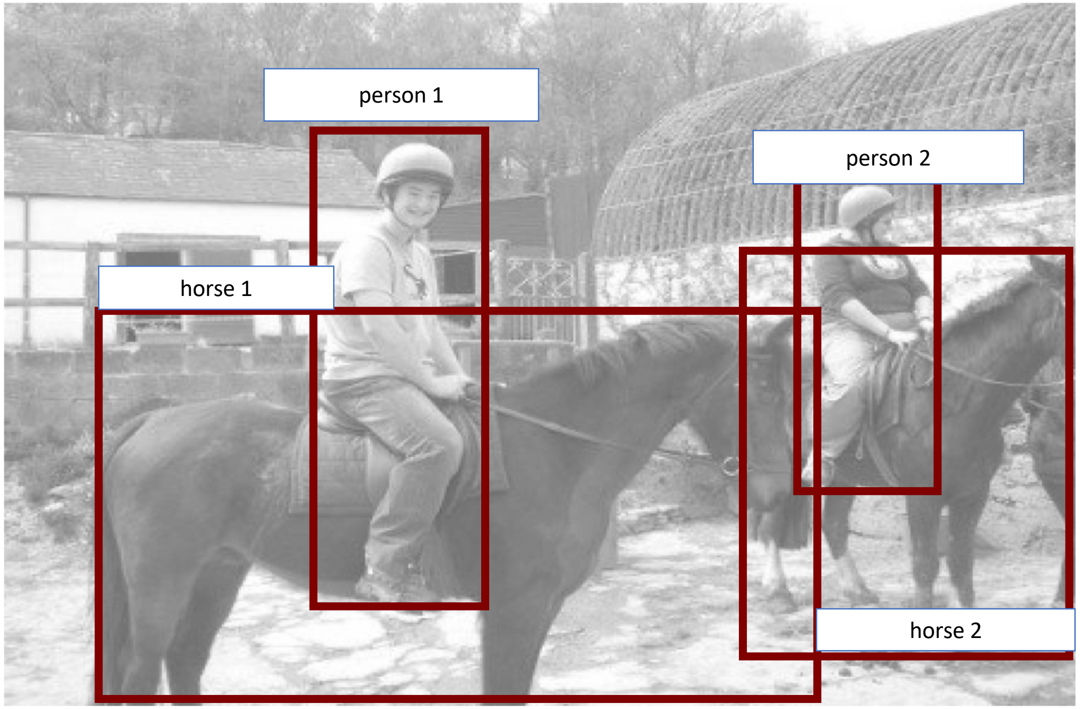
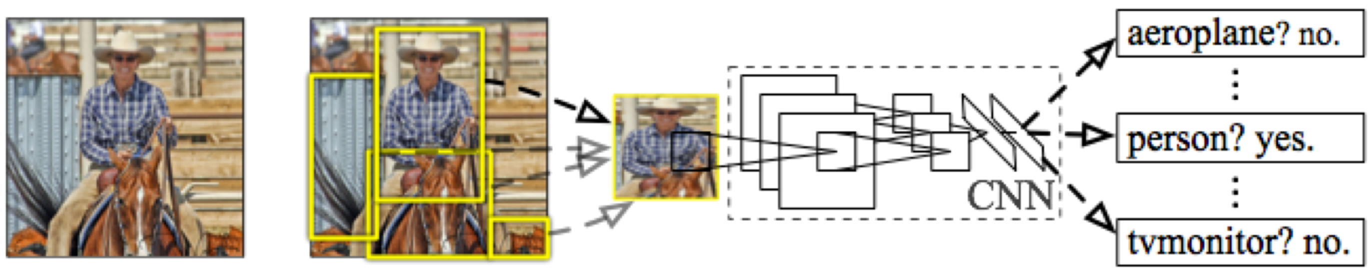
Object detection

The Task



R-CNN: Regions with CNN features



Input image

Extract region proposals (~2k / image)

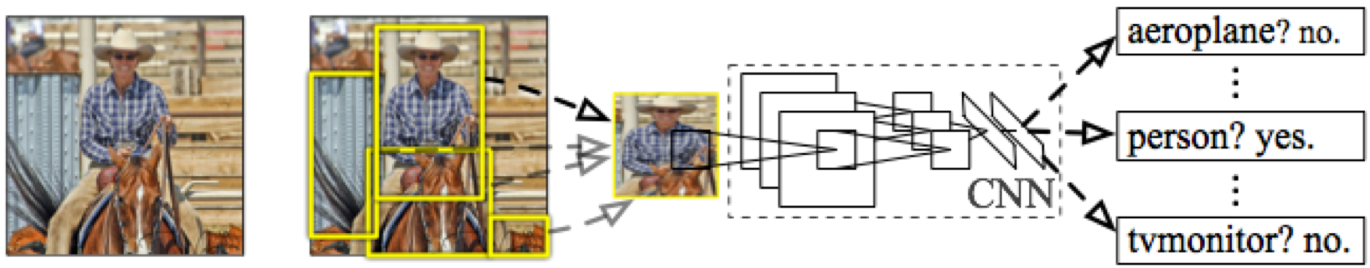
Compute CNN features

Classify regions (linear SVM)

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

Slide credit : Ross Girshick

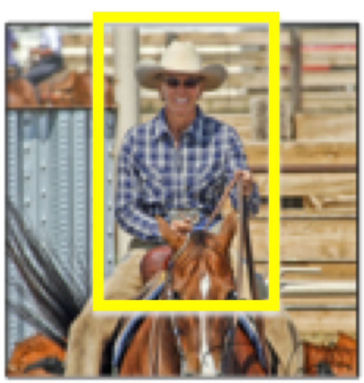
R-CNN at test time: Step 2



Input image

Extract region proposals (~2k / image)

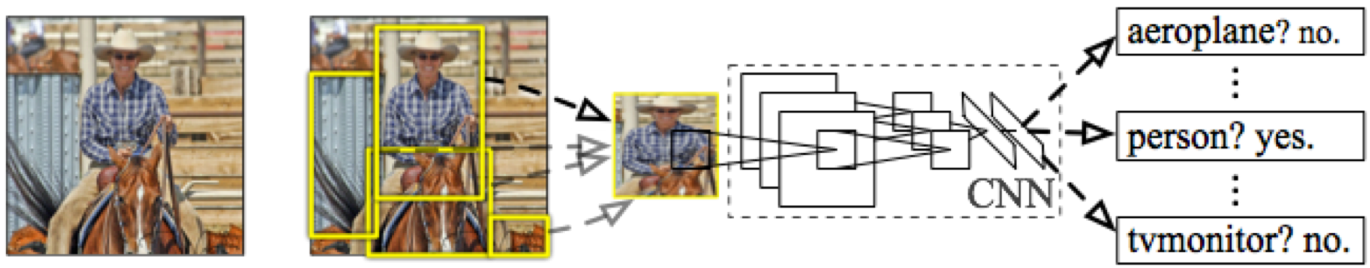
Compute CNN features



a. Crop

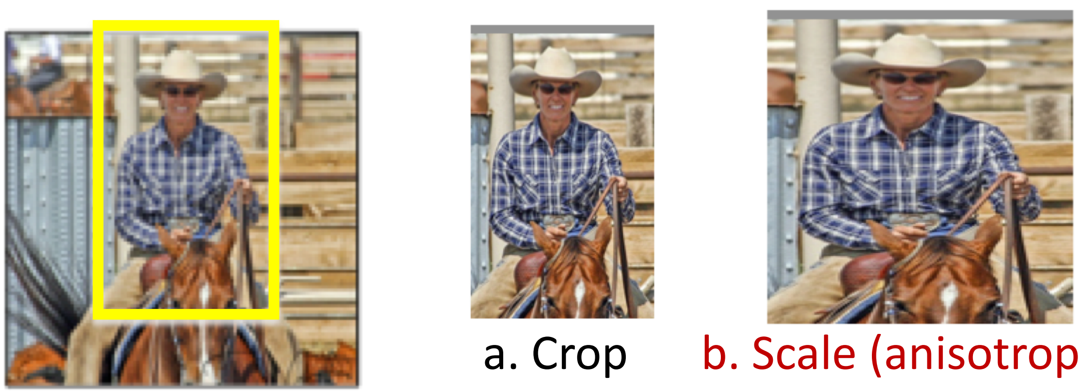
Slide credit : Ross Girshick

R-CNN at test time: Step 2



Input image

Extract region proposals (~2k / image) → Compute CNN features



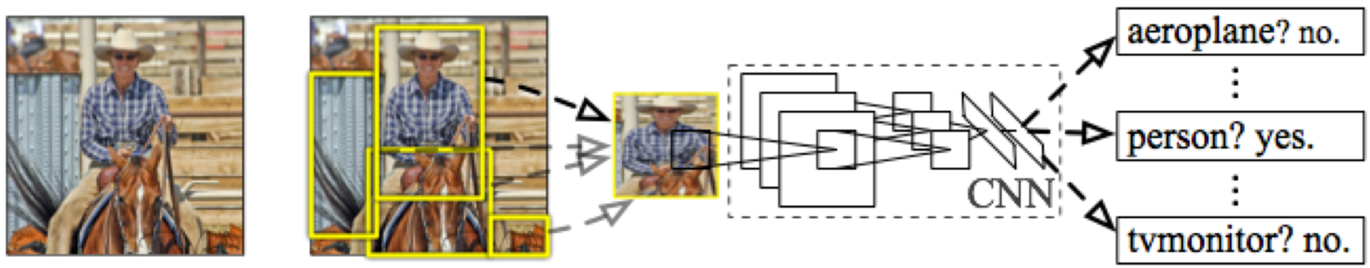
227 x 227

a. Crop

b. Scale (anisotropic)

Slide credit : Ross Girshick

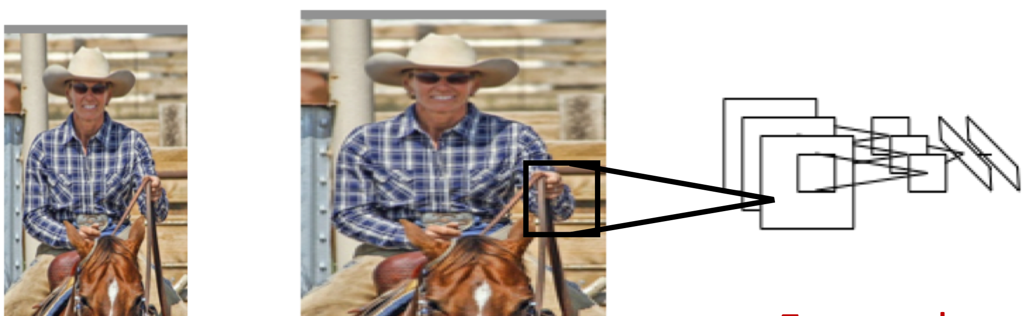
R-CNN at test time: Step 2



Input image

Extract region proposals (~2k / image)

→ Compute CNN features



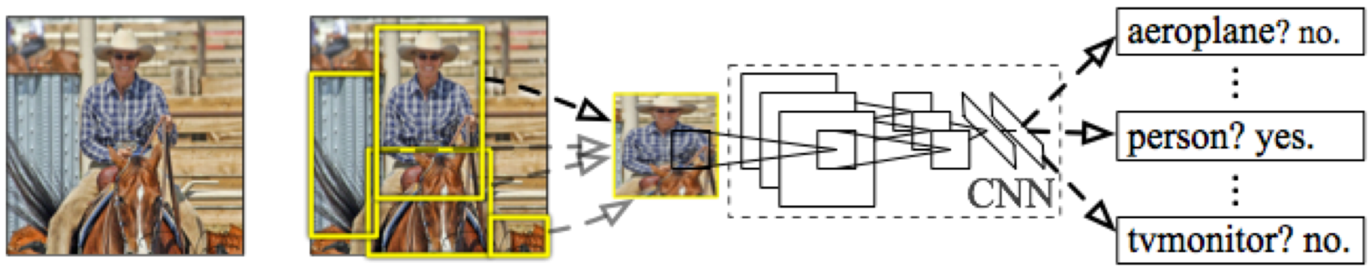
1. Crop

b. Scale (anisotropic)

c. Forward propagate
Output: “fc_7” features

Slide credit : Ross Girshick

R-CNN at test time: Step 3

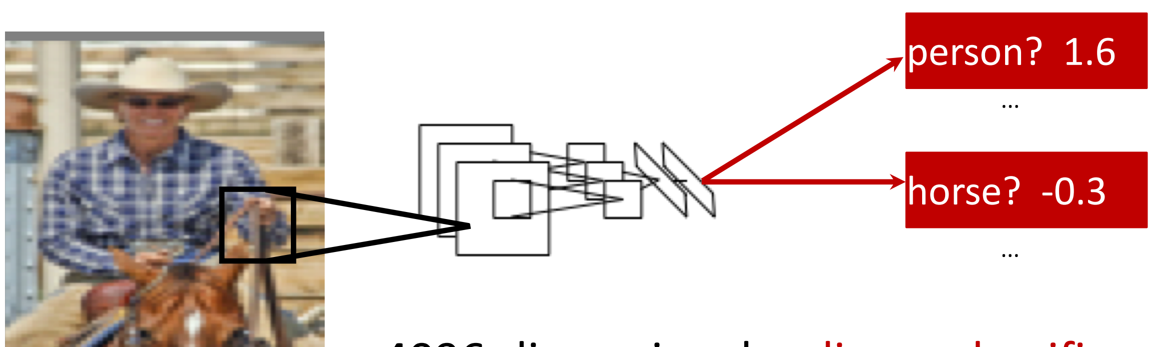


Input image

Extract region proposals (~2k / image)

Compute CNN features

Classify regions



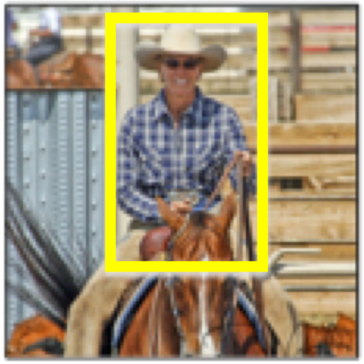
Warped proposal

4096-dimensional fc_7 feature vector

linear classifiers (SVM or softmax)

Slide credit : Ross Girshick

Step 4: Object proposal refinement



Linear regression

on CNN features

Original
proposal

Predicted
object bounding box

Bounding-box regression

Slide credit : Ross Girshick

R-CNN results on PASCAL

	VOC 2007	VOC 2010
DPM v5 (Girshick et al. 2011)	33.7%	29.6%
UVA sel. search (Uijlings et al. 2013)		35.1%
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)		40.4%

Reference systems

metric: mean average precision (higher is better)

Slide credit : Ross Girshick

R-CNN results on PASCAL

	VOC 2007	VOC 2010
DPM v5 (Girshick et al. 2011)	33.7%	29.6%
UVA sel. search (Uijlings et al. 2013)		35.1%
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)		40.4%
R-CNN	54.2%	50.2%
R-CNN + bbox regression	58.5%	53.7%

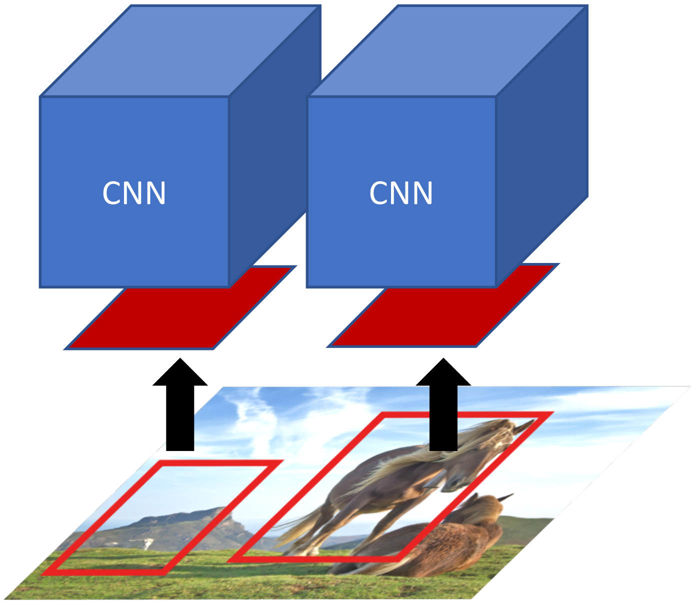
metric: mean average precision (higher is better)

Slide credit : Ross Girshick

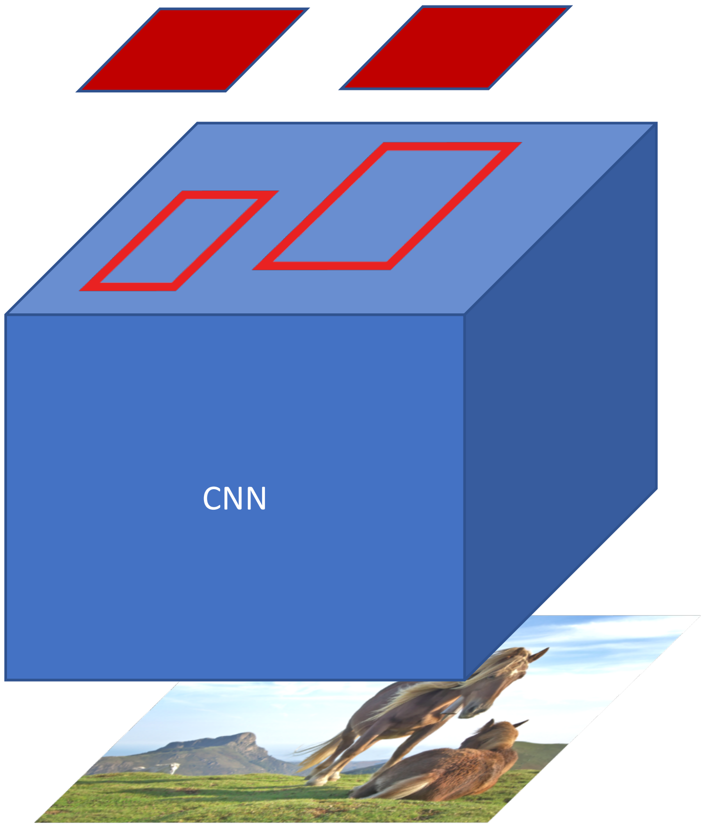
Training R-CNN

- Train convolutional network on ImageNet classification
- *Finetune* on detection
  - Classification problem!
  - Proposals with IoU > 50% are positives
  - Sample fixed proportion of positives in each batch because of imbalance

Speeding up R-CNN

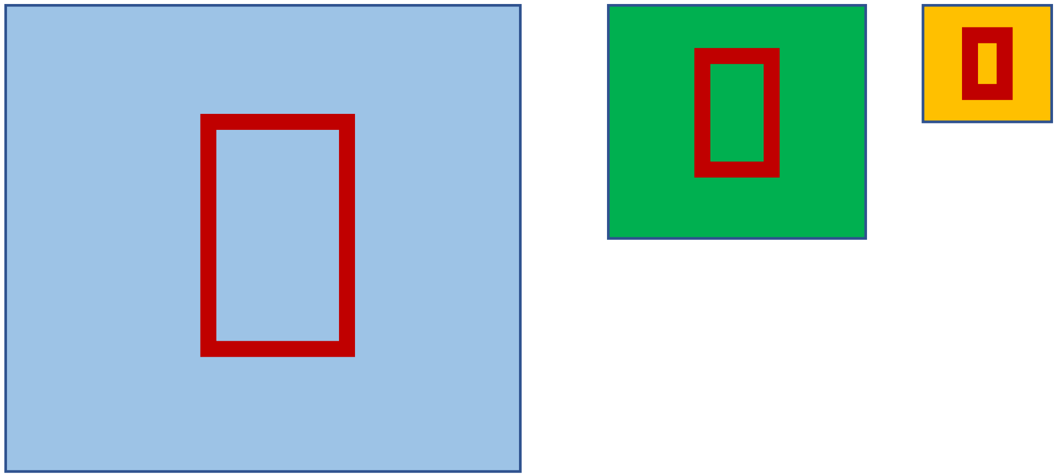


Speeding up R-CNN



ROI Pooling

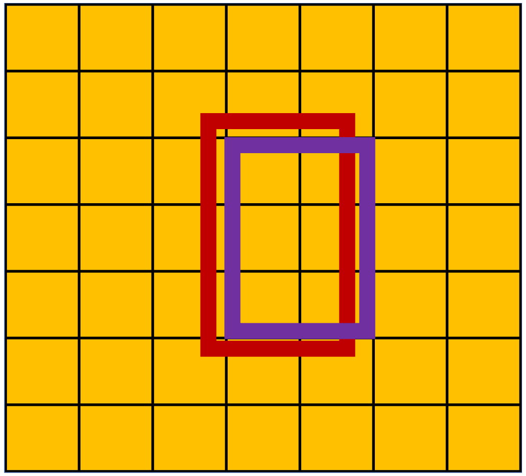
- How do we crop from a feature map?
- Step 1: Resize boxes to account for subsampling



Fast R-CNN. Ross Girshick. In ICCV 2015

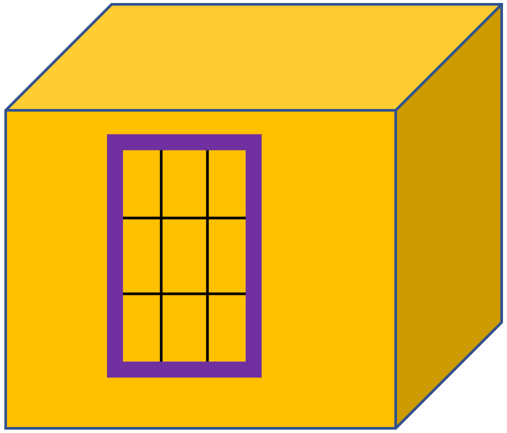
ROI Pooling

- How do we crop from a feature map?
- Step 2: Snap to feature map grid



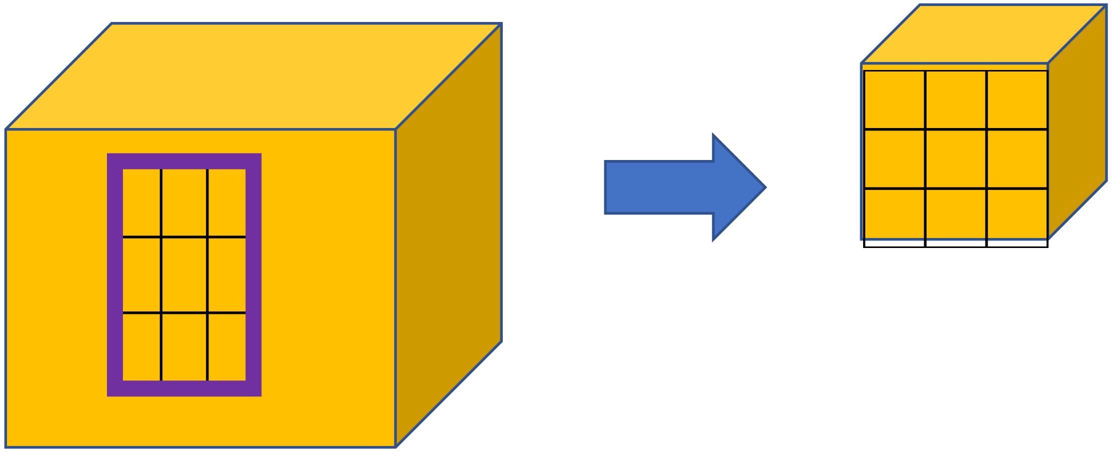
ROI Pooling

- How do we crop from a feature map?
- Step 3: Place a grid of fixed size



ROI Pooling

- How do we crop from a feature map?
- Step 4: Take max in each cell



Fast R-CNN

	Fast R-CNN	R-CNN
Train time (h)	9.5	84
Speedup	8.8x	1x
Test time / image	0.32s	47.0s
Speedup	146x	1x
mean AP	66.9	66.0

Fast R-CNN

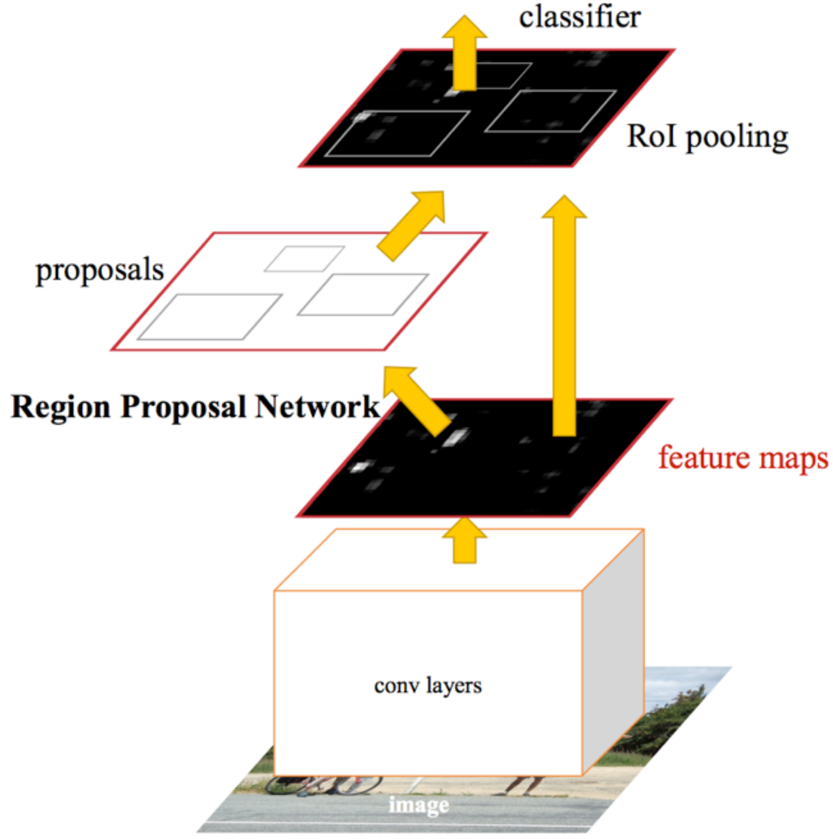
- Bottleneck remaining (not included in time):
  - Object proposal generation
- Slow
  - Requires segmentation
  - O(1s) per image

Fast*er* R-CNN

- Can we produce *object proposals* from convolutional networks?
- A change in intuition
  - Instead of using grouping
  - Recognize likely objects?
- For every possible box, score if it is likely to correspond to an object

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. S. Ren, K. He, R. Girshick, J. Sun. In *NIPS* 2015.

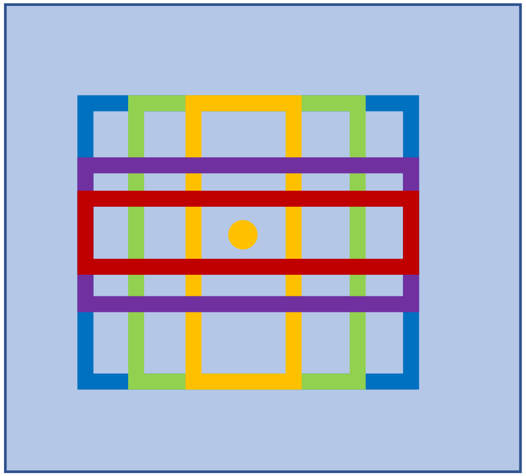
Faster R-CNN



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

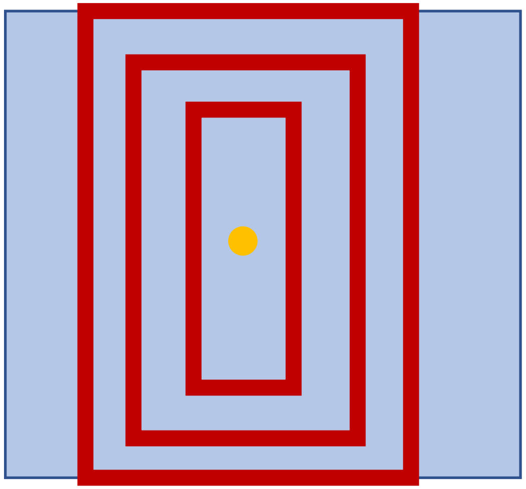
Faster R-CNN

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



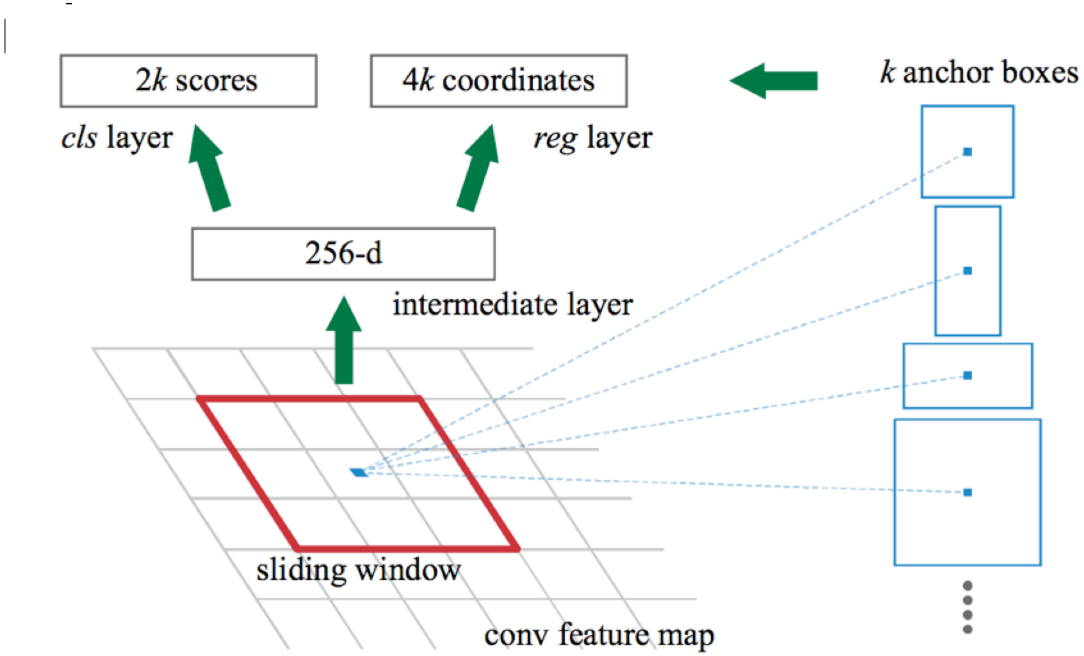
Faster R-CNN

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



Faster R-CNN

- At each location, consider boxes of many different sizes a



Faster R-CNN

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

Faster R-CNN

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

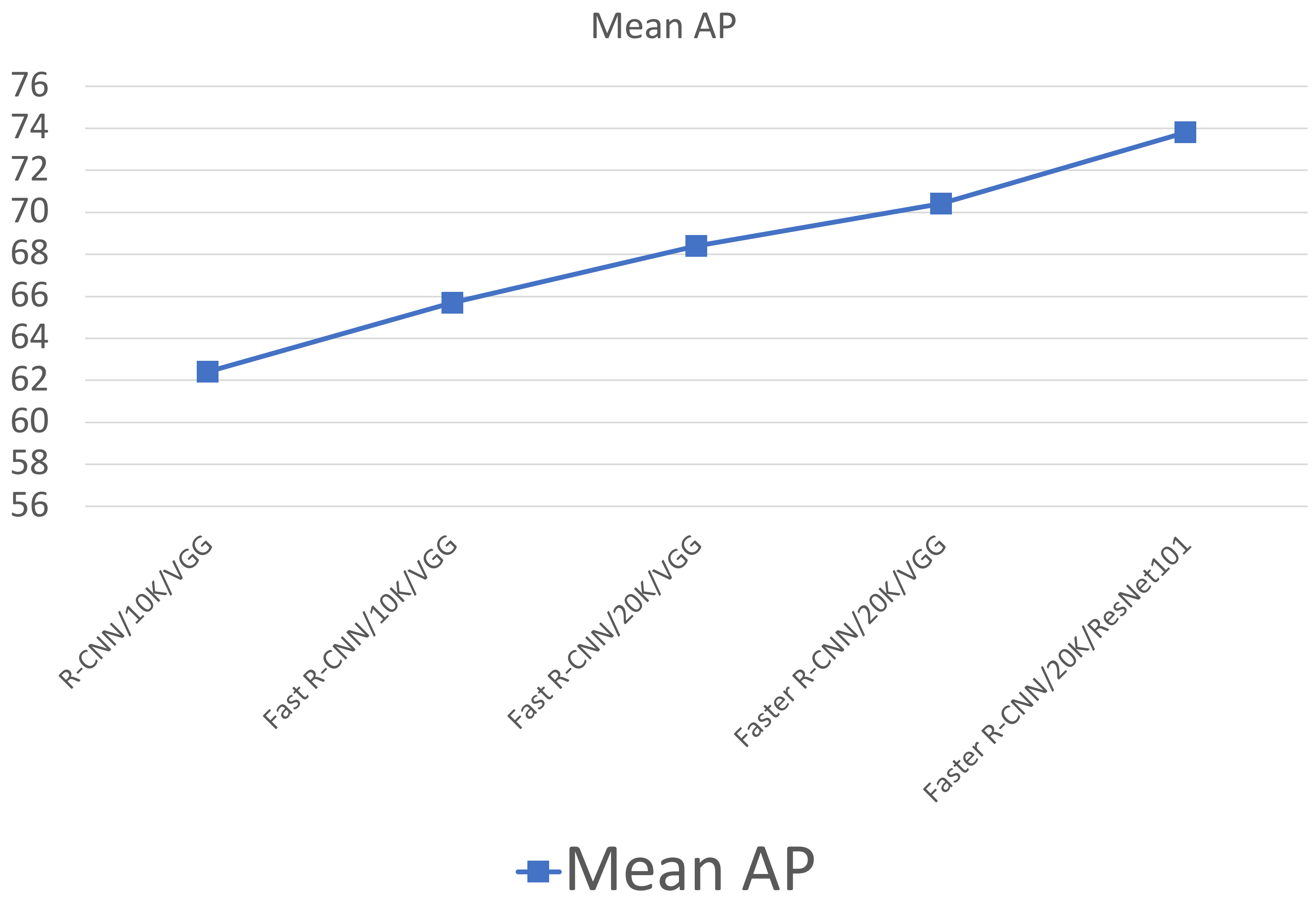
Impact of Feature Extractors

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

Impact of Additional Data

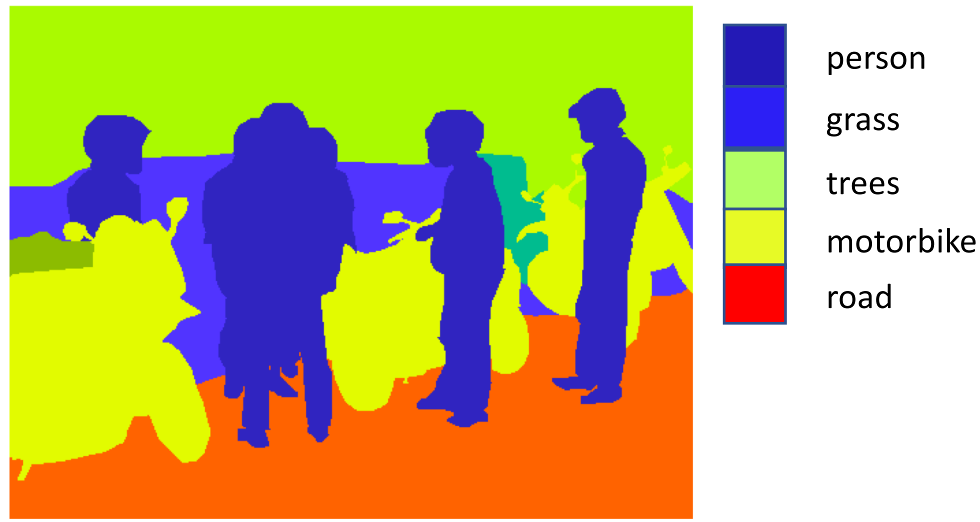
Method	Training data	mean AP (PASCAL VOC 2012 Test)
Fast R-CNN	VOC 12 Train (10K)	65.7
Fast R-CNN	VOC07 Trainval + VOC 12 Train	68.4
Faster R-CNN	VOC 12 Train (10K)	67.0
Faster R-CNN	VOC07 Trainval + VOC 12 Train	70.4

The R-CNN family of detectors



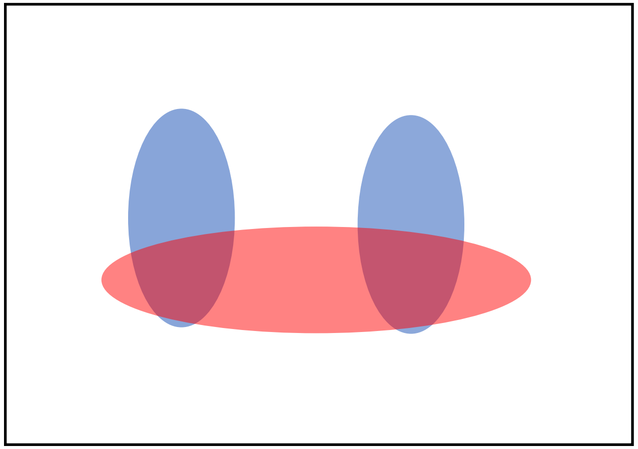
Semantic Segmentation

The Task



Evaluation metric

- Pixel classification!
- Accuracy?
  - Heavily unbalanced
  - Common classes are over-emphasized
- *Intersection over Union*
  - Average across classes and images
- Per-class accuracy
  - Compute accuracy for every class and then average



Things vs Stuff

THINGS
- Person, cat, horse, etc
- Constrained shape
- Individual instances with separate identity
- May need to look at objects

STUFF
- Road, grass, sky etc
- Amorphous, no shape
- No notion of instances
- Can be done at pixel level
- “texture”

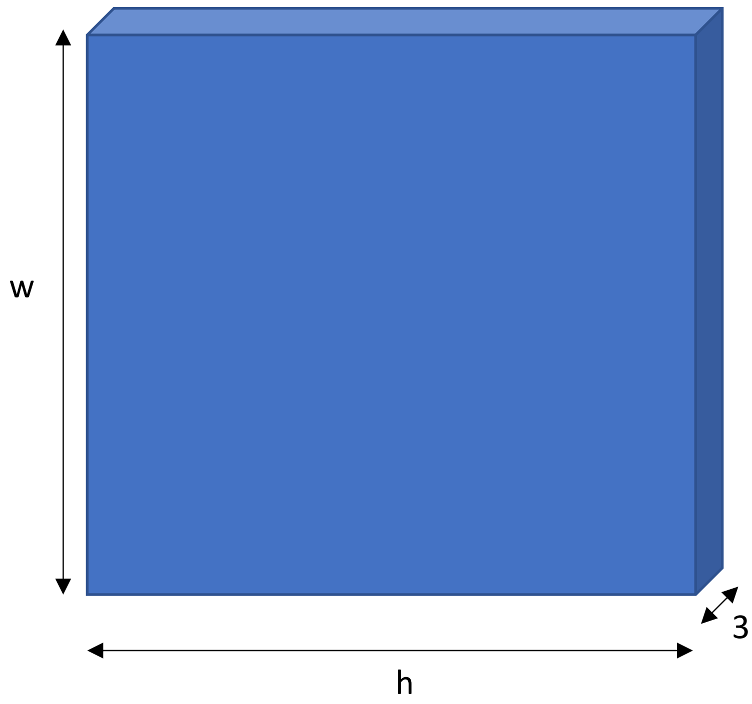
Challenges in data collection

- Precise localization is hard to annotate
- Annotating every pixel leads to heavy tails
- Common solution: annotate few classes (often things), mark rest as "Other"
- Common datasets: PASCAL VOC 2012 (~1500 images, 20 categories), COCO (~100k images, 20 categories)

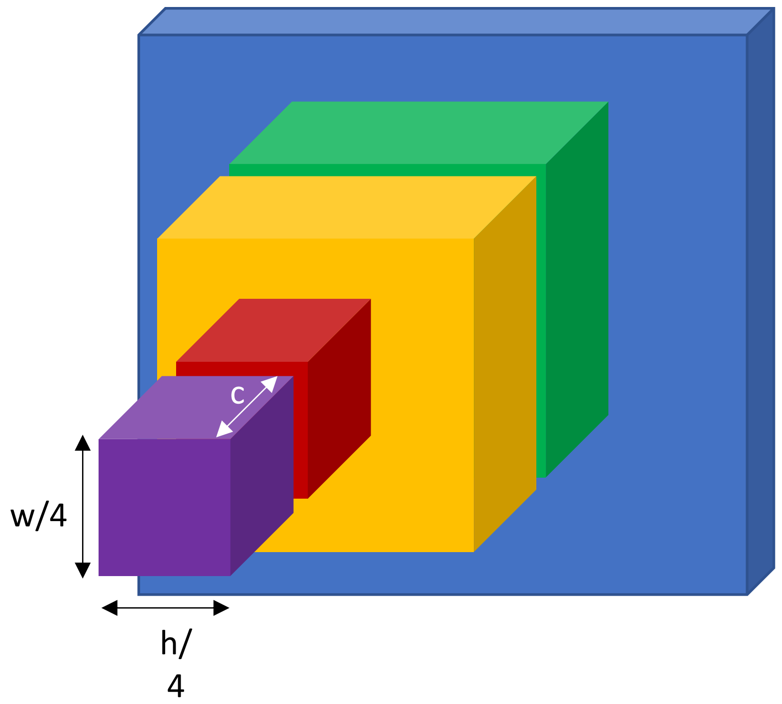
Pre-convnet semantic segmentation

- Things
  - Do object detection, then segment out detected objects
- Stuff
  - ”Texture classification”
  - Compute histograms of filter responses
  - Classify local image patches

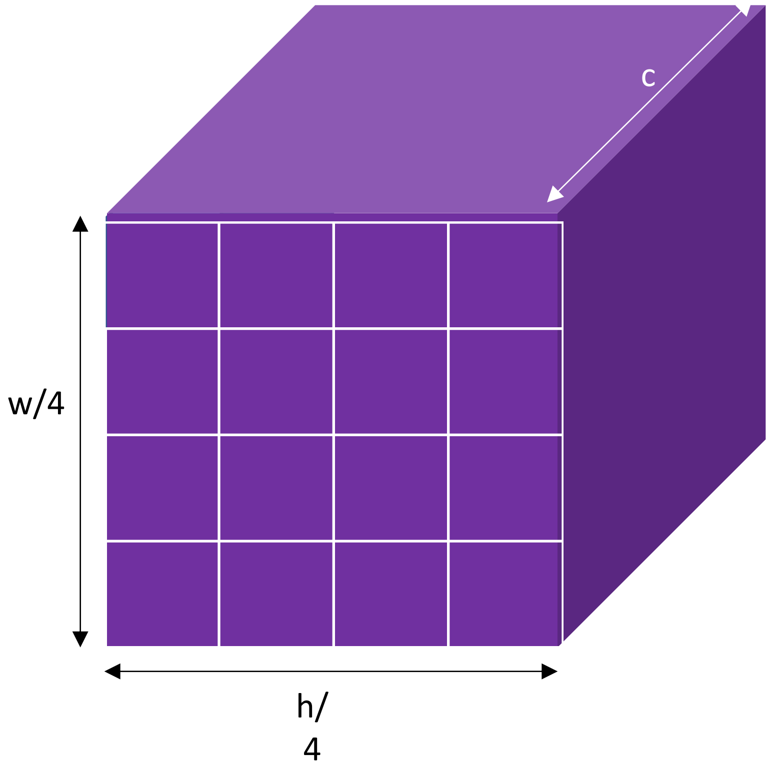
Semantic segmentation using convolutional networks



Semantic segmentation using convolutional networks

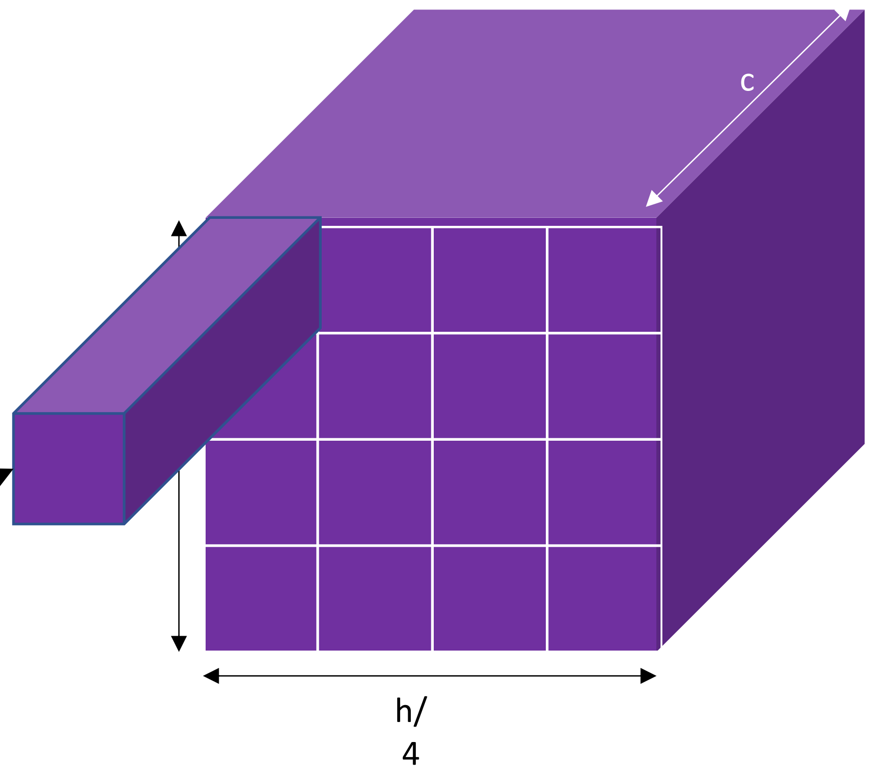


Semantic segmentation using convolutional networks

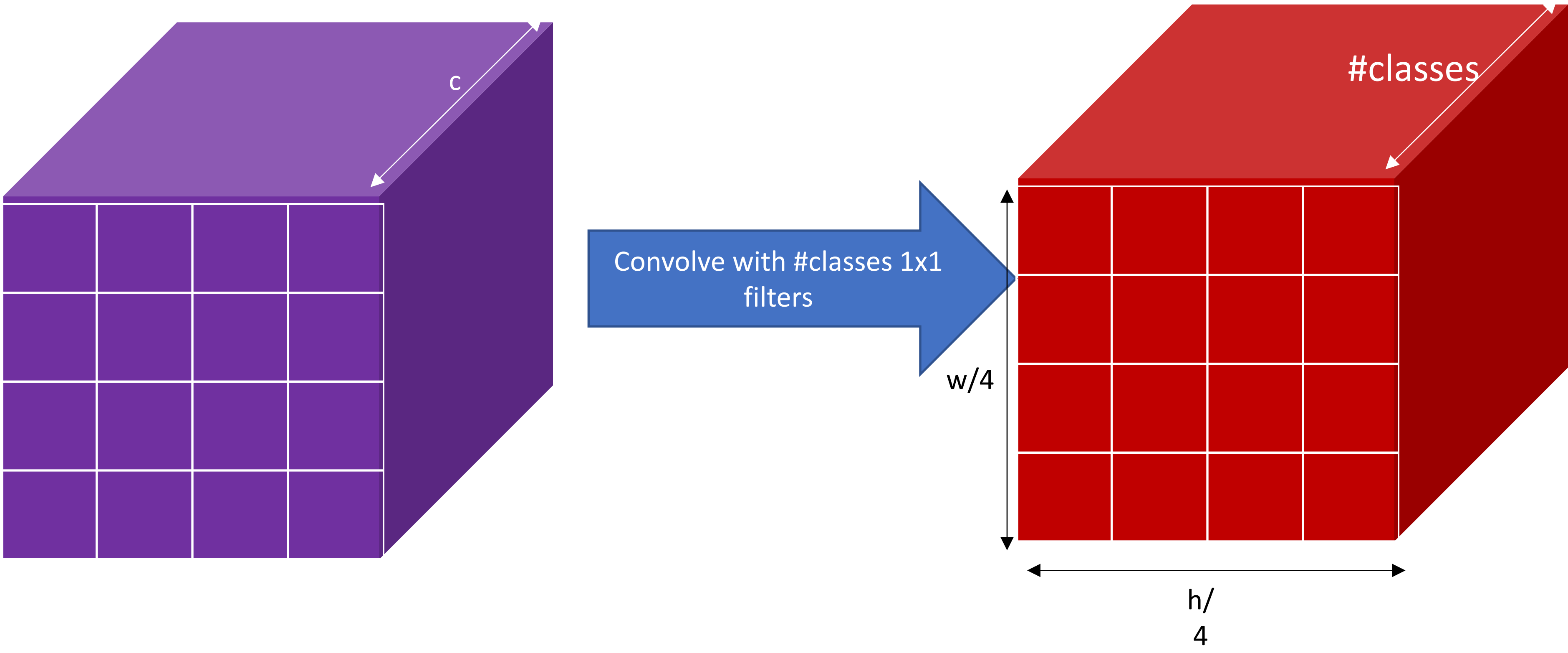


Semantic segmentation using convolutional networks

Can be considered as a feature vector for a pixel



Semantic segmentation using convolutional networks



Semantic segmentation using convolutional networks

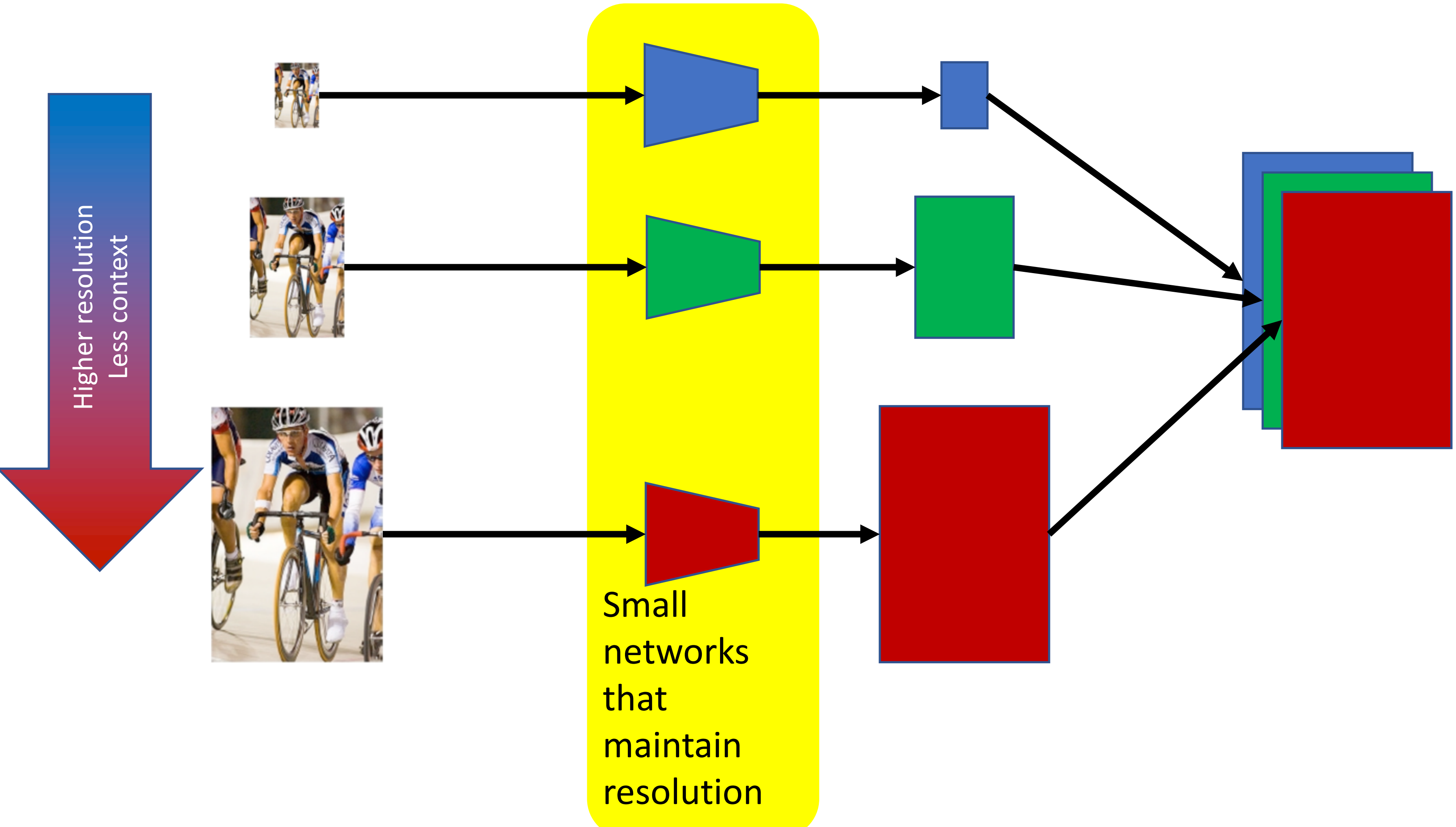
- Pass image through convolution and subsampling layers
- Final convolution with #classes outputs
- Get scores for *subsampled* image
- Upsample back to original size

Semantic segmentation using convolutional networks

The resolution issue

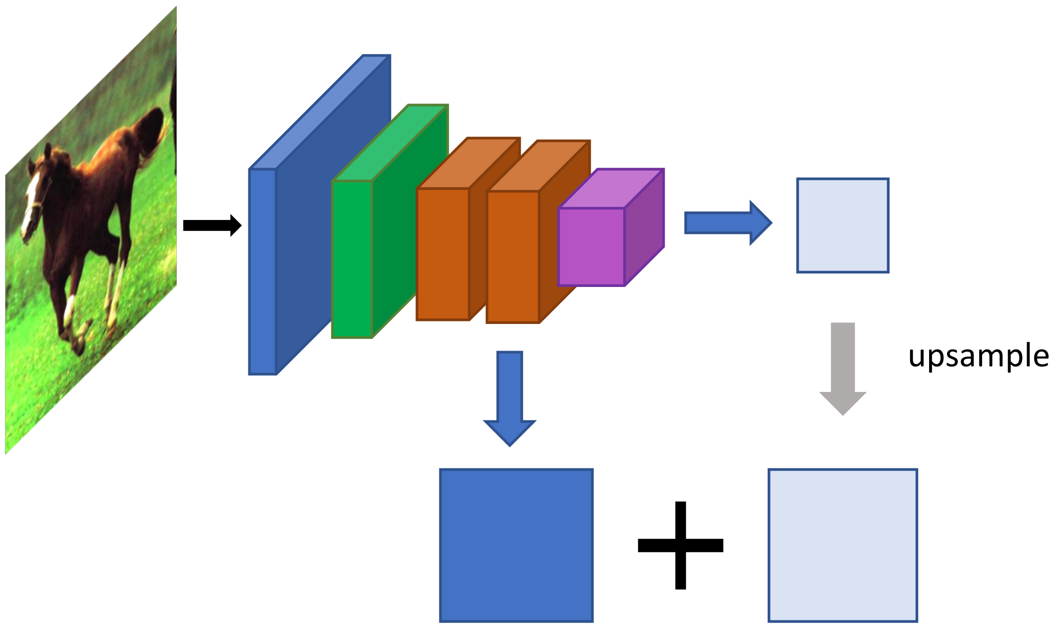
- Problem: Need fine details!
- Shallower network / earlier layers?
  - Deeper networks work better: more abstract concepts
  - Shallower network => Not very semantic!
- Remove subsampling?
  - Subsampling allows later layers to capture larger and larger patterns
  - Without subsampling => Looks at only a small window!

Solution 1: Image pyramids



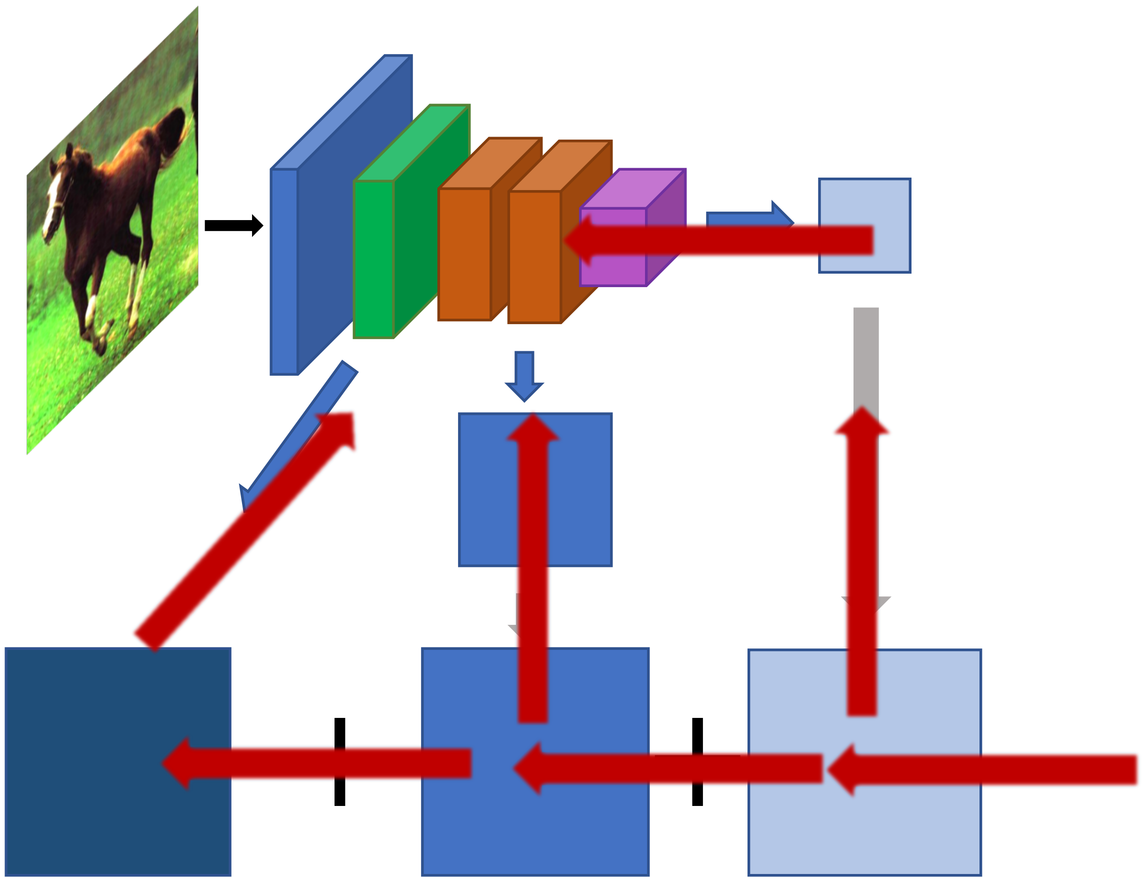
Learning Hierarchical Features for Scene Labeling. Clement Farabet, Camille Couprie, Laurent Najman, Yann LeCun. In *TPAMI,* 2013

Solution 2: Skip connections



Compute class scores at multiple layers, then upsample and add

Solution 2: Skip connections



Red arrows indicate backpropagation

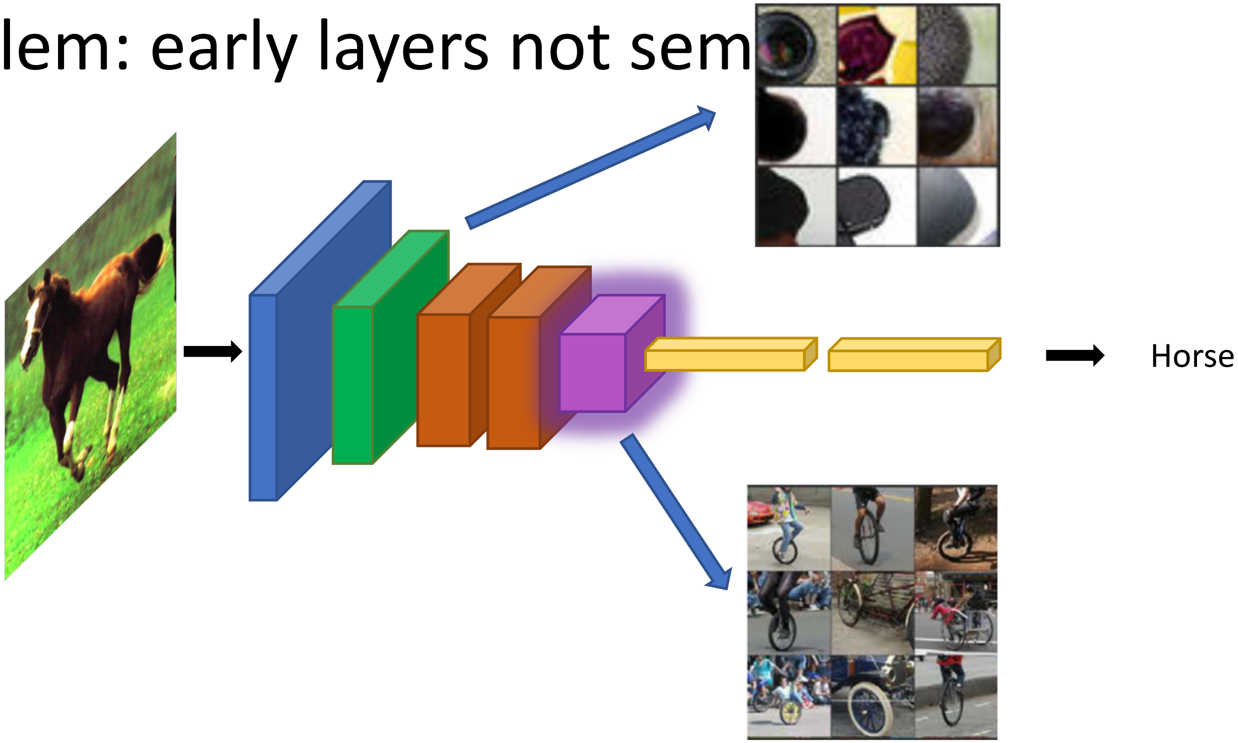
Skip connections

without skip with skip

Fully convolutional networks for semantic segmentation. Evan Shelhamer, Jon Long, Trevor Darrell. In *CVPR* 2015

Skip connections

- Problem: early layers not sem



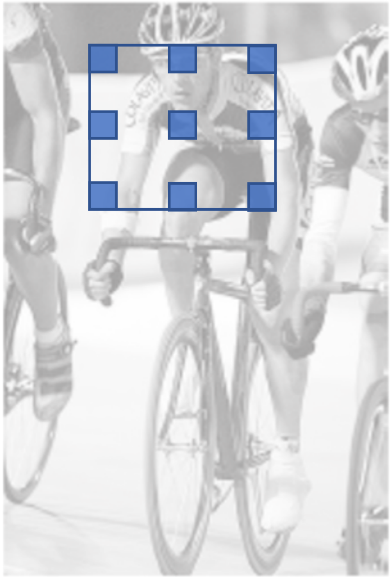
Visualizations from : M. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *ECCV* 2014.

Solution 3: Dilation

- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?

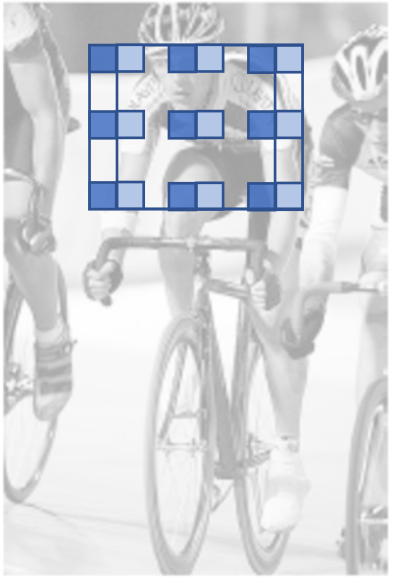
Solution 3: Dilation

- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?



Solution 3: Dilation

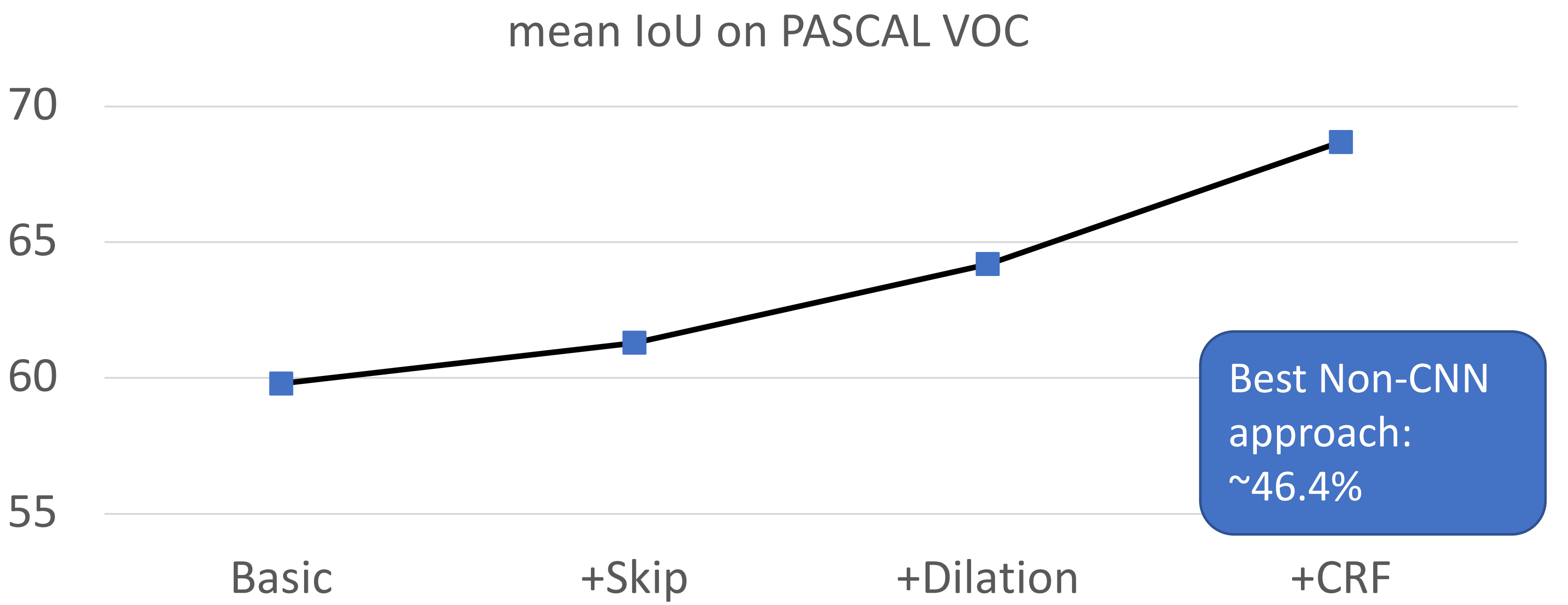
- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?



Solution 3: Dilation

- Instead of subsampling by factor of 2: dilate by factor of 2
- Dilation can be seen as:
  - Using a much larger filter, but with most entries set to 0
  - Taking a small filter and “exploding”/ “dilating” it
- Not panacea: without subsampling, feature maps are much larger: memory issues

Putting it all together



Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan Yuille. In *ICLR,* 2015.

Other additions

Method	mean IoU (%)
VGG16 + Skip + Dilation	65.8
ResNet101	68.7
ResNet101 + Pyramid	71.3
ResNet101 + Pyramid + COCO	74.9

DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan Yuille. Arxiv 2016.

Image-to-image translation problems

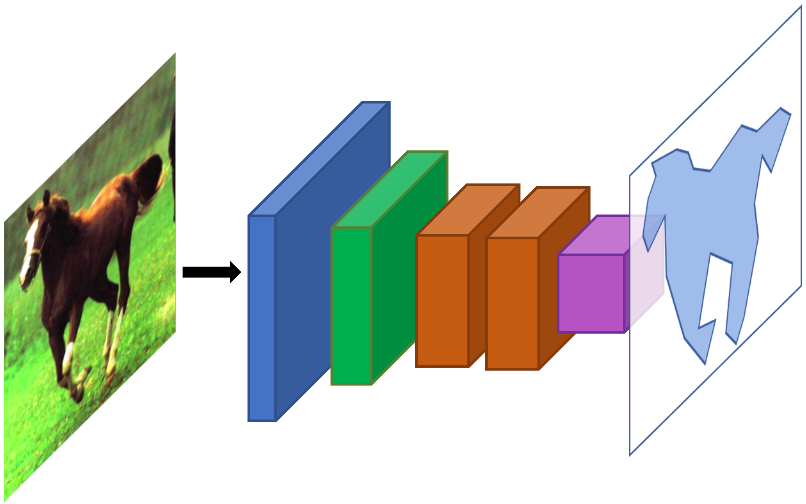


Image-to-image translation problems

- Segmentation
- Optical flow estimation
- Depth estimation
- Normal estimation
- Boundary detection
- …