Insight
Centre for Data Analytics

DEEP
LEARNING
WORKSHOP

Dublin City University
21-22 May 2018

Day 2 Lecture 6

Segmentation

Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

Segmentation

Segmentation

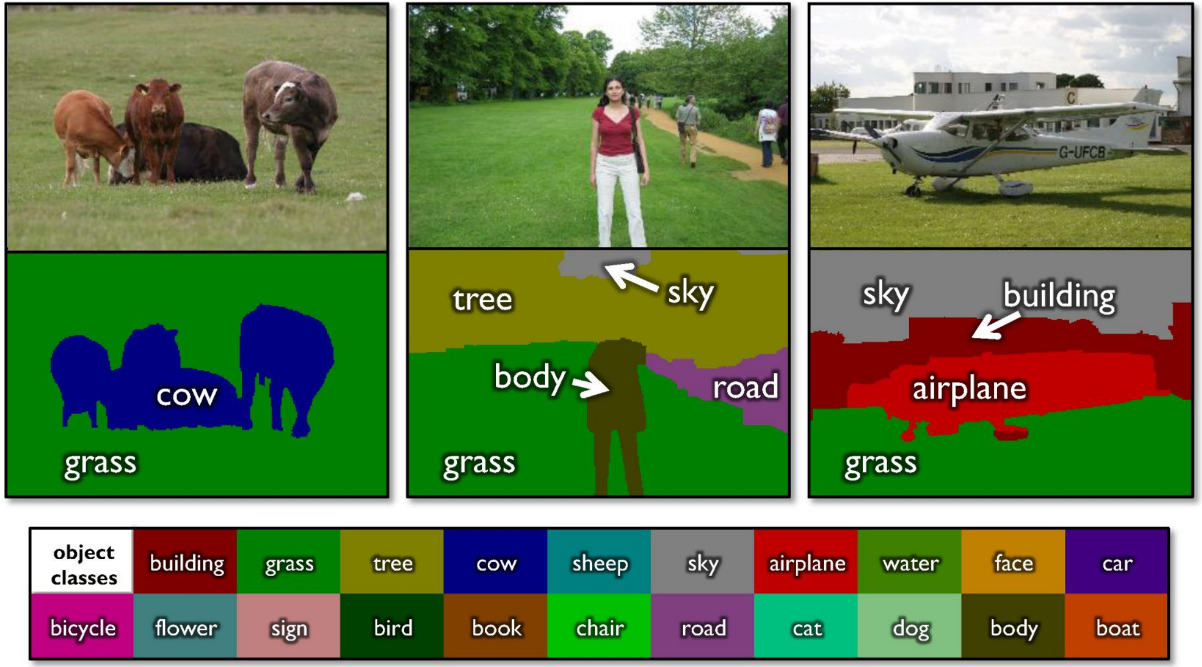
Define the accurate boundaries of all objects in an image

Semantic Segmentation

Label every pixel!

Don’t differentiate instances (cows)

Classic computer vision problem



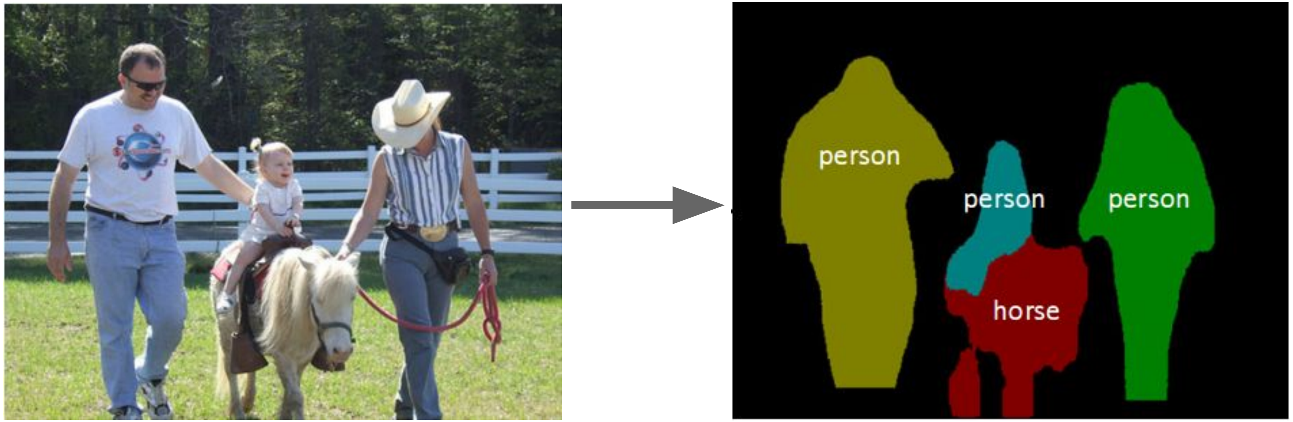
Slide Credit: CS231n

Instance Segmentation

Detect instances, give category, label pixels

“simultaneous detection and segmentation” (SDS)

Label are class-aware and instance-aware



Slide Credit: [CS231n]

Outline

Segmentation Datasets

Semantic Segmentation Methods

- Deconvolution (or transposed convolution)
- Dilated Convolution
- Skip Connections

Instance Segmentation Methods

- Proposal-Based
- Recurrent
- Metric Learning

Outline

Segmentation Datasets

Semantic Segmentation Methods

- Deconvolution (or transposed convolution)
- Dilated Convolution
- Skip Connections

Instance Segmentation Methods

- Proposal-Based
- Recurrent
- Metric Learning

Segmentation: Datasets

Pascal Visual Object Classes

- 20 categories
- +10,000 images
- Semantic segmentation GT
- Instance segmentation GT

Pascal Context

- Real indoor & outdoor scenes
- 540 categories
- +10,000 images
- Dense annotations
- Semantic segmentation GT
- Objects + stuff

Segmentation: Datasets

ADE20K

- Real general scenes
- +150 categories
- +22,000 images
- Semantic segmentation GT
- Instance + parts segmentation GT
- Objects and stuff

COCO Common Objects in Context

- Real indoor & outdoor scenes
- 80 categories
- +300,000 images
- 2M instances
- Partial annotations
- Semantic segmentation GT
- Instance segmentation GT
- Objects, but no stuff

Segmentation: Datasets

CityScapes

- Real driving scenes
- 30 categories
- +25,000 images
- 20,000 partial annotations
- 5,000 dense annotations
- Semantic segmentation GT
- Instance segmentation GT
- Depth, GPS and other metadata
- Objects and stuff

Mapillary Vistas Dataset

- Real driving scenes
- 100 categories
- 25,000 images
- Semantic segmentation GT
- Instance + parts segmentation GT
- Objects and stuff

Outline

Segmentation Datasets

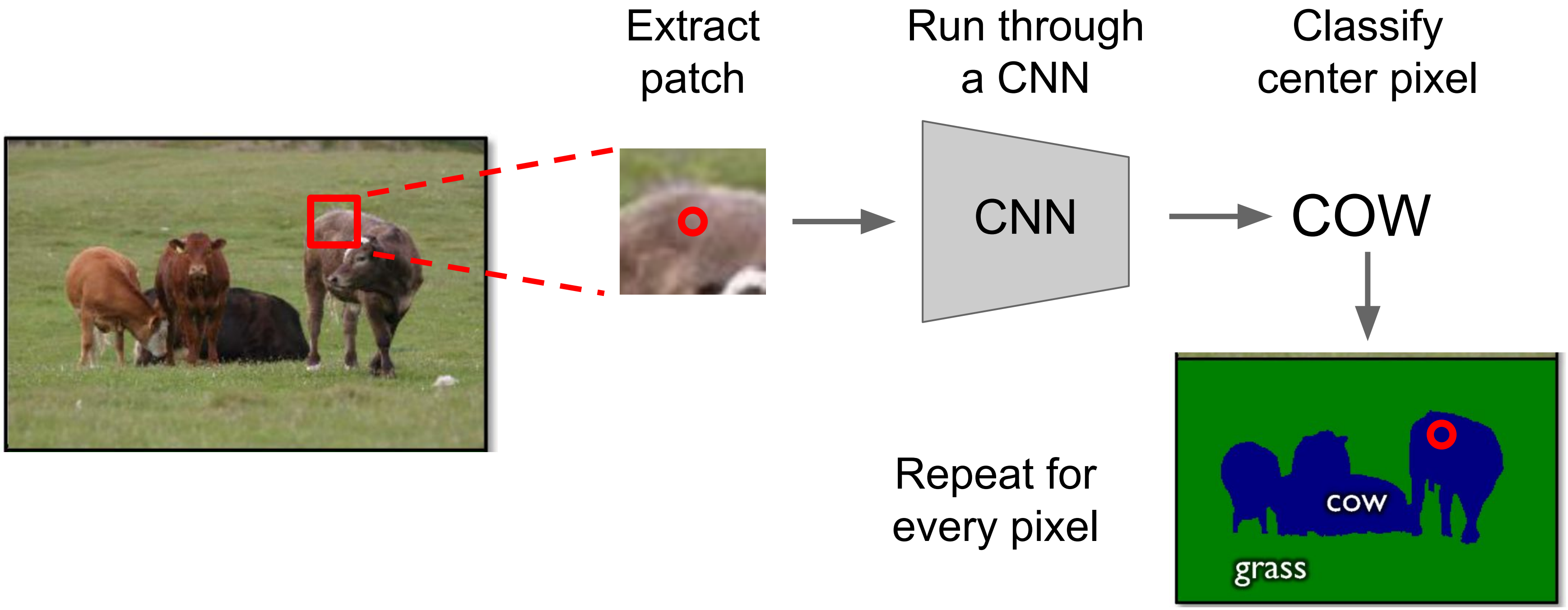
Semantic Segmentation Methods

- Deconvolution (or transposed convolution)
- Dilated Convolution
- Skip Connections

Instance Segmentation Methods

- Proposal-Based
- Recurrent
- Metric Learning

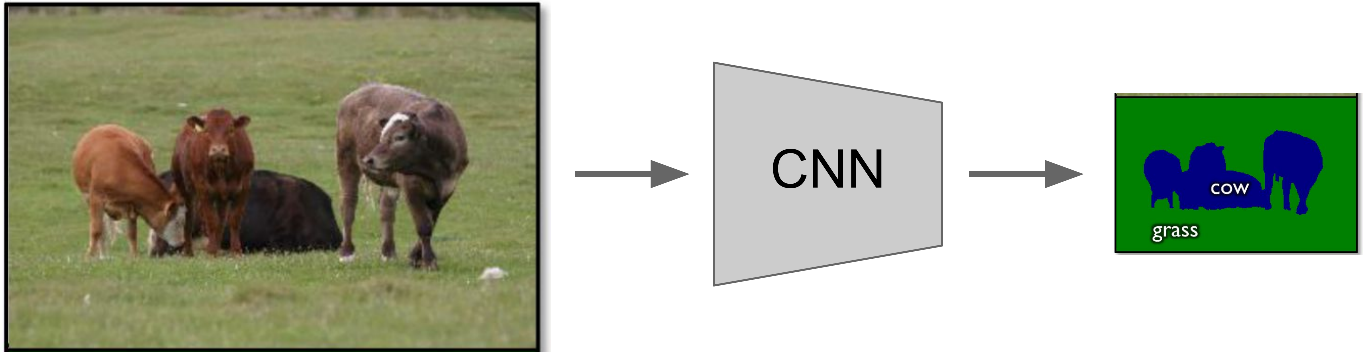
From Classification to Segmentation



Slide Credit: CS231n

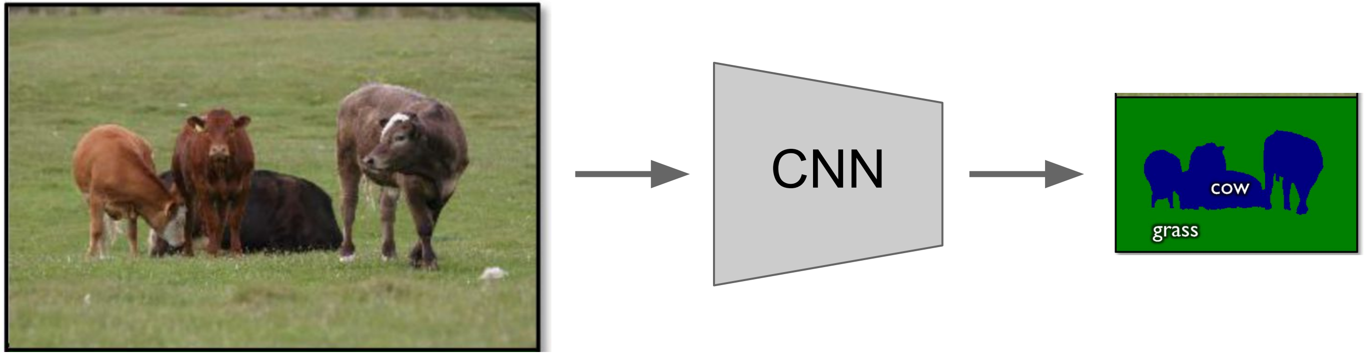
From Classification to Segmentation

Run “fully convolutional” network
to get all pixels at once



Slide Credit: CS231n

Semantic Segmentation

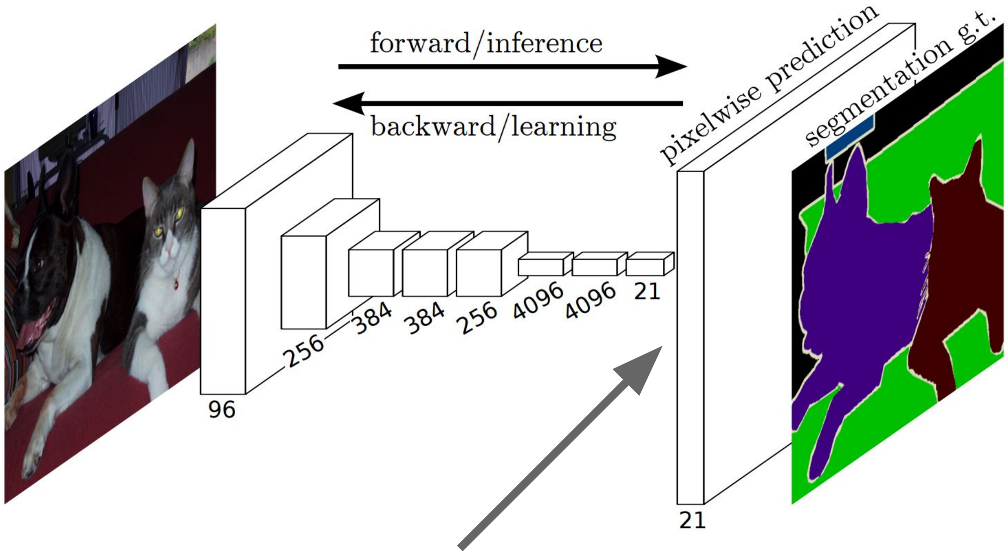


Problem 1:

Smaller output due to pooling

Slide Credit: [CS231n]

Learnable upsampling

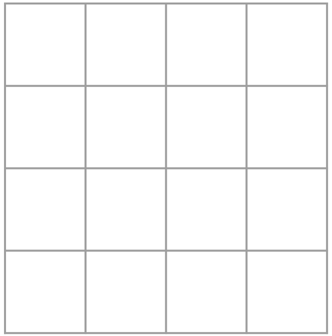


Long et al. [Fully Convolutional Networks for Semantic Segmentation.](#) CVPR 2015

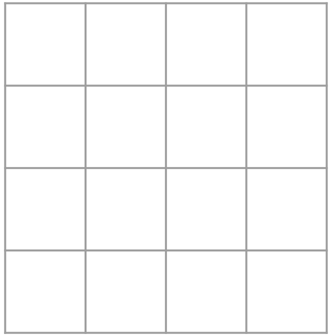
Slide Credit: [CS231n](#)

Reminder: Convolutional Layer

Typical 3 x 3 convolution, stride 1 pad 1



Input: 4 x 4

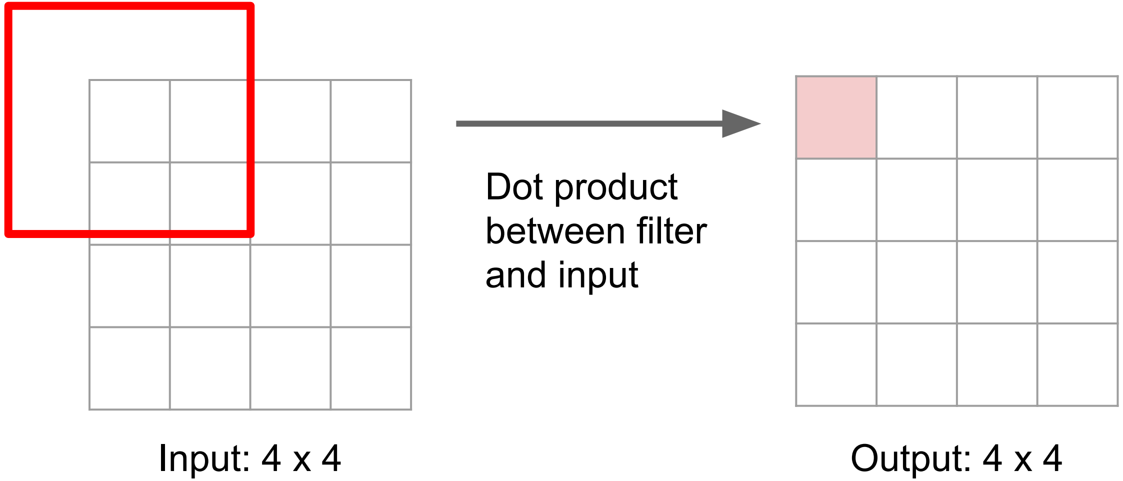


Output: 4 x 4

Slide Credit: CS231n

Reminder: Convolutional Layer

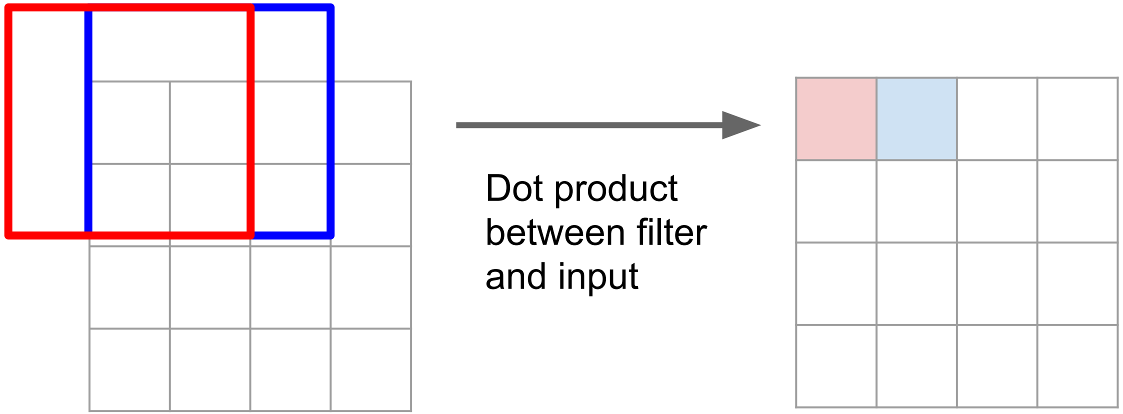
Typical 3 x 3 convolution, stride 1 pad 1



Slide Credit: [CS231n]

Reminder: Convolutional Layer

Typical 3 x 3 convolution, stride 1 pad 1



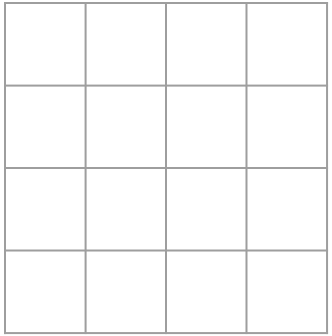
Input: 4 x 4

Output: 4 x 4

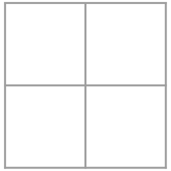
Slide Credit: [CS231n]

Reminder: Convolutional Layer

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



Input: 4 x 4

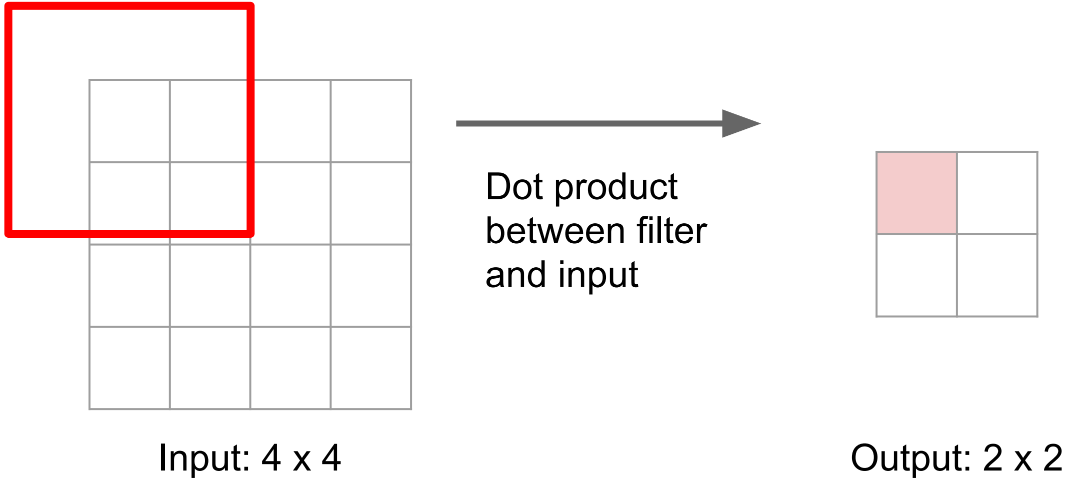


Output: 2 x 2

Slide Credit: CS231n

Reminder: Convolutional Layer

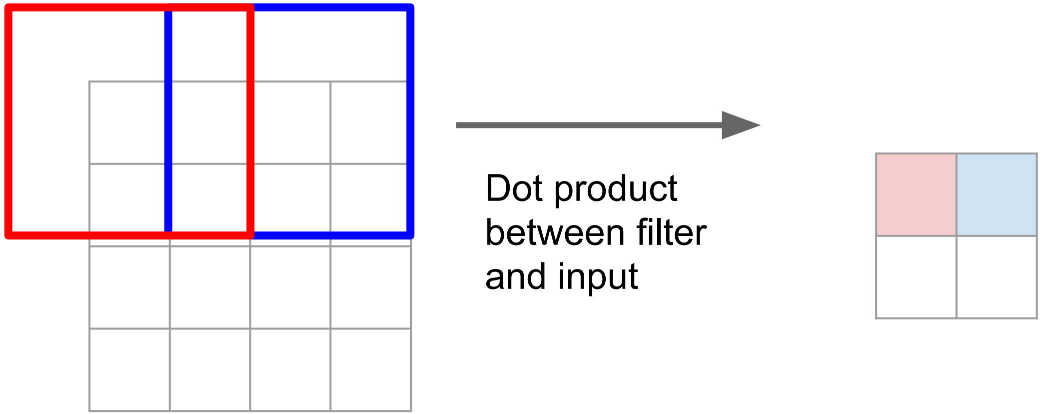
Typical 3 x 3 convolution, stride 2 pad 1



Slide Credit: [CS231n]

Reminder: Convolutional Layer

Typical 3 x 3 convolution, stride 2 pad 1



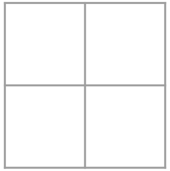
Input: 4 x 4

Output: 2 x 2

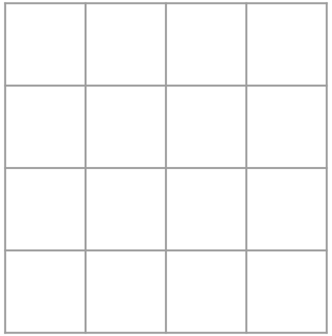
Slide Credit: CS231n

Learnable Upsample: Transposed Convolution

3 x 3 “deconvolution”, stride 2 pad 1



Input: 2 x 2

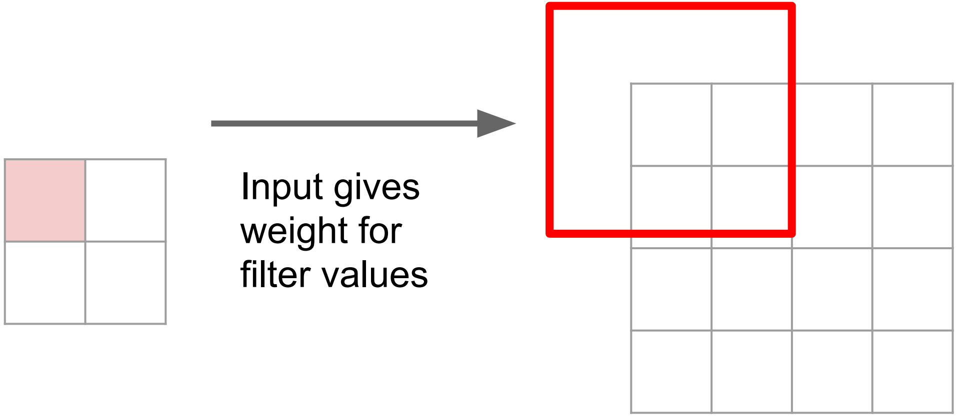


Output: 4 x 4

Slide Credit: [CS231n]

Learnable Upsample: Transposed Convolution

3 x 3 “deconvolution”, stride 2 pad 1



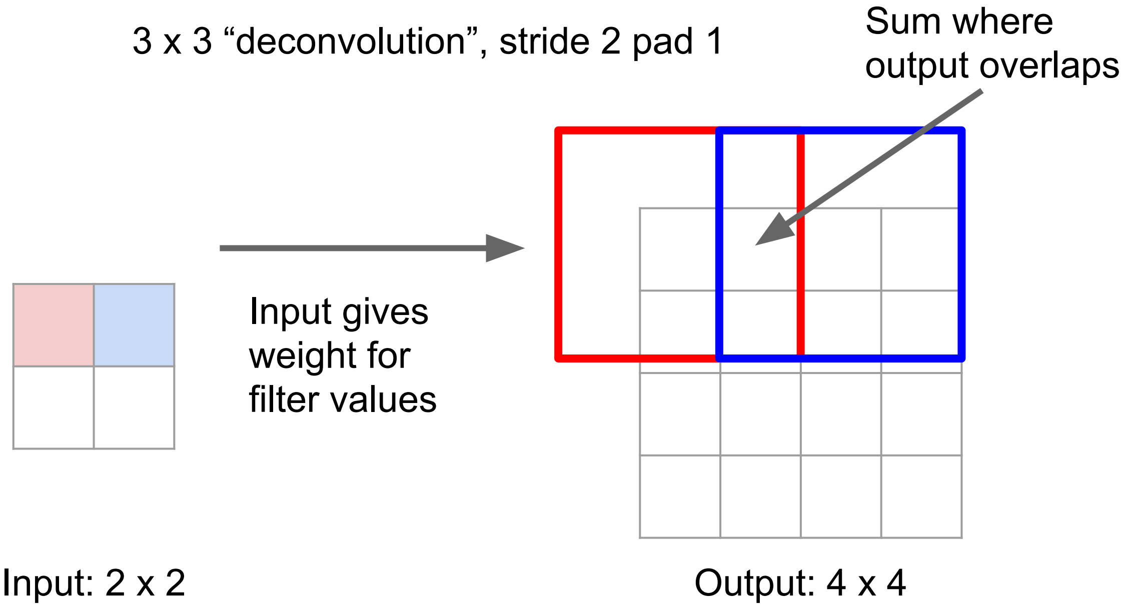
Input: 2 x 2

Output: 4 x 4

Slide Credit: [CS231n]

Learnable Upsample: Transposed Convolution

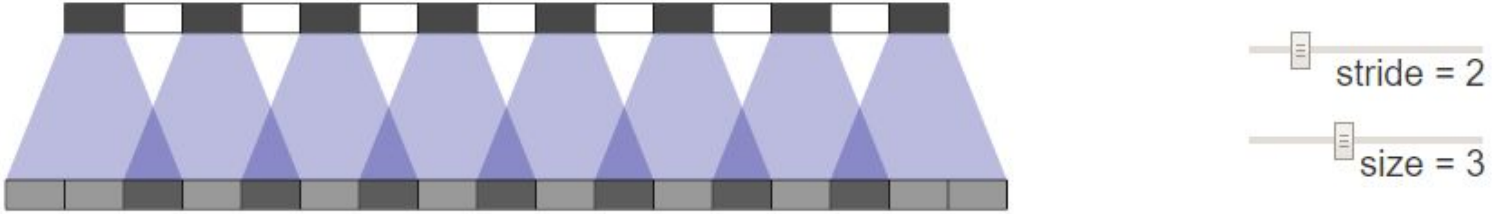
3 x 3 "deconvolution", stride 2 pad 1

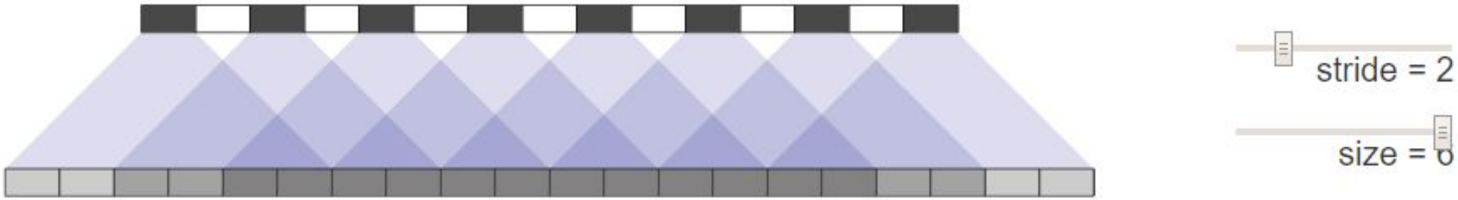


Slide Credit: [CS231n]

Learnable Upsample: Transposed Convolution

Warning: Checkerboard effect when kernel size is not divisible by the stride

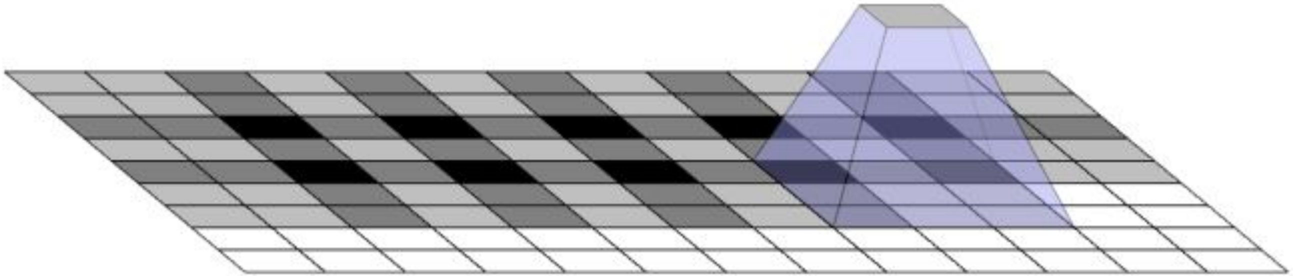




Source: [distill.pub]

Learnable Upsample: Transposed Convolution

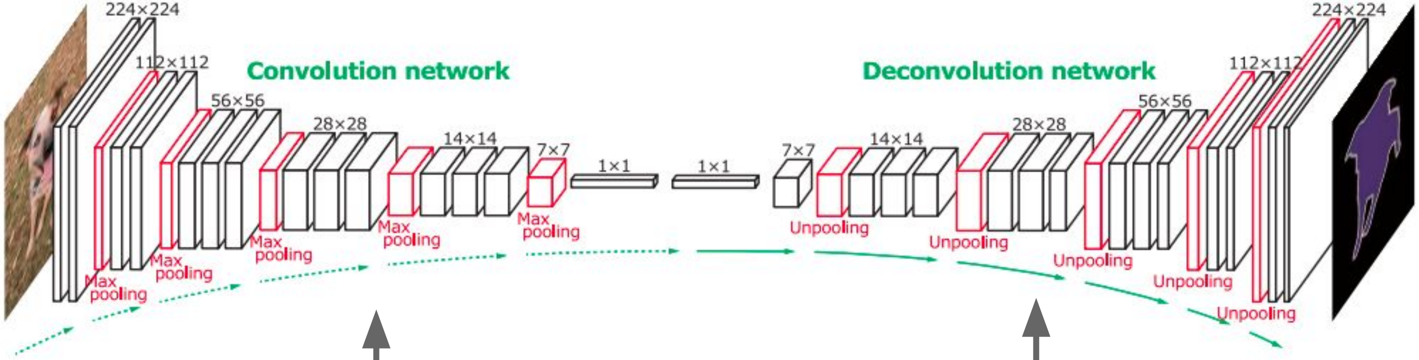
Warning: Checkerboard effect when kernel size is not divisible by the stride



stride = 2, kernel_size = 3

Source: distill.pub

Learnable Upsample: Transposed Convolution



“Regular” VGG

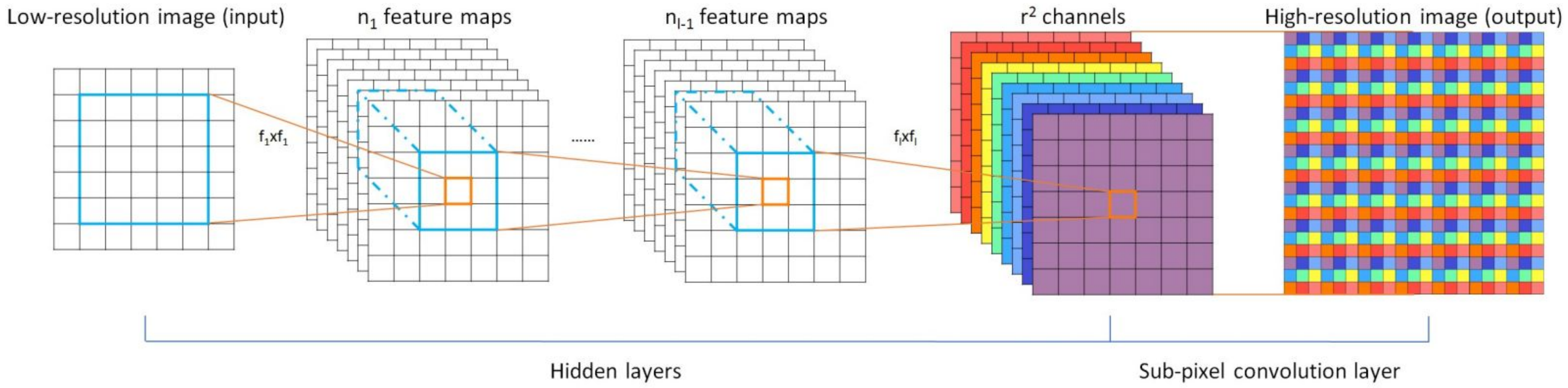
“Upside down” VGG

Noh et al. [Learning Deconvolution Network for Semantic Segmentation](). ICCV 2015

Slide Credit: [CS231n]()

Alternative to Transposed Convolution: Subpixel

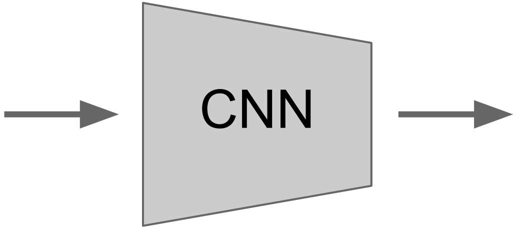
Rearrange features in previous convolutional layer to form a higher resolution output

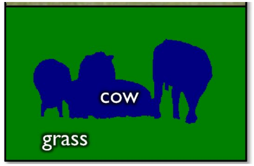


Shi et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. CVPR 2016

Semantic Segmentation

Problem 2:



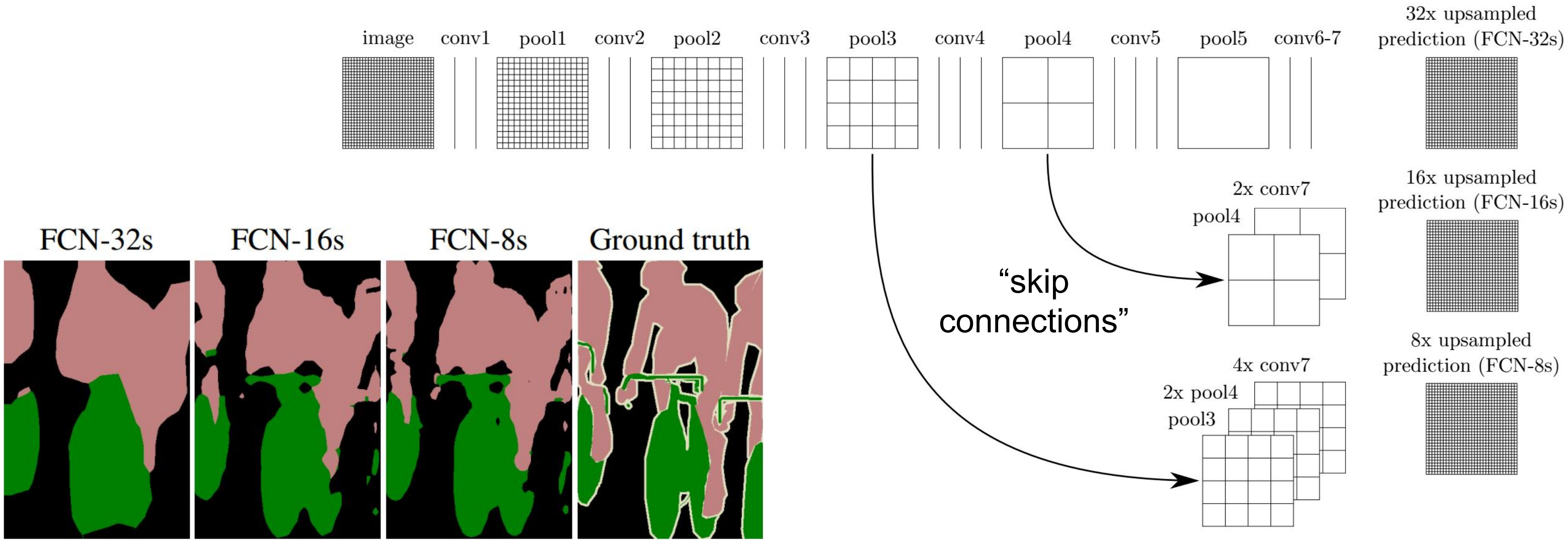


Coarse output

High-level features (e.g. conv5 layer) from a pretrained classification network are the input for the segmentation branch

Skip Connections

Recovering low level features from early layers



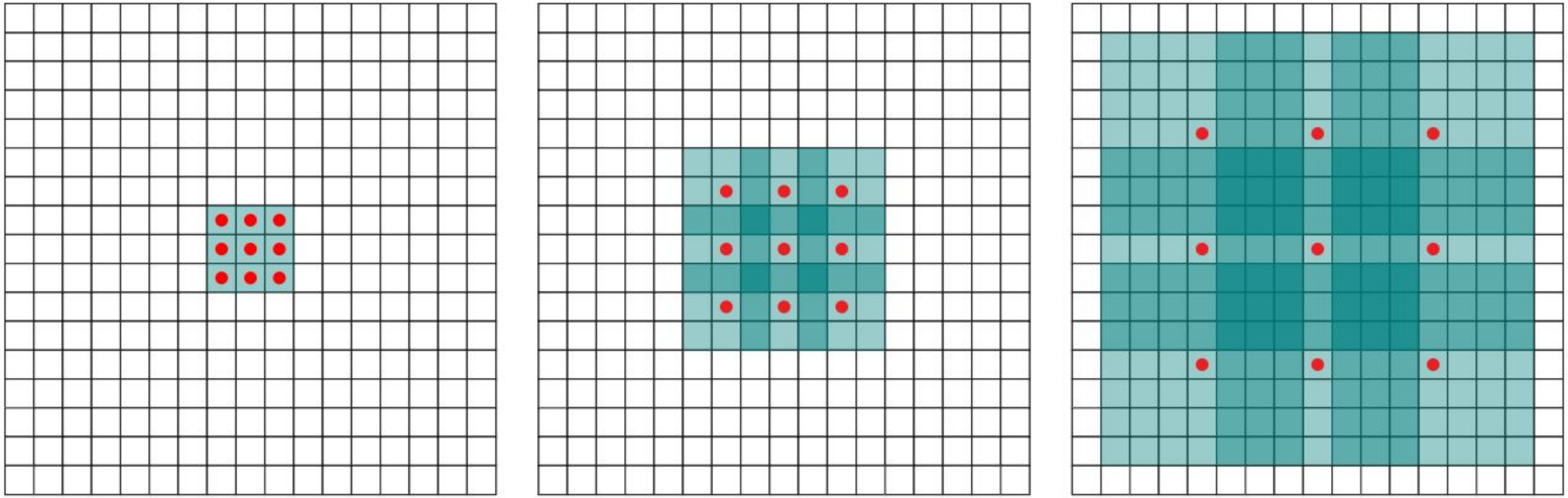
Skip connections = Better results

Long et al. [Fully Convolutional Networks for Semantic Segmentation.](#) CVPR 2015

Slide Credit: [CS231n](#)

Dilated Convolutions

Structural change in convolutional layers for dense prediction problems (e.g. image segmentation)

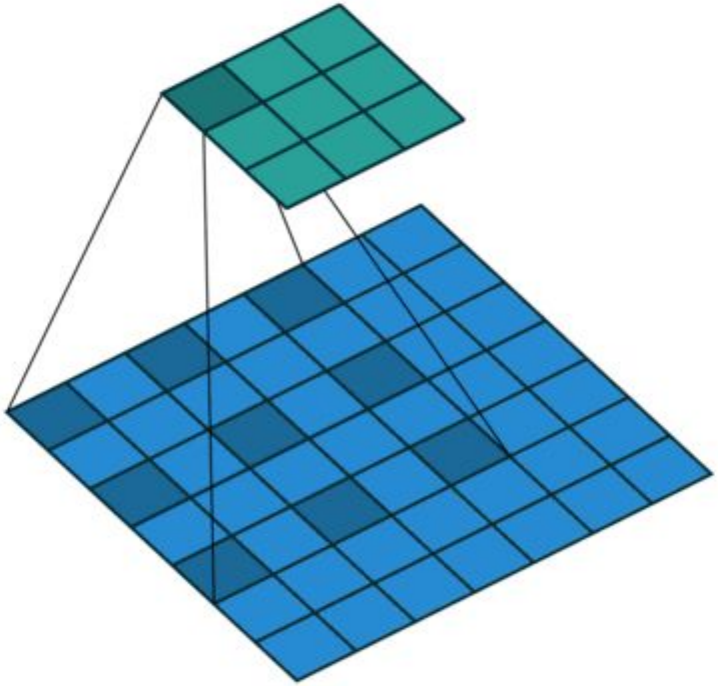


(a) (b) (c)

- The receptive field grows exponentially as you add more layers → more context information in deeper layers wrt regular convolutions
- Number of parameters increases linearly as you add more layers

Yu & Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. ICLR 2016

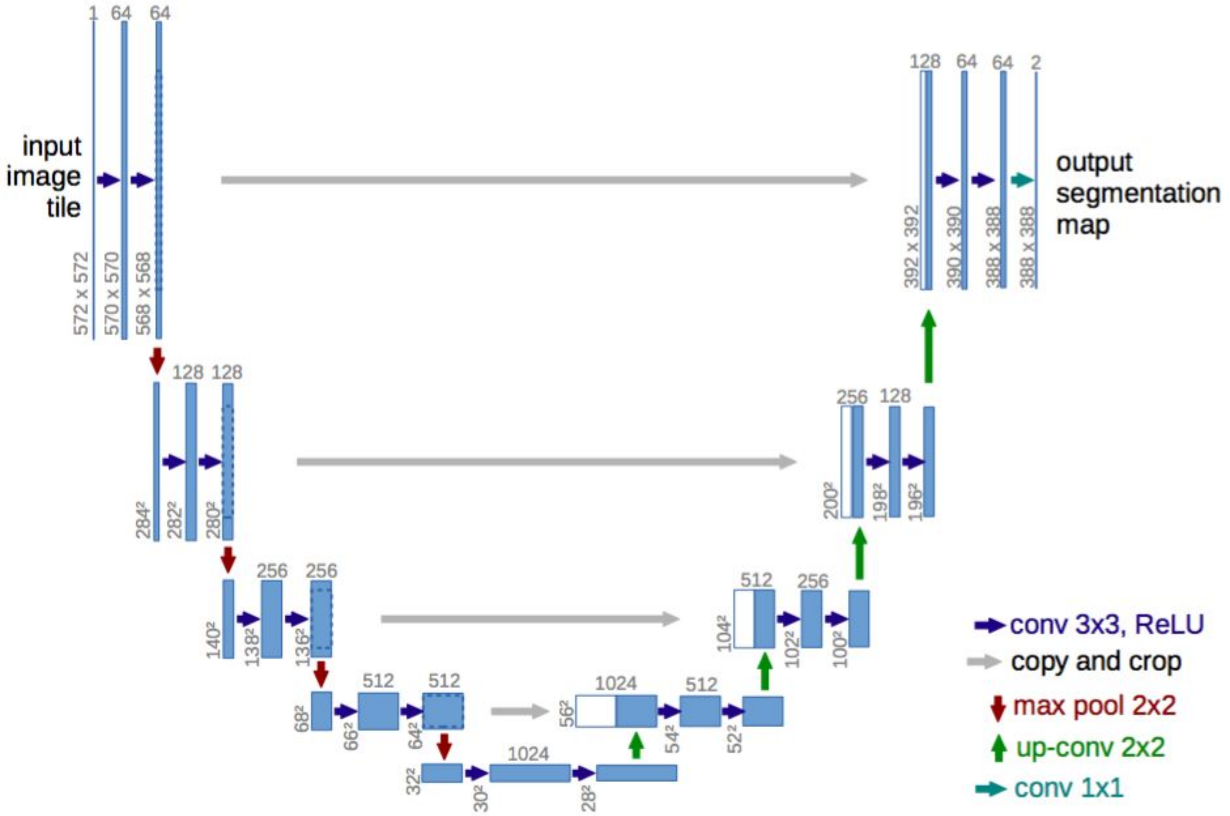
Dilated Convolutions



Source: https://github.com/vdumoulin/conv_arithmetic

State-of-the-art models

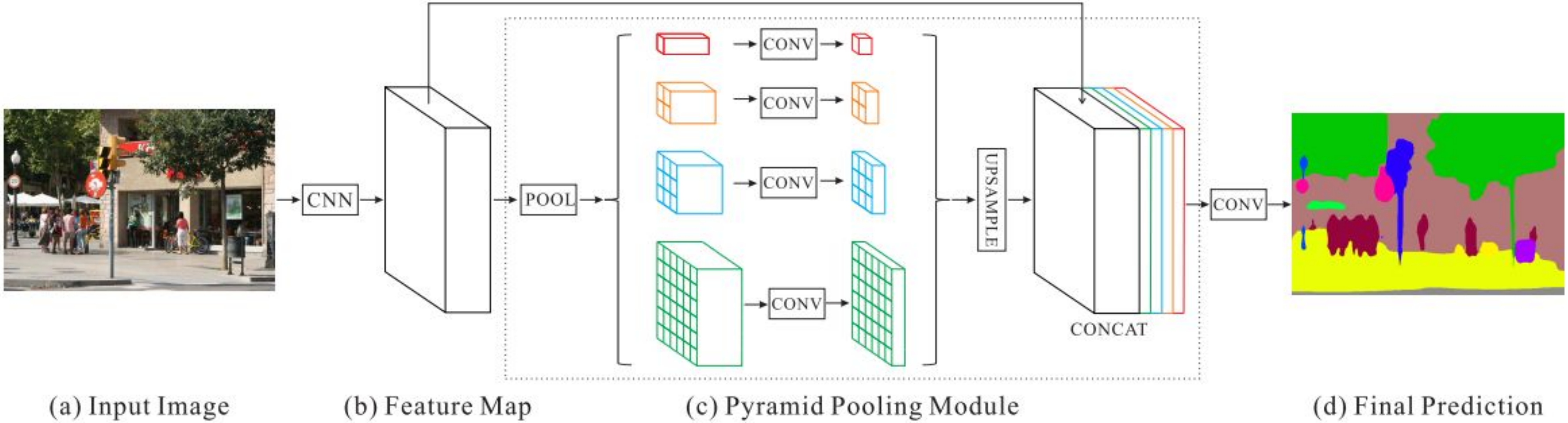
- U-Net
  - Deconvolutions
  - skip connections



Ronneberger et al. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015

State-of-the-art models

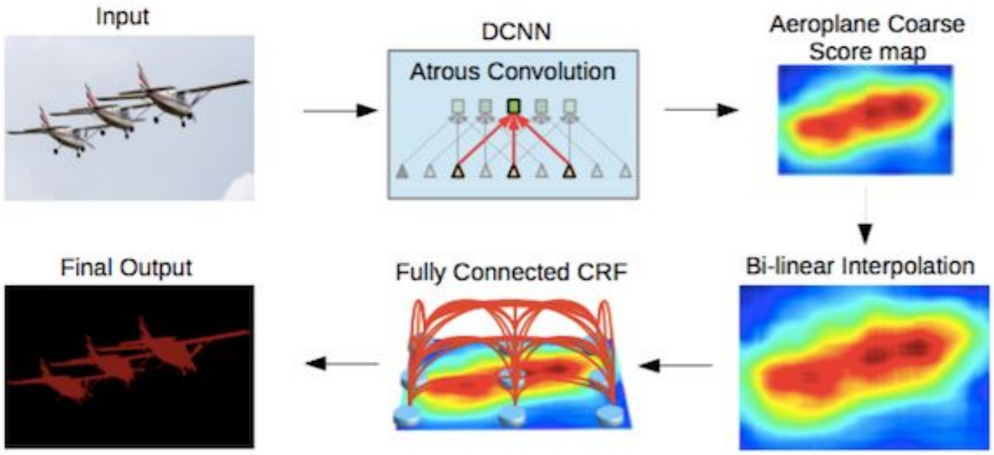
- PSPNet (dilated convolutions + pyramid pooling)



Zhao et al. [Pyramid Scene Parsing Network](). CVPR 2017

State-of-the-art models

- DeepLab v2 (dilated convolutions + CRF)



- DeepLab v3 (added pyramid pooling. Removed CRF)

Chen et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. TPAMI 2017
Chen et al. Rethinking Atrous Convolution for Semantic Image Segmentation. TPAMI 2017

Outline

Segmentation Datasets

Semantic Segmentation Methods

- Deconvolution (or transposed convolution)
- Dilated Convolution
- Skip Connections

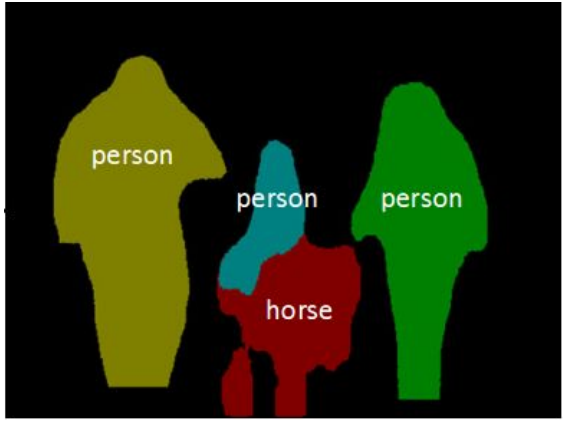
Instance Segmentation Methods

- Proposal-Based
- Recurrent
- Metric Learning

Instance Segmentation

Detect instances, give category, label pixels

“simultaneous detection and segmentation” (SDS)



Slide Credit: [CS231n]

Instance Segmentation

More challenging than Semantic Segmentation

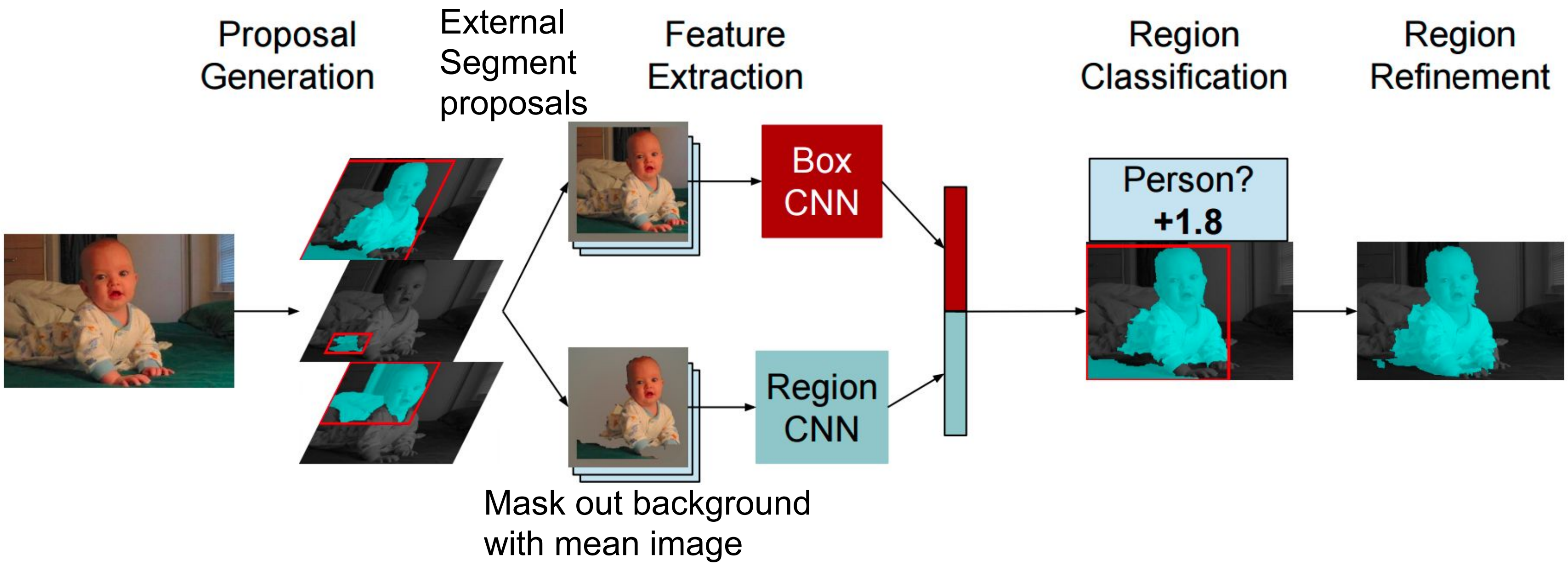
- Number of objects is variable
- No unique match between predicted and ground truth objects (cannot use instance IDs)

Several attack lines:

- Proposal-based methods
- Recurrent Neural Networks
- Metric Learning

Proposal-based

Similar to R-CNN, but with segment proposals

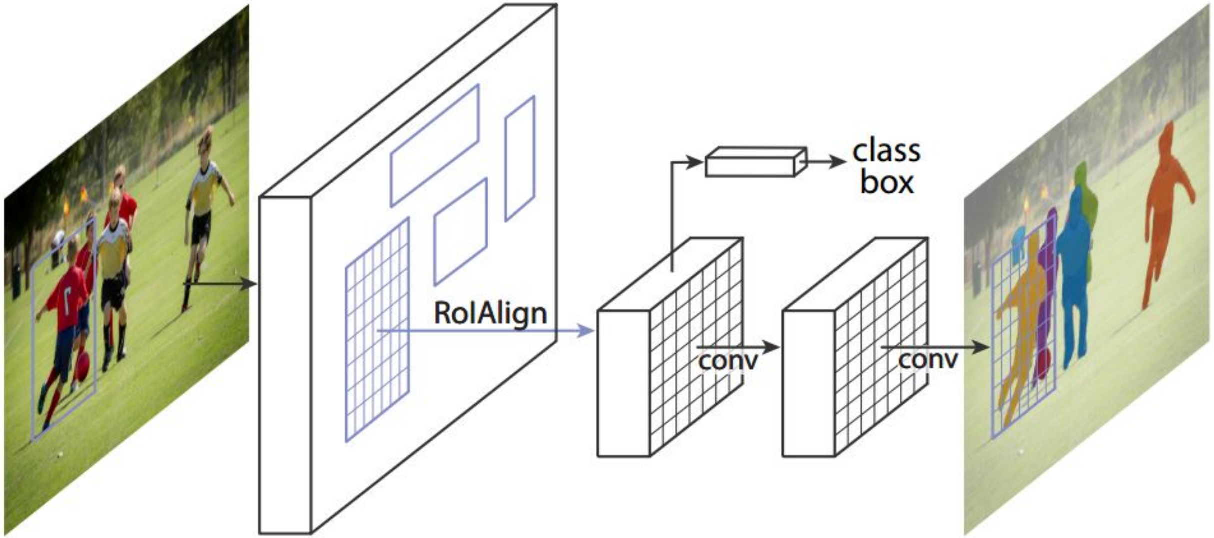


Hariharan et al. Simultaneous Detection and Segmentation. ECCV 2014

Slide Credit: CS231n

Proposal-based Instance Segmentation: Mask R-CNN

Faster R-CNN for Pixel Level Segmentation as a **parallel prediction of masks and class labels**

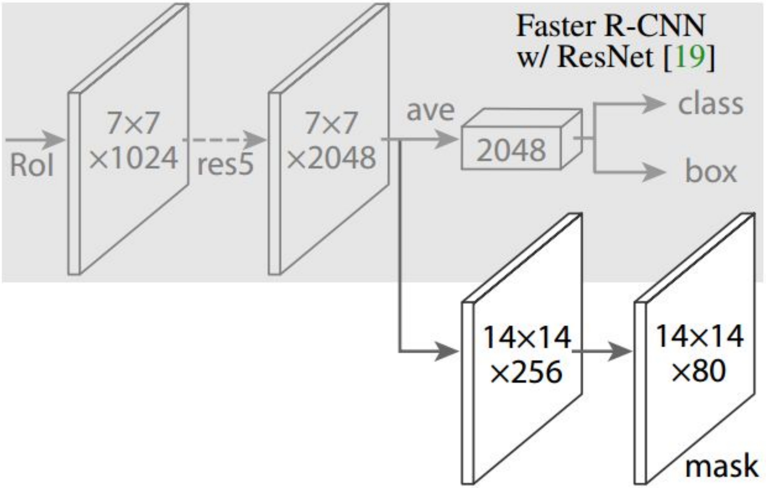


He et al. [Mask R-CNN](). ICCV 2017

Mask R-CNN

- Classification & box detection losses are identical to those in Faster R-CNN
- Addition of a new loss term for mask prediction:

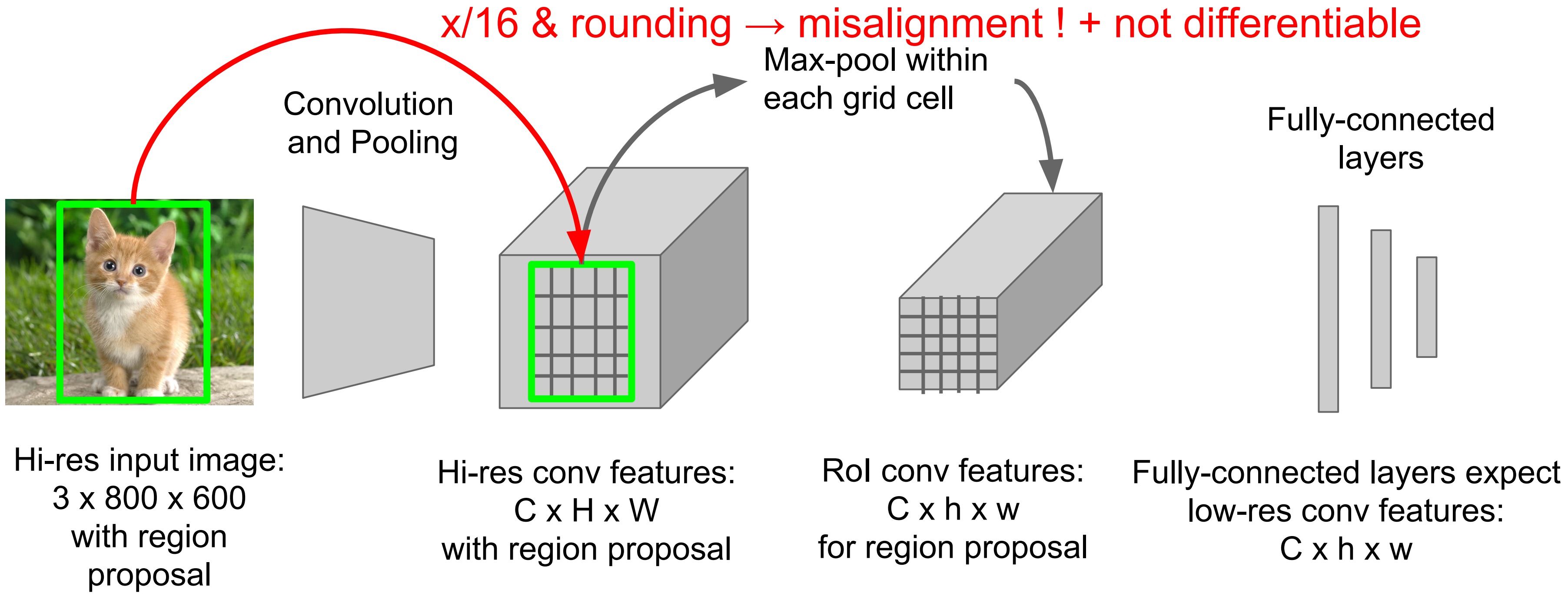
The network outputs a *K x m x m* volume for mask prediction, where *K* is the number of categories and *m* is the size of the mask (square)



He et al. Mask R-CNN. ICCV 2017

Mask R-CNN: RoI Align

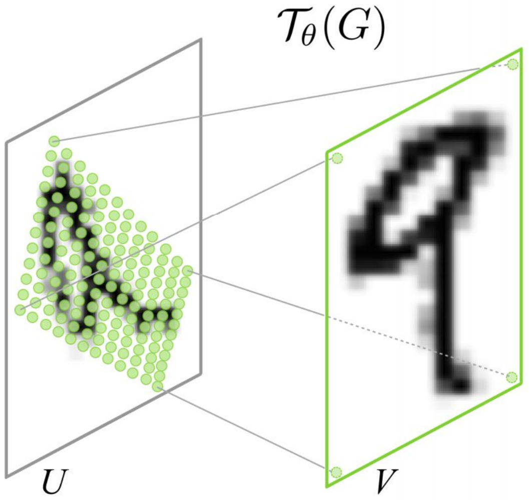
Reminder: RoI Pool from Fast R-CNN



He et al. Mask R-CNN. ICCV 2017

Mask R-CNN: RoI Align

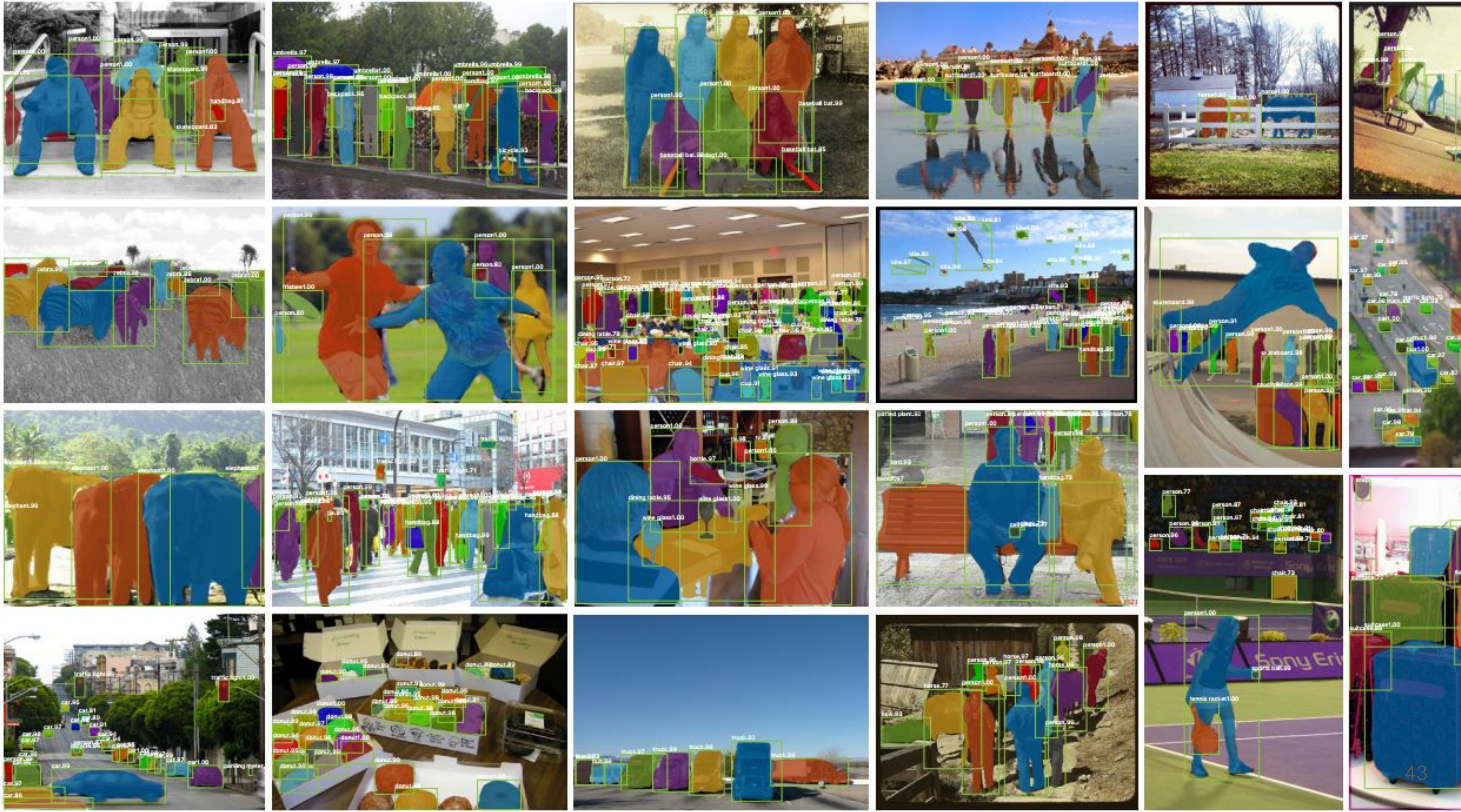
Use bilinear interpolation instead of cropping + maxpool



Mapping given by box coordinates
(θ_{12} and θ_{21} = 0 translation + scale)

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

Jaderberg et al. Spatial Transformer Networks. NIPS 2015

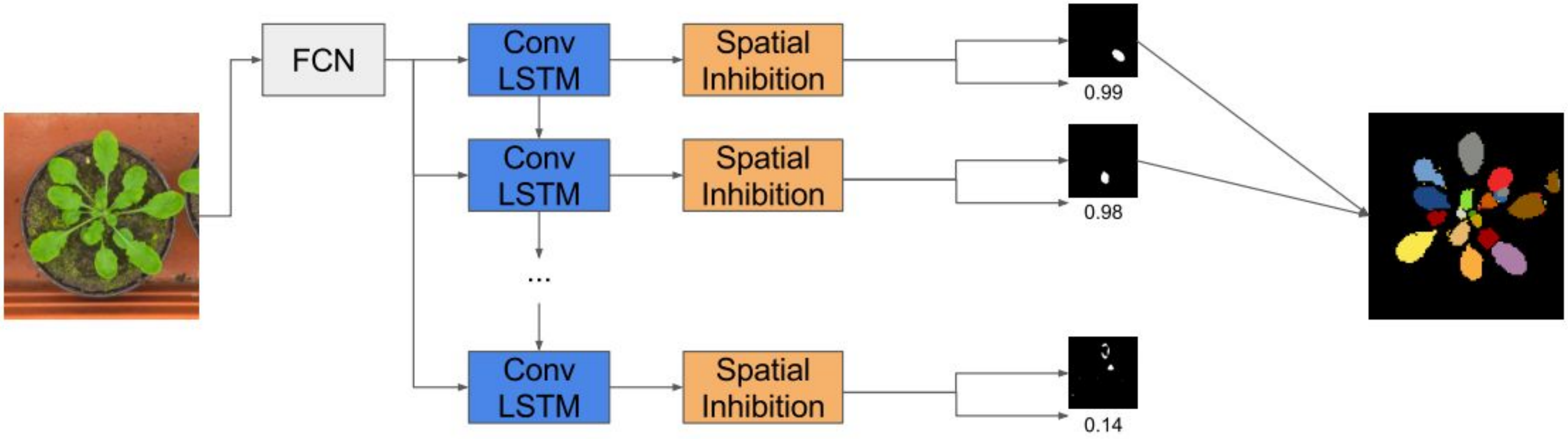


Limitations of Proposal-based models

1. Two objects might share the same bounding box: Only one will be kept after NMS step.
2. Choice of NMS threshold is application dependant
3. Choice of anchor boxes is application dependant
4. Same pixel can be assigned to multiple instances
5. Number of predictions is limited by the number of proposals.

Recurrent Instance Segmentation

Sequential mask generation



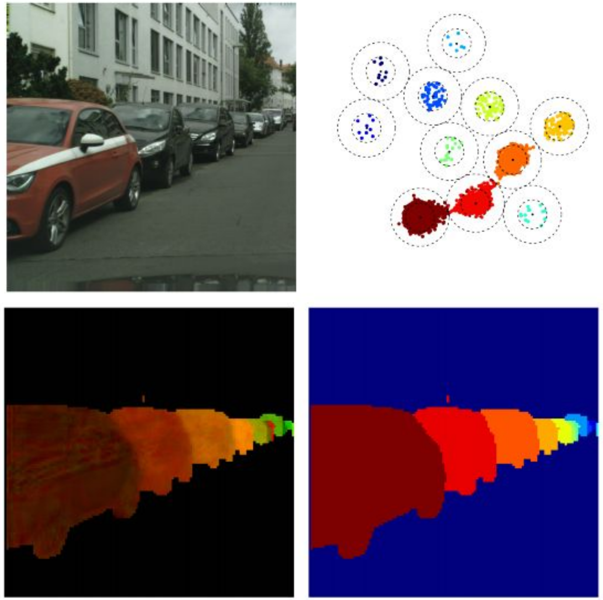
Romera-Paredes & H.S. Torr. [Recurrent Instance Segmentation](#) ECCV 2016

Recurrent Instance Segmentation

Romera-Paredes & H.S. Torr. [Recurrent Instance Segmentation]() ECCV 2016

Metric Learning

Mapping pixels to a N-dimensional space where pixels belonging to the same object are close to each other.



	AP	AP0.5	AP100m	AP50m
R-CNN+MCG	4.6	12.9	7.7	10.3
FCN+Depth	8.9	21.1	15.3	16.7
JGD	9.8	23.2	16.8	20.3
InstanceCut	13.0	27.9	22.1	26.1
Boundary-aware	17.4	36.7	29.3	34.0
DWT	19.4	35.3	31.4	36.8
Pixelwise DIN	20.0	38.8	32.6	37.6
Mask R-CNN	26.2	49.9	37.6	40.1
Ours	17.5	35.9	27.8	31.0

Results on Cityscapes

Brabandere et al. Semantic Instance Segmentation with a Discriminative Loss Function. CVPRW 2017

Outline

Segmentation Datasets

Semantic Segmentation Methods

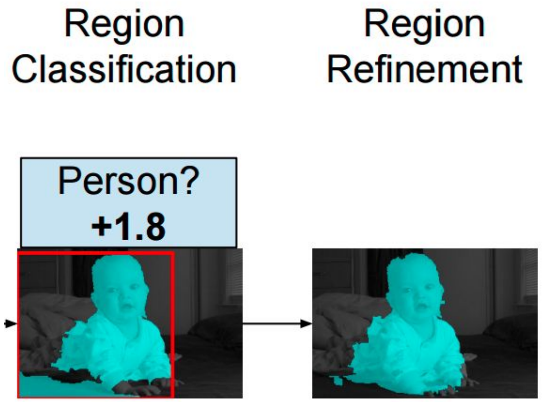
- Deconvolution (or transposed convolution)
- Dilated Convolution
- Skip Connections

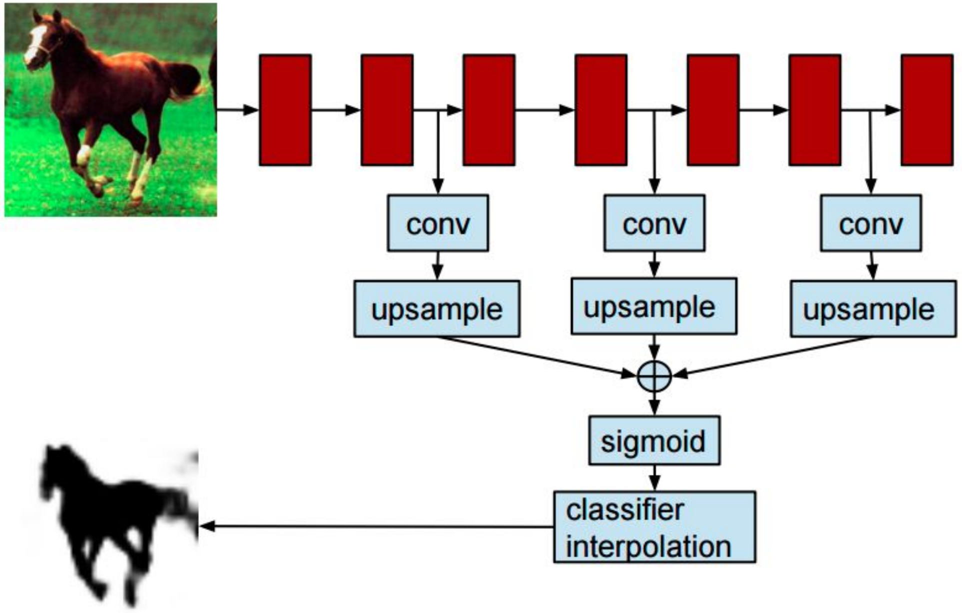
Instance Segmentation Methods

- Proposal-Based
- Recurrent
- Metric Learning

Questions ?

Proposal-based





Hariharan et al. [Hypercolumns for Object Segmentation and Fine-grained Localization](). CVPR 2015

Slide Credit: [CS231n]()

Proposal-based Instance Segmentation: MNC

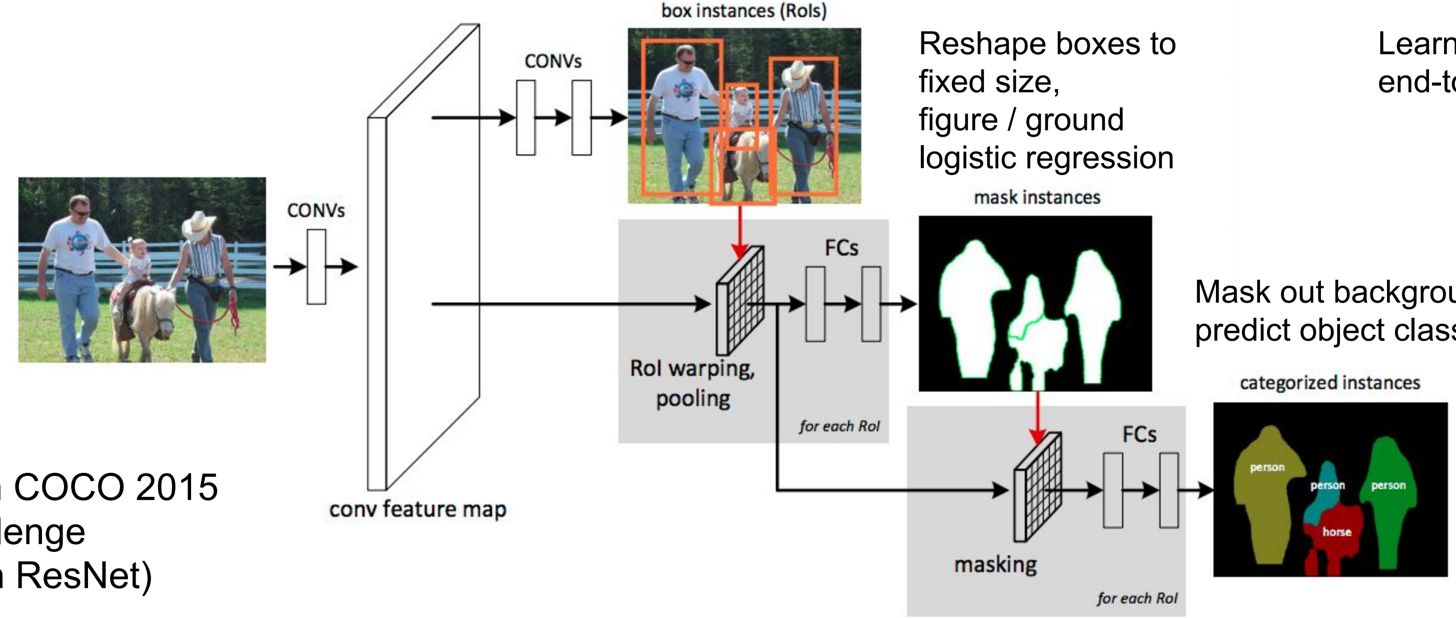
Faster R-CNN for Pixel Level Segmentation in a multi-stage cascade strategy

Region proposal network (RPN)

Reshape boxes to fixed size, figure / ground logistic regression

Learn entire model end-to-end!

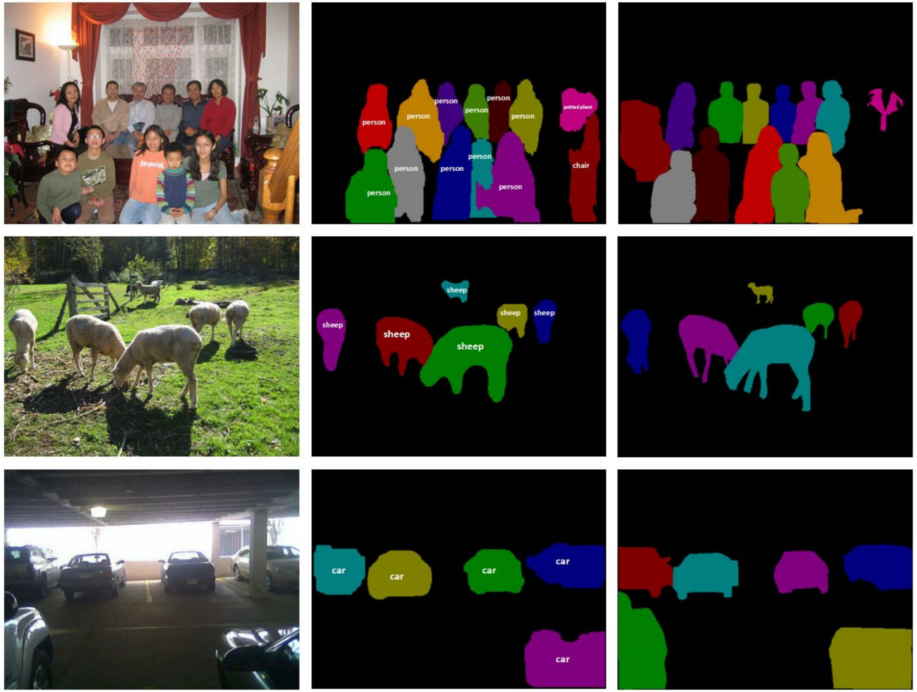
Mask out background, predict object class



Won COCO 2015 challenge (with ResNet)

Dai et al. [Instance-aware Semantic Segmentation via Multi-task Network Cascades](). CVPR 2016

Proposal-based Instance Segmentation: MNC



Predictions **Ground truth**

Dai et al. [Instance-aware Semantic Segmentation via Multi-task Network Cascades](). CVPR 2016

Mask R-CNN

Instance Segmentation

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

Object Detection

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_S	AP^{bb}_M	AP^{bb}_L
Faster R-CNN+++ [19]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [27]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [21]	Inception-ResNet-v2 [37]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [36]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3	19.8	40.2	48.8
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**	**22.1**	**43.2**	51.2

He et al. Mask R-CNN. arXiv Mar 2017