Course: Deep Learning

Unit 2: Computer Vision

Encoder-decoder architectures. Applications.

Luis Baumela

Universidad Politécnica de Madrid

Computer Vision applications

1. Introduction
2. Encoder-decoder architectures. Semantic Segmentation
   - Sliding window approach
   - Fully Convolutional Neural Net (FCNN)
   - Encoder-decoder architectures
   - U-Net
   - Stacked hourglass
3. Object detection
   - Single Step approaches
   - Two-step approaches
   - Evaluating object detection
4. Instance segmentation

Introduction

- A deep NN is a powerful image description model

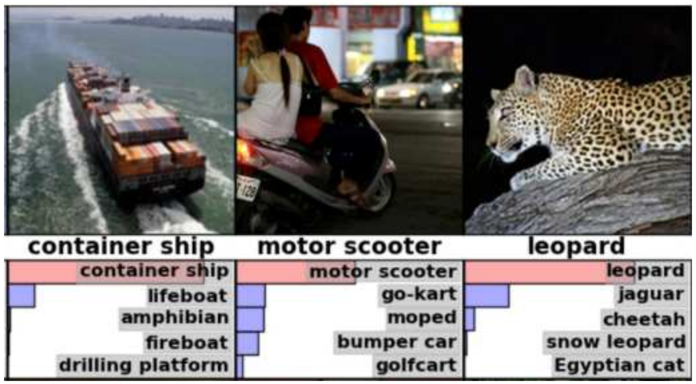


Image classification task

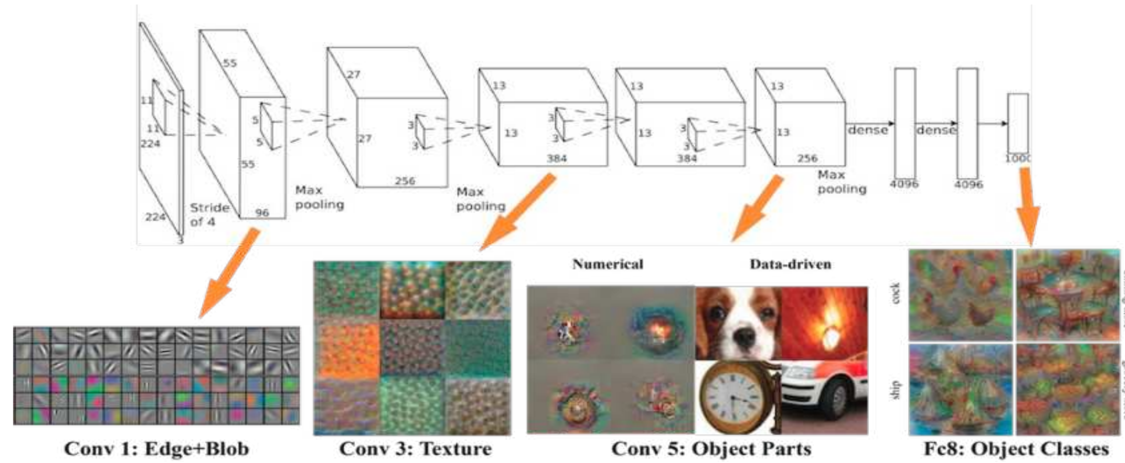
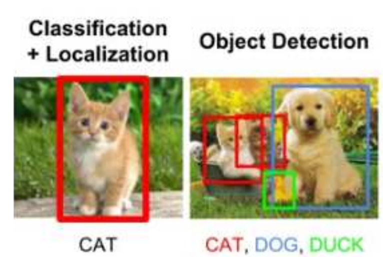
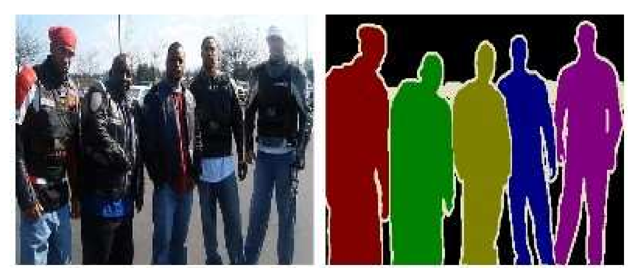


Image classification CNN: AlexNet

AlexNet, VGG, ResNet, … are
- hiearchical models
- composed of various layers
- each layer extracts image features at different levels of abstraction

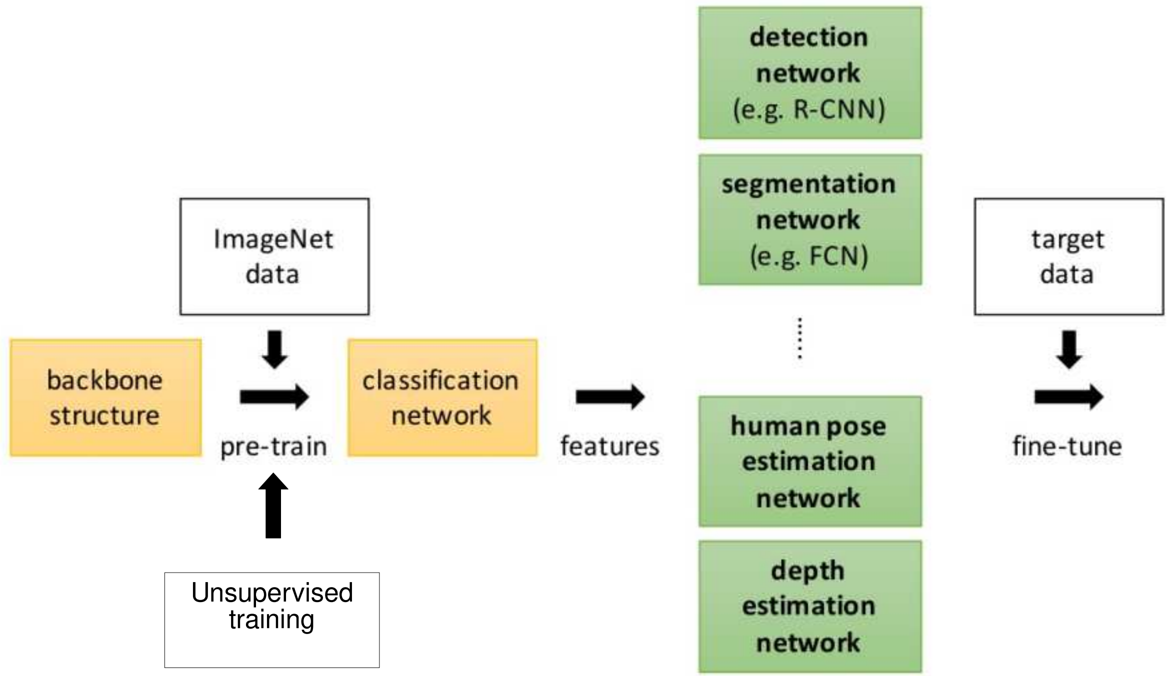
These features may be used for solving many other CV problems





Introduction

- General CNN computer vision pipeline



The performance of the final system will depend on both:
- Selected backbone architecture (VGG, GoogLeNet, ResNet, …)
- Task-specific model

K. He

Introduction

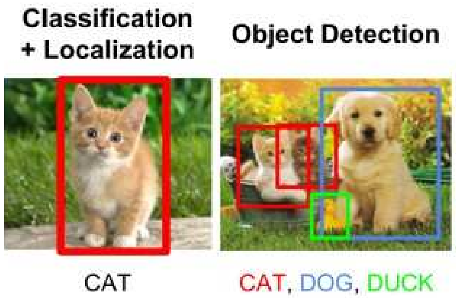
- Problems considered

Semantic segmentation

Pixels + labels

Object localization and detection

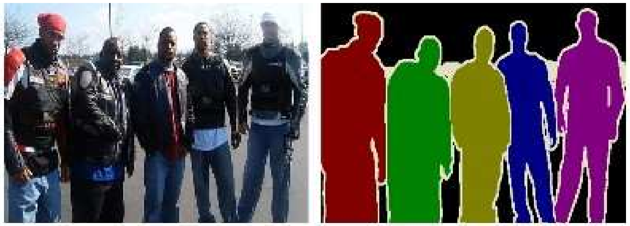
Single object + localization
+ class label



Multiple object +
localizations + class labels

Instance segmentation

Pixels + instance class labels



Semantic segmentation

- Problem statement

Attach to each pixel in an image a label from a set of predefined classes.

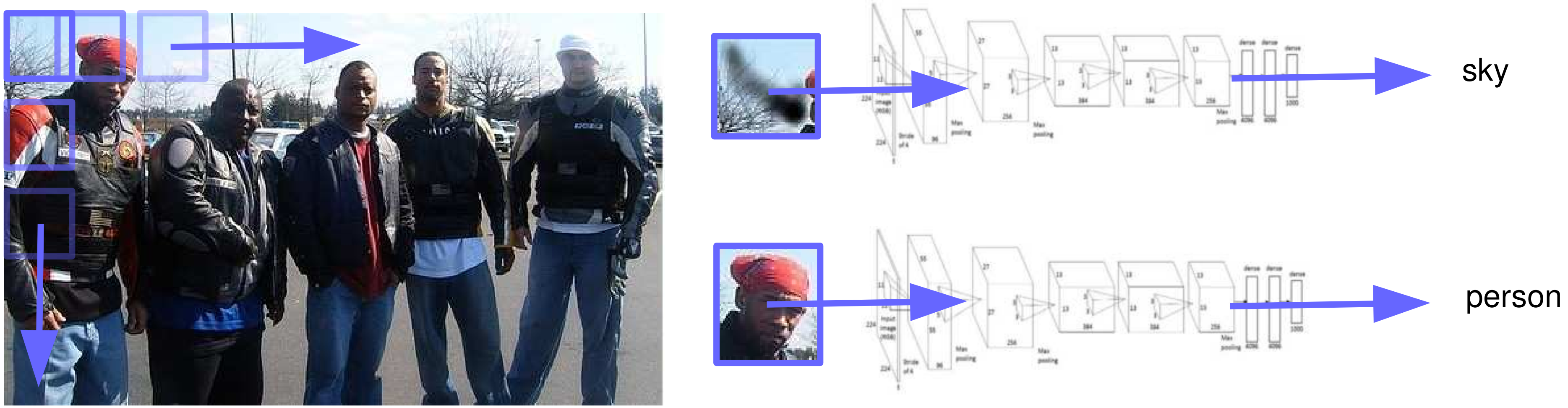
$\boxed{\text{Semantic segmentation}}$ ~~Instance segmentation~~

All pixels from the same class have the same label!

Semantic segmentation

- Naive approach: sliding window

The model only classifies one pixel per run.
Center a sliding window onto each pixel and push it thought the net to establish its label.

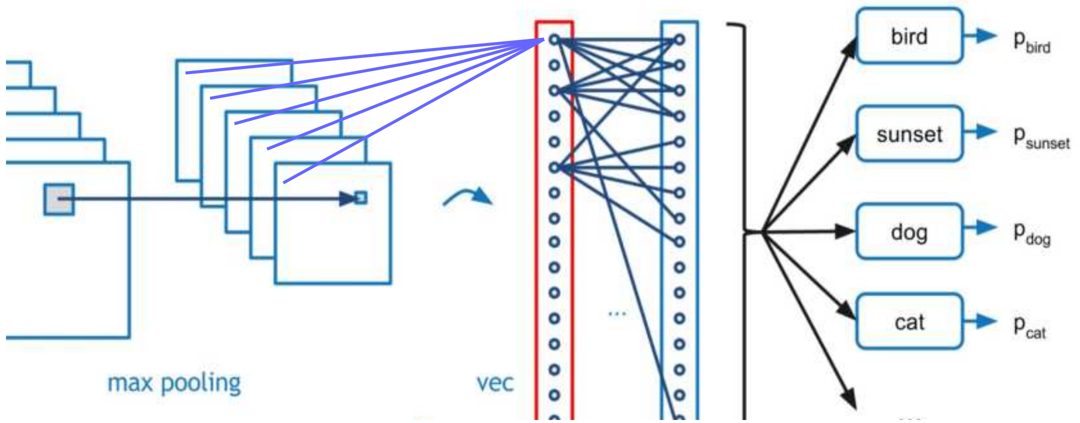


+ Lots of training data
- computationally very inefficient
- unable to use large neighbourhoods
- no parameter reuse

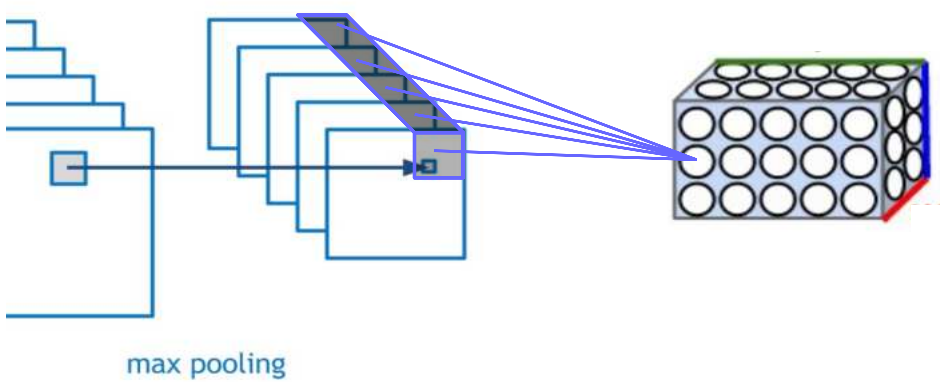
Semantic segmentation

- Better approach: use a fully convolutional NN

A fully connected layer is a convolutional layer with a receptive field whose size is the full spatial extent of the previous layer



standard fully connected layer

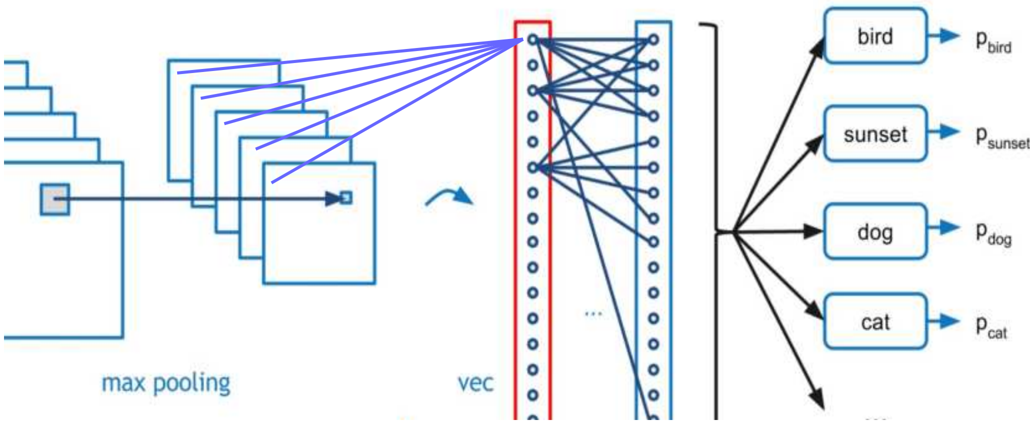


standard convolutional layer

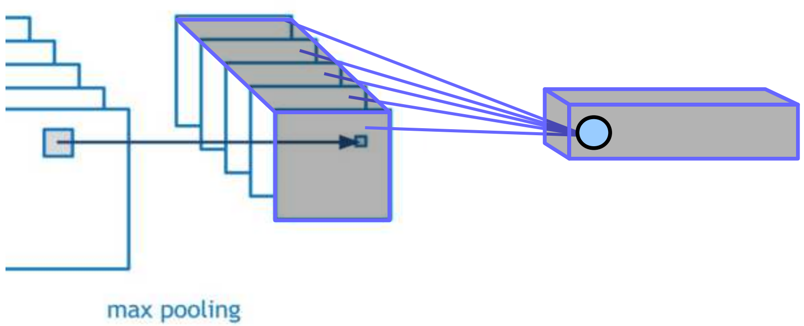
Semantic segmentation

- Better approach: use a fully convolutional NN

A fully connected layer is a convolutional layer with a receptive field whose size is the full spatial extent of the previous layer



standard fully connected layer

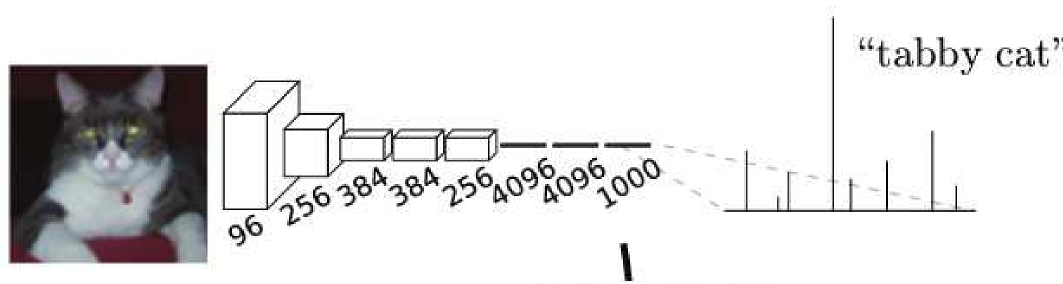


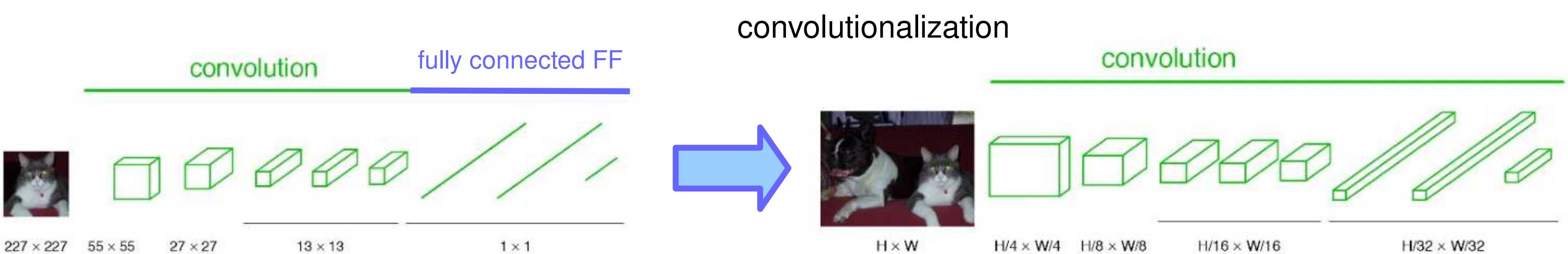
convolutional layer equivalent to a fully connected layer

Semantic segmentation

- Better approach: use a fully convolutional NN

What is the result of the convolutionalization?





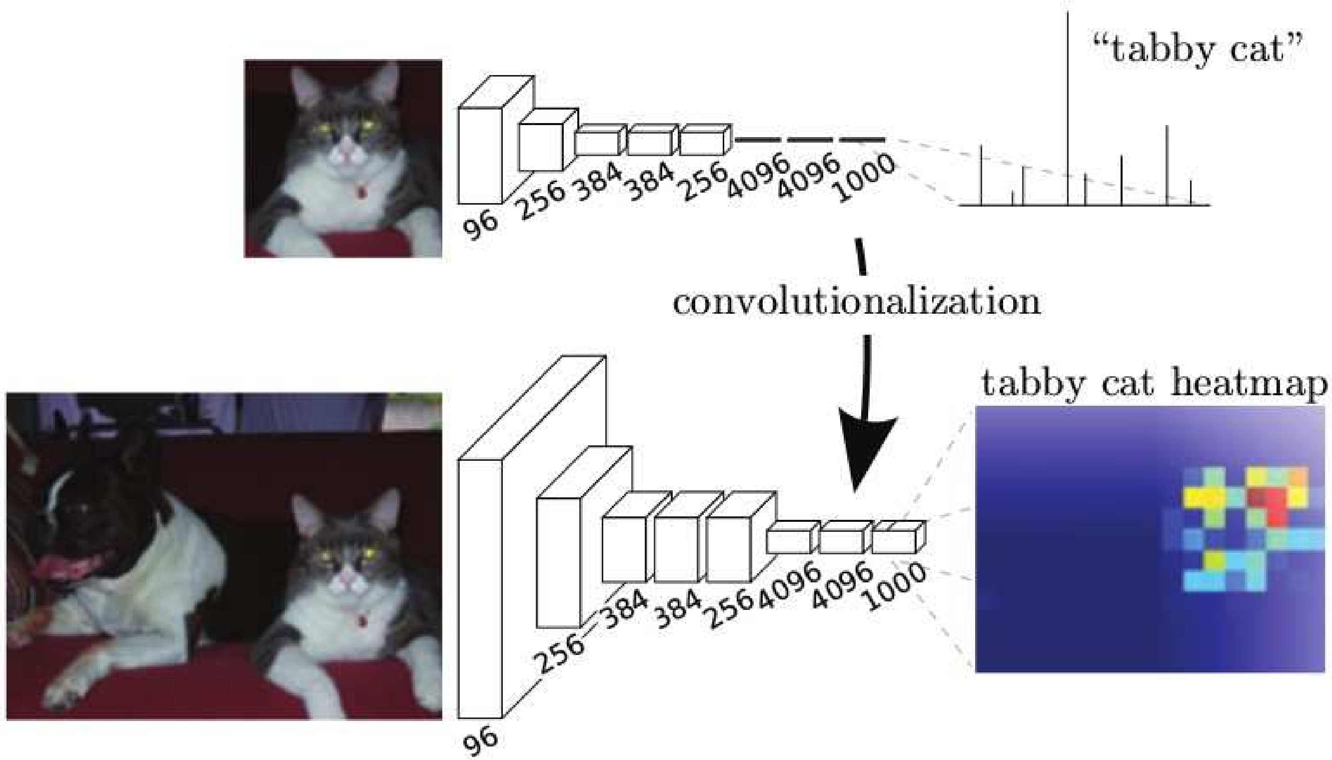
A fully CNN behaves like a huge filter:
- input image size is arbitrary,
- output size depends on input.

Long et al. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.

Semantic segmentation

- Better approach: use a fully convolutional NN

Why is a fully-CNN better for segmentation?

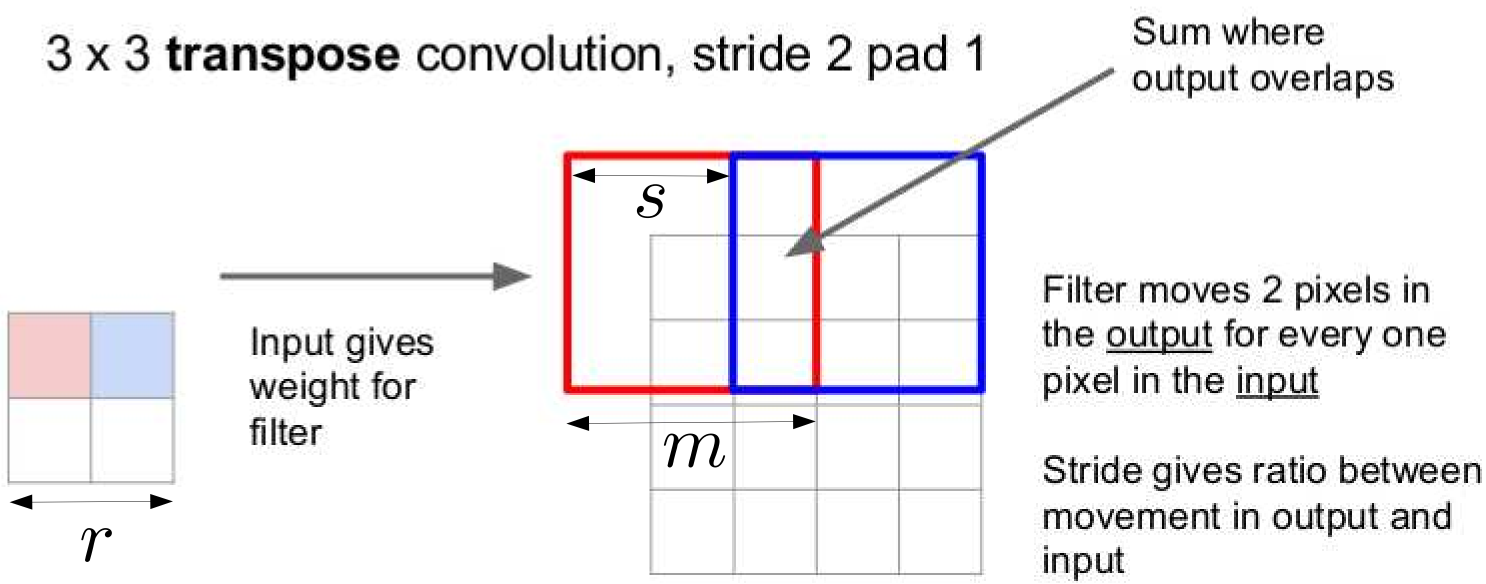


A fCNN provides a "heat map" for each class, and
- efficient evaluation and end-to-end training of a large images,
- large receptive field (depending on CNN depth)
- re-use of shared parameters.
- less parameters.

Long et al. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.

Semantic segmentation

- Transposed convolution



The size of the upsampled matrix

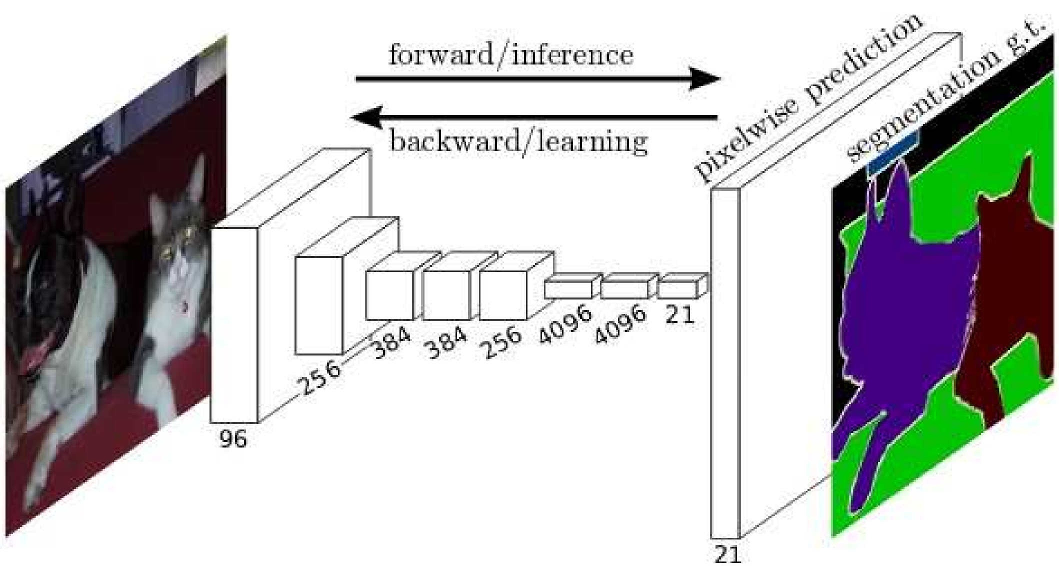
$$M = (r-1) * s + m$$

Semantic segmentation

- Better approach: use a fully convolutional NN

Recover initial image resolution with **transposed convolution**.

New network architecture



Results

no skips

Long et al. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.

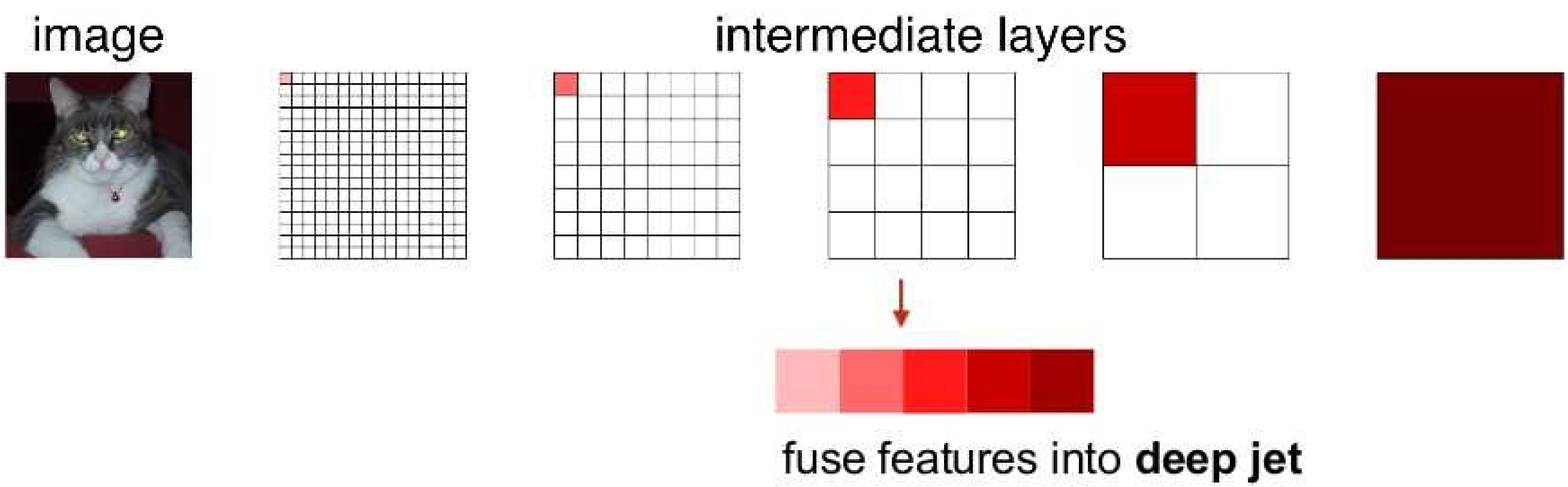
Semantic segmentation

- Better approach: use a fully convolutional NN

What is the problem?

spectrum of deep features

combine *where* (local, shallow) with *what* (global, deep)

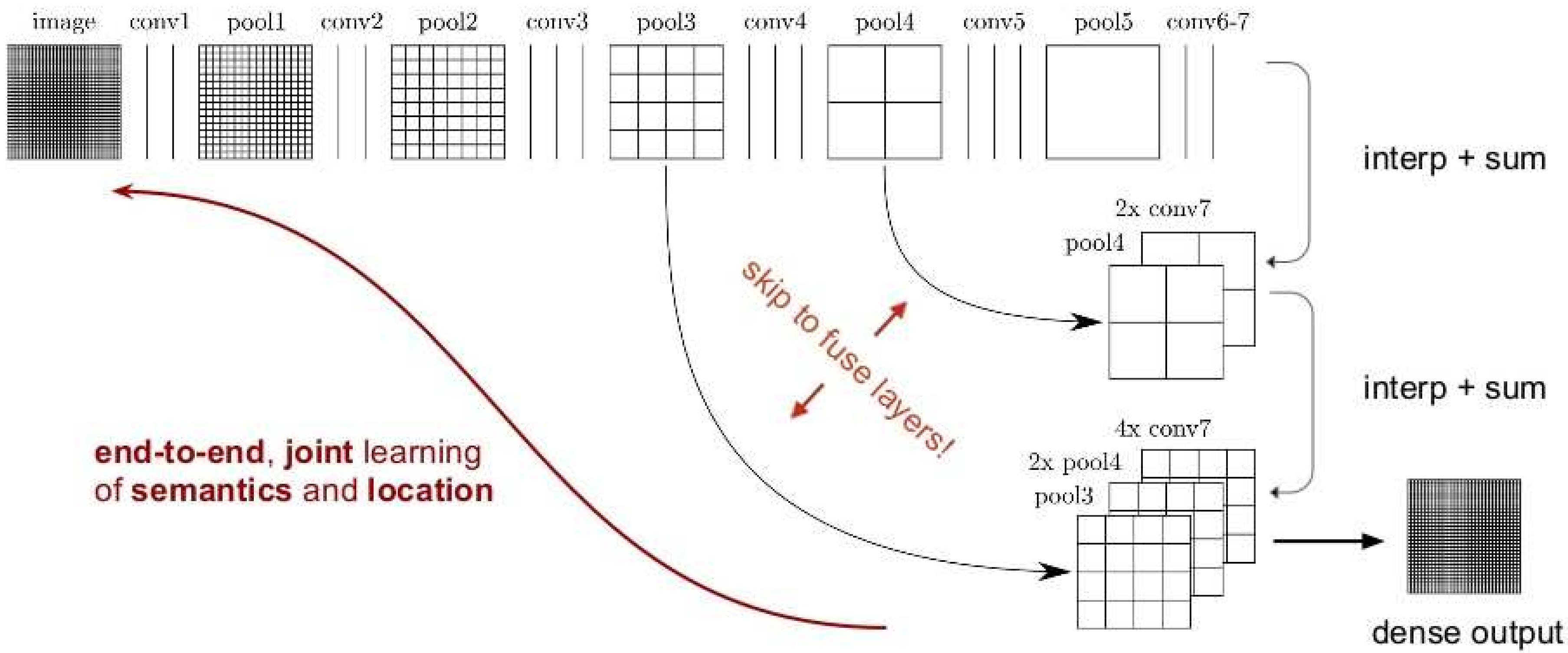


Long et al. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.

Semantic segmentation

- Better approach: use a fully convolutional NN

Solution: add “skip connections” from finer convolutional layers



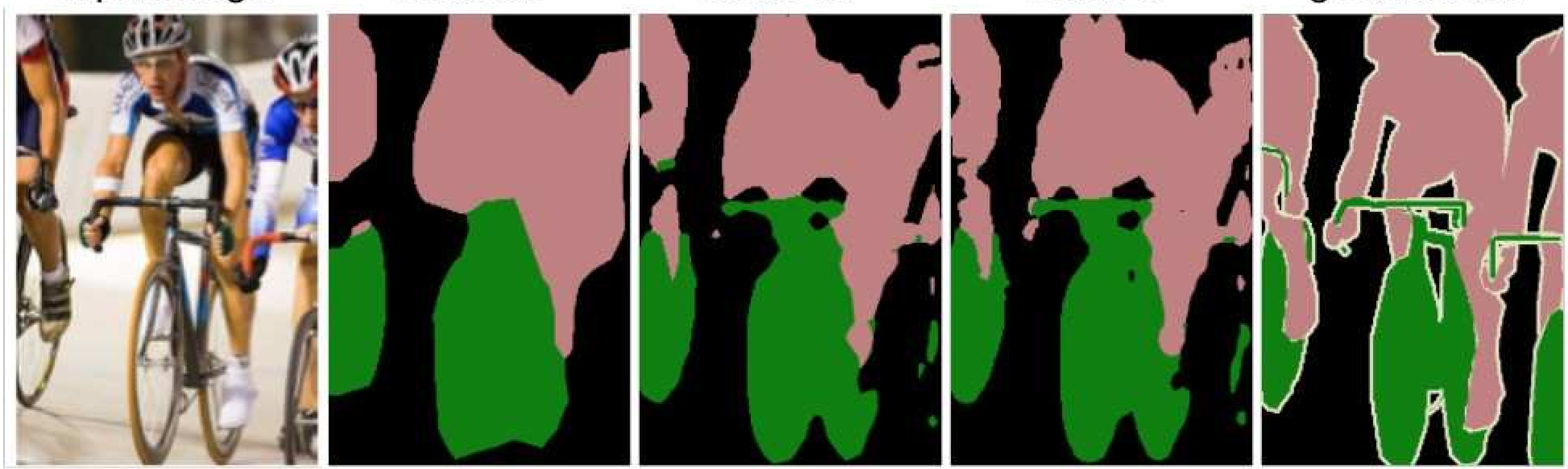
Long et al. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.

Semantic segmentation

- Better approach: use a fully convolutional NN

Solution: add “skip connections” from finer convolutional layers

Results:

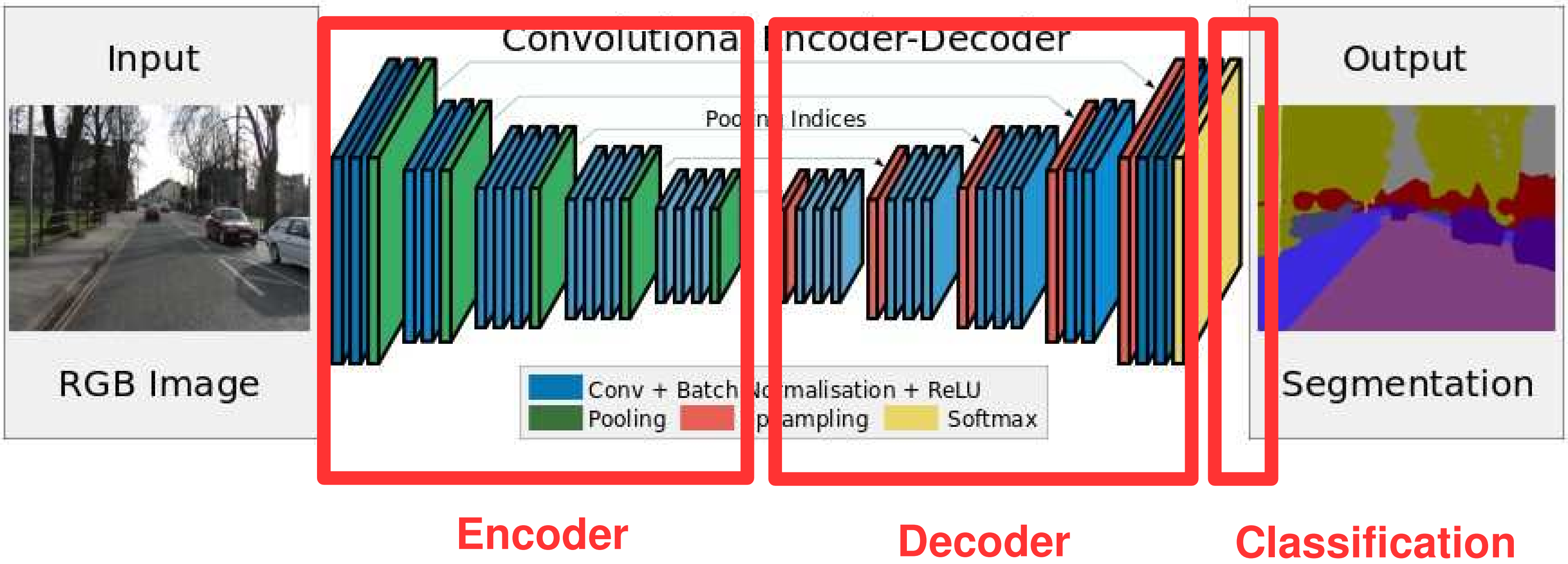


Long et al. Fully Convolutional Networks for Semantic Segmentation, CVPR 2015.

Semantic segmentation

- Alternative approach: Encoder-decoder

Symmetric encoder-decoder type of architecture



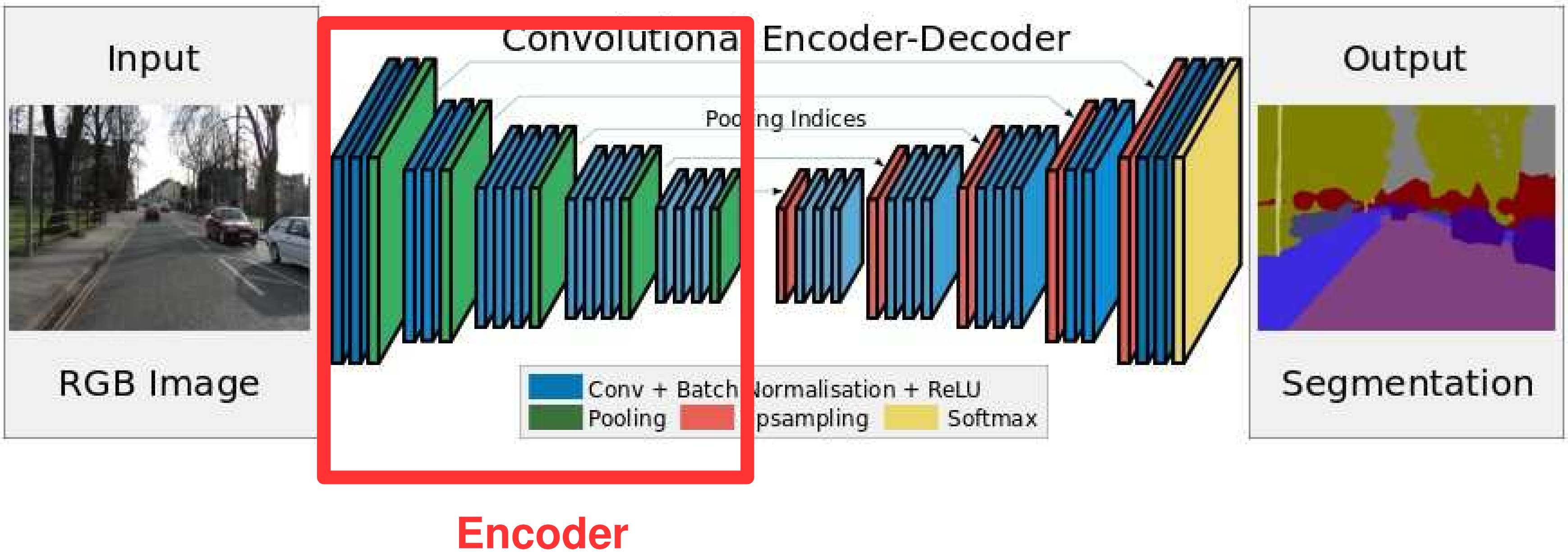
Three components:
- encoder
- decoder
- pixelwise classifier

Badrinarayana et al. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" PAMI 17

Semantic segmentation

- Alternative approach: Encoder-decoder

Symmetric encoder-decoder type of architecture



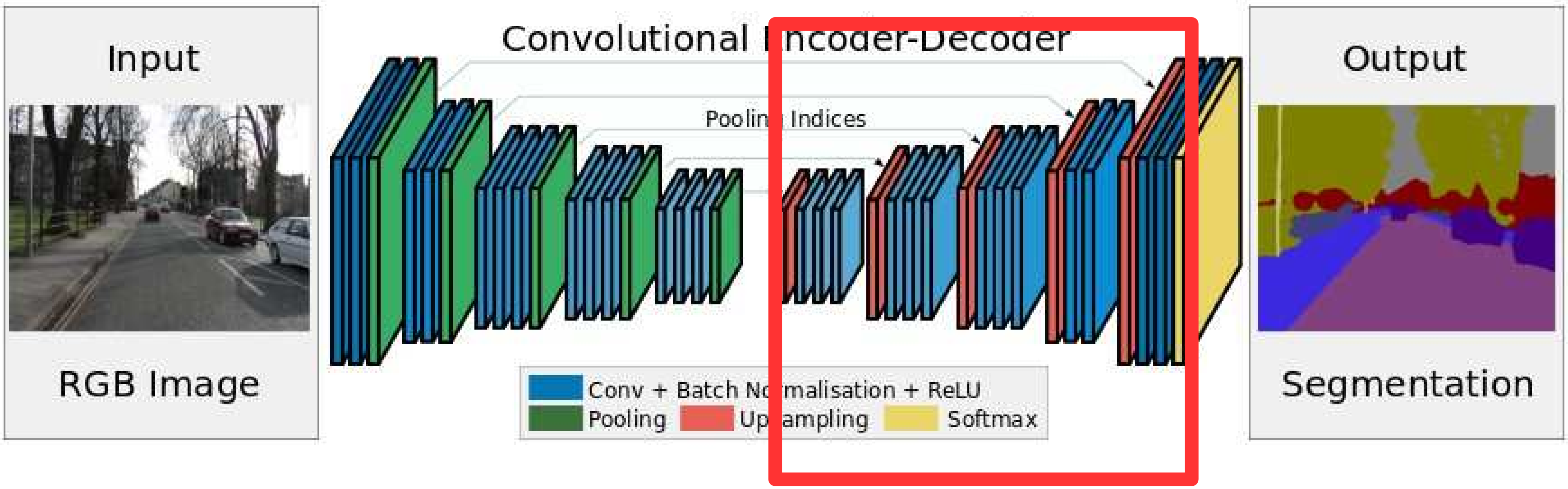
Encoder
- VGG16-based (13 conv layers)
- conv layer 3x3, stride 1 + batch normalization + ReLU
- max pooling 2x2, stride 2
- stored max pool indices (for later upsampling)

Badrinarayana et al. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" PAMI 17

Semantic segmentation

- Alternative approach: Encoder-decoder

Symmetric encoder-decoder type of architecture



Decoder

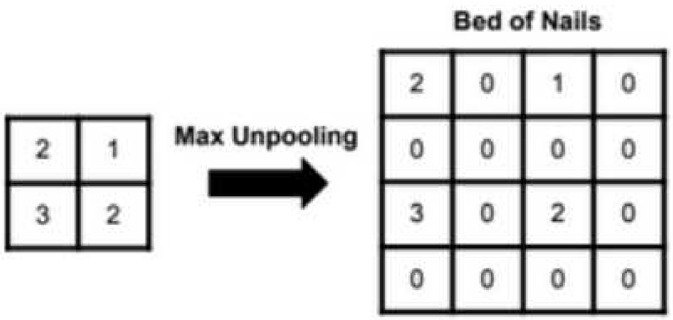
Decoder:
- unpooling sparse feature map (from memorized indices)
- batch normalization + ReLU

Badrinarayana et al. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" PAMI 17

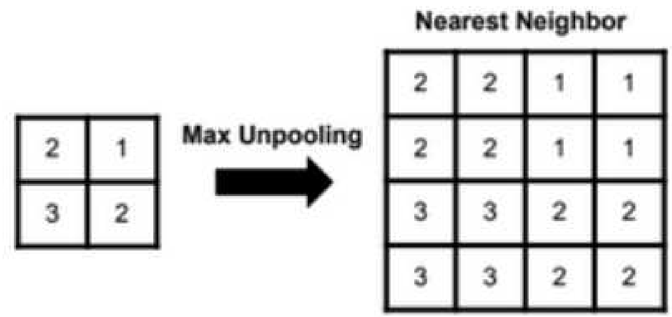
Semantic segmentation

- Unpooling

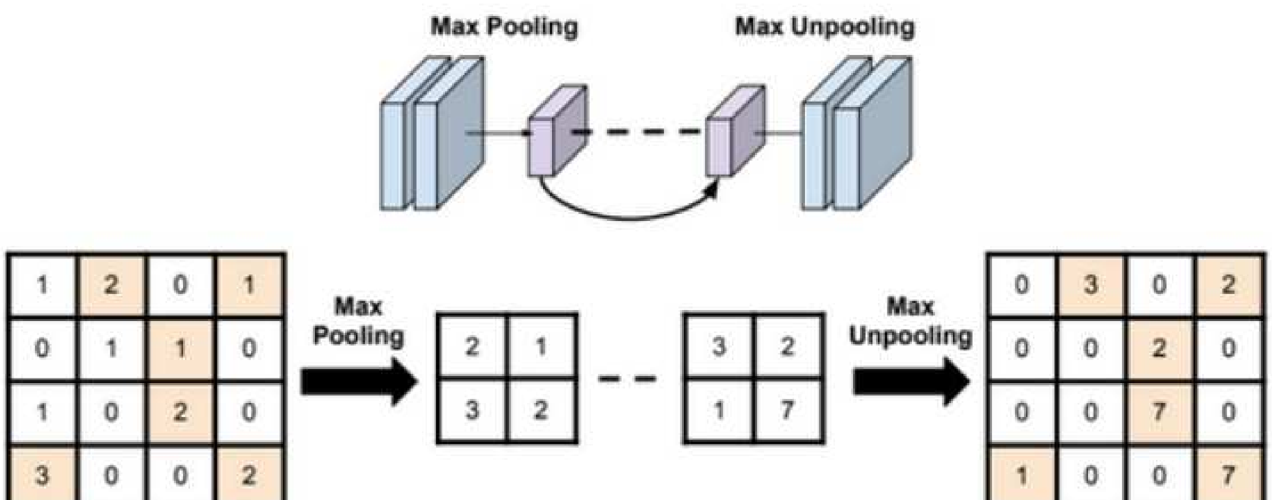
Bed of nails unpooling



Nearest neighbor unpooling



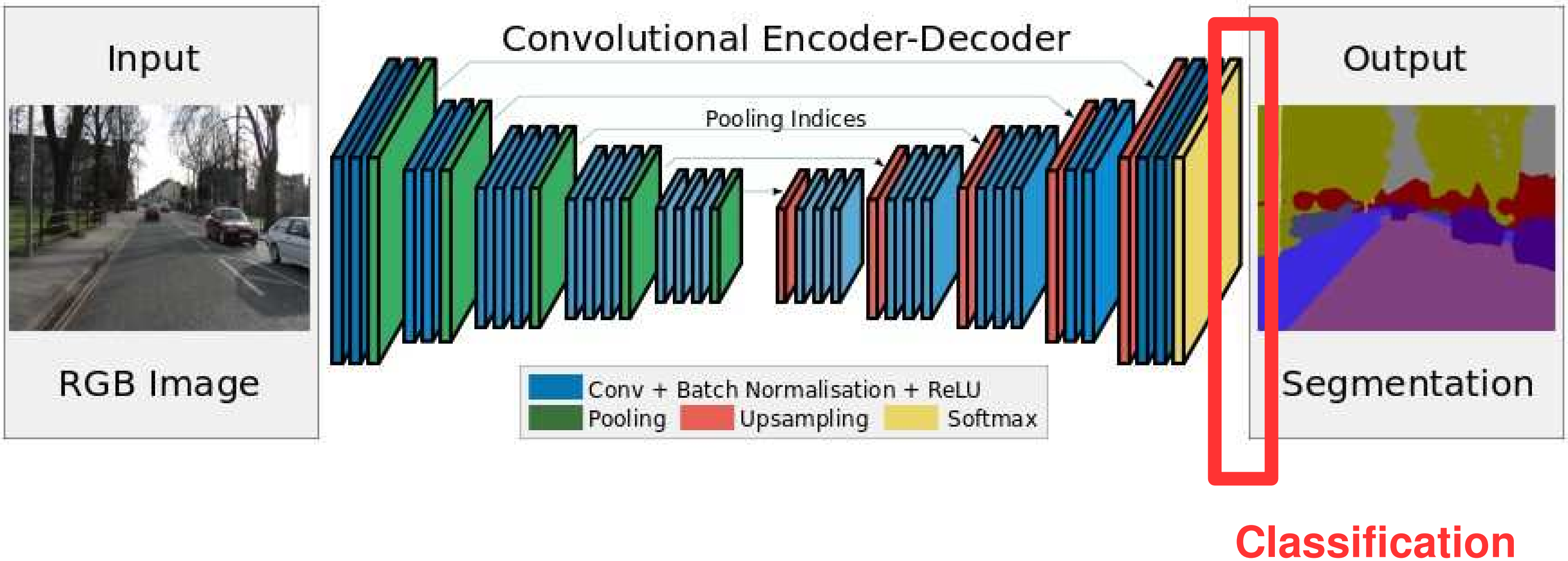
Max unpooling with memory



Semantic segmentation

- Alternative approach: Encoder-decoder

Symmetric encoder-decoder type of architecture



Classification:
- multiclass soft-max trainable classifier (each pixel is a soft-max!).
- class frequency balancing

Badrinarayana et al. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" PAMI 17

Semantic segmentation

- Alternative approach: Encoder-decoder

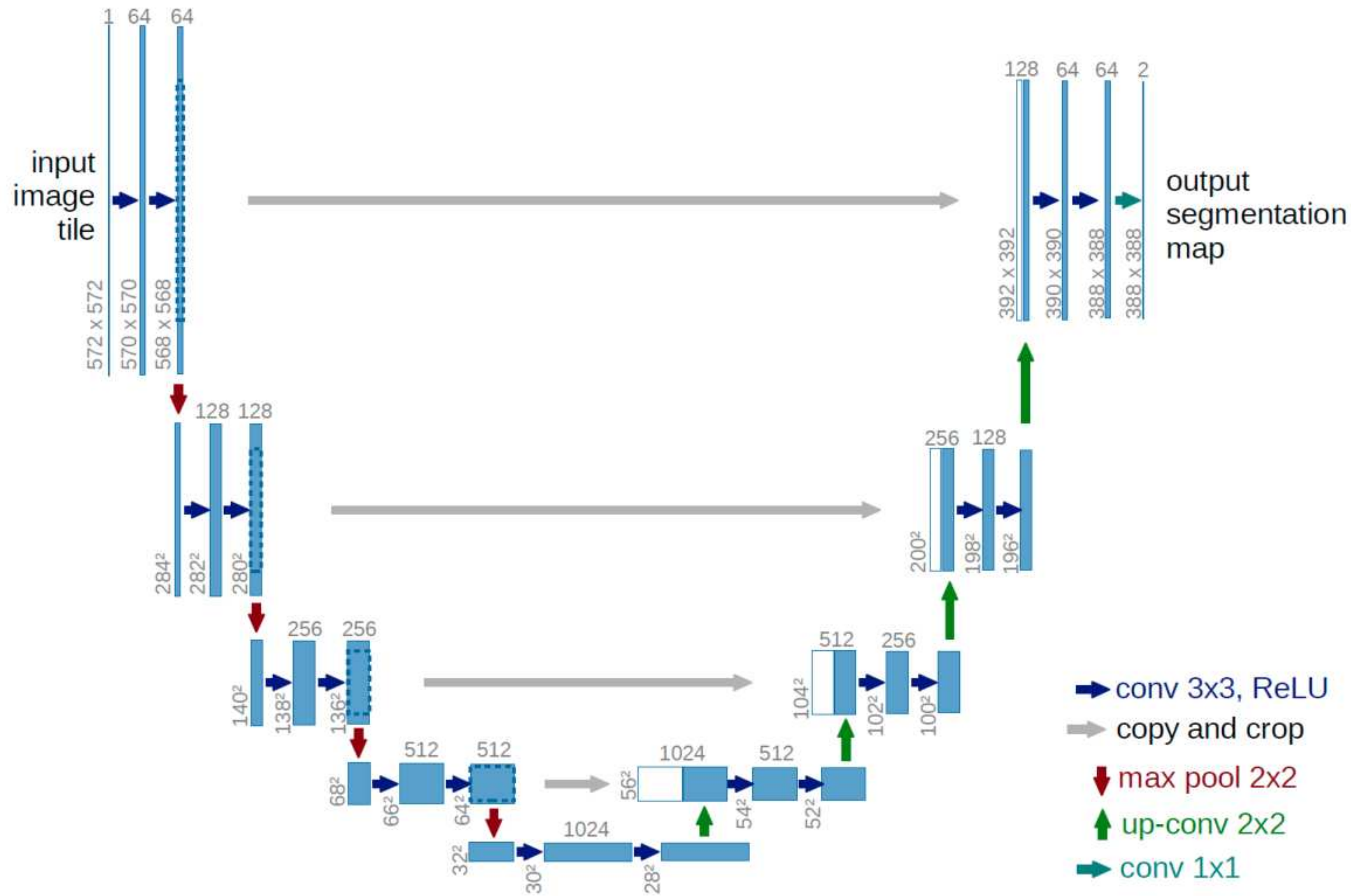
Segmentation results

Badrinarayana et al. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" PAMI 17

Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

U-Net

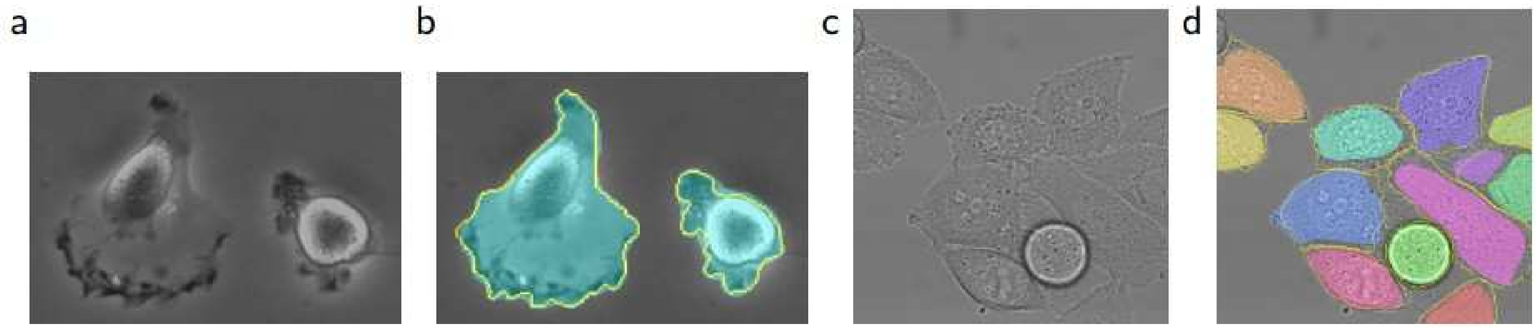


Ronnenberget et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation" MICCAI 2015

Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

U-Net. Results



Result on the ISBI cell tracking challenge. (**a**) part of an input image of the "PhC-U373" data set. (**b**) Segmentation result (cyan mask) with manual ground truth (yellow border) (**c**) input image of the "DIC-HeLa" data set. (**d**) Segmentation result (random colored masks) with manual ground truth (yellow border).

Segmentation results (IOU) on the ISBI cell tracking challenge 2015.

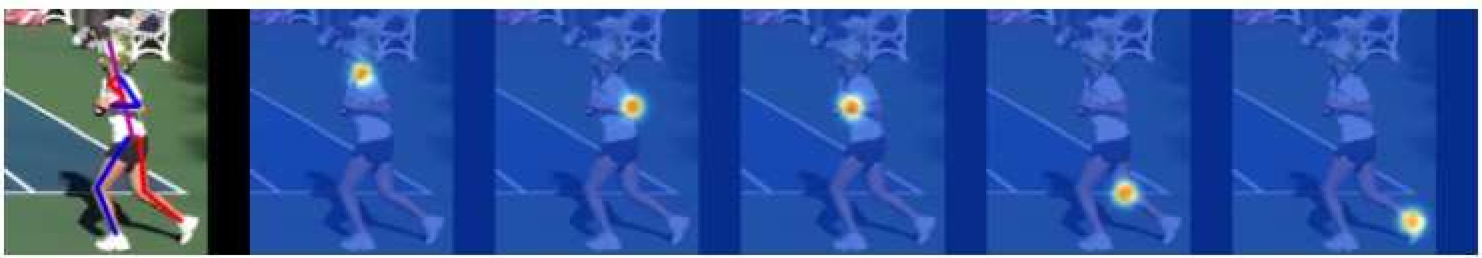
Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

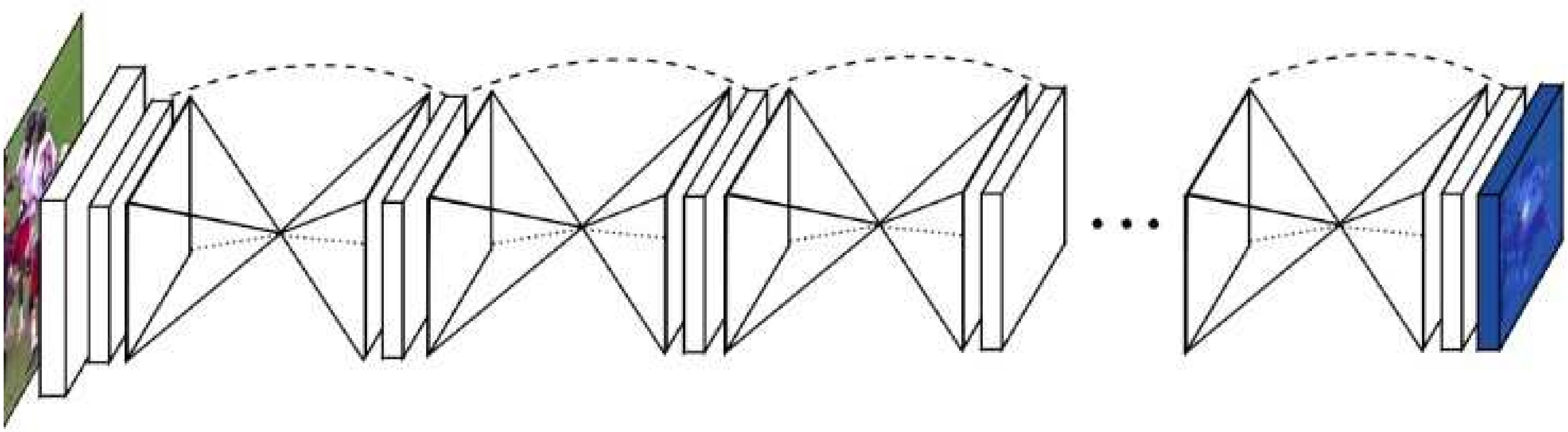
Ronnenberget et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation" MICCAI 2015

Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

Stacked Hourglass
- encoder-decoder network architecture for pose estimation
- features are processed across all scales to capture spatial relationships
- repeated bottom-up, top-down processing
- intermediate supervision





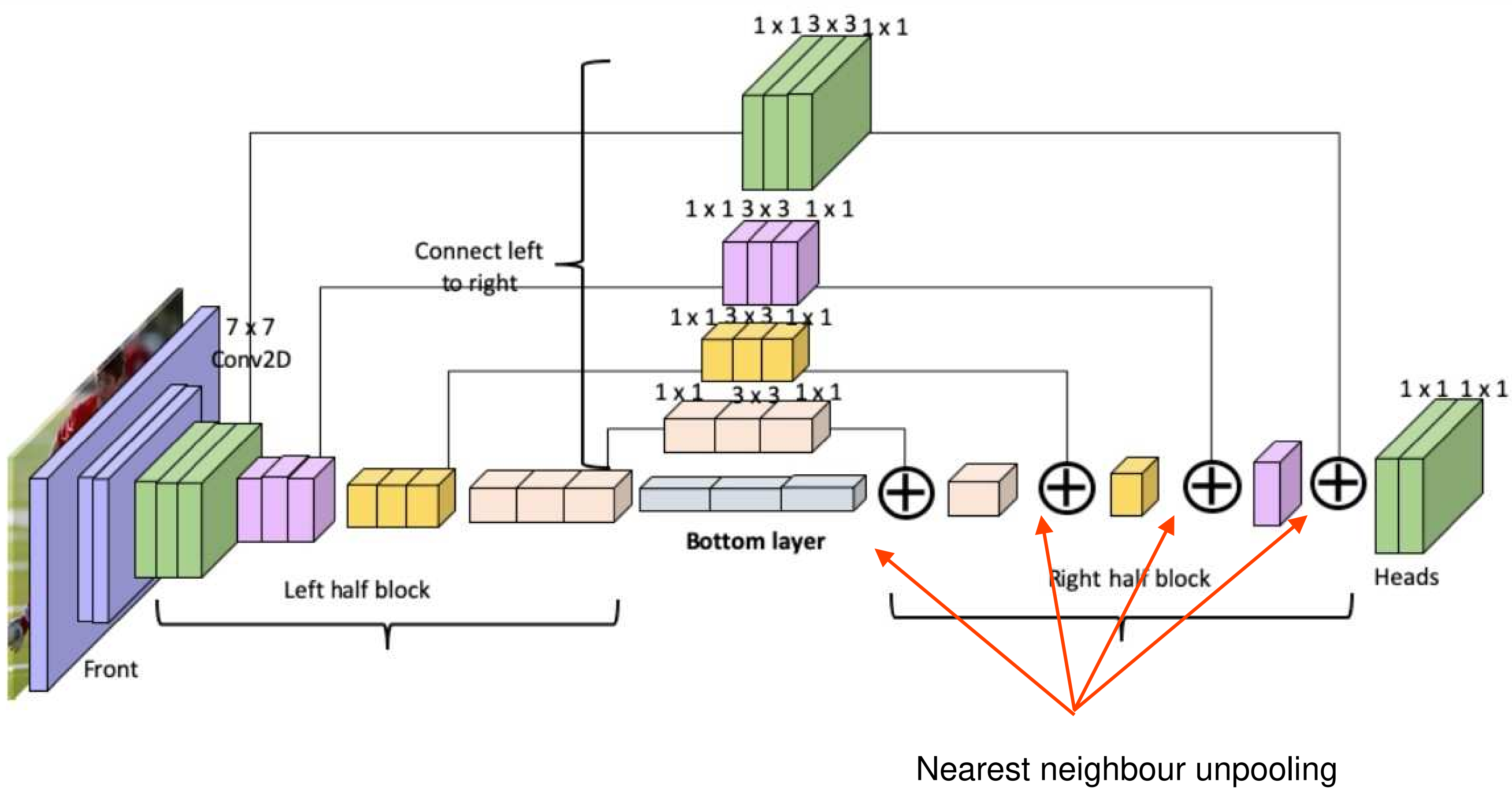
Newell. Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

Stacked Hourglass

- hourglass module



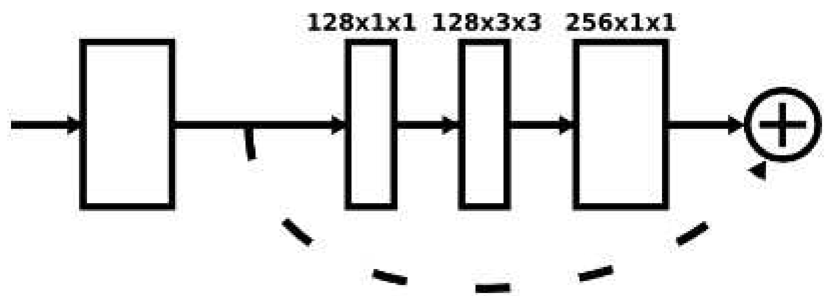
Newell. Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

Semantic segmentation

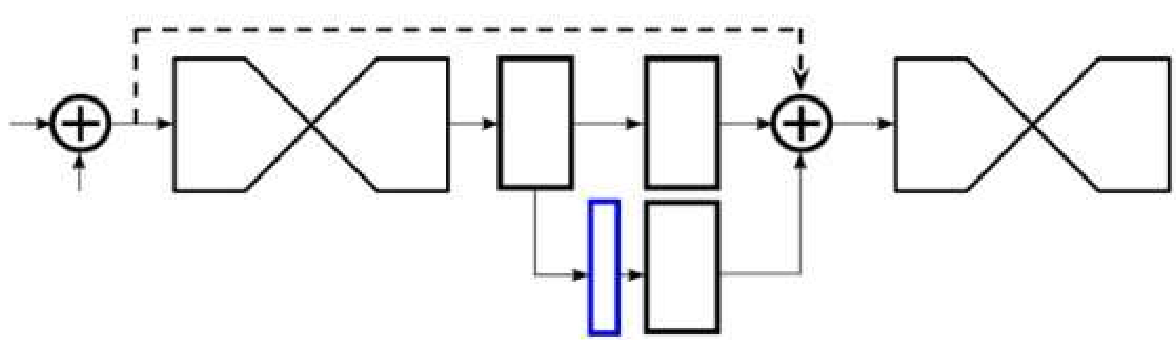
- Eclectic approach: Encoder-decoder + skip connections

Stacked Hourglass

- Residual module
  Each architecture block is a residual module



- Intermediate supervision



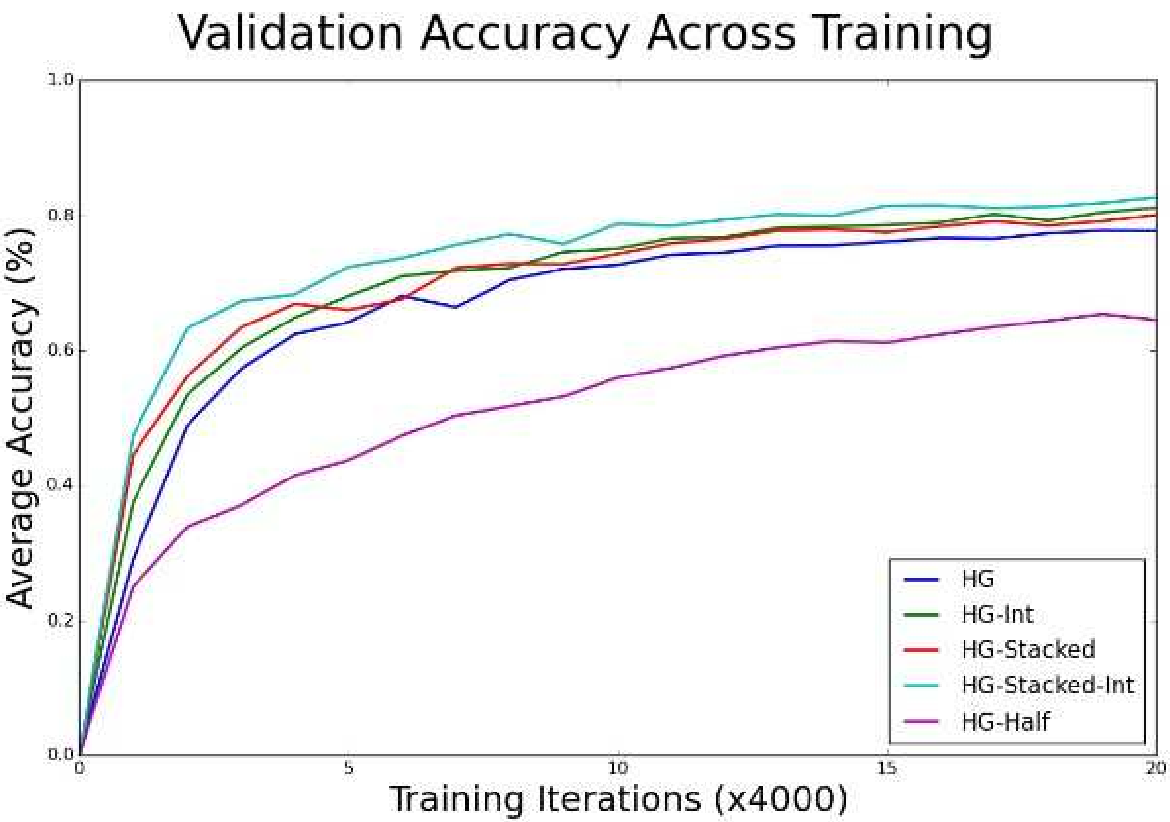
Newell. Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

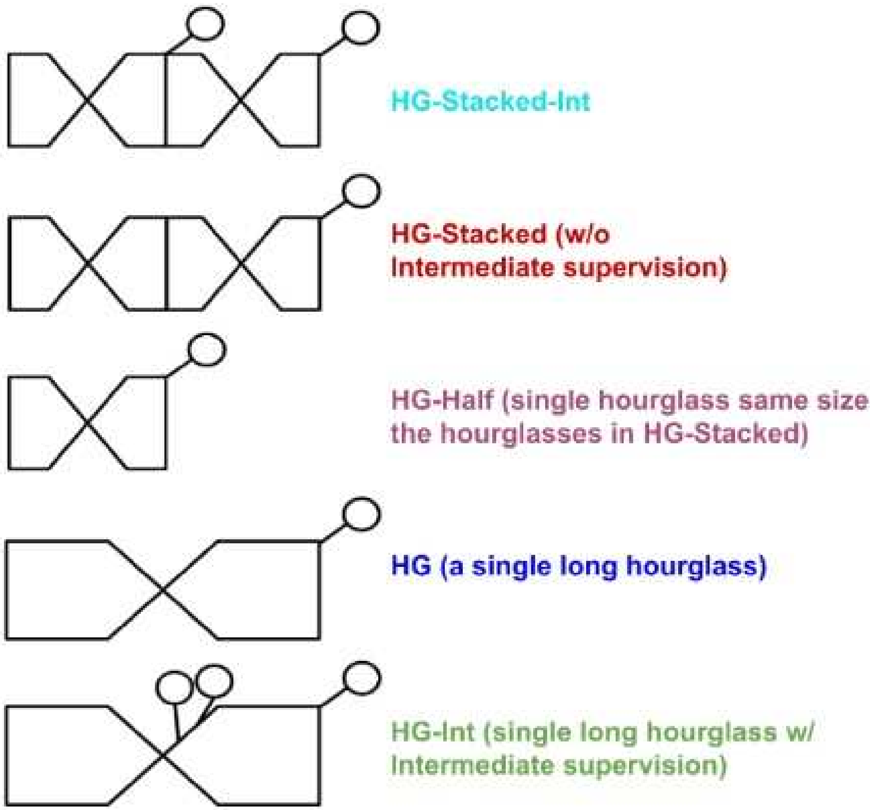
Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

Stacked Hourglass

- Ablation





Newell. Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

Stacked Hourglass

- Results

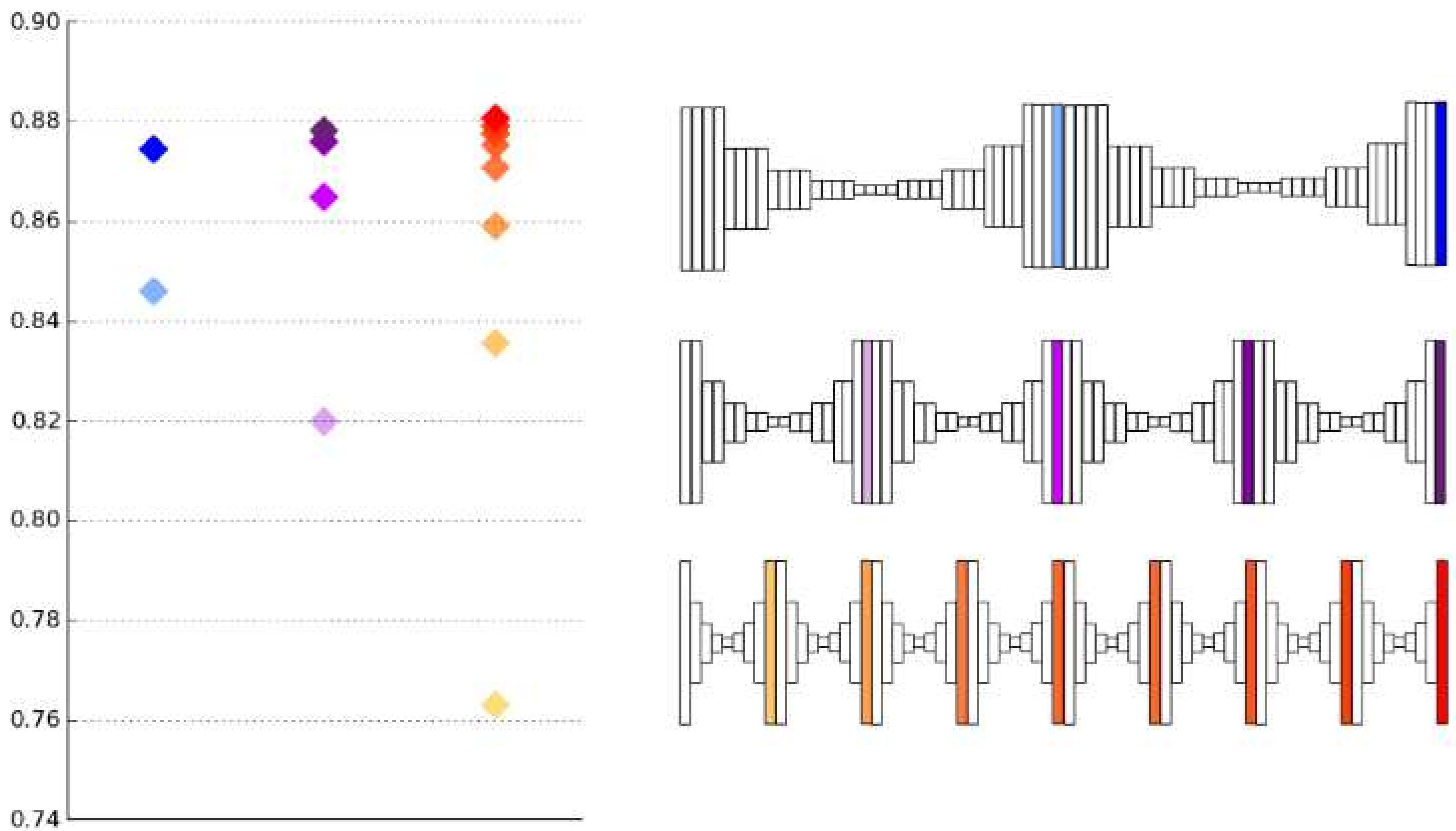
Newell. Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

Semantic segmentation

- Eclectic approach: Encoder-decoder + skip connections

Stacked Hourglass

- Results



Newell. Stacked Hourglass Networks for Human Pose Estimation, ECCV 2016

Object localization and detection

- Problem statement

Object localization

Classify the single object in the image and locate it with a bounding box.

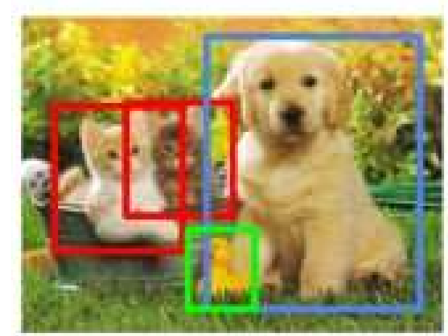
Object detection

Detect multiple objects in the image, locate each of them with a bounding box and attach to each bounding box an object label.

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

Object localization

- Multi-task solution (classification + regression)