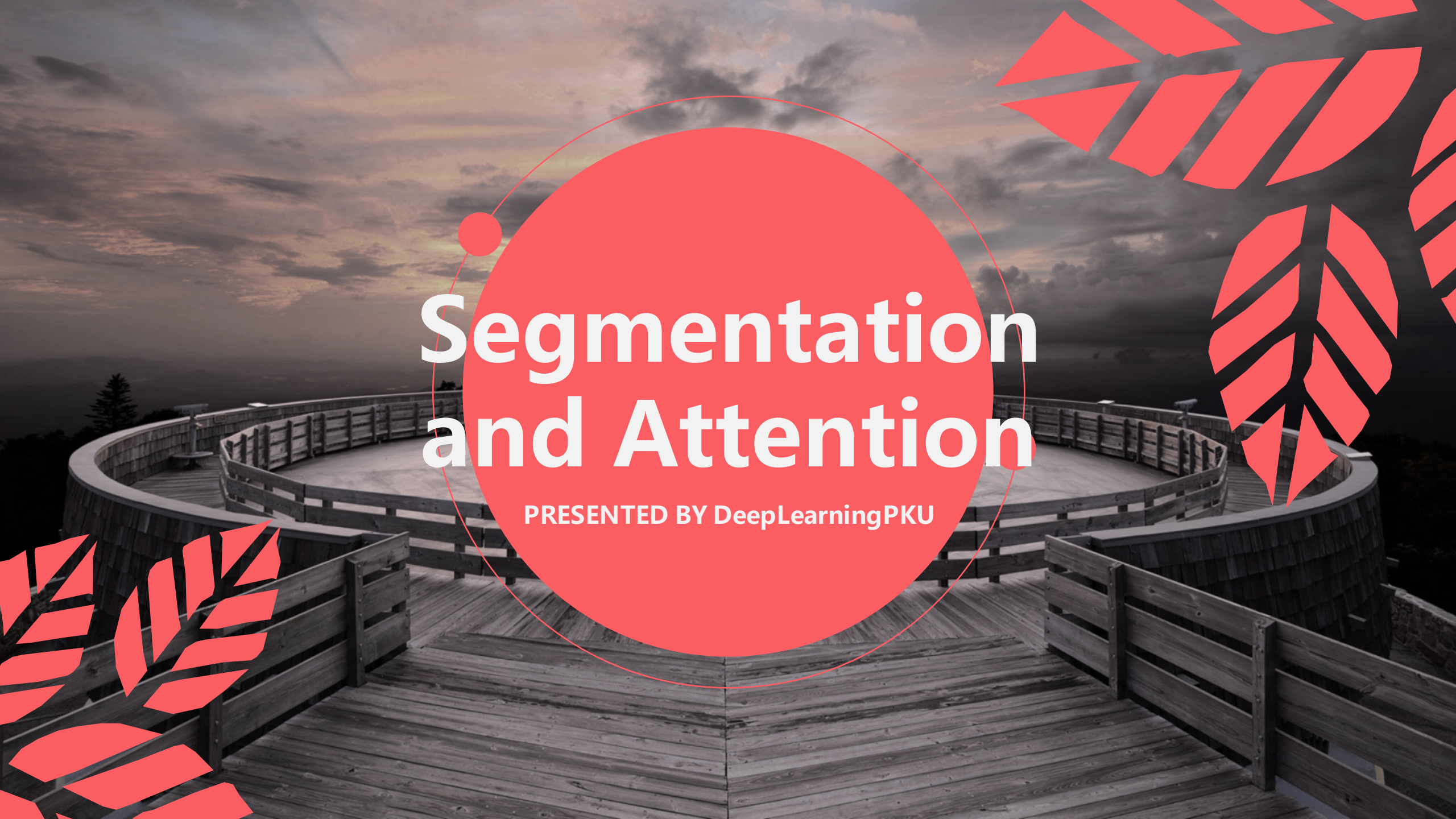


Segmentation and Attention

PRESENTED BY DeepLearningPKU

Segmentation & Attention

01 **PART ONE**
Semantic Segmentation

02 **PART TWO**
Instance Segmentation

03 **PART THREE**
Discrete locations

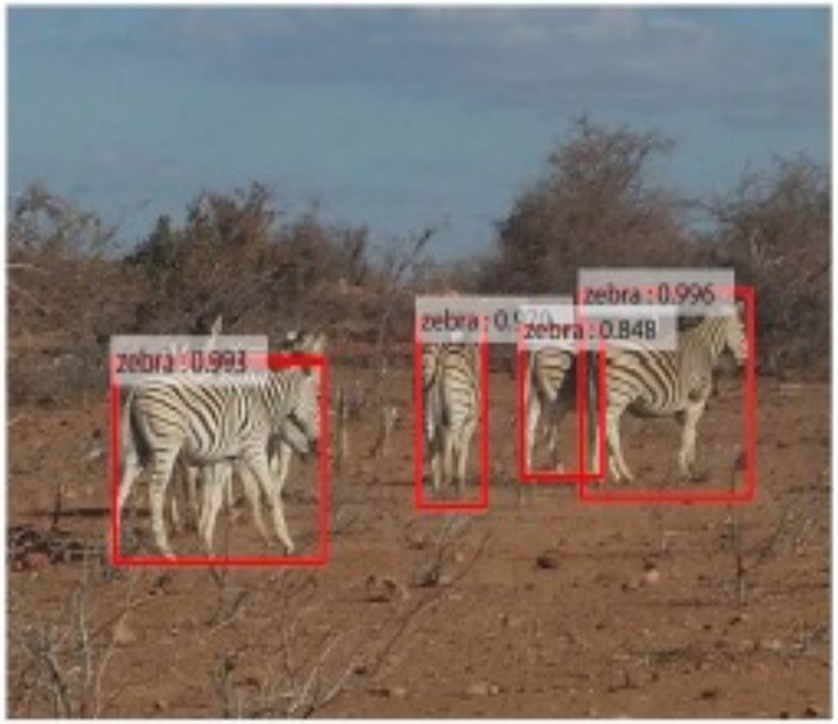
04 **PART FOUR**
Continuous locations

Segmentation

Semantic Segmentation

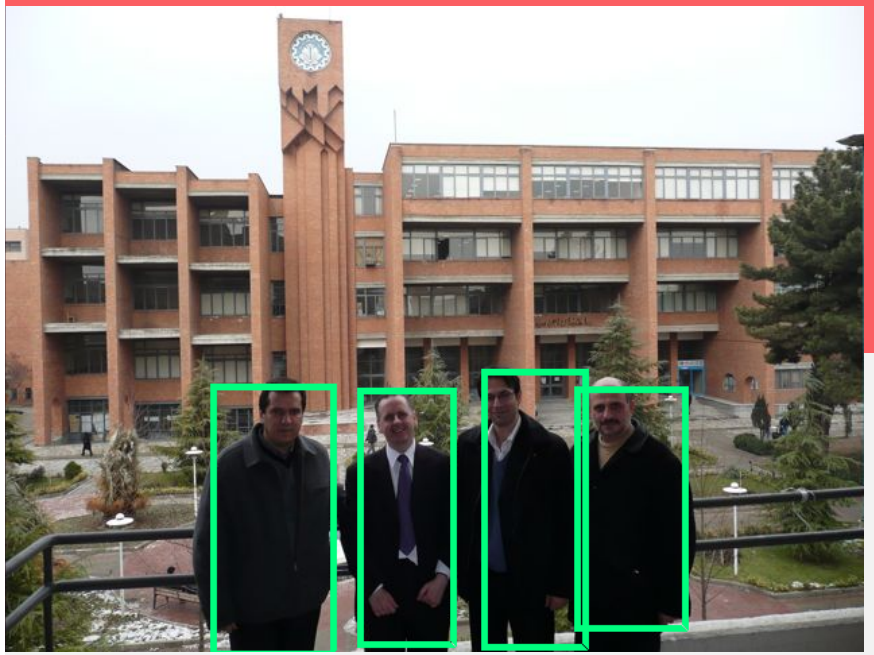
Instance Segmentation

基本概况



Object Recognition:

Localized them with a bounding box and label that bounding box with a label.

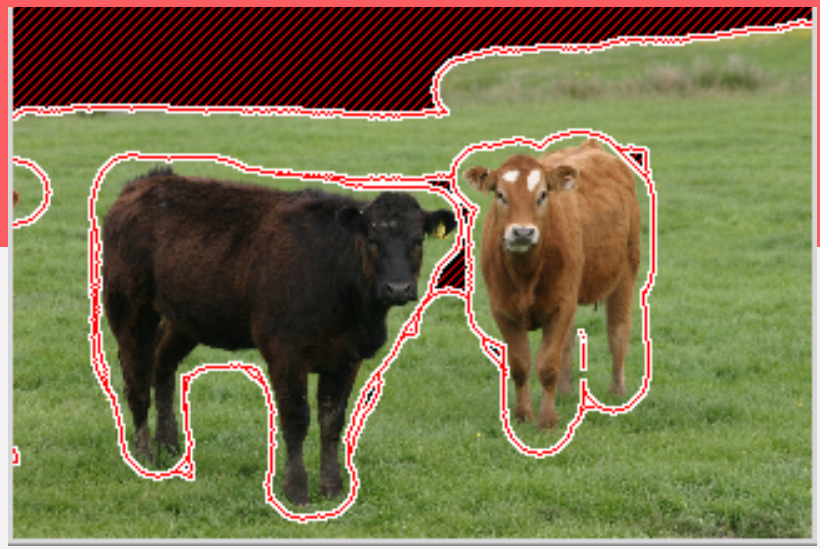


Object Detection:

Object recognition

two class of object classification

which means object bounding boxes and non-object bounding boxes.

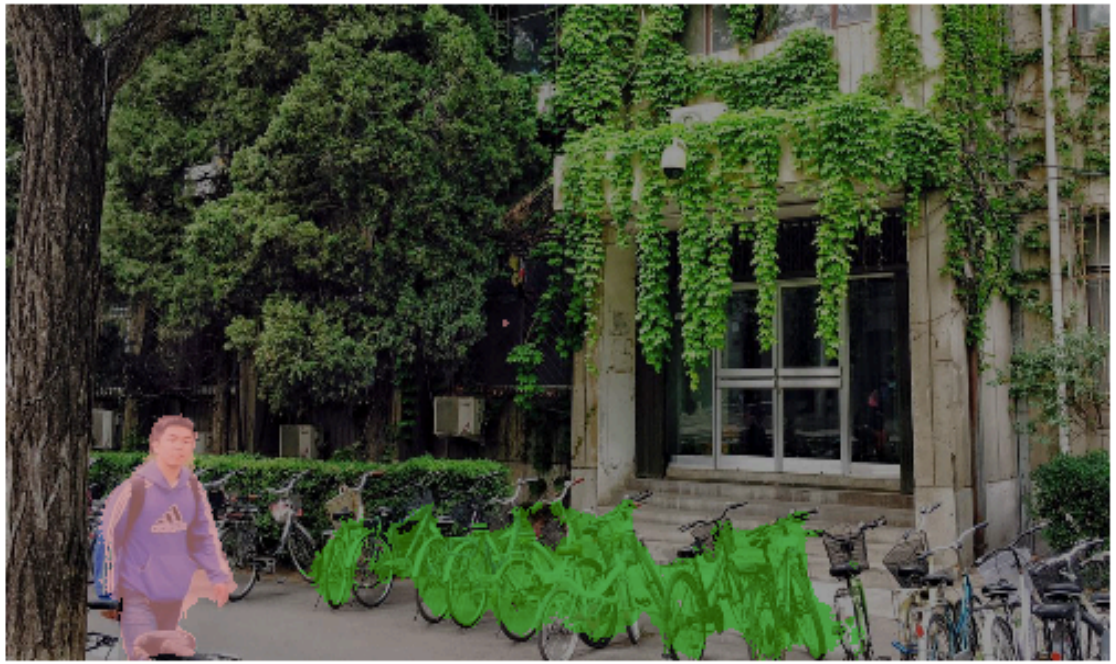


Object Segmentation:

Like object recognition you will recognize all objects in an image but your output should show this object classifying pixels of the image.

Semantic Segmentation

Original image (hover to highlight segmented parts)



Semantic segmentation

Objects appearing in the image:

Bicycle | Person

Objects not appearing in the image:

Aeroplane | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow

Dining table | Dog | Horse | Motorbike | Potted plant | Sheep | Sofa | Train | TV/Monitor

Computer Vision Tasks

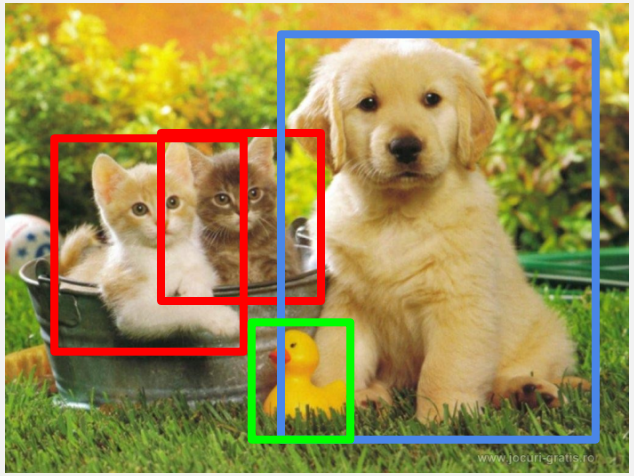
Classification

CAT

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

Segmentation

CAT, DOG, DUCK

Single object

Multiple objects

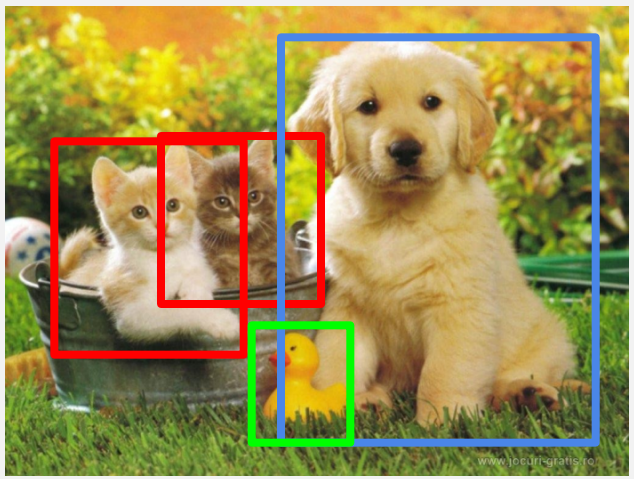
Computer Vision Tasks

Classification

Classification + Localization

Object Detection

Segmentation



Lecture 8

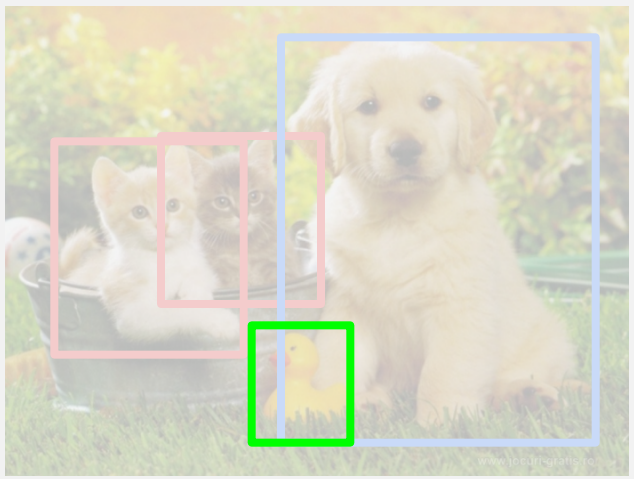
Computer Vision Tasks

Classification

Classification + Localization

Object Detection

Segmentation



Today

Semantic Segmentation

Label every pixel!

Don’ t differentiate instances (cows)

Classic computer vision problem

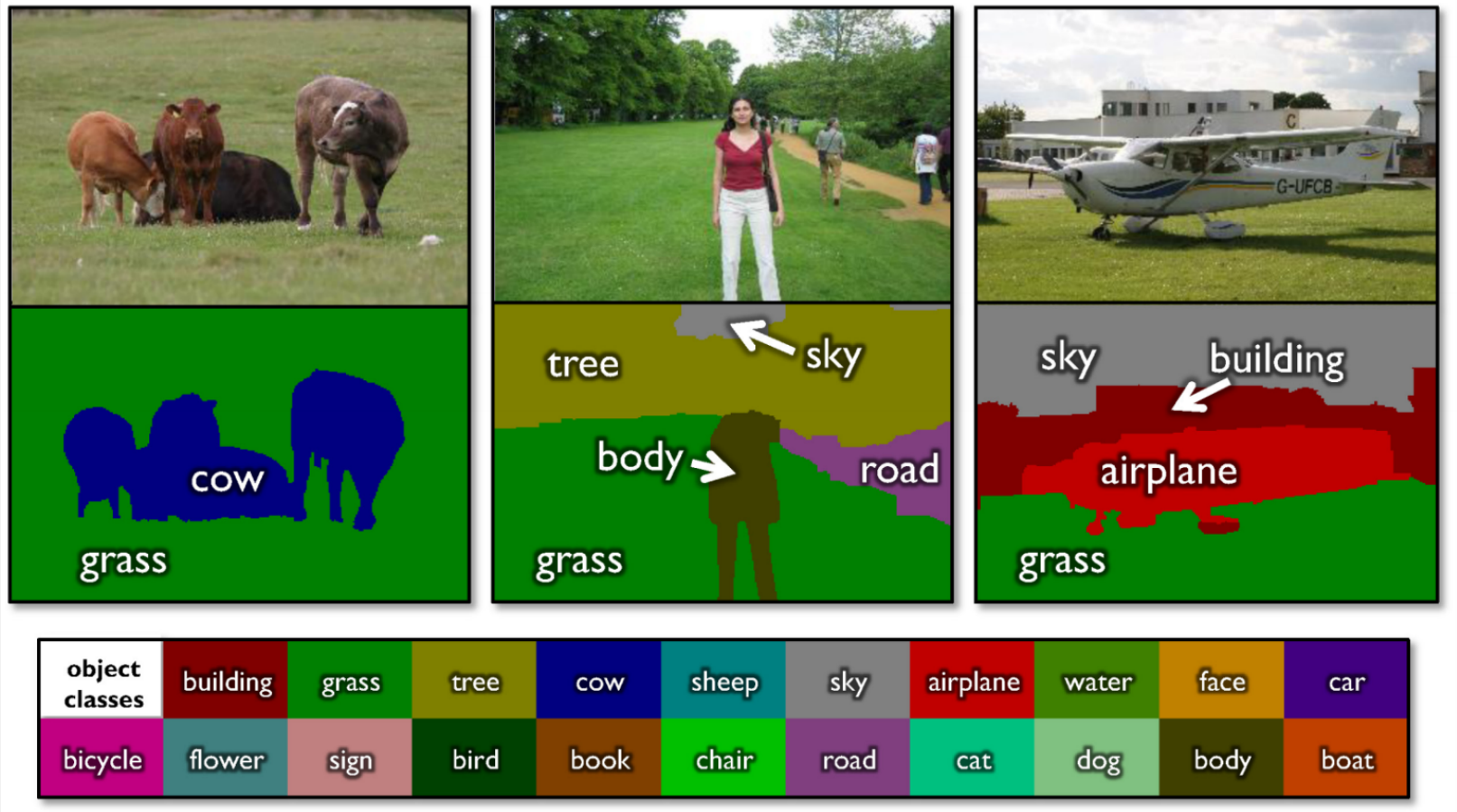


Figure credit: Shotton et al, “TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context”, IJCV 2007

Instance Segmentation

Detect instances, give category, label pixels

"simultaneous detection and segmentation" (SDS)

Lots of recent work (MS-COCO)

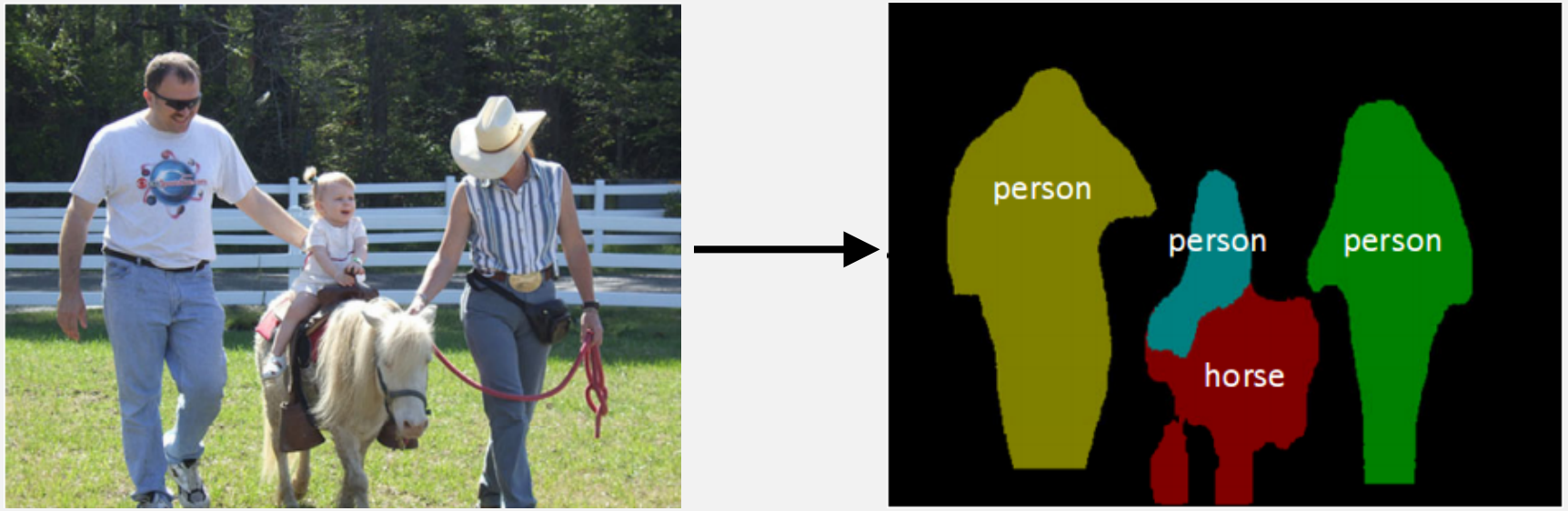


Figure credit: Dai et al, "Instance-aware Semantic Segmentation via Multi-task Network Cascades", arXiv 2015

PART 1

Semantic Segmentation

Semantic segmentation faces an inherent tension between semantics and location: global information resolves what while local information resolves where.

基本概况

Semantic segmentation

Scene parsing, or semantic segmentation, consists in labeling each pixel in an image with the category of the object it belongs to. It is a challenging task that involves the simultaneous detection, segmentation and recognition of all the objects in the image.

1

Semantic Segmentation

The goal of the scene labeling task is to assign a class label to each pixel in an image.

Semantic Segmentation: Multi-Scale

2

uses purely supervised training from fully-labeled images to learn appropriate low-level and mid-level features

3

Refinement

For a good visual coherence and a high class accuracy, capture long range (pixel) label dependencies in images.

Upsampling

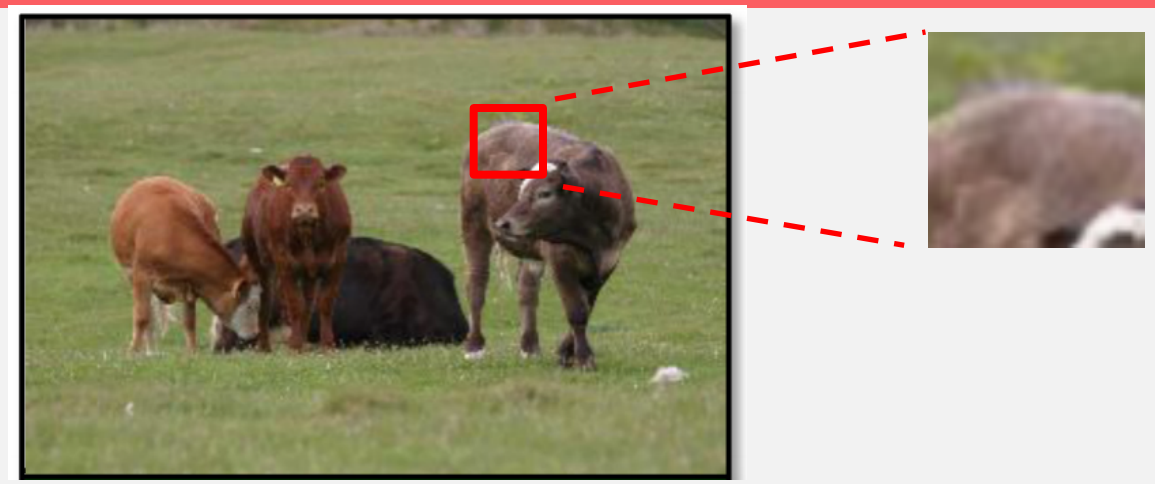
4

semantic information + appearance information
a deep, coarse layer + a shallow, fine layer
to produce accurate and detailed segmentations.

Semantic Segmentation

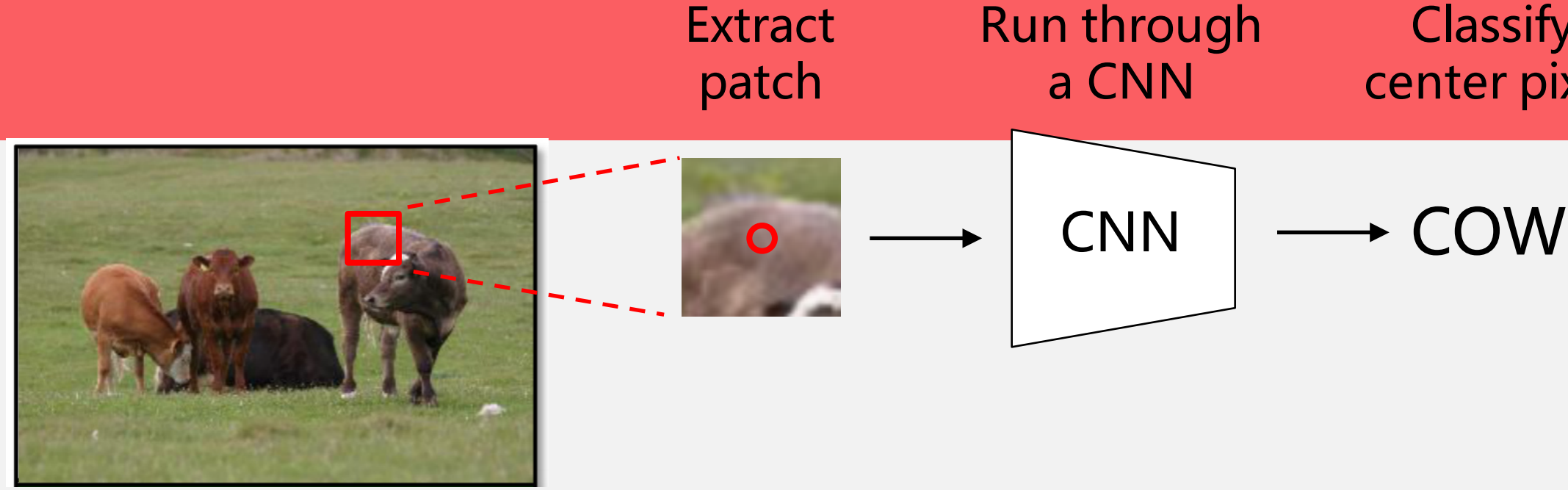
Semantic Segmentation

Extract patch

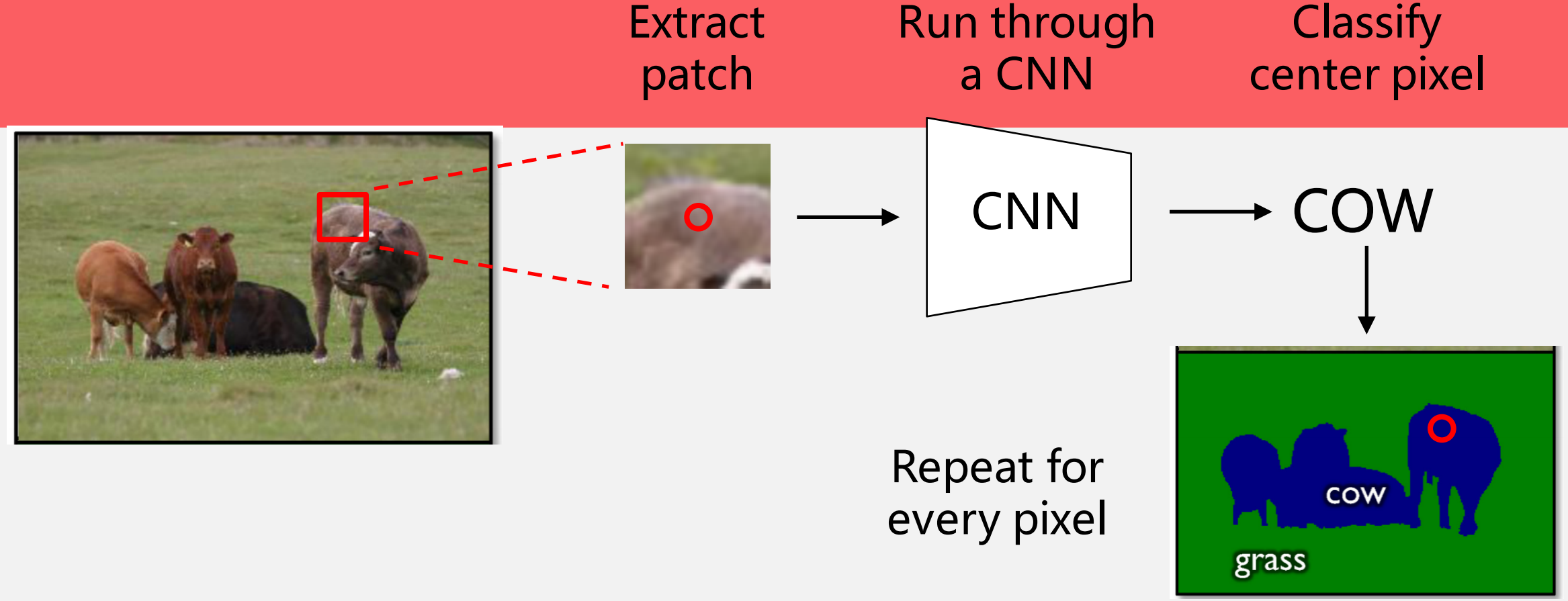


Semantic Segmentation

Semantic Segmentation

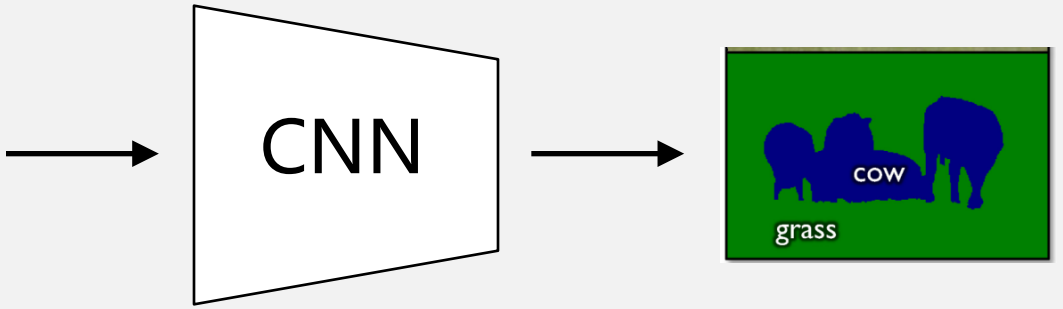


Semantic Segmentation



Semantic Segmentation

Run "fully convolutional" network to get all pixels at once



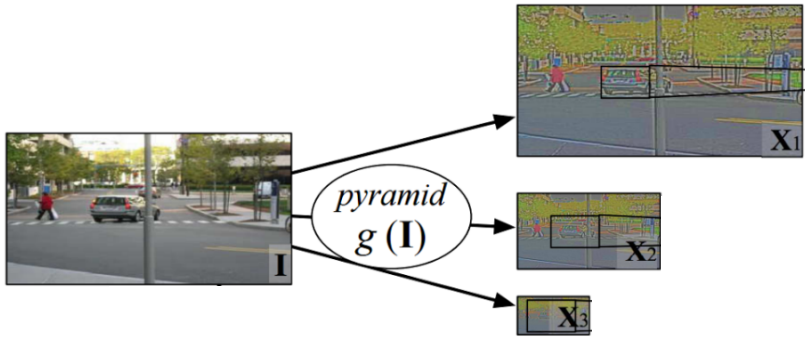
Smaller output due to pooling

Semantic Segmentation: Multi-Scale

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013

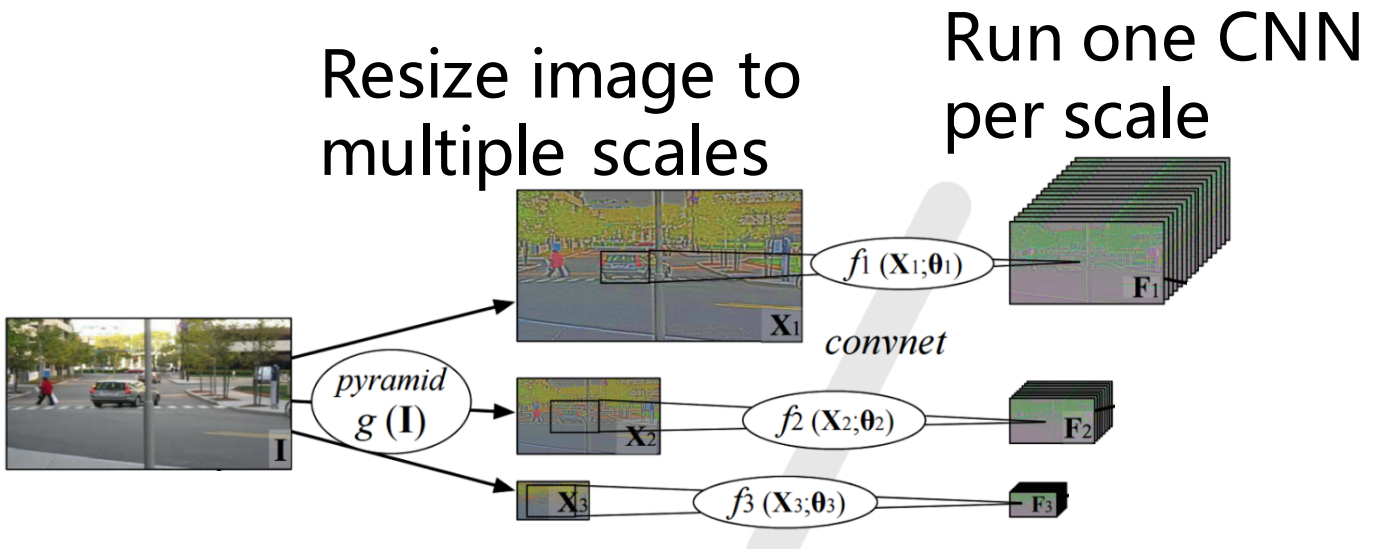
Semantic Segmentation: Multi-Scale

Resize image to multiple scales



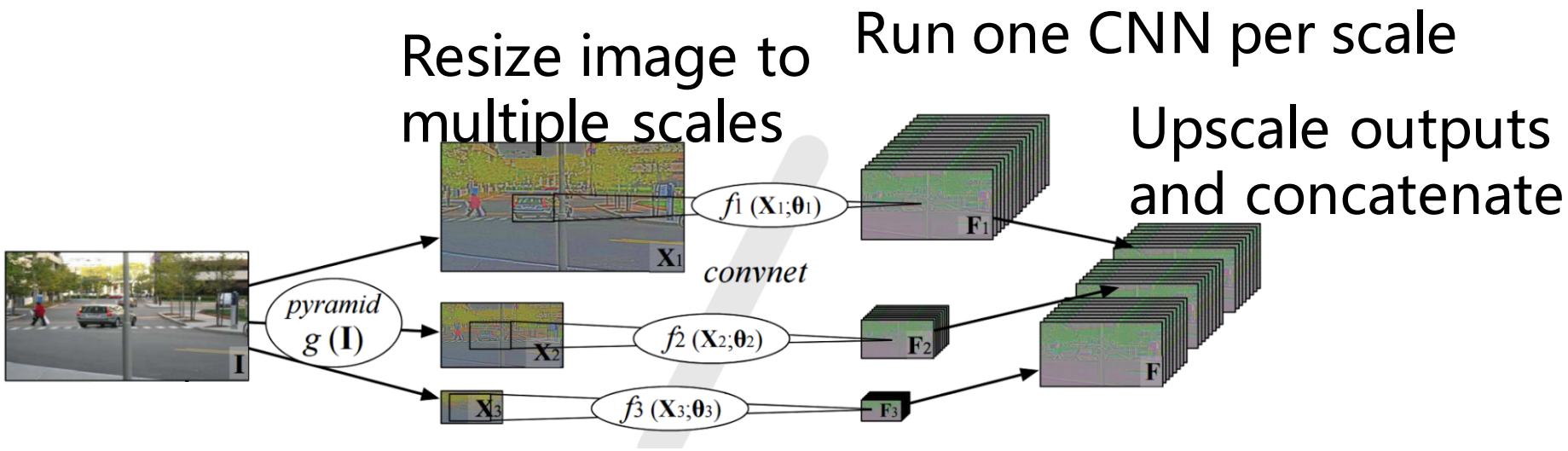
Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013

Semantic Segmentation: Multi-Scale



Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013

Semantic Segmentation: Multi-Scale



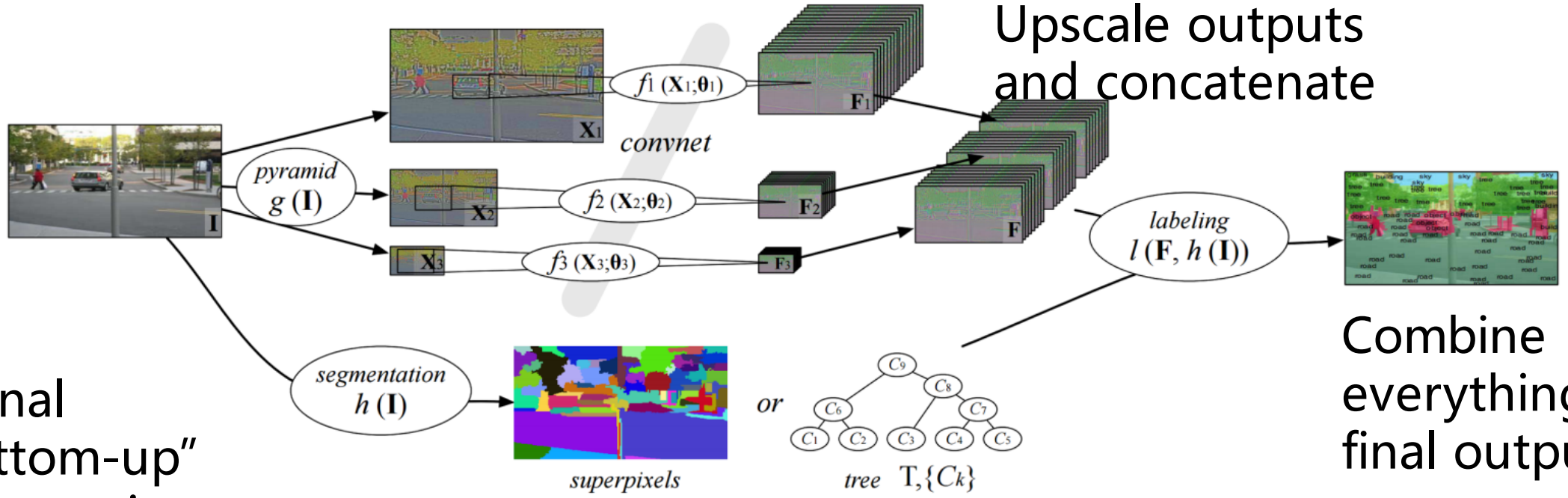
Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013

Semantic Segmentation: Multi-Scale

Resize image to multiple scales

Run one CNN per scale

Upscale outputs and concatenate

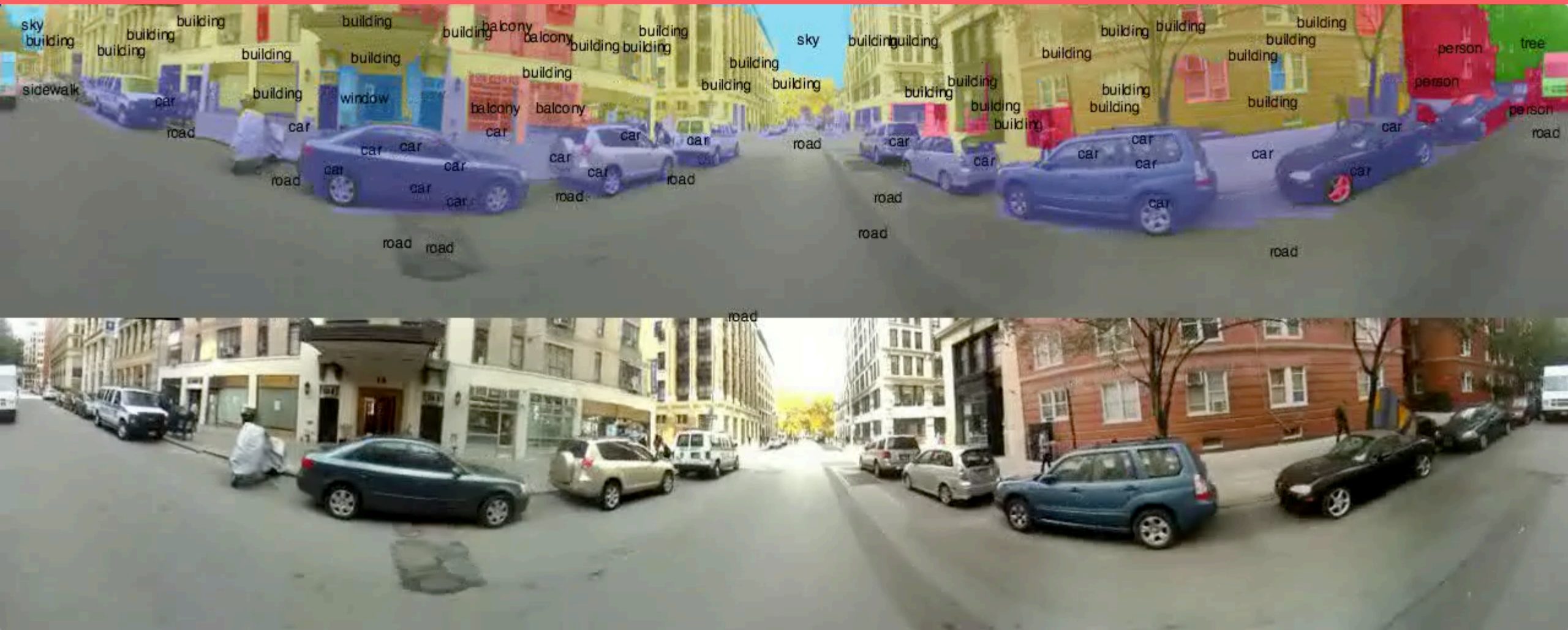


Combine everything for final outputs

External "bottom-up" segmentation

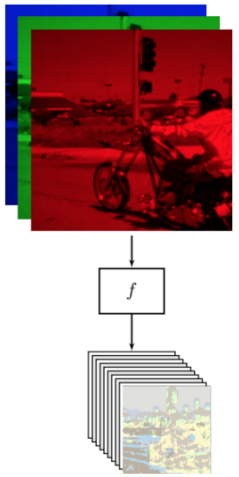
Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013

Multi-Scale



Semantic Segmentation: Refinement

Apply CNN once to get labels

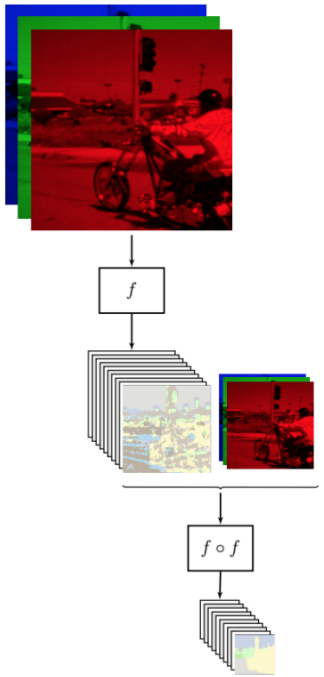


Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation: Refinement

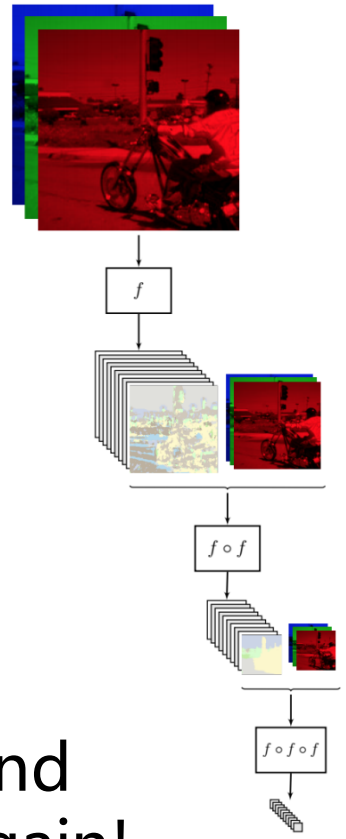
Apply CNN once to get labels

Apply AGAIN to refine labels



Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation: Refinement



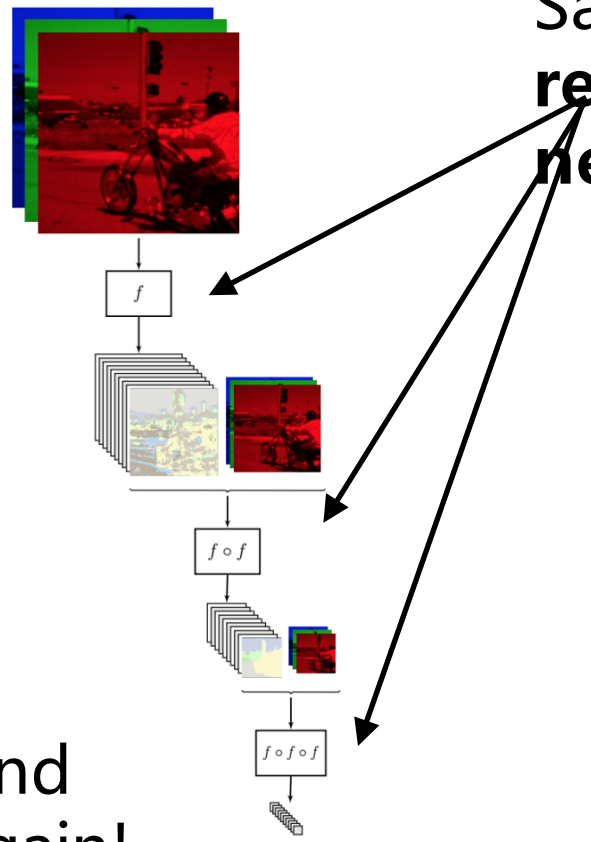
Apply CNN once to get labels

Apply AGAIN to refine labels

And again!

Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation: Refinement



Same CNN weights:
recurrent convolutional network

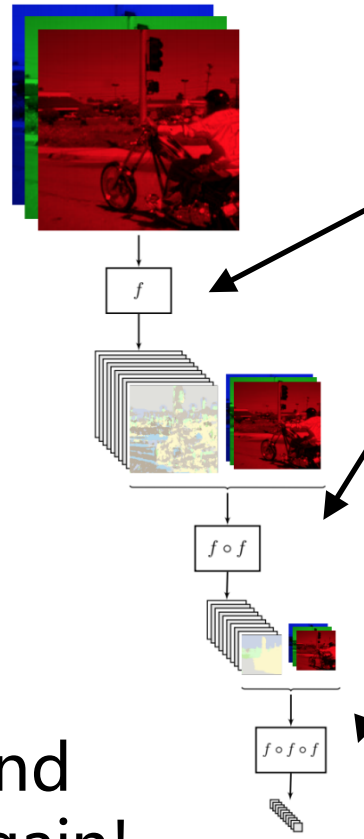
Apply CNN once to get labels

Apply AGAIN to refine labels

And again!

Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation: Refinement

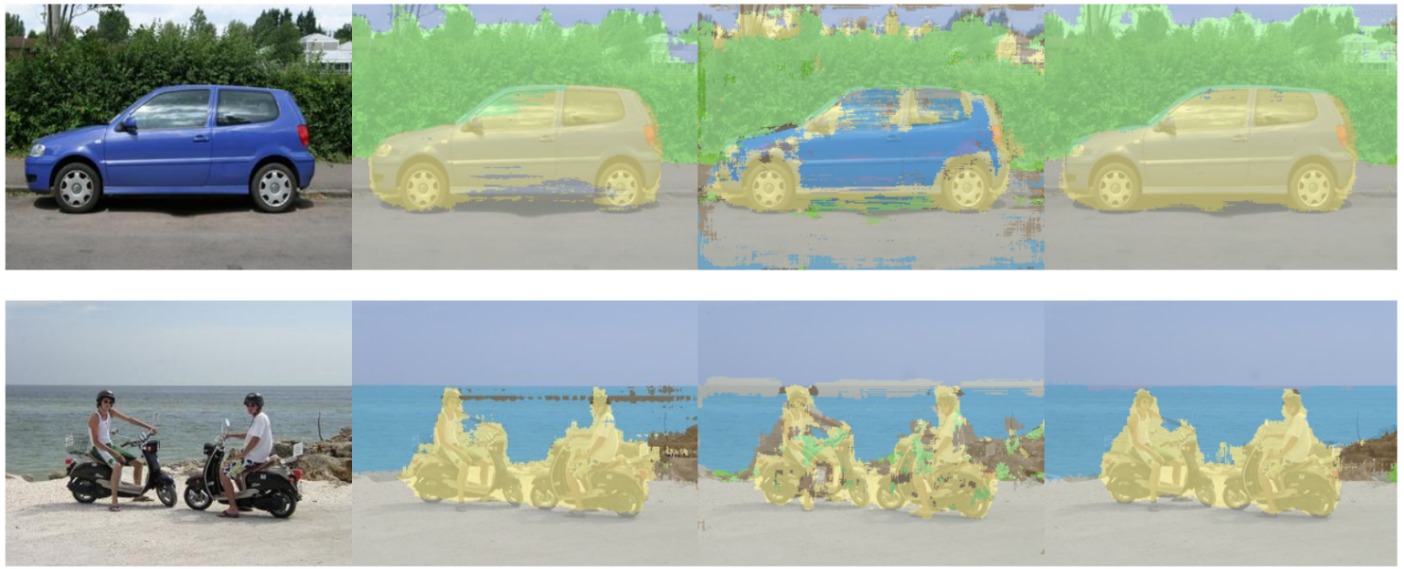


Same CNN weights:
recurrent convolutional network

Apply CNN once to get labels

Apply AGAIN to refine labels

And again!

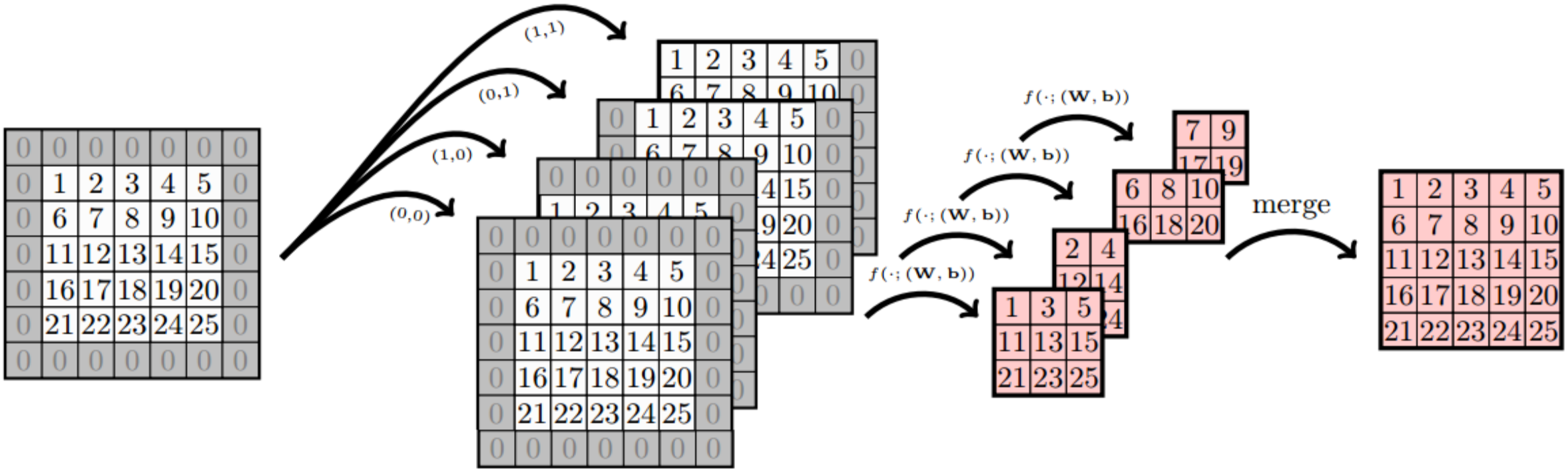


More iterations improve results

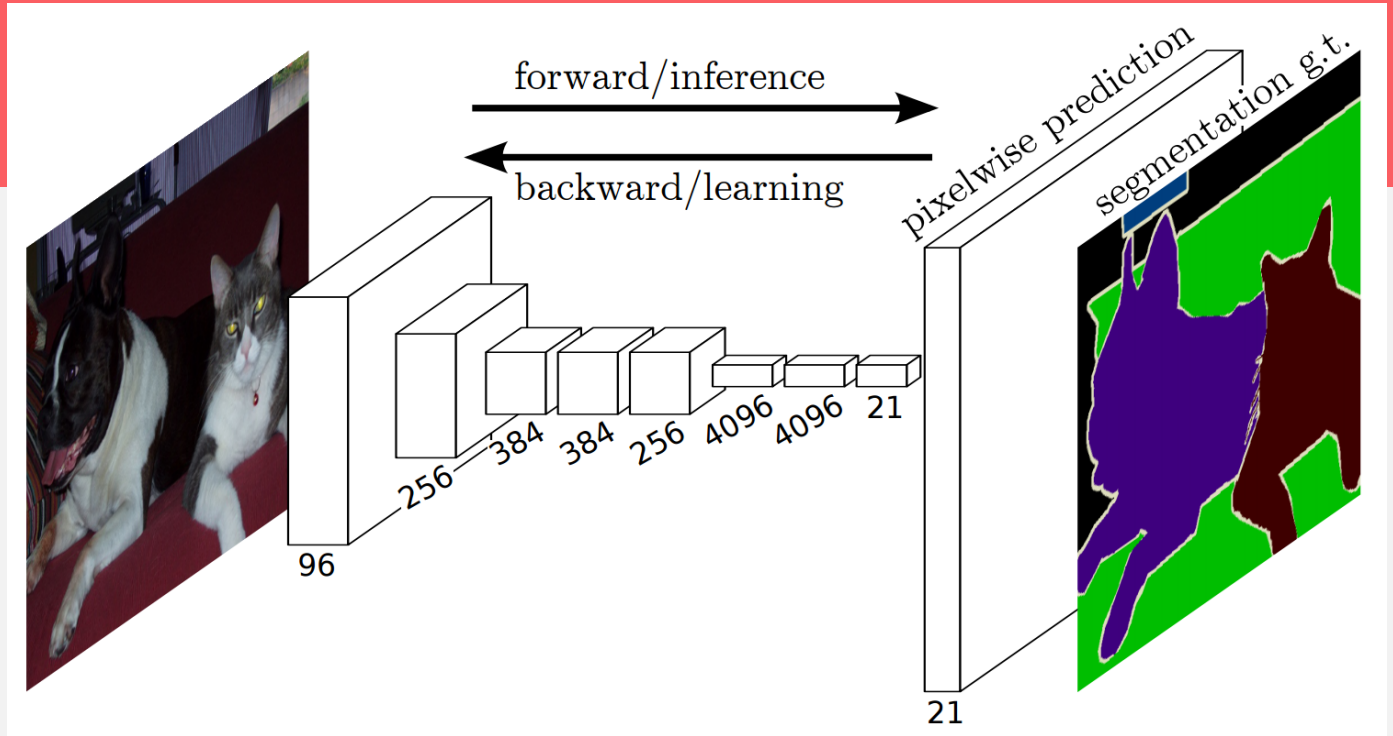
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Refinement

- downscaled label planes (compared to the input image) due to pooling layers.
- Downscaled predicted label planes (here in red) are then merged to get back the full resolution label plane in an efficient manner

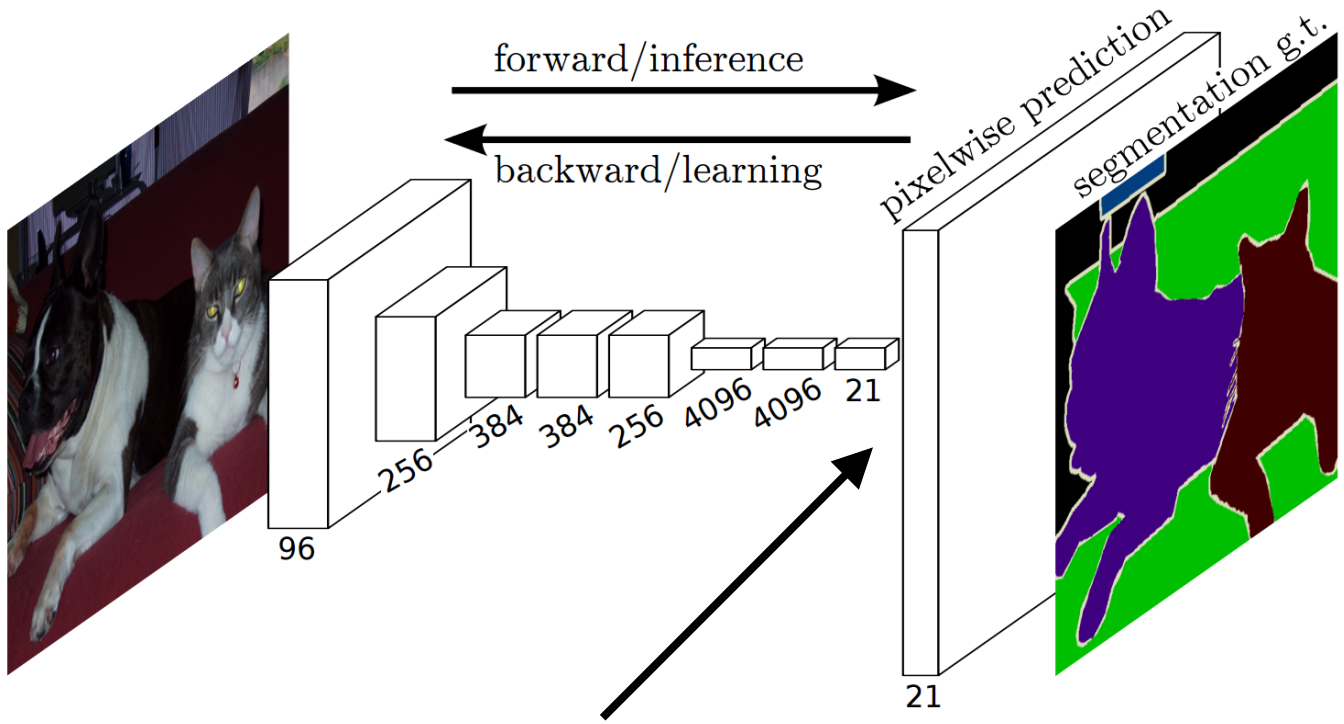


Semantic Segmentation:Upsampling



Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

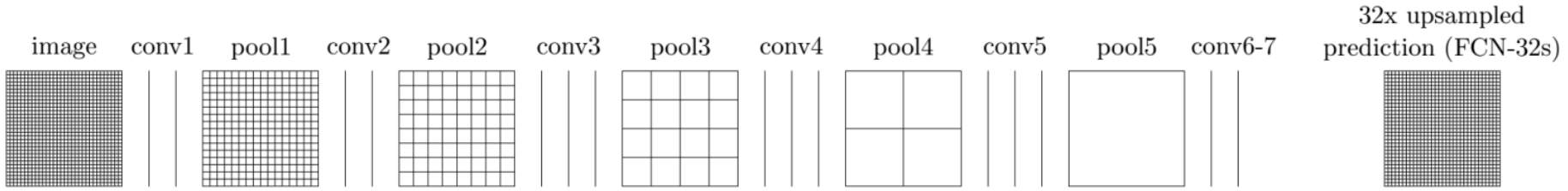
Semantic Segmentation: Upsampling



Learnable upsampling!

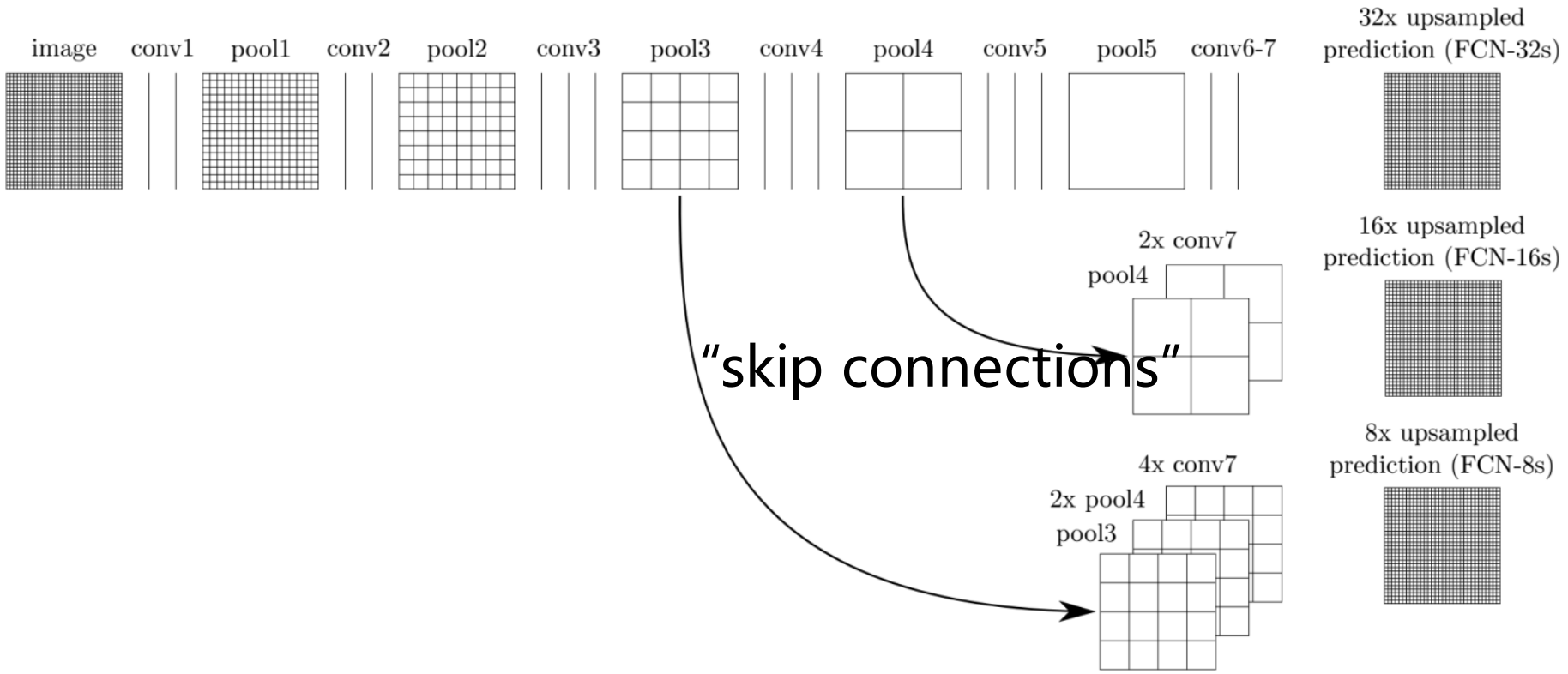
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

Semantic Segmentation: Upsampling



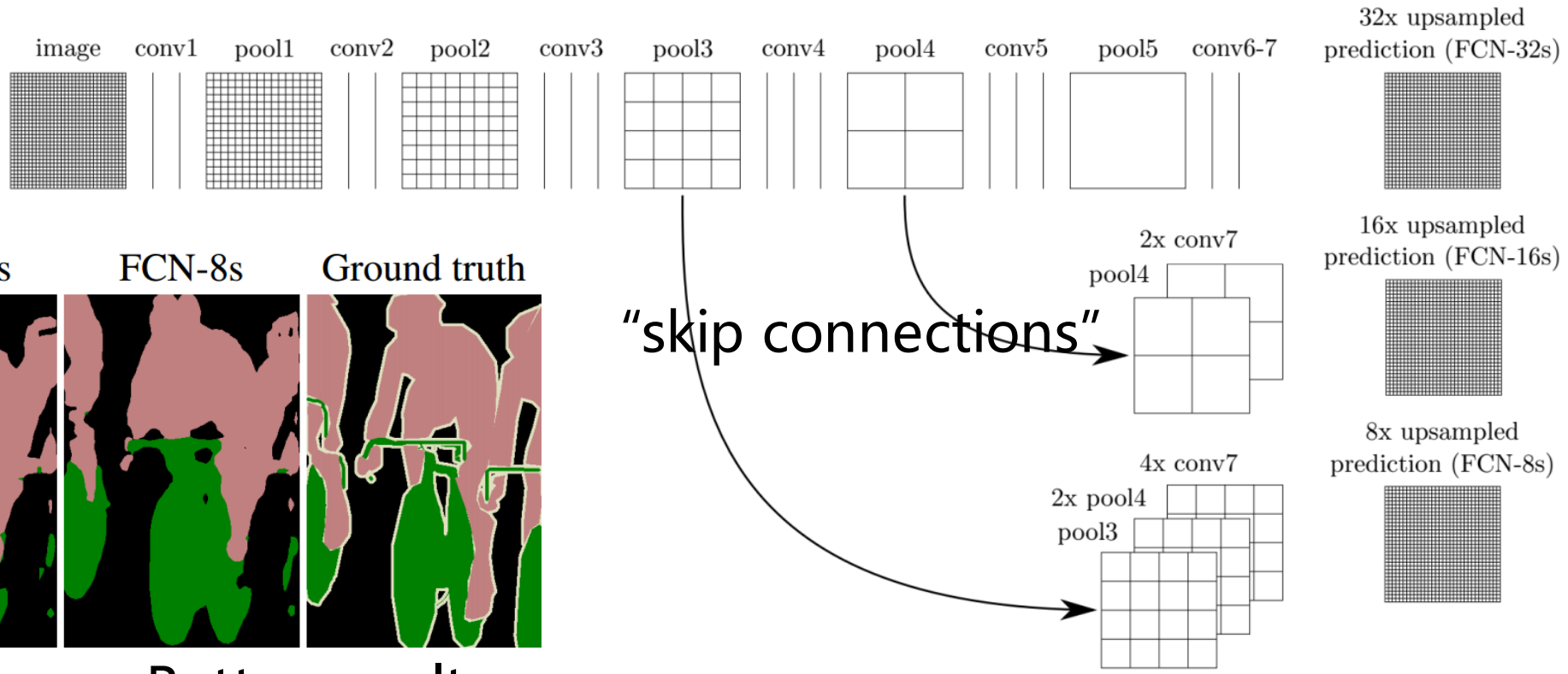
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

Semantic Segmentation: Upsampling



Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

Semantic Segmentation: Upsampling

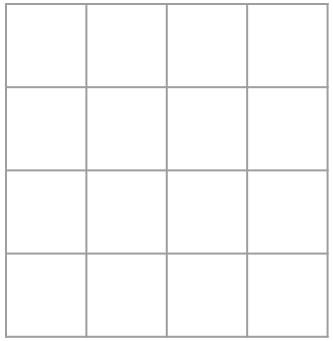


Skip connections = Better results

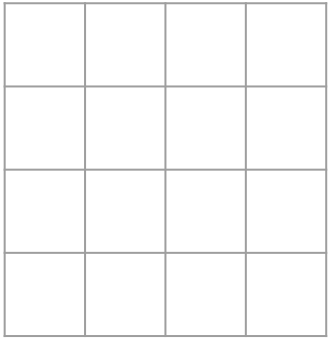
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

Learnable Upsampling: “Deconvolution”

Typical 3 x 3 convolution, stride 1 pad 1



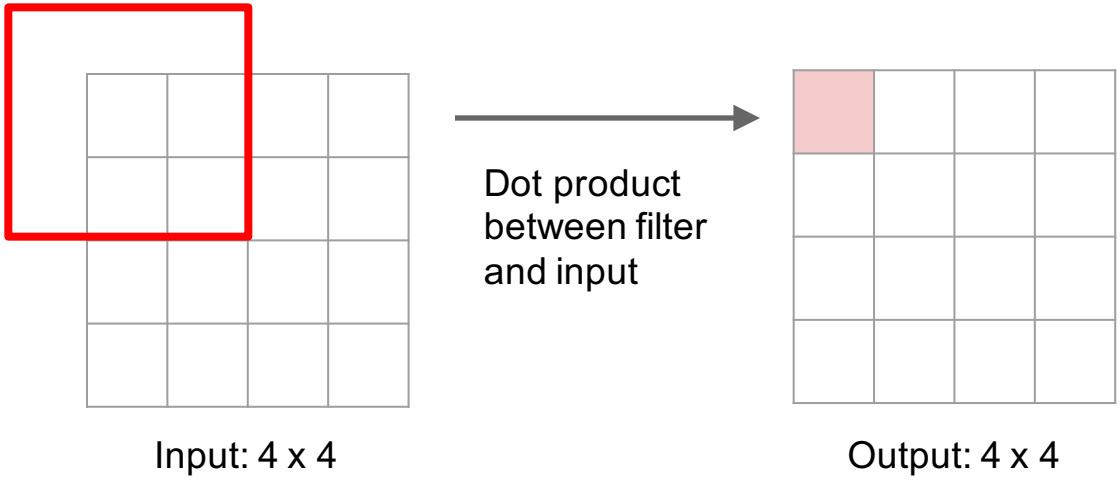
Input: 4 x 4



Output: 4 x 4

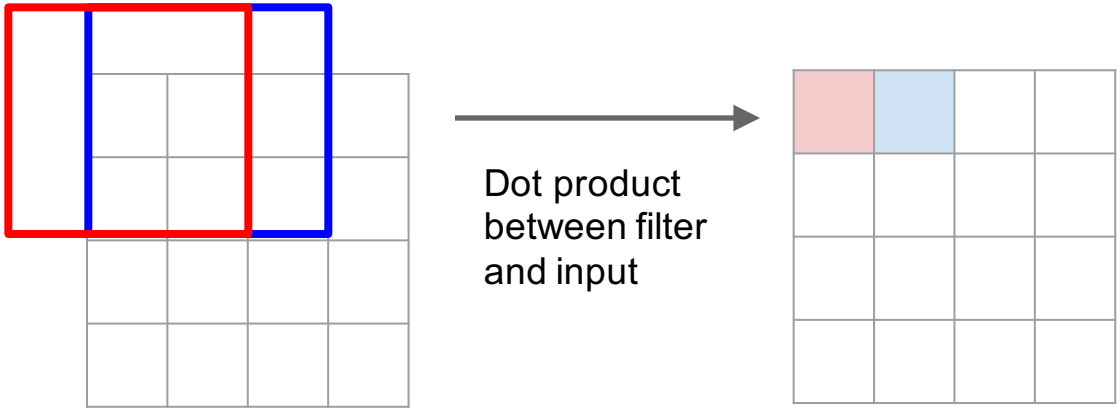
Learnable Upsampling: “Deconvolution”

Typical 3 x 3 convolution, stride 1 pad 1



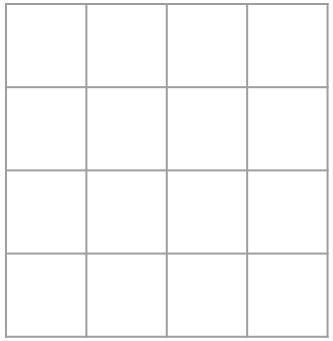
Learnable Upsampling: “Deconvolution”

Typical 3 x 3 convolution, stride 1 pad 1

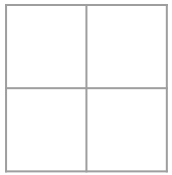


Learnable Upsampling: “Deconvolution”

Typical 3 x 3 convolution, **stride 2** pad 1



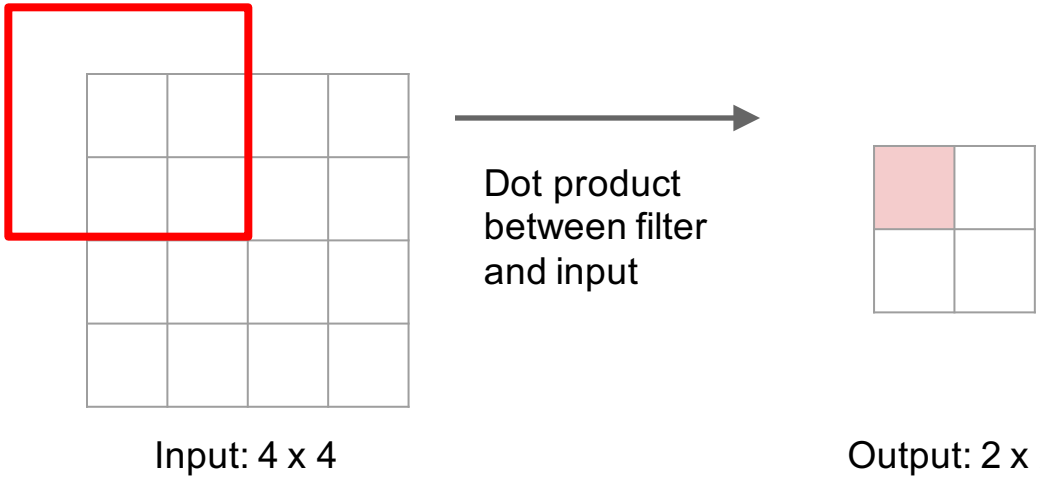
Input: 4 x 4



Output: 2 x 2

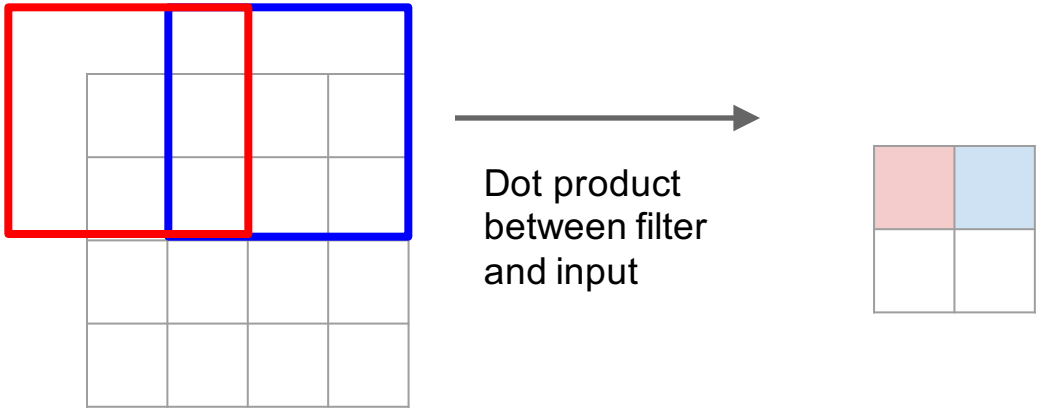
Learnable Upsampling: “Deconvolution”

Typical 3 x 3 convolution, stride 2 pad 1



Learnable Upsampling: “Deconvolution”

Typical 3 x 3 convolution, stride 2 pad 1

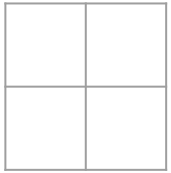


Input: 4 x 4

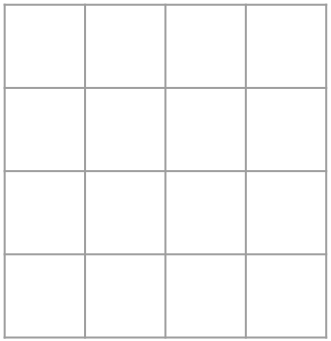
Output: 2 x 2

Learnable Upsampling: “Deconvolution”

3 x 3 “deconvolution”, stride 2 pad 1



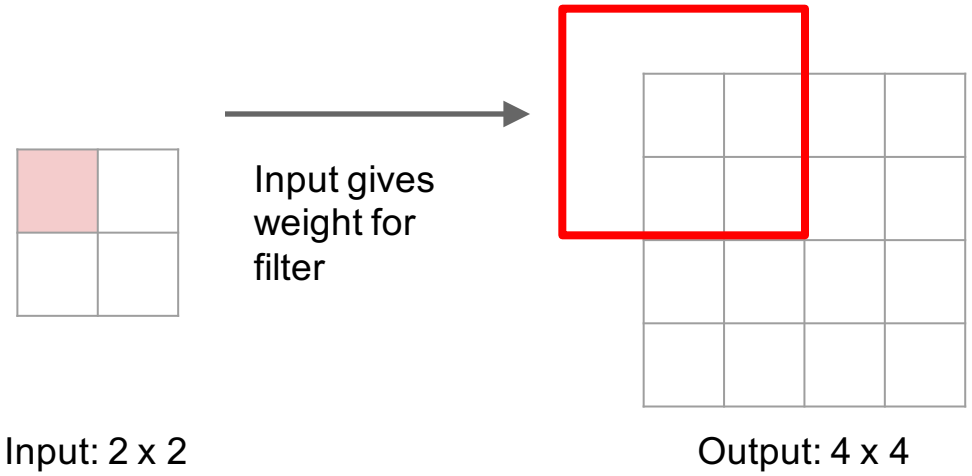
Input: 2 x 2



Output: 4 x 4

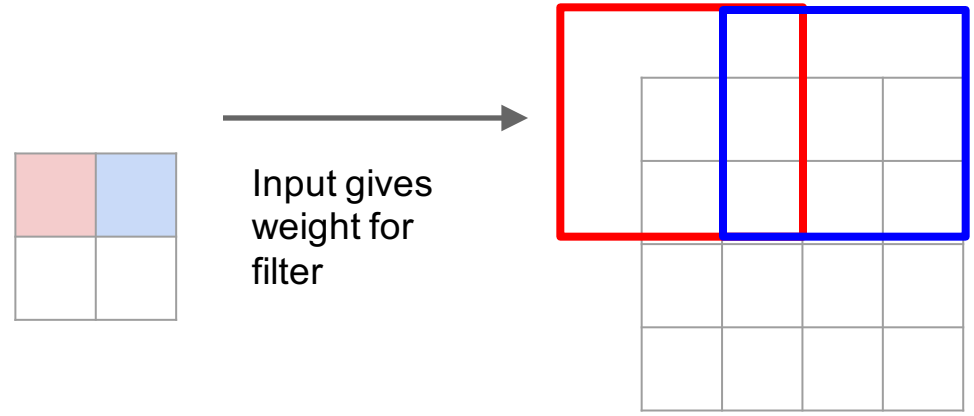
Learnable Upsampling: “Deconvolution”

3 x 3 “deconvolution”, stride 2 pad 1

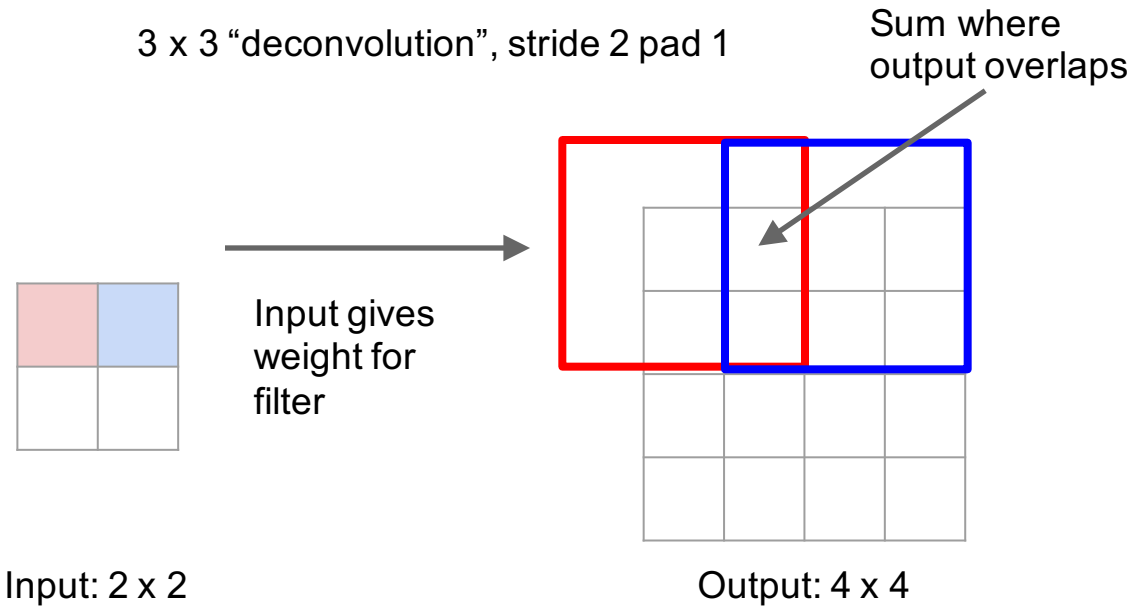


Learnable Upsampling: “Deconvolution”

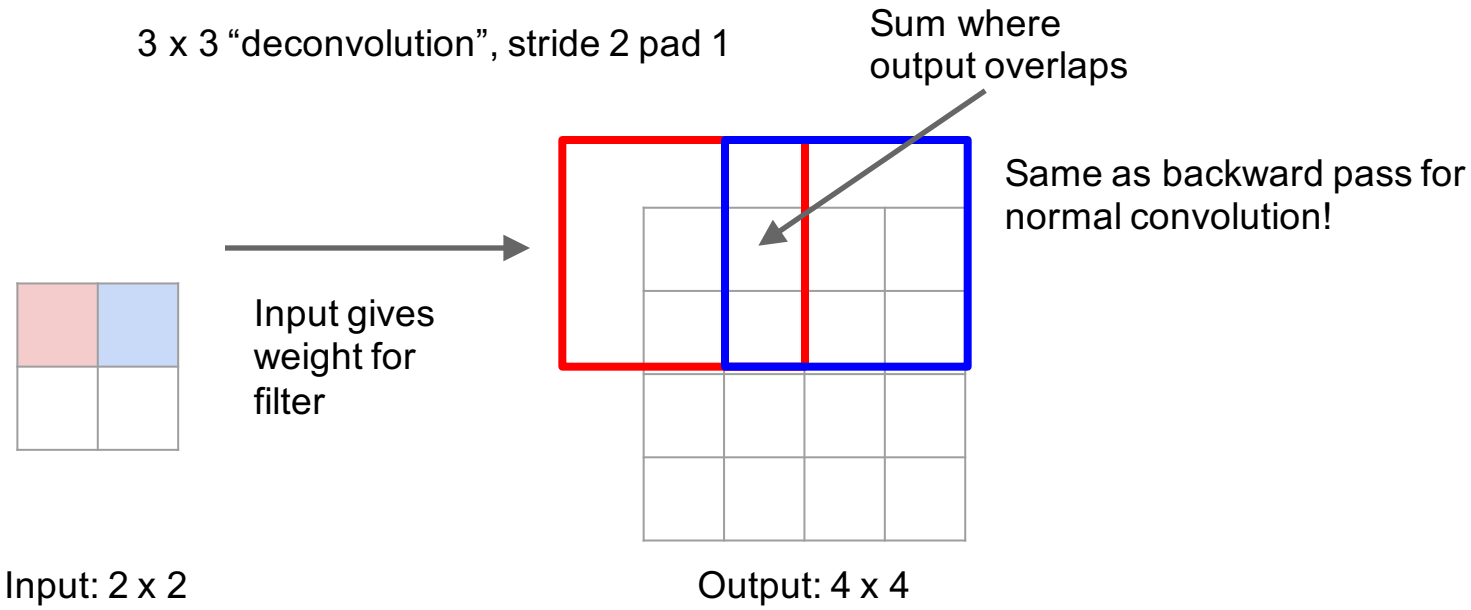
3 x 3 “deconvolution”, stride 2 pad 1



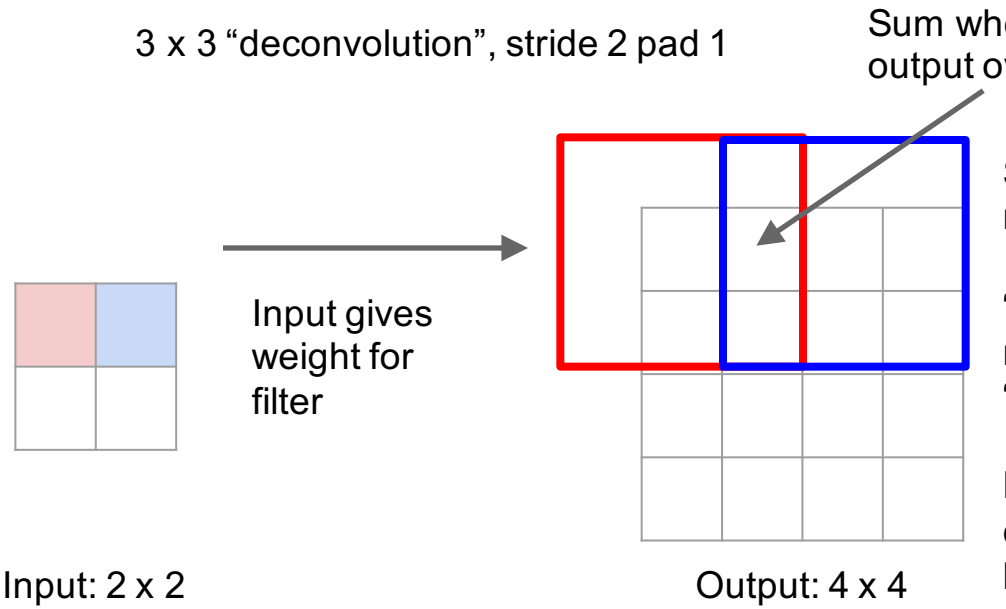
Learnable Upsampling: “Deconvolution”



Learnable Upsampling: “Deconvolution”



Learnable Upsampling: “Deconvolution”



Same as backward pass for normal convolution!

“Deconvolution” is a bad name, already defined as “inverse of convolution”

Better names:
convolution transpose, backward strided convolution, 1/2 strided convolution, upconvolution

Learnable Upsampling: "Deconvolution"

[1]It is more proper to say "convolutional transpose operation" rather than "deconvolutional" operation. Hence, we will be using the term "convolutional transpose" from now.

Im et al, "Generating images with recurrent adversarial networks", arXiv 2016

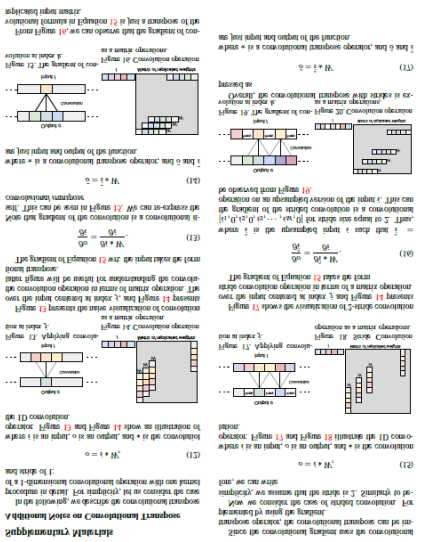
A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions)

Radford et al, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", ICLR 2016

"Deconvolution" is a bad name, already defined as "inverse of convolution"

Better names:
convolution transpose,
backward strided convolution,
1/2 strided convolution,
upconvolution

Learnable Upsampling: "Deconvolution"



Great explanation in appendix

[1]It is more proper to say "convolutional transpose operation" rather than "deconvolutional" operation. Hence, we will be using the term "convolutional transpose" from now.

Im et al, "Generating images with recurrent adversarial networks", arXiv 2016

A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions)

Radford et al, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", ICLR 2016

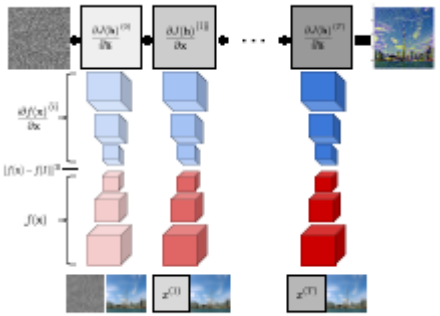


Figure 21. The gradient of convolution at index k.

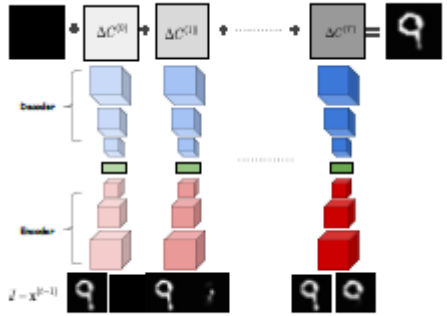
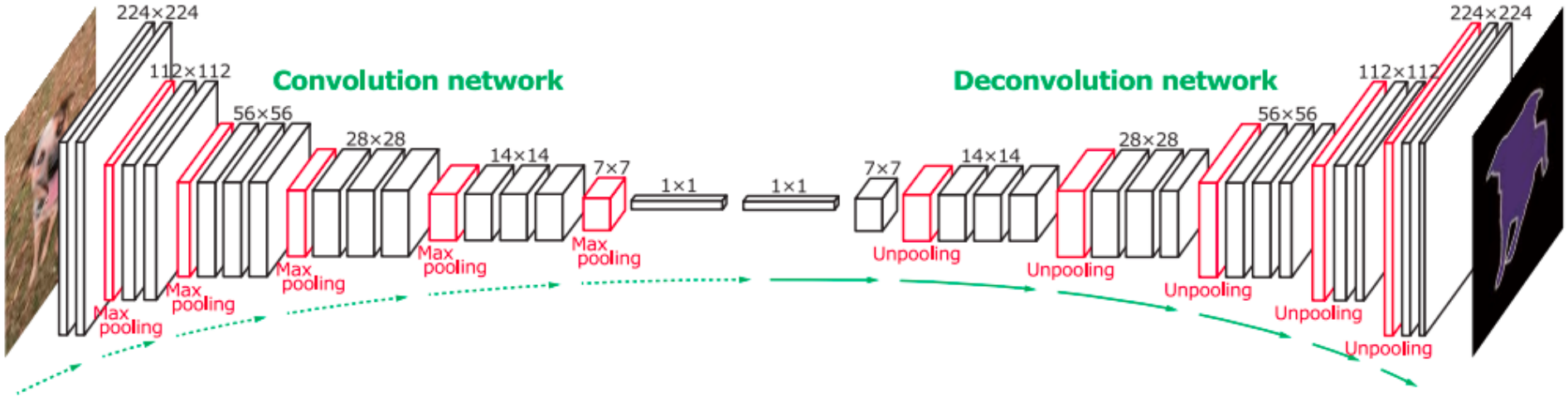


Figure 22. The abstract view of DRAW architecture is delineated.

"Deconvolution" is a bad name, already defined as "inverse of convolution"

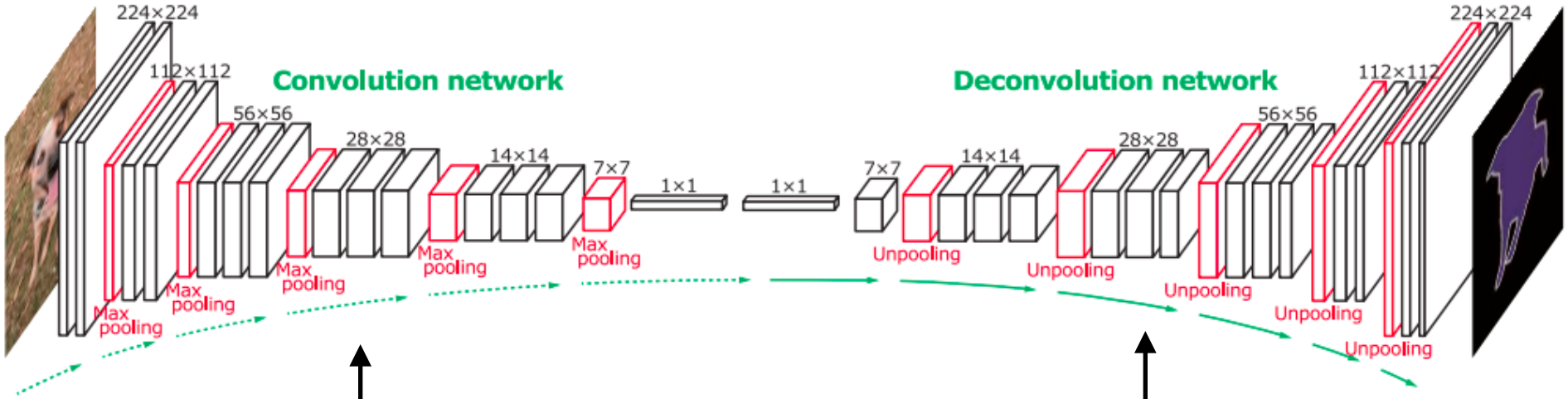
Better names:
convolution transpose,
backward strided convolution,
1/2 strided convolution,
upconvolution

Semantic Segmentation: Upsampling



Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Semantic Segmentation: Upsampling



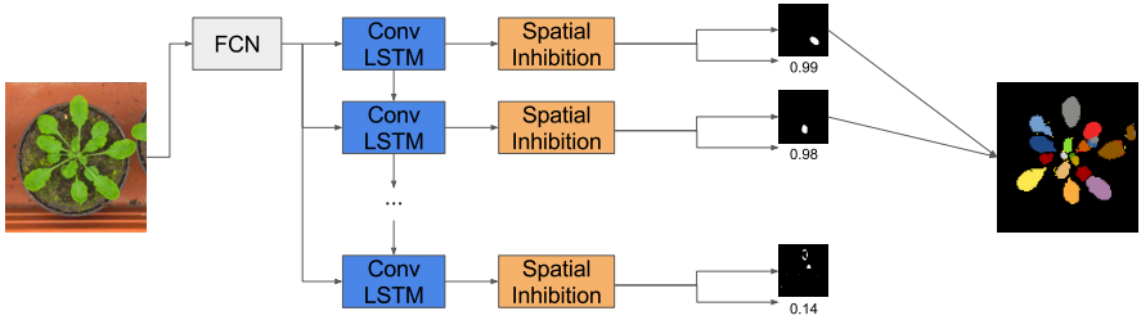
Normal VGG

"Upside down" VGG

Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

6 days of training on Titan X...

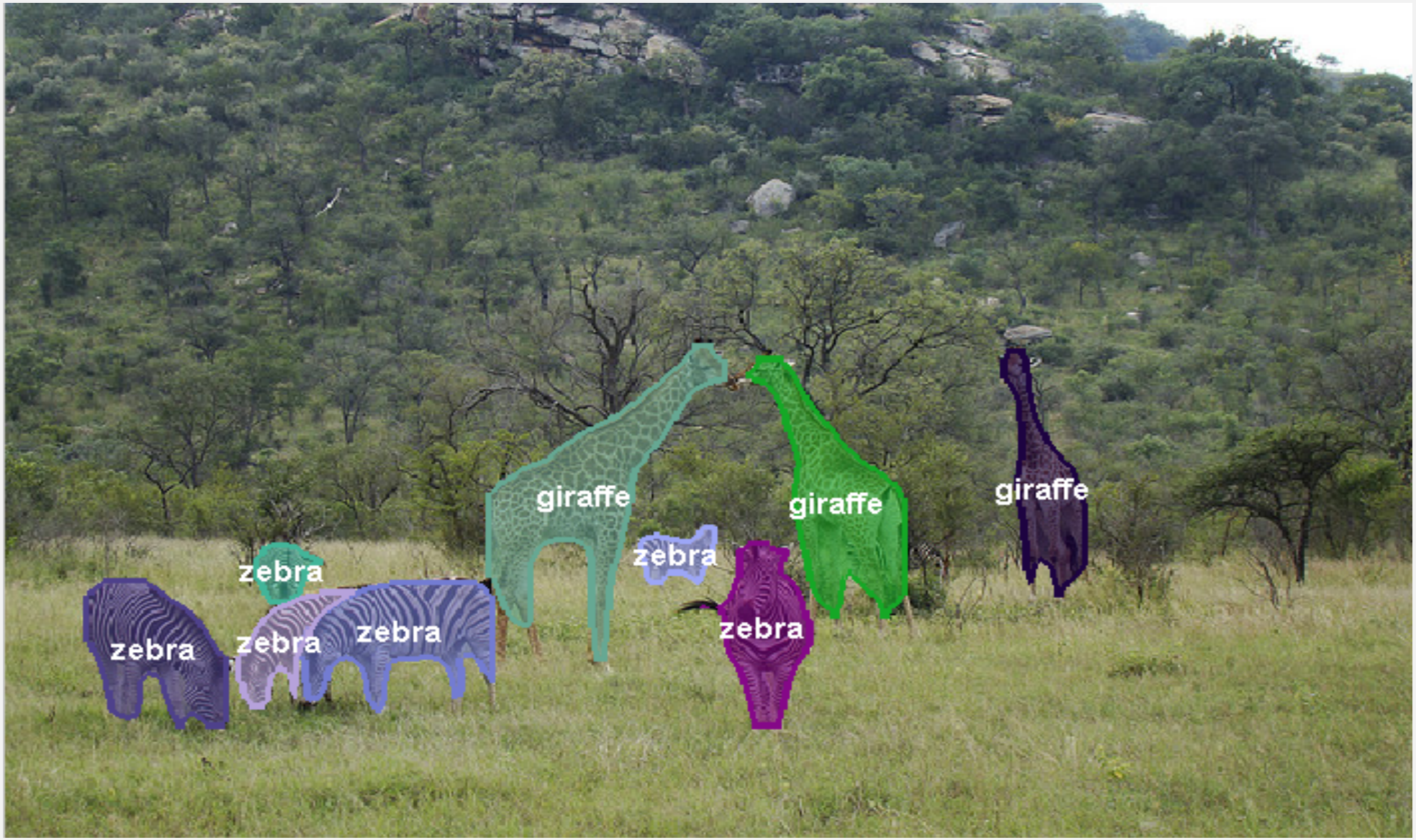
PART 2



Instance Segmentation

- Semantic segmentation research has recently witnessed rapid progress, but many leading methods are unable to identify object instances.

Difference


giraffe
giraffe
giraffe
zebra
zebra
zebra
zebra
zebra
zebra
zebra

Instance Segmentation

Detect instances, give category, label pixels

"simultaneous detection and segmentation" (SDS)

Lots of recent work (MS-COCO)

detect all instances of a category in an image and, for each instance, mark the pixels that belong to it

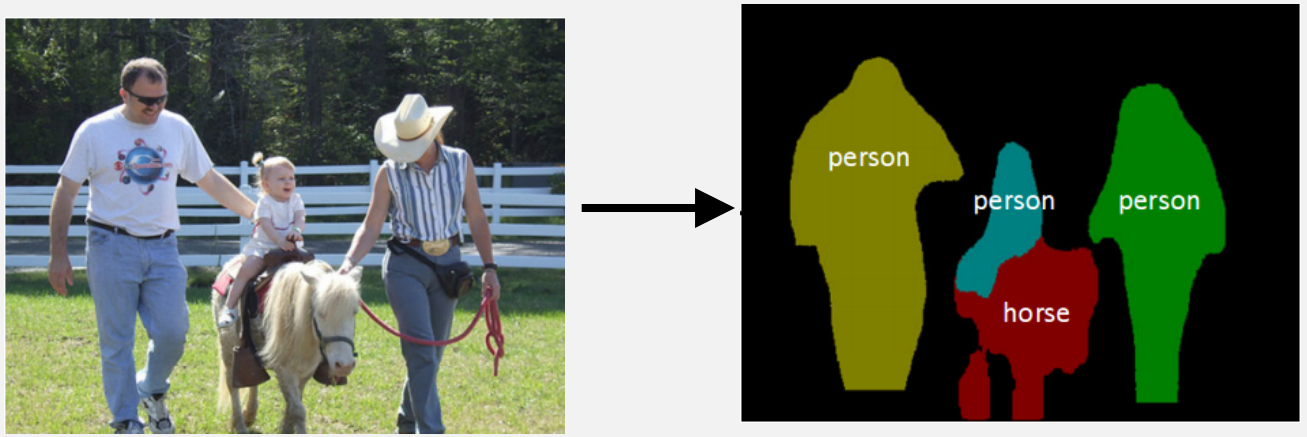
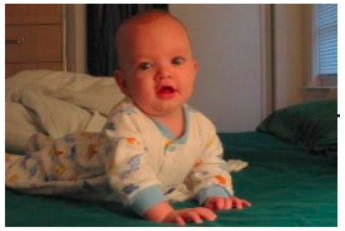


Figure credit: Dai et al, "Instance-aware Semantic Segmentation via Multi-task Network Cascades", arXiv 2015

Instance Segmentation

Similar to R-CNN, but with segments



Hariharan et al, "Simultaneous Detection and Segmentation", ECCV 2014

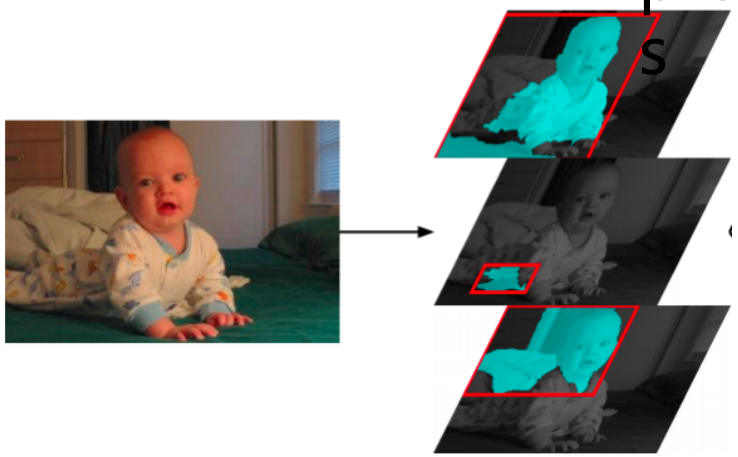
Instance Segmentation

Similar to R-CNN, but with segments

region proposals: 2000 region candidates

Proposal Generation

External Segment proposal s



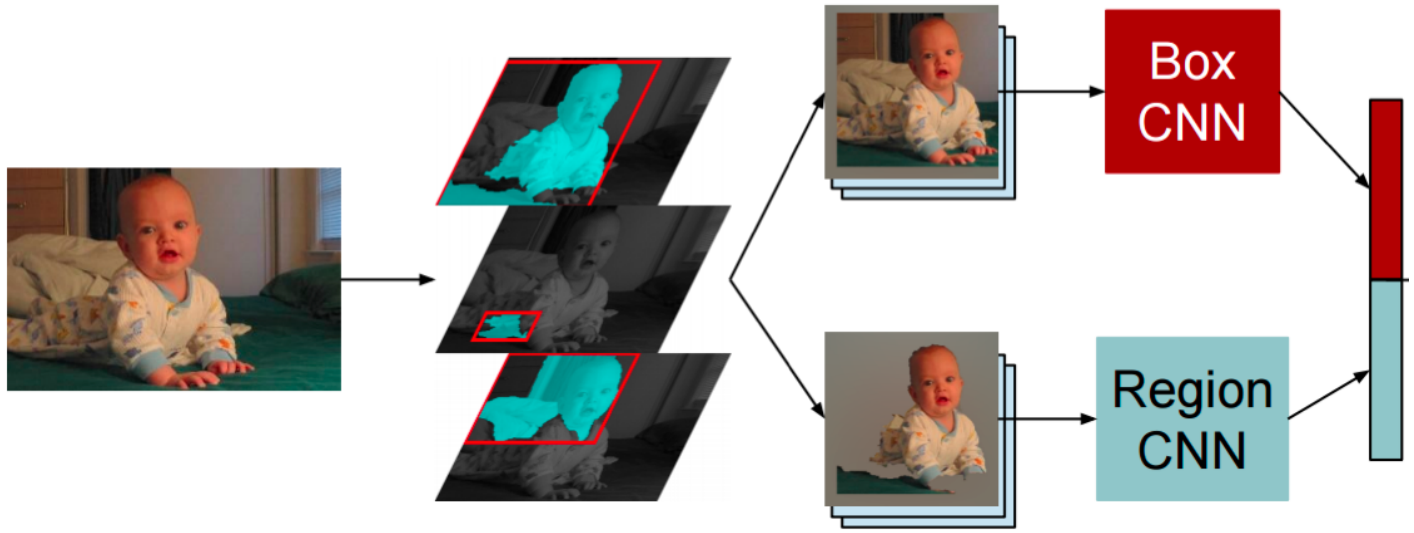
Hariharan et al, "Simultaneous Detection and Segmentation", ECCV 2014

Instance Segmentation

Similar to R-CNN, but with segments

External Segment proposals

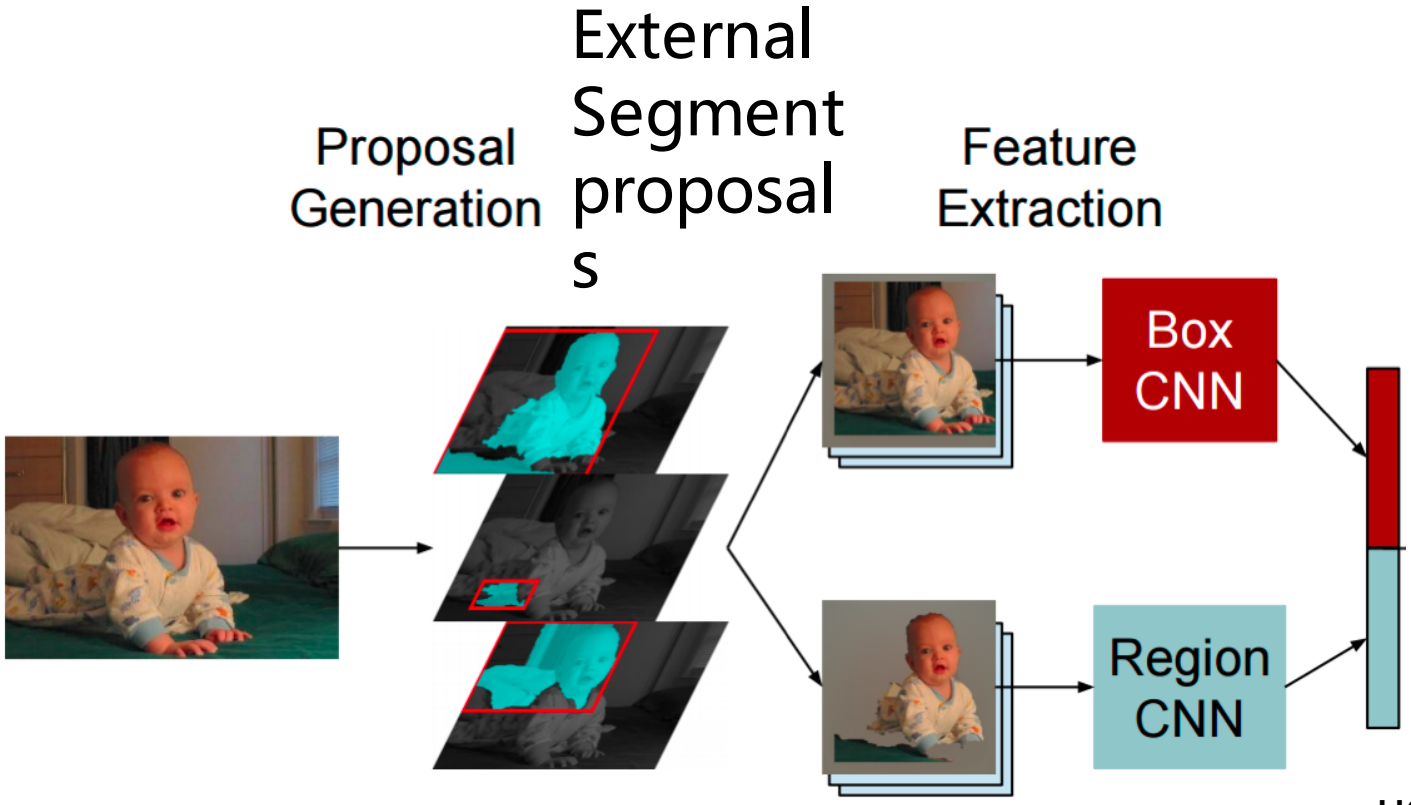
both the bounding box of the region as well as from the region foreground.



Hariharan et al, "Simultaneous Detection and Segmentation", ECCV 2014

Instance Segmentation

Similar to R-CNN, but with segments



Disjoint but same architecture different weights but train together

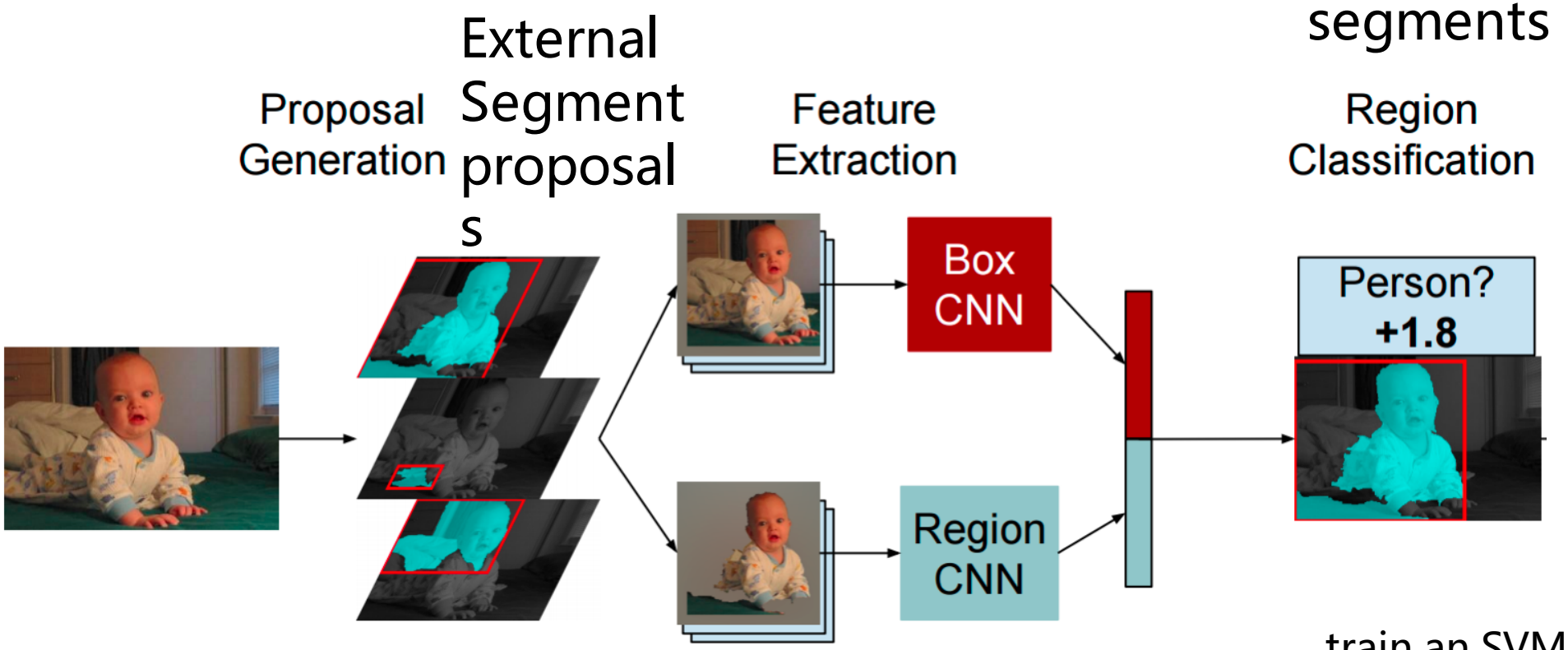
Mask out background with mean image

using separate networks where each network is finetuned for its respective role dramatically improves performance.

Hariharan et al, "Simultaneous Detection and Segmentation", ECCV 2014

Instance Segmentation

Similar to R-CNN, but with segments



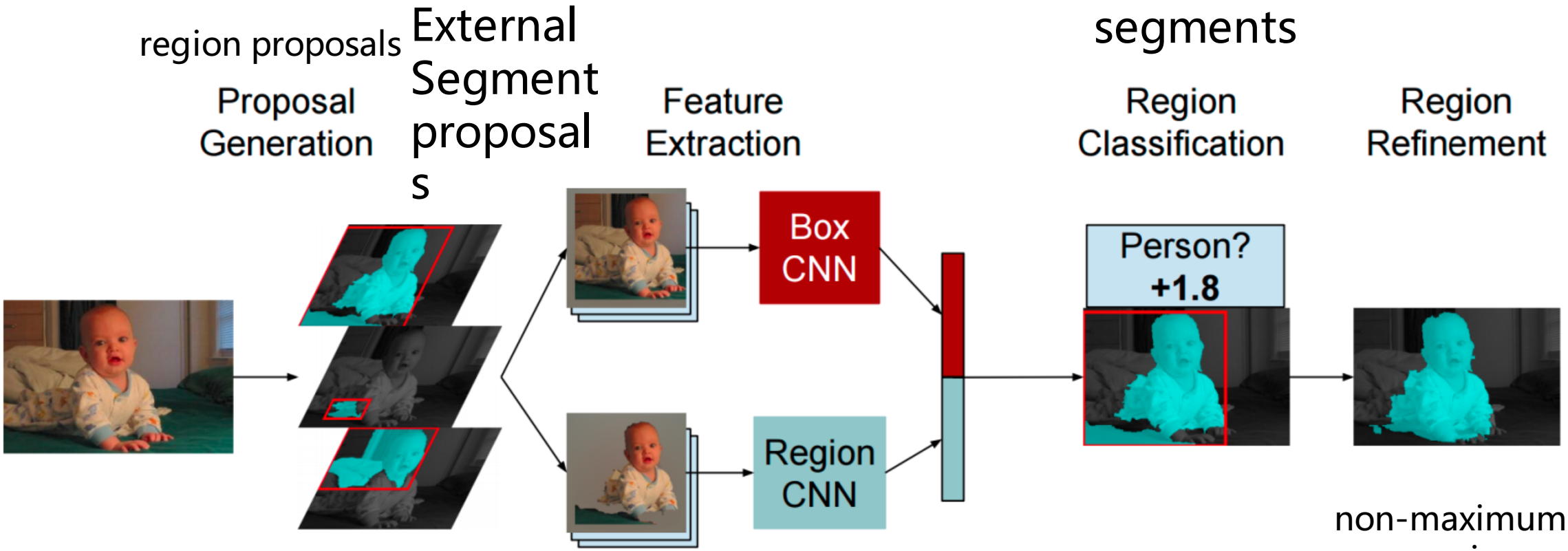
Mask out background with mean image

train an SVM

Hariharan et al, "Simultaneous Detection and Segmentation", ECCV 2014

Instance Segmentation

Similar to R-CNN, but with segments

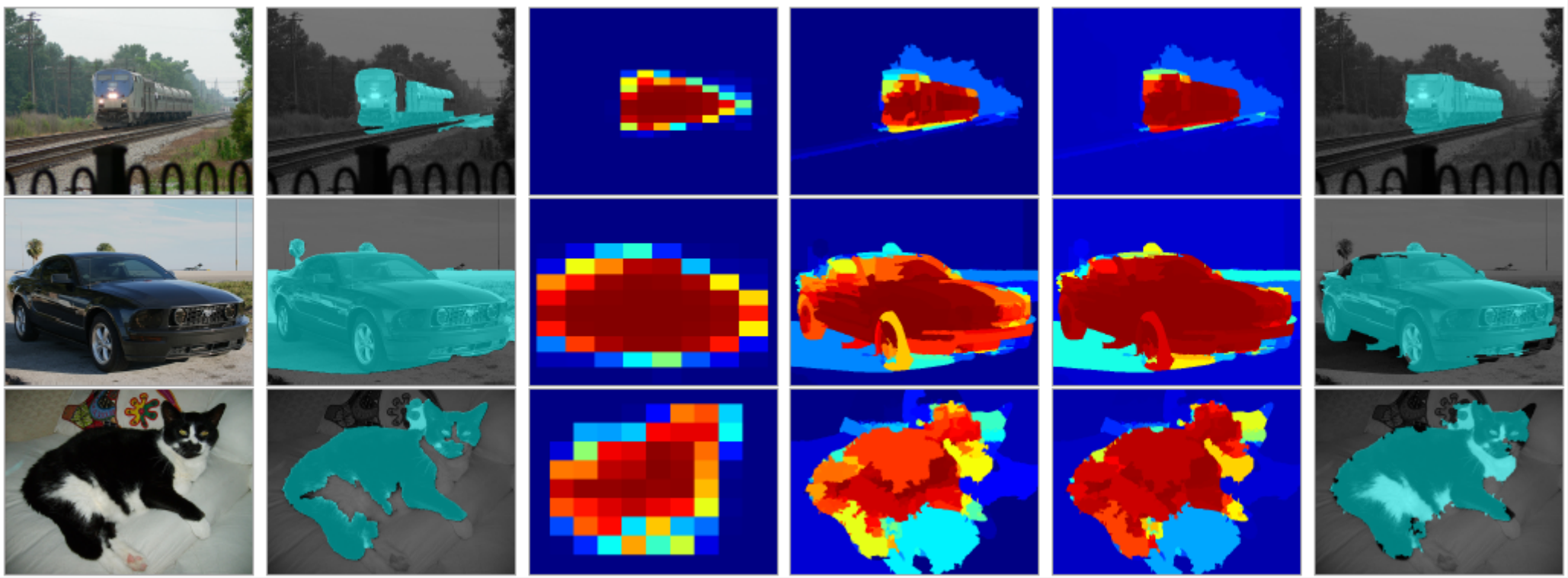


Hariharan et al, "Simultaneous Detection and Segmentation", ECCV 2014

region refinement

examples

Refinement uses top-down category specific information to fill in the body of the train and the cat and remove the road from the car.



image

original region

coarse 10 × 10 mask

projected to superpixels

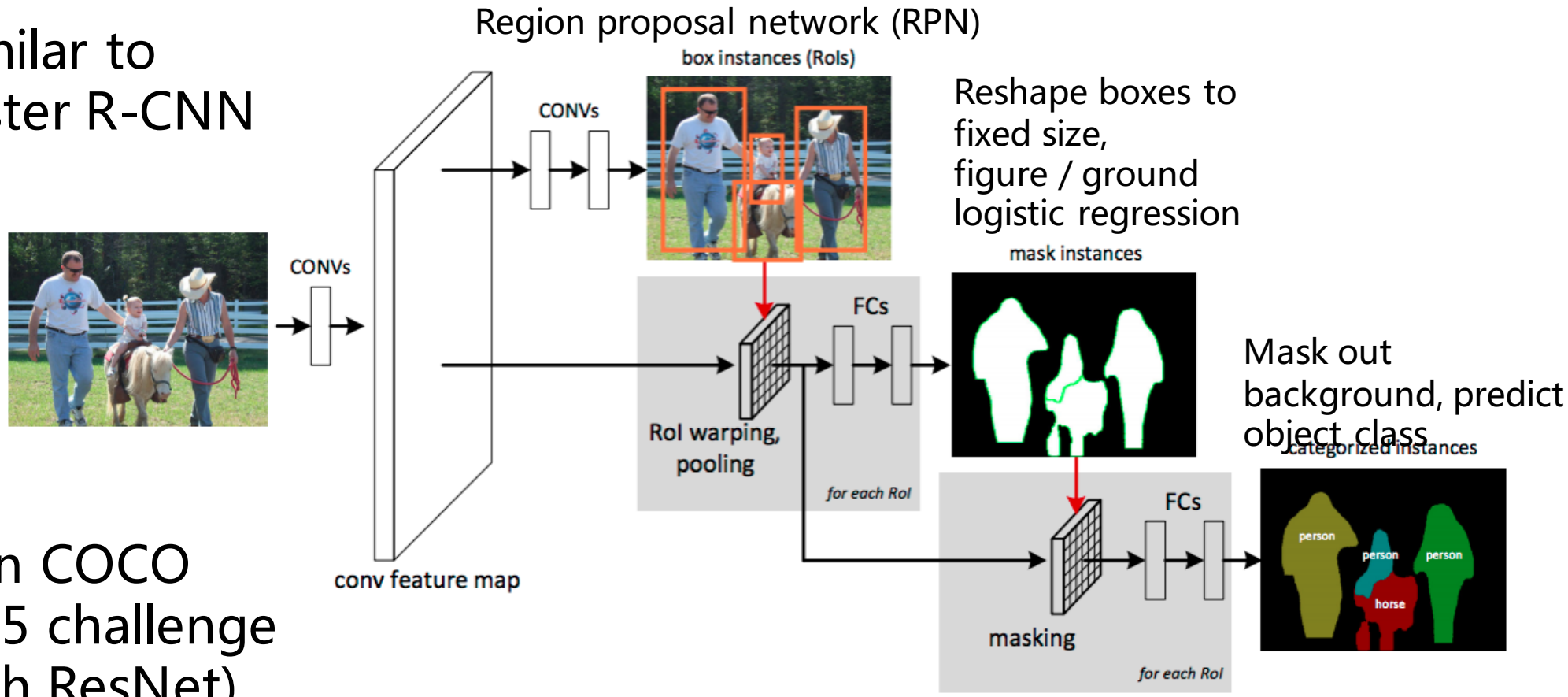
superpixels

final classifier

after thresholding

Instance Segmentation: Cascades

Similar to
Faster R-CNN

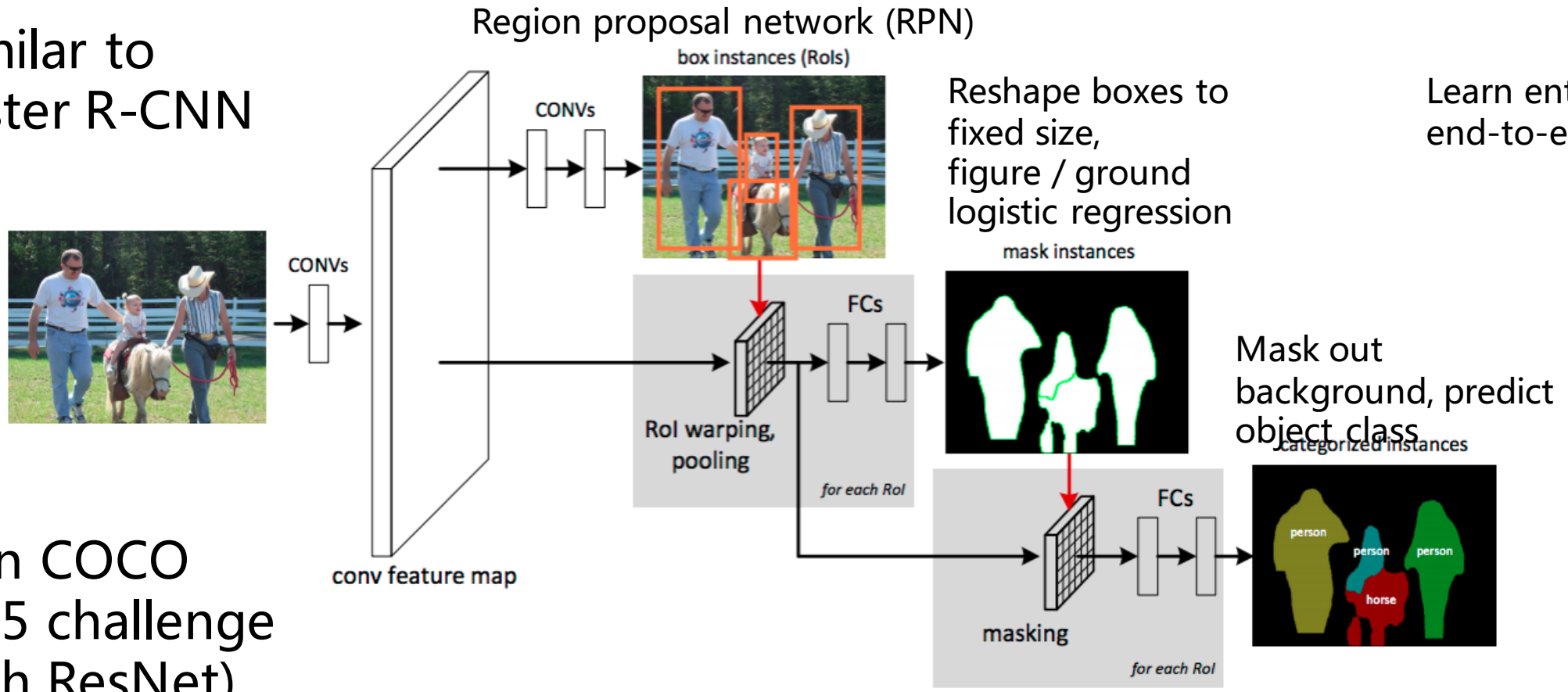


Won COCO
2015 challenge
(with ResNet)

Dai et al, "Instance-aware Semantic Segmentation via Multi-task Network Cascades", arXiv 2015

Instance Segmentation: Cascades

Similar to Faster R-CNN

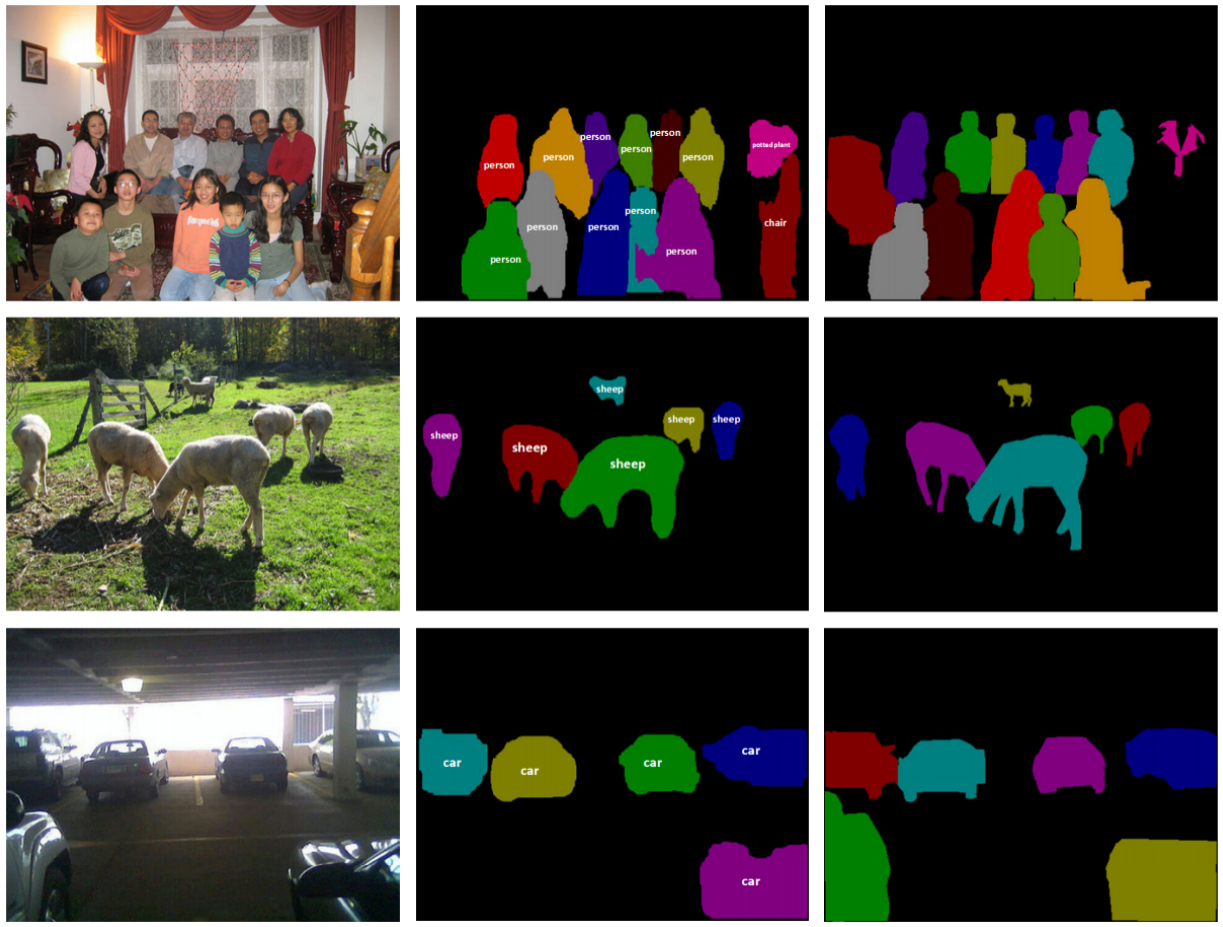


Learn entire model end-to-end!

Won COCO 2015 challenge (with ResNet)

Dai et al, "Instance-aware Semantic Segmentation via Multi-task Network Cascades", arXiv 2015

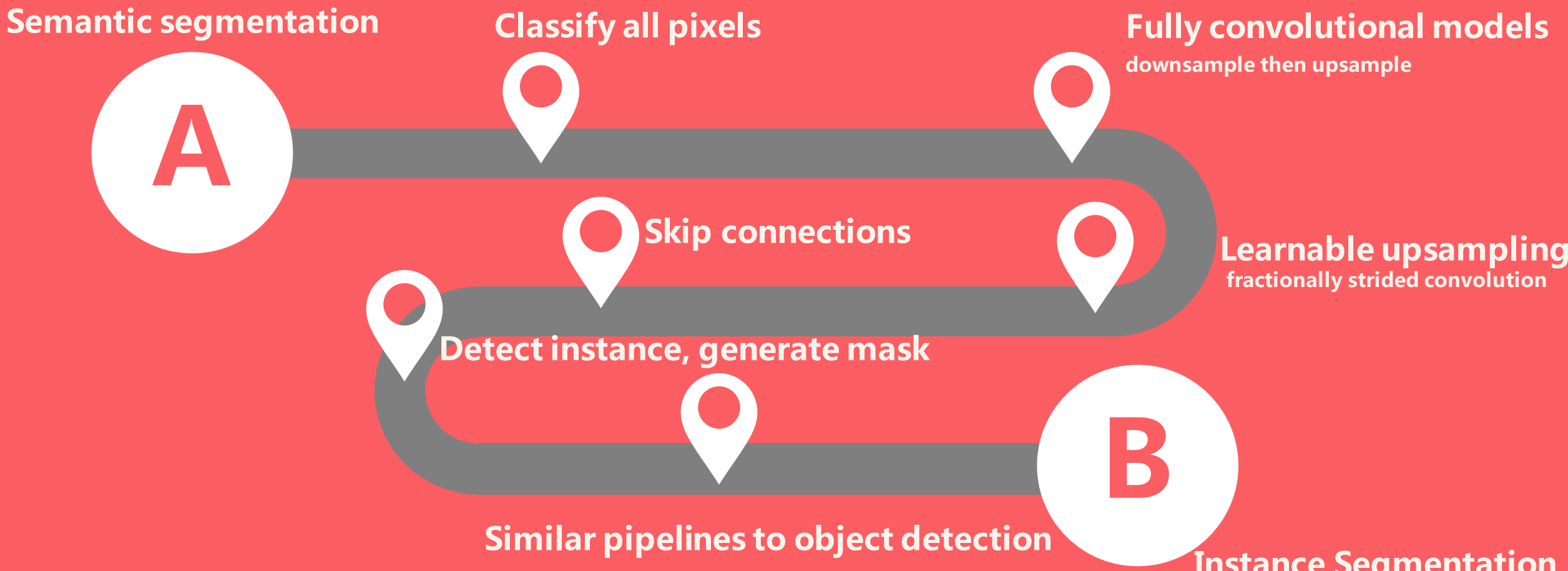
Instance Segmentation: Cascades



Dai et al, "Instance-aware Semantic Segmentation via Multi-task Network Cascades", arXiv 2015

Predictions **Ground truth**

Segmentation Overview




THANK YOU!
PRESENTED BY DeepLearningPKU