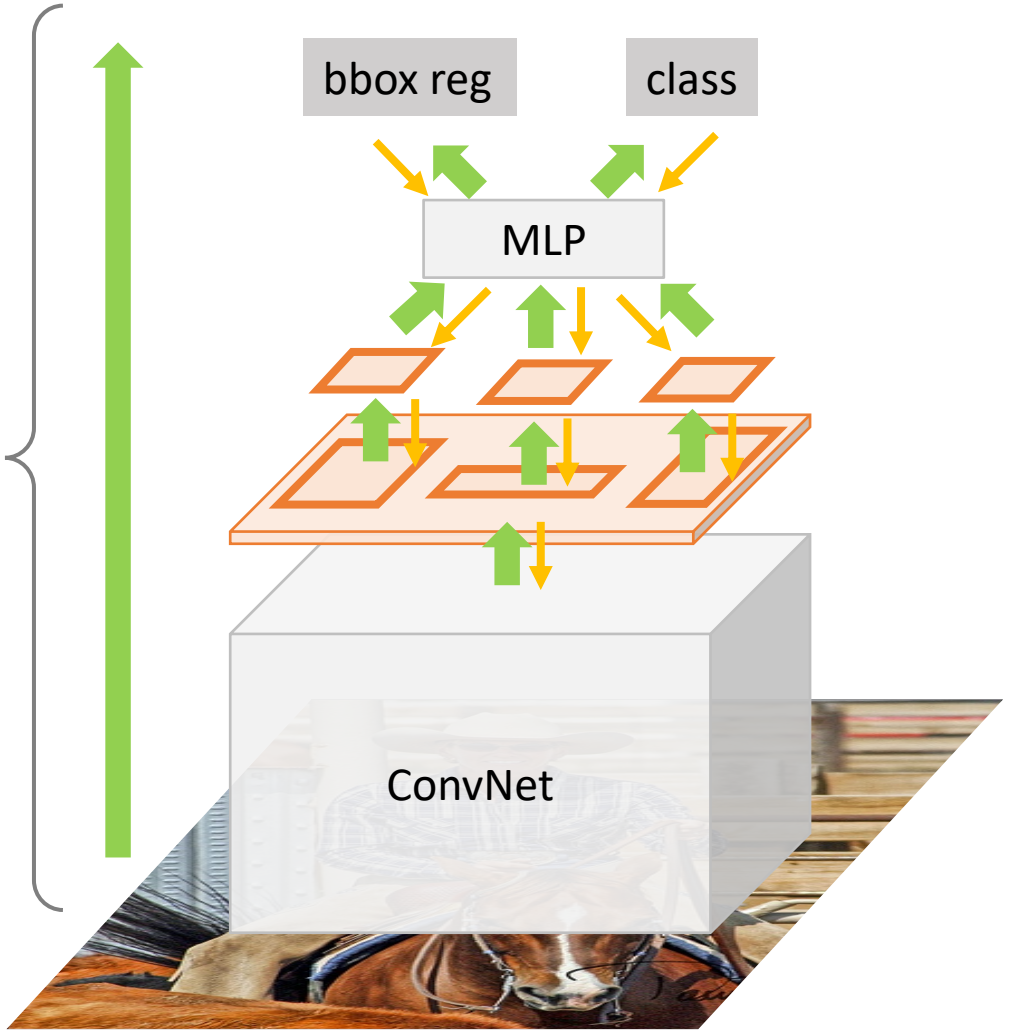


Slides adapted from: Ross Girshick, ICCV 2015
Ross Girshick, “Fast R-CNN”, ICCV 2015

Fast R-CNN

- **Differentiable Programming**

forward path
(a program)

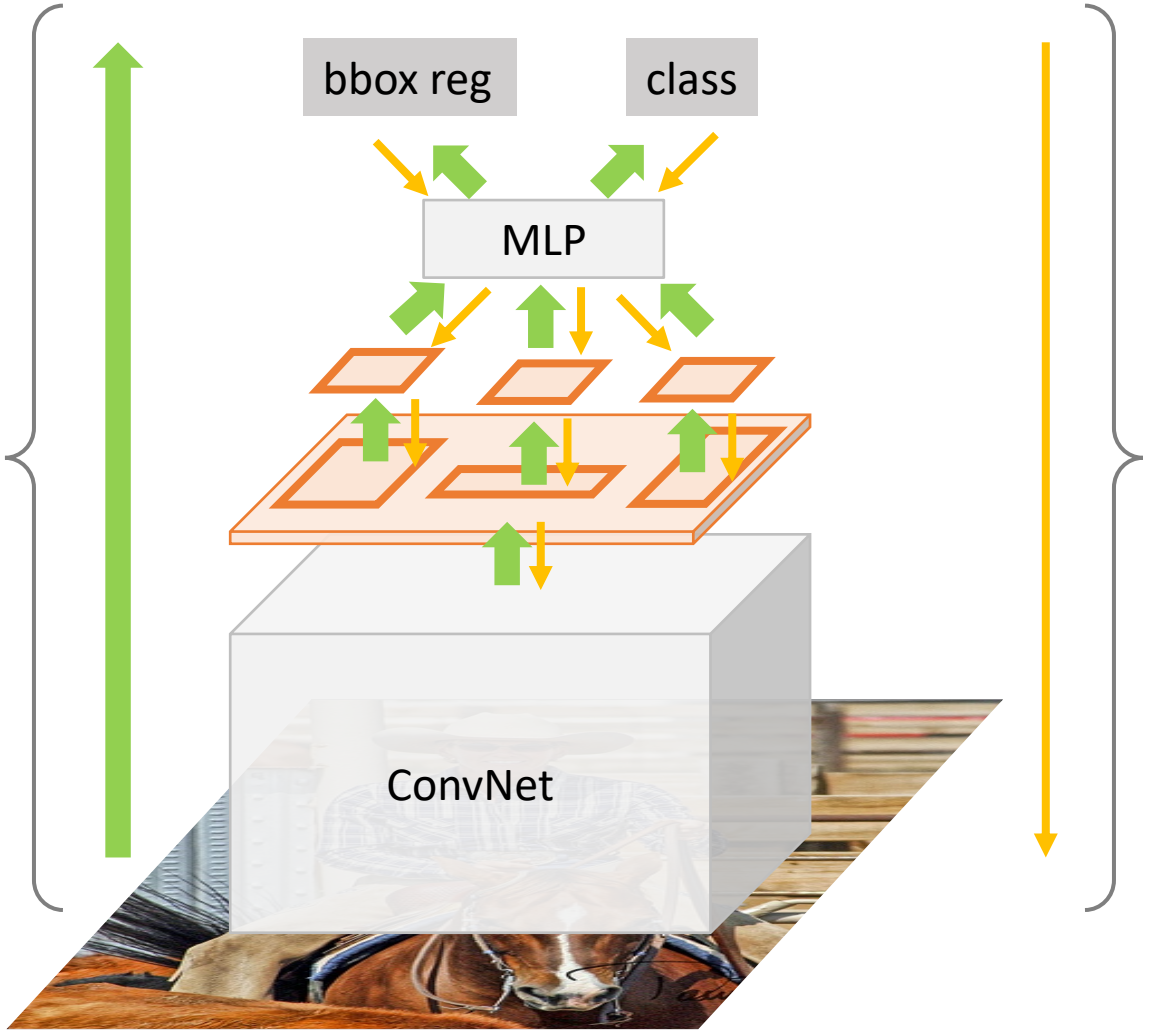


Slides adapted from: Ross Girshick, ICCV 2015
Ross Girshick, “Fast R-CNN”, ICCV 2015

Fast R-CNN

- **Differentiable Programming**

forward path
(a program)



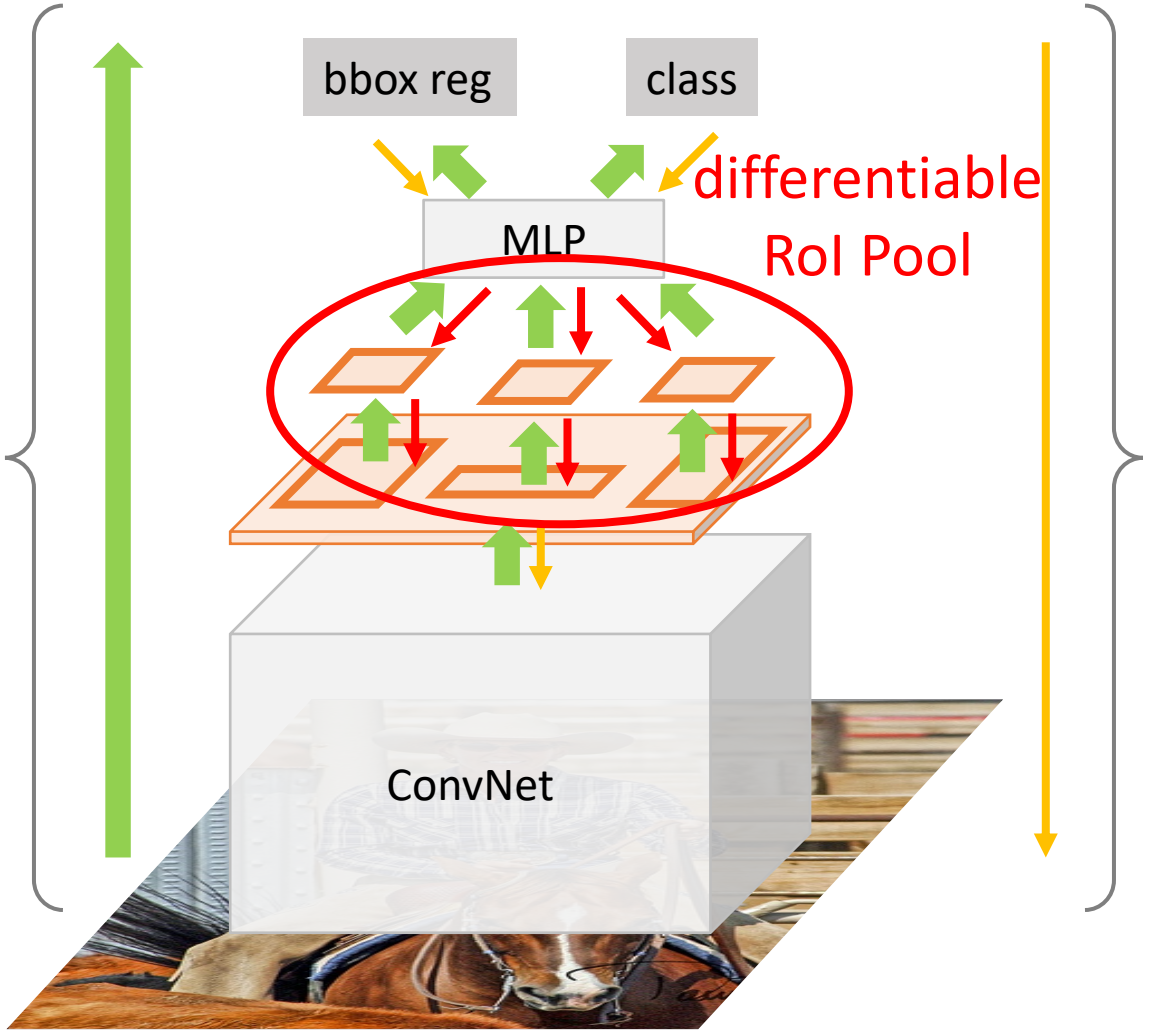
backward path
(differentiable)

Slides adapted from: Ross Girshick, ICCV 2015
Ross Girshick, “Fast R-CNN”, ICCV 2015

Fast R-CNN

- **Differentiable Programming**

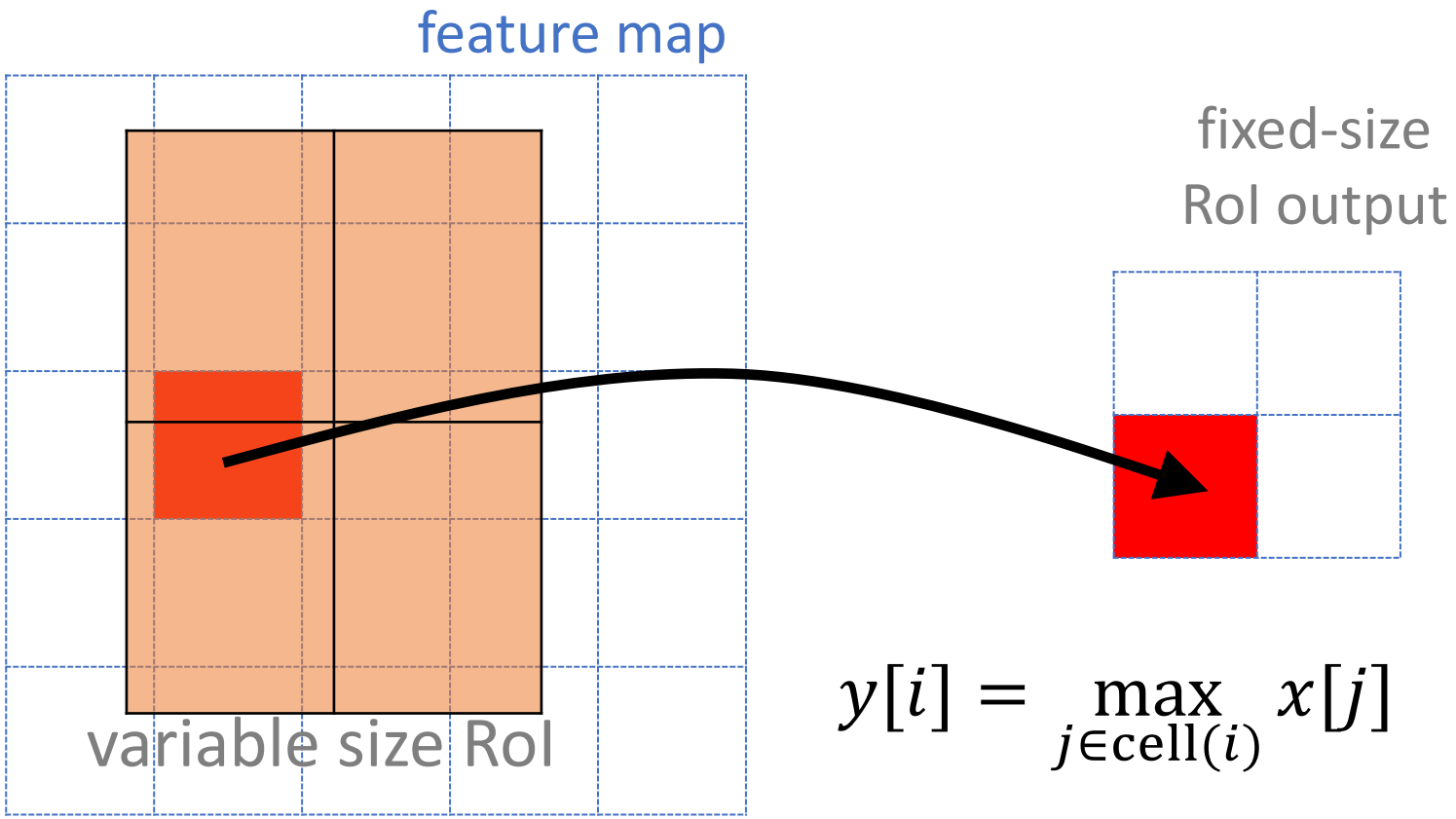
forward path
(a program)



backward path
(differentiable)

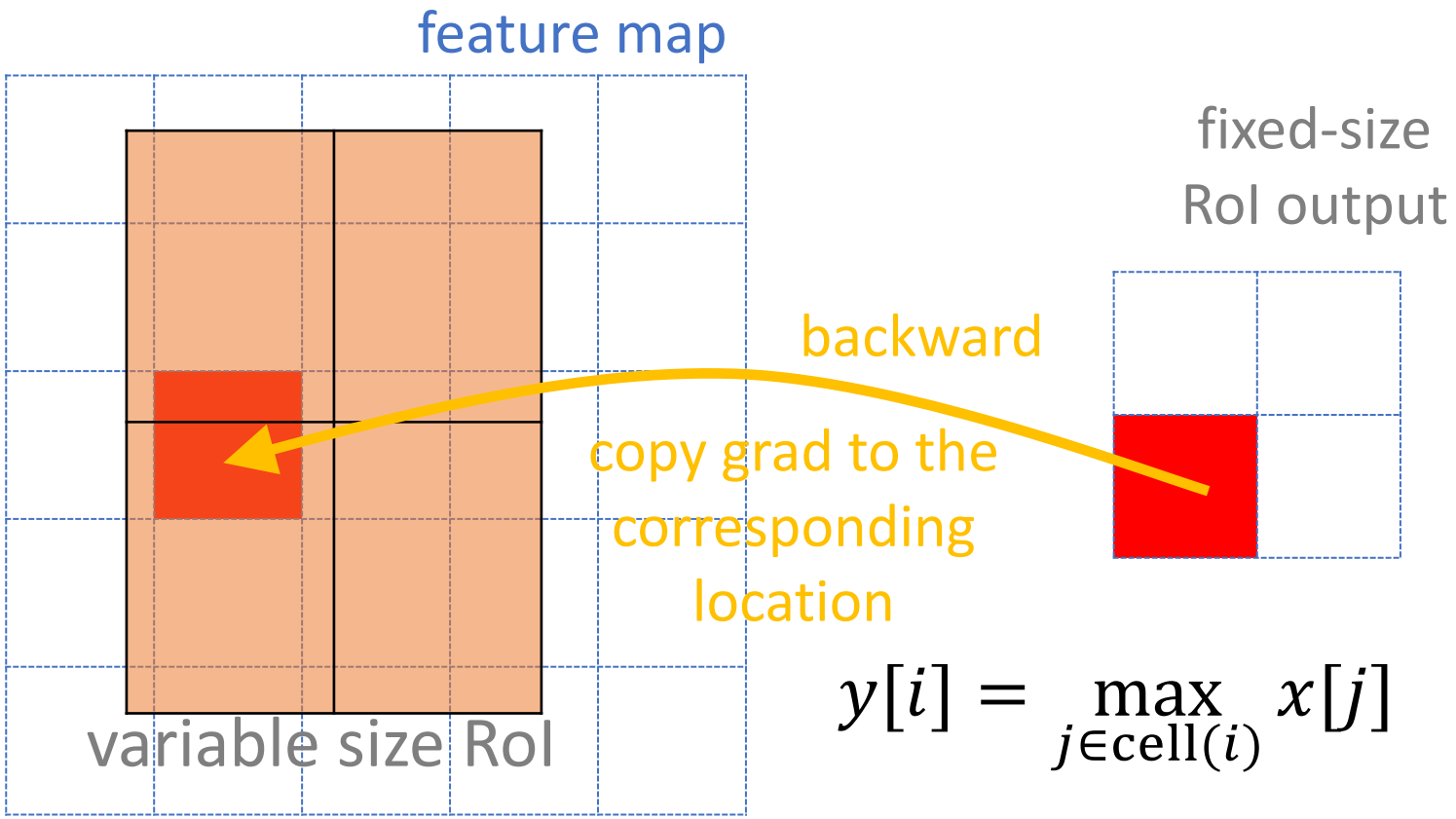
Slides adapted from: Ross Girshick, ICCV 2015
Ross Girshick, “Fast R-CNN”, ICCV 2015

Differentiable RoI Pooling



Ross Girshick, “Fast R-CNN”, ICCV 2015

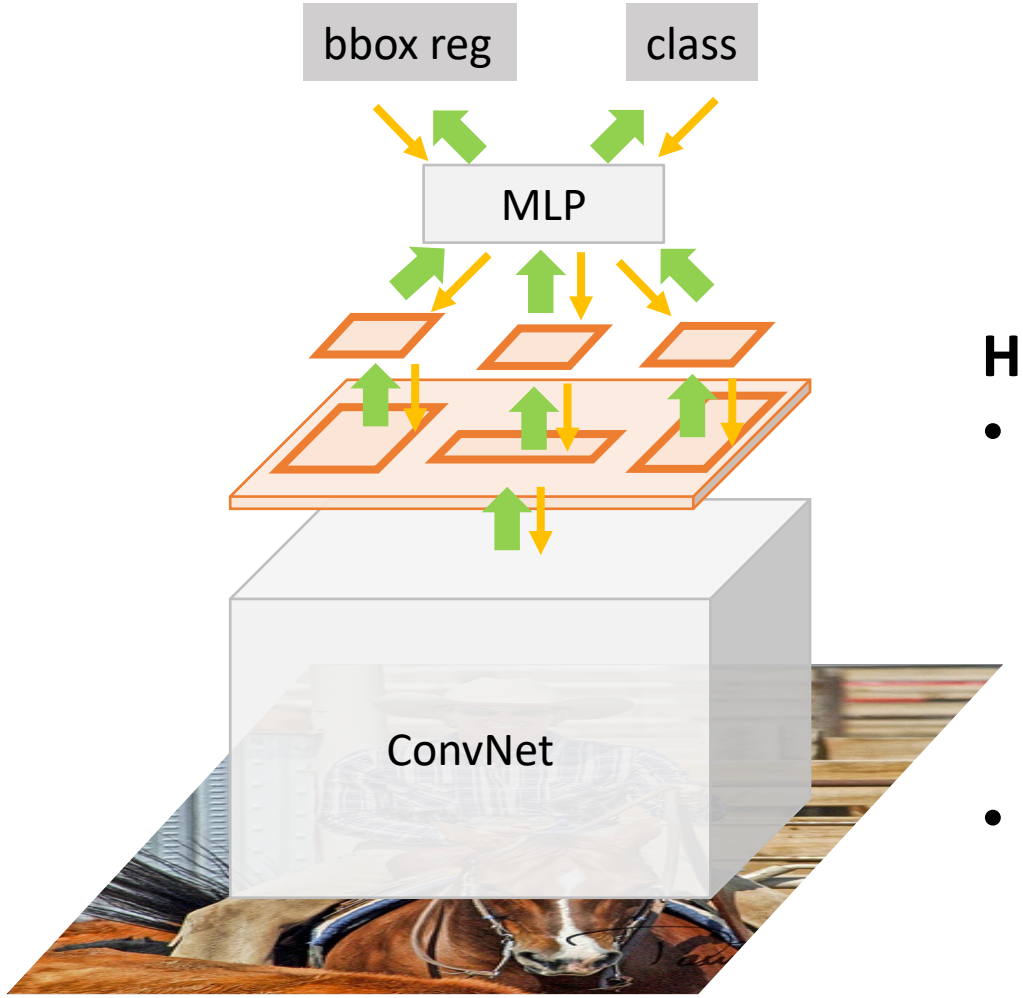
Differentiable RoI Pooling



Ross Girshick, “Fast R-CNN”, ICCV 2015

Fast R-CNN

- **Differentiable Programming**



Hierarchical batching

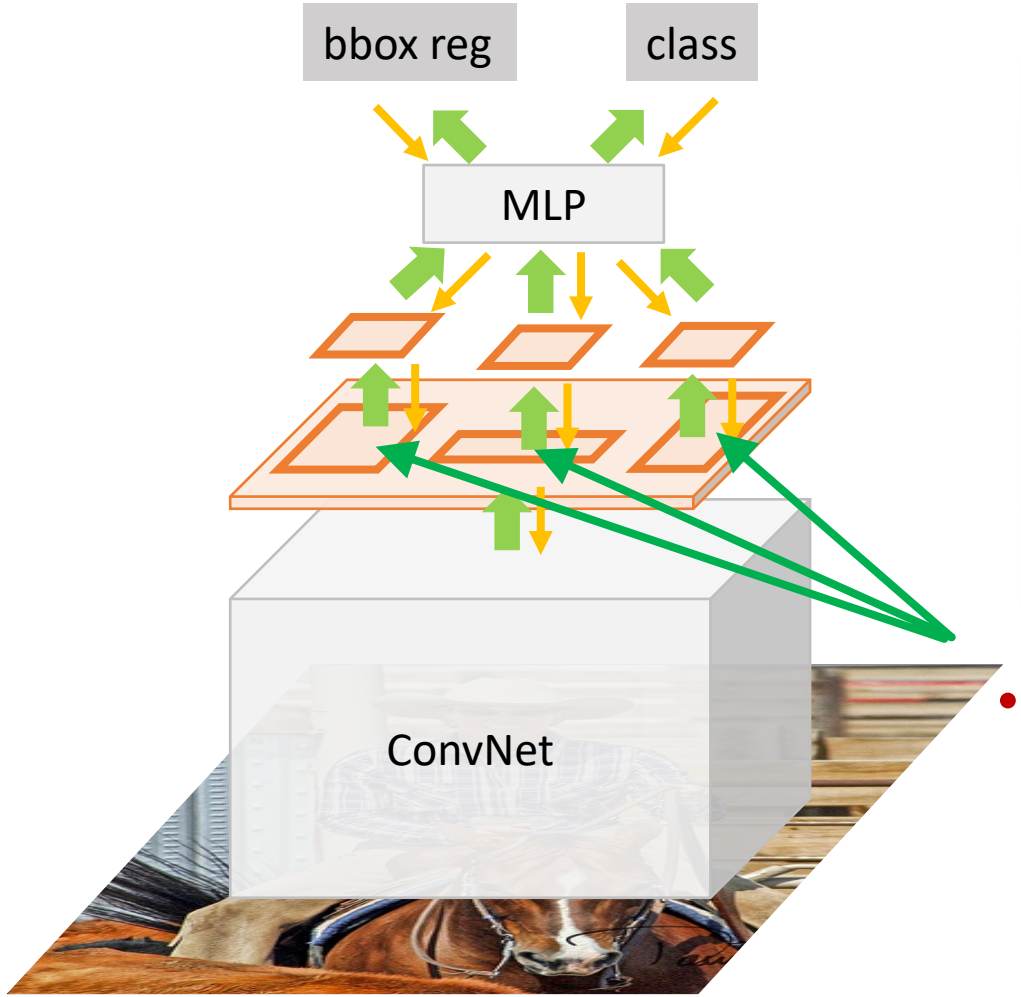
- for each image, sample a batch of RoIs
- sample a batch of images

Slides adapted from: Ross Girshick, ICCV 2015

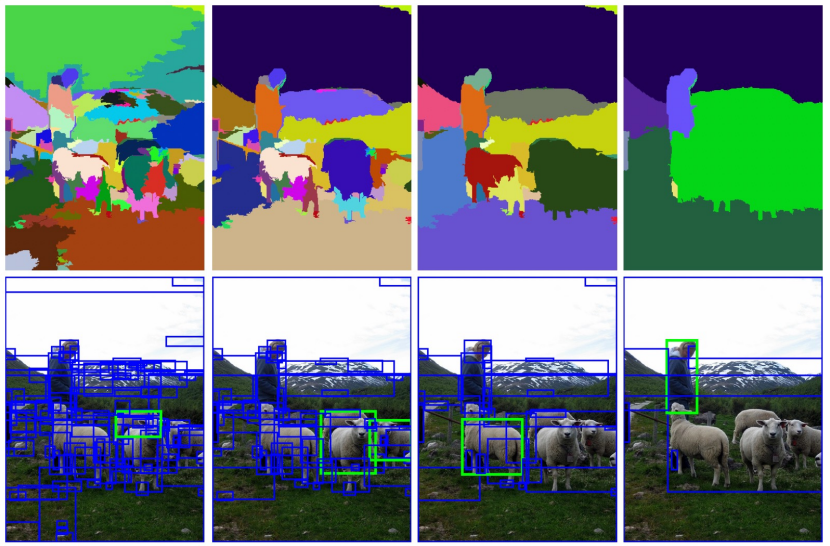
Ross Girshick, “Fast R-CNN”, ICCV 2015

Fast R-CNN

- **What's still wrong?** Region proposals are not part of the program



recap: Selective Search



- Propose Regions

Slides adapted from: Ross Girshick, ICCV 2015

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

Faster R-CNN

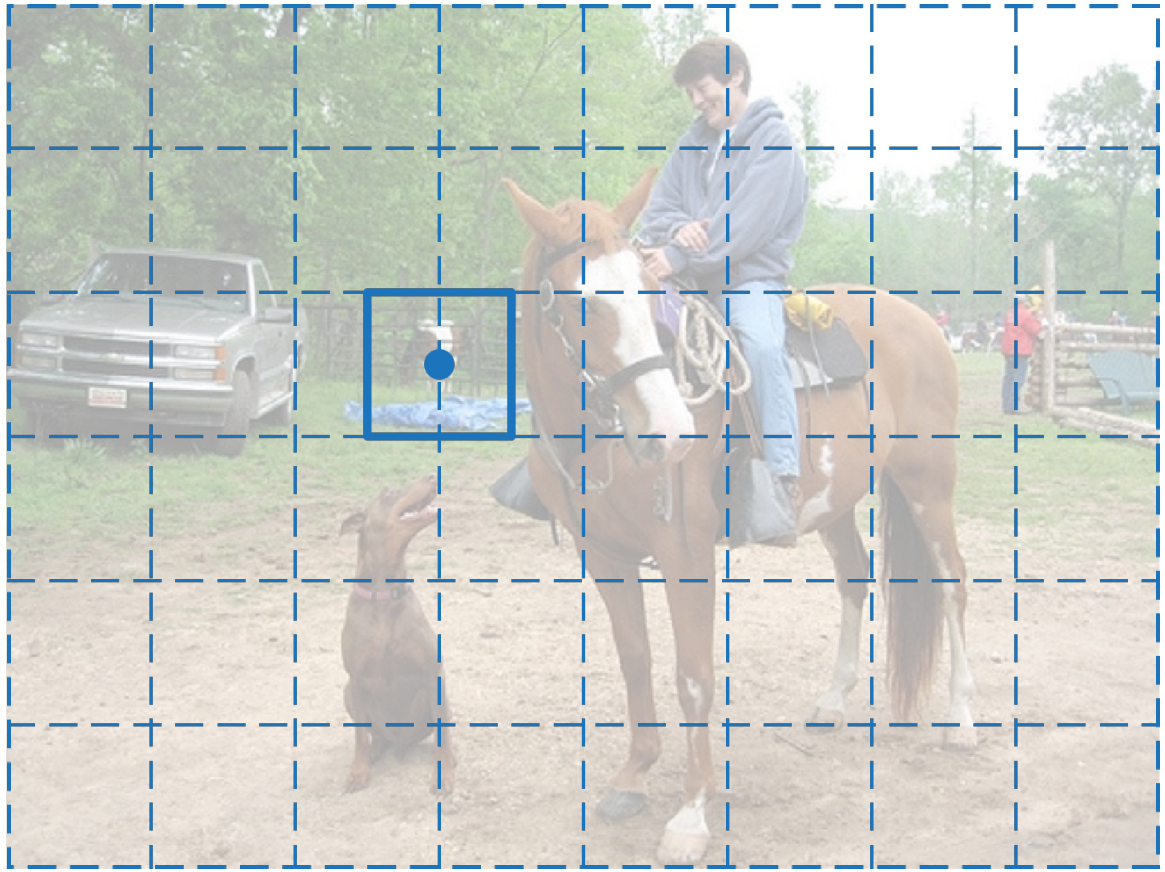
- **How can neural nets produce regions?**

- **ConvNets are Fully Convolutional!**
- Search objects w/ sliding windows

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN

- **How can neural nets produce regions?**

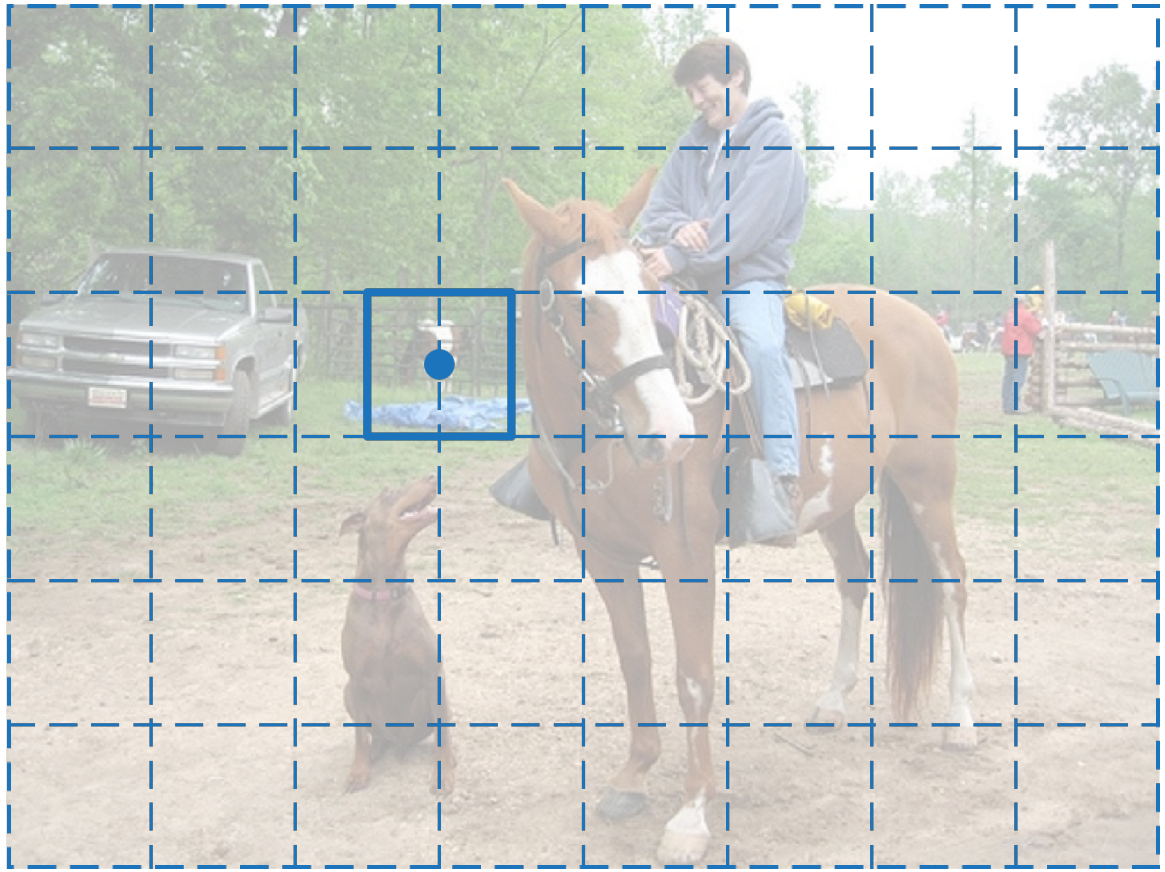


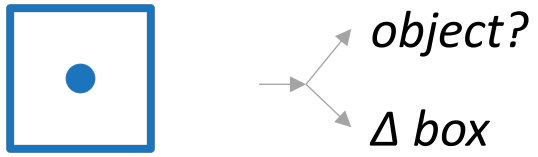
- **ConvNets are Fully Convolutional!**
- Search objects w/ sliding windows

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN

- **How can neural nets produce regions?**





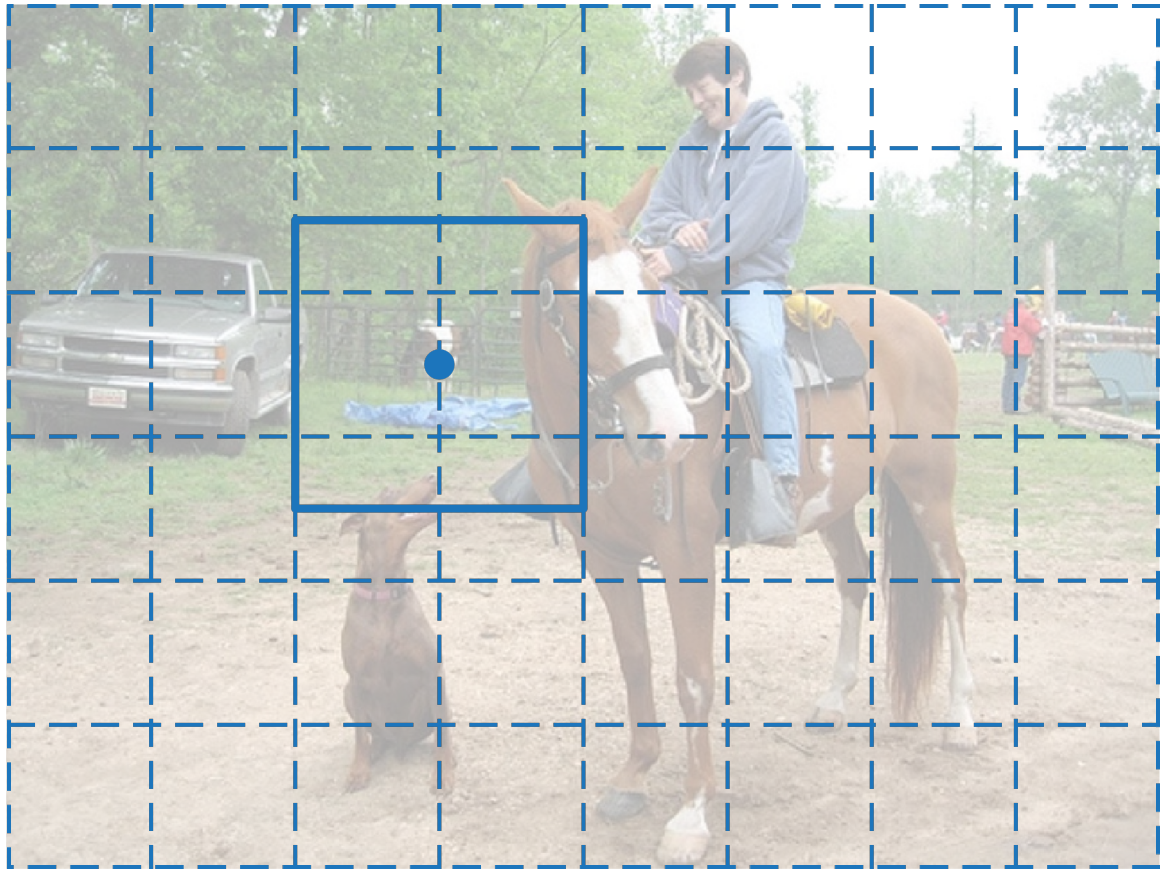
For each window, predict:

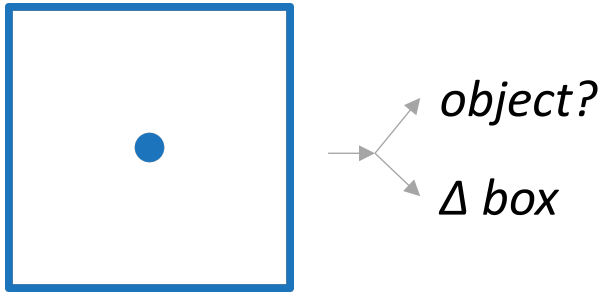
- **Is it an object?**
  - binary classification
- **Object location w.r.t. window?**
  - bbox regression

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN

- How can neural nets produce regions?





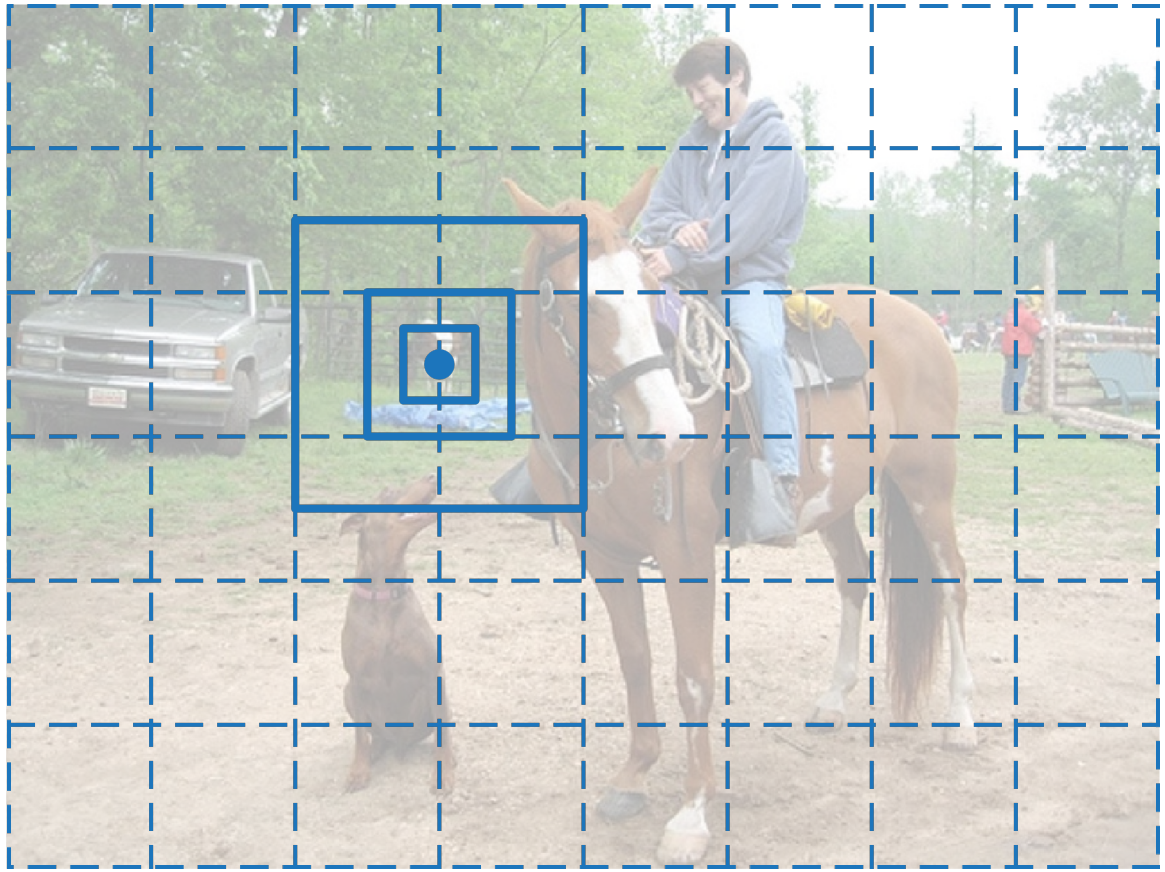
Anchor

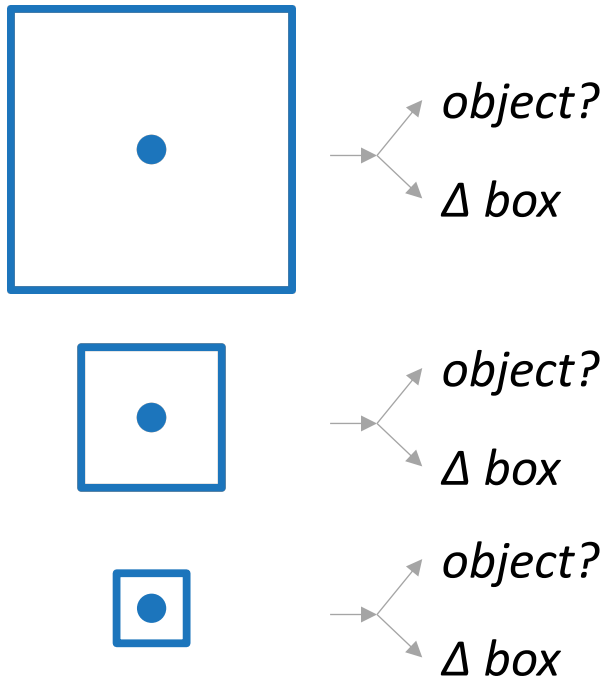
- pred w.r.t. a reference window
- reference is called an “anchor”
- be different w/ sliding window

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN

- How can neural nets produce regions?





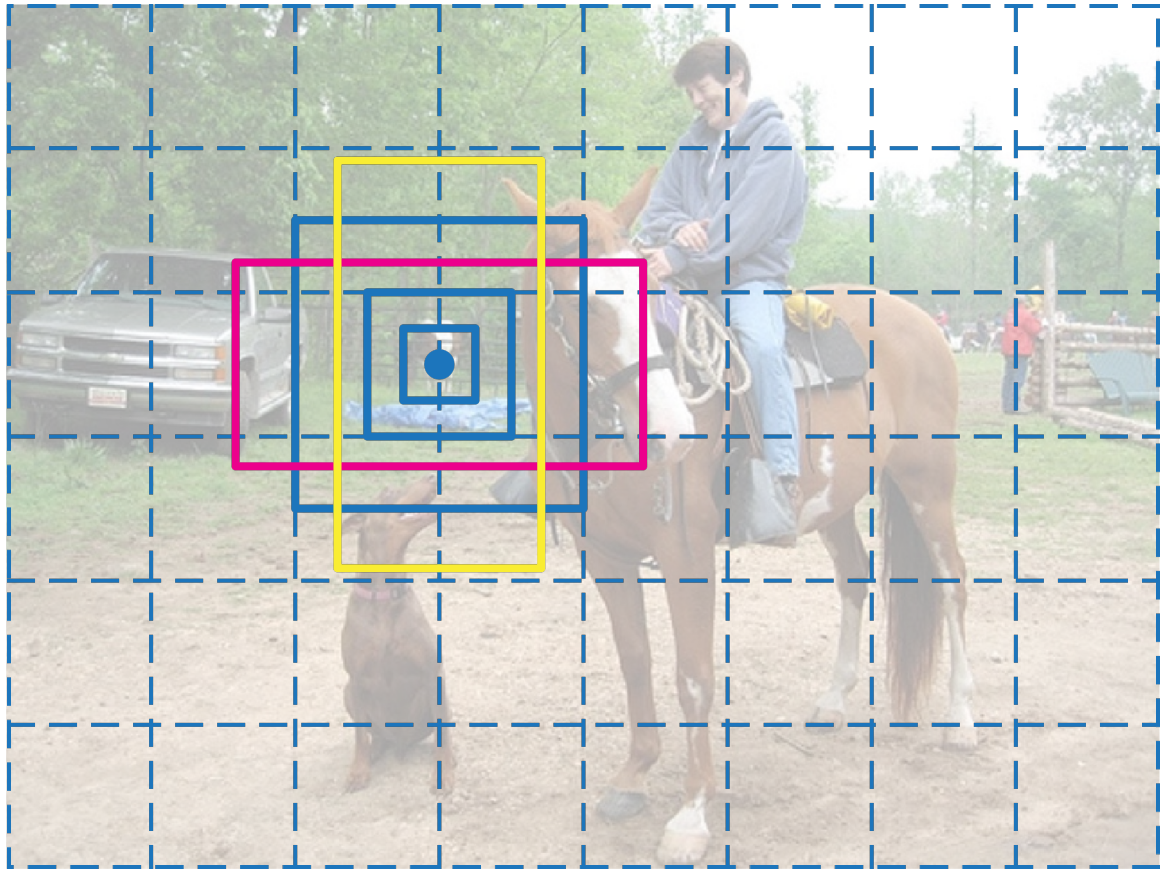
Anchors

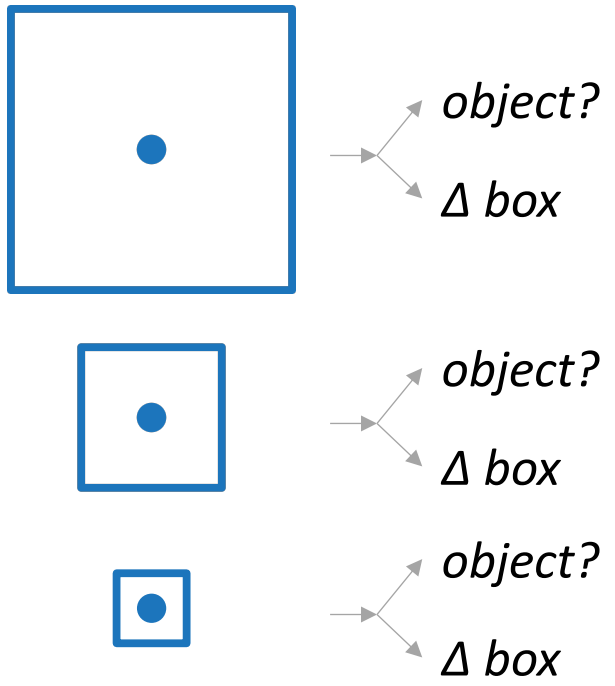
- one sliding window, many anchors
- multiple scales

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN

- How can neural nets produce regions?





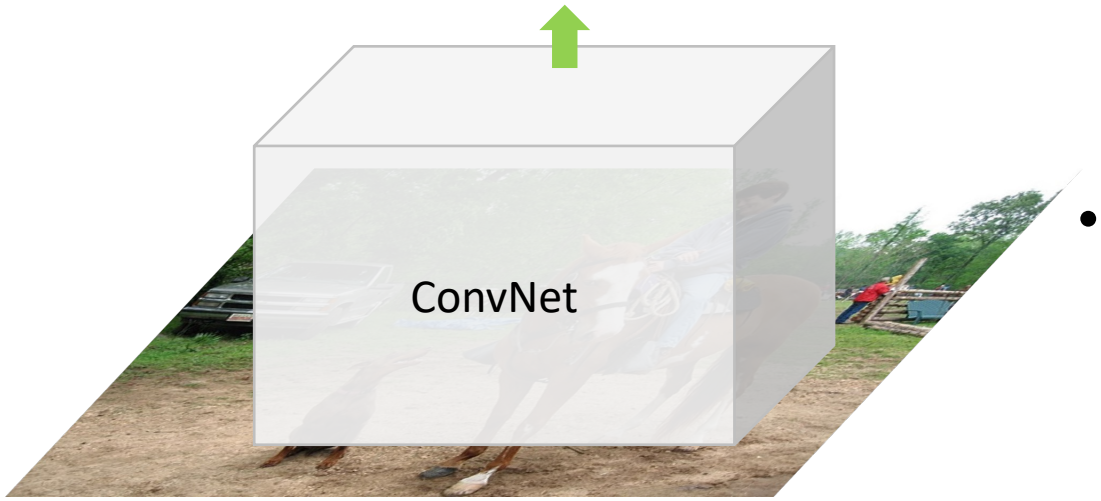
Anchors

- one sliding window, many anchors
- multiple scales
- multiple aspect ratios

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: Region Proposal Network (RPN)

- **How can neural nets produce regions?**

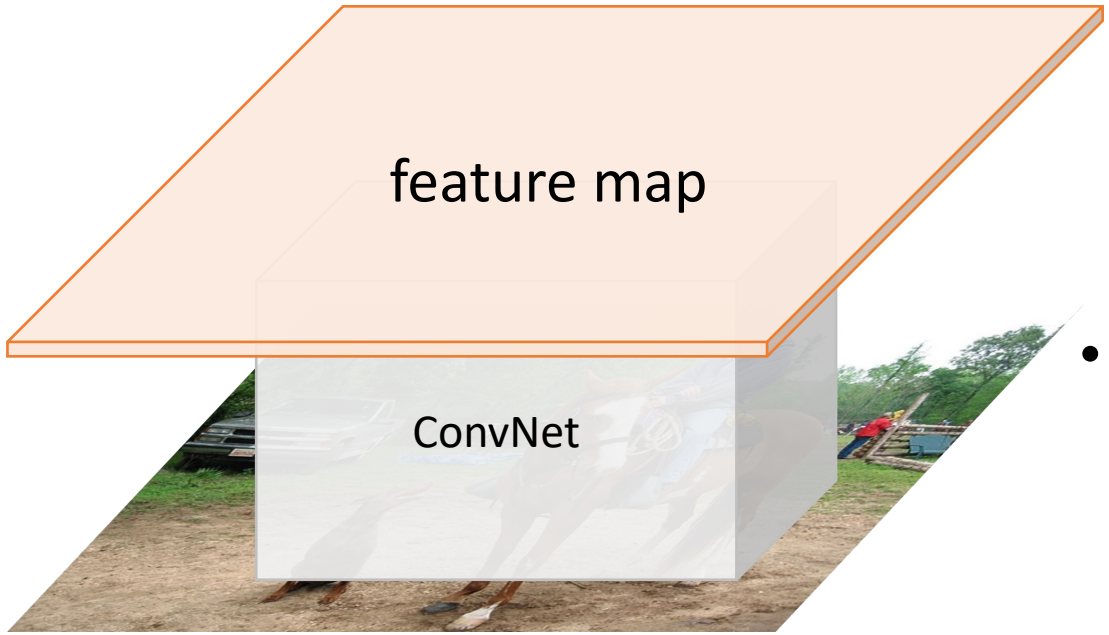


- Apply ConvNet to **whole** image

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN: Region Proposal Network (RPN)

- How can neural nets produce regions?

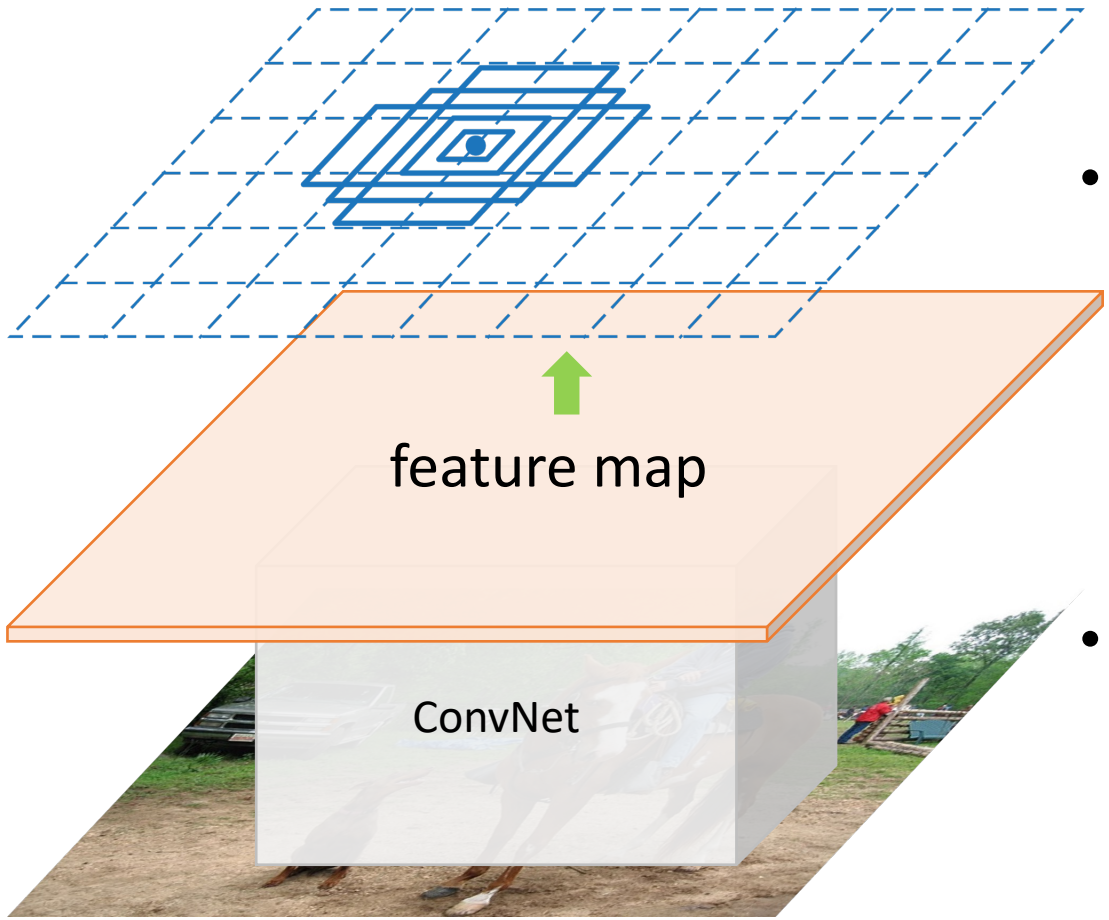


- Apply ConvNet to **whole** image

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: Region Proposal Network (RPN)

- **How can neural nets produce regions?**



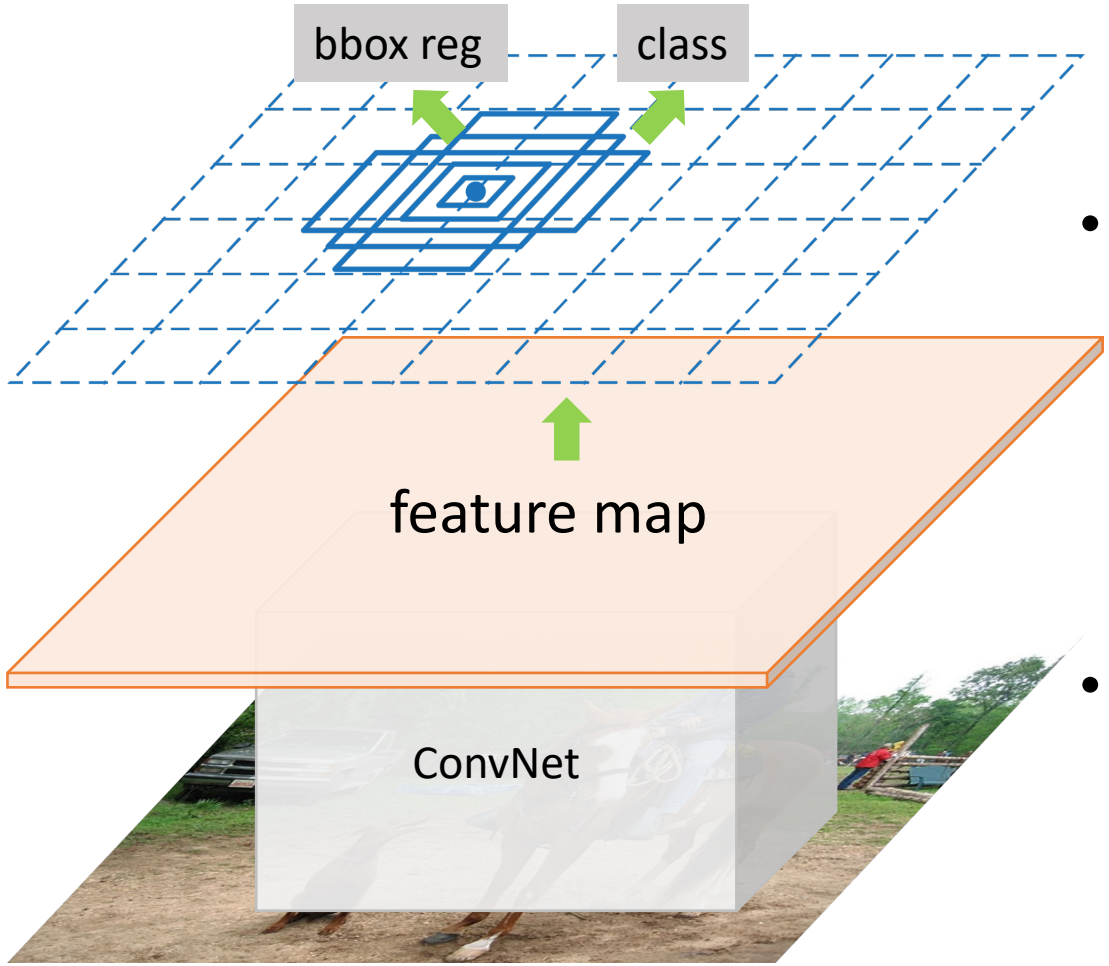
- Convs for k anchors
  - $2k$-d: binary classification
  - $4k$-d: bbox regression

- Apply ConvNet to **whole** image

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: Region Proposal Network (RPN)

- **How can neural nets produce regions?**

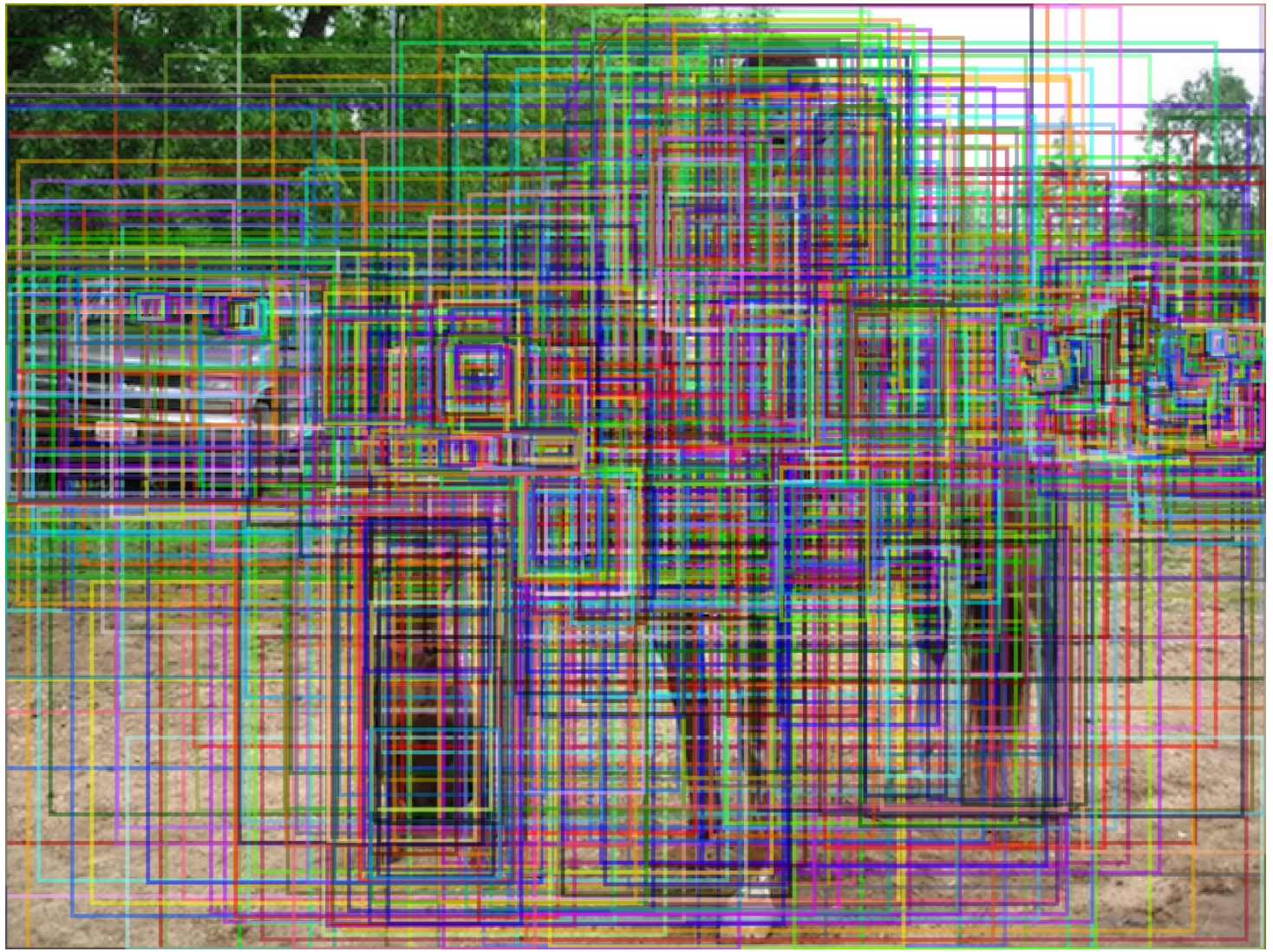


- Loss: reg + class
  - over all locations
  - over all anchors
- Convs for k anchors
  - $2k$-d: binary classification
  - $4k$-d: bbox regression
- Apply ConvNet to **whole** image

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: Region Proposal Network (RPN)

Example: 1000 regions proposed by RPN



Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN: End-to-End Object Detection

- **End-to-End Differentiable Programming**

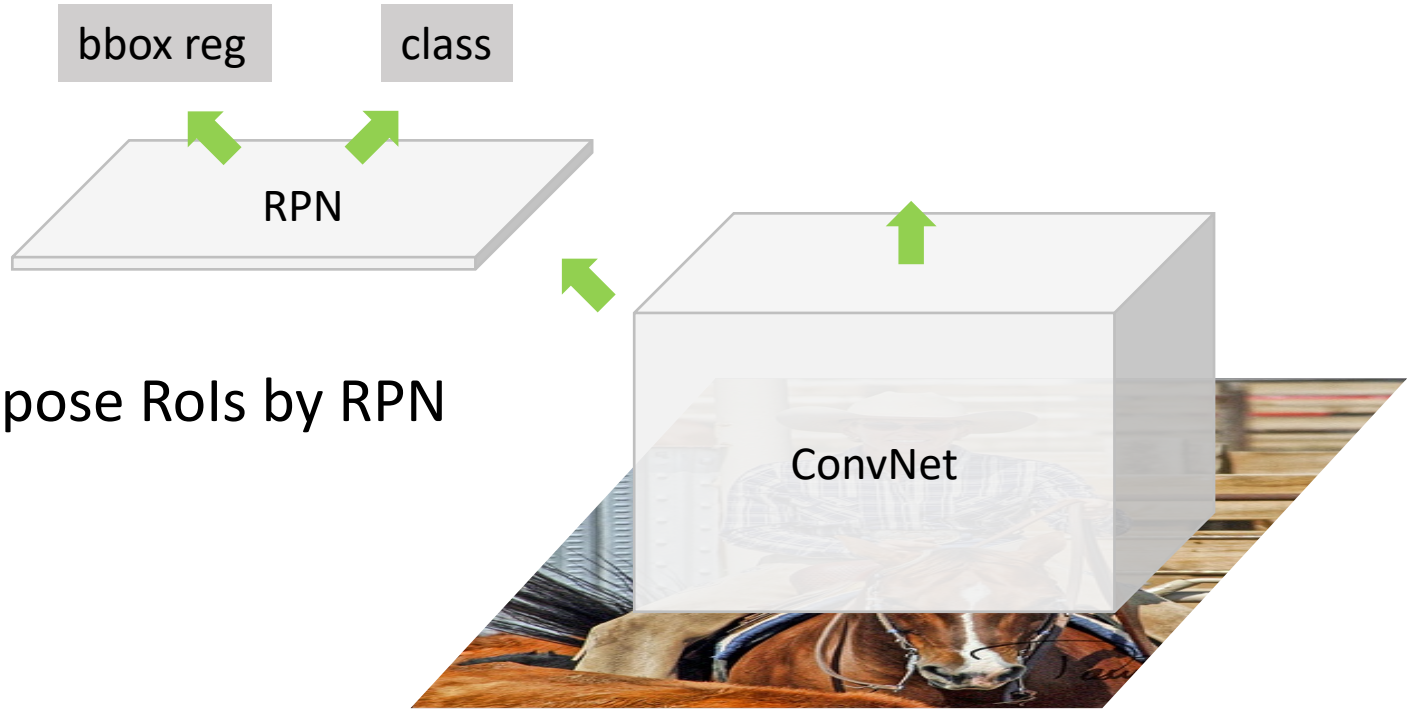
- Shared ConvNet **backbone** on whole image

Slides adapted from: Ross Girshick, ICCV 2015

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN: End-to-End Object Detection

- **End-to-End Differentiable Programming**



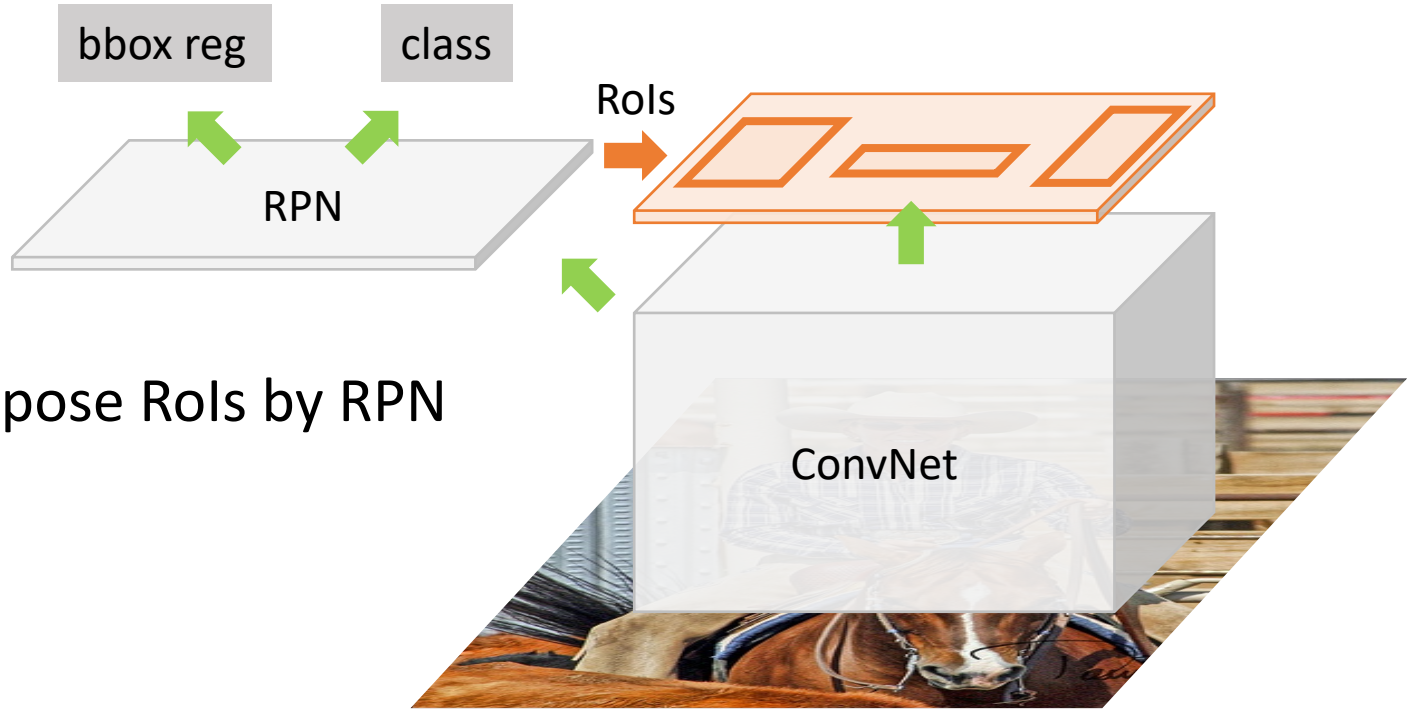
- Propose RoIs by RPN
- Shared ConvNet **backbone** on whole image

Slides adapted from: Ross Girshick, ICCV 2015

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: End-to-End Object Detection

- **End-to-End Differentiable Programming**



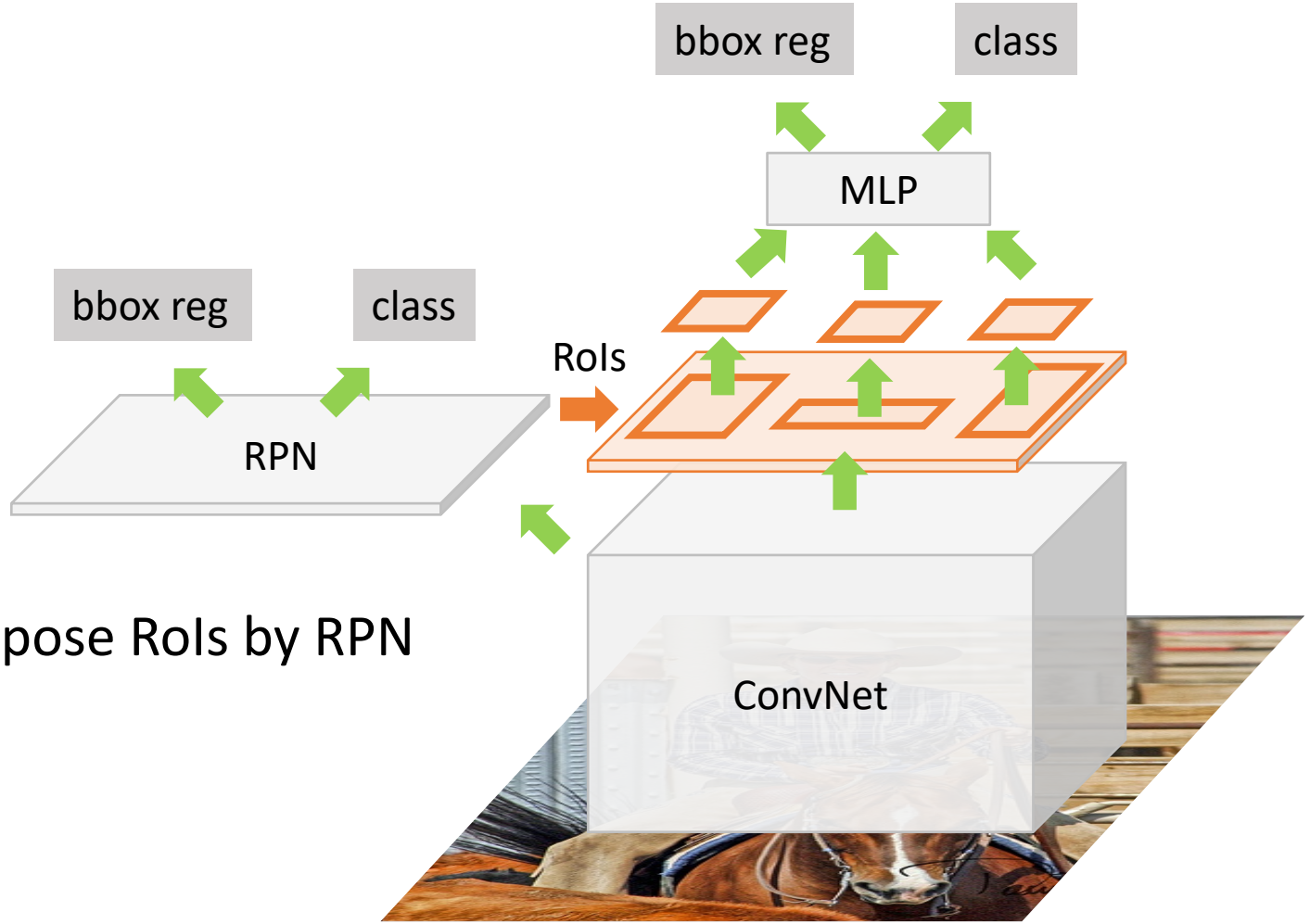
- Propose RoIs by RPN
- Apply RPN's RoIs
- Shared ConvNet **backbone** on whole image

Slides adapted from: Ross Girshick, ICCV 2015

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: End-to-End Object Detection

- **End-to-End Differentiable Programming**



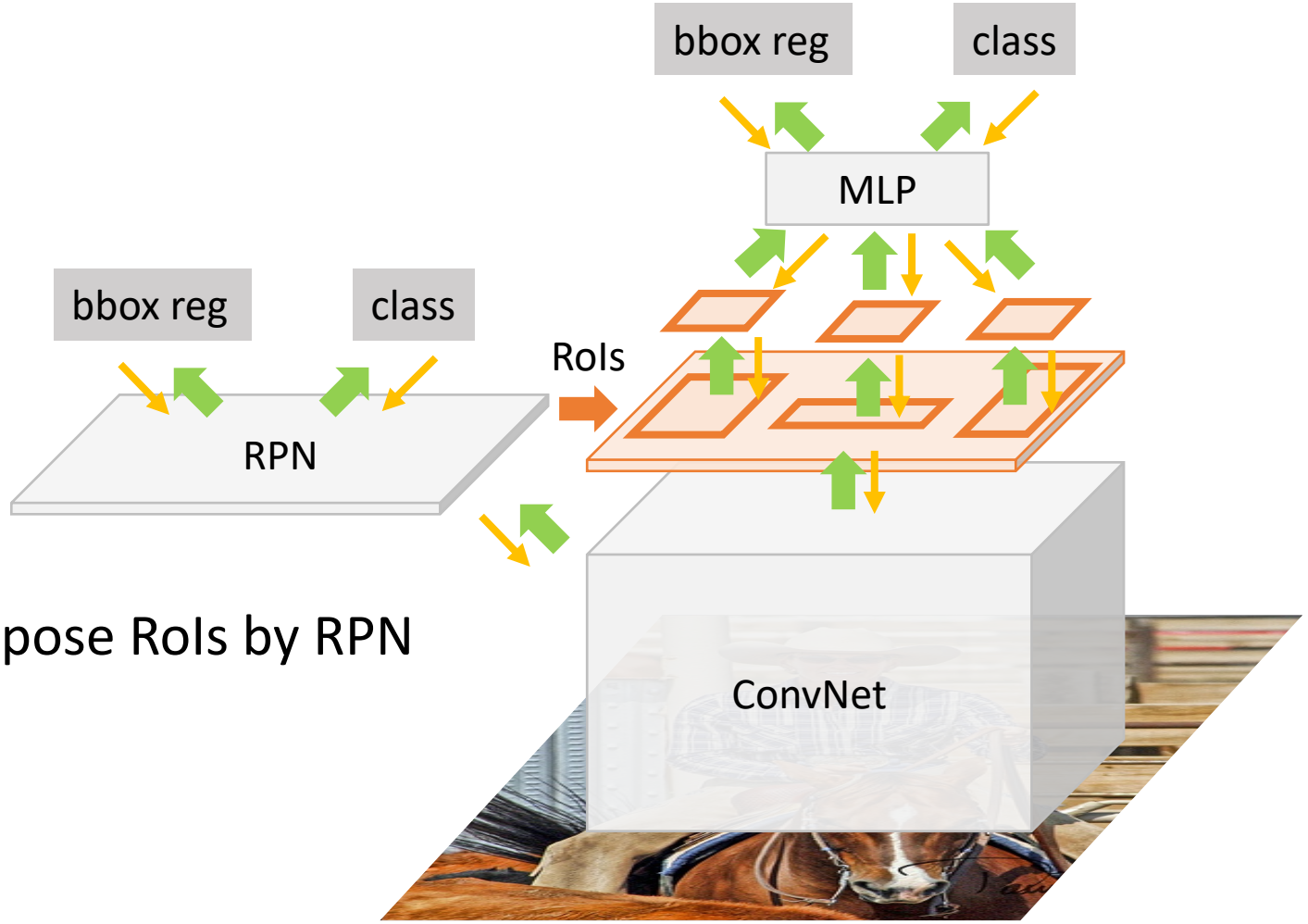
- Same as Fast R-CNN
- Apply RPN's RoIs
- Shared ConvNet **backbone** on whole image
- Propose RoIs by RPN

Slides adapted from: Ross Girshick, ICCV 2015

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: End-to-End Object Detection

- **End-to-End Differentiable Programming**



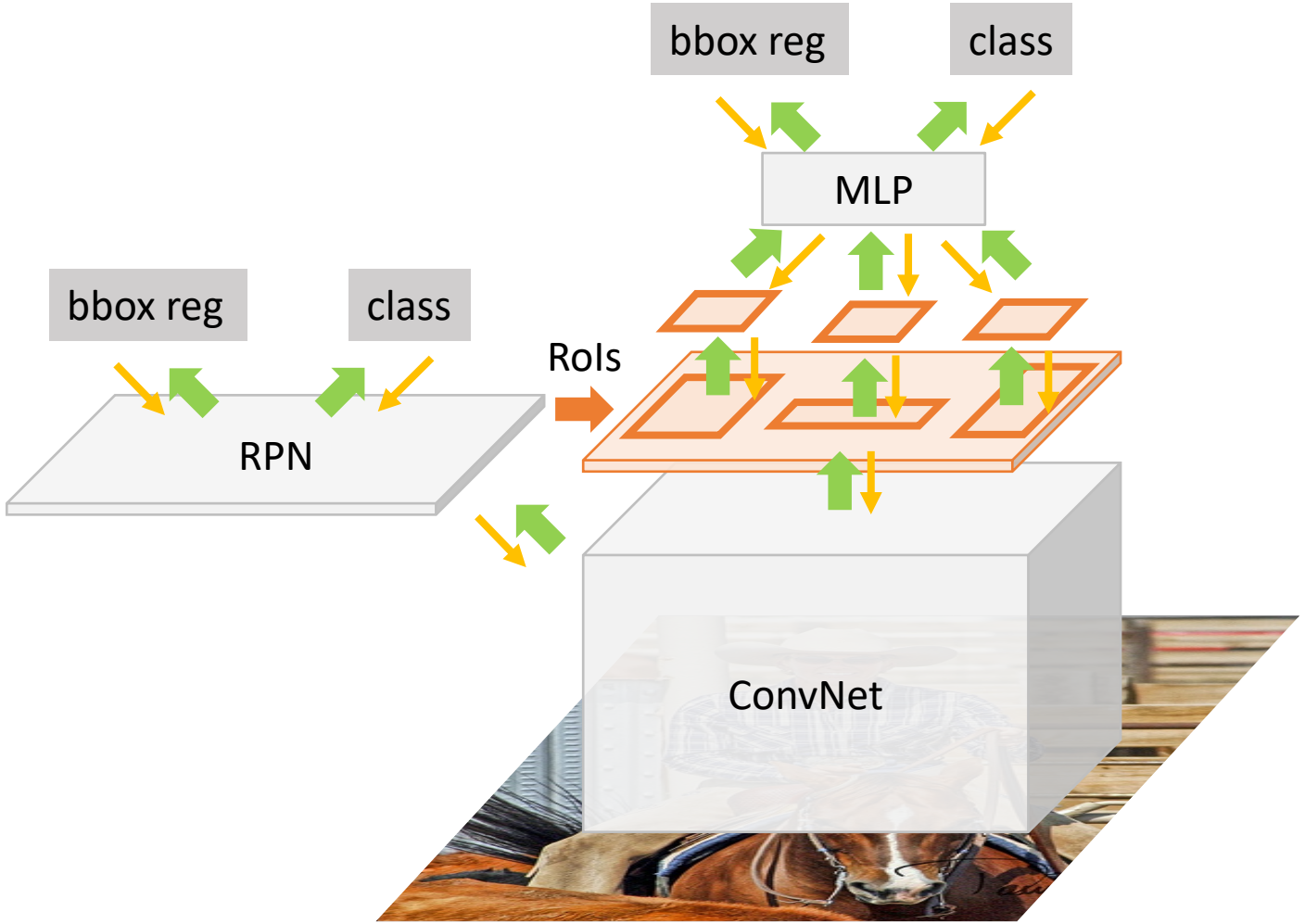
- Same as Fast R-CNN
- Apply RPN's RoIs
- Shared ConvNet **backbone** on whole image
- Propose RoIs by RPN

Slides adapted from: Ross Girshick, ICCV 2015

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN: End-to-End Object Detection

- **End-to-End Differentiable Programming**



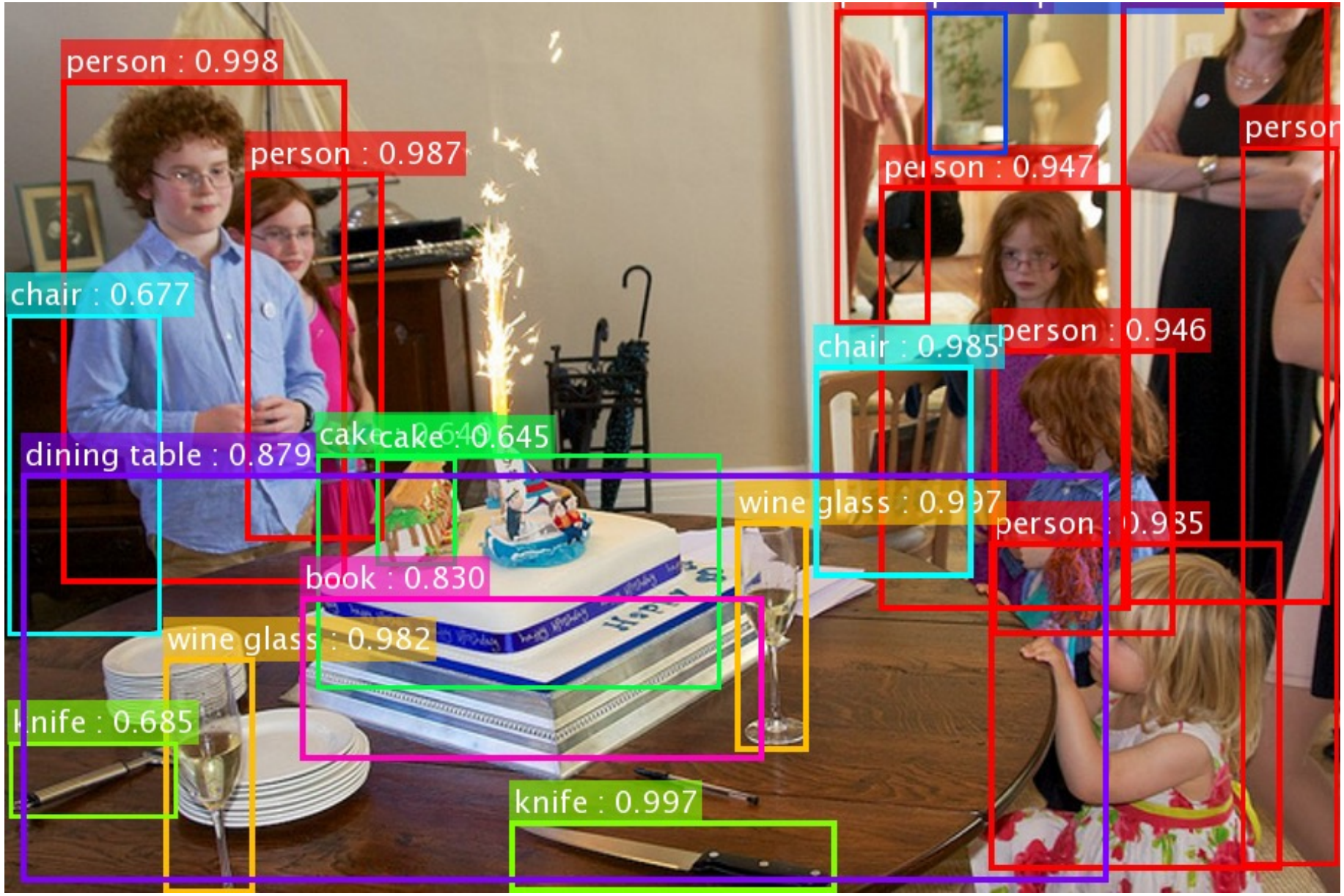
Output end:
bboxes + classes

Input end:
image

Slides adapted from: Ross Girshick, ICCV 2015

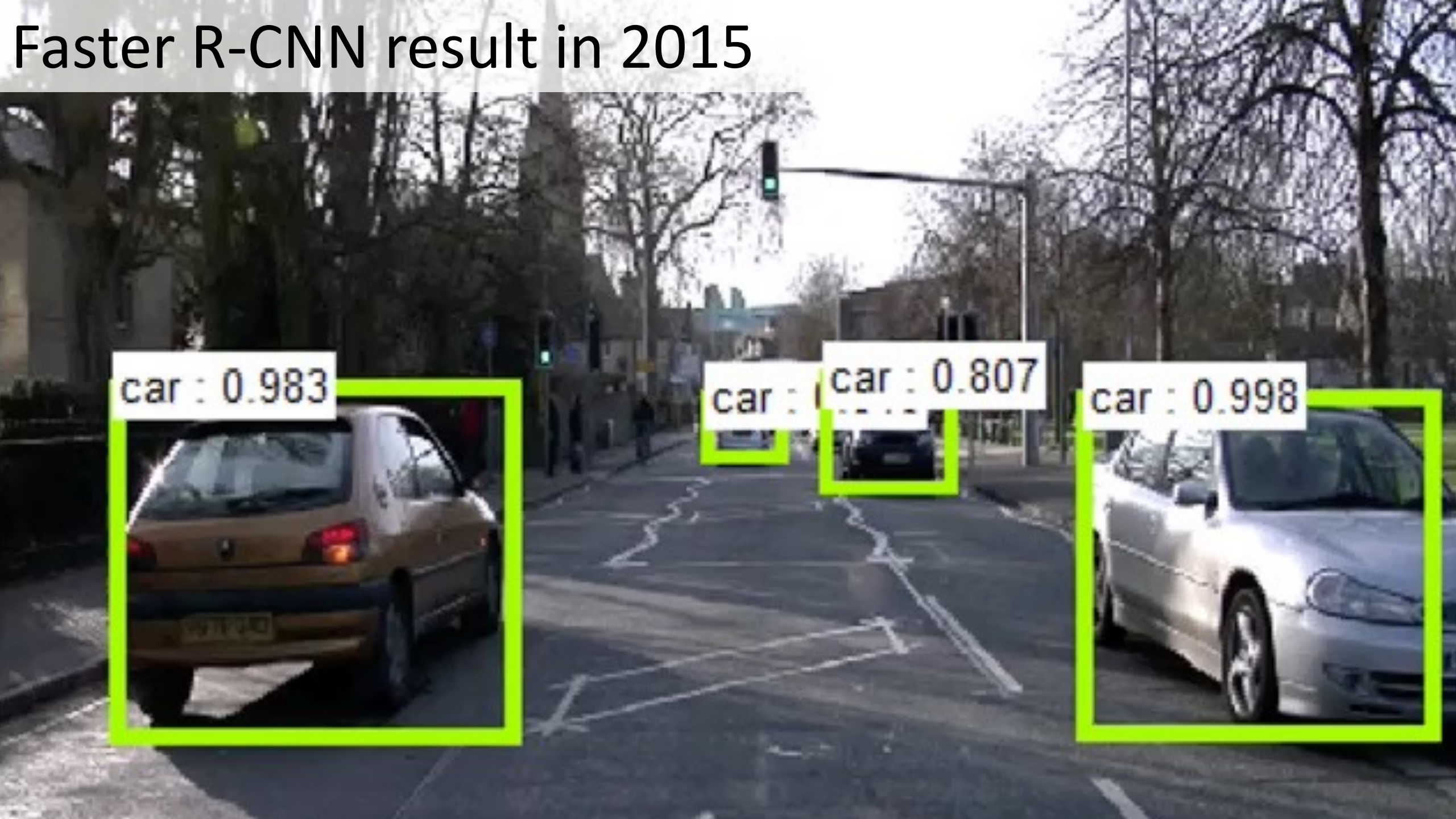
Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

Faster R-CNN result in 2015



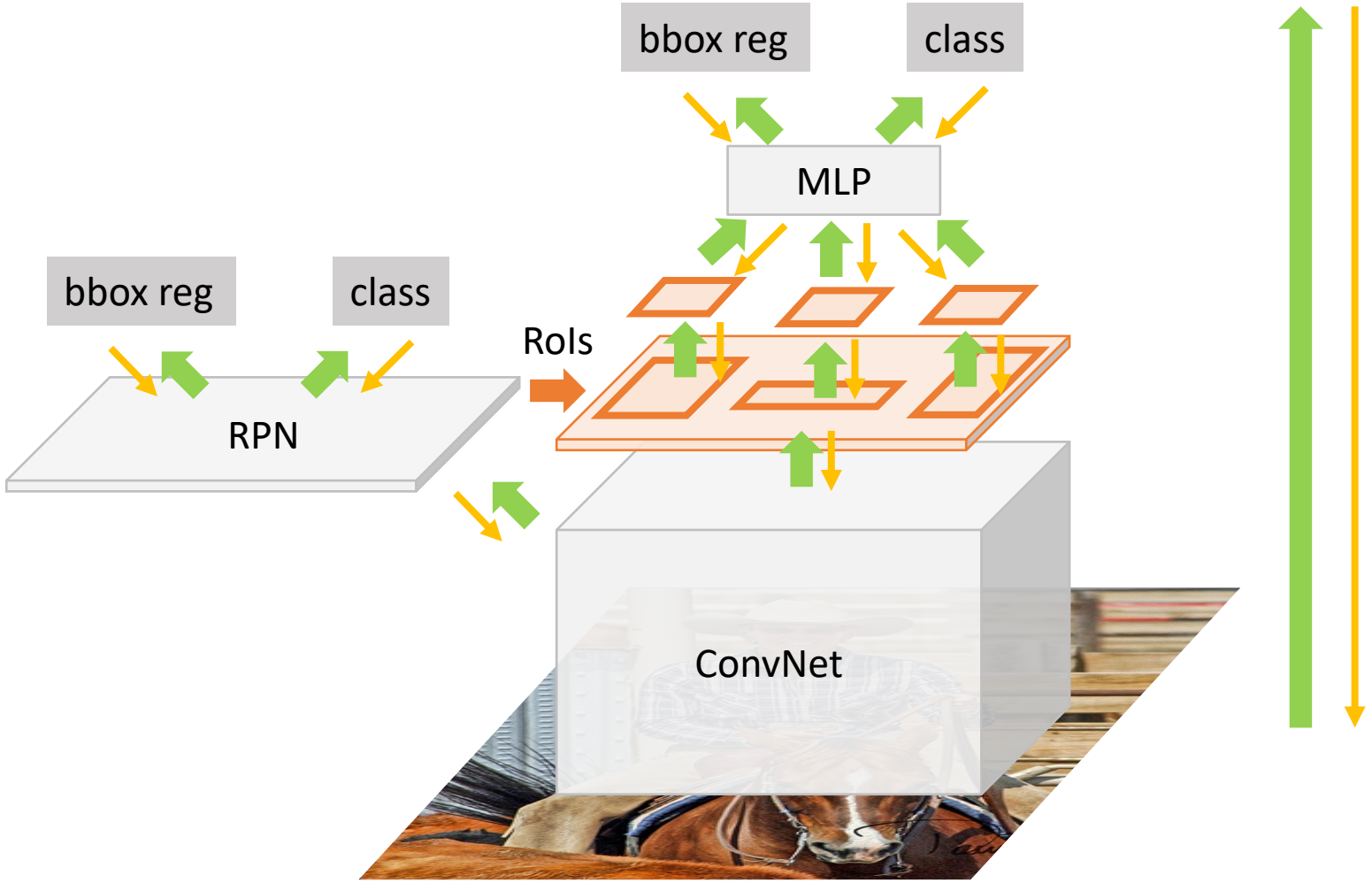
Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NeurIPS 2015

Faster R-CNN result in 2015



Faster R-CNN: End-to-End Object Detection

- **What's still wrong?** Single-scale feature maps for multi-scale objects



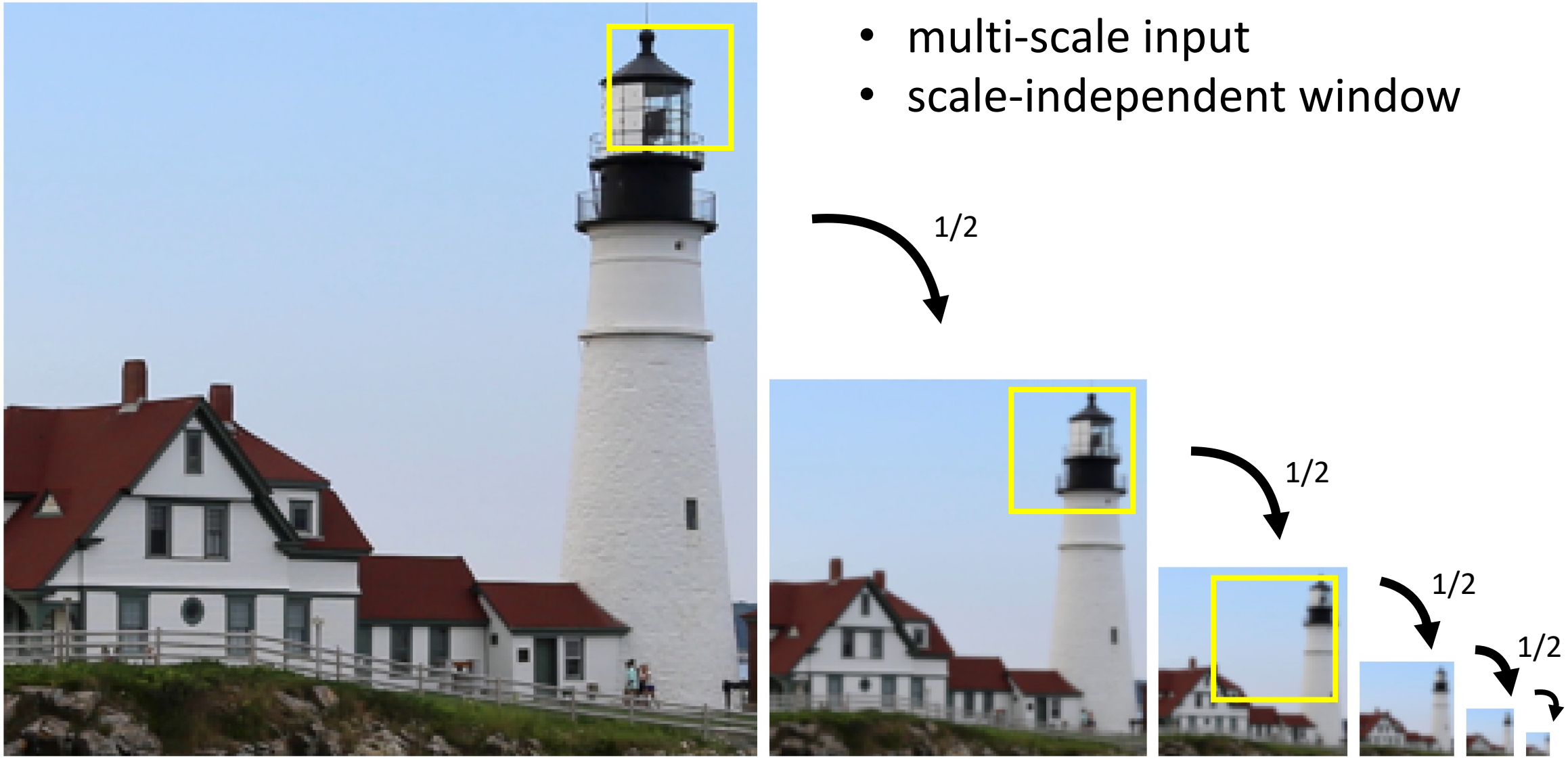
Slides adapted from: Ross Girshick, ICCV 2015

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NeurIPS 2015

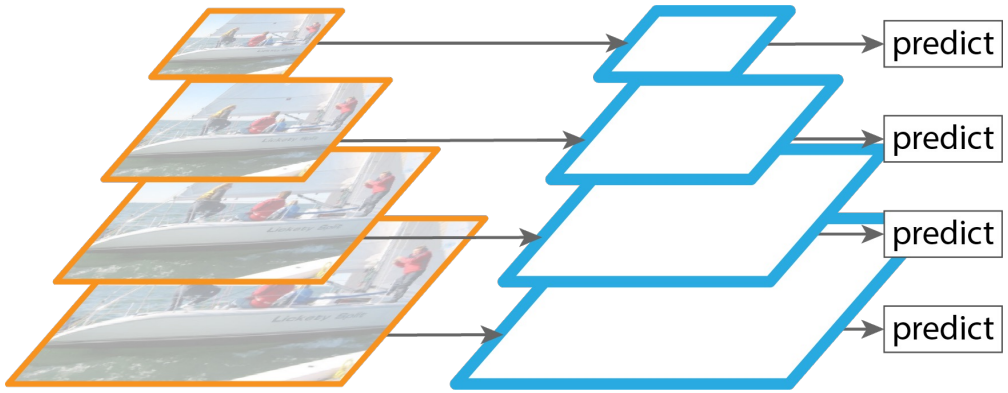
Image Pyramid

Scale-invariance

- multi-scale input
- scale-independent window



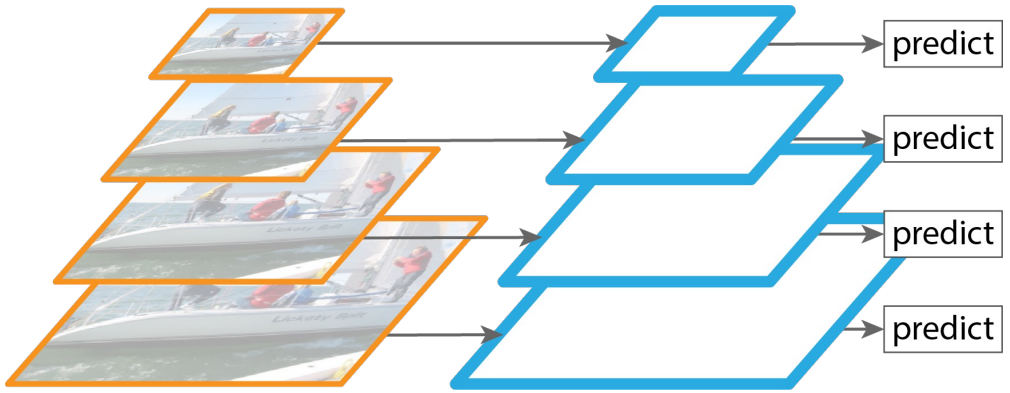
Pyramid and Neural Networks



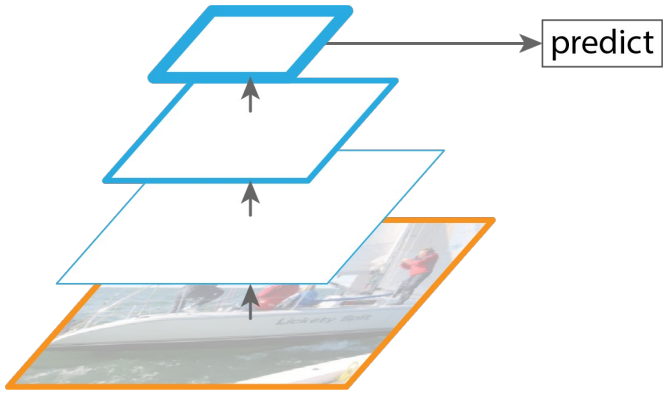
(a) image pyramid

- Apply the same model (e.g., Faster R-CNN) to multi-scale input images

Pyramid and Neural Networks



(a) image pyramid

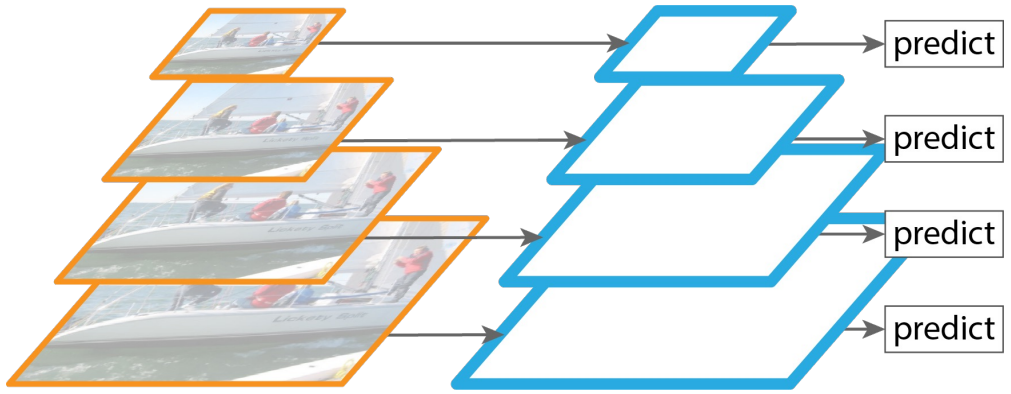


(b) single feature map

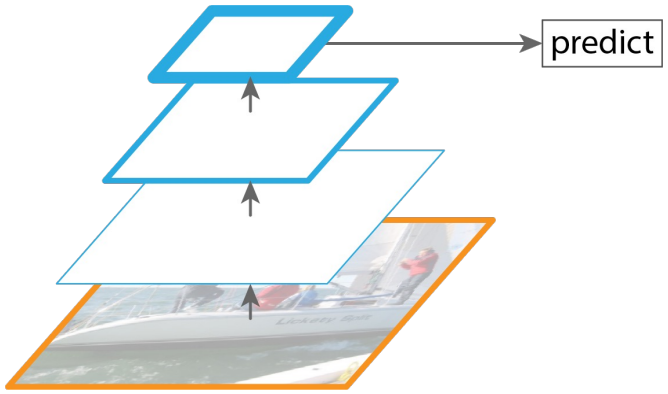
- Apply the model to the single, last (lowest-resolution) feature map

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature Pyramid Networks for Object Detection”, CVPR 2017

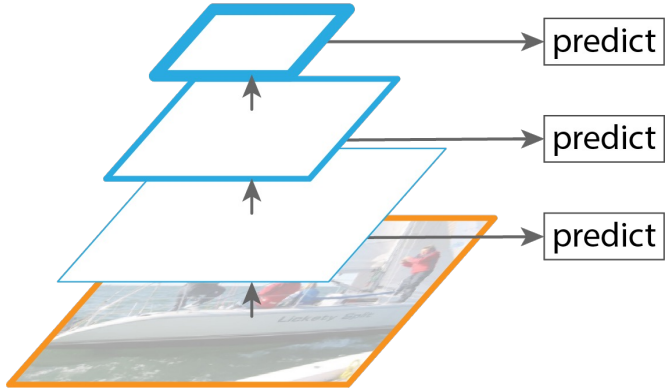
Pyramid and Neural Networks



(a) image pyramid



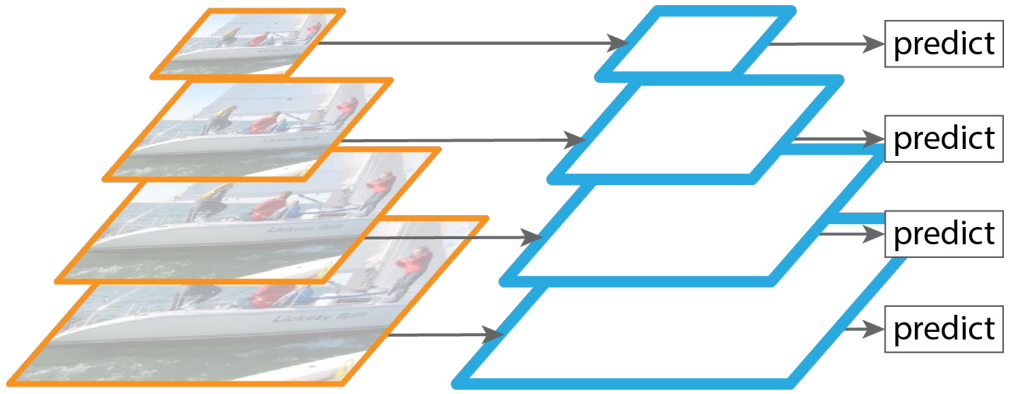
(b) single feature map



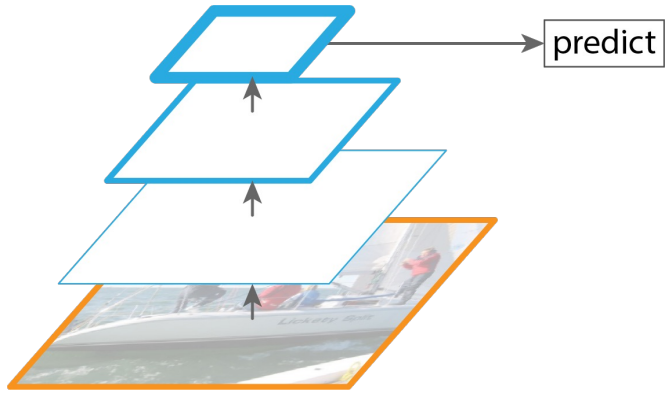
(c) pyramid feature hierarchy

- Reuse the “feature hierarchy” of a ConvNet as a pyramid (e.g., 1/2×, 1/4×, 1/8×)

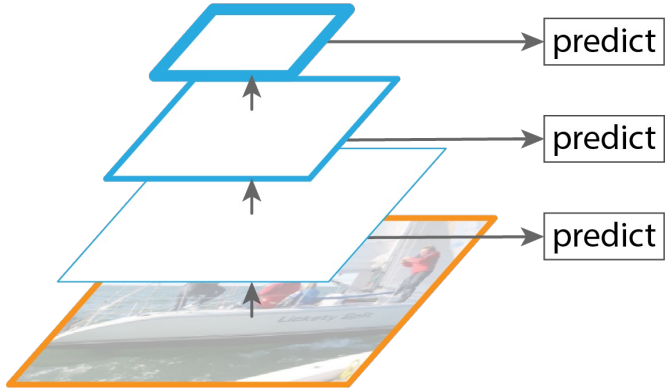
Pyramid and Neural Networks



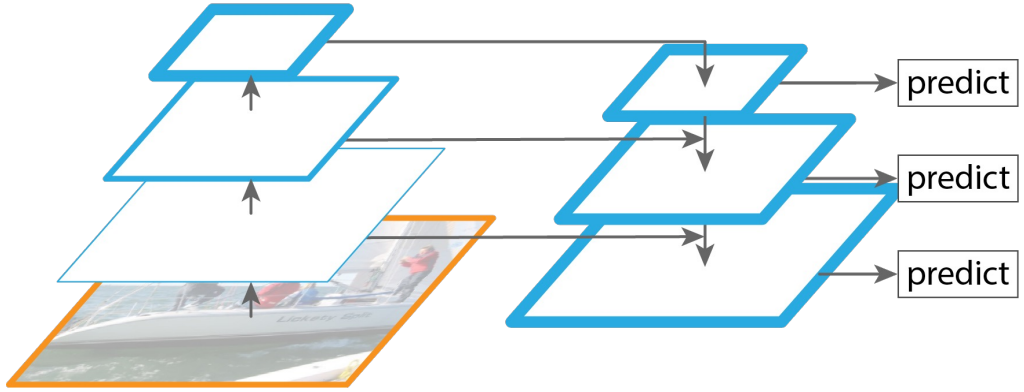
(a) image pyramid



(b) single feature map



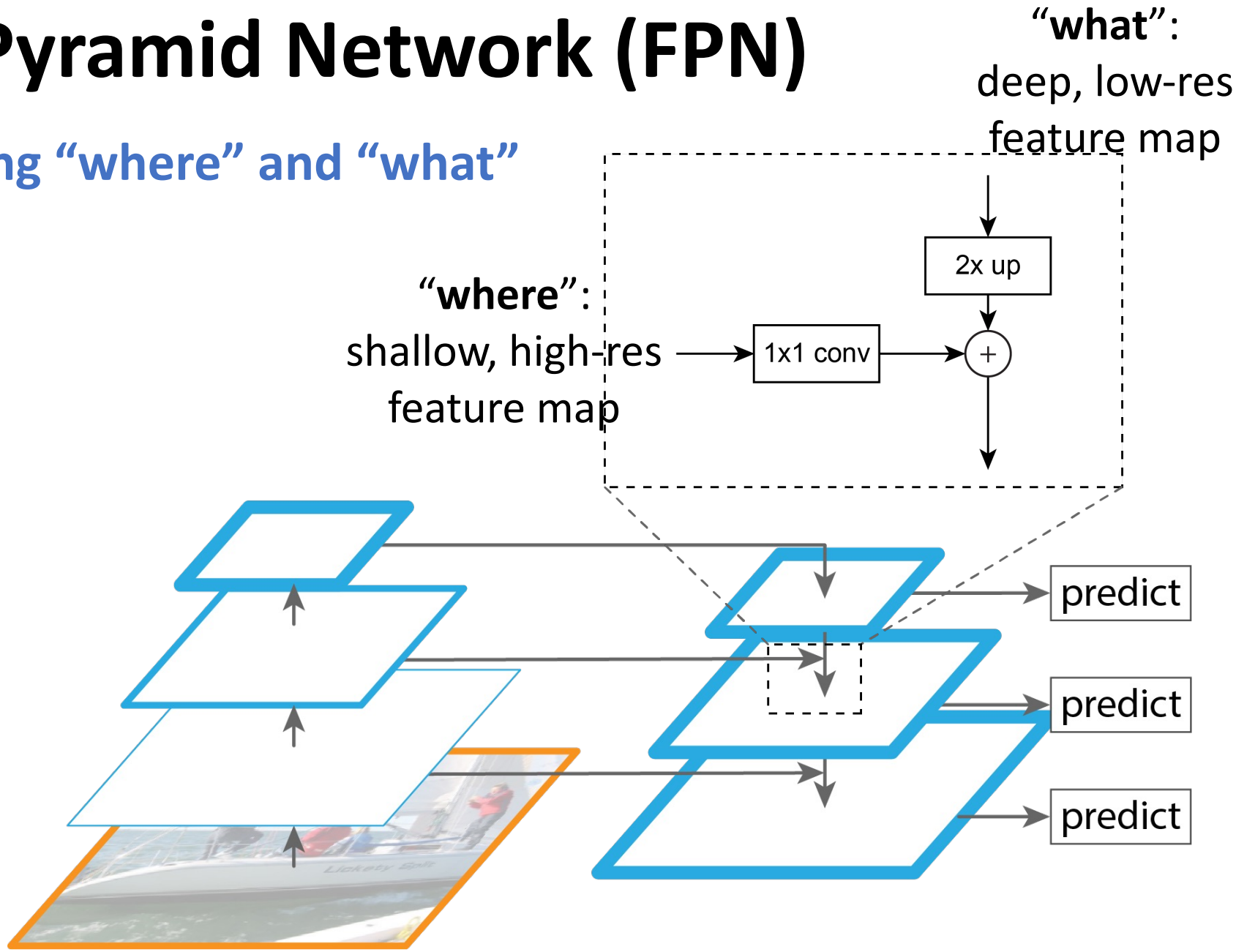
(c) pyramid feature hierarchy



(d) **Feature Pyramid Network**

Feature Pyramid Network (FPN)

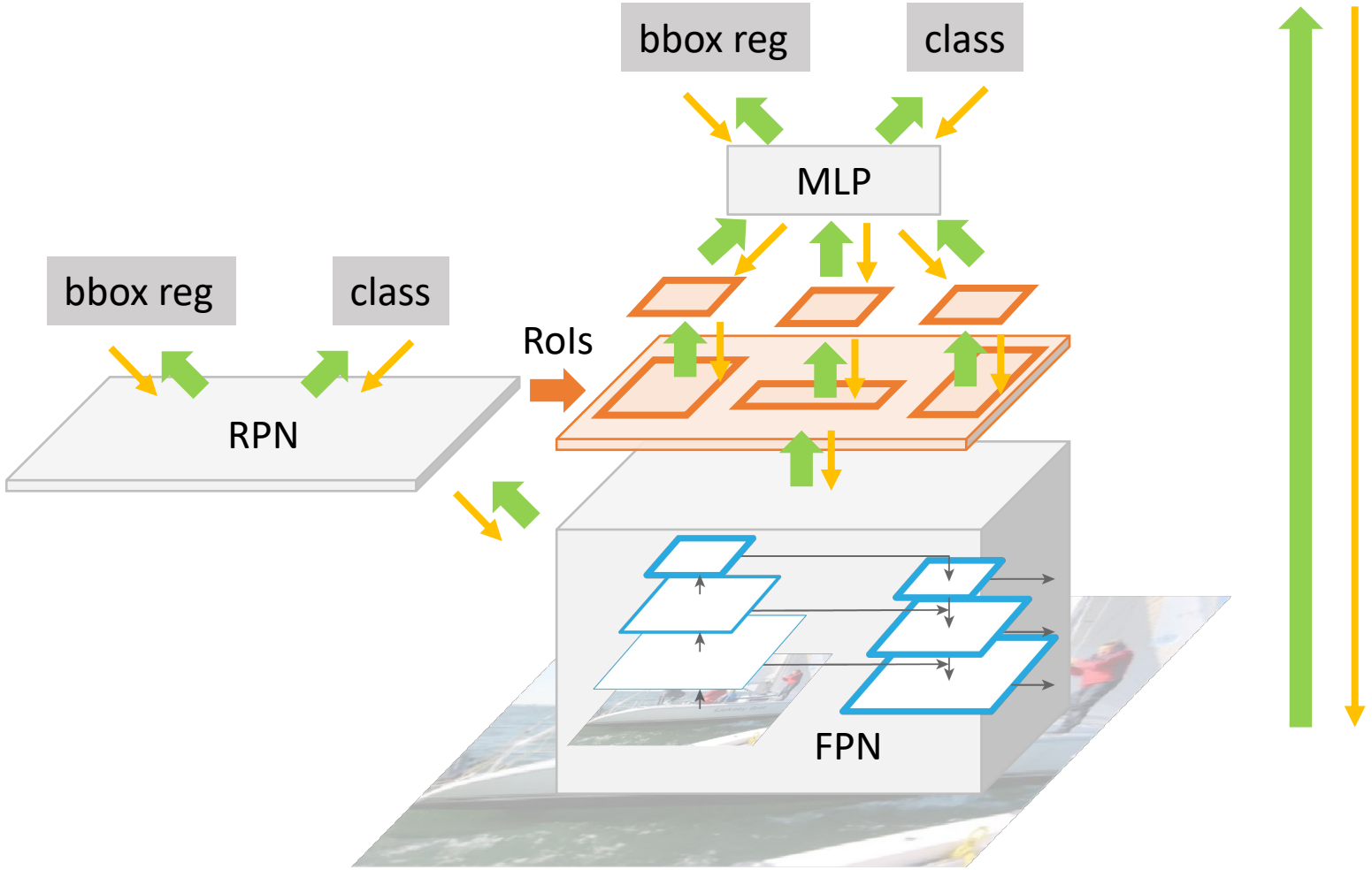
- **Combining "where" and "what"**



Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature Pyramid Networks for Object Detection", CVPR 2017

Feature Pyramid Network for Faster R-CNN

- Apply RPN and RoIs to Feature Pyramid

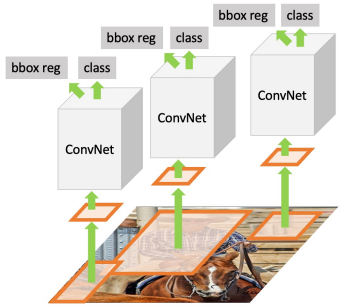


Slides adapted from: Ross Girshick, ICCV 2015

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature Pyramid Networks for Object Detection", CVPR 2017

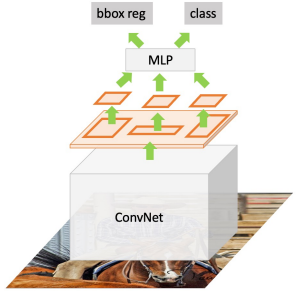
Evolution of R-CNN

R-CNN



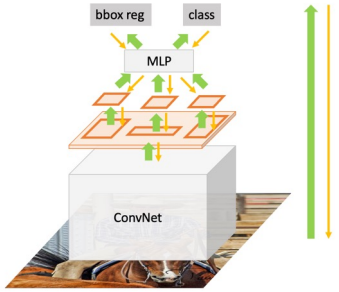
DL for detection made work!

Fast R-CNN "v0.1" (SPPnet)



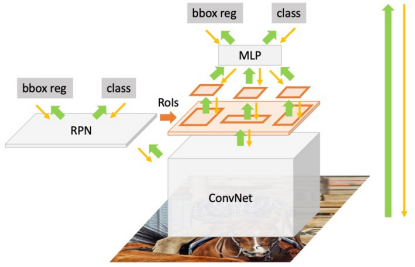
Fully convolutional feature maps

Fast R-CNN



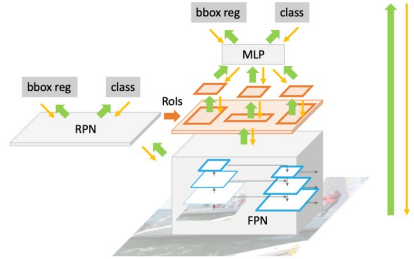
Differentiable Programming

Faster R-CNN



RPN & End-to-End detection

Faster R-CNN w/ FPN

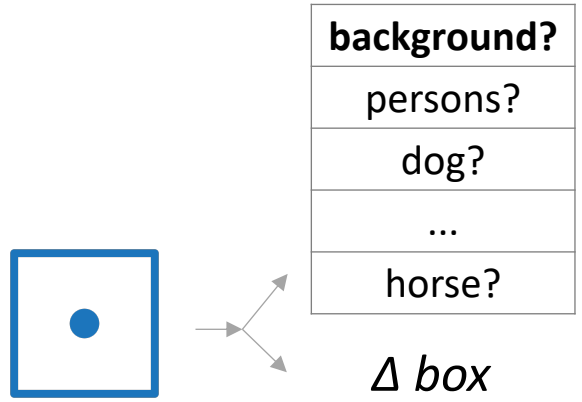


Feature Pyramid

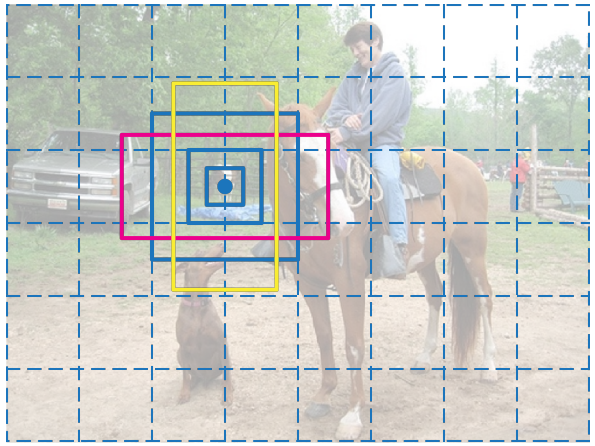
Single-Stage Object Detection

What could still be wrong? Do we need RoI operations?

- "Two-Stage" Detectors: RPN + RoI nets
- "Single-Stage" Detectors
  - YOLO [Redmon+], SSD [Liu+], RetinaNet [Lin+]



- Analogous to "**class-aware** RPN"
- What are still needed?
  - Anchors
  - Feature Pyramid Networks



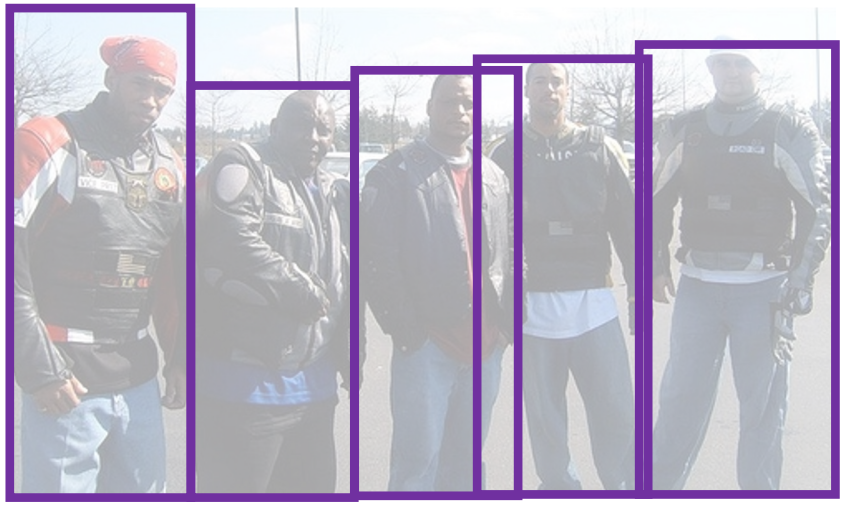
Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You only look once: Unified, real-time object detection", CVPR 2016

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

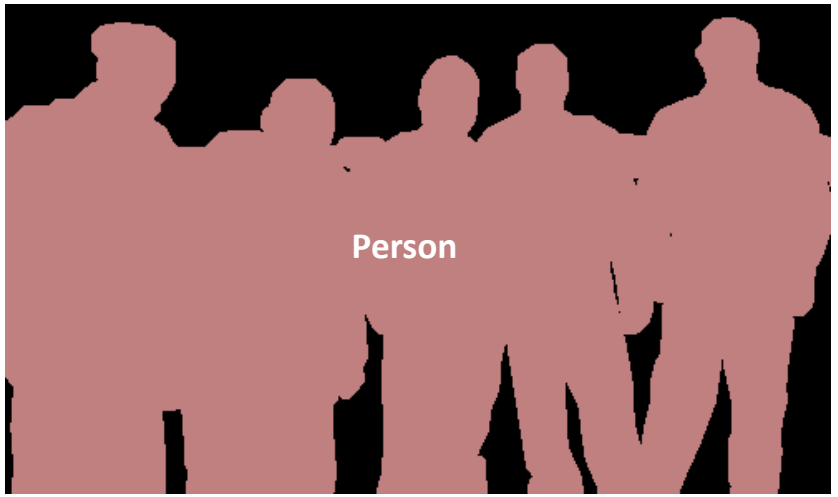
Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, "Focal Loss for Dense Object Detection", ICCV 2017

Instance Segmentation

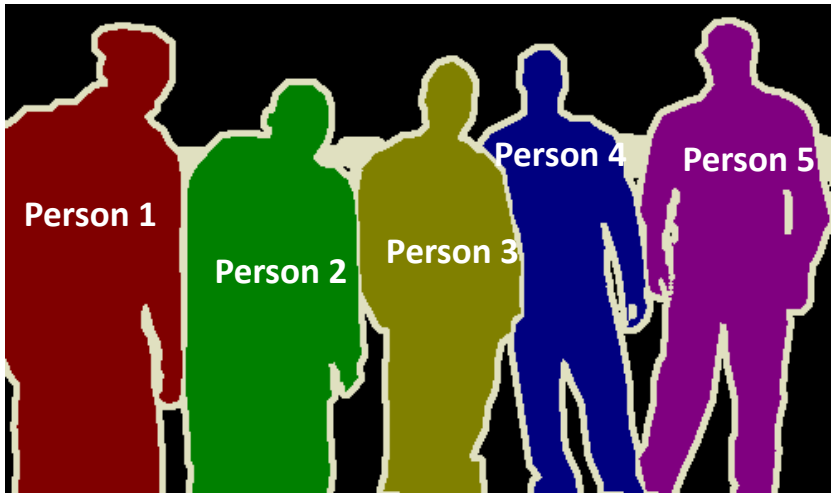
Instance Segmentation: Problem Definition



Object Detection



Semantic Segmentation

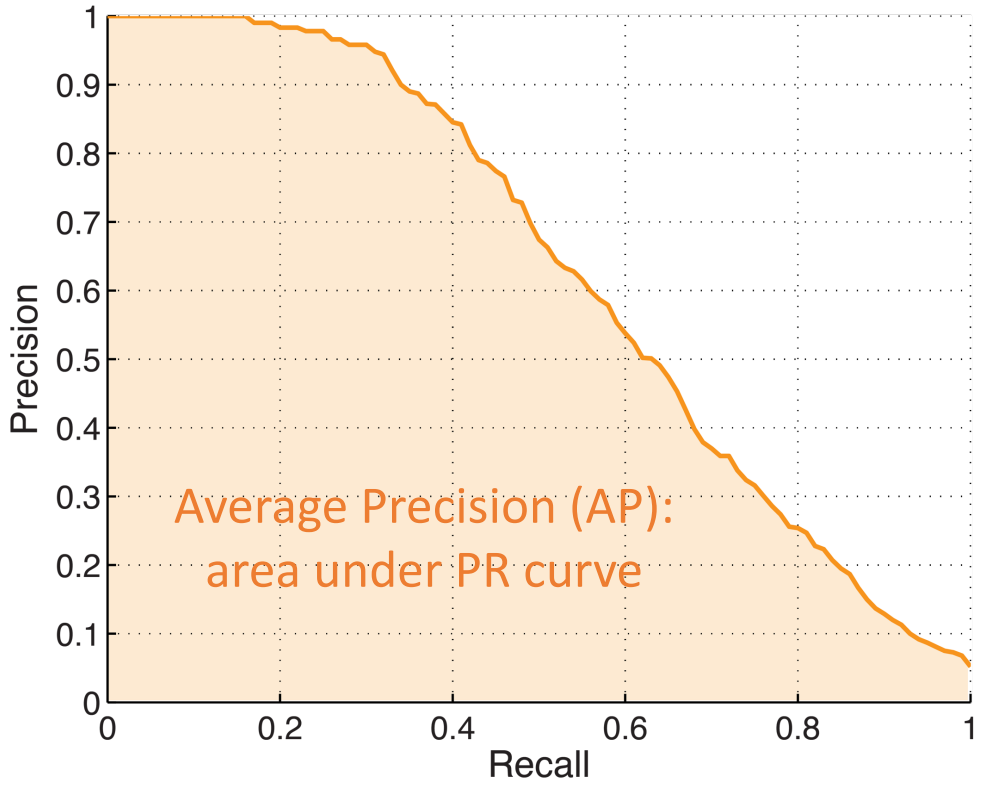


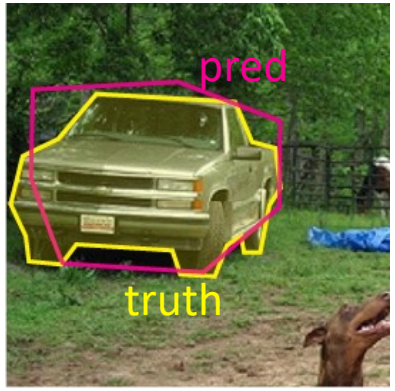
Instance Segmentation

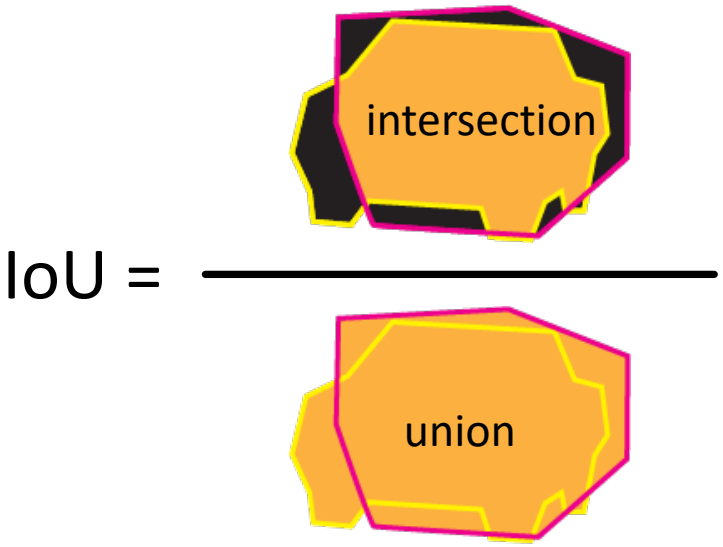
(Detect and segment each object)

Instance Segmentation: Evaluation Metric

- Similar to Object Detection: Average Precision (AP)
- But localization is determined by mask IoU



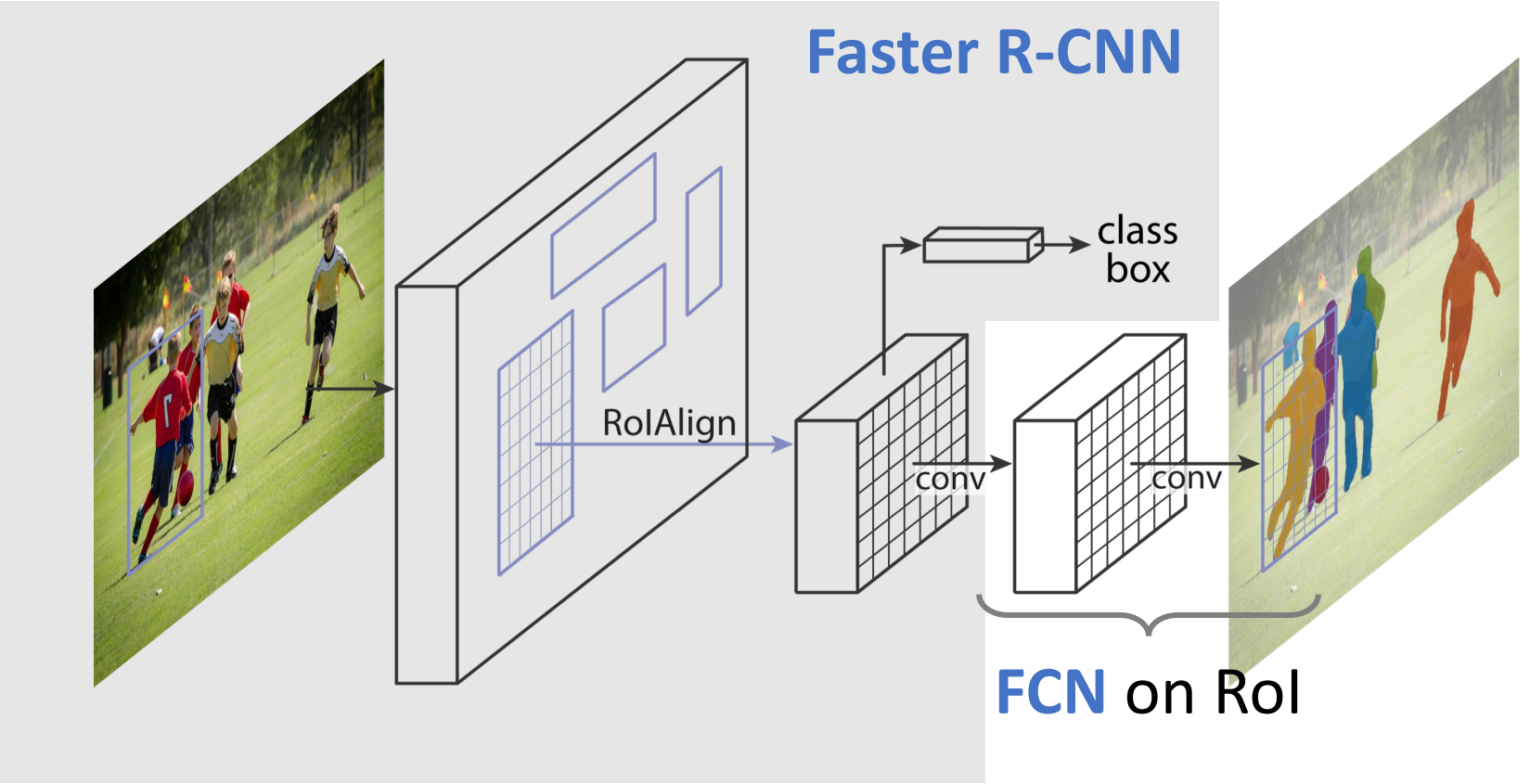




correct localization:
mask IoU > threshold

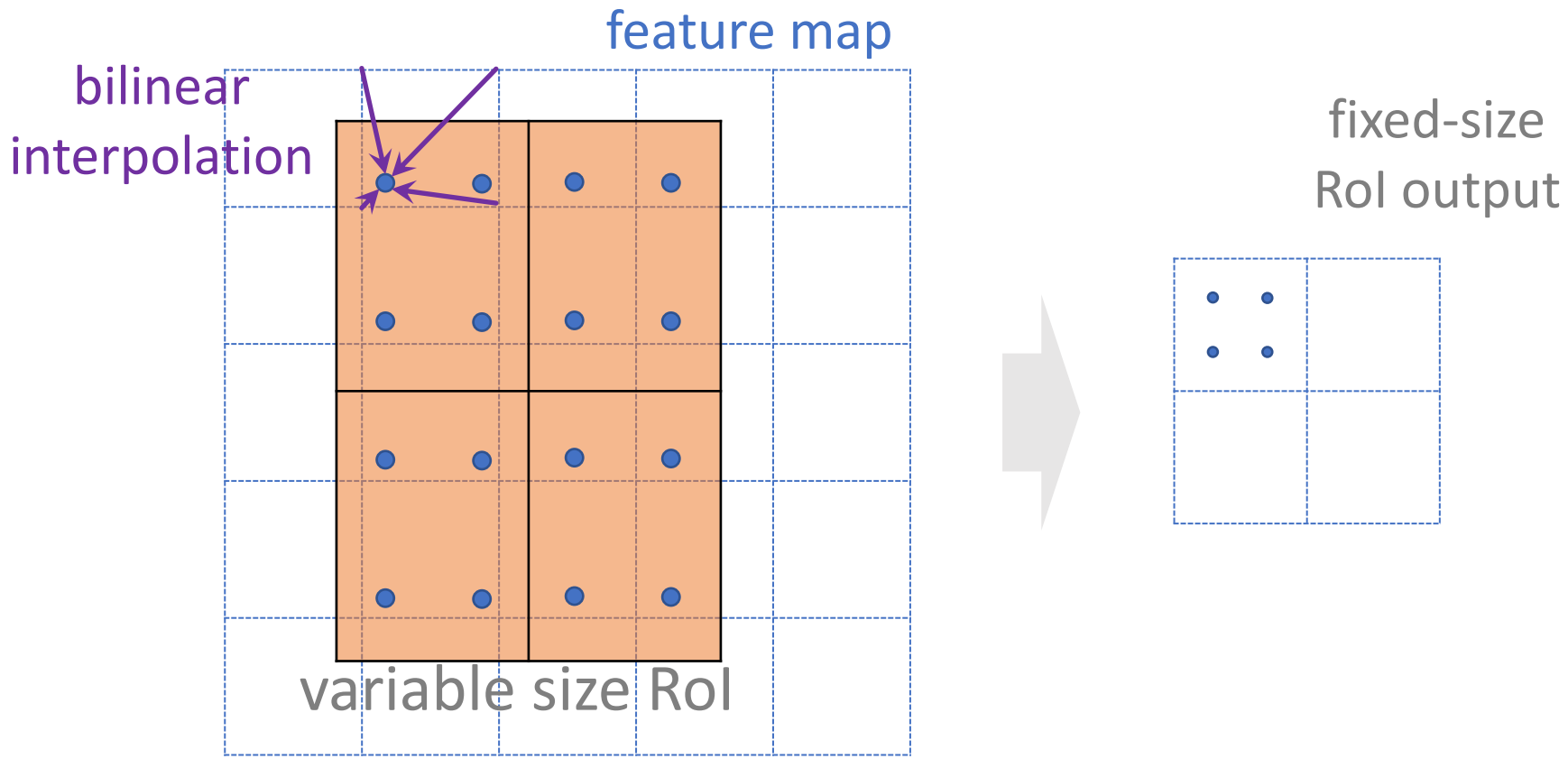
Mask R-CNN

- **Difficult problem, simple solution: Faster R-CNN + FCN**



Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, "Mask R-CNN", ICCV 2017

Mask R-CNN: RoI Align

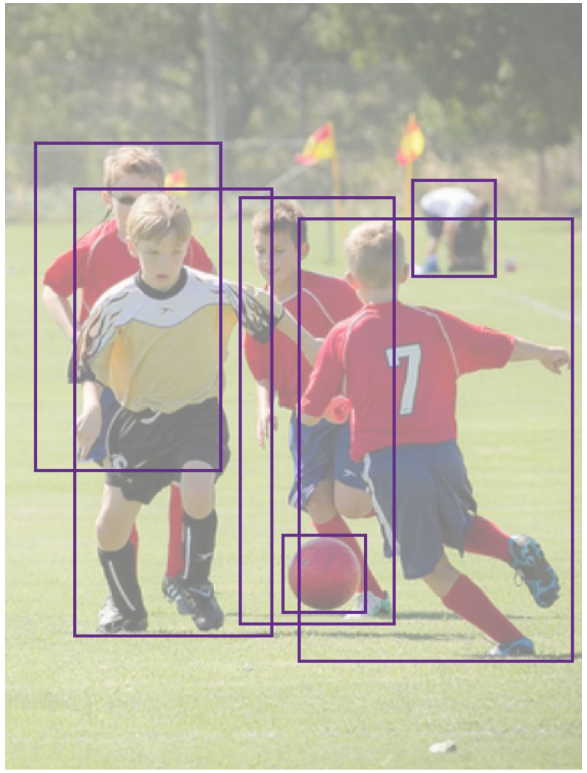


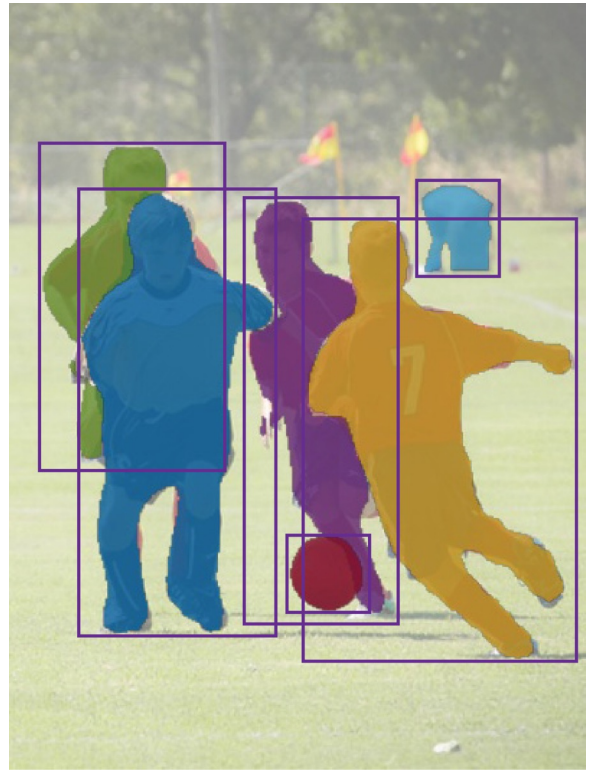
- **To apply FCN on each region, we need "aligned" features**
- Manipulate feature maps just like images: crop and interpolate

Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, "Mask R-CNN", ICCV 2017

Mask R-CNN

- **Difficult problem, simple solution: Faster R-CNN + FCN**





Detect individual objects:
Faster R-CNN

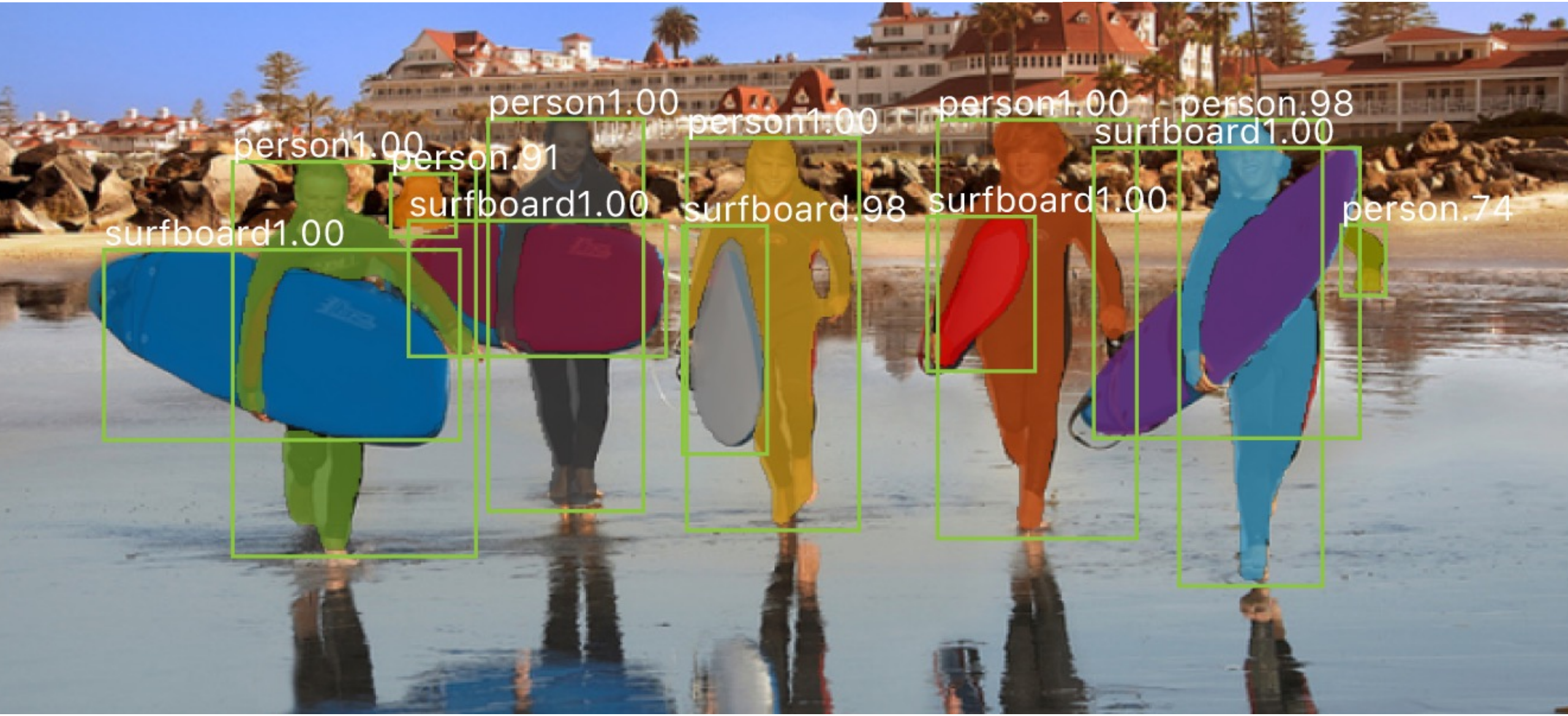
For each object, extra an aligned RoI feature and apply a tiny **FCN**

Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, “Mask R-CNN”, ICCV 2017

Mask R-CNN result in 2017

Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, “Mask R-CNN”, ICCV 2017

Mask R-CNN result in 2017



Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, “Mask R-CNN”, ICCV 2017

TEXTS IN COMPUTER SCIENCE

Computer Vision

Algorithms and Applications
Second Edition

Richard Szeliski

Springer

Lecture 13: Object Detection and Image Segmentation

- Visual Recognition Tasks
- Methodology: ConvNets, Fully Conv Nets
- Semantic Segmentation
- Object Detection
- Instance Segmentation